University of Tennessee, Knoxville

TRACE: Tennessee Research and Creative Exchange

Doctoral Dissertations

Graduate School

5-2019

Network and multi-scale signal analysis for the integration of large omic datasets: applications in *Populus trichocarpa*

Deborah Ann Weighill
University of Tennessee, dweighil@vols.utk.edu

Follow this and additional works at: https://trace.tennessee.edu/utk_graddiss

Recommended Citation

Weighill, Deborah Ann, "Network and multi-scale signal analysis for the integration of large omic datasets: applications in *Populus trichocarpa*. " PhD diss., University of Tennessee, 2019.
https://trace.tennessee.edu/utk_graddiss/5380

This Dissertation is brought to you for free and open access by the Graduate School at TRACE: Tennessee Research and Creative Exchange. It has been accepted for inclusion in Doctoral Dissertations by an authorized administrator of TRACE: Tennessee Research and Creative Exchange. For more information, please contact trace@utk.edu.

To the Graduate Council:

I am submitting herewith a dissertation written by Deborah Ann Weighill entitled "Network and multi-scale signal analysis for the integration of large omic datasets: applications in *Populus trichocarpa*." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Energy Science and Engineering.

Daniel Jacobson, Major Professor

We have read this dissertation and recommend its acceptance:

Gerald Tuskan, Timothy Tschaplinkski, Wellington Muchero

Accepted for the Council:

Dixie L. Thompson

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

Network and multi-scale signal analysis for the integration of large omic datasets: applications in *Populus trichocarpa*

A Dissertation Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Deborah Ann Weighill

May 2019

© by Deborah Ann Weighill, 2019
All Rights Reserved.

For my inspirational, kind father, Greg Grobler, who got me into science.

Acknowledgments

I would like to express my deep gratitude to Dr. Daniel Jacobson for his support and guidance during my PhD studies. I would like to thank Dr. Gerald Tuskan, Dr. Timothy Tschaplinski and Dr. Wellington Muchero for serving on my doctoral committee and providing valuable input and collaboration. A special, affectionate thanks to members of the Jacobson CompBio Team, especially David “Kainerd” Kainer, Piet Jones, Armin Geiger, Anna “SeaFurchin” Furches, Ben Garcia, Jarad Streich, Ashley Cliff and Jonathon Romero for input, support, advice and friendship.

This research was funded 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. This research was also supported by the Plant-Microbe Interfaces Scientific Focus Area (http://pmi.ornl.gov) in the Genomic Science Program, the Office of Biological and Environmental Research (BER) in the U.S. Department of Energy Office of Science, and by the Department of Energy, Laboratory Directed Research and Development funding (7758), at the Oak Ridge National Laboratory. Oak Ridge National Laboratory is managed by UT-Battelle, LLC, for the US DOE under contract DE-AC05-00OR22725. An award of computer time was provided by the INCITE program. This research used resources of the Oak Ridge Leadership Computing Facility (OLCF) and 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. Support for DOI 10.13139/OLCF/1411410 dataset is provided by the U.S. Department of Energy, project BIF102 under Contract DE-AC05-00OR22725. Project BIF102 used resources of the Oak Ridge Leadership Computing Facility 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.

Abstract

Poplar species are promising sources of cellulosic biomass for biofuels because of their fast growth rate, high cellulose content and moderate lignin content. There is an increasing movement on integrating multiple layers of 'omics data in a systems biology approach to understand gene-phenotype relationships and assist in plant breeding programs. This dissertation involves the use of network and signal processing techniques for the combined analysis of these various data types, for the goals of (1) increasing fundamental knowledge of *P. trichocarpa* and (2) facilitating the generation of hypotheses about target genes and phenotypes of interest. A data integration "Lines of Evidence" method is presented for the identification and prioritization of target genes involved in functions of interest. A new post-GWAS method, Pleiotropy Decomposition, is presented, which extracts pleiotropic relationships between genes and phenotypes from GWAS results, allowing for identification of genes with signatures favorable to genome editing. Continuous wavelet transform signal processing analysis is applied in the characterization of genome distributions of various features (including variant density, gene density, and methylation profiles) in order to identify chromosome structures such as the centromere. This resulted in the approximate centromere locations on all *P. trichocarpa* chromosomes, which had previously not been adequately reported in the scientific literature. Discrete wavelet transform signal processing followed by correlation analysis was applied to genomic features from various data types including transposable element density, methylation density, SNP density, gene density, centromere position and putative ancestral centromere position. Subsequent correlation analysis of the resulting wavelet coefficients identified scale-specific relationships between these genomic features, and provide insights into the evolution of the genome structure of *P. trichocarpa*. These methods have provided strategies to both increase fundamental knowledge about the *P. trichocarpa* system, as well as to identify new target genes related to biofuels targets. We intend that these approaches will ultimately be used in the designing of better plants for more efficient and sustainable production of bioenergy.

Table of Contents

List of Tables

List of Figures

Attachments

Lignin-related metabololites (table_S2.3.xlsx)

LOE Scores and annotations of genes for high LOE genes (table_S2.4.xlsx)

Annotations of MPA genes (table_S3.1.xlsx)

GO enrichment results (table_S3.2.xlsx)

Annotation of chaperone-related MPA genes (table_S3.3.xlsx)

Bingo results (Supplementary_File_S3.1.cys)

PlantCyc metabolic pathway map for *P. trichocarpa* (Supplementary_File_S3.2.pdf)

Number of SNPs in *P.trichocarpa* histone genes (table_S4.2.xlsx)

Genes co-expressing with Potri.014G096400 (table_S4.3.xlsx)

Chapter 1

Introduction

This chapter has been submitted for publication in *Frontiers in Genetics* and contains contributions from other authors:

Deborah Weighill, Timothy Tschaplinski, Gerald Tuskan and Daniel Jacobson

Author Contributions: DW and DJ planned the manuscript. DW performed the research and wrote the manuscript. DJ, TT and GT provided editorial feedback.

Abstract

Populus trichocarpa is an important biofuels feedstock which has been the target of extensive research, and is emerging as model organism for plants, especially woody perennials. This research has generated several large ’omics datasets. However, only few studies in poplar have attempted to integrate various data types. This review will summarize various ’omics data layers, focusing on their application in poplar species. Subsequently, network and signal processing techniques for the integration and analysis of these data types will be discussed, with particular reference to examples in poplar.

1.1 Introduction

Poplar species are promising sources of cellulosic biomass for biofuels because of their fast growth rate, high cellulose content and moderate lignin content [1]. Ragauskas *et. al* [2] outline areas of research needed “to increase the impact, efficiency, and sustainability of biorefinery facilities” [2], such as research into modifying plants to enhance favorable traits, including altered cell wall structure leading to increased sugar release, as well as resilience to biotic and abiotic stresses. One particular research target in poplar is the decrease/alteration of the lignin content of cell walls.

There is an increasing movement on integrating multiple layers of ’omics data in a systems biology approach to understand gene-phenotype relationships and assist in plant breeding programs (see Ingvarsson *et al.* (2016) [3], Weckwerth (2011) [4] and Valledor *et al.* (2018) [5] for reviews).

This chapter will review different sources of 'omics data layers with particular reference to *P. trichocarpa* or other poplar species where studies are available. Subsequently, we will review network and signal processing approaches to representing, analysing and integrating multiple 'omics data layers, again providing examples in poplar species were possible. Lastly we will conclude by forming the aims of this dissertation.

1.2 Sources of 'Omics Data Layers

1.2.1 Genome and Annotation

The genome sequence of Black Cottonwood, *Populus trichocarpa* (Torr. & Gray) was released in 2006 [6]. This genome, a single female genome "Nisqually-1", was the first tree to have its complete genome sequenced, and it became a model system for studies on woody perennial plants [7, 8]. The *P. trichocarpa* genome consists of 19 chromosomes, with chromosome 19 found to be evolving into a sex chromosome [9]. Analysis of homologous regions of the genome showed evidence for several genome duplication events; the most recent being the salicoid duplication event which is contained within the family *Salicaceae*, the next termed the *Eurosid* duplication shared among *Eurosids*, and an ancient duplication event [6].

Since initial sequencing, the genome assembly has gone through several revisions and is now in its 4th version. Furthermore, a genome wide association study (GWAS) population of ∼1,000 natural accessions from the U.S and Canada were propagated in multiple common gardens and resequenced, providing a rich resource for studies of the variation in natural *P. trichocarpa* populations as well as GWAS studies [10, 11, 12].

The genome sequence is available on Phytozome [13], and the genome along with gene and functional annotation such as Gene Ontology (GO) terms and PFams can be viewed and interacted with using the JBrowse [14] plugin on Phytozome.

1.2.2 Gene Expression (Transcriptomics)

Transcriptomic analysis involves the measuring of the expression levels of transcripts within a sample. Various study designs have been implemented in *P. trichocarpa* to investigate a variety of properties of the cellular system. Several studies have focused on the response of the poplar transcriptome, or a subset of the transcriptome, to drought stress. The study by Shuai *et al.* (2013) used RNA-Seq to identify microRNAs responsive to drought stress[15], and subsequently Shuai *et al.* (2014) performed RNA-Seq on control and drought leaf samples of *P. trichocarpa* to identify long intergenic non-coding RNAs (lincRNAs) which were responsive to drought stress [16]. Tang *et al.* (2015) used RNA-Seq to identify genes differentially expressed between well-watered and water-limited samples, and several differentially expressed genes and functions were identified. Genes related to energy metabolism and growth (cell division and tissue expansion) were significantly downregulated, and a particular gene previously found to improve drought and salt tolerance in several plants was significantly upregulated [17]. Another transcriptomic drought study used Affymetric microarrays for expression measurements of *Populus tremula* x *Populus alba* roots for 6 time points under drought stress. Differential expression and network analysis identified two interesting genes (PtaJAZ3 and PtaRAP2.6) which, when overexpressed under drought conditions, increased root growth [18].

Other transcriptomic studies in poplar have focused on variation in gene expression across tissues or across a population. In the study by Quesada *et al.* (2008), gene expression levels in *P. trichocarpa* were measured across 5 different tissues (roots, young leaves, mature leaves, nodes and internodes) using NimbleGen microarrays [19]. Genes with tissue-specific gene expression were identified, with stem samples having the highest number of tissue-specific genes. GO enrichment was used to determine the enriched functions of organ-specific genes. The expression of *P. trichocarpa* genes across organs was also compared to the expression of their *Arabidopsis thaliana* orthologs across equivalent tissues, and the authors concluded that, while there were some similarities between expression patterns across these two species, significant diversification in gene expression regulation has occurred between orthologs. Shi *et al.* (2009) used quantitative real-time PCR (qPCR) to determine the expression level of 95 genes in the phenylpropanoid pathway in xylem, leaf, shoot and phloem tissues, in order to determine the abundance and tissue specificity of genes potentially involved in monolignol biosyntehsis [20]. Bao *et al.* (2013) performed RNA-Seq of xylem tissue from 20 *P. trichocarpa* individuals from different populations, identified a set of sets expressed in xylem across all individuals, and found several instances of alternative splicing, particularly in cell wall-related genes and that these alternative splicing events differed significantly across individuals [21].

An increasingly common study design is the construction of a gene expression atlas for a species, which involves determining the expression level of every gene in the genome in various different tissues and/or conditions. Gene expression atlas studies have been performed in various plant species (see for example Table 1.1), and several expression atlas datasets are available on Phytozome.

The *P. trichocarpa* RNA-Seq gene expression atlas (Sreedasyam et al., unpublished) consists of genome-wide gene expression measurements across several different samples of tissue and condition combinations, including root, root tip, stem, node, internode, bud, leaf and flower tissues. Root and stem tissues included several samples varied by nitrogen source. Bud, leaf, male and female flowers included several samples of different stages of maturity. Gene expression values for 40 of these samples are currently publically available in PhytoMine on the Phytozome web interface [13]. To our knowledge, this is the largest RNA-Seq expression study performed in poplar.

Table 1.1: Examples of gene expression atlas studies in plants.

Species	**Samples**	**Method**
Arabidopsis thaliana [22]	79 samples from various tissues and developmental stages	Affymetrix GeneChip
Sorghum bicolor [23]	47 combinations of tissues (roots, leaves, stems, panicles) and developmental stages (juvenile, vegetative, reproductive)	RNASeq
Glycine max [24]	14 tissues from different developmental stages	RNASeq
Lotus japonicus [25]	237 samples of 8 tissues across various conditions	Affymetrix GeneChip
Medicago truncatula [26]	18 samples from tissues across different developmental stages	Affymetrix GeneChip
Barley [27]	15 tissues identified from eight developmental stages	Affymetrix GeneChip
Rice [28]	31 tissues spanning life cycle of rice plant for 2 rice varieties, 8 samples from stages in the tissue culture process	Affymetrix GeneChip
Panicum virgatum L (Switchgrass) [29]	Tissues (roots, shoots, and panicle) and developmental stages (leaf development, stem elongation and reproduction)	ESTs
Vitis vinifera [30]	54 samples from tissues spanning different developmental stages	NimbleGen microarray and RNASeq

1.2.3 Metabolomics

Metabolomics studies involve measuring the quantities of metabolites within a sample. While targeted metabolomics studies aim to only measure and identify a select few metabolites within a sample (for instance using standards), untargeted metabolomics involves the measuring as many metabolites as possible within a sample [31]. Identification of metabolites in untargeted metabolomics studies is much harder than that of targeted metabolomics studies. While the candidate identities of many metabolite peaks can be determined though database matching or manual inspection of mass spectra with the necessary expertise, many metabolites will remain unidentified or partially identified.

Several targeted and untargeted metabolomics studies have been performed in poplar. In a study by Morreel *et al* (2006), metabolite levels of 15 flavonoids were measured using high performance liquid chromatography (HPCL), and subsequntly mQTL (metabolite quantitative trait loci) based on amplified fragment length polymorphisms (AFLPs) was used to identify potential genes involved in rate limiting steps of flavonoid biosynthesis [32]. Kaling *et al* (2015) performed untargetted metabolomics on UV-B treated vs. control *P. alba x P. tremula* plants using Fourier transform ion cyclotron resonance mass spectrometry (FT-ICR-MS). This allowed for the investigation of the effect of UV radiation on the metabolome[33]. Tuskan *et al.* (2012) performed gas chromatography-mass spectrometry (GC-MS) analysis of 16 individual trees in *Populus deltoides* and *Populus nigra*, and showed gender-specific accumulations of metabolites in floral buds [34]. In Hamanishi *et al.* (2015), transcriptomic and metabolomic data of six *Populus balsamifera* were collected using Affymetrix microarrays and gas chromatography-mass spectrometry (GC/MS), respectively, to investigate the response of the metabolome and transcriptome to drought stress [35]. Tschaplinski *et al* (2014) used GC/MS-based metabolomics on samples of *P. trichocarpa* and *P. deltoides* roots colonized with *Laccaria bicolor* as well as control samples to investigate the different metabolic responses to colonization [36]. One interesting result was that increased levels of defense-related compounds were found in the incompatible host, *P. deltoides*, whereas some defense compounds were significantly lower in the compatible host, *P. trichocarpa*. An recent study by Veach *et al.* (2018) investigated the effects on the metabolome of *P. deltoides* when downregulating *PdKOR1*, a glycosyl hydrolase gene involved in cellulose biosynthesis [37]. GC/MS analysis of root tissue from *PdKOR1* RNAi lines vs. control lines showed that caffeic acid derivatives, metabolites involved in fatty acid metabolism as well as salicylates and flavonoids were upregulated in RNAi lines when compared to control lines [37].

A genome-wide association study using SNPs from 917 *P. trichocarpa* accessions as well as GC/MS-based metabolomics and RNA-Seq-based gene expression measurement identified hydroxycinnamoyl- CoA:shikimate hydroxycinnamoyl transferase 2 (PtHCT2) as a gene which is significantly associated with the levels of 3-O-caffeoylquinic acid, and also identified transcription factors which regulate this gene [38].

A comprehensive study of a GWAS analysis of metabolomics phenotypes in a population of $\sim$1,000 *P. trichocarpa* genotypes is currently being performed by Timothy Tschaplinski and his team at the Oak Ridge National Laboratory (Timothy Tschaplinski, personal communication).

1.2.4 Variant Data

Different individuals in a population can accumulate different kinds of variation in their genome, such as Single Nucleotide Polymorphisms (SNPs) involving a nucleotide change at a single position, insertions/deletions of a single nucleotide or larger pieces of DNA, copy number variations (CNVs) of DNA segments or translocations (the movement of a section DNA from one location to another) [39].

There are two major approaches to calling SNPs in a given sample in a relatively high-throughput manner, namely a genotyping SNP array and SNP calling from Next Generation Sequencing (NGS) data. A genotyping SNP array involves hybridizing extracted DNA to an array containing probes with known SNPs [40], and is thus limited by the SNPs chosen to appear on the array. For example, the *P. trichocarpa* genotyping array is based on 34,131 SNPs located near/within around 3,500 selected candidate genes [41]. SNP calling through NGS involves whole genome sequencing of all individuals, aligning of all reads to a common reference genome, and then calling variants [42] using software such as GATK [43]. The advantage of SNP calling from NGS data is that one is not limited by the set of SNPs available on an array. A larger number of SNPs can be detected and the discovery of new SNPs is possible. SNP genotyping arrays are also not able to detect other classes of genome variants such as translocations and inversions [40].

A population of 1,100 natural *P. trichocarpa* accessions have been clonally propagated in 4 common gardens [10] and resequenced in order to provide NGS data for SNP calling. Several studies have been published making use of SNPs called across parts of this population [12, 44, 11].

Slavov *et al.* (2012) performed a study involving SNPs called from resequenced genomes of 16 of the genotypes within this *P. trichocarpa* population. PCA analysis of SNP genotypes revealed clear separation based on the geographic origin of the genotypes, and linkage disequilibrium was reported to decay to $r^2 \leq 0.2$ within 3-6kb. It is important to note that this is based on the resequencing of only 16 genotypes (LD calculations of the whole population are presented in later chapters, and find that the decay is in fact faster).

A set of $\sim$28 million bi-allelic SNPs called across 882 genotypes from this population have been publicly released and are available from online from DOI 10.13139/OLCF/1411410.

1.2.5 Genome Wide Association Studies

Phenotypes are often complex traits, in that they are influenced or controlled by a number of genes [45]. Genome Wide Association Studies (GWAS) attempt to associate the presence/absence of SNPs with these complex traits [45, 46]. This involves genotyping a large sample of individuals of a population, measuring phenotypes across all of these individuals and statistically determining the association between the presence/absence of the genotyped markers or SNPs and the phenotypes across the population [47]. A general concern when conducting GWAS studies is that individuals within a population that are genetically related can share both causal alleles, which cause the phenotype [46], and non-causal alleles [47]. These causal and non-causal alleles could be located nearby to each other on the chromosome and could thus be in linkage disequilibrium (alleles which are correlated across a population and thus co-inherited [48]). This linkage disequilibrium (LD) between causal and non-causal alleles across related individuals could result in non-causal alleles being correlated with a phenotype when they have no actual effect on the phenotype.

GWAS analyses generally require that the individuals in the population are unrelated. However, some level of population structure due to some shared ancestures, which can cause sperious associations between genotype and phenotype, and accounting for population structure is this important in order to remove variance that is due solely to the relatedness of individuals (for a useful review, see Astle and Balding (2009) [49]). It is thus important to account for population structure in association models. However, there is the possibility of masking true associations which happen to correlate with population structure because they are local adaptations of clades to local environments.

EMMAX [50] is one particular GWAS method which attempts to correct for the effect of individual relatedness within the population. It is a faster version of the EMMA method [51]. GWAS methods such as EMMAX model the relationship between measured phenotypes and SNPs as a linear model:

$$y_i = \beta_0 + \sum_{k=1}^{M} \beta_k X_{ik} + \epsilon_i \tag{1.1}$$

where y_i is a measured phenotype for individual i, β_k is the effect of SNP k on the phenotype, X is a matrix of fixed effects (SNPs) in which X_{ik} is the minor allele count of SNP k in individual i, and ϵ_i represents environmental variation on the phenotype y_i [50]. The aim is to determine which of the β_k are significantly different from zero, thus identifying which SNPs have a significant effect on the phenotype [50]. EMMAX accounts for sample structure by calculating a kinship matrix K which contains pairwise genetic similarities of the of the individuals under consideration. A variance component model is used, partitioning the phenotypic variance into variance due to environmental factors σ_e^2, and variance due to the additive effect of genetic factors σ_a^2 [50]. This variance component model includes the kinship matrix, modeling the variance-covariance structure of the phenotype in terms of the genetic similarity of pairs of individuals defined in the kinship matrix [50]:

$$\text{Var}(Y) = \sigma_a^2 K + \sigma_e^2 I \tag{1.2}$$

where Var(Y) is the variance-co-variance structure of the phenotype and I is the identity matrix. The β_k are then estimated using Generalized Least Squares and an F-test is used to determine which of these β_k are statistically different from zero [50]. Each SNP k corresponding with a β_k statistically different from zero thus potentially affects the phenotype. Thus, for a given measured phenotype, EMMAX produces a list of all SNPs and their respective p-values for their association with the phenotype. A p-value threshold can then be applied to determine which of the associations are significant.

Performing a GWAS involves testing multiple hypotheses, each asking "is SNP k associated with the phenotype p?" for each SNP in the dataset. When testing multiple hypotheses, or a so-called family of m hypotheses, the quantity called the Family-wise Error Rate (FWER) becomes inflated [52]. The FWER is defined as the probability that at least one null hypothesis was rejected when it should not have been, or, the probability of achieving at least one false positive. When a statistical test is performed and a p-value is generated, the p-value represents the Type-1 error rate (or false-positive rate), which is the probability that the null hypothesis was incorrectly rejected [52]. Let the p-value threshold chosen be α. Then, for each true null hypothesis, the probability that it was incorrectly rejected is α. Given that a null hypothesis is true, the probability it was not rejected is thus $1 - \alpha$. If we assume that all m null hypotheses are true, the probability that all m null hypotheses were not rejected (i.e. the probability of obtaining no false positives) is $(1 - \alpha)^m$. Therefore, given that all null hypotheses are true, the probability of obtaining at least one false positive (also known as the FWER) is [52]:

$$\text{FWER} = 1 - (1 - \alpha)^m \tag{1.3}$$

As can be seen from Equation 1.3, the FWER increases with the number of hypotheses tested. The probability of obtaining false positives thus increases with the number of hypothesis tests performed. Methods for multiple hypothesis correction attempt to control this FWER.

Let $H_1, H_2 ... H_m$ be a family of m hypotheses and let $P_1, P_2 ... P_m$ be their respective p-values. Bonferonni Correction [53] is a simple method which rejects null hypothesis H_i if [53]:

$$P_i \leq \frac{\alpha}{m} \tag{1.4}$$

This has been proven to control the FWER, ensuring that FWER $\leq \alpha$.

An adaptation to this method known is as Sequential Bonferonni Correction or Holm-Bonferonni Correction [54]. This method orders the hypotheses such that $P_1 \leq P_2 \leq ... \leq P_m$. The index k is then determined such that k is the largest index for which the following holds [54]:

$$P_k \leq \frac{\alpha}{m + 1 - k} \tag{1.5}$$

Hypotheses $H_1, H_2 ... H_k$ are then rejected and hypotheses $H_{k+1}, H_{k+2} ... H_m$ are not rejected.

Another type of multiple hypothesis correction attempts to control the False Discovery Rate (FDR), which is defined as the proportion of rejected null hypotheses which were incorrectly rejected, or, the proportion of Type-1 errors made [55]. This is performed by ordering p-values in a similar fashion to Holm-Bonferonni Correction. The index k is then determined such that k is the largest index for which the following holds [55]:

$$P_k \leq \frac{k\alpha}{m} \tag{1.6}$$

Hypotheses $H_1, H_2 ... H_k$ are then rejected and hypotheses $H_{k+1}, H_{k+2} ... H_m$ are not rejected. This procedure ensures that the FDR is below α.

Several studies involving GWAS analyses in *P. trichocarpa* have been published. Mcknown *et al.* (2014) genotyped 448 individuals from the *P. trichocarpa* GWAS population using a SNP array containing ~34,000 SNPs, and performed GWAS on 40 different traits measured in the population [44]. These traits included biomass phenotypes such as height, volume, and height:diameter ratio, ecophysiological traits such as leaf shape, chlorophyll content and carbon:nitrogen ratio ad phenology traits such as bud set, growth period and leaf drop. A set of 1118 significant GWAS associations were identified involving 410 unique SNPs, 78% of which occurred in non-coding regions and 28% occurred in coding regions. This resulted in 275 genes having significant trait associations, many of which were transcription factors or regulators of some kind. [44]. A subset of 42 of the 275 genes exhibited multiple GWAS associations with traits in different trait categories, exhibiting potential pleiotropy.

Evans *et al.* (2014) performed whole genome sequencing of 544 individuals from the *P. trichocarpa* GWAS population [12], and subsequent variant calling identified 17,902,740 SNPs. They found that nucleotide diversity was twice as high in intergenic space than in genic space, and that diversity was even lower in coding space, and that a large proportion of the SNPs had a minor allele frequency (MAF) leq 0.01 and were thus considered rare alleles. Metrics of natural selection such as F_{ST} were used to identify candidate regions under strong selection, and suggest that this could be driven by climate.

Tuskan *et al.* (2018) tested callus induction in 280 genotypes from within the *P. trichocarpa* GWAS population, and performed a GWAS analysis to identify SNPs potentially affecting callus formation [56]. Eight genes potentially associated with callus formation were identified. Combining GWAS results with co-expression information allowed for a putative regulatory network for callus formation to be constructed [56].

In a recent study by Liu *et al.* (2018), 64 individuals from a full-sib family from a cross between *Populus deltoides* and *Populus euramericana*, were genotyped using real-time PCR. Phenotypes used were stem heights and diameters over 24 years. Both a standard GWAS and distance correlation sure independence screening (DC-SIS) association tests are performed. DC-SIS is an association method which allows for a multi-dimensional phenotype (diameter measurements over time) as opposed to a single phenotype measurement [57].

1.2.6 DNA Methylation

Epigenetics involves the the study of additions of chemical groups to chromatin (either the DNA or histones) which do not change the underlying DNA sequence. These modifications consist of histone methylation [58], histone acetylation [59] and DNA methylation [60]. Histone methylation occurs on lysine and argenine residues in histones, and can have a silencing or activating effect on gene expression, depending on which lysine residue is methylated [58]. Histone acetylation involes the addition of an acetyl group to the ϵ amino group of lysine residues in the N-terminal tails of histones which protrude from the histone octamer complex [59]. While histones are usually positively charged and DNA is negatively charged, acetylation can neutralize the positive charge of the histones, resulting in a weaker association between the DNA and the histone complex. This can allow for greater access for transcription factors to the DNA, and can thus impact gene expression [59]. DNA methylation involves the addition of a methyl group to a cytosine residues [61]. This is known to have a gene-silencing affect. DNA methylation in plants occurs mostly in repetitive DNA and transposable elements. This is thought to be a protective mechanism to silence transposons. DNA methylation is also found within the transcribed regions of genes in plants [62]. This gene body methylation does not have a silencing effect like promotor methylation does, but appears to lead to stable gene expression across many tissues [62, 63]. Epigenetic modifications can be inherited from parents, or occur as a result of a stress response [64, 65].

Two major whole genome sequencing-based approaches for determining DNA methylation across a genome are Methyl-DNA immunoprecipitation (MeDIP) followed by sequencing (MeDIP-Seq) [66] or treatment of DNA with bisulfite followed by sequencing [67, 68]. MeDIP-Seq involves shearing of DNA into small fragments of 300-600bp and subsequent immunoprecipitaion of methylated DNA using an antibody raised against 5-methylcytidine. The resulting immunoprecipitated fragments are sequenced, and mapping of the reads to the reference genome reveals the regions of the genome which contain methylated cytosines. It is important to note that the resolution of the methylation results cannot exceed the fragment size to which the DNA was sheared. In bisulphite sequencing, DNA is pre-treated with sodium bisulfite which converted un-methylated cytosine residues to uracil residues, while methylated cytosine residues remain unchanged. Subsequent sequencing provides single-base resolution of methylated cytosines [67, 68]. See published papers of Laird *et al.* (2010) [69] and Bock *et al* (2012) [70] for useful reviews on DNA methylation analysis.

Vining *et al.* (2012) investigated DNA cytosine methylation in seven different tissues in *P. trichocarpa*, including bud, male catkin, female catkin, leaf, root, xylem and phloem [71]. DNA methylation was determined using MeDIP-Seq followed by mapping of reads to the *P. trichocarpa* reference genome. Reads mapped most frequently to intergenic regions and repeat sequences, although promotor methylation and gene body methylation was observed. Variation in methylation across tissues was observed at certain chromosomal locations . A surprising result was that when looking at gene expression of methylated genes, gene body methylation apeared to be a stronger repressor of transcription than promotor methylation [71]. Slavov *et al.* (2012) made use of the methylation data generated by Vining *et al.* (2012) in investigating the correlates of recombination in the *P. trichocarpa* genome, and found that DNA within recombination hotspots was significantly less methylated than non-hotspots [11].

A follow-up study by Vining *et al.* (2013) used MeDIP-Seq to examine methylation levels in another three tissues, focusing on regeneration and dedifferentiation tissue types, namely internode stem from propagated explants, callus and internodes from regenerated plants [72]. The MeDIP-Seq reads for the 10 different *P. trichocarpa* tissues from these two studies have been mapped to the version 3 genome assembly and are available on Phytozome [13].

A stress response methylome study was performed in *P. trichocarpa* in which DNA methylation was measured in drought stress and control plants using bisulphite sequencing [73]. The number of methylated cytosines increased significantly under drought stress, and the genes differentially methylated in drought stress vs control plants were enriched for regulatory Gene Ontology (GO) terms. This study also performed the first investigation of alternative splicing in *P. trichocarpa* and identified multiple forms of alternative splicing. An interesting finding was also that all fusion genes identified were methylated [73].

Lafon-Placette *et al.* (2013) investigated the component of the *P. trichocarpa* methylome in open chromatin by isolating chromatin sensitive to DNase I, and performing MeDIP-Seq on the resulting DNA [74]. Extensive gene body methylation was found

The two studies by Vining *et al* (2012 and 2013) provide the best available methylation dataset to use as an association network layer as it covers the broadest range of sample types.

1.2.7 ATAC-Seq

The assay of transposase-accessible chromatin (ATAC-Seq) uses a transposase to insert sequencing adaptors into accessible regions of chromatin (i.e. the areas in between nucleosomes) [75, 76]. The resulting fragments are then PCR-amplifed and sequenced. This results in nucleotide-resolution of open chromatin. There were challenges in applying this method to plant cells due to contaminating DNA from chloroplasts and mitochondria. This is because these elements of the genome are very accessible to the transposase and thus lower the efficiency of the technique. Lu *et al.* (2016) developed a technique, fluorescence-activated nuclei sorting (FANS)-ATAC-Seq which involves sorting of nuclei using flow cytometry prior to ATAC-Seq analysis [77]. Bajic *et al.* (2018) describe protocols for the isolation of plant nuclei from different cell types for further analysis using ATAC-Seq [78]. A recent study by Maher *et al.* (2018) applied ATAC-Seq to *Arabidopsis thaliana*, *Medicago truncatula*, *Solanum lycopersicum* (tomato), and *Oryza sativa* (rice). An interesting finding was that in all four species, most open chromatin sites were in non-transcribed regions [79].

ATAC-Seq is a relatively new technology, and, to date, no study has been published on the application of ATAC-Seq in *P. trichocarpa*.

1.2.8 DAP-Seq

DNA affinity purification sequencing (DAP-seq) is a technique used to determine transcription factor binding sites in DNA developed by O'Malley *et al.* (2016)[80]. This technique involves coupling a particular transcription factor of interest to affinity beads. Fragmented genomic DNA is eluted over the beads, retaining only DNA fragments which bind to the transcription factor. Subsequently, the retained fragments are sequenced. The first study describing this technique demonstrated its use in identifying the *Arabidopsis* "cistrome" - the binding location/motifs of 1,812 transcription factors [80]. The DAP-seq protocol was published by Bartlett *et al.* (2017) [81].

To date, no DAP-seq study has been performed in *P. trichocarpa*. This would be an incredibly valuable data layer to investigate the transcription factor regulatory network of *P. trichocarpa*.

1.2.9 Transposable Elements and Repeats

Transposable elements (TEs) are segments of DNA which are mobile, in the sense that they can move from one genomic location to another. Type I elements, or retrotransposons, require an RNA intermediate, and are then reverse-transcribed into the genome at a different location [82, 83]. This is thus a "copy and paste" mechanism. Type II TEs are called DNA transposons, and involve the excision of the DNA TE and subsequent integration elsewhere. This can thus be described as a “cut and paste” mechanism [82, 83]. Many TEs are no longer active because mutations have inhibited their ability to transpose. However, some TEs are silenced by the host. This can include mechanisms such as silencing by RNAi or though DNA/histone methylation [82].

Different TEs show preference for insertion at different locations in the genome, and thus exhibit very different distributions across the genome [84]. TEs have large impacts on genome characteristics and evolution [85]. Firstly, they have a significant impact on genome size, comprising a large part of many plant genomes [84], ranging from 10% of the genome of *Medicago truncatual*, 42% of *P. trichocarpa* and 80% of *Pinus taeda* (loblolly pine) [86]. Unequal homologous recombination can also result from the presence of multiple TEs of a given family. This can cause various genome rearrangements including duplications, inversions, deletions and translocations [84, 87]. TEs which insert into gene regions can cause the gene to become non-functional. In addition, TEs which insert near genes can impact the expression pattern of the genes, especially since some TEs contain regulatory sequences [84, 83]. Application of stress to a organism has been shown to activate TEs, leading to the hypothesis that TEs create variability in the genome which could be useful under times of stress [88].

Since the genome release, several investigations of repeats and transposable elements have been performed in *P. trichocarpa*. Soon after the release of the *P. trichocarpa* genome, Zhou *et al.* (2009) annotated repeat sequences in the genome and made them publically available in a database called RepPop [89]. Cossu *et al.* (2012) identified LTR repeats in *P. trichocarpa* and investigated their distribution across the genome, finding Gypsy LTRs to be enriched in putative centromeric regions [90]. Soon after, Natali *et al.* (2015) surveyed LTR-retrotransposons in an updated version of the *P. trichocarpa* genome [91]. Vining *et al.* (2012) investigated the number of repeats and genes which were methylated vs non-methylated in *P. trichocarpa*, and found that methylated retroelements, LTRs, hAT elemets, Cacta elements and certain LINEs were overrepresented when compared to their un-methylated versions [71]. It was also found that the methylation patterns of TEs differed significantly across tissues [71]. Usai *et al.* performed an investigation into the repetitive DNA content of seven different *Populus* species, including *P. deltoides*, *P. nigra*, *P. tremula*, *P. tremoloides*, *P. balsamifera*, *P. simonii* and *P. trichocarpa* [92]. LTR repeats were the dominant repeat type across all species, although the total repeat content varied from 33.8% in *P. nigra* to 46.5% in *P. tremuloides*.

In a recent study by Mascagni *et al.* (2018), insertion ages of LTR TEs were determined in *P. trichocarpa* by comparing the sequences of the 3' and 5' ends of LTRs. This provides an indication of the time since insertion because at the time of insertion, the 3' and 5' LTRs are identical, and subsequently accumulate mutations independently after insertion [93]. Insertion time was also determined by comparing the sequences of paralogous RTs from the same lineages. The two methods provided conflicting results, with the LTR comparison method suggesting that Gypsy TEs were older than Copia TEs, whereas the RT comparison method did not find a significant difference in the age of these classes [93]. Yi *et al.* (2018) recently published a database (SPTEdb) of transposable elements in *P. trichocarpa*, *P. euphratica* and *Salix suchowensis*. This database provides TE annotation for these organisms using multiple TE identification methods and presents these in a database format as well as a JBrowse interface [94].

1.3 Data Integration

1.3.1 Multi-Omic Studies and Data Integration

The current era has an extensive suite of technologies capable of measuring and characterizing several aspects of a cellular system, such as next generation sequencing technologies for genomics, transcriptomics and epigenomics as well as metabolomics and other phenotypes. An untargeted approach is often favored over a targeted approach as this attempts to capture information about the entire system and understand the organism as a whole. In the review by Weckwerth (2011), it is highlighted that the next step in understanding complex systems will involve the integration of these different data layers [4]. An important and challending task which data integration can help solve is the identification of new candidate genes involved in complex phenotypes[95, 5], which can then be validated using genetic/molecular biology tools. It is particularly difficult to generate hypotheses that suggest the mechanism of a gene's affect on a particular phenotype. Prioritizing candidate genes and hypothesizing the mechanism of the effect requires multiple data types, such as gene-phenotype associations, expression/co-expression information, knowledge from literature, annotation information, protein-protein/protein-DNA interactions, epigenetic modifications, to name a few [95]. This presents a challenge because of the heterogenous nature of these data types, and the fact that they are often distributed across different databases and represented as different structures [95]. nodes and edges have different types. There is thus an increasing value in databases which integrate various layers of data from various sources [96], for example Knetminer [97, 98], and String [99, 100, 101, 102].

Data integration requires that the various data layers be coerced into a uniform data structure. The data collected from various techniques can each be represented as a matrix/table of samples and variables, as illustrated in the review by Weckwerth (2011) [4]. Once represented as a matrix, there are various data structures/analysis approaches which can be used to integrate and analyze the data. This can range from multivariate analysis such as Orthogonal Projections to Latent Structures (OPLS) [103], to networks [95, 104] and signal processing, such as that seen in the study by Spencer *et al.* (2006) [105].

This section will describe the theory behind different data structure/approaches such as networks and signal processing, discuss examples in which these data structures/methods are used in the analysis of multiple biological data types and show use cases of data integration using these strategies, focusing on examples in poplar where possible.

1.3.2 Networks

Network Theory

Networks are useful mathematical structures which represent a system in terms of its components, and pairwise interactions between the components [106]. The field of Network Theory has its origins in Graph Theory. Intuitively, a graph (or network) is a set of objects (nodes) connected by lines (edges) as shown in Figure 1.1A. Mathematically, a graph G is an ordered pair defined as $G = (V, E)$ where V is a set of nodes and E is a set of edges [107]. Each edge $e_{ij} \in E$ is defined as a set of two nodes:

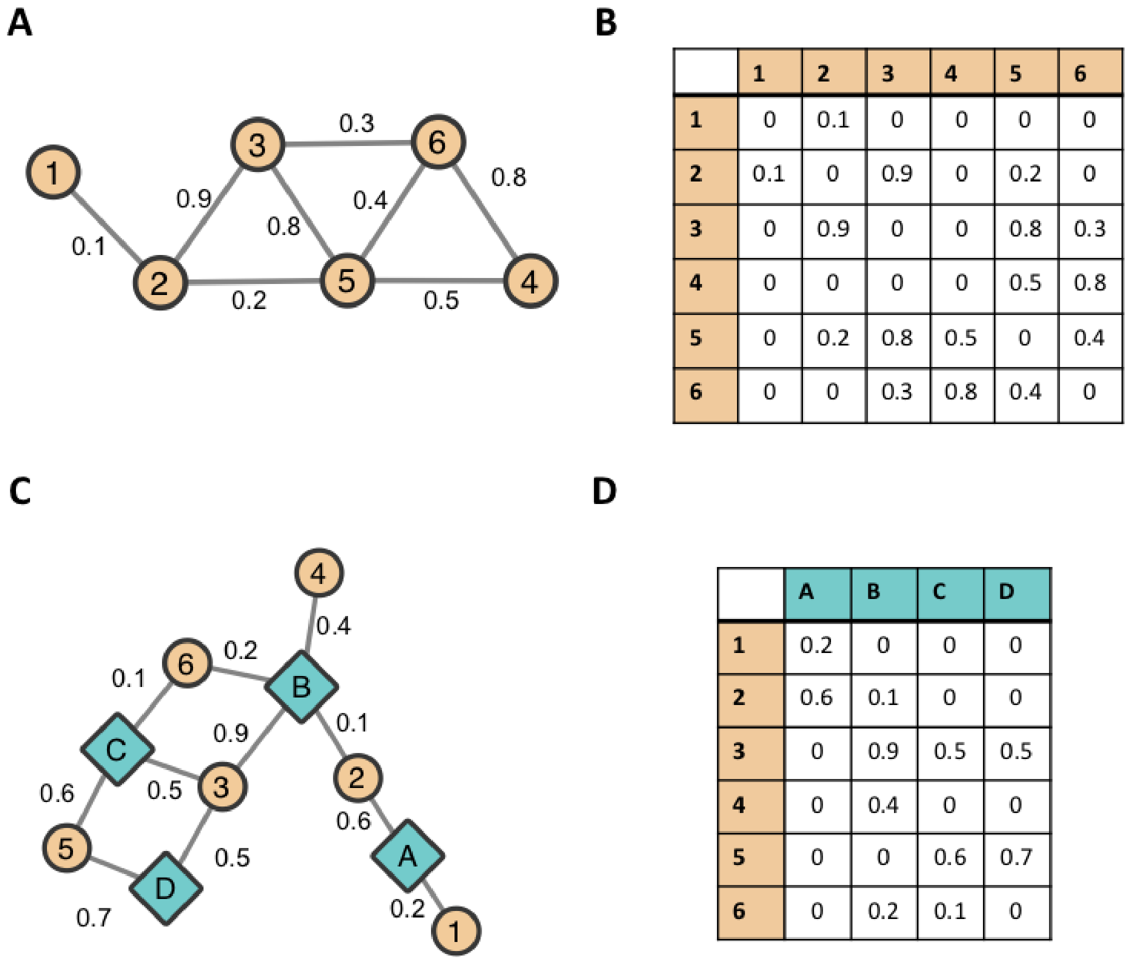
$$e_{ij} = \{i, j\} \tag{1.7}$$

where $i \in V$ and $j \in V$. In biological network applications, nodes represent a biological object of interest and edges will represent associations/interactions/similarities between these biological objects.

A graph can be represented numerically as a matrix, namely an Adjacency Matrix [107]. The Adjacency Matrix A is an $n \times n$ matrix where $n = |V|$, the number of nodes in the network. Each entry a_{ij} in an Adjacency Matrix associated with a graph is defined as [107]:

$$a_{ij} = \begin{cases} 1 & \text{if } \{i, j\} \in E \\ 0 & \text{otherwise.} \end{cases} \tag{1.8}$$

The Adjacency Matrix associated with the small example graph in Figure 1.1A is shown in Figure 1.1B. Each edge e_{ij} in a graph can be assigned a real number weight w_{ij} which represents the strength of the relationship between the two nodes it connects. A weighted graph can be mathematically represented as a Weighted Adjacency Matrix. This matrix



	1	2	3	4	5	6
1	0	0.1	0	0	0	0
2	0.1	0	0.9	0	0.2	0
3	0	0.9	0	0	0.8	0.3
4	0	0	0	0	0.5	0.8
5	0	0.2	0.8	0.5	0	0.4
6	0	0	0.3	0.8	0.4	0

	A	B	C	D
1	0.2	0	0	0
2	0.6	0.1	0	0
3	0	0.9	0.5	0.5
4	0	0.4	0	0
5	0	0	0.6	0.7
6	0	0.2	0.1	0

Figure 1.1: **Example Network.** A small network represented (a) set theoretically, (b) visually and (c) as a matrix.

is constructed in a similar manner to the normal Adjacency Matrix. Each entry a_{ij} of the Weighted Adjacency Matrix is defined as:

$$a_{ij} = \begin{cases} w_{ij} & \text{if } \{i, j\} \in E \\ 0 & \text{otherwise.} \end{cases} \tag{1.9}$$

where w_{ij} is the weight associated with edge e_{ij} [107].

A bipartite graph $G = (V, E)$ is a graph in which the nodes of V can be partitioned into two non-overlapping sets, V_1 and V_2 and each edge $e_{ij} \in E$ is defined as:

$$e_{ij} = \{v_i, v_j\} \tag{1.10}$$

where $v_i \in V_1$ and $v_j \in V_2$ [108]. Intuitively, this means that a bipartite graph (or a bipartite network) consists of two classes of nodes in which nodes of one class can only be connected to nodes of the other class. An example of a bipartite network is shown in Figure 1.1C, and it's matrix representation in Figure 1.1D.

Networks are useful tools for modeling and analyzing complex biological systems by representing biological molecules/components as nodes, (e.g. genes, proteins or metabolites) and representing the relationships/interactions/similarities between them as edges [106]. For example, networks can model co-expression relationships between genes, sequence similarity between genes, physical interactions between proteins or correlations between metabolites. Networks allow for biological datasets to be visualized in an intuitive manner and network visualization packages such as Cytoscape [109] provide an interactive environment for network visualization. However, networks are not simply useful as a visualization tool. Networks provide a data structure which can be computed upon, allowing further analysis to be performed on a dataset represented as a network. Examples of such analysis methods include network-based clustering algorithms such as

Markov Clustering (MCL) [110] and Weighted Gene Co-expression Network Analysis [111] which cluster the nodes of a network into groups based on the topology of the underlying network. Datasets represented as networks are also very easily merged with each other. This feature makes networks a useful tool for combining information from different data sources to create a combined and holistic environment for data interpretation.

GWAS Networks

Network approaches have been applied to GWAS analyses in order to interpret or further analyse the resulting lists of SNPs and p-values. These often involve mapping the resulting SNPs associated with phenotypes to their respective genes, and then projecting these genes into protein-protein interaction networks [112] or co-expression networks [113] in order to identify other putative causal genes, or to form sets or subnetworks of genes putatively affecting the same phenotype [114].

The results of a GWAS study can be viewed as bipartite network $G = (V, E)$ where V can be partitioned into a set of SNPs (V_1), and a set of measured phenotypes (V_2), and E is a set of edges connecting SNP nodes to the phenotype nodes they are associated with. SNPs can be connected to multiple phenotypes, and phenotypes can be connected to multiple SNPs. A toy example of such a network can be seen in Figure 1.1(a). Pink, diamond shaped nodes represent measured phenotypes A-D and blue, circular nodes represent SNPs 1-10. Each edge represents GWAS associations between SNPs and phenotypes. This representation of GWAS results has been used to estimate pleiotropy within a Human-Phenotype-Network, calculated as the average degree of the gene nodes within a gene-phenotype bipartite network [115]. Another example can be seen in a study by Fagny *et al.* (2017) in which the results of an expression quantitative trait loci (eQTL) were represented as a bipartite network, connecting SNPs to genes if the expression level of the gene was significantly associated with the SNP [116].

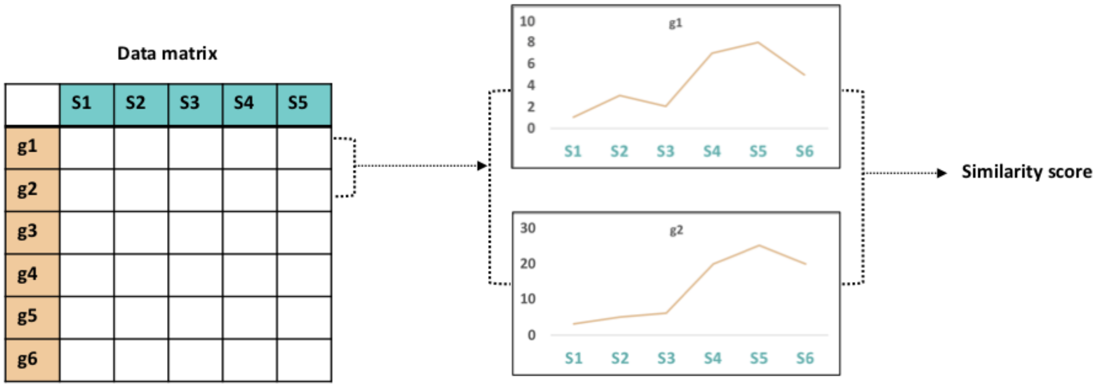


Figure 1.2: **Vector Similarity.** Gene association network comparison involves construction of a data matrix of measurements (e.g. gene expression) for all genes in a genome across various samples. Calculation of the similarity between all pairs of gene vectors results in a similarity score.

Co-expression, Co-methylation, and Correlation Networks

Several of the ’omics data layers discussed in Section 1.2 can be used to construct gene networks, such as gene co-expression networks and gene co-methylation networks. These networks require some quantity, such as gene expression, to be measured for every gene across multiple samples representing different conditions, tissues or perturbations. A common way to construct gene association networks is to calculate the similarity between the profiles of all pairs of genes (Figure 1.2), and then apply a threshold [117, 118]. The choice of similarity metric can have a large impact on the resulting network topology, as shown in a study by Weighill and Jacobson (2017) [118].

Co-expression networks have been used for various applications, including gene function investigations, gene module and regulatory hub gene investigations, as well as comparative co-expression network analysis across different species [117, 119, 120, 121, 122]. Movahedi *et al.* (2012) described an approach for incorporating gene homology information in order to compare gene co-expression modules across plant species to identify clusters which are conserved across species [123]. The overal functional impact of modules of sets of co-expressed genes can be investigated using enrichment of functional ontologies such as Gene Ontology [124] and MapMan [125] (see for example Emamjomeh *et al* (2017) [122]).

Horvath *et al.* (2008) presented a useful set of network topology measures to characterize the structure of a co-expression network (or any network) [126]. These measures ranged from global network measures such as centralization, density and heterogeneity to node-based metrics such as connectivity and the clustering coefficient. Yip and Horvath (2007) also developed a new network measure/transformation called Topological overlap, which calculates the connectedness" of two nodes based on direct connections as well as indirect connections via their neighbors[127]. This provides an extra transformation which can be performed on a similarity network, which considers not only the similarity between expression profiles of two genes in question, but also the expression profiles of their network neighbours, and thus can help address the problem arbitrary thresholds missing important edge connections. This Topological Overlap measure is an integral part of a popular gene co-expression pipeline called Weighted Gene Co-expression Network Analysis (WGCNA) developed by Langfelder and Horvath (2008) [128]. An extension of the Topological Overlap measure, called Cross-Network Topological Overlap, was developed by Weighill *et al.* (2017), which can be used to compare the similarity in the neighborhoods of a given node in two distinct networks [118].

Several studies in *Populus* species have involved co-expression networks, some focusing on co-expression networks as the main aspect of the investigation, and others using co-expression networks as a supplementary investigation surrounding the functions of a specific set of genes. Netotea *et al.* (2014) investigated differences in the genome-wide co-expression networks of *P. trichocarpa*, *Oryza sativa* and *A. thaliana* constructed from publicly availabe expression data [129]. It was found that while individual gene-gene co-expression relationships were different between the three species, overal neighbourhoods of genes were significantly conserved across species. Another interesting finding was that orthologs with the most sequence similarity did not have the most similar expression pattern ("expressolog" [130]).

An interesting co-expression study by Grönlund *et al.* (2009) constructed co-expression networks from 1024 publicly available microarray datasets for *P. tremuloides* by jack-knife re-sampling half of the number of samples 100 times, calculating the Pearson correlation between all pairs of genes in each jack-knife re-sample, converting the Pearson correlations to distance metrics, and subsequently constructing 100 minimum spanning trees (MSTs) and merging the resulting networks [131]. This approach of re-sampling allowed for the identification of rarer interactions between genes that would not have been identified through only looking at the dataset as a whole. Another whole genome co-expression study in poplar was performed by Ogata *et al.* (2009) in which 95 publicly available *P. trichocarpa* microarray expression datasets were used to construct a co-expression network and extracted co-expression modules, which were released in a publicly availabe database [132].

Several studies of specific genes in poplar incorporated co-expression elements into their analysis. Tian *et al.* (2017) investigated the role of *P. trichocarpa* Na+/H+ antiporters in stress responses, as well as potential functional divergences within the family of these NHX genes [133]. Using a co-expression network from publicly available data on Phytozome, they showed divergence in the expression pattern of members of this family. Several studies in poplar performed Weighted Gene Co-expression Network Analysis (WGCNA) of genes responding to certain stresses/conditions, including control vs drought conditions [134] in *Populus tremula x alba,* controls vs Jasmonic acid and Salicylic acid treatments in a *Populus deltoides x P. euramericana* hybrid[135], and also a developmental gradient of stem tissue [136]. In a characterization of DWARF14 genes in *P. trichocarpa*, co-expression networks showed divergent expression between the two DWARF14 [137]. In another recent study by *Tuskan et al.* (2018), genes having a GWAS association with callus formation were identified [56], and the co-expression patterns of these genes were investigated using a co-expression network constructed from the *P. trichocarpa* gene expression atlas, and identified interesting clusters of positive and negative co-expression relationships between these genes, showing a clear regulatory pattern. It is evident that co-expression networks are a well-developed and widely-used data layer in various organisms including *Populus* species.

Co-methylation networks are a newer approach looking at the similarity between the methylation patterns of genes, and a more limited number of studies using co-methylation networks were found. However, they are a valid and useful data layer which carries information not present in co-expression datasets.

Van Eijk *et al.* (2012) collected methylation and gene expression data for several human individuals to investigate the relationship between these two data layers [138]. WGCNA was used to construct co-expression and co-methylation networks, and subsequently to identify co-expression and co-methylation modules. In general, co-expression and co-methylation modules had very few overlapping genes, although both co-expression and co-methylation modules showed significant functional enrichment for various GO terms. Linear regression was also used to identify relationships between methylation and expression across individuals in which both positive and negative relationships were identified [138]. Various other co-methylation network analyses have been performed in human cancer investigations [139, 140, 141].

SNP correlation

SNP correlation networks involve calculating the correlation/co-occurrence between SNPs across a population, and can be converted to gene-gene networks by mapping SNPs to the genes in which they reside. The Custom Correlation Coefficient (CCC) is an allele-specific correlation metric proven to be useful in identifying sets of SNPs ("blocs") which can be tested against complex phenotypes to uncover combinatorial genetic associations which affect the phenotype [142, 143]. The edges in the SNP correlation network can also be interpreted as potential co-evolutionary relationships, particularly when the variants in question reside on different chromosomes. The CCC is defined for bi-allelic SNPs. For a given pair of sites i and j, the CCC is calculated 4 times, once for each pair of alleles x and y between the two sites. The CCC between alleles x and y at positions i and j, respectively, is defined as:

$$CCC_{i_x j_y} = \frac{9}{2} R_{i_x j_y} \left(1 - \frac{1}{f_{i_x}}\right) \left(1 - \frac{1}{f_{j_y}}\right) \tag{1.11}$$

where $R_{i_x j_y}$ represents the relative co-occurrence of x and y at positions i and j, f_{i_x} represents the frequency of allele x at position i and f_{j_y} represents the frequency of allele y at position j [142, 143].

CCC has been used to investigate the genetic underpinnings of heart disease [142], psoriasis [143] and genes implicated in various other diseases [144]. This metric was also applied in a study by Bryan *et al.* (2018) in *P. trichocarpa* [145]. The positioning of the two DWARF14 paralogs in the *P. trichocarpa* SNP correlation network were investigated, indicating that they appeared to have different co-evolution partners, potentially indicating functional divergence [145]. The CCC metric has been ported to run on graphics processing units (GPUs) providing a significant increase in speed [146, 147].

Kogelman *et al.* (2014) constructed SNP correlation modules by calculating the Pearson correlation between pairs of variants across individuals followed by topological overlap clustering using WGCNA [148]. This method was termed "WISH" (Weighted Interaction SNP Hub) and was considered an extension of WGCNA to genotype data. Later, in 2017, the developers of WGCNA published an extension of the method to construct SNP correlation networks from GWAS associations, termed "WSCNA" (Weighted SNP Correlation Network Analysis) which involves clustering SNPs based on beta coefficients from a GWAS analysis [149], and describe the use of these networks in calculating polygenic risk scores.

Network-based Data Integration

A useful review by Gligorijević *et al.* (2015) classifies network-based data integration into two categories, namely homogeneous and heterogeneous integration [150]. Homogeneous integration involves integrating networks with the same type of nodes, but different edge types, for example, a gene co-expression network and a gene interaction network. Heterogeneous data integration involves integrating networks with both different node types and edge types. These strategies for data integration are then subdivided into groups based on the stage at which data integration occurs. Early integration involves integration of the datasets, and a single model is built on a combined dataset. This appears similar to the definition of *Concatenation-based integration* as described by Ritchie *et al* (2015) [151]. Late integration involves building separate models from each individual dataset and subsequently combines the information in the separate models. This is similar to *Model-based integration* as described by Ritchie *et al.* (2015) [151]. A third integration strategy as described by Ritchie *et al.* (2015), *transformation-based integration*, involves transforming multiple datasets into an intermediate, common structure, such as a network, which are then merged before the constructing further models [151].

Two of the most exhaustive network-based data integration tools are String (Search Tool for the Retrieval of Interacting Genes) and KnetMiner, both of which are online, freely accessible resources.

STRING is an online, publicly available database of protein interactions, incorporating various data types and data sources, including co-expression, co-occurrence, physical interactions, sequence homology, and associations from textmining [99, 100, 101]. The user can search for genes, and resulting network neighborhoods can also be clustered using K-means clustering and MCL [110]. Protein 3D structure as well as functional enrichment information is also displayed. The SRING database can also be queried through the Cytoscape network visualization app [101, 109]. Certain sets of publicly available data for *P. trichocarpa* is available in STRING.

KnetMiner is a publicly available tool/database consisting of heterogeneous “knowledge” networks for 11 species, including *P. trichocarpa*, and includes layers of information of different types and sources represented as networks, such as GWAS data, sequence homology relationships, annotation information, metabolic pathways, protein interactions and occurence in scientific literature [97, 98]. KnetMiner allows the user to search not only for genes, but also concepts, phenotypes or pathways. A score (KNETscore) is then calculated to rank genes based on their relevance of the neighborhood to the search terms. KnetMiner provides useful network visualizations as well as a chromosomal view indicating the location on the chromosomes in which the genes occur and an “evidence view" indicating the number of nodes/concepts of different types in the neighborhood of the genes in question.

The Mergeomics R package and webserver allows one to integrate GWAS summary statistics with other biological pathways and gene networks, and perform enrichment analyses as well as Weighted Key Driver Analysis [152]. This involves identifying hub genes in a selected/uploaded gene network, and subsequently overlaying phenotype-associated genes from uploaded GWAS analyses, and reports key drivers for each of these genes [152]. Key drivers and their neighborhoods can then be visualized using Cytoscape Web.

Mizrachi *et al.* (2017) developed an interesting network-based data integration approach to combine pathway information from KEGG, eQTL associations and gene expression data in *Eucalyptus*[153]. The network-based integration approach involves constructing a gene interaction network based on information in KEGG as well as eQTL associations with biomass and wood traits. The adjacency matrix of this network is then multiplied with a gene expression matrix, which results in a "network-diffused gene expression" matrix. This adjusts gene expression values based on those of neighboring genes in the gene interaction network. These new gene expression profiles are then correlated with each trait to identify genes of relevance to wood properties and biomass[153].

Walley *et al.* (2016) performed a study which compared the topologies of various gene association networks in Maize [154]. A gene co-expression network, a protein-co-expression network and a phosphoprotein co-expression network were constructed and clustered into modules using WGCNA, and the edge conservation between the networks were calculated using the Jaccard index, and found that 6.1% of the edges were shared between the protein co-expression and gene co-expression networks. Functional enrichment using MapMan [125] terms was performed on modules of co-expressed genes/proteins from the two networks, and similar enriched functions were found in both networks.

NetICS (Network-based Integration of Multi-omics Data) is a data integration strategy based on graph diffusion [155]. This method was developed by Dimitrakopoulos *et al.* (2018) in order to prioritize cancer genes. A directed gene interation network was constructed from publicly available data which included multiple types of relationships, including phosphorylation, co-expression, activation and inhibition. Aberrant genes (i.e. those found to be differentially impacted in a case/control experiment) are marked, and uses network diffusion to predict "mediator genes" which link upstream "genetically aberrant" genes to downstream gene expression changes [155]. This approach successfully identified many known cancer genes.

Gutiérrez *et al.* (2007) investigated gene expression in *A. thaliana* under Carbon and Nitrogen treatments [156]. A separate gene interaction network was constructed using publicly available protein-protein and protein-DNA interactions, as well as miRNA-RNA interactions and the *Arabidopsis* metabolic pathway. A subnetwork consisting of C/N responsive genes and their neighbours in the multi-network was constructed. Clustering of this subnetwork revealed interesting regulatory subnetworks.

Bunyavanich *et al.* (2014) used a multi-omic network-based approach to investigate allergic rhinitis [157]. GWAS was performed on 5633 genotyped individuals, and gene expression was measured in 200 of these individuals. Gene co-expression network and modules were constructed using WGCNA. Co-expression modules that contained genes which harbored or were near to GWAS associated genes were considered candidate modules associated with allergic rhinitis. Associations between SNPs and gene expression was determined (called "eSNPs") which are SNPs within 1MB of a gene which is also associated with the expression of the gene. identifying loci associated both with the allergic rhinitis and the expression of a gene. Modules enriched in eSNPs were also identified, and it was found that the candidate allergic rhinitis modules were enriched in eSNPs associated with allergic rhinitis, and mitochondrial pathways were identified as important components of allergic rhinitis using functional enrichment [157].

Calabrese *et al.* (2017) integrated GWAS and co-expression data in an investigation into genes affecting bone mineral density [158]. Genes identified as associated with bone mineral density in a GWAS analysis were mapped onto a co-expression network, which was subsequently clustered into modules. Co-expression modules which were enriched for GWAS hits were then identified [158].

There have thus been several efforts to integrate various data layers, sometimes for the goal of prioritizing candidate genes, and others for providing biological context for the interpretation of GWAS results.

1.3.3 Signal Processing

Data Representation

In the previous section, we discussed the representation of biological data at network structures, which focuses on relationships between pairs of objects. Here, we discuss the representation of biological data as "signals", and subsequent analysis techniques.

A biological signal represents the response of a variable over some range of input values, which usually have some longitudinal feature, such as a response over increasing time, or a response over increasing distance. Classic examples of biological signals are feature density signals across chromosomes, such as SNP density, gene density, recombination density, GC content, to name a few [159, 105, 23].

These signals have variation at different scales, (i.e. are composed of multiple signals of different frequencies) and signal processing techniques can be used to extract frequency information. McCormick *et al.* (2018) who used the Fourier Transform to identify a prominent periodicity in SNP density, finding that SNP density peaked with a period of 3 base pairs downstream of coding sequence start sites [23], which was explained by the positions in the third "wobble" base being under lower selective pressure.

The Fourier Transform represents a signal as a linear combination of sine and cosine waves. These are infinite waves and thus the Fourier Transform provides no information as to which frequencies are observed at different locations in the signal. The Wavelet Transform is a newer signal processing technique which addresses this limitation [160].

Continuous Wavelet Transform

The Continuous Wavelet Transform (CWT) is a signal processing technique which expresses a signal as a linear combination of special functions called wavelets. These functions are scaled translations of a mother wavelet function, i.e. different widths and different x-axis locations of a particular function. A wavelet w is required to have oscillations and is required to "die out", i.e. the function $\lim_{x\to\infty} w(x) = 0$. An example of a wavelet function called the Ricker Wavelet can be seen in Figure 1.3A.

What results from a wavelet transform is a wavelet coefficient $W(s, \tau)$ (Equation 5.1), for every scale s and translation (shift along the x-axis) τ [160].

$$W(s, \tau) = \frac{1}{\sqrt{s}} \int f(t)\psi^{*}\left(\frac{t-\tau}{s}\right) dt \tag{1.12}$$

This essentially can be interpreted as "sliding" the wavelet of a certain width over the signal, and at each position calculating the integral of the product of the wavelet and the signal over the entire x-axis, producing a vector of coefficients. This process is then repeated for multiple widths of the mother wavelet. An example of the CWT applied to the SNP density of *P. trichocarpa* chromosome 1 can be seen in Figure 1.3B. Other visual examples various mother wavelets and CWT coefficient outputs can be seen in references [161, 105, 160, 162].

Discrete Wavelet Transform

The Discrete Wavelet Transform (DWT) is a sampled version of the CWT, and involves sampling of the x dimention of the signal and scale dimension of the wavelet [160]. This is a dyadic sampling, which results in low frequency, large scales being sampled sparsely and high frequency, small scales being sampled densely [163]. The DWT uses discrete wavelet functions (for example, see Figure 1.3C) and produces a series of sets of coefficients with one set of coefficients for each scale computed (Figure 1.3D). DWT coefficients for Palmer

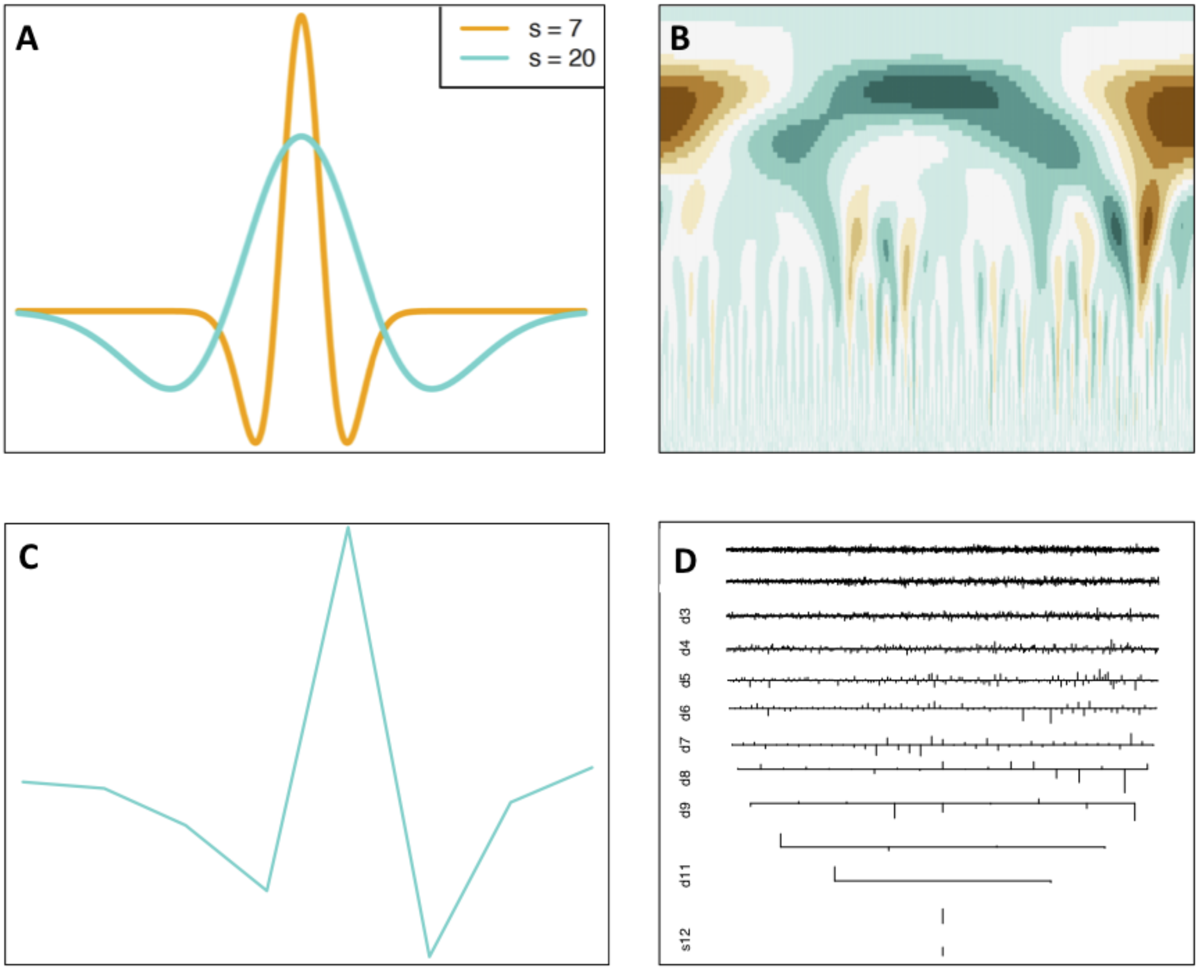


Figure 1.3: **Continuous and Discrete Wavelet Transforms.** (A) Continuous Ricker Wavelet, (B), CWT coefficient matrix heatmap, (C) discrete s8 wavelet, (D) DWT coefficients.

Drought Severity Index data across time can be seen in the tutorial by Dong *et al.* (2008) [161].

1.3.4 Wavelet-Based Analysis and Integration of Biological Data

A useful overview of the wavelet transform and previous biological applications prior to 2003, including sequence analysis, protein structure investigation and expression data analysis to identify periodicities, can be found in the review by Liò (2003) [164]. More recent applications of the wavelet transform in biological data analysis is discussed below.

Thurman *et al.* (2007) performed an investigation to detect "functional domains" of a scale larger than that of within a gene, in the human genome [165]. The wavelet transform was used to smooth density signals of various ENCODE data over various scales. This included transcriptional data, histone acetylation, histone methylation and DNA replication time. A Hidden Markov Model was then used to segment the genome into one of 2 states, namely state 0 ("repressed") and state 1 ("active"), particular signal [165]. This was performed separately for each data type and also in a combined fashion. Domains with the state 1 ("active") classification were enriched in characteristics of "active" chromatin, for example, transcriptional stop/start sites, mRNAs, CpG islands, among others. However, domains with the state 0 ("repressed") classification were significantly enriched in signal transduction genes as determined using GO enrichment. Transposable elements in general were evenly distributed across active and repressed domain, however, certain classes of repeats, such as L1 LINE repeats and LTR elements were enriched in state 0 domains ("repressed" domains) [165].

Shim *et al.* (2015) determined variants which are associated with open chromatin using DNase-seq data from 70 genotyped individuals. Chromatin accessibility vectors are transformed using the DWT prior to associating them to phenotypes [166]. The advantage of this method is that it takes into account the read profile, without having to resort to "artificial" boundaries such as known exon boundaries or sliding windows of a set size [166].

Machado performed wavelet analysis of sequence data by transforming DNA sequence into a vector of numbers, with each base pair mapped to a point on one of the axes of the complex plane [167]. The wavelet transform is applied to these sequence vectors and various wavelets are tested. However, no functional interpretations of results were discussed.

Biological signals can have different relationships with each other depending on the scale at which one is looking. While two features may be correlated at certain scales, they may be anti-correlated at others. Keitt *et al.* (2005) introduced *wavelet-coefficient regression*, in which wavelet transforms are applied to dependent and independant variables before performing regression analysis, allowing for scale-specific inference [168]. Spencer *et al.* (2006) used this kind of approach, applying the DWT and linear model analysis to investigate scale-specific relationships between various genomic features [105]. Genomic signals of recombination, divergence, diversity, GC content and gene content in 1kb regions across human chromosome 20. The DWT was performed on each of these transforms, and calculated the correlation between the wavelet coefficients of features at each scale to identify scale-specific correlations [105]. Paape *et al.* (2012) applied the same approach as Spencer *et al.* (2006), using the wavelet transform followed by linear model analysis to identify genomic features which correlate with recombination in *Medicago truncatula* [159]. The wavelet correlation results revealed a negative correlation between recombination and the distance to the centromere, which had not been found in several other organisms [159].

Very recently, Fernandéz *et al.* (2018) applied the wavelet transform in an application for visualizing DNA methylation data at various scales/resolutions [169].

1.4 Concluding Remarks and Aims

In this review, we have discussed large scale omics data types, multi-omics studies, as well as network based analysis/integration techniques and wavelet-based multi-scale analysis and comparisons, all with a particular focus on investigations performed in poplar. Table

1.2 summarizes examples of multi-omic/data integration studies in poplar. While many such studies have been performed over the last decade, few studies involve the integration of multiple data types in a combined analysis, as opposed to a sequential analysis.

A vast collection of different data types has been generated for *Populus trichocarpa*. As described in this review, the genome has been sequenced and annotated [6], and the assembly is currently in its fourth version of revision. Approximately ∼1,300 *P. trichocarpa* genotypes have been propagated in four different common gardens [10, 11, 12] and have been resequenced. This has provided a large set of ∼ 28,000,000 Single Nucleotide Polymorphisms (SNPs) that have recently been publicly released (DOI 10.13139/OLCF/1411410). Many molecular phenotypes measured through untargetted metabolomics, RNA-Seq, ionomics, pyMBMS, as well as physical properties [170] measured in this population have provided an unparalleled resource for Genome Wide Association Studies (for example, see [44]). DNA methylation data in the form of MeDIP (Methyl-DNA immunoprecipitation)-seq has been performed on 10 different *P. trichocarpa* tissues [71].

The availability of public data as well as access to high performance computing resources at the Oak Ridge Leadership Computing Facility (OLCF) and The Compute and Data Environment for Science (CADES) provides an opportunity for the large-scale, concurrent analysis of all of this data in order to profile and characterize the *P. trichocarpa* genome.

Gaps in the investigations in poplar include (1), comprehensive extraction and investigations of pleiotropic signatures, (2) comprehensive analysis of the positions and signatures of the centromere, and (3) large scale target gene identification from integrated multi-omics datasets. This dissertation will involve the use of network and signal processing techniques for the combined analysis of these various data types, for the goals of (1) increasing fundamental knowledge of *P. trichocarpa* and (2) facilitating the generation of hypotheses about target genes and phenotypes of interest. The research in this dissertation will be divided into four chapters:

Table 1.2: Examples of multi-omic/data integration studies in poplar species.

Species	**Data Types/Layers**	**Year [Reference]**
P. trichocarpa	transcriptomics, metabolomics, biomass/-sugar release	2019 [171]
P. trichocarpa	genomic, transcriptomic, proteomic, flux-omic, wood chemical property phenotypes	2018 [172]
P. tremula x P. tremuloides	transcriptome, proteome, GC-MS metabolome, LC-MS metabolome, pyrolysis-GC MS metabolome	2018 [173]
P. trichocarpa	transcriptomics, co-expression, genotype, callus phenotype (GWAS)	2018 [56]
P. trichocarpa	metabolomics, genotype, transcriptomics, GWAS, eQTL, co-expression	2018 [38]
P. deltoides	metabolomics, microbiome	2018 [37]
P. trichocarpa	co-expression, protein-protein interaction, population genotype	2017 [133]
P. trichocarpa	methylation, transcript expression, miRNAs	2016 [174]
P. tremuloides and Laccaria	transcriptomics, protein-protein interactions,	2016 [175]
P. balsamifera	transcriptomics, metabolomics	2015 [35]
P. trichocarpa and*P. deltoides*	metabolomics, transcriptomics	2014 [36]
P. trichocarpa	genotype, phenotype (GWAS)	2014 [44]
P. trichocarpa	methylome (bisulfite sequencing), transcriptomics	2014 [73]
P. trichocarpa	genotype, phenotype (GWAS)	2014 [12]

Continued on next page.

Table 1.2: (continued)

Species	**Data Types/Layers**	**Year [Reference]**
P. trichocarpa	methylome (MeDIP-seq), transcriptomics	2013 [72]
P. trichocarpa	open chromatin, methylome	2013 [74]
P. trichocarpa	methylome (MeDIP-seq), transcriptomics, transposable elements,	2012 [71]
P. trichocarpa	genotype, repeat elements, methylation, recombination	2012 [11]
Populus euphratica and *Populus x canescens*	transcriptomics, metabolomics	2010 [176]
P. tremula x P. tremuloides	transcriptomics, metabolomics, proteomics	2008 [177]
P. tremula x P. tremuloides	transcriptomics, metabolomics	2007 [103]
P. deltoides x P. nigra and *P. deltoides x P. trichocarpa*	genotypes, metabolites (mQTLs)	2006 [32]

Chapter 2: A data integration "Lines of Evidence" method is presented for the identification and prioritization of target genes involved in functions of interest.

Chapter 3: A new post-GWAS method, Pleiotropy Decomposition, is presented, which extracts pleiotropic relationships between genes and phenotypes from GWAS results, allowing genes to be clustered based on their pleiotropic signatures.

Chapter 4: Continuous wavelet transform signal processing analysis is applied in the characterization genome distributions of various features (e.g. gene density and methylation profiles) in order to identify chromosome structures such as the centromere.

Chapter 5: Synteny analysis and discrete wavelet transform signal processing of various genomic features is performed, followed by correlation analysis, in order to identify scale-specific relationships between various genomic features.

Bibliography

[1] Poulomi Sannigrahi, Arthur J Ragauskas, and Gerald A Tuskan. Poplar as a feedstock for biofuels: a review of compositional characteristics. *Biofuels, Bioproducts and Biorefining*, 4(2):209–226, 2010. 2

[2] Arthur J Ragauskas, Charlotte K Williams, Brian H Davison, George Britovsek, John Cairney, Charles A Eckert, William J Frederick, Jason P Hallett, David J Leak, Charles L Liotta, et al. The path forward for biofuels and biomaterials. *science*, 311(5760):484–489, 2006. 2

[3] Pär K Ingvarsson, Torgeir R Hvidsten, and Nathaniel R Street. Towards integration of population and comparative genomics in forest trees. *New Phytologist*, 212(2):338–344, 2016. 3

[4] Wolfram Weckwerth. Green systems biology - from single genomes, proteomes and metabolomes to ecosystems research and biotechnology. *Journal of proteomics*, 75(1):284–305, 2011. 3, 20

[5] Luis Valledor, María Carbó, Laura Lamelas, Mónica Escandón, Francisco Javier Colina, María Jesús Cañal, and Mónica Meijón. When the tree let us see the forest: Systems biology and natural variation studies in forest species. 2018. 3, 20

[6] Gerald A Tuskan, S Difazio, Stefan Jansson, J Bohlmann, I Grigoriev, U Hellsten, N Putnam, S Ralph, Stephane Rombauts, A Salamov, et al. The Genome of Black Cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science*, 313(5793):1596–1604, 2006. 3, 38

[7] Stan D Wullschleger, DJ Weston, Stephen P DiFazio, and Gerald A Tuskan. Revisiting the sequencing of the first tree genome: Populus trichocarpa. *Tree physiology*, 33(4):357–364, 2012. 3

[8] Stefan Jansson and Carl J Douglas. Populus: a model system for plant biology. *Annu. Rev. Plant Biol.*, 58:435–458, 2007. 3

[9] Tongming Yin, Stephen P DiFazio, Lee E Gunter, Xinye Zhang, Michell M Sewell, Scott A Woolbright, Gery J Allan, Collin T Kelleher, Carl J Douglas, Mingxiu Wang, et al. Genome structure and emerging evidence of an incipient sex chromosome in populus. *Genome research*, 18(3):422–430, 2008. 3

[10] Gerald Tuskan, Gancho Slavov, Steve DiFazio, Wellington Muchero, Ranjan Pryia, Wendy Schackwitz, Joel Martin, Daniel Rokhsar, Robert Sykes, Mark Davis, et al. *Populus* resequencing: towards genome-wide association studies. In *BMC Proceedings*, volume 5, page I21. BioMed Central Ltd, 2011. 3, 9, 38

[11] Gancho T Slavov, Stephen P DiFazio, Joel Martin, Wendy Schackwitz, Wellington Muchero, Eli Rodgers-Melnick, Mindie F Lipphardt, Christa P Pennacchio, Uffe Hellsten, Len A Pennacchio, et al. Genome resequencing reveals multiscale geographic structure and extensive linkage disequilibrium in the forest tree *Populus trichocarpa*. *New Phytologist*, 196(3):713–725, 2012. 3, 9, 15, 38, 40

[12] Luke M Evans, Gancho T Slavov, Eli Rodgers-Melnick, Joel Martin, Priya Ranjan, Wellington Muchero, Amy M Brunner, Wendy Schackwitz, Lee Gunter, Jin-Gui Chen, et al. Population genomics of *Populus trichocarpa* identifies signatures of selection and adaptive trait associations. *Nature Genetics*, 46(10):1089–1096, 2014. 3, 9, 13, 38, 39

[13] David M Goodstein, Shengqiang Shu, Russell Howson, Rochak Neupane, Richard D Hayes, Joni Fazo, Therese Mitros, William Dirks, Uffe Hellsten, Nicholas Putnam, et al. Phytozome: a comparative platform for green plant genomics. *Nucleic acids research*, 40(D1):D1178–D1186, 2012. 4, 5, 16

[14] Mitchell E Skinner, Andrew V Uzilov, Lincoln D Stein, Christopher J Mungall, and Ian H Holmes. JBrowse: A next-generation genome browser. *Genome Research*, 19(9):1630–1638, 2009. 4

[15] Peng Shuai, Dan Liang, Zhoujia Zhang, Weilun Yin, and Xinli Xia. Identification of drought-responsive and novel populus trichocarpa micrornas by high-throughput sequencing and their targets using degradome analysis. *Bmc Genomics*, 14(1):233, 2013. 4

[16] Peng Shuai, Dan Liang, Sha Tang, Zhoujia Zhang, Chu-Yu Ye, Yanyan Su, Xinli Xia, and Weilun Yin. Genome-wide identification and functional prediction of novel and drought-responsive lincrnas in populus trichocarpa. *Journal of experimental botany*, 65(17):4975–4983, 2014. 4

[17] Sha Tang, Yan Dong, Dan Liang, Zhoujia Zhang, Chu-Yu Ye, Peng Shuai, Xiao Han, Ying Zhao, Weilun Yin, and Xinli Xia. Analysis of the drought stress-responsive transcriptome of black cottonwood (populus trichocarpa) using deep rna sequencing. *Plant Molecular Biology Reporter*, 33(3):424–438, 2015. 4

[18] Madhumita Dash, Yordan S Yordanov, Tatyana Georgieva, Hairong Wei, and Victor Busov. Gene network analysis of poplar root transcriptome in response to drought stress identifies a ptajaz3ptarap2. 6-centered hierarchical network. *PloS one*, 13(12):e0208560, 2018. 4

[19] Tania Quesada, Zhen Li, Christopher Dervinis, Yao Li, Philip N Bocock, Gerald A Tuskan, George Casella, John M Davis, and Matias Kirst. Comparative analysis of the transcriptomes of populus trichocarpa and arabidopsis thaliana suggests extensive evolution of gene expression regulation in angiosperms. *New Phytologist*, 180(2):408–420, 2008. 4

[20] Rui Shi, Ying-Hsuan Sun, Quanzi Li, Steffen Heber, Ronald Sederoff, and Vincent L Chiang. Towards a systems approach for lignin biosynthesis in populus trichocarpa: transcript abundance and specificity of the monolignol biosynthetic genes. *Plant and Cell Physiology*, 51(1):144–163, 2009. 5

[21] Hua Bao, Eryang Li, Shawn D Mansfield, Quentin CB Cronk, Yousry A El-Kassaby, and Carl J Douglas. The developing xylem transcriptome and genome-wide analysis of alternative splicing in populus trichocarpa (black cottonwood) populations. *BMC genomics*, 14(1):359, 2013. 5

[22] Markus Schmid, Timothy S Davison, Stefan R Henz, Utz J Pape, Monika Demar, Martin Vingron, Bernhard Schölkopf, Detlef Weigel, and Jan U Lohmann. A gene expression map of arabidopsis thaliana development. *Nature genetics*, 37(5):501, 2005. 6

[23] Ryan F McCormick, Sandra K Truong, Avinash Sreedasyam, Jerry Jenkins, Shengqiang Shu, David Sims, Megan Kennedy, Mojgan Amirebrahimi, Brock D Weers, Brian McKinley, et al. The sorghum bicolor reference genome: improved assembly, gene annotations, a transcriptome atlas, and signatures of genome organization. *The Plant Journal*, 93(2):338–354, 2018. 6, 33

[24] Andrew J Severin, Jenna L Woody, Yung-Tsi Bolon, Bindu Joseph, Brian W Diers, Andrew D Farmer, Gary J Muehlbauer, Rex T Nelson, David Grant, James E Specht, et al. Rna-seq atlas of glycine max: a guide to the soybean transcriptome. *BMC plant biology*, 10(1):160, 2010. 6

[25] Jerome Verdier, Ivone Torres-Jerez, Mingyi Wang, Andry Andriankaja, Stacy N Allen, Ji He, Yuhong Tang, Jeremy D Murray, and Michael K Udvardi. Establishment of the lotus japonicus gene expression atlas (ljgea) and its use to explore legume seed maturation. *The Plant Journal*, 74(2):351–362, 2013. 6

[26] Vagner A Benedito, Ivone Torres-Jerez, Jeremy D Murray, Andry Andriankaja, Stacy Allen, Klementina Kakar, Maren Wandrey, Jérôme Verdier, Hélène Zuber, Thomas Ott, et al. A gene expression atlas of the model legume medicago truncatula. *The Plant Journal*, 55(3):504–513, 2008. 6

[27] Arnis Druka, Gary Muehlbauer, Ilze Druka, Rico Caldo, Ute Baumann, Nils Rostoks, Andreas Schreiber, Roger Wise, Timothy Close, Andris Kleinhofs, et al. An atlas

of gene expression from seed to seed through barley development. *Functional & integrative genomics*, 6(3):202–211, 2006. 6

[28] Lei Wang, Weibo Xie, Ying Chen, Weijiang Tang, Jiangyi Yang, Rongjian Ye, Li Liu, Yongjun Lin, Caiguo Xu, Jinghua Xiao, et al. A dynamic gene expression atlas covering the entire life cycle of rice. *The Plant Journal*, 61(5):752–766, 2010. 6

[29] Ji-Yi Zhang, Yi-Ching Lee, Ivone Torres-Jerez, Mingyi Wang, Yanbin Yin, Wen-Chi Chou, Ji He, Hui Shen, Avinash C Srivastava, Christa Pennacchio, et al. Development of an integrated transcript sequence database and a gene expression atlas for gene discovery and analysis in switchgrass (panicum virgatum l.). *The Plant Journal*, 74(1):160–173, 2013. 6

[30] Marianna Fasoli, Silvia Dal Santo, Sara Zenoni, Giovanni Battista Tornielli, Lorenzo Farina, Anita Zamboni, Andrea Porceddu, Luca Venturini, Manuele Bicego, Vittorio Murino, et al. The grapevine expression atlas reveals a deep transcriptome shift driving the entire plant into a maturation program. *The Plant Cell*, pages tpc–112, 2012. 6

[31] Gary J Patti, Oscar Yanes, and Gary Siuzdak. Innovation: Metabolomics: the apogee of the omics trilogy. *Nature reviews Molecular cell biology*, 13(4):263, 2012. 7

[32] Kris Morreel, Geert Goeminne, Véronique Storme, Lieven Sterck, John Ralph, Wouter Coppieters, Peter Breyne, Marijke Steenackers, Michel Georges, Eric Messens, et al. Genetical metabolomics of flavonoid biosynthesis in populus: a case study. *The Plant Journal*, 47(2):224–237, 2006. 7, 40

[33] Moritz Kaling, Basem Kanawati, Andrea Ghirardo, Andreas Albert, Jana Barbro Winkler, Werner Heller, Csengele Barta, Francesco Loreto, PHILIPPE SCHMITT-KOPPLIN, and JÖRG-PETER SCHNITZLER. Uv-b mediated metabolic rearrangements in poplar revealed by non-targeted metabolomics. *Plant, cell & environment*, 38(5):892–904, 2015. 7

[34] Gerald A Tuskan, Steve DiFazio, Patricia Faivre-Rampant, Muriel Gaudet, Antoine Harfouche, Véronique Jorge, Jessy L Labbé, Priya Ranjan, Maurizio Sabatti, Gancho Slavov, et al. The obscure events contributing to the evolution of an incipient sex chromosome in populus: a retrospective working hypothesis. *Tree genetics & genomes*, 8(3):559–571, 2012. 7

[35] Erin T Hamanishi, Genoa LH Barchet, Rebecca Dauwe, Shawn D Mansfield, and Malcolm M Campbell. Poplar trees reconfigure the transcriptome and metabolome in response to drought in a genotype-and time-of-day-dependent manner. *BMC genomics*, 16(1):329, 2015. 7, 39

[36] Timothy J Tschaplinski, Jonathan M Plett, Nancy L Engle, Aurelie Deveau, Katherine C Cushman, Madhavi Z Martin, Mitchel J Doktycz, Gerald A Tuskan, Annick Brun, Annegret Kohler, et al. *Populus trichocarpa* and *Populus deltoides* Exhibit Different Metabolomic Responses to Colonization by the Symbiotic Fungus *Laccaria bicolor*. *Molecular Plant-Microbe Interactions*, 27(6):546–556, 2014. 7, 39

[37] Allison M Veach, Daniel Yip, Nancy L Engle, Zamin K Yang, Amber Bible, Jennifer Morrell-Falvey, Timothy J Tschaplinski, Udaya C Kalluri, and Christopher W Schadt. Modification of plant cell wall chemistry impacts metabolome and microbiome composition in populus pdkor1 rnai plants. *Plant and Soil*, pages 1–13, 2018. 8, 39

[38] Jin Zhang, Yongil Yang, Kaijie Zheng, Meng Xie, Kai Feng, Sara S Jawdy, Lee E Gunter, Priya Ranjan, Vasanth R Singan, Nancy Engle, et al. Genome-wide association studies and expression-based quantitative trait loci analyses reveal roles of hct 2 in caffeoylquinic acid biosynthesis and its regulation by defense-responsive transcription factors in populus. *New Phytologist*, 220(2):502–516, 2018. 8, 39

[39] Haley J Abel and Eric J Duncavage. Detection of structural dna variation from next generation sequencing data: a review of informatic approaches. *Cancer genetics*, 206(12):432–440, 2013. 8

[40] Thomas LaFramboise. Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic acids research*, 37(13):4181–4193, 2009. 8, 9

[41] Armando Geraldes, SP Difazio, GT Slavov, P Ranjan, W Muchero, J Hannemann, LE Gunter, AM Wymore, CJ Grassa, N Farzaneh, et al. 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, 2013. 8

[42] Rasmus Nielsen, Joshua S Paul, Anders Albrechtsen, and Yun S Song. Genotype and snp calling from next-generation sequencing data. *Nature Reviews Genetics*, 12(6):443, 2011. 9

[43] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, et al. The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome research*, 2010. 9

[44] Athena D McKown, Jaroslav Klápště, Robert D Guy, Armando Geraldes, Ilga Porth, Jan Hannemann, Michael Friedmann, Wellington Muchero, Gerald A Tuskan, Jürgen Ehlting, et al. Genome-wide association implicates numerous genes underlying ecological trait variation in natural populations of *Populus trichocarpa*. *New Phytologist*, 203(2):535–553, 2014. 9, 13, 38, 39

[45] Nadia Solovieff, Chris Cotsapas, Phil H Lee, Shaun M Purcell, and Jordan W Smoller. Pleiotropy in complex traits: challenges and strategies. *Nature Reviews Genetics*, 14(7):483–495, 2013. 9

[46] Peter M Visscher, Matthew A Brown, Mark I McCarthy, and Jian Yang. Five years of gwas discovery. *The American Journal of Human Genetics*, 90(1):7–24, 2012. 9

[47] Arthur Korte and Ashley Farlow. The advantages and limitations of trait analysis with gwas: a review. *Plant methods*, 9(1):29, 2013. 9

[48] Sherry A Flint-Garcia, Jeffry M Thornsberry, and Buckler IV. Structure of linkage disequilibrium in plants*. *Annual review of plant biology*, 54(1):357–374, 2003. 10

[49] William Astle, David J Balding, et al. Population structure and cryptic relatedness in genetic association studies. *Statistical Science*, 24(4):451–471, 2009. 10

[50] Hyun Min Kang, Jae Hoon Sul, Susan K Service, Noah A Zaitlen, Sit-yee Kong, Nelson B Freimer, Chiara Sabatti, Eleazar Eskin, et al. Variance component model to account for sample structure in genome-wide association studies. *Nature genetics*, 42(4):348–354, 2010. 10, 11

[51] Hyun Min Kang, Noah A Zaitlen, Claire M Wade, Andrew Kirby, David Heckerman, Mark J Daly, and Eleazar Eskin. Efficient control of population structure in model organism association mapping. *Genetics*, 178(3):1709–1723, 2008. 10

[52] Randall C Johnson, George W Nelson, Jennifer L Troyer, James A Lautenberger, Bailey D Kessing, Cheryl A Winkler, and Stephen J O'Brien. Accounting for multiple comparisons in a genome-wide association study (gwas). *BMC genomics*, 11(1):724, 2010. 11

[53] Shawn R Narum. Beyond bonferroni: less conservative analyses for conservation genetics. *Conservation Genetics*, 7(5):783–787, 2006. 12

[54] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979. 12

[55] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995. 12, 13

[56] Gerald A Tuskan, Ritesh Mewalal, Lee E Gunter, Kaitlin J Palla, Kelsey Carter, Daniel A Jacobson, Piet C Jones, Benjamin J Garcia, Deborah A Weighill, Philip D Hyatt, et al. Defining the genetic components of callus formation: A gwas approach. *PloS one*, 13(8):e0202519, 2018. 14, 27, 39

[57] Jingyuan Liu, Meixia Ye, Sheng Zhu, Libo Jiang, Mengmeng Sang, Jingwen Gan, Qian Wang, Minren Huang, and Rongling Wu. Two-stage identification of snp effects on dynamic poplar growth. *The Plant Journal*, 93(2):286–296, 2018. 14

[58] Chunyan Liu, Falong Lu, Xia Cui, and Xiaofeng Cao. Histone methylation in higher plants. *Annual review of plant biology*, 61:395–420, 2010. 14

[59] Alexandra Lusser, Doris Kölle, and Peter Loidl. Histone acetylation: lessons from the plant kingdom. *Trends in plant science*, 6(2):59–65, 2001. 14

[60] EJ Finnegan, RK Genger, WJ Peacock, and ES Dennis. Dna methylation in plants. *Annual review of plant biology*, 49(1):223–247, 1998. 14

[61] Julie A Law and Steven E Jacobsen. Establishing, maintaining and modifying dna methylation patterns in plants and animals. *Nature Reviews Genetics*, 11(3):204, 2010. 14

[62] Miho M Suzuki and Adrian Bird. Dna methylation landscapes: provocative insights from epigenomics. *Nature Reviews Genetics*, 9(6):465, 2008. 14, 15

[63] Daniel Zilberman, Mary Gehring, Robert K Tran, Tracy Ballinger, and Steven Henikoff. Genome-wide analysis of arabidopsis thaliana dna methylation uncovers an interdependence between methylation and transcription. *Nature genetics*, 39(1):61, 2007. 15

[64] Viswanathan Chinnusamy and Jian-Kang Zhu. Epigenetic regulation of stress responses in plants. *Current opinion in plant biology*, 12(2):133–139, 2009. 15

[65] Jörn Lämke and Isabel Bäurle. Epigenetic and chromatin-based mechanisms in environmental stress adaptation and stress memory in plants. *Genome biology*, 18(1):124, 2017. 15

[66] Filipe V Jacinto, Esteban Ballestar, and Manel Esteller. Methyl-dna immunoprecipitation (medip): hunting down the dna methylome. *Biotechniques*, 44(1):35–43, 2008. 15

[67] Marianne Frommer, Louise E McDonald, Douglas S Millar, Christina M Collis, Fujiko Watt, Geoffrey W Grigg, Peter L Molloy, and Cheryl L Paul. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual dna strands. *Proceedings of the National Academy of Sciences*, 89(5):1827–1831, 1992. 15

[68] Felix Krueger, Benjamin Kreck, Andre Franke, and Simon R Andrews. Dna methylome analysis using short bisulfite sequencing data. *Nature methods*, 9(2):145, 2012. 15

[69] Peter W Laird. Principles and challenges of genome-wide dna methylation analysis. *Nature Reviews Genetics*, 11(3):191, 2010. 15

[70] Christoph Bock. Analysing and interpreting dna methylation data. *Nature Reviews Genetics*, 13(10):705, 2012. 15

[71] Kelly J Vining, Kyle R Pomraning, Larry J Wilhelm, Henry D Priest, Matteo Pellegrini, Todd C Mockler, Michael Freitag, and Steven H Strauss. Dynamic DNA cytosine methylation in the *Populus trichocarpa* genome: tissue-level variation and relationship to gene expression. *BMC Genomics*, 13(1):1, 2012. 15, 18, 19, 38, 40

[72] Kelly Vining, Kyle R Pomraning, Larry J Wilhelm, Cathleen Ma, Matteo Pellegrini, Yanming Di, Todd C Mockler, Michael Freitag, and Steven H Strauss. Methylome reorganization during in vitro dedifferentiation and regeneration of populus trichocarpa. *BMC plant biology*, 13(1):92, 2013. 16, 40

[73] Dan Liang, Zhoujia Zhang, Honglong Wu, Chunyu Huang, Peng Shuai, Chu-Yu Ye, Sha Tang, Yunjie Wang, Ling Yang, Jun Wang, Weilun Yin, and Xinli Xia. Single-base-resolution methylomes of populus trichocarpa reveal the association between dna methylation and drought stress. *BMC Genetics*, 15(1):S9, Jun 2014. 16, 39

[74] Clément Lafon-Placette, Patricia Faivre-Rampant, Alain Delaunay, Nathaniel Street, Franck Brignolas, and Stéphane Maury. Methylome of dnase i sensitive chromatin in populus trichocarpa shoot apical meristematic cells: a simplified approach

revealing characteristics of gene-body dna methylation in open chromatin state. *New Phytologist*, 197(2):416–430, 2013. 16, 40

[75] Jason D Buenrostro, Paul G Giresi, Lisa C Zaba, Howard Y Chang, and William J Greenleaf. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, dna-binding proteins and nucleosome position. *Nature methods*, 10(12):1213, 2013. 16

[76] Jason D Buenrostro, Beijing Wu, Howard Y Chang, and William J Greenleaf. Atac-seq: a method for assaying chromatin accessibility genome-wide. *Current protocols in molecular biology*, 109(1):21–29, 2015. 16

[77] Zefu Lu, Brigitte T Hofmeister, Christopher Vollmers, Rebecca M DuBois, and Robert J Schmitz. Combining atac-seq with nuclei sorting for discovery of cis-regulatory regions in plant genomes. *Nucleic acids research*, 45(6):e41–e41, 2016. 17

[78] Marko Bajic, Kelsey A Maher, and Roger B Deal. Identification of open chromatin regions in plant genomes using atac-seq. In *Plant Chromatin Dynamics*, pages 183–201. Springer, 2018. 17

[79] Kelsey A Maher, Marko Bajic, Kaisa Kajala, Mauricio Reynoso, Germain Pauluzzi, Donnelly A West, Kristina Zumstein, Margaret Woodhouse, Kerry Bubb, Michael W Dorrity, et al. Profiling of accessible chromatin regions across multiple plant species and cell types reveals common gene regulatory principles and new control modules. *The Plant Cell*, 30(1):15–36, 2018. 17

[80] Ronan C OḾalley, Shao-shan Carol Huang, Liang Song, Mathew G Lewsey, Anna Bartlett, Joseph R Nery, Mary Galli, Andrea Gallavotti, and Joseph R Ecker. Cistrome and epicistrome features shape the regulatory dna landscape. *Cell*, 165(5):1280–1292, 2016. 17

[81] Anna Bartlett, Ronan C O'Malley, Shao-shan Carol Huang, Mary Galli, Joseph R Nery, Andrea Gallavotti, and Joseph R Ecker. Mapping genome-wide transcription-factor binding sites using dap-seq. *Nature protocols*, 12(8):1659, 2017. 17

[82] R Keith Slotkin and Robert Martienssen. Transposable elements and the epigenetic regulation of the genome. *Nature reviews genetics*, 8(4):272, 2007. 17, 18

[83] Thomas Wicker, François Sabot, Aurélie Hua-Van, Jeffrey L Bennetzen, Pierre Capy, Boulos Chalhoub, Andrew Flavell, Philippe Leroy, Michele Morgante, Olivier Panaud, et al. A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics*, 8(12):973, 2007. 17, 18

[84] Jeffrey L Bennetzen and Hao Wang. The contributions of transposable elements to the structure, function, and evolution of plant genomes. *Annual review of plant biology*, 65:505–530, 2014. 18

[85] Savannah J Klein and Rachel J O'Neill. Transposable elements: genome innovation, chromosome diversity, and centromere conflict. *Chromosome Research*, pages 1–19, 2018. 18

[86] Eduard Kejnovsky, Jennifer S Hawkins, and Cédric Feschotte. Plant transposable elements: biology and evolution. In *Plant Genome Diversity Volume 1*, pages 17–34. Springer, 2012. 18

[87] Brandon S Gaut, Stephen I Wright, Carène Rizzon, Jan Dvorak, and Lorinda K Anderson. Recombination: an underappreciated factor in the evolution of plant genomes. *Nature Reviews Genetics*, 8(1):77, 2007. 18

[88] Pierre Capy, Giuliano Gasperi, Christian Biémont, and Claude Bazin. Stress and transposable elements: co-evolution or useful parasites? *Heredity*, 85(2):101, 2000. 18

[89] Fengfeng Zhou and Ying Xu. Reppop: a database for repetitive elements in populus trichocarpa. *BMC genomics*, 10(1):14, 2009. 18

[90] Rosa Maria Cossu, Matteo Buti, Tommaso Giordani, Lucia Natali, and Andrea Cavallini. A computational study of the dynamics of LTR retrotransposons in the *Populus trichocarpa* genome. *Tree Genetics & Genomes*, 8(1):61–75, 2012. 18

[91] Lucia Natali, Rosa Maria Cossu, Flavia Mascagni, Tommaso Giordani, and Andrea Cavallini. A survey of gypsy and copia ltr-retrotransposon superfamilies and lineages and their distinct dynamics in the populus trichocarpa (l.) genome. *Tree genetics & genomes*, 11(5):107, 2015. 18

[92] Gabriele Usai, Flavia Mascagni, Lucia Natali, Tommaso Giordani, and Andrea Cavallini. Comparative genome-wide analysis of repetitive dna in the genus populus l. *Tree Genetics & Genomes*, 13(5):96, 2017. 19

[93] F Mascagni, G Usai, L Natali, A Cavallini, and T Giordani. A comparison of methods for ltr-retrotransposon insertion time profiling in the populus trichocarpa genome. *Caryologia*, 71(1):85–92, 2018. 19

[94] Fei Yi, Zirui Jia, Yao Xiao, Wenjun Ma, and Junhui Wang. Sptedb: a database for transposable elements in salicaceous plants. *Database*, 2018, 2018. 19

[95] Keywan Hassani-Pak and Christopher Rawlings. Knowledge discovery in biological databases for revealing candidate genes linked to complex phenotypes. *Journal of integrative bioinformatics*, 14(1), 2017. 20

[96] Haifei Hu, Armin Scheben, and David Edwards. Advances in integrating genomics and bioinformatics in the plant breeding pipeline. *Agriculture*, 8(6):75, 2018. 20

[97] Keywan Hassani-Pak, Martin Castellote, Maria Esch, Matthew Hindle, Artem Lysenko, Jan Taubert, and Christopher Rawlings. Developing integrated crop knowledge networks to advance candidate gene discovery. *Applied & translational genomics*, 11:18–26, 2016. 20, 30

[98] Keywan Hassani-Pak. Knetminer-an integrated data platform for gene mining and biological knowledge discovery [phd thesis]. 2017. 20, 30

[99] Damian Szklarczyk, Andrea Franceschini, Michael Kuhn, Milan Simonovic, Alexander Roth, Pablo Minguez, Tobias Doerks, Manuel Stark, Jean Muller, Peer Bork, et al. The string database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic acids research*, 39(suppl_1):D561–D568, 2010. 20, 30

[100] Christian von Mering, Martijn Huynen, Daniel Jaeggi, Steffen Schmidt, Peer Bork, and Berend Snel. String: a database of predicted functional associations between proteins. *Nucleic acids research*, 31(1):258–261, 2003. 20, 30

[101] Damian Szklarczyk, John H Morris, Helen Cook, Michael Kuhn, Stefan Wyder, Milan Simonovic, Alberto Santos, Nadezhda T Doncheva, Alexander Roth, Peer Bork, et al. The string database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic acids research*, page gkw937, 2016. 20, 30

[102] Atsushi Fukushima and Miyako Kusano. A network perspective on nitrogen metabolism from model to crop plants using integrated 'omics' approaches. *Journal of experimental botany*, 65(19):5619–5630, 2014. 20

[103] Max Bylesjö, Daniel Eriksson, Miyako Kusano, Thomas Moritz, and Johan Trygg. Data integration in plant biology: the o2pls method for combined modeling of transcript and metabolite data. *The Plant Journal*, 52(6):1181–1191, 2007. 20, 40

[104] Arjun Krishnan, Jaclyn N Taroni, and Casey S Greene. Integrative networks illuminate biological factors underlying gene–disease associations. *Current Genetic Medicine Reports*, 4(4):155–162, 2016. 20

[105] Chris CA Spencer, Panos Deloukas, Sarah Hunt, Jim Mullikin, Simon Myers, Bernard Silverman, Peter Donnelly, David Bentley, and Gil McVean. The Influence of Recombination on Human Genetic Diversity. *PLoS Genet*, 2(9):e148, 2006. 20, 33, 34, 37

[106] Albert-Laszlo Barabasi and Zoltan N Oltvai. Network biology: understanding the cell's functional organization. *Nature reviews genetics*, 5(2):101–113, 2004. 21, 23

[107] Martin Charles Golumbic. *Algorithmic graph theory and perfect graphs*, volume 57. Elsevier, 2004. 21, 23

[108] Daniel Marcus. *Graph theory: A problem oriented approach*. Maa, 2008. 23

[109] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, 13(11):2498–2504, 2003. 23, 30

[110] Anton J Enright, Stijn Van Dongen, and Christos A Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucleic acids research*, 30(7):1575–1584, 2002. 24, 30

[111] Bin Zhang and Steve Horvath. A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, 4(1), 2005. 24

[112] Nirmala Akula, Ancha Baranova, Donald Seto, Jeffrey Solka, Michael A Nalls, Andrew Singleton, Luigi Ferrucci, Toshiko Tanaka, Stefania Bandinelli, Yoon Shin Cho, et al. A network-based approach to prioritize results from genome-wide association studies. *PloS one*, 6(9):e24220, 2011. 24

[113] Charles R Farber. Systems-level analysis of genome-wide association data. *G3: Genes— Genomes— Genetics*, 3(1):119–129, 2013. 24

[114] Mark DM Leiserson, Jonathan V Eldridge, Sohini Ramachandran, and Benjamin J Raphael. Network analysis of gwas data. *Current opinion in genetics & development*, 23(6):602–610, 2013. 24

[115] CHRISTIAN Darabos, SAMANTHA H Harmon, and JASON H Moore. Using the bipartite human phenotype network to reveal pleiotropy and epistasis beyond the

gene. In *Pac Symp Biocomput*, volume 19, pages 188–199. World Scientific, 2014. 24

[116] Maud Fagny, Joseph N Paulson, Marieke L Kuijjer, Abhijeet R Sonawane, Cho-Yi Chen, Camila M Lopes-Ramos, Kimberly Glass, John Quackenbush, and John Platig. Exploring regulation in tissues with eqtl networks. *Proceedings of the National Academy of Sciences*, 114(37):E7841–E7850, 2017. 24

[117] Yupeng Li, Stephanie A Pearl, and Scott A Jackson. Gene networks in plant biology: approaches in reconstruction and analysis. *Trends in plant science*, 20(10):664–675, 2015. 25

[118] Deborah A. Weighill and Daniel Jacobson. *Network Metamodeling: Effect of Correlation Metric Choice on Phylogenomic and Transcriptomic Network Topology*, pages 143–183. Springer International Publishing, Cham, 2017. 25, 26

[119] Koh Aoki, Yoshiyuki Ogata, and Daisuke Shibata. Approaches for extracting practical information from gene co-expression networks in plant biology. *Plant and Cell Physiology*, 48(3):381–390, 2007. 25

[120] Robert J Schaefer, Jean-Michel Michno, and Chad L Myers. Unraveling gene function in agricultural species using gene co-expression networks. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1860(1):53–63, 2017. 25

[121] Elise AR Serin, Harm Nijveen, Henk WM Hilhorst, and Wilco Ligterink. Learning from co-expression networks: possibilities and challenges. *Frontiers in plant science*, 7:444, 2016. 25

[122] Abbasali Emamjomeh, Elham Saboori Robat, Javad Zahiri, Mahmood Solouki, and Pegah Khosravi. Gene co-expression network reconstruction: a review on computational methods for inferring functional information from plant-based expression data. *Plant biotechnology reports*, 11(2):71–86, 2017. 25

[123] Sara Movahedi, Michiel Van Bel, Ken S Heyndrickx, and Klaas Vandepoele. Comparative co-expression analysis in plant biology. *Plant, cell & environment*, 35(10):1787–1798, 2012. 25

[124] Gene Ontology Consortium. The gene ontology (go) database and informatics resource. *Nucleic acids research*, 32(suppl_1):D258–D261, 2004. 25

[125] Oliver Thimm, Oliver Bläsing, Yves Gibon, Axel Nagel, Svenja Meyer, Peter Krüger, Joachim Selbig, Lukas A Müller, Seung Y Rhee, and Mark Stitt. MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *The Plant Journal*, 37(6):914–939, 2004. 25, 31

[126] Steve Horvath and Jun Dong. Geometric interpretation of gene coexpression network analysis. *PLoS computational biology*, 4(8):e1000117, 2008. 26

[127] Andy M Yip and Steve Horvath. Gene network interconnectedness and the generalized topological overlap measure. *BMC bioinformatics*, 8(1):22, 2007. 26

[128] Peter Langfelder and Steve Horvath. Wgcna: an r package for weighted correlation network analysis. *BMC bioinformatics*, 9(1):559, 2008. 26

[129] Sergiu Netotea, David Sundell, Nathaniel R Street, and Torgeir R Hvidsten. Complex: conservation and divergence of co-expression networks in a. thaliana, populus and o. sativa. *BMC genomics*, 15(1):106, 2014. 26

[130] Rohan V Patel, Hardeep K Nahal, Robert Breit, and Nicholas J Provart. Bar expressolog identification: expression profile similarity ranking of homologous genes in plant species. *The Plant Journal*, 71(6):1038–1050, 2012. 26

[131] Andreas Grönlund, Rishikesh P Bhalerao, and Jan Karlsson. Modular gene expression in poplar: a multilayer network approach. *New Phytologist*, 181(2):315–322, 2009. 27

[132] Yoshiyuki Ogata, Hideyuki Suzuki, and Daisuke Shibata. A database for poplar gene co-expression analysis for systematic understanding of biological processes, including stress responses. *Journal of wood science*, 55(6):395, 2009. 27

[133] Fengxia Tian, Ermei Chang, Yu Li, Pei Sun, Jianjun Hu, and Jin Zhang. Expression and integrated network analyses revealed functional divergence of nhx-type na+/h+ exchanger genes in poplar. *Scientific Reports*, 7(1):2607, 2017. 27, 39

[134] Liang-Jiao Xue, Christopher J Frost, Chung-Jui Tsai, and Scott A Harding. Drought response transcriptomes are altered in poplar with reduced tonoplast sucrose transporter expression. *Scientific Reports*, 6:33655, 2016. 27

[135] Jie Luo, Wenxiu Xia, Pei Cao, Zheng'ang Xiao, Yan Zhang, Meifeng Liu, Chang Zhan, Nian Wang, et al. Integrated transcriptome analysis reveals plant hormones jasmonic acid and salicylic acid coordinate growth and defense responses upon fungal infection in poplar. *Biomolecules*, 9(1):12, 2019. 27

[136] Qing Chao, Zhi-Fang Gao, Dong Zhang, Biligen-Gaowa Zhao, Feng-Qin Dong, Chun-Xiang Fu, Li-Jun Liu, and Bai-Chen Wang. The developmental dynamics of the populus stem transcriptome. *Plant biotechnology journal*, 17(1):206–219, 2019. 27

[137] Kaijie Zheng, Xiaoping Wang, Deborah A Weighill, Hao-Bo Guo, Meng Xie, Yongil Yang, Jun Yang, Shucai Wang, Daniel A Jacobson, Hong Guo, et al. Characterization of dwarf14 genes in populus. *Scientific reports*, 6:21593, 2016. 27

[138] Kristel R van Eijk, Simone de Jong, Marco PM Boks, Terry Langeveld, Fabrice Colas, Jan H Veldink, Carolien GF de Kovel, Esther Janson, Eric Strengman, Peter Langfelder, et al. Genetic analysis of dna methylation and gene expression levels in whole blood of healthy human subjects. *BMC genomics*, 13(1):636, 2012. 28

[139] Thomas E Bartlett, Sofia C Olhede, and Alexey Zaikin. A dna methylation network interaction measure, and detection of network oncomarkers. *PloS one*, 9(1):e84573, 2014. 28

[140] Min Jin Ha, Veerabhadran Baladandayuthapani, and Kim-Anh Do. Dingo: differential network analysis in genomics. *Bioinformatics*, 31(21):3413–3420, 2015. 28

[141] Ruslan Akulenko and Volkhard Helms. Dna co-methylation analysis suggests novel functional associations between gene pairs in breast cancer samples. *Human molecular genetics*, 22(15):3016–3022, 2013. 28

[142] Sharlee Climer, Wei Yang, Lisa Fuentes, Victor G Dávila-Román, and C Charles Gu. A custom correlation coefficient (ccc) approach for fast identification of multi-snp association patterns in genome-wide snps data. *Genetic epidemiology*, 38(7):610–621, 2014. 28, 29

[143] Sharlee Climer, Alan R Templeton, and Weixiong Zhang. Allele-specific network reveals combinatorial interaction that transcends small effects in psoriasis gwas. *PLOS Computional Biology*, 10(9):e1003766, 2014. 28, 29

[144] Sharlee Climer, Alan R Templeton, and Weixiong Zhang. Human gephyrin is encompassed within giant functional noncoding yin–yang sequences. *Nature communications*, 6:6534, 2015. 29

[145] Anthony C Bryan, Jin Zhang, Jianjun Guo, Priya Ranjan, Vasanth Singan, Kerrie Barry, Jeremy Schmutz, Deborah Weighill, Daniel Jacobson, Sara Jawdy, et al. A variable polyglutamine repeat affects subcellular localization and regulatory activity of a populus angustifolia protein. *G3: Genes, Genomes, Genetics*, 8(8):2631–2641, 2018. 29

[146] Wayne Joubert, James Nance, Sharlee Climer, Deborah Weighill, and Daniel Jacobson. Parallel accelerated custom correlation coefficient calculations for genomics applications. *arXiv preprint arXiv:1705.08213*, 2017. 29

[147] Wayne Joubert, Deborah Weighill, David Kainer, Sharlee Climer, Amy Justice, Kjiersten Fagnan, and Daniel Jacobson. Attacking the opioid epidemic: determining the epistatic and pleiotropic genetic architectures for chronic pain and opioid addiction. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis*, page 57. IEEE Press, 2018. 29

[148] Lisette JA Kogelman and Haja N Kadarmideen. Weighted interaction snp hub (wish) network method for building genetic networks for complex diseases and traits using whole genome genotype data. *BMC systems biology*, 8(2):S5, 2014. 29

[149] Morgan E Levine, Peter Langfelder, and Steve Horvath. A weighted snp correlation network method for estimating polygenic risk scores. In *Biological Networks and Pathway Analysis*, pages 277–290. Springer, 2017. 29

[150] Vladimir Gligorijević and Nataša Pržulj. Methods for biological data integration: perspectives and challenges. *Journal of the Royal Society Interface*, 12(112):20150571, 2015. 29

[151] Marylyn D Ritchie, Emily R Holzinger, Ruowang Li, Sarah A Pendergrass, and Dokyoon Kim. Methods of integrating data to uncover genotype–phenotype interactions. *Nature Reviews Genetics*, 16(2):85, 2015. 30

[152] Douglas Arneson, Anindya Bhattacharya, Le Shu, Ville-Petteri Mäkinen, and Xia Yang. Mergeomics: a web server for identifying pathological pathways, networks, and key regulators via multidimensional data integration. *BMC genomics*, 17(1):722, 2016. 31

[153] Eshchar Mizrachi, Lieven Verbeke, Nanette Christie, Ana C Fierro, Shawn D Mansfield, Mark F Davis, Erica Gjersing, Gerald A Tuskan, Marc Van Montagu, Yves Van de Peer, et al. 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, 2017. 31

[154] Justin W Walley, Ryan C Sartor, Zhouxin Shen, Robert J Schmitz, Kevin J Wu, Mark A Urich, Joseph R Nery, Laurie G Smith, James C Schnable, Joseph R Ecker, et al. Integration of omic networks in a developmental atlas of maize. *Science*, 353(6301):814–818, 2016. 31

[155] Christos Dimitrakopoulos, Sravanth Kumar Hindupur, Luca Häfliger, Jonas Behr, Hesam Montazeri, Michael N Hall, and Niko Beerenwinkel. Network-based

integration of multi-omics data for prioritizing cancer genes. *Bioinformatics*, 34(14):2441–2448, 2018. 31, 32

[156] Rodrigo A Gutiérrez, Laurence V Lejay, Alexis Dean, Francesca Chiaromonte, Dennis E Shasha, and Gloria M Coruzzi. Qualitative network models and genome-wide expression data define carbon/nitrogen-responsive molecular machines in arabidopsis. *Genome biology*, 8(1):R7, 2007. 32

[157] Supinda Bunyavanich, Eric E Schadt, Blanca E Himes, Jessica Lasky-Su, Weiliang Qiu, Ross Lazarus, John P Ziniti, Ariella Cohain, Michael Linderman, Dara G Torgerson, et al. Integrated genome-wide association, coexpression network, and expression single nucleotide polymorphism analysis identifies novel pathway in allergic rhinitis. *BMC medical genomics*, 7(1):48, 2014. 32

[158] Gina M Calabrese, Larry D Mesner, Joseph P Stains, Steven M Tommasini, Mark C Horowitz, Clifford J Rosen, and Charles R Farber. Integrating gwas and co-expression network data identifies bone mineral density genes sptbn1 and mark3 and an osteoblast functional module. *Cell systems*, 4(1):46–59, 2017. 32

[159] Timothy Paape, Peng Zhou, Antoine Branca, Roman Briskine, Nevin Young, and Peter Tiffin. Fine-scale population recombination rates, hotspots, and correlates of recombination in the medicago truncatula genome. *Genome biology and evolution*, 4(5):726–737, 2012. 33, 37

[160] CM Leavey, MN James, J Summerscales, and R Sutton. An introduction to wavelet transforms: a tutorial approach. *Insight-Non-Destructive Testing and Condition Monitoring*, 45(5):344–353, 2003. 33, 34

[161] Xuejun Dong, Paul Nyren, Bob Patton, Anne Nyren, Jim Richardson, and Thomas Maresca. Wavelets for agriculture and biology: a tutorial with applications and outlook. *BioScience*, 58(5):445–453, 2008. 34, 36

[162] Xiangcheng Mi, Haibao Ren, Zisheng Ouyang, Wei Wei, and Keping Ma. The use of the mexican hat and the morlet wavelets for detection of ecological patterns. *Plant Ecology*, 179(1):1–19, 2005. 34

[163] Bjørn K Alsberg, Andrew M Woodward, and Douglas B Kell. An introduction to wavelet transforms for chemometricians: A time-frequency approach. *Chemometrics and intelligent laboratory systems*, 37(2):215–239, 1997. 34

[164] Pietro Lio. Wavelets in bioinformatics and computational biology: state of art and perspectives. *Bioinformatics*, 19(1):2–9, 2003. 36

[165] Robert E Thurman, Nathan Day, William S Noble, and John A Stamatoyannopoulos. Identification of higher-order functional domains in the human encode regions. *Genome research*, 17(6):917–927, 2007. 36

[166] Heejung Shim and Matthew Stephens. Wavelet-based genetic association analysis of functional phenotypes arising from high-throughput sequencing assays. *The annals of applied statistics*, 9(2):655, 2015. 36

[167] JA Tenreiro Machado, António C Costa, and Maria Dulce Quelhas. Wavelet analysis of human DNA. *Genomics*, 98(3):155–163, 2011. 37

[168] Timothy H Keitt and Dean L Urban. Scale-specific inference using wavelets. *Ecology*, 86(9):2497–2504, 2005. 37

[169] Lisardo Fernández, Mariano Pérez, and Juan M Orduña. Visualization of dna methylation results through a gpu-based parallelization of the wavelet transform. *The Journal of Supercomputing*, pages 1–14, 2018. 37

[170] Ilga Porth, Jaroslav Klápště, Oleksandr Skyba, Ben SK Lai, Armando Geraldes, Wellington Muchero, Gerald A Tuskan, Carl J Douglas, Yousry A El-Kassaby, and Shawn D Mansfield. *Populus trichocarpa* cell wall chemistry and ultrastructure trait variation, genetic control and genetic correlations. *New Phytologist*, 197(3):777–790, 2013. 38

[171] Jin Zhang, Mi Li, Anthony C Bryan, Chang Geun Yoo, William Rottmann, Kimberly A Winkeler, Cassandra M Collins, Vasanth Singan, Erika A Lindquist, Sara S Jawdy, et al. Overexpression of a serine hydroxymethyltransferase increases biomass production and reduces recalcitrance in the bioenergy crop populus. *Sustainable Energy & Fuels*, 3(1):195–207, 2019. 39

[172] Jack P Wang, Megan L Matthews, Cranos M Williams, Rui Shi, Chenmin Yang, Sermsawat Tunlaya-Anukit, Hsi-Chuan Chen, Quanzi Li, Jie Liu, Chien-Yuan Lin, et al. Improving wood properties for wood utilization through multi-omics integration in lignin biosynthesis. *Nature communications*, 9(1):1579, 2018. 39

[173] Ogonna Obudulu, Niklas Mähler, Tomas Skotare, Joakim Bygdell, Ilka N Abreu, Maria Ahnlund, Madhavi Latha Gandla, Anna Petterle, Thomas Moritz, Torgeir R Hvidsten, et al. A multi-omics approach reveals function of secretory carrier-associated membrane proteins in wood formation of populus trees. *BMC genomics*, 19(1):11, 2018. 39

[174] Brigitte Schönberger, Xiaochao Chen, Svenja Mager, and Uwe Ludewig. Site-dependent differences in dna methylation and their impact on plant establishment and phosphorus nutrition in populus trichocarpa. *PloS one*, 11(12):e0168623, 2016. 39

[175] Peter E Larsen, Avinash Sreedasyam, Geetika Trivedi, Shalaka Desai, Yang Dai, Leland J Cseke, and Frank R Collart. Multi-omics approach identifies molecular mechanisms of plant-fungus mycorrhizal interaction. *Frontiers in plant science*, 6:1061, 2016. 39

[176] Dennis Janz, Katja Behnke, Jörg-Peter Schnitzler, Basem Kanawati, Philippe Schmitt-Kopplin, and Andrea Polle. Pathway analysis of the transcriptome and metabolome of salt sensitive and tolerant poplar species reveals evolutionary adaption of stress tolerance mechanisms. *BMC Plant Biology*, 10(1):150, 2010. 40

[177] Max Bylesjo, Robert Nilsson, Vaibhav Srivastava, Andreas Gronlund, Annika I Johansson, Stefan Jansson, Jan Karlsson, Thomas Moritz, Gunnar Wingsle, and Johan Trygg. Integrated analysis of transcript, protein and metabolite data to study lignin biosynthesis in hybrid aspen. *Journal of proteome research*, 8(1):199–210, 2008. 40

Chapter 2

Pleiotropic and Epistatic Network-Based Discovery: Integrated Networks for Target Gene Discovery

This chapter contains contributions from multiple authors and has been published as:

Weighill D, Jones P, Shah M, Ranjan P, Muchero W, Schmutz J, Sreedasyam A, Macaya-Sanz D, Sykes R, Zhao N, Martin MZ, DiFazio S, Tschaplinski TJ, Tuskan G and Jacobson D (2018) Pleiotropic and Epistatic Network-Based Discovery: Integrated Networks for Target Gene Discovery. Front. Energy Res. 6:30. doi: 10.3389/fenrg.2018.00030

Author Contributions: DW calculated methylation TPM values, constructed the networks, developed the scoring technique, performed the data layering and scoring analysis and interpreted the results, PJ performed the outlier analysis and GWAS and wrote the part of the GWAS Network Construction section describing the outlier analysis and GWAS methods, MS mapped gene expression atlas reads and calculated gene expression TPM values, SD, GT and WM lead the effort on constructing the GWAS population, TJT led the leaf sample collection for GCMS-based metabolomic analyses, identified the peaks, and summarized the metabolomics data and wrote the Metabolomics Phenotype Data section and provided lists of input lignin-related genes and metabolites, PR did automated extraction of metabolite intensity from GCMS, MZM collected the leaf samples and manually extracted the metabolite data, NZ conducted leaf sample preparation, extracted and derivatized and analyzed the metabolites by GCMS, JS and AS generated the gene expression atlas data and contributed the Gene Expression Data section, SD and DMS generated the SNP calls and contributed the Single Nucleotide Polymorphism Data section, RS generated the pyMBMS data and contributed the pyMBMS Phenotype Data section, DJ conceived of and supervised the project, generated MapMan annotations and edited the manuscript, DW, PJ, SD, DMS, RS, TJT, JS and AS wrote the manuscript. Nancy Engle, David Weston, Ryan Aug, KC Cushman, Lee Gunter and Sara Jawdy collected the metabolomics samples, Carissa Bleker contributed to the GWAS and outlier analysis, Mark Davis generated the pyMBMS data and the Department of Energy Joint Genome Institute (JGI) performed the sequencing.

Abstract

Biological organisms are complex systems that are composed of functional networks of interacting molecules and macro-molecules. Complex phenotypes are the result of orchestrated, hierarchical, heterogeneous collections of expressed genomic variants. However, the effects of these variants are the result of historic selective pressure and current environmental and epigenetic signals, and, as such, their co-occurrence can be seen as genome-wide correlations in a number of different manners. Biomass recalcitrance (i.e., the resistance of plants to degradation or deconstruction, which ultimately enables access to a plant's sugars) is a complex polygenic phenotype of high importance to biofuels initiatives. This study makes use of data derived from the re-sequenced genomes from over 800 different *Populus trichocarpa* genotypes in combination with metabolomic and pyMBMS data across this population, as well as co-expression and co-methylation networks in order to better understand the molecular interactions involved in recalcitrance, and identify target genes involved in lignin biosynthesis/degradation. A Lines Of Evidence (LOE) scoring system is developed to integrate the information in the different layers and quantify the number of lines of evidence linking genes to target functions. This new scoring system was applied to quantify the lines of evidence linking genes to lignin-related genes and phenotypes across the network layers, and allowed for the generation of new hypotheses surrounding potential new candidate genes involved in lignin biosynthesis in *P. trichocarpa*, including various AGAMOUS-LIKE genes. The resulting Genome Wide Association Study networks, integrated with Single Nucleotide Polymorphism (SNP) correlation, co-methylation and co-expression networks through the LOE scores are proving to be a powerful approach to determine the pleiotropic and epistatic relationships underlying cellular functions and, as such, the molecular basis for complex phenotypes, such as recalcitrance.

2.1 Introduction

Populus species are promising sources of cellulosic biomass for biofuels because of their fast growth rate, high cellulose content and moderate lignin content [1]. Ragauskas *et al.* (2006) outline areas of research needed "to increase the impact, efficiency, and sustainability of bio-refinery facilities" [2], such as research into modifying plants to enhance favorable traits, including altered cell wall structure leading to increased sugar release, as well as resilience to biotic and abiotic stress. One particular research target in *Populus* species is the decrease/alteration of the lignin content of cell walls.

A large collection of different data types has been generated for *Populus trichocarpa*. The genome has been sequenced and annotated [3], and the assembly is currently in its third version of revision. A collection of 1,100 accessions of *P. trichocarpa* that have been clonally propagated in four different common gardens [4, 5, 6] have been resequenced, which has provided a large set of $\sim$ 28,000,000 Single Nucleotide Polymorphisms (SNPs) that has recently been publicly released (http://bioenergycenter.org/besc/gwas/). Many molecular phenotypes, such as untargetted metabolomics and pyMBMS phenotypes, that have been measured in this population provide an unparalleled resource for Genome Wide Association Studies (for example, see McKnown *et al.* (2014) [7]). DNA methylation data in the form of MeDIP (Methyl-DNA immunoprecipitation)-seq has been performed on 10 different *P. trichocarpa* tissues [8], and gene expression has been measured across various tissues and conditions.

This study involved the development of a method to integrate these various data types in order to identify new possible candidate genes involved in target functions of interest. The importance of *P. trichocarpa* as a bioenergy crop, the availability of the high density SNP data in a GWAS population, as well as the increasing amount of genomic/phenotypic data being generated for *P. trichocarpa* made it an excellent species in which to demonstrate the method. Integrating Genome Wide Association Study (GWAS) data with other data types has previously been done to help provide context and identify relevant subnetworks/modules [9, 10]. Ritchie *et al.* (2015) reviewed techniques for integrating various data types for the aim of investigating gene-phenotype associations [11]. Integrating multiple lines of evidence is a useful strategy as the more lines of evidence that connect a gene to a phenotype lowers the chance of false positives. Ritchie *et al.* (2015) categorized data integration approaches into two main classes, namely multi-staged analysis and meta-dimensional analysis [11]. Multi-staged analysis analyses aims to enrich a biological signal through various steps of analysis. Meta-dimensional analysis involves the concurrent analysis of various data types, and is divided into three subcategories [11]: Concatenation-based integration concatenates the data matrices of different data types into a single matrix on which a model is constructed (for example, see Fridley *et al.* (2012) [12]). Model-based integration involves constructing a separate model for each dataset and then constructing a final model from the results of the separate models (for example, see Kim *et al.* (2013) [13]). Transformation-based integration involves transforming transforming each data type into a common form (e.g. a network) before combining them (for example, see Kim *et al.* (2012) [14]).

This study presents a new transformation-based integration technique: the calculation of Lines Of Evidence (LOE) scores across SNP correlation, GWAS, co-methylation and co-expression networks for *P. trichocarpa*. Association networks for the various different data types were constructed, including a pyMBMS GWAS network, a metabolomics GWAS network, as well as co-expression, co-methylation and SNP correlation networks, and subsequently the information in the different networks was integrated through the calculation of the newly developed Lines Of Evidence (LOE) scores. These scores quantify the number of lines of evidence connecting each gene to target functions of interest. In this work, we apply this data integration technique to the wealth of *P. trichocarpa* data in order to identify new potential genes involved in lignin biosynthesis/degradation/regulation in *P. trichocarpa*. The LOE scores represent the number of lines of evidence that exist connecting genes to lignin-related genes and phenotypes across the network layers. This is a novel multi-omic data integration approach which provides easily interpretable scores, and allows for the identification of new possible candidate genes involved in lignin biosynthesis/regulation through multiple lines of evidence. This is also the first time all of these *P. trichocarpa* datasets have been integrated on a genome-scale in a network-based manner, allowing for the easy identification of new target genes through their respective connections across network layers.

2.2 Methods

2.2.1 Overview

This approach involved combining various data types in order to identify new possible target genes involved in lignin biosynthesis/degradation/regulation. Figure 2.1 summarizes the overall approach. First, association networks were constructed including metabolomics and pyMBMS GWAS networks, co-expression, co-methylation and SNP correlation networks. Known lignin-related genes and phenotypes were then identified, and used as seeds to select lignin-related subnetworks from these various networks. The Lines Of Evidence (LOE) scoring technique was developed, and each gene was then scored based on its Lines Of Evidence linking it to lignin-related genes and phenotypes.

2.2.2 Metabolomics Phenotype Data

The *P. trichocarpa* leaf samples for 851 unique clones were collected over three consecutive sunny days in July 2012. For 200 of those clones, a second biological replicate was also sampled. Typically, leaves (leaf plastocron index 9 plus or minus 1) on a south facing branch from the upper canopy of each tree were quickly collected, wiped with a wet tissue to clean both surfaces and the leaf then fast frozen under dry ice. Leaves were kept on dry ice and shipped back to the lab and stored at -80 until processed for analyses. Metabolites from leaf samples were lyophilized and then ground in a micro-Wiley mill (1 mm mesh size). Approximately 25 mg of each sample was twice extracted in 2.5 mL 80% ethanol (aqueous) for 24 hr with the extracts combined, and 0.5 mL dried in a helium stream. "Sorbitol [($75 \mu l$ of a 1 mg/mL aqueous solution)] was added ... before extraction as an internal standard to correct for differences in

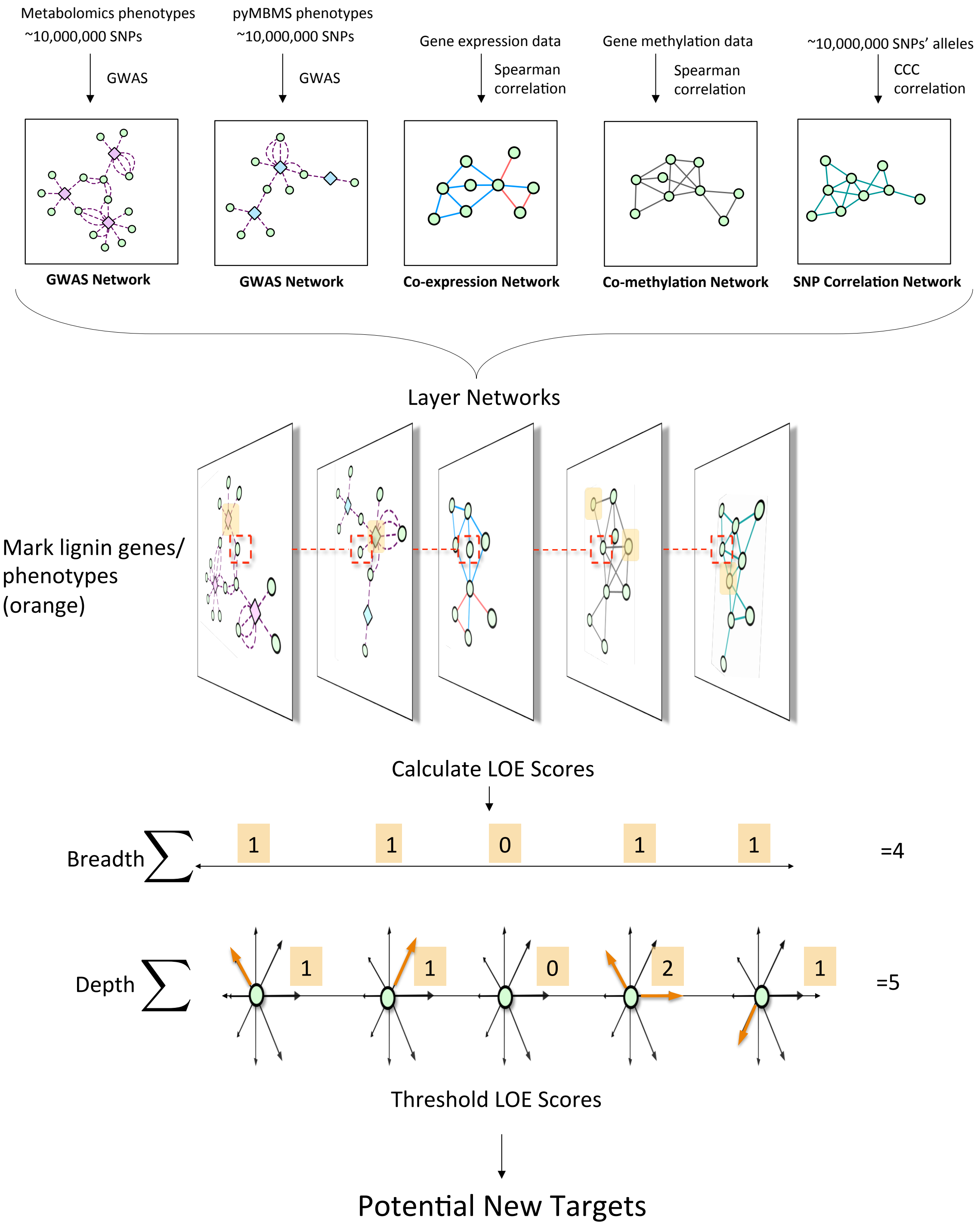


Figure 2.1: **Overview of LOE approach.** Overview of pipeline for data layering and score calcualtion. First, the different network layers are constructed. Networks are layered, and lignin-related genes and phenotypes (orange) are identified. LOE scores are calculated for each gene. An example of the LOE score calculation for the red-boxed gene is shown. Thresholding the LOE scores results in a set of new potential target genes involved in lignin biosynthesis/degradation/regulation.

extraction efficiency, subsequent differences in derivatization efficiency and changes in sample volume during heating" [15]. Metabolites in the dried sample extracts were converted to their trimethylsilyl (TMS) derivatives, and analyzed by gas chromatography-mass spectrometry, as described previously [16, 17], and also described by Timm *et al.* (2016) [18]: Briefly, dried extracts of metabolites "were dissolved in acetonitrile followed by the addition of N-methyl-N-trimethylsilyltrifluoroacetamide (MSTFA) with 1% trimethylchlorosilane (TMCS), and samples then heated for 1 h at 70 to generate trimethylsilyl (TMS) derivatives [16, 17]. After 2 days, aliquots were injected into an Agilent 5975C inert XL gas chromatograph-mass spectrometer (GC-MS). The standard quadrupole GC-MS is operated in the electron impact (70 eV) ionization mode, targeting 2.5 full-spectrum (50-650 Da) scans per second, as described previously [16]. Metabolite peaks were extracted using a key selected ion, characteristic m/z fragment, rather than the total ion chromatogram, to minimize integrating co-eluting metabolites" (quotation from Timm *et al.* (2016) [18]). As described in Zhao *et al.* (2015) [15]: "...[The peak areas were] normalized to the quantity of the internal standard (sorbitol) [injected, and the] amount of sample extracted... A large user-created database [($>$2400 spectra)] of mass spectral electron impact ionization (EI) fragmentation patterns of TMS-derivatized metabolites, as well as the Wiley Registry [10th] Edition combined with NIST [2014] mass spectral database, were used to identify the metabolites of interest to be quantified" [15]. (Brackets indicate deviations from quoted text.)

2.2.3 pyMBMS Phenotype Data

The pyMBMS phenotype data was generated using the method as described in Biswal *et al.* (2015) [19]: "A commercially available molecular beam mass spectrometer (MBMS) designed specifically for biomass analysis was used for pyrolysis vapor analysis [[20, 21, 22]]. Approximately 4 mg of air dried 20 mesh biomass was introduced into the quartz pyrolysis reactor via 80 uL deactivated stainless steel Eco-Cups provided with the autosampler. Mass spectral data from m/z 30-450 were acquired on a Merlin Automation data system version 3.0 using 17 eV electron impact ionization."

The pyMBMS mz peaks were annotated as described by Sykes *et al.* (2009) [21], as done previously by Muchero *et al.* (2015) [23].

2.2.4 Single Nucleotide Polymorphism Data

A dataset consisting of 28,342,758 SNPs called across 882 *P. trichocarpa* [3] genotypes was obtained from http://bioenergycenter.org/besc/gwas/. This dataset is derived from whole genome sequencing of undomesticated *P. trichocarpa* genotypes collected from the U.S. and Canada, and clonally replicated in common gardens [4]. Genotypes from this population have previously been used for population genomics [6] and GWAS studies in *P. trichocarpa* [7] as well as for investigating linkage disequilibrium in the population [5].

Whole genome resequencing was carried out on a sample 882 *P. trichocarpa* natural individuals to an expected median coverage of 15x using Illumina Genome Analyzer, HiSeq 2000, and HiSeq 2500 sequencing platforms at the DOE Joint Genome Institute. Alignments to the *P. trichocarpa* Nisqually-1 v.3.0 reference genome were performed using BWA v0.5.9-r16 with default parameters, followed by post-processing with the picard FixMateInformation and MarkDuplicates tools. Genetic variants were called by means of the Genome Analysis Toolkit v. 3.5.0 (GATK; Broad Institute, Cambridge, MA, USA) [24, 25]. Briefly, variants were called independently for each individual using the concatenation of RealignerTargetCreator, IndelRealigner and HaplotypeCaller tools, and the whole population was combined using GenotypeGVCFs, obtaining a dataset with all the variants detected across the sample population. Biallelic SNPs were extracted using the SelectVariants tool and quality-filtered using the GATK's machine-learning implementation Variant Quality Score Recalibration (VQSR). To this end, the tool VariantRecalibrator was used to create the recalibration file and the sensitivity tranches file. As a "truth" dataset, we used SNP calls from a population of seven female and seven male *P. trichocarpa* that had been crossed in a half diallel design. "True" SNPs were identified by the virtual absence of segregation distortion and Mendelian violations in the progeny of these 49 crosses (ca. 500 offspring in total). As a "non-true" dataset, we used the SNP calls of seven open-pollinated crosses from these 7 females (n $= 90$), filtered using hard-filtering methods recommended in the GATK documentation (tool: VariantFiltration; quality thresholds: QD < 1.5, FS > 75.0, MQ < 35.0, missing alleles < 0.5 and MAF > 0.05). The prior likelihoods for the true and non-true datasets were Q = 15 and Q = 10, respectively, and the variant quality annotations to define the variant recalibration space were DP, QD, MQ, MQRankSum, ReadPosRankSum, FS, SOR and InbreedingCoeff. Finally, we used the ApplyRecalibration tool on the full GWAS dataset to assign SNPs to tranches representing different levels of confidence. We selected SNPs in the tranche with true sensitivity < 90, which minimizes false positives, but at an expected cost of 10% false negatives. The final filtered dataset had a transition/transversion ratio of 2.07, compared to 1.88 for the unfiltered SNPs. To further validate the quality of these SNP calls, we compared them to an Illumina Infinium BeadArray that had been generated from a subset of this population dataset [26]. The average match rate was 96% (±2% SD) for 641 individuals across 20,723 loci.

SNPs in this dataset were divided into different Tranches, indicating the percentage of "true" SNPs recovered. For further analysis in this study, we made use of the PASS SNPs, corresponding to the most stringent Tranche, recovering 90% of the true SNPs [see *http://gatkforums.broadinstitute.org/gatk/discussion/39/variant-quality-score-recalibration-vqsr*]. VCFtools [27] was used to extract the desired Tranche of SNPs from the VCF file and reformat it into .tfam and .tped files.

2.2.5 GWAS Network Construction

The metabolomics and pyMBMS data was used as phenotypes in a genome wide association analysis. The respective phenotype measured over all the genotypes were analyzed to account for potential outliers. A median absolute deviation (MAD) from the median [28] cutoff was applied to determine if a particular measurement of a given phenotype was an outlier with respect to all measurements of that phenotype across the population. To account for asymmetry, the deviation values were estimated separately for values below and above the median, respectively. The distribution of the measured values together with the distribution of their estimated deviation was analyzed and a cutoff of 5 was determined to identify putative outlier values. Phenotypes that had non-outlier measurements in at least 20 percent of the population were retained for further analysis, this was to ensure sufficient signal for the genome wide association model. This resulted in 1262 pyMBMS derived phenotypes and 818 metabolomics derived phenotypes.

To estimate the statistical significant associations between the respective phenotypes and the SNPs called across the population, we applied a linear mixed model using EMMAX [29]. Taking into account population structure estimated from a kinship matrix, we tested each of the respective 2080 phenotypes against the high-confidence SNPs and corrected for multiple hypotheses bias using the Benjamini-Hochberg control for false-discovery rate of 0.1 [30]. This was done in parallel with a python wrapper that utilized the schwimmbad python package [31].

SNP-Phenotype GWAS networks were then pruned to only include SNPs that resided within genes, and SNPs were mapped to their respective genes, resulting in a gene-phenotype network. SNPs were determined to be within genes using the gene boundaries defined in the `Ptrichocarpa_210_v3.0.gene.gff3` from the *P. trichocarpa* version 3.0 genome assembly on Phytozome [32].

2.2.6 Gene Expression Data

P. trichocarpa (Nisqually-1) RNA-seq dataset from JGI Plant Gene Atlas project (Sreedasyam et al., unpublished) was obtained from Phytozome. This dataset consists of samples for standard tissues (leaf, stem, root and bud tissue) and libraries generated from nitrogen source study. List of sample descriptions was accessed from: https://phytozome.jgi.doe.gov/phytomine/aspect.do?name=Expression.

P. trichocarpa (Nisqually-1) cuttings were potted in $4''$ X $4''$ X $5''$ containers containing 1:1 mix of peat and perlite. Plants were grown under 16-h-light/8-h-dark conditions, maintained at 20-23 and an average of 235 μmol m^{-2}s^{-1} to generate tissue for (1) standard tissues and (2) nitrogen source study. Plants for standard tissue experiment were watered with McCown's woody plant nutrient solution and plants for nitrogen experiment were supplemented with either 10mM KNO3 (NO3− plants) or 10mM NH4Cl (NH4+ plants) or 10 mM urea (urea plants). Once plants reached leaf plastochron index 15 (LPI-15), leaf, stem, root and bud tissues were harvested and immediately flash frozen in liquid nitrogen and stored at -80 until further processing was done. Every harvest involved at least three independent biological replicates for each condition and a biological replicate consisted of tissue pooled from 3 plants.

RNA extraction and sequencing was performed as previously described in McCormick *et al.* (2018) [33]. Tissue was ground under liquid nitrogen and high quality RNA was extracted using standard Trizol-reagent based extraction [34]. The integrity and concentration of the RNA preparations were checked initially using Nano-Drop ND-1000 (Nano-Drop Technologies) and then by BioAnalyzer (Agilent Technologies). "Plate-based RNA sample prep was performed on the PerkinElmer Sciclone NGS robotic liquid handling system using Illumina's TruSeq Stranded mRNA HT sample prep kit utilizing poly-A selection of mRNA following the protocol outlined by Illumina in their user guide: http://support.illumina.com/sequencing/sequencing_kits/truseq_stranded_mrna_ht_sample_prep_kit.html. The quantified libraries were then prepared for sequencing on the Illumina HiSeq sequencing platform utilizing a TruSeq paired-end cluster kit, v4, and Illumina's cBot instrument to generate a clustered flowcell for sequencing. Sequencing of the flowcell was performed on the Illumina HiSeq2500 sequencer using HiSeq TruSeq SBS sequencing kits, v4, following a 2x150 indexed run recipe" [33].

2.2.7 Co-expression Network Construction

Gene expression atlas data for *P. trichocarpa* consisting of 63 different samples were used to construct a co-expression network. Reads were trimmed using Skewer [35]. Star [36] was then used to align the reads to the *P. trichocarpa* reference genome [3] obtained from Phytozome [32]. TPM (Transcripts Per Million) expression values [37] were then calculated for each gene. This resulted in a gene expression matrix E in which rows represented genes, columns represented samples and each entry ij represented the expression (TPM) of gene i in sample j. The Spearman correlation coefficient was then calculated between the expression profiles of all pairs of genes (i.e. all pairs of rows of the matrix E) using the mcxarray and mcxdump programs from the MCL-edge package [38, 39] available from http://micans.org/mcl/. This was performed in parallel using Perl wrappers making use of the Parallel::MPI::Simple Perl module, (Alex Gough, http://search.cpan.org/~ajgough/Parallel-MPI-Simple-0.03/Simple.pm) using compute resources at the Oak Ridge Leadership Computing Facility (OLCF).

Supplementary Figure S2.1A shows the distribution of Spearman correlation values for the co-expression network. An absolute threshold of 0.85 was applied.

2.2.8 Co-Methylation Network Construction

Methylation data for *P. trichocarpa* [8] re-aligned to the version 3.0 assembly of *P. trichocarpa* was obtained from Phytozome [32]. This data consisted of MeDIP-seq (Methyl-DNA immunoprecipitation-seq) reads from 10 different *P. trichocarpa* tissues, including bud, callus, female catkin, internode explant, leaf, male catkin, phloem, regenerated internode, root and xylem tissue.

BamTools stats [40] was used to determine basic properties of the reads in each .bam file. Samtools [41] was then used to extract only mapped reads. The number of reads which mapped to each gene feature was determined using htseq-count [42]. These read counts were then converted to TPM values [37], providing a methylation score for each gene in each tissue. The TPM value for a gene g in a given sample was defined as:

$$TPM_g = \frac{\frac{c_g}{l_g} \times 10^6}{\sum_g \frac{c_g}{l_g}} \tag{2.1}$$

where c_g is the number of reads mapped to gene g and l_g is the length of gene g in kb, calculated by subtracting the gene start position from the gene end position, and dividing the resulting difference by 1,000. A methylation matrix M was then formed, in which rows represented genes, columns represented tissues and each entry ij represented the methylation score (TPM) of gene i in tissue j. A co-methylation network (see Busch *et al.* (2016) [43], Akulenko *et al.* (2013) [44] and Davies *et al.* (2012) [45]) was then constructed by calculating the Spearman correlation coefficient between the methylation profiles of all pairs of genes using mcxarray and mcxdump programs from the MCL-edge package [38, 39] `http://micans.org/mcl/`. Supplementary Figure S2.1B shows the distribution of Spearman Correlation values. An absolute threshold of 0.95 was applied.

Read counting using htseq-count, as well as Spearman correlation calculations were performed in parallel using Perl wrappers making use of the Parallel::MPI::Simple Perl module, developed by Alex Gough and available on The Comprehensive Perl Archive Network (CPAN) at `www.cpan.org` and used compute resources at the Oak Ridge Leadership Computing Facility (OLCF).

2.2.9 SNP Correlation Network Construction

The Custom Correlation Coefficient (CCC) [46, 47] was used to calculate the correlation between the occurrence of pairs of SNPs across the 882 genotypes. The CCC between allele x at position i and allele y and position j is defined as:

$$CCC_{i_x j_y} = \frac{9}{2} R_{i_x j_y} \left(1 - \frac{1}{f_{i_x}}\right) \left(1 - \frac{1}{f_{j_y}}\right) \tag{2.2}$$

where $R_{i_x j_y}$ is the relative co-occurrence of allele x at position i and allele y at position j, f_{i_x} is the frequency of allele x at position i and f_{j_y} is the frequency of allele y at position j.

This was performed in a parallel fashion using similar computational approaches as described for the co-expression network above. The set of ~10 million SNPs was divided

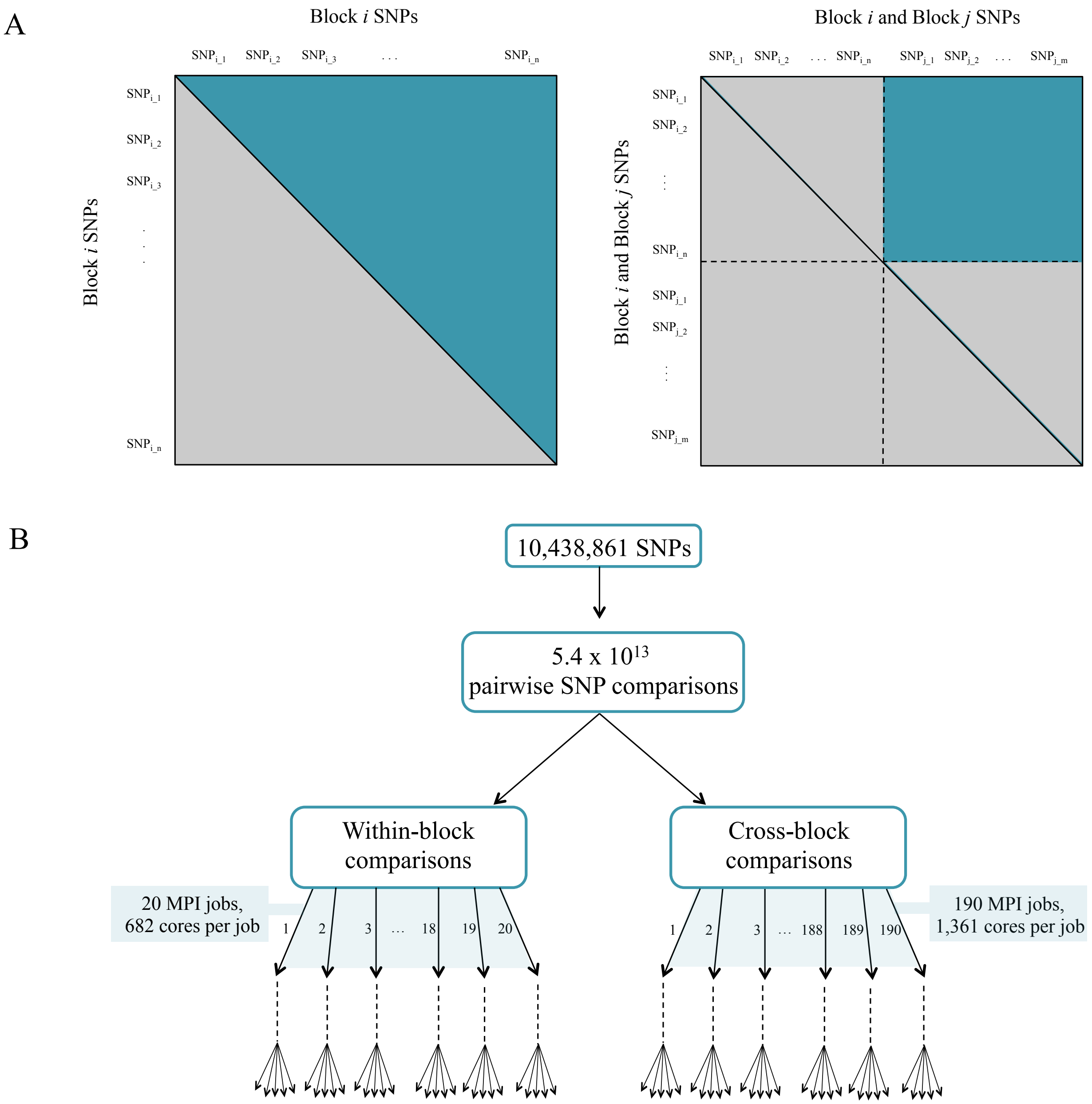


Figure 2.2: **Computing SNP correlations.** (A) Parallelization strategy for ccc calculation between all pairs of SNPs. (B) MPI jobs for within and cross-block comparisons.

into 20 different blocks, and the CCC was calculated for each within-block and cross-block SNPs in separate jobs, to a total of 210 MPI jobs (Figure 2.2). A threshold of 0.7 was then applied. The resulting SNP correlation network was pruned to only include SNPs that resided within genes. Gene boundaries used were defined in the `Ptrichocarpa_210_v3.0.gene.gff3` from the *P. trichocarpa* version 3.0 genome assembly on Phytozome [32]. A local LD filter was then set, retaining correlations between SNPs greater than 10kb apart. The distribution of CCC values can be seen in Supplementary Figure S2.1C (Supplementary Text S2.1).

2.2.10 Target Lignin Genes/Phenotypes

A scoring system was developed in order to quantify the Lines Of Evidence (LOE) linking each gene to lignin-related genes/phenotypes. The LOE scores quantify the number of lines linking each gene to lignin-related genes and phenotypes across the different network data layers. Thus, the method requires as input a list of known lignin-related genes/phenotypes.

P. trichocarpa gene annotations in the `Ptrichocarpa_210_v3.0.annotation_info.txt` file from the version 3.0 genome assembly were used, available on Phytozome [32]. This included *Arabidopsis* best hits and corresponding gene descriptions, as well as GO terms [48, 49] and Pfam domains [50]. Genes were also assigned MapMan annotations using the Mercator tool [51].

Lignin building blocks (monolignols) are derived from phenylalanine in the phenylpropanoid and monolignol pathways, and phenylalanine itself is produced from the shikimate pathway [52]. To compile a list of *P. trichocarpa* genes which are related to the biosynthesis of lignin, *P. trichocarpa* genes were assigned MapMan annotations using the Mercator tool [51]. Genes in the Shikimate (MapMan bins 13.1.6.1, 13.1.6.3 and 13.1.6.4), Phenylpropanoid (MapMan bin 16.2) and Lignin/Lignan (MapMan bin 16.2.1) pathways were then selected. A list of these lignin-related genes and their MapMan annotations can be seen in Supplementary Table S2.1.

Lignin-related pyMBMS peaks, as described by Sykes *et al.* (2009) [21], Davis *et al.* (2006) [53] and Muchero *et al.* (2015) [23] were identified among the pyMBMS GWAS hits, and are shown in Supplementary Table S2.2. Lignin-related metabolites and metabolites in the lignin pathway were also identified among the metabolomics GWAS hits, a list of which can be seen in Supplementary Table S2.3. For partially identified metabolites, additional RT and mz information can be seen in Supplementary Table S2.3.

2.2.11 Extraction of Lignin-Related Subnetworks

Let L_G, L_M and L_P represent our sets of lignin-related genes, metabolites and pyMBMS peaks, respectively (Supplementary Tables S2.1, S2.2 and S2.3). A network can be defined as $N = (V, E)$ where V is the set of nodes and E is the set of edges connecting nodes in V. In particular, let the co-expression network be represented by $N_{coex} = (V_{coex}, E_{coex})$, the co-methylation network by $N_{cometh} = (V_{cometh}, E_{cometh})$ and the SNP correlation network by $N_{snp} = (V_{snp}, E_{snp})$. The GWAS networks can be represented as bipartite networks $N = (U, V, E)$ where U is the set of phenotype nodes, V is the set of gene nodes, and E is the set of edges, with each edge e_{ij} connecting node $i \in U$ with node $j \in V$. Let the metabolomics GWAS network be represented by $N_{metab} = (U_{metab}, V_{metab}, E_{metab})$ and the pyMBMS GWAS network by $N_{pymbms} = (U_{pymbms}, V_{pymbms}, E_{pymbms})$. We construct the *guilt by association* subnetworks of genes connected to lignin-related genes/phenotypes as follows:

N^L_{coex} is the subnetwork of N_{coex} including the lignin related genes $l \in L_G$ and their direct neighbors:

$$N^L_{coex} = (V^L_{coex}, E^L_{coex}) \text{ where} \tag{2.3}$$

$$V^L_{coex} = \{g | g \in (L_G \cap V_{coex})\} \cup \{g | (g \in V_{coex}) \wedge (\exists l \in L_G | \{l, g\} \in E_{coex})\} \tag{2.4}$$

$$E^L_{coex} = \{e = \{i, j\} \in E_{coex} | i \in V^L_{coex} \wedge j \in V^L_{coex}\} \tag{2.5}$$

N^L_{cometh} is the subnetwork of N_{cometh} including the lignin related genes $l \in L_G$ and their direct neighbors:

$$N^L_{cometh} = (V^L_{cometh}, E^L_{cometh}) \text{ where} \tag{2.6}$$

$$V^L_{cometh} = \{g|g \in (L_G \cap V_{cometh})\} \cup \{g|\,(g \in V_{cometh}) \wedge (\exists l \in L_G|\{l,g\} \in E_{cometh})\} \tag{2.7}$$

$$E^L_{cometh} = \{e = \{i,j\} \in E_{cometh}|i \in V^L_{cometh} \wedge j \in V^L_{cometh}\} \tag{2.8}$$

N^L_{snp} is the subnetwork of N_{snp} including the lignin related genes $l \in L_G$ and their direct neighbors:

$$N^L_{snp} = (V^L_{snp}, E^L_{snp}) \text{ where} \tag{2.9}$$

$$V^L_{snp} = \{g|g \in (L_G \cap V_{snp})\} \cup \{g|\,(g \in V_{snp}) \wedge (\exists l \in L_G|\{l,g\} \in E_{snp})\} \tag{2.10}$$

$$E^L_{snp} = \{e = \{i,j\} \in E_{snp}|i \in V^L_{snp} \wedge j \in V^L_{snp}\} \tag{2.11}$$

N^L_{metab} is the subnetwork of N_{metab} including the lignin related metabolites $m \in L_M$ and their direct neighboring genes:

$$N^L_{metab} = (U^L_{metab}, V^L_{metab}, E^L_{metab}) \text{ where} \tag{2.12}$$

$$U^L_{metab} = \{m|m \in (L_M \cap U_{metab})\} \tag{2.13}$$

$$V^L_{metab} = \{g|\,(g \in V_{metab}) \wedge (\exists m \in L_M|(m,g) \in E_{metab})\} \tag{2.14}$$

$$E^L_{metab} = \{e = (i,j) \in E_{metab}|i \in U^L_{metab} \wedge j \in V^L_{metab}\} \tag{2.15}$$

N^L_{pymbms} is the subnetwork of N_{pymbms} including the lignin related pyMBMS peaks $p \in L_P$ and their direct neighboring genes:

$$N^L_{pymbms} = (U^L_{pymbms}, V^L_{pymbms}, E^L_{pymbms}) \text{ where} \tag{2.16}$$

$$U^L_{pymbms} = \{p|p \in (L_P \cap U_{pymbms})\} \tag{2.17}$$

$$V^L_{pymbms} = \{g | (g \in V_{pymbms}) \wedge (\exists p \in L_P | (p, g) \in E_{pymbms})\} \tag{2.18}$$

$$E^L_{pymbms} = \{e = (i, j) \in E_{pymbms} | i \in U^L_{pymbms} \wedge j \in V^L_{pymbms}\} \tag{2.19}$$

2.2.12 Calculating LOE Scores

For a given gene g, the *degree* of that gene $D(g)$ indicates the number of connections that the gene has in a given network. Let $D_{coex}(g)$, $D_{cometh}(g)$, $D_{snp}(g)$, $D_{metab}(g)$, $D_{pymbms}(g)$ represent the degrees of gene g in the lignin subnetworks N^L_{coex}, N^L_{cometh}, N^L_{snp}, N^L_{metab} and N^L_{pymbms}, respectively. The LOE *breadth* score $LOE_{breadth}(g)$ is then defined as

$$LOE_{breadth}(g) = \text{bin}\,(D_{coex}(g)) + \text{bin}\,(D_{cometh}(g)) + \text{bin}\,(D_{snp}(g)) + \text{bin}\,(D_{metab}(g)) + \text{bin}\,(D_{pymbms}(g)) \tag{2.20}$$

where

$$\text{bin}(x) = \begin{cases} 1 \text{ if } x \geq 1 \\ 0 \text{ otherwise} \end{cases} \tag{2.21}$$

The $LOE_{breadth}(g)$ score indicates the number of different types of lines of evidence that exist linking gene g to lignin-related genes/phenotypes.

The LOE *depth* score $LOE_{depth}(g)$ represents the total number of lines of evidence exist linking gene g to lignin-related genes/phenotypes, and is defined as

$$LOE_{depth}(g) = D_{coex}(g) + D_{cometh}(g) + D_{snp}(g) + D_{metab}(g) + D_{pymbms}(g) \tag{2.22}$$

The GWAS LOE score $LOE_{gwas}(g)$ indicates the number of lignin-related phenotypes (metabolomic or pyMBMS) that a gene is connected to, and is defined as:

$$LOE_{gwas}(g) = D_{metab}(g) + D_{pymbms}(g) \tag{2.23}$$

Distributions of the LOE scores can be seen in Supplementary Figure S2.2. Cytoscape version 3.4.0 [54] was used for network visualization. Expression, methylation, SNP correlation and GWAS diagrams were created using R [55] and various R libraries [56, 57, 58, 59, 60]. Data parsing, wrappers and LOE score calculation was performed using Perl. Diagrams were edited to overlay certain text using Microsoft PowerPoint.

2.3 Results and Discussion

2.3.1 Layered Networks, LOE Scores and New Potential Targets

This study involved the construction of a set of networks providing different layers of information about the relationships between genes, and between genes and phenotypes, and the development of a Lines Of Evidence scoring system (LOE scores) which integrate the information in the different network layers and quantify the number of lines of evidence connecting genes to lignin-related genes/phenotypes. The GWAS network layers provide information as to which genes are potentially involved in certain functions because they contain genomic variants significantly associated with measured phenotypes. The co-methylation and co-expression networks provide information on different layers of regulatory mechanisms within the cell. The SNP correlation network provides information about possible co-evolution relationships between genes, through correlated variants across a population.

Marking known genes and phenotypes involved in lignin biosynthesis in these networks allowed for the calculation of a set of LOE (Lines Of Evidence) scores for each gene, indicating the strength of the evidence linking each gene to lignin-related functions. The breadth LOE score indicates the number of types of lines of evidence (number of layers) which connect the gene to lignin-related genes/phenotypes, whereas the depth LOE score indicates the total number of lignin-related genes/phenotypes the gene is associated with. Individual layer LOE scores (e.g. co-expression LOE score or GWAS LOE score) indicate the number of lignin-related associations the gene has within that layer.

This data layering approach differs from previous data integration methods. Mizrachi *et al.* (2017) integrate gene expression data with eQTN data and gene relationships from KEGG though matrix multiplication, before correlating genes' NBDI (Network Based Data Integration)-transformed values with measured traits, allowing the ranking of genes [61]. The Mergeomics method [62] performs Marker Set Enrichment Analysis, ranking predefined sets of molecular markers based on their enrichment in a disease phenotype. Knetminer [63, 64] is a web server which allows the user to search for keywords, producing lists of genes and the associations they have to annotations, genes, phenotypes, publications etc. which match the keywords and that are available in public databases. Knetminer can also produce a network view of the results. While Knetminer is also an approach which utilizes multiple lines of evidence, the main approach and the scoring systems differ. LOE requires input lists of genes and phenotypes of interest to the user, Knetminer uses gene lists and keyword searching. In terms of lines of evidence, Knetminer counts the number of "concepts" (nodes, including publications, phenotypes, annotations etc) a gene has linking it to a keyword [63, 64]. However, LOE scores (particularly, breadth LOE scores) count the number of types of relationships (e.g. GWAS association, co-expression, co-methylation, variant correlation *edges*) connecting a gene to specific input genes and phenotypes related to the user's function of interest. This is thus a valuable approach to identify new target genes based on the *relationships* of a gene to target genes/phenotypes of interest in custom-made association network layers where publically available data is not available.

To select the top set of potential new candidate genes involved in lignin biosynthesis, genes which showed a number of different lines of evidence connecting them to lignin-related functions were identified by selecting genes with a LOE breadth score $>= 3$. Since the GWAS networks provide the highest resolution, most direct connections to lignin-related functions, it was also required that our potential new targets had a GWAS score $>= 1$.

This provides a set of 375 new candidate genes potentially involved in lignin biosynthesis, identified through multiple lines of evidence (Supplementary Table S2.4). This set of Potential New Target genes will be referred to as set of PNTs. A selection of these potential new candidates below and their annotations, derived from their *Arabidopsis* best hits, will be discussed below.

2.3.2 Agamous-like Genes

Genes in the AGAMOUS-LIKE gene family are MADS-box transcription factors, many of which which have been found to play important roles in floral development [65, 66, 67, 68, 69, 70]. Three potential AGAMOUS-LIKE (AGL) genes are found in the set of PNTs, in particular, a homolog of *Arabidopsis* AGL8 (AT5G60910, also known as FRUITFUL), a homolog of *Arabidopsis* AGL12 (AT1G71692), and a homolog of *Arabidopsis* AGL24 (AT4G24540) and AGL22 (AT2G22540).

The first potential AGL gene in our set of PNTs is Potri.012G062300, with a breadth score of 3 and a GWAS score of 2 (Figure 2.3A), whose best *Arabidopsis thaliana* hit is AGL8 (AT5G60910).

It has GWAS associations with a lignin-related metabolite (quinic acid) and a lignin pyMBMS peak (syringol) (Figure 2.3C, Table 2.1) and is co-methylated with three lignin-related genes (Figure 2.3B, Table 2.2). There is thus strong evidence for the involvement of *P. trichocarpa* AGL8 in the regulation of lignin-related functions. There is literature evidence that supports the hypothesis of AGL8's involvement in the regulation of lignin biosynthesis. A patent exists for the use of AGL8 expression in reducing the lignin content of plants [71]. The role of AGL8 (FUL) was described by Ferrandiz *et al.* (2000) [72], in which they investigated the differences in lignin deposition in transgenic plants in which AGL8 is constitutively expressed, loss-of-function AGL8 mutants and wild-type *Arabidopsis* plants [72]. In wild-type plants, a single layer of valve cells were lignified. In loss-of-function AGL8 mutants, all valve mesophyl cell layers were lignified, while in the transgenic plants, constitutive expression of AGL8 resulted in loss of lignified cells

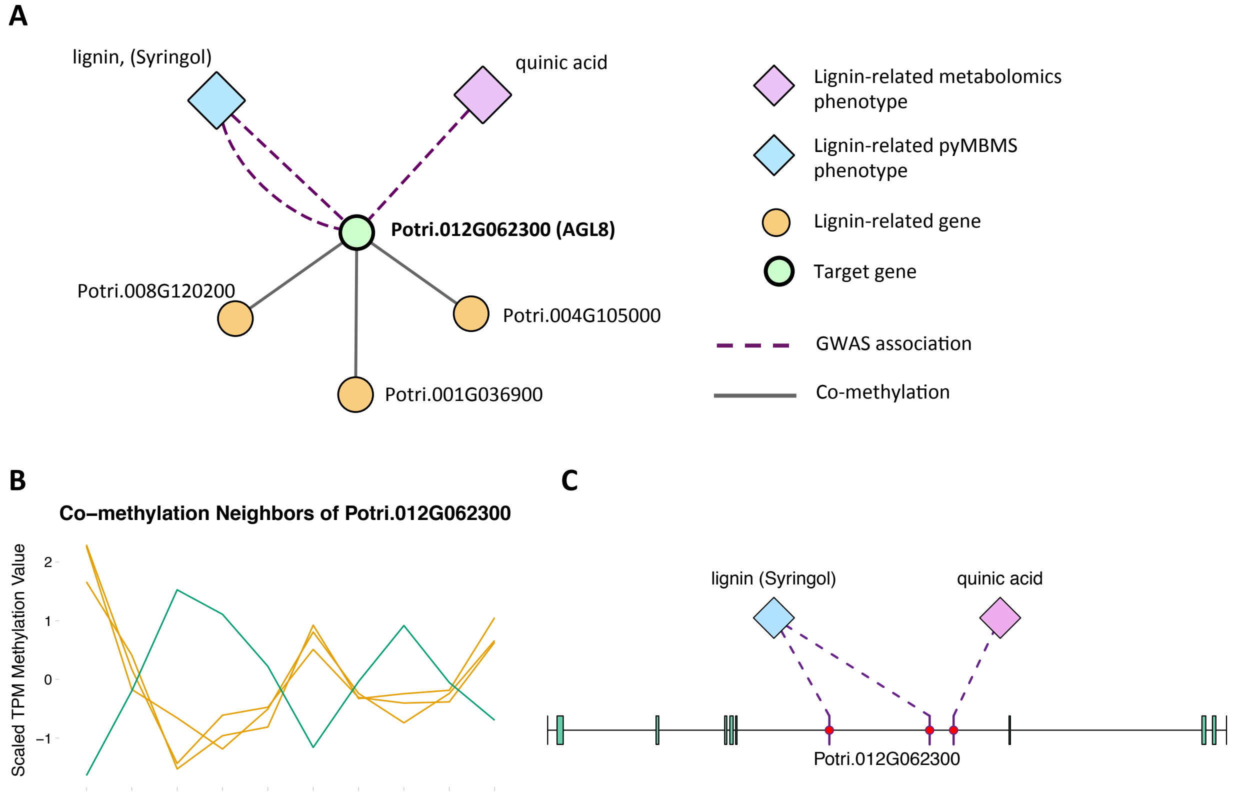


Figure 2.3: **Lines of Evidence for AGL8.** (A) Lines of Evidence for Potri.012G062300 (homolog of *Arabidopsis* AGL8). (B) Co-methylation of Potri.012G062300 with three lignin-related genes (Table 2.2) The green line represents potential target Potri.012G062300 and yellow lines represent lignin-related genes. (C) GWAS associations of Potri.012G062300 with a lignin-related metabolite and a lignin-related pyMBMS peak (Table 2.1).

Table 2.1: GWAS associations for select new potential target genes, indicating the SNP(s) within the potential new target gene which are associated with the lignin-related phenotype(s). Additional RT and mz information for partially identified metabolites can be seen in Supplementary Table S2.3.

Source SNP	**Source Gene**	**Target Phenotype**
GWAS Associations for Potri.012G062300 (AGL8, AT5G60910)		
12:6952245	Potri.012G062300	quinic acid
12:6948543	Potri.012G062300	lignin (Syringol)
12:6951532	Potri.012G062300	lignin (Syringol)
GWAS Associations for Potri.013G102600 (AGL12, AT1G71692)		
13:11604094	Potri.013G102600	3-O-caffeoyl-quinate
13:11606331	Potri.013G102600	coumaroyl-tremuloidin
13:11600422	Potri.013G102600	coumaroyl-tremuloidin
13:11601236	Potri.013G102600	hydroxyphenyl lignan glycoside
GWAS Associations for Potri.007G115100 (AGL22, AT2G22540/AGL24, AT4G24540)		
07:13650194	Potri.007G115100	caffeoyl conjugate
07:13651354	Potri.007G115100	caffeoyl conjugate
07:13642539	Potri.007G115100	caffeoyl conjugate
07:13639923	Potri.007G115100	lignin, syringyl (Syringaldehyde)
GWAS Associations for Potri.009G053900 (MYB46, AT5G12870)		
09:5768381	Potri.009G053900	hydroxyphenyl lignan glycoside
GWAS Associations for Potri.010G141000 (MYB111, AT5G49330)		
10:15273000	Potri.010G141000	benzoyl-salicylate caffeic acid conjugate

Continued on next page.

Table 2.1: (continued)

Source SNP	**Source Gene**	**Target Phenotype**
GWAS Associations for Potri.006G170800 (MYB36, AT5G57620)		
06:17847162	Potri.006G170800	mz 297, RT 17.14

GWAS Associations for Potri.016G078600 (CPSRP54, AT5G03940)		
16:5995136	Potri.016G078600	caffeoyl conjugate
16:5995136	Potri.016G078600	feruloyl conjugate
16:5996083	Potri.016G078600	salicyl-coumaroyl-glucoside
16:5999408	Potri.016G078600	salicyl-coumaroyl-glucoside
16:5999474	Potri.016G078600	salicyl-coumaroyl-glucoside
16:6000236	Potri.016G078600	salicyl-coumaroyl-glucoside

Table 2.2: Co-methylation associations for select new potential target genes. Annotations are derived from best *Arabidopsis* hit descriptions and GO terms and in some cases MapMan annotations.

Target Gene	**Target *Arabidopsis* best hit**	**Annotation**
Co-methylation Associations for Potri.012G062300 (AGL8, AT5G60910)		
Potri.001G036900	AT3G21240	4-coumarate:CoA ligase 2
Potri.008G120200	AT1G68540	Cinnamoyl CoA reductase-like 6
Potri.004G105000	AT5G14700	(NAD(P)-binding Rossmann-fold superfamily protein, cinnamoyl-CoA reductase activity/CCR1
Co-methylation Associations for Potri.013G102600 (AGL12, AT1G71692)		
Potri.001G334400	AT5G63380	4-coumarate-CoA ligase activity /4CL
Potri.001G365300	AT3G26300	cytochrome P450, family 71, subfamily B, polypeptide 34/F5H
Potri.006G265500	AT5G10820	Major facilitator superfamily protein/Phenylpropanoid pathway
Potri.006G165200	AT2G19070	spermidine hydroxycinnamoyl transferase
Co-methylation Associations for Potri.009G053900 (MYB46, AT5G12870)		
Potri.008G196100	AT3G06350	bi-functional dehydroquinate-shikimate dehydrogenase enzyme
Potri.002G018300	AT4G39330	cinnamyl alcohol dehydrogenase 9
Potri.004G102000	AT4G05160	4-coumarate-CoA ligase activity/4CL)
Potri.008G136600	AT1G67980	caffeoyl-CoA 3-O-methyltransferase

Continued on next page.

Table 2.2: (continued)

Target Gene	**Target *Arabidopsis* best hit**	**Annotation**
Co-methylation Associations for Potri.010G141000 (MYB111, AT5G49330)		
Potri.008G196100	AT3G06350	bi-functional dehydroquinate-shikimate dehydrogenase enzyme
Potri.004G102000	AT4G05160	4-coumarate-CoA ligase activity/4CL
Potri.008G074500	AT5G34930	arogenate dehydrogenase
Potri.005G028000	AT5G48930	hydroxycinnamoyl-CoA shikimate/quinate hydroxycinnamoyl transferase
Potri.018G100500	AT2G23910	NAD(P)-binding Rossmann-fold superfamily protein, cinnamoyl-CoA reductase activity/CCR1
Potri.010G230200	AT1G20510	OPC-8:0 CoA ligase1, 4-coumarate-CoA ligase activity/4CL
Co-methylation Associations for Potri.006G170800 (MYB36, AT5G57620)		
Potri.016G093700	AT4G05160	AMP-dependent synthetase and ligase family, 4-coumarate-CoA ligase activity/4CL
Co-methylation Associations for Potri.016G078600 (CPSRP54, AT5G03940)		
Potri.014G135500	AT3G06350	bi-functional dehydroquinate-shikimate dehydrogenase enzyme

[72]. This study thus showed the involvement of AGL8 in fruit lignification during fruit development.

There is evidence of other AGAMOUS-LIKE genes affecting lignin content. A study by Gimenez *et al.* (2010) investigated TALG1, an AGAMOUS-LIKE gene in tomato, and found that TAGL1 RNAi-silenced fruits showed increased lignin content, and increased expression levels of lignin biosynthesis genes [73]. A recent study by Cosio *et al.* (2017) showed that AGL15 in *Arabidopsis* is also involved in regulating lignin-related functions, in that AGL15 binds to the promotor of peroxidase PRX17, and regulates its expression [74]. In addition, PRX17 loss of function mutants had reduced lignin content [74].

There is thus compelling evidence that various AGAMOUS-LIKE genes are involved in regulating lignin biosynthesis/deposition in plants. Two other AGAMOUS-like genes are seen in the set of PNTs, namely a homolog of *Arabidopsis* AGL12 (Potri.013G102600) and a homolog of *Arabidopsis* AGL22/AGL24 (Potri.007G115100). Potri.013G102600 (AGL12) has GWAS associations with three lignin-related metabolites, namely hydroxyphenyl lignan glycoside, coumaroyl-tremuloidin and 3-O-caffeoyl-quinate (Figure 2.4A, Figure 2.4B, Table 2.1). It is co-expressed with four lignin-related genes including two caffeoyl coenzyme A O-methyltransferases, a caffeate O-methyltransferase and a ferulic acid 5-hydroxylase (Figure 2.4A, Figure 2.4C, Table 2.3) and it is co-methylated with four other lignin-related genes (Figure 2.4A, Figure 2.4D, Table 2.2). Potri.007G115100 (AGL22/AGL24) has GWAS associations with the syringaldehyde pyMBMS phenotype and a caffeoyl conjugate metabolite (Figure 2.5A, Figure 2.5B, Table 2.1). It also has SNP correlations with a laccase and a nicotinamidase (Figure 2.5A, Figure 2.5C, Figure 2.5D, Table 2.4, Supplementary Table S2.5). The combination of the multiple lines of multi-omic evidence thus suggest the involvement of *P. trichocarpa* homologs of *A. thaliana* AGL22/AGL24 and AGL12 in regulating lignin biosynthesis.

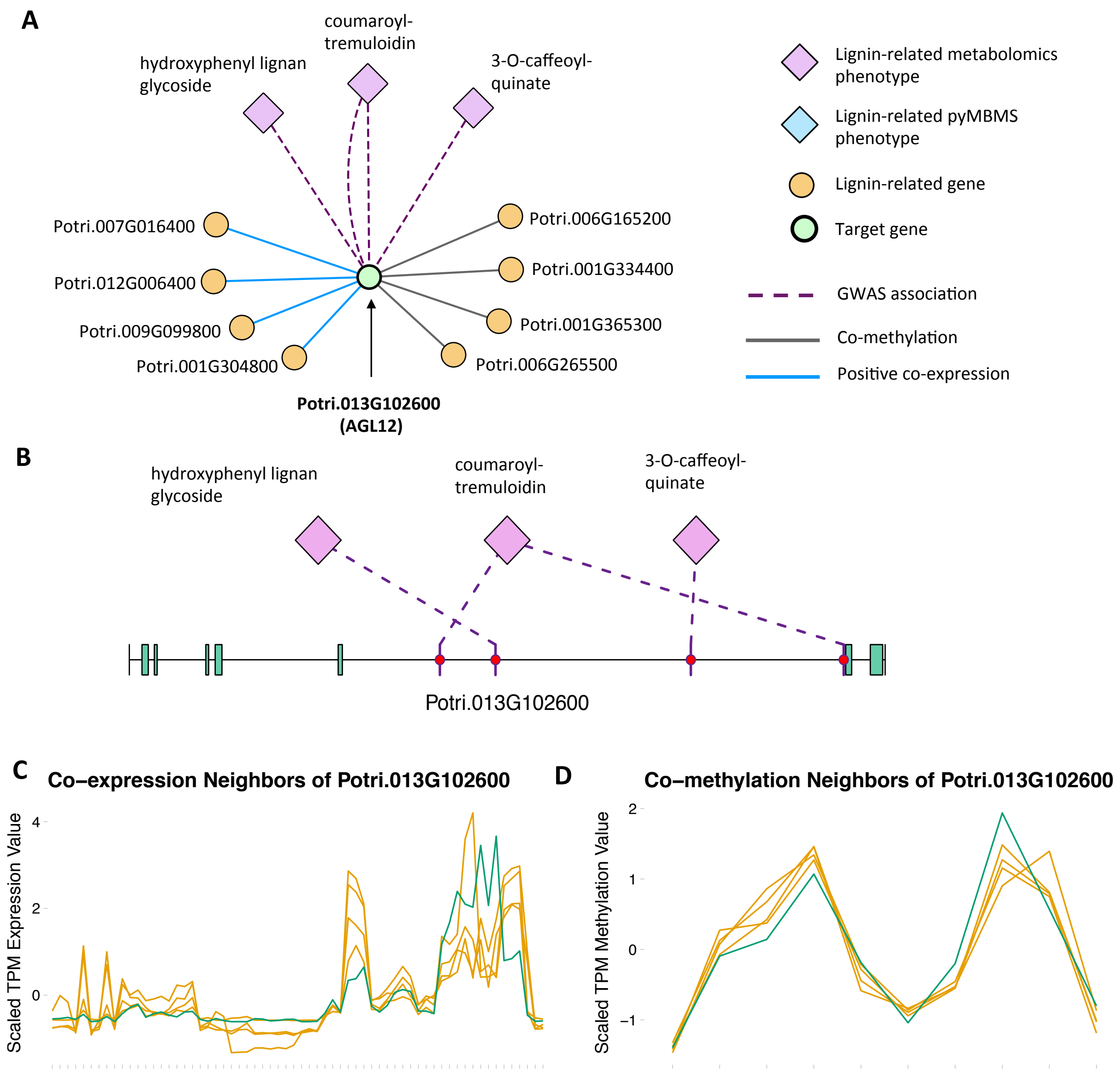


Figure 2.4: **Lines of Evidence for AGL12.** (A) Lines of Evidence for Potri.013G102600 (homolog of *Arabidopsis* AGL12). (B) GWAS associations of Potri.013G102600 with three lignin-related metabolites (Table 2.1). (C) Co-expression of Potri.013G102600 with three lignin-related genes (Table 2.3). (D) Co-methylation of Potri.013G102600 with four lignin-related genes (Table 2.2). In line plots, the green lines represent potential target Potri.013G102600 and yellow lines represent lignin-related genes.

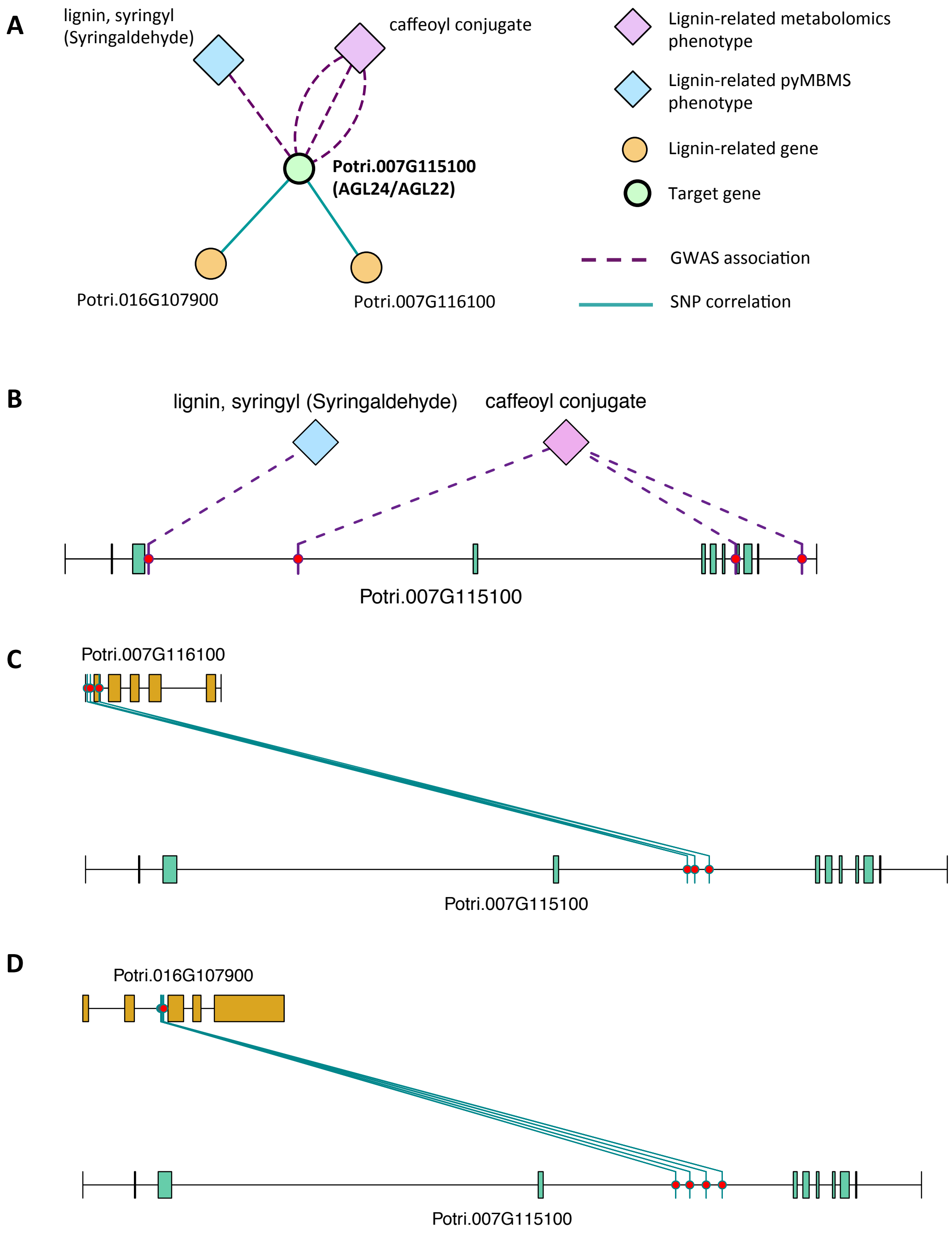


Figure 2.5: **Lines of Evidence for AGL22/24.** (A) Lines of Evidence for Potri.007G115100 (homolog of *Arabidopsis* AGL22/24). (B) GWAS associations of Potri.007G115100 with a lignin-related metabolite and a lignin-related pyMBMS peak (Table 2.1). (C,D) Correlations of SNPs in Potri.007G115100 with SNPs in two lignin-related genes (Table 2.4, Supplementary Table S2.5).

Table 2.3: Co-expression associations for select new potential target genes. Annotations are derived from best *Arabidopsis* hit descriptions and GO terms and in some cases MapMan annotations.

Target Gene	**Target *Arabidopsis* best hit**	**Annotation**
Co-expression Associations for Potri.013G102600 (AGL12, AT1G71692)		
Potri.001G304800	AT4G34050	Caffeoyl Coenzyme A O-Methyltransferase 1
Potri.009G099800	AT4G34050	Caffeoyl Coenzyme A O-Methyltransferase 1
Potri.012G006400	AT5G54160	Caffeate O-Methyltransferase 1
Potri.007G016400	AT4G36220	Ferulic acid 5-hydroxylase 1
Co-expression Associations for Potri.009G053900 (MYB46, AT5G12870)		
Potri.003G100200	AT1G32100	pinoresinol reductase 1
Potri.012G006400	AT5G54160	Caffeate O-Methyltransferase 1
Co-expression Associations for Potri.010G141000 (MYB111, AT5G49330)		
Potri.007G030300	AT3G50740	UDP-glucosyl transferase 72E1
Co-expression Associations for Potri.006G170800 (MYB36, AT5G57620)		
Potri.001G362800	AT3G26300	cytochrome P450, family 71, subfamily B, polypeptide 34/F5H
Potri.016G106100	AT3G09220	laccase 7
Potri.013G120900	AT4G35160	N-acetylserotonin O-methyltransferase
Co-expression Associations for Potri.016G078600 (CPSRP54, AT5G03940)		
Potri.003G096600	AT2G35500	shikimate kinase like 2
Potri.017G062800	AT3G26900	shikimate kinase like 1

Table 2.4: SNP correlation associations for select new potential target genes. Annotations are derived from best *Arabidopsis* hit descriptions and GO terms and in some cases MapMan annotations.

Target gene	**Target *Arabidopsis* best hit**	**Annotation**
SNP Correlations for Potri.007G115100 (AGL22, AT2G22540/AGL24, AT4G24540)		
Potri.007G116100	AT2G22570	nicotinamidase 1
Potri.016G107900	AT3G09220	laccase 7
SNP Correlations for Potri.016G078600 (CPSRP54, AT5G03940)		
Potri.016G078300	AT4G37970	cinnamyl alcohol dehydrogenase 6

2.3.3 MYB Transcription Factors

MYB proteins contain the conserved MYB DNA-binding domain, and usually function as transcription factors. R2R3-MYBs have been found to regulate various functions, including flavonol biosynthesis, anthocyanin biosynthesis, lignin biosynthesis, cell fate and developmental functions [75]. The set of PNTs contains several genes which are homologs of *Arabidopsis* MYB transcription factors, including homologs of *Arabidopsis* MYB66/MYB3, MYB46, MYB36 and MYB111.

There is already existing literature evidence for how some of these MYBs affect lignin biosynthesis. Liu *et al.* (2015) [76] review the involvement of MYB transcription factors in the regulation of phenylpropanoid metabolism. MYB3 in *Arabidopsis* is known to repress phenylpropanoid biosynthesis [77], and a *P. trichocarpa* homolog of MYB3 is found in our set of potential new targets. Another potential new target is the *P. trichocarpa* homolog of *Arabidopsis* MYB36 (Potri.006G170800) which is connected to lignin-related functions through multiple lines of evidence (Figure 2.6). In *Arabidopsis*, MYB36 has been found to regulate the local deposition of lignin during casparian strip formation, and *myb36* mutants exhibit incorrectly localized lignin deposition [78].

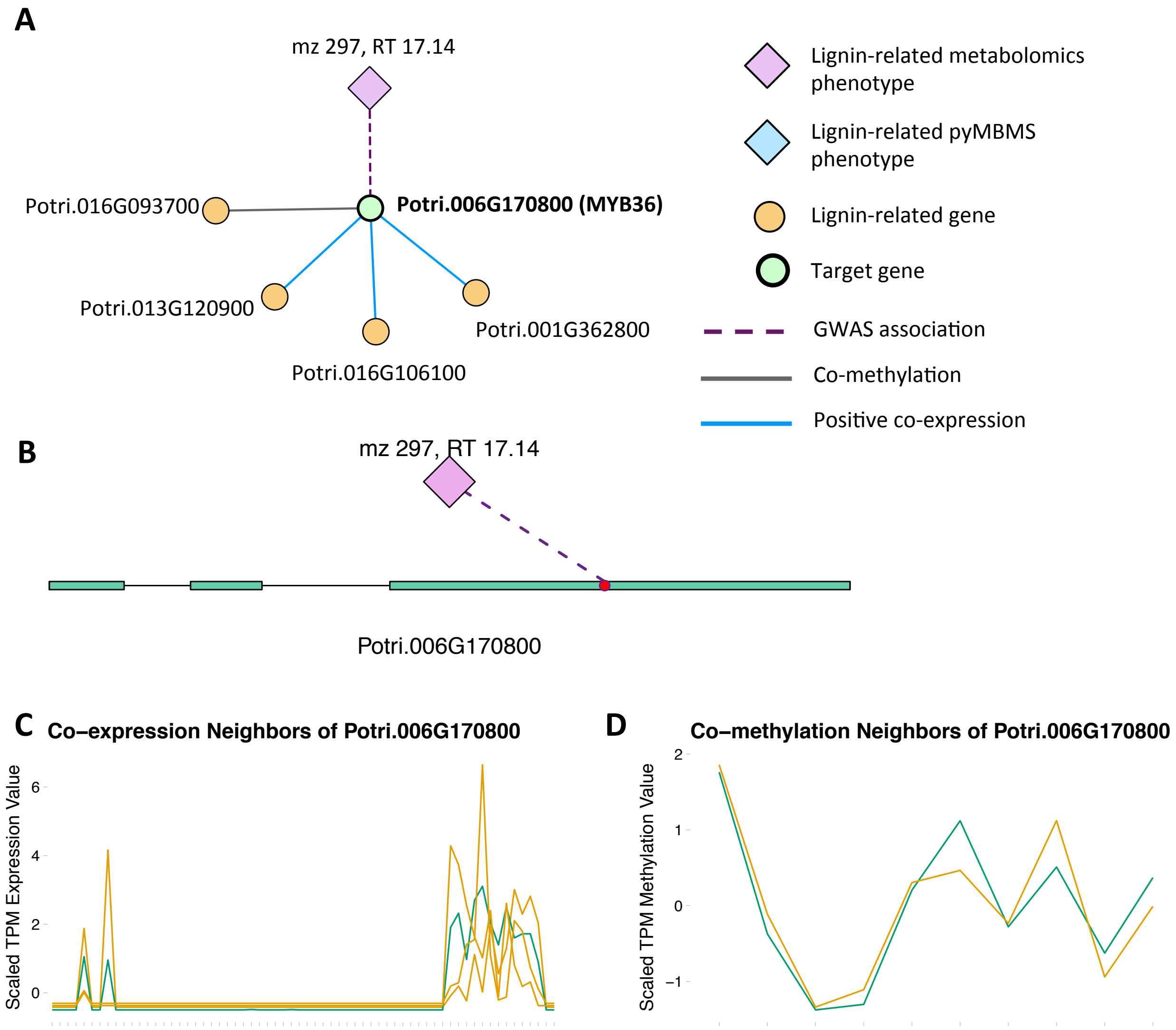


Figure 2.6: **Lines of Evidence for MYB36.** (A) Lines of Evidence for Potri.006G170800 (homolog of *Arabidopsis* MYB36). (B) GWAS associations of Potri.006G170800 with a lignin-related metabolite (Table 2.1). (C) Co-expression of Potri.006G170800 with three lignin-related genes (Table 2.3). (D) Co-methylation of Potri.006G170800 with a lignin-related gene (Table 2.2). In line plots, the green lines represent potential target Potri.006G170800 and yellow lines represent lignin-related genes.

MYB46 is known to be a regulator of secondary cell wall formation [79]. Overexpression of MYB46 in *Arabidopsis* activates lignin, cellulose and xylan biosynthesis pathways [79]. The MYB46 homolog in *P. trichocarpa*, Potri.009G053900, is connected to lignin-related functions through multiple lines of evidence (Figure 2.7A), including a GWAS association with a hydroxyphenyl lignan glycoside (Figure 2.7E, Table 2.1), co-expression with pinoresinol reductase 1 and caffeate O-methyltransferase 1 (Figure 2.7F, Table 2.3) and co-methylation with dehydroquinate-shikimate dehydrogenase enzyme, cinnamyl alcohol dehydrogenase 9, 4-coumarate-CoA ligase activity/4CL) and caffeoyl-CoA 3-O-methyltransferase (Figure 2.7G, Table 2.2).

A MYB transcription factor in the set of PNTs which has, to our knowledge, not yet been directly associated with lignin biosynthesis is MYB111 (Figure 2.7A-D). However, with existing literature evidence, one can hypothesize that MYB111 can alter lignin content by redirecting carbon flux from flavonoids to monolignols. There is evidence that MYB111 is involved in crosstalk between lignin and flavonoid pathways. Monolignols and flavonoids are both derived from phenylalanine through the phenylpropanoid pathway [76]. There is crosstalk between the signalling pathways of ultraviolet-B (UV-B) stress and biotic stress pathways [80]. In the study by Schenke *et al.* (2011), it was shown that under UV-B light stress, *Arabidopsis* plants produce flavonols as a UV protectant [80]. Also, simultaniously applying the bacterial elicitor flg22, which simulates biotic stress, repressed flavonol biosynthesis genes and induced production of defense compounds including camalexin and scopoletin, as well as lignin, which provides a physical barrier preventing pathogens' entry [80]. This crosstalk involved regulation by MYB12 and MYB4 [80]. This study by Schenke *et al.* (2011) was performed using cell cultures. A second study by Zhou *et al.* (2017) used *Arabidopsis* seedlings, and found that MYB111 may be involved in the crosstalk in planta [81]. The multiple lines of evidence connecting the *P. trichocarpa* homolog of *Arabidopsis* MYB111 (Potri.010G141000) to lignin related functions, in combination with the above literature evidence suggests the involvement this gene in the regulation of lignin biosynthesis by redirecting carbon flux from flavonol biosynthesis to monolignol biosynthesis, as part of the crosstalk between UV-B protection and biotic stress signalling pathways.

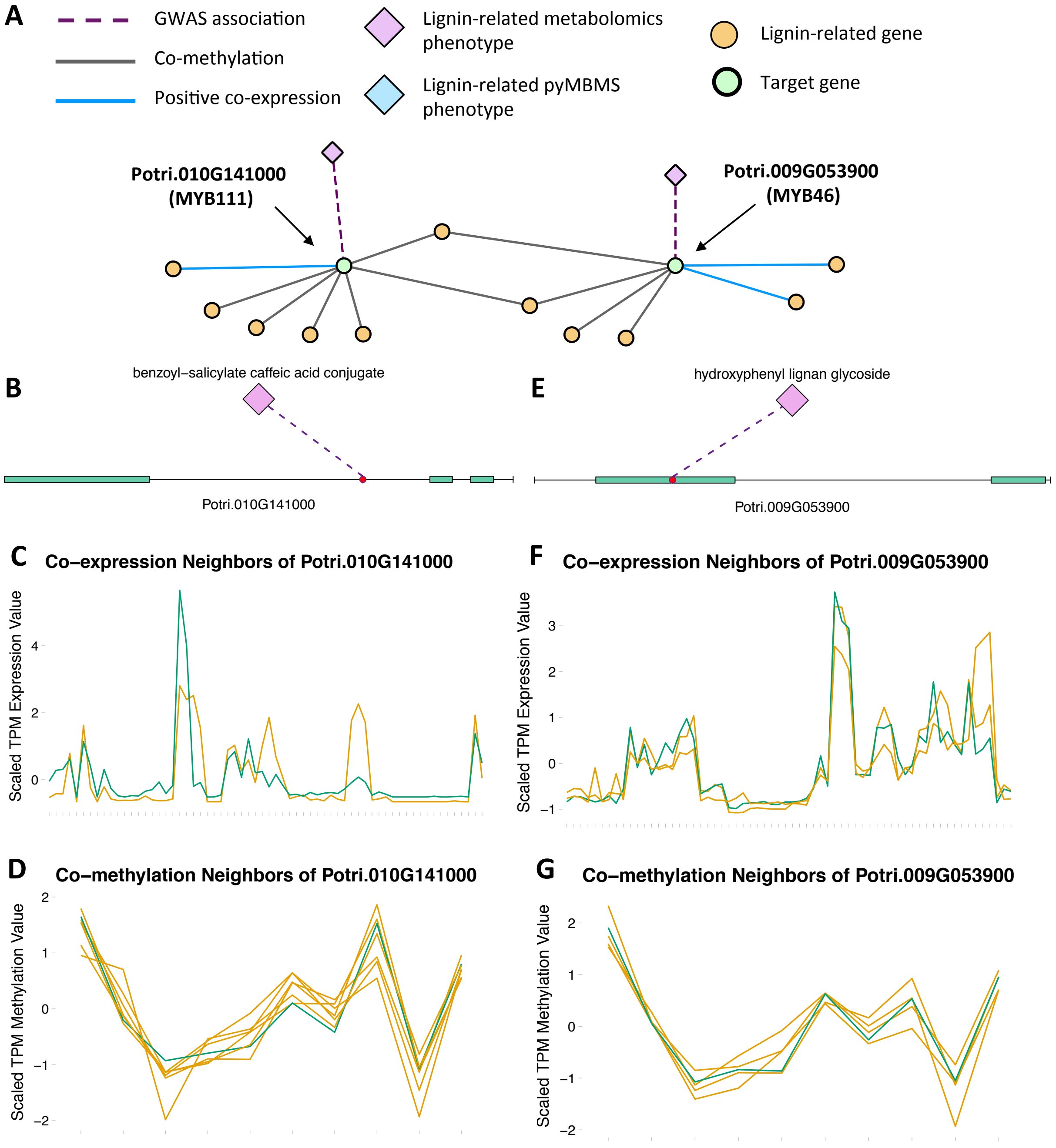


Figure 2.7: **Lines of Evidence for MYB46 and MYB111.** (A) Lines of Evidence for Potri.009G053900 (homolog of *Arabidopsis* MYB46) and Potri.010G141000 (homolog of *Arabidopsis* MYB111). (B) GWAS associations of Potri.010G141000 with a lignin-related metabolite (Table 2.1). (C) Co-expression of Potri.010G141000 with a lignin-related gene (Table 2.3). (D) Co-methylation of Potri.010G141000 with six lignin-related genes (Table 2.2). (E) GWAS associations of Potri.009G053900 with a lignin-related metabolite (Table 2.1). (F) Co-expression of Potri.009G053900 with two lignin-related genes (Table 2.3). (G) Co-methylation of Potri.009G053900 with four lignin-related genes (Table 2.2). In line plots, the green lines represent potential targets Potri.009G053900/Potri.010G141000 and yellow lines represent lignin-related genes.

2.3.4 Chloroplast Signal Recognition Particle

Potri.016G078600, a homolog of the *Arabidopsis* chloropast signal recognition particle cpSRP54 occurs in the set of PNTs (Figure 2.8). It has a GWAS LOE score of 3, through GWAS associations with salicyl-coumaroyl-glucoside, a caffeoyl conjugate and a feruloyl conjugate (Figure 2.8B, Table 2.1, Supplementary Table S2.4). It also has a breadth score of 4, indicating that it is linked to lignin-related genes/phenotypes though 4 different types of associations (Figure 2.8). CpSRP54 gene has been found to regulate carotenoid accumulation in *Arabidopsis* [82]. CpSRP54 and cpSRP43 form a "transit complex" along with a light-harvesting chlorophyll a/b-binding protein (LHCP) family member to transport it to the thylakoid membrane [83, 84]. A study in *Arabidopsis* found that cpSRP43 mutants had reduced lignin content [85]. Since CpSRP54 regulates carotenoid accumulation, and cpSRP43 appears to affect lignin content, it is possible that chloroplast signal recognition particles affect lignin and carotenoid content through flux through the phenylpropanoid pathway, the common origin of both of these compounds. In fact, a gene mutation *cue1* which causes LHCP underexpression also results in reduced aromatic amino acid biosynthesis [86]. These multiple lines of evidence, combined with the above cited literature suggests that chloroplast signal recognition particles in *P. trichocarpa* could potentially influence lignin content.

2.3.5 Practical Implications

The LOE method of data integration provides a useful way for biologists to identify new target genes. Any genes and phenotypes of interest that are present in the networks can be used as input to the method, and thus, the results can be tailored to the particular function of interest of the biologist. The collection of LOE scores will allow the user to rank genes in the genome based on the particular lines of evidence most appropriate to function under investigation, and in so doing, provides a shortlist of genes as targets for genetic modification (knockout/knockdown/overexpression) in order to alter the phenotype of interest. For example, AGL genes, MYB transcription factors and CpSRP genes discussed

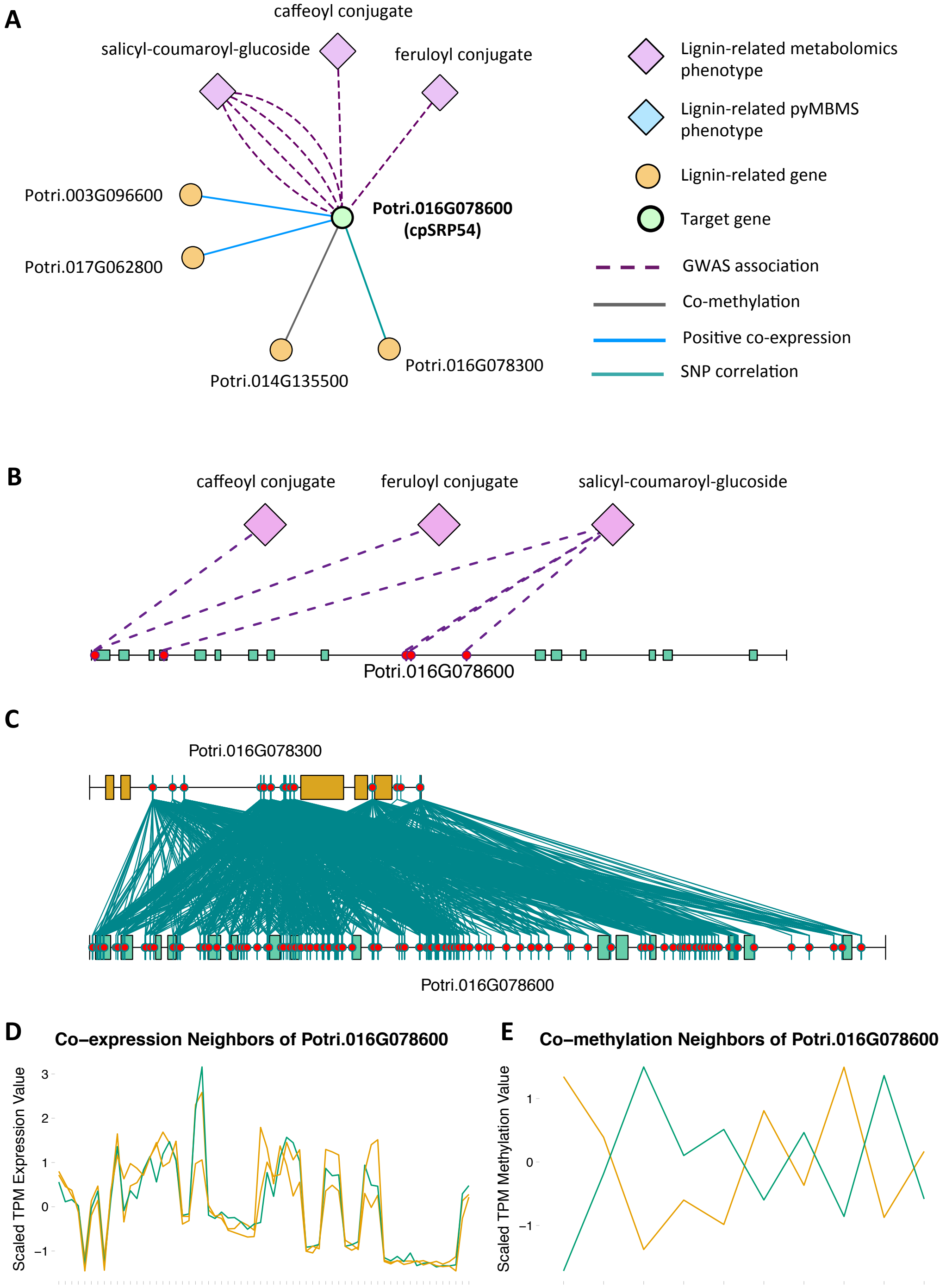


Figure 2.8: **Lines of Evidence for cpSRP54.** (A) Lines of Evidence for Potri.016G078600 (homolog of *Arabidopsis* cpSRP54). (B) GWAS associations of Potri.016G078600 with three lignin-related metabolite (Table 2.1). (C) Correlations of SNPs within Potri.016G078600 with SNPs in a lignin-related gene (Table 2.4). (D) Co-expression of Potri.016G078600 with two lignin-related genes (Table 2.3). (E) Co-methylation of Potri.016G078600 with a lignin-related gene (Table 2.2). In line plots, the green lines represent potential target Potri.016G078600 and yellow lines represent lignin-related genes.

above could be seen as potential new targets for knockout/knockdown/overexpression in order to alter the lignin content of *P. trichocarpa*.

The LOE scoring method can be applied to any species for which there is multiple data types that can be represented as association networks which the scientist wishes to integrate in order to identify new candidate genes involved in a particular function. This method will be particularly useful for the analysis of new, unpublished datasets where publically available datasets/web servers would not necessarily be able to be used.

2.3.6 Concluding Remarks

This study made use of high-resolution GWAS data, combined with co-expression, co-methylation and SNP correlation networks in a multi-omic, data layering approach which has allowed the identification of new potential target genes involved in lignin biosynthesis/regulation. Various literature evidence supports the involvement of many of these new target genes in lignin biosynthesis/regulation, and these are suggested for future validation for involvement in the regulation of lignin biosynthesis. The data layering technique and LOE scoring system developed can be applied to other omic data types to assist in the generation of new hypotheses surrounding various functions of interest.

2.4 Funding

Funding provided by The BioEnergy Science (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.

This research was also supported by the Department of Energy Laboratory Directed Research and Development funding (7758), at the Oak Ridge National Laboratory. Oak Ridge National Laboratory is managed by UT-Battelle, LLC, for the US DOE under contract DE-AC05-00OR22725.

This research used resources of the Oak Ridge Leadership Computing Facility (OLCF) and 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 is provided by the U.S. Department of Energy, Office of Science Biological and Environmental Research (BER) via the Bioenergy Science Center (BESC) under Contract No. DE-PS02-06ER64304. 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. Full Gene Atlas data sets are available at http://phytozome.jgi.doe.gov.

2.5 Supplementary Material

2.5.1 Text S2.1: Constructing Samples CCC Distribution

Printing out the complete result set of all possible pairwise comparisons of ∼10,000,000 SNPs would require more disk space than was possibly available. In order to construct an approximate distribution of the CCC values, we selected a random subset of 100,000 SNPs and calculated the CCC correlation between all pairs of these SNPs, storing all correlation values. This sampled set of correlations was used to compute the CCC distribution. Thereafter, the CCC was calculated between all pairs of all ∼10,000,000 SNPs. Only correlations meeting a threshold of 0.7 were stored.

2.5.2 Supplementary Figures

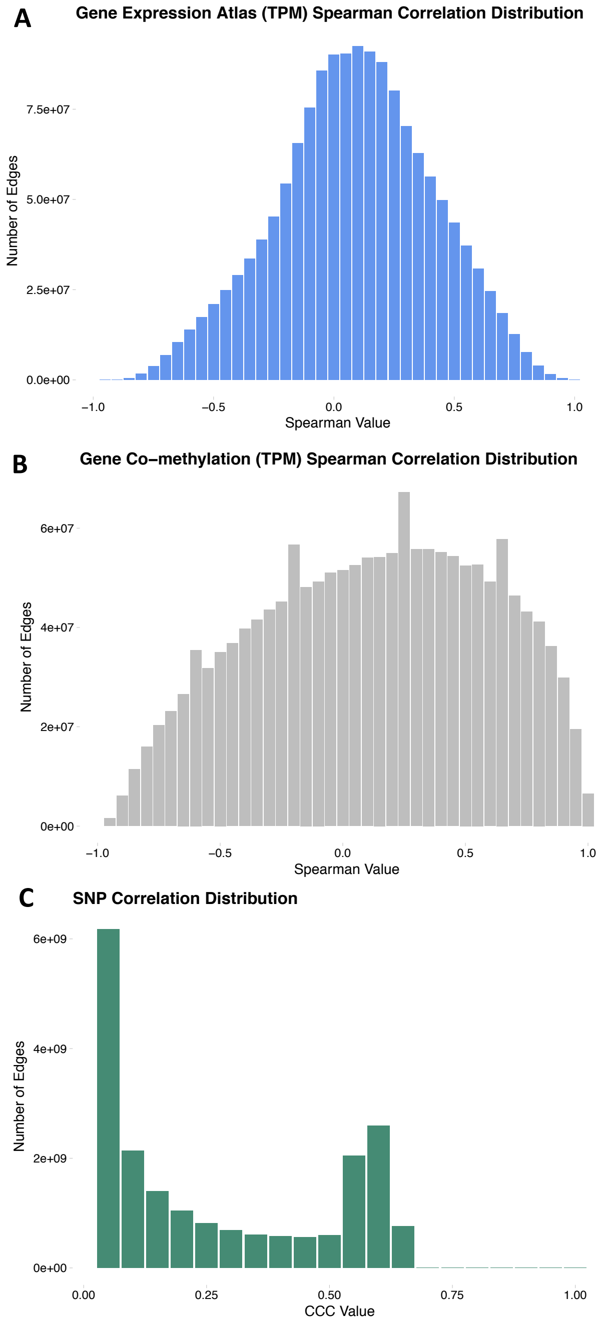
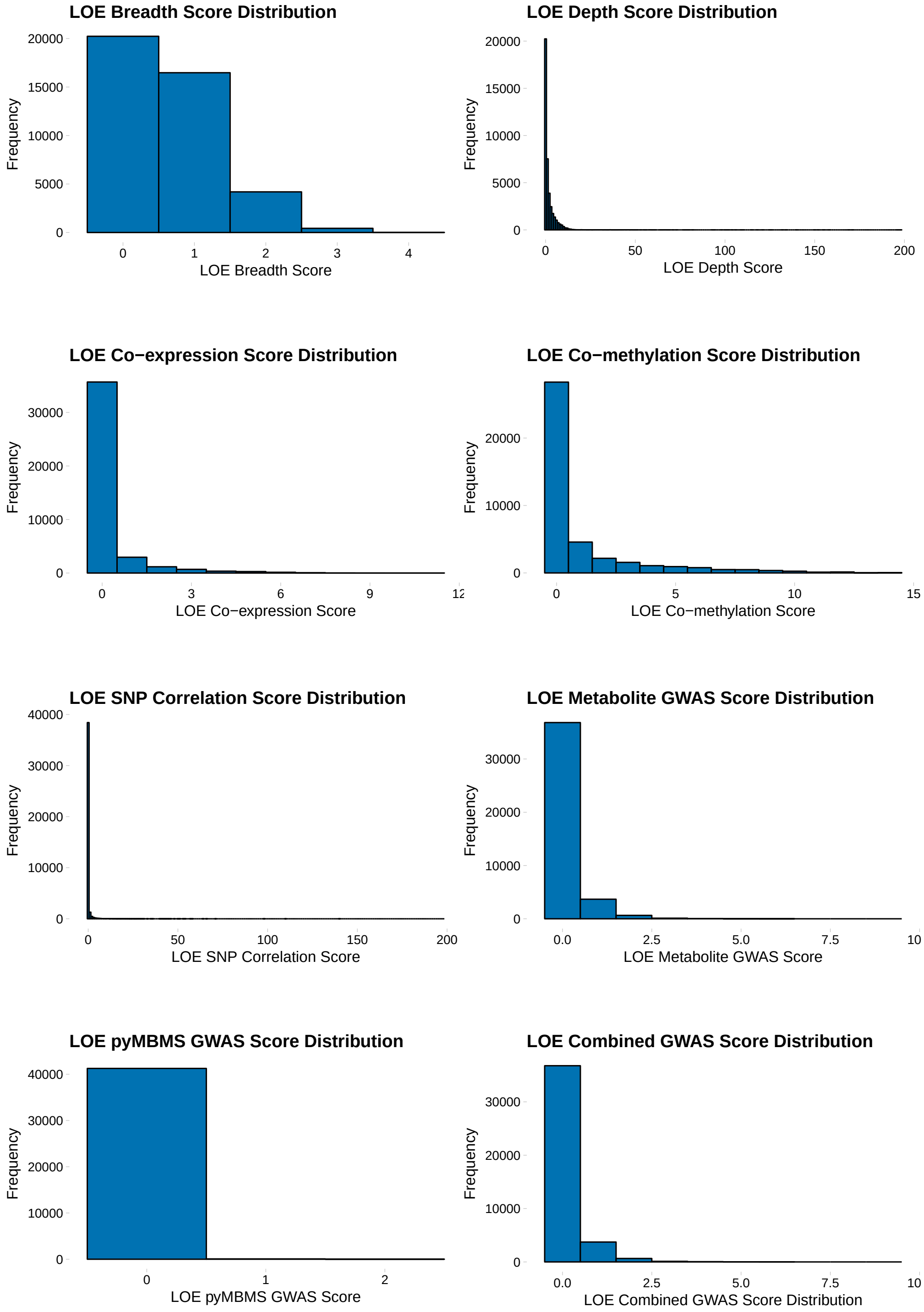


Figure S2.1: **Edge weight distributions.** (A) Distribution of Spearman Correlation values in the co-expression network. (B) Distribution of Spearman Correlation values in the co-methylation network. (C) Sampled distribution of the CCC SNP correlation network. See Supplementary Text S2.1 for details on the construction of the sampled distribution.



LOE Breadth Score Distribution

LOE Depth Score Distribution

LOE Co−expression Score Distribution

LOE Co−methylation Score Distribution

LOE SNP Correlation Score Distribution

LOE Metabolite GWAS Score Distribution

LOE pyMBMS GWAS Score Distribution

LOE Combined GWAS Score Distribution

Figure S2.2: **Score distributions.** Distributions of the various categories of Lines Of Evidence (LOE) score.

2.5.3 Supplementary Tables

Table S2.1: MapMan annotations of lignin genes.

Gene	Mapman Name
Potri.001G133200.v3.0	secondary metabolism.flavonoids.isoflavones.isoflavone reductase : secondary metabolism.phenylpropanoids
Potri.003G196700.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.CAD
Potri.012G094900.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.4CL
Potri.001G372400.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.CAD
Potri.006G097500.v3.0	secondary metabolism.phenylpropanoids : secondary metabolism.flavonoids.anthocyanins
Potri.T178300.v3.0	secondary metabolism.phenylpropanoids : secondary metabolism.unspecified
Potri.001G304800.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.CCoAOMT
Potri.001G045000.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.CCR1
Potri.001G045100.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.CCR1
Potri.001G268600.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.CAD
Potri.004G230900.v3.0	secondary metabolism.phenylpropanoids
Potri.007G029800.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis : misc.UDP glucosyl and glucoronyl transferases

Continued on next page.

Table S2.1: (continued)

Gene	Mapman Name
Potri.003G183900.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.HCT
Potri.005G243700.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.CAD
Potri.002G025700.v3.0	misc.cytochrome P450 : secondary metabolism.phenylpropanoids.lignin biosynthesis.C3H
Potri.007G083000.v3.0	misc.cytochrome P450 : secondary metabolism.phenylpropanoids.lignin biosynthesis.F5H : secondary metabolism.flavonoids.dihydroflavonols.flavonoid 3-monooxygenase
Potri.003G096600.v3.0	amino acid metabolism.synthesis.aromatic aa.chorismate.shikimate kinase
Potri.017G033600.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.4CL
Potri.013G029800.v3.0	amino acid metabolism.synthesis.aromatic aa.chorismate.dehydroquinate/shikimate dehydrogenase
Potri.011G148100.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.CAD
Potri.016G107900.v3.0	secondary metabolism.simple phenols : secondary metabolism.phenylpropanoids
Potri.019G078100.v3.0	secondary metabolism.flavonoids.isoflavones.isoflavone reductase : secondary metabolism.phenylpropanoids
Potri.007G030300.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis : misc.UDP glucosyl and glucoronyl transferases
Potri.002G003200.v3.0	secondary metabolism.phenylpropanoids

Continued on next page.

Table S2.1: (continued)

Gene	Mapman Name
Potri.010G224100.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.PAL
Potri.009G062800.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.CAD
Potri.014G025500.v3.0	secondary metabolism.unspecified : secondary metabolism.phenylpropanoids
Potri.004G188100.v3.0	amino acid metabolism.synthesis.aromatic aa.phenylalanine.arogenate dehydratase / prephenate dehydratase
Potri.001G045800.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.CCR1
Potri.006G199100.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.CAD
Potri.001G046400.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.CCR1
Potri.007G083500.v3.0	misc.cytochrome P450 : secondary metabolism.flavonoids.dihydroflavonols.flavonoid 3-monooxygenase : secondary metabolism.phenylpropanoids.lignin biosynthesis.F5H
Potri.014G041900.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis : secondary metabolism.flavonoids.dihydroflavonols : misc.UDP glucosyl and glucoronyl transferases
Potri.003G057000.v3.0	secondary metabolism.phenylpropanoids
Potri.008G038200.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.PAL

Continued on next page.

Table S2.1: (continued)

Gene	Mapman Name
Potri.010G019000.v3.0	amino acid metabolism.synthesis.aromatic aa.chorismate.dehydroquinate/shikimate dehydrogenase
Potri.001G307200.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.CAD
Potri.016G065300.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.CAD
Potri.018G100500.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.CCR1 : secondary metabolism.phenylpropanoids
Potri.008G040700.v3.0	amino acid metabolism.synthesis.aromatic aa.chorismate.chorismate synthase
Potri.007G003800.v3.0	secondary metabolism.phenylpropanoids
Potri.001G045900.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.CCR1
Potri.004G053500.v3.0	secondary metabolism.phenylpropanoids
Potri.005G248500.v3.0	secondary metabolism.phenylpropanoids
Potri.010G020600.v3.0	amino acid metabolism.synthesis.aromatic aa.chorismate.dehydroquinate/shikimate dehydrogenase
Potri.013G157900.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.C4H
Potri.002G004100.v3.0	secondary metabolism.phenylpropanoids
Potri.014G124100.v3.0	secondary metabolism.phenylpropanoids
Potri.008G195500.v3.0	amino acid metabolism.synthesis.aromatic aa.phenylalanine.arogenate dehydratase / prephenate dehydratase
Potri.017G035100.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.4CL : secondary metabolism.phenylpropanoids

Continued on next page.

Table S2.1: (continued)

Gene	Mapman Name
Potri.017G112800.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.4CL
Potri.002G012800.v3.0	secondary metabolism.phenylpropanoids
Potri.016G091100.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.PAL
Potri.007G116100.v3.0	secondary metabolism.phenylpropanoids
Potri.001G036900.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.4CL
Potri.T107000.v3.0	amino acid metabolism.synthesis.aromatic aa.chorismate.dehydroquinate/shikimate dehydrogenase
Potri.007G085000.v3.0	misc.cytochrome P450 : secondary metabolism.flavonoids.dihydroflavonols.flavonoid 3-monooxygenase : secondary metabolism.phenylpropanoids.lignin biosynthesis.F5H
Potri.007G083200.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.F5H : secondary metabolism.flavonoids.dihydroflavonols.flavonoid 3-monooxygenase : misc.cytochrome P450
Potri.006G265500.v3.0	secondary metabolism.phenylpropanoids
Potri.003G030600.v3.0	amino acid metabolism.synthesis.aromatic aa.tyrosine.prephenate dehydrogenase : amino acid metabolism.synthesis.aromatic aa.tyrosine.arogenate dehydrogenase & prephenate dehydrogenase
Potri.009G063300.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.CAD

Continued on next page.

Table S2.1: (continued)

Gene	Mapman Name
Potri.016G112400.v3.0	secondary metabolism.phenylpropanoids : secondary metabolism.flavonoids.anthocyanins
Potri.014G068300.v3.0	amino acid metabolism.synthesis.aromatic aa.chorismate.5-enolpyruvylshikimate-3-phosphate synthase
Potri.013G120900.v3.0	secondary metabolism.phenylpropanoids
Potri.015G127000.v3.0	secondary metabolism.phenylpropanoids
Potri.005G028400.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.HCT
Potri.007G083300.v3.0	misc.cytochrome P450 : secondary metabolism.phenylpropanoids.lignin biosynthesis.F5H : secondary metabolism.flavonoids.dihydroflavonols.flavonoid 3-monooxygenase
Potri.007G084700.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.F5H : secondary metabolism.flavonoids.dihydroflavonols.flavonoid 3-monooxygenase : misc.cytochrome P450
Potri.005G110900.v3.0	amino acid metabolism.synthesis.aromatic aa.chorismate.3-dehydroquinate synthase
Potri.004G102000.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.4CL
Potri.006G048200.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis : misc.UDP glucosyl and glucoronyl transferases
Potri.014G135500.v3.0	amino acid metabolism.synthesis.aromatic aa.chorismate.dehydroquinate/shikimate dehydrogenase

Continued on next page.

Table S2.1: (continued)

Gene	Mapman Name
Potri.010G230200.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.4CL : secondary metabolism.phenylpropanoids
Potri.007G030200.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis : secondary metabolism.flavonoids.dihydroflavonols : misc.UDP glucosyl and glucoronyl transferases
Potri.008G136600.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.CCoAOMT
Potri.016G031100.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.C3H
Potri.004G105000.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.CCR1
Potri.002G061100.v3.0	amino acid metabolism.synthesis.aromatic aa.chorismate.shikimate kinase
Potri.007G084800.v3.0	misc.cytochrome P450 : secondary metabolism.flavonoids.dihydroflavonols.flavonoid 3-monooxygenase : secondary metabolism.phenylpropanoids.lignin biosynthesis.F5H
Potri.005G028200.v3.0	secondary metabolism.phenylpropanoids
Potri.007G049200.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.CAD
Potri.001G451100.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.COMT : misc.O-methyl transferases
Potri.007G082900.v3.0	misc.cytochrome P450 : secondary metabolism.flavonoids.dihydroflavonols.flavonoid 3-monooxygenase : secondary metabolism.phenylpropanoids.lignin biosynthesis.F5H

Continued on next page.

Table S2.1: (continued)

Gene	**Mapman Name**
Potri.003G003300.v3.0	secondary metabolism.phenylpropanoids
Potri.005G162800.v3.0	amino acid metabolism.synthesis.aromatic aa.chorismate.3-deoxy-D-arabino-heptulosonate 7-phosphate synthase
Potri.015G003100.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.COMT
Potri.018G104700.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.HCT
Potri.005G043400.v3.0	amino acid metabolism.synthesis.aromatic aa.chorismate.dehydroquinate/shikimate dehydrogenase
Potri.001G140700.v3.0	secondary metabolism.phenylpropanoids : secondary metabolism.phenylpropanoids.lignin biosynthesis.CCR1
Potri.008G196100.v3.0	amino acid metabolism.synthesis.aromatic aa.chorismate.dehydroquinate/shikimate dehydrogenase
Potri.002G018300.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.CAD
Potri.016G101500.v3.0	secondary metabolism.phenylpropanoids
Potri.002G076800.v3.0	misc.O-methyl transferases : secondary metabolism.phenylpropanoids.lignin biosynthesis.COMT
Potri.001G334400.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.4CL
Potri.003G093700.v3.0	secondary metabolism.phenylpropanoids : secondary metabolism.phenylpropanoids.lignin biosynthesis.CCR1

Continued on next page.

Table S2.1: (continued)

Gene	Mapman Name
Potri.001G167800.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.C3H : secondary metabolism.flavonoids.dihydroflavonols.flavonoid 3-monooxygenase : misc.cytochrome P450
Potri.012G095000.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.4CL
Potri.005G175400.v3.0	secondary metabolism.phenylpropanoids
Potri.001G300000.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.CAD
Potri.T134100.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.CCR1
Potri.016G057300.v3.0	misc.UDP glucosyl and glucoronyl transferases : secondary metabolism.flavonoids.flavonols.flavonol 3-O-glycosyltransferase : stress.biotic : secondary metabolism.phenylpropanoids.lignin biosynthesis
Potri.007G095700.v3.0	amino acid metabolism.synthesis.aromatic aa.chorismate.3-deoxy-D-arabino-heptulosonate 7-phosphate synthase
Potri.010G104400.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.CCoAOMT
Potri.006G169700.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.4CL
Potri.006G094100.v3.0	secondary metabolism.simple phenols : secondary metabolism.phenylpropanoids
Potri.007G081000.v3.0	amino acid metabolism.synthesis.aromatic aa.chorismate.shikimate kinase

Continued on next page.

Table S2.1: (continued)

Gene	Mapman Name
Potri.016G023300.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.CAD
Potri.001G032800.v3.0	hormone metabolism.brassinosteroid.synthesis-degradation.BRs.metabolic regulation : misc.cytochrome P450 : secondary metabolism.phenylpropanoids.lignin biosynthesis.F5H : secondary metabolism.isoprenoids.carotenoids.carotenoid epsilon ring hydroxylase
Potri.009G099800.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.CCoAOMT
Potri.011G004700.v3.0	amino acid metabolism.synthesis.aromatic aa.phenylalanine.arogenate dehydratase / prephenate dehydratase
Potri.008G074500.v3.0	amino acid metabolism.synthesis.aromatic aa.tyrosine.prephenate dehydrogenase : amino acid metabolism.synthesis.aromatic aa.tyrosine.arogenate dehydrogenase & prephenate dehydrogenase
Potri.005G084600.v3.0	amino acid metabolism.synthesis.aromatic aa.chorismate.shikimate kinase
Potri.006G024400.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.CAD
Potri.006G169600.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.4CL
Potri.003G099700.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.4CL

Continued on next page.

Table S2.1: (continued)

Gene	Mapman Name
Potri.T161300.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.CAD
Potri.012G006400.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.COMT
Potri.001G128100.v3.0	secondary metabolism.phenylpropanoids
Potri.013G079500.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.CCR1 : secondary metabolism.phenylpropanoids
Potri.016G106100.v3.0	secondary metabolism.phenylpropanoids : secondary metabolism.simple phenols
Potri.010G125400.v3.0	secondary metabolism.phenylpropanoids
Potri.015G092300.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.4CL
Potri.T149600.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.CAD
Potri.005G043300.v3.0	amino acid metabolism.synthesis.aromatic aa.chorismate.dehydroquinate/shikimate dehydrogenase
Potri.018G017400.v3.0	secondary metabolism.phenylpropanoids
Potri.009G148800.v3.0	amino acid metabolism.synthesis.aromatic aa.phenylalanine.arogenate dehydratase / prephenate dehydratase
Potri.006G165200.v3.0	secondary metabolism.phenylpropanoids
Potri.T071600.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.4CL
Potri.004G161600.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.F5H

Continued on next page.

Table S2.1: (continued)

Gene	Mapman Name
Potri.001G133300.v3.0	secondary metabolism.flavonoids.isoflavones.isoflavone reductase : secondary metabolism.phenylpropanoids
Potri.006G062600.v3.0	amino acid metabolism.synthesis.aromatic aa.tyrosine.arogenate dehydrogenase & prephenate dehydrogenase
Potri.002G146400.v3.0	amino acid metabolism.synthesis.aromatic aa.chorismate.5-enolpyruvylshikimate-3-phosphate synthase
Potri.016G106300.v3.0	secondary metabolism.simple phenols : secondary metabolism.phenylpropanoids
Potri.005G147400.v3.0	secondary metabolism.phenylpropanoids
Potri.010G221600.v3.0	amino acid metabolism.synthesis.aromatic aa.chorismate.chorismate synthase
Potri.018G104800.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.HCT
Potri.001G042900.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.HCT
Potri.005G028000.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.HCT
Potri.004G017900.v3.0	secondary metabolism.phenylpropanoids
Potri.005G028100.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.HCT
Potri.008G120200.v3.0	secondary metabolism.phenylpropanoids
Potri.010G186300.v3.0	secondary metabolism.phenylpropanoids
Potri.018G109900.v3.0	secondary metabolism.phenylpropanoids

Continued on next page.

Table S2.1: (continued)

Gene	Mapman Name
Potri.010G224200.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.PAL
Potri.007G016400.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.F5H
Potri.001G362800.v3.0	misc.cytochrome P450 : secondary metabolism.flavonoids.dihydroflavonols.flavonoid 3-monooxygenase : secondary metabolism.phenylpropanoids.lignin biosynthesis.F5H
Potri.006G126800.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.PAL
Potri.008G082300.v3.0	secondary metabolism.flavonoids.dihydroflavonols.flavonoid 3-monooxygenase : secondary metabolism.phenylpropanoids.lignin biosynthesis.F5H : misc.cytochrome P450
Potri.T149400.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.CAD
Potri.010G054200.v3.0	secondary metabolism.phenylpropanoids
Potri.002G004500.v3.0	secondary metabolism.phenylpropanoids
Potri.001G150500.v3.0	amino acid metabolism.synthesis.aromatic aa.chorismate.3-deoxy-D-arabino-heptulosonate 7-phosphate synthase
Potri.004G013400.v3.0	amino acid metabolism.synthesis.aromatic aa.phenylalanine.arogenate dehydratase / prephenate dehydratase
Potri.016G093700.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.4CL

Continued on next page.

Table S2.1: (continued)

Gene	Mapman Name
Potri.009G095800.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.CAD
Potri.009G063400.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.CAD
Potri.003G100200.v3.0	secondary metabolism.phenylpropanoids : secondary metabolism.flavonoids.isoflavones.isoflavone reductase
Potri.018G021200.v3.0	secondary metabolism.phenylpropanoids
Potri.018G105400.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.HCT
Potri.002G183600.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.CCoAOMT
Potri.007G083700.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.F5H : secondary metabolism.flavonoids.dihydroflavonols.flavonoid 3-monooxygenase : misc.cytochrome P450
Potri.009G062900.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.CAD
Potri.018G146100.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.C4H
Potri.001G045300.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.CCR1
Potri.018G094200.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.4CL
Potri.017G034900.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.4CL : secondary metabolism.phenylpropanoids

Continued on next page.

Table S2.1: (continued)

Gene	Mapman Name
Potri.019G048200.v3.0	amino acid metabolism.synthesis.aromatic aa.chorismate.shikimate kinase
Potri.005G257700.v3.0	secondary metabolism.phenylpropanoids
Potri.018G070300.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.CCoAOMT
Potri.002G086000.v3.0	secondary metabolism.phenylpropanoids
Potri.008G031500.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.4CL : secondary metabolism.phenylpropanoids
Potri.001G201100.v3.0	amino acid metabolism.synthesis.aromatic aa.tyrosine.prephenate dehydrogenase
Potri.007G083600.v3.0	secondary metabolism.flavonoids.dihydroflavonols.flavonoid 3-monooxygenase : secondary metabolism.phenylpropanoids.lignin biosynthesis.F5H : misc.cytochrome P450
Potri.012G094800.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.4CL
Potri.016G078300.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.CAD
Potri.007G030400.v3.0	misc.UDP glucosyl and glucoronyl transferases : secondary metabolism.phenylpropanoids.lignin biosynthesis
Potri.003G057200.v3.0	secondary metabolism.phenylpropanoids
Potri.008G071200.v3.0	secondary metabolism.phenylpropanoids
Potri.016G031000.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.C3H
Potri.003G188500.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.4CL

Continued on next page.

Table S2.1: (continued)

Gene	Mapman Name
Potri.008G136700.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.CCoAOMT
Potri.001G045500.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.CCR1
Potri.006G141400.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.F5H : secondary metabolism.flavonoids.dihydroflavonols.flavonoid 3-monooxygenase : misc.cytochrome P450
Potri.005G073300.v3.0	amino acid metabolism.synthesis.aromatic aa.chorismate.3-deoxy-D-arabino-heptulosonate 7-phosphate synthase
Potri.003G181400.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.CCR1
Potri.014G106600.v3.0	misc.O-methyl transferases : secondary metabolism.phenylpropanoids.lignin biosynthesis.COMT
Potri.002G026000.v3.0	misc.cytochrome P450 : secondary metabolism.phenylpropanoids.lignin biosynthesis.C3H : secondary metabolism.flavonoids.dihydroflavonols.flavonoid 3-monooxygenase
Potri.019G126400.v3.0	polyamine metabolism : secondary metabolism.phenylpropanoids
Potri.006G033300.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.C3H
Potri.017G062800.v3.0	amino acid metabolism.synthesis.aromatic aa.chorismate.shikimate kinase

Continued on next page.

Table S2.1: (continued)

Gene	Mapman Name
Potri.018G105500.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.HCT
Potri.001G363900.v3.0	misc.cytochrome P450 : secondary metabolism.phenylpropanoids.lignin biosynthesis.F5H : secondary metabolism.flavonoids.dihydroflavonols.flavonoid 3-monooxygenase
Potri.007G084400.v3.0	secondary metabolism.flavonoids.dihydroflavonols.flavonoid 3-monooxygenase : secondary metabolism.phenylpropanoids.lignin biosynthesis.F5H : misc.cytochrome P450
Potri.T149300.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.CAD
Potri.006G078100.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.C4H
Potri.001G365300.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.F5H : misc.cytochrome P450
Potri.019G130700.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.C4H
Potri.008G024800.v3.0	secondary metabolism.flavonoids.dihydroflavonols : misc.UDP glucosyl and glucoronyl transferases : secondary metabolism.phenylpropanoids.lignin biosynthesis : hormone metabolism.salicylic acid.synthesis-degradation : secondary metabolism.flavonoids.anthocyanins.anthocyanidin 3-O-glucosyltransferase

Continued on next page.

Table S2.1: (continued)

Gene	Mapman Name
Potri.009G076300.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.CCR1
Potri.019G049500.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.4CL
Potri.005G175600.v3.0	secondary metabolism.phenylpropanoids
Potri.003G210700.v3.0	secondary metabolism.phenylpropanoids : secondary metabolism.phenylpropanoids.lignin biosynthesis.4CL
Potri.014G025600.v3.0	secondary metabolism.phenylpropanoids : secondary metabolism.unspecified
Potri.009G123600.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.F5H
Potri.001G045700.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.CCR1
Potri.001G046100.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.CCR1
Potri.003G057100.v3.0	secondary metabolism.phenylpropanoids
Potri.003G210600.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.4CL : secondary metabolism.phenylpropanoids : lipid metabolism.FA synthesis and FA elongation.acyl coa ligase
Potri.001G045600.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.CCR1
Potri.006G178700.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.CCR1 : secondary metabolism.phenylpropanoids
Potri.005G117500.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.F5H

Continued on next page.

Table S2.1: (continued)

Gene	**Mapman Name**
Potri.006G024300.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.CAD
Potri.001G365100.v3.0	misc.cytochrome P450 : secondary metabolism.flavonoids.dihydroflavonols.flavonoid 3-monooxygenase : secondary metabolism.phenylpropanoids.lignin biosynthesis.F5H
Potri.010G057000.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.4CL
Potri.001G045400.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.CCR1
Potri.007G030500.v3.0	misc.UDP glucosyl and glucoronyl transferases : secondary metabolism.flavonoids.dihydroflavonols : secondary metabolism.phenylpropanoids.lignin biosynthesis
Potri.013G029900.v3.0	amino acid metabolism.synthesis.aromatic aa.chorismate.dehydroquinate/shikimate dehydrogenase
Potri.017G110500.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.CCR1
Potri.002G099200.v3.0	amino acid metabolism.synthesis.aromatic aa.chorismate.3-deoxy-D-arabino-heptulosonate 7-phosphate synthase
Potri.001G055700.v3.0	secondary metabolism.phenylpropanoids
Potri.019G084300.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.COMT
Potri.011G148200.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.CAD

Continued on next page.

Table S2.1: (continued)

Gene	Mapman Name
Potri.001G349600.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.CCR1
Potri.009G063100.v3.0	secondary metabolism.phenylpropanoids.lignin biosynthesis.CAD

Table S2.2: Mass/Charge (mz) ratio for Lignin pyMBMS Peaks.

mz	Annotation [21]
120	lignin (vinylphenol)
124	lignin, guaiacyl
137	lignin,guaiacyl (Ethylguaiacol, homovanillin,coniferyl alcohol)
138	lignin,guaiacyl (Methylguaiacol)
150	lignin,guaiacyl (Vinylguaiacol)
152	lignin
154	lignin,syringyl (Syringol)
168	syringyl (4-Methyl-2,6-dimethoxyphenol)
180	lignin (Coniferyl alcohol, syringylethene)
182	lignin,syringyl (Syringaldehyde)
210	lignin,syringyl (Sinapylalcohol)

Table S2.3: Lignin-related metababololites from the metabolomics analysis. For partially identified metabolites, additional RT and mz information is provided.

See attached excel file.

Table S2.4: LOE Scores, *Arabidopsis* best hits and MapMan annotations of genes for which $LOE_{breadth}(g) \geq 3$ and $LOE_{gwas}(g) \geq 1$.

See attached excel file.

Table S2.5: Positions of SNPs involved in SNP correlations in select portential new target genes.

Source	**Target**
SNP Correlations between Potri.007G115100 (AGL22) and Potri.016G107900 (laccase 7)	
SNP 07:13647758	SNP 16:11083690
SNP 07:13647758	SNP 16:11083708
SNP 07:13647758	SNP 16:11083712
SNP 07:13647758	SNP 16:11083737
SNP 07:13647978	SNP 16:11083690
SNP 07:13647978	SNP 16:11083708
SNP 07:13647978	SNP 16:11083712
SNP 07:13647978	SNP 16:11083737
SNP 07:13648235	SNP 16:11083690
SNP 07:13648235	SNP 16:11083708
SNP 07:13648235	SNP 16:11083712
SNP 07:13648235	SNP 16:11083737
SNP 07:13648488	SNP 16:11083690
SNP 07:13648488	SNP 16:11083708
SNP 07:13648488	SNP 16:11083712
SNP 07:13648488	SNP 16:11083737
SNP Correlations between Potri.007G115100 (AGL22) and Potri.007G116100 (nicotinamidase 1)	
SNP 07:13647645	SNP 07:13706654
SNP 07:13647645	SNP 07:13706699
SNP 07:13647645	SNP 07:13706834
SNP 07:13647758	SNP 07:13706654

Continued on next page.

Table S2.5: (continued)

Source	**Target**
SNP 07:13647758	SNP 07:13706699
SNP 07:13647978	SNP 07:13706654
SNP 07:13647978	SNP 07:13706699
SNP 07:13647978	SNP 07:13706834

Bibliography

[1] Poulomi Sannigrahi, Arthur J Ragauskas, and Gerald A Tuskan. Poplar as a feedstock for biofuels: A review of compositional characteristics. *Biofuels, Bioproducts and Biorefining*, 4(2):209–226, 2010. 69

[2] Arthur J Ragauskas, Charlotte K Williams, Brian H Davison, George Britovsek, John Cairney, Charles A Eckert, William J Frederick, Jason P Hallett, David J Leak, Charles L Liotta, et al. The Path Forward for Biofuels and Biomaterials. *science*, 311(5760):484–489, 2006. 69

[3] Gerald A Tuskan, S Difazio, Stefan Jansson, J Bohlmann, I Grigoriev, U Hellsten, N Putnam, S Ralph, Stephane Rombauts, A Salamov, et al. The Genome of Black Cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science*, 313(5793):1596–1604, 2006. 69, 74, 77

[4] Gerald Tuskan, Gancho Slavov, Steve DiFazio, Wellington Muchero, Ranjan Pryia, Wendy Schackwitz, Joel Martin, Daniel Rokhsar, Robert Sykes, Mark Davis, et al. *Populus* resequencing: towards genome-wide association studies. In *BMC Proceedings*, volume 5, page I21. BioMed Central Ltd, 2011. 69, 74

[5] Gancho T Slavov, Stephen P DiFazio, Joel Martin, Wendy Schackwitz, Wellington Muchero, Eli Rodgers-Melnick, Mindie F Lipphardt, Christa P Pennacchio, Uffe Hellsten, Len A Pennacchio, et al. Genome resequencing reveals multiscale geographic structure and extensive linkage disequilibrium in the forest tree *Populus trichocarpa*. *New Phytologist*, 196(3):713–725, 2012. 69, 74

[6] Luke M Evans, Gancho T Slavov, Eli Rodgers-Melnick, Joel Martin, Priya Ranjan, Wellington Muchero, Amy M Brunner, Wendy Schackwitz, Lee Gunter, Jin-Gui Chen, et al. Population genomics of *Populus trichocarpa* identifies signatures of selection and adaptive trait associations. *Nature Genetics*, 46(10):1089–1096, 2014. 69, 74

[7] Athena D McKown, Jaroslav Klápště, Robert D Guy, Armando Geraldes, Ilga Porth, Jan Hannemann, Michael Friedmann, Wellington Muchero, Gerald A Tuskan, Jürgen Ehlting, et al. Genome-wide association implicates numerous genes underlying ecological trait variation in natural populations of *Populus trichocarpa*. *New Phytologist*, 203(2):535–553, 2014. 69, 74

[8] Kelly J Vining, Kyle R Pomraning, Larry J Wilhelm, Henry D Priest, Matteo Pellegrini, Todd C Mockler, Michael Freitag, and Steven H Strauss. Dynamic DNA cytosine methylation in the *Populus trichocarpa* genome: tissue-level variation and relationship to gene expression. *BMC Genomics*, 13(1):1, 2012. 69, 78

[9] Gina M Calabrese, Larry D Mesner, Joseph P Stains, Steven M Tommasini, Mark C Horowitz, Clifford J Rosen, and Charles R Farber. Integrating GWAS and Co-expression Network Data Identifies Bone Mineral Density Genes SPTBN1 and MARK3 and an Osteoblast Functional Module. *Cell systems*, 4(1):46–59, 2017. 69

[10] Supinda Bunyavanich, Eric E Schadt, Blanca E Himes, Jessica Lasky-Su, Weiliang Qiu, Ross Lazarus, John P Ziniti, Ariella Cohain, Michael Linderman, Dara G Torgerson, et al. Integrated genome-wide association, coexpression network, and expression single nucleotide polymorphism analysis identifies novel pathway in allergic rhinitis. *BMC medical genomics*, 7(1):48, 2014. 69

[11] Marylyn D Ritchie, Emily R Holzinger, Ruowang Li, Sarah A Pendergrass, and Dokyoon Kim. Methods of integrating data to uncover genotype-phenotype interactions. *Nature reviews. Genetics*, 16(2):85, 2015. 70

[12] Brooke L Fridley, Steven Lund, Gregory D Jenkins, and Liewei Wang. A Bayesian integrative genomic model for pathway analysis of complex traits. *Genetic epidemiology*, 36(4):352–359, 2012. 70

[13] Dokyoon Kim, Ruowang Li, Scott M Dudek, and Marylyn D Ritchie. ATHENA: Identifying interactions between different levels of genomic data associated with cancer clinical outcomes using grammatical evolution neural network. *BioData mining*, 6(1):23, 2013. 70

[14] Dokyoon Kim, Hyunjung Shin, Young Soo Song, and Ju Han Kim. Synergistic effect of different levels of genomic data for cancer clinical outcome prediction. *Journal of biomedical informatics*, 45(6):1191–1198, 2012. 70

[15] Qiao Zhao, Yining Zeng, Yanbin Yin, Yunqiao Pu, Lisa A Jackson, Nancy L Engle, Madhavi Z Martin, Timothy J Tschaplinski, Shi-You Ding, Arthur J Ragauskas, et al. Pinoresinol reductase 1 impacts lignin distribution during secondary cell wall biosynthesis in arabidopsis. *Phytochemistry*, 112:170–178, 2015. 73

[16] Timothy J Tschaplinski, Robert F Standaert, Nancy L Engle, Madhavi Z Martin, Amandeep K Sangha, Jerry M Parks, Jeremy C Smith, Reichel Samuel, Nan Jiang, Yunqiao Pu, Arthur J Ragauskas, Choo Y Hamilton, Chunxiang Fu, Zeng-Yu Wang, Brian H Davidson, Richard A Dixon, and Jonathan R Mielenz. Down-regulation of the caffeic acid *O*-methyltransferase gene in switchgrass reveals a novel monolignol analog. *Biotechnology for Biofuels*, 5(1):1, 2012. 73

[17] Yongchao Li, Timothy J Tschaplinski, Nancy L Engle, Choo Y Hamilton, Miguel Rodriguez, James C Liao, Christopher W Schadt, Adam M Guss, Yunfeng Yang, and David E Graham. Combined inactivation of the *Clostridium cellulolyticum* lactate and malate dehydrogenase genes substantially increases ethanol yield from cellulose and switchgrass fermentations. *Biotechnology for biofuels*, 5(1):2, 2012. 73

[18] Collin M Timm, Dale A Pelletier, Sara S Jawdy, Lee E Gunter, Jeremiah A Henning, Nancy Engle, Jayde Aufrecht, Emily Gee, Intawat Nookaew, Zamin Yang, et al. Two poplar-associated bacterial isolates induce additive favorable responses in a constructed plant-microbiome system. *Frontiers in plant science*, 7:497, 2016. 73

[19] Ajaya K Biswal, Zhangying Hao, Sivakumar Pattathil, Xiaohan Yang, Kim Winkeler, Cassandra Collins, Sushree S Mohanty, Elizabeth A Richardson, Ivana Gelineo-Albersheim, Kimberly Hunt, et al. Downregulation of gaut12 in populus deltoides by rna silencing results in reduced recalcitrance, increased growth and reduced xylan and pectin in a woody biofuel feedstock. *Biotechnology for biofuels*, 8(1):41, 2015. 73

[20] Robert J Evans and Thomas A Milne. Molecular Characterization of the Pyrolysis of Biomass. *Energy & Fuels*, 1(2):123–137, 1987. 73

[21] Robert Sykes, Matthew Yung, Evandro Novaes, Matias Kirst, Gary Peter, and Mark Davis. High-Throughput Screening of Plant Cell-Wall Composition Using Pyrolysis Molecular Beam Mass Spectroscopy. *Biofuels: Methods and protocols*, pages 169–183, 2009. 73, 74, 82, 127

[22] Gerald Tuskan, Darrell West, Harvey D Bradshaw, David Neale, Mitch Sewell, Nick Wheeler, Bob Megraw, Keith Jech, Art Wiselogel, Robert Evans, et al. Two High-Throughput Techniques for Determining Wood Properties as Part of a Molecular Genetics Analysis of Hybrid Poplar and Loblolly Pine. In *Twentieth Symposium on Biotechnology for Fuels and Chemicals*, pages 55–65. Springer, 1999. 73

[23] Wellington Muchero, Jianjun Guo, Stephen P DiFazio, Jin-Gui Chen, Priya Ranjan, Gancho T Slavov, Lee E Gunter, Sara Jawdy, Anthony C Bryan, Robert Sykes, et al. High-resolution genetic mapping of allelic variants associated with cell wall chemistry in *Populus*. *BMC genomics*, 16(1):24, 2015. 74, 82

[24] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, 20(9):1297–1303, 2010. 74

[25] Geraldine A Van der Auwera, Mauricio O Carneiro, Christopher Hartl, Ryan Poplin, Guillermo del Angel, Ami Levy-Moonshine, Tadeusz Jordan, Khalid Shakir, David

Roazen, Joel Thibault, et al. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Current protocols in bioinformatics*, pages 11–10, 2013. 74

[26] Armando Geraldes, SP Difazio, GT Slavov, P Ranjan, W Muchero, J Hannemann, LE Gunter, AM Wymore, CJ Grassa, N Farzaneh, et al. 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, 2013. 75

[27] Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A Albers, Eric Banks, Mark A DePristo, Robert E Handsaker, Gerton Lunter, Gabor T Marth, Stephen T Sherry, et al. The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158, 2011. 75

[28] Christophe Leys, Christophe Ley, Olivier Klein, Philippe Bernard, and Laurent Licata. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, 49(4):764–766, 2013. 75

[29] Hyun Min Kang, Jae Hoon Sul, Noah A Zaitlen, Sit-yee Kong, Nelson B Freimer, Chiara Sabatti, Eleazar Eskin, et al. Variance component model to account for sample structure in genome-wide association studies. *Nature genetics*, 42(4):348–354, 2010. 76

[30] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300, 1995. 76

[31] Adrian M. Price-Whelan and Daniel Foreman-Mackey. schwimmbad: A uniform interface to parallel processing pools in Python. *The Journal of Open Source Software*, 2(17), sep 2017. 76

[32] David M Goodstein, Shengqiang Shu, Russell Howson, Rochak Neupane, Richard D Hayes, Joni Fazo, Therese Mitros, William Dirks, Uffe Hellsten, Nicholas Putnam,

et al. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Research*, 40(D1):D1178–D1186, 2012. 76, 78, 81

[33] Ryan F McCormick, Sandra K Truong, Avinash Sreedasyam, Jerry Jenkins, Shengqiang Shu, David Sims, Megan Kennedy, Mojgan Amirebrahimi, Brock D Weers, Brian McKinley, et al. The sorghum bicolor reference genome: improved assembly, gene annotations, a transcriptome atlas, and signatures of genome organization. *The Plant Journal*, 93(2):338–354, 2018. 77

[34] Zhiwu Li and Harold N Trick. Rapid method for high-quality RNA isolation from seed endosperm containing high levels of starch. *Biotechniques*, 38(6):872, 2005. 77

[35] Hongshan Jiang, Rong Lei, Shou-Wei Ding, and Shuifang Zhu. Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC bioinformatics*, 15(1):182, 2014. 77

[36] Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, 2013. 77

[37] Günter P Wagner, Koryu Kin, and Vincent J Lynch. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory in biosciences*, 131(4):281–285, 2012. 78

[38] Stijn Van Dongen. Graph Clustering Via a Discrete Uncoupling Process. *SIAM Journal on Matrix Analysis and Applications*, 30(1):121–141, 2008. 78, 79

[39] Stijn Marinus Van Dongen. Graph clustering by flow simulation. 2001. 78, 79

[40] Derek W Barnett, Erik K Garrison, Aaron R Quinlan, Michael P Strömberg, and Gabor T Marth. BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics*, 27(12):1691–1692, 2011. 78

[41] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009. 78

[42] Simon Anders, Paul Theodor Pyl, and Wolfgang Huber. HTSeq–a Python framework to work with high-throughput sequencing data. *Bioinformatics*, page btu638, 2014. 78

[43] Robert Busch, Weiliang Qiu, Jessica Lasky-Su, Jarrett Morrow, Gerard Criner, and Dawn DeMeo. Differential DNA methylation marks and gene comethylation of COPD in African-Americans with COPD exacerbations. *Respiratory Research*, 17(1):143, 2016. 79

[44] Ruslan Akulenko and Volkhard Helms. DNA co-methylation analysis suggests novel functional associations between gene pairs in breast cancer samples. *Human molecular genetics*, page ddt158, 2013. 79

[45] Matthew N Davies, Manuela Volta, Ruth Pidsley, Katie Lunnon, Abhishek Dixit, Simon Lovestone, Cristian Coarfa, R Alan Harris, Aleksandar Milosavljevic, Claire Troakes, et al. Functional annotation of the human brain methylome identifies tissue-specific epigenetic variation across brain and blood. *Genome biology*, 13(6):R43, 2012. 79

[46] Sharlee Climer, Wei Yang, Lisa Fuentes, Victor G Dávila-Román, and C Charles Gu. A Custom Correlation Coefficient (CCC) Approach for Fast Identification of Multi-SNP Association Patterns in Genome-Wide SNPs Data. *Genetic Epidemiology*, 38(7):610–621, 2014. 79

[47] Sharlee Climer, Alan R Templeton, and Weixiong Zhang. Allele-Specific Network Reveals Combinatorial Interaction that Transcends Small Effects in Psoriasis GWAS. *PLoS Comput Biol*, 10(9):e1003766, 2014. 79

[48] Gene Ontology Consortium. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Research*, 45(D1):D331–D338, 2017. 81

[49] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, Midori A Harris, David P Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C Matese, Joel E Richardson, Martin Ringwald, Gerald M Rubin, and Gavin

Sherlock. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000. 81

[50] Robert D. Finn, Penelope Coggill, Ruth Y. Eberhardt, Sean R. Eddy, Jaina Mistry, Alex L. Mitchell, Simon C. Potter, Marco Punta, Matloob Qureshi, Amaia Sangrador-Vegas, Gustavo A. Salazar, John Tate, and Alex Bateman. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Research*, 44(D1):D279–D285, 2016. 81

[51] Marc Lohse, Axel Nagel, Thomas Herter, Patrick May, Michael Schroda, Rita Zrenner, Takayuki Tohge, Alisdair R Fernie, Mark Stitt, and Björn Usadel. Mercator: a fast and simple web server for genome scale functional annotation of plant sequence data. *Plant, cell & environment*, 37(5):1250–1258, 2014. 81

[52] Ruben Vanholme, Brecht Demedts, Kris Morreel, John Ralph, and Wout Boerjan. Lignin Biosynthesis and Structure. *Plant physiology*, 153(3):895–905, 2010. 81

[53] Mark F Davis, Gerald A Tuskan, Peggy Payne, Timothy J Tschaplinski, and Richard Meilan. Assessment of *Populus* wood chemistry following the introduction of a Bt toxin gene. *Tree physiology*, 26(5):557–564, 2006. 82

[54] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, 13(11):2498–2504, 2003. 85

[55] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017. 85

[56] Andrie de Vries and Brian D. Ripley. *ggdendro: Create Dendrograms and Tree Diagrams Using 'ggplot2'*, 2016. R package version 0.1-20. 85

[57] Baptiste Auguie. *gridExtra: Miscellaneous Functions for "Grid" Graphics*, 2017. R package version 2.3. 85

[58] Hadley Wickham. Reshaping data with the reshape package. *Journal of Statistical Software*, 21(12), 2007. 85

[59] Jeffrey B. Arnold. *ggthemes: Extra Themes, Scales and Geoms for 'ggplot2'*, 2017. R package version 3.4.0. 85

[60] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009. 85

[61] Eshchar Mizrachi, Lieven Verbeke, Nanette Christie, Ana C Fierro, Shawn D Mansfield, Mark F Davis, Erica Gjersing, Gerald A Tuskan, Marc Van Montagu, Yves Van de Peer, et al. 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, 2017. 86

[62] Le Shu, Yuqi Zhao, Zeyneb Kurt, Sean Geoffrey Byars, Taru Tukiainen, Johannes Kettunen, Luz D Orozco, Matteo Pellegrini, Aldons J Lusis, Samuli Ripatti, et al. Mergeomics: multidimensional data integration to identify pathogenic perturbations to biological systems. *BMC genomics*, 17(1):874, 2016. 86

[63] Keywan Hassani-Pak. *KnetMiner - An integrated data platform for gene mining and biological knowledge discovery*. PhD thesis, Universität Bielefeld, 2017. 86

[64] Keywan Hassani-Pak, Martin Castellote, Maria Esch, Matthew Hindle, Artem Lysenko, Jan Taubert, and Christopher Rawlings. Developing integrated crop knowledge networks to advance candidate gene discovery. *Applied & translational genomics*, 11:18–26, 2016. 86

[65] Seung Kwan Yoo, Jong Seob Lee, and Ji Hoon Ahn. Overexpression of *AGAMOUS-LIKE 28* (*AGL28*) promotes flowering by upregulating expression of floral promoters within the autonomous pathway. *Biochemical and biophysical research communications*, 348(3):929–936, 2006. 87

[66] Donna E Fernandez, Chieh-Ting Wang, Yumei Zheng, Benjamin J Adamczyk, Rajneesh Singhal, Pamela K Hall, and Sharyn E Perry. 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 physiology*, 165(4):1591–1603, 2014. 87

[67] Xiaohui Yu, Guoping Chen, Xuhu Guo, Yu Lu, Jianling Zhang, Jingtao Hu, Shibing Tian, and Zongli Hu. Silencing *SlAGL6*, a tomato *AGAMOUS-LIKE6* lineage gene, generates fused sepal and green petal. *Plant Cell Reports*, pages 1–11, 2017. 87

[68] Hao Yu, Toshiro Ito, Frank Wellmer, and Elliot M Meyerowitz. Repression of AGAMOUS-LIKE 24 is a crucial step in promoting flower development. *Nature genetics*, 36(2):157, 2004. 87

[69] Hao Yu, Yifeng Xu, Ee Ling Tan, and Prakash P Kumar. AGAMOUS-LIKE 24, a dosage-dependent mediator of the flowering signals. *Proceedings of the National Academy of Sciences*, 99(25):16336–16341, 2002. 87

[70] Horim Lee, Sung-Suk Suh, Eunsook Park, Euna Cho, Ji Hoon Ahn, Sang-Gu Kim, Jong Seob Lee, Young Myung Kwon, and Ilha Lee. The AGAMOUS-LIKE 20 MADS domain protein integrates floral inductive pathways in *Arabidopsis*. *Genes & Development*, 14(18):2366–2376, 2000. 87

[71] Martin F Yanofsky, Sarah Liljegren, and Cristina Ferrandiz. Selective control of lignin biosynthesis in transgenic plants, July 27 2004. US Patent 6,768,042. 87

[72] Cristina Ferrándiz, Sarah J Liljegren, and Martin F Yanofsky. Negative regulation of the *SHATTERPROOF* genes by FRUITFULL during *Arabidopsis* fruit development. *Science*, 289(5478):436–438, 2000. 87, 93

[73] Estela Giménez, Benito Pineda, Juan Capel, María Teresa Antón, Alejandro Atarés, Fernando Pérez-Martín, Begoña García-Sogo, Trinidad Angosto, Vicente Moreno, and Rafael Lozano. Functional Analysis of the *Arlequin* Mutant Corroborates the Essential Role of the *Arlequin/TAGL1* Gene during Reproductive Development of Tomato. *PLoS One*, 5(12):e14427, 2010. 93

[74] Claudia Cosio, Philippe Ranocha, Edith Francoz, Vincent Burlat, Yumei Zheng, Sharyn E Perry, Juan-Jose Ripoll, Martin Yanofsky, and Christophe Dunand. 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(1):250–263, 2017. 93

[75] Christian Dubos, Ralf Stracke, Erich Grotewold, Bernd Weisshaar, Cathie Martin, and Loïc Lepiniec. MYB transcription factors in *Arabidopsis*. *Trends in plant science*, 15(10):573–581, 2010. 97

[76] Jingying Liu, Anne Osbourn, and Pengda Ma. MYB Transcription Factors as Regulators of Phenylpropanoid Metabolism in Plants. *Molecular plant*, 8(5):689–708, 2015. 97, 99

[77] Meiliang Zhou, Kaixuan Zhang, Zhanmin Sun, Mingli Yan, Cheng Chen, Xinquan Zhang, Yixiong Tang, and Yanmin Wu. LNK1 and LNK2 Corepressors Interact with the MYB3 Transcription Factor in Phenylpropanoid Biosynthesis. *Plant Physiology*, 174(3):1348–1358, 2017. 97

[78] Takehiro Kamiya, Monica Borghi, Peng Wang, John MC Danku, Lothar Kalmbach, Prashant S Hosmani, Sadaf Naseer, Toru Fujiwara, Niko Geldner, and David E Salt. The MYB36 transcription factor orchestrates Casparian strip formation. *Proceedings of the National Academy of Sciences*, 112(33):10533–10538, 2015. 97

[79] Ruiqin Zhong, Elizabeth A Richardson, and Zheng-Hua Ye. The MYB46 Transcription Factor Is a Direct Target of SND1 and Regulates Secondary Wall Biosynthesis in *Arabidopsis*. *The Plant Cell*, 19(9):2776–2792, 2007. 99

[80] Dirk Schenke, Christoph Boettcher, and Dierk Scheel. Crosstalk between abiotic ultraviolet-B stress and biotic (flg22) stress signalling in *Arabidopsis* prevents flavonol accumulation in favor of pathogen defence compound production. *Plant, cell & environment*, 34(11):1849–1864, 2011. 99

[81] Zheng Zhou, Dirk Schenke, Ying Miao, and Daguang Cai. Investigation of the crosstalk between the flg22 and the UV-B-induced flavonol pathway in *Arabidopsis thaliana* seedlings. *Plant, cell & environment*, 40(3):453–458, 2017. 99

[82] Bianyun Yu, Margaret Y Gruber, George G Khachatourians, Rong Zhou, Delwin J Epp, Dwayne D Hegedus, Isobel AP Parkin, Ralf Welsch, and Abdelali Hannoufa. Arabidopsis cpSRP54 regulates carotenoid accumulation in *Arabidopsis* and Brassica napus. *Journal of experimental botany*, 63(14):5189–5202, 2012. 101

[83] Matthew R Groves, Alexandra Mant, Audrey Kuhn, Joachim Koch, Stefan Dübel, Colin Robinson, and Irmgard Sinning. Functional Characterization of Recombinant Chloroplast Signal Recognition Particle. *Journal of Biological Chemistry*, 276(30):27778–27786, 2001. 101

[84] Danja Schünemann. Structure and function of the chloroplast signal recognition particle. *Current genetics*, 44(6):295–304, 2004. 101

[85] Markus Klenell, Shigeto Morita, Mercedes Tiemblo-Olmo, Per Mühlenbock, Stanislaw Karpinski, and Barbara Karpinska. Involvement of the Chloroplast Signal Recognition Particle cpSRP43 in Acclimation to Conditions Promoting Photooxidative Stress in *Arabidopsis*. *Plant and cell physiology*, 46(1):118–129, 2005. 101

[86] Stephen J Streatfield, Andreas Weber, Elizabeth A Kinsman, Rainer E Häusler, Jianming Li, Dusty Post-Beittenmiller, Werner M Kaiser, Kevin A Pyke, Ulf-Ingo Flügge, and Joanne Chory. The Phosphoenolpyruvate/Phosphate Translocator Is Required for Phenolic Metabolism, Palisade Cell Development, and Plastid-Dependent Nuclear Gene Expression. *The Plant Cell*, 11(9):1609–1621, 1999. 101

Chapter 3

Multi-Phenotype Association Decomposition: Unraveling complex gene-phenotype relationships

This chapter has been submitted for publication in *Frontiers in Genetics* and is currently under review. This chapter contains contributions from other authors:

Deborah Weighill, Piet Jones, Carissa Bleker Priya Ranjan, Manesh Shah, Nan Zhao, Madhavi Martin, Stephen DiFazio, David Macaya-Sanz, Jeremy Schmutz, Avinash Sreedasyam, Timothy Tschaplinski, Gerald Tuskan and Daniel Jacobson

Author Contributions: DJ conceived of the study and supervised the project. DAW developed MPA decomposition and signature clustering, implemented the method, generated and interpreted results and wrote the manuscript. PJ and CB performed the GWAS and outlier analysis and contributed to the corresponding methods sections. PR, NZ, MM and TT performed the metabolomics. JS and AS contributed the genome sequence and transcriptome expression analysis. MS mapped gene expression atlas reads and calculated gene expression TPM values. GT led the sequencing of *Populus* genotypes. SD and DMS performed the SNP calling and validation. DJ, GT, TT, PJ, DMS and SD provided editorial feedback on the manuscript. TT, PJ and CB contributed Metabolomics, GWAS and Outlier Analysis sections. TT contributed metabolomics methods and interpretations of metabolite associations in separate pathways. Nancy Engle, David Weston, Ryan Ahg, KC Cushman, Lee Gunter and Sara Jawdy collected metabolomics samples, Sara Jawdy performed library preparation for the gene atlas experiment and Lee Gunter performed preparation of the GWAS genomic samples.

Abstract

Various patterns of multi-phenotype associations (MPAs) exist in the results of Genome Wide Association Studies (GWAS) involving different topologies of single nucleotide polymorphism (SNP)-phenotype associations. These can provide interesting information about the different impacts of a gene on closely related phenotypes or disparate phenotypes (pleiotropy). In this work we present MPA Decomposition, a new network-based approach which decomposes the results of a multi-phenotype GWAS study into three bipartite networks, which, when used together, unravel the multi-phenotype signatures

of genes on a genome-wide scale. The decomposition involves the construction of a phenotype powerset space, and subsequent mapping of genes into this new space. Clustering of genes in this powerset space groups genes based on their detailed MPA signatures. We show that this method allows us to find multiple different MPA and pleiotropic signatures within individual genes and to classify and cluster genes based on these SNP-phenotype association topologies. We demonstrate the use of this approach on a GWAS analysis of a large population of 882 *Populus trichocarpa* genotypes using untargeted metabolomics phenotypes. This method should prove invaluable in the interpretation of large GWAS datasets and aid in future synthetic biology efforts designed to optimize phenotypes of interest.

3.1 Introduction

Unravelling the complex genetic patterns underlying complex phenotypes has previously been challenging. While individual Genome-Wide Association Studies (GWAS) can provide insight into the genetic underpinnings of measured phenotypes, they typically involved associations of genetic variants with only one or a few phenotypes. The field of phenomics involves the collection of high-dimensional phenotype data of an organism, with the aim of capturing the overall, comprehensive phenotype (the "Phenome") of the organism [1]. Association studies involving many measured phenotypes, for example, Phenome-Wide Association Studies (PheWAS) present many advantages, in that they allow for the complex interconnected networks between phenotypes and their genetic underpinnings to be elucidated, and also allow for the detection of pleiotropy [2, 3, 4, 5].

Pleiotropy is the phenomenon in which a gene affects multiple phenotypes [6]. One can also have a locus-centric view of pleiotropy involving a single SNP affecting multiple phenotypes [7]. While pleiotropy used to be considered an exception to the rules of Mendelian genetics, it has since been proposed to be a common, central property inherent to biological systems [6]. Multi-phenotype associations (MPAs) can be detected in the results of Genome Wide Association Studies (GWASs) as Single

Nucleotide Polymorphisms (SNPs) within genes/functional regions having multiple significant phenotype associations. This can be considered to be a pleiotropic pattern when the two phenotypes are seemingly unrelated. Two main MPA patterns exist within GWAS results. Type 1 MPAs occur when a single SNP within a functional region (such as a gene) is associated with more than one phenotype, whereas Type 2 MPAs occur when two different SNPs within a single functional region have different phenotype associations [7, 8] (Figure 3.1A and 3.1B).

Multivariate analysis of the results of GWAS studies across many phenotypes have allowed for the investigation of complex relationships between genes and phenotypes, including pleiotropic relationships and the clustering of variants based on their phenotype associations. Many of these studies have involved the analysis of SNP associations with complex human disease traits. Some studies have considered pleiotropy as genes and SNPs associated with more than one phenotype, and found that pleiotropic genes tended to be longer, and that SNPs within pleiotropic genes were more likely to be exonic [9]. Levine *et al.* (2017) extended Weighted Gene Co-expression Network Analysis (WGCNA) to cluster SNPs based on their phenotype associations using a matrix of beta coefficients, followed by hierarchical clustering of the Topological Overlap Matrix [10], and show how the resulting clusters can be used to produce polygenic scores. Gupta *et al.* (2011) introduced a biclustering algorithm, simultaneously clustering SNPs and phenotypes in a matrix of regression coefficients [11]. Network-based approaches have been developed which construct bipartite networks of gene-disease phenotype associations from GWAS, and constructed network projections of this bipartite network resulting in disease similarity and gene-similarity networks [12]. Though these studies provide a baseline of the use of multivariate and network approaches for the analysis of GWAS results, there is, to our knowledge, no method which characterizes detailed MPA signatures of genes and no method which clusters genes based on these detailed signatures. Simply clustering genes based on their phenotype associations will not capture the vast amount of combinatorial possibilities of type 1 and type 2 signatures any given gene can harbor (Figure 3.1C), especially when the multi-phenotype GWAS study involves millions of variants and hundreds of phenotypes.

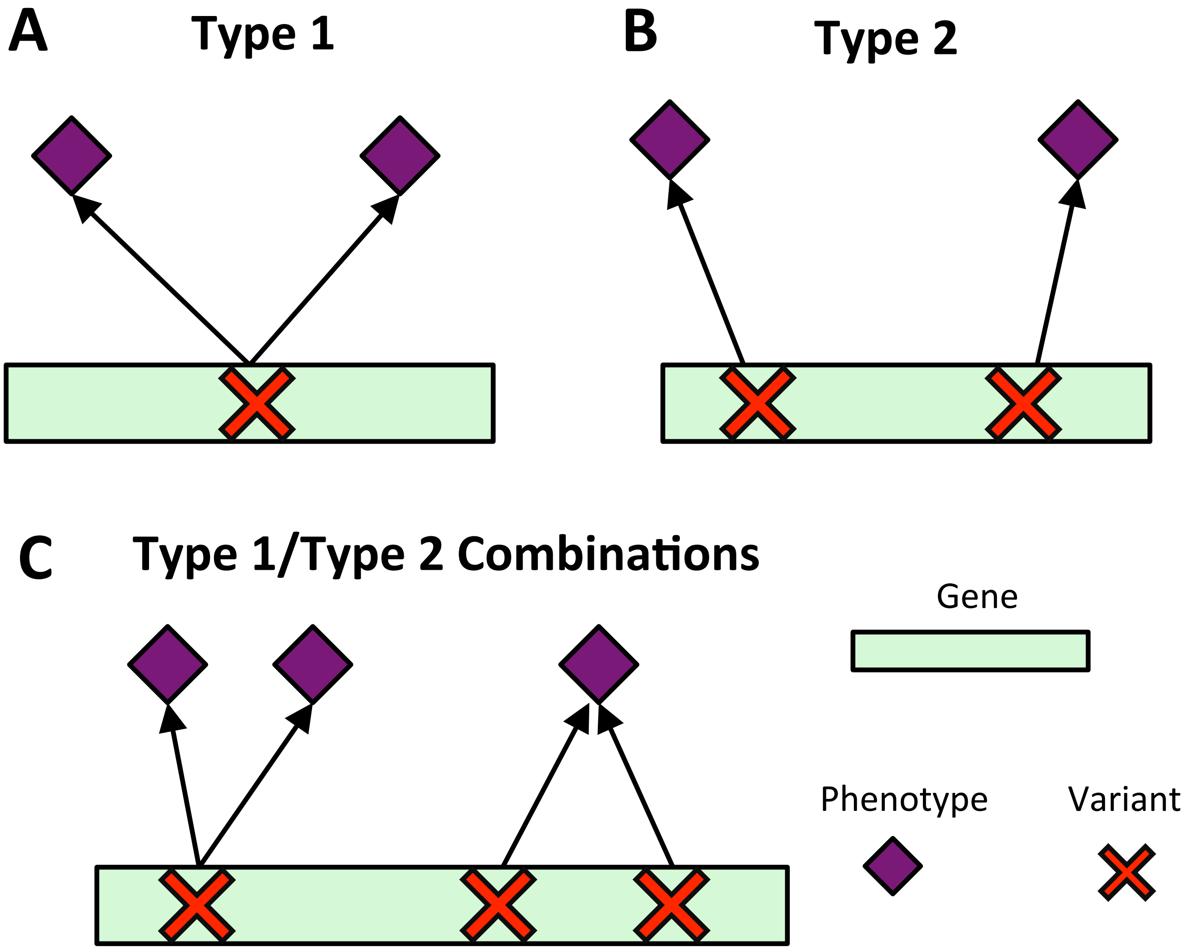


Figure 3.1: **MPA signatures.** (a) Type 1 MPA: a gene is associated with more than one phenotype due to a single variant within the gene associating with multiple phenotypes. (b) Type 2 MPA: a gene is associated with more than one phenotype because of alternate SNPs within the gene having different phenotypic associations. (Figure created from information presented in Solovieff *et al.* (2013) [7].) (c) Complex combinations of Type 1 and Type 2 signatures.

Methods for multi-trait GWAS have also been developed, associating variants to groups of phenotypes (see, for example Porter *et al.* (2017) [13], Thoen *et al.* (2017) [14], Furlotte *et al.* (2015) [15] and Stephens *et al.* (2013) [16]). Though these methods have value in their own right for identifying variants which affect a group of traits, they do not address the question of the architecture of variant-phenotype associations within a region, such as a gene, nor do they automatically allow for the clustering of regions based on the architecture of the variant-phenotype associations (MPA/pleiotropic signatures) they harbor.

To this end, we present MPA Decomposition and Signature Clustering, a network-based approach involving a constructed powerset space, in which clustering distinguishes between genes based on the detailed topology of their unique MPA signature. MPA decomposition provides a framework allowing the precise mathematical representation of the architecture of variant-phenotype associations within regions (MPA/pleiotropic signatures), and thus allows these regions (such as genes) to be clustered based on these complex signatures.

3.2 Methods and Materials

3.2.1 Overview

MPA decomposition involves the mathematical characterization of each gene's MPA signature in a network-based context. This process begins in phenotype space. In this multi-dimensional space, each axis represents a phenotype and genes are represented as points, with points close together representing genes with similar phenotype associations and points far apart representing genes with very different phenotype associations. This phenotype space provides no information on the topology of associations within each gene. MPA decomposition maps genes to a newly constructed *powerset space*, which is constructed though clustering of SNP association vectors (Figure 3.2A-E). This clustering produces discrete sets of SNPs/overlapping sets of phenotypes called *association modules* which form the axes of powerset space, which provides the detailed structure of phenotype associations within a gene. The second stage - signature clustering - groups genes based on their detailed MPA signature (Figure 3.2F). Clustering of genes in this space results in groups of genes with identical MPA signatures (Figure 3.2G-I). These genes grouped by MPA signatures provide a useful tool for the researcher planning genetic modification experiments, easily highlighting groups of genes with favorable signatures for modification to influence a particular phenotype.

The approach of MPA decomposition and its application are described below. We apply and demonstrate this method on GWAS results from a densely genotyped *Populus trichocarpa* GWAS population involving approximately 10 million SNPs and over 400 untargetted metabolomics phenotypes measured across the population.

3.2.2 *Populus trichocarpa* SNPs

P. trichocarpa [17] SNP data (DOI 10.13139/OLCF/1411410) obtained from [`https://doi.ccs.ornl.gov/ui/doi/55`] was derived from the whole genome resequencing of a Genome Wide Association Study (GWAS) population clonally replicated in common gardens [18]. This dataset consists of 28,342,758 SNPs called across 882 *P. trichocarpa* genotypes. Details on the generation of this SNP dataset can be found in [19]. VCFtools [20] was used to extract the most reliable set of SNPs corresponding to the 90% tranche, resulting in a set of 10,438,861 bi-allelic SNPs.

3.2.3 Metabolomics

Untargetted metabolomics was conducted on *P. trichocarpa* genotypes using GC-MS. The metabolite analysis used is described in Tschaplinski *et al.* (2014) [21]. Briefly, samples were freeze dried for 48 h and then ground with a microWiley mill with a 20 mesh screen, with samples then twice extracted in 80% ethanol (aqueous) and the extracts combined before an aliquot was dried under nitrogen. Dried extracts were dissolved in acetonitrile

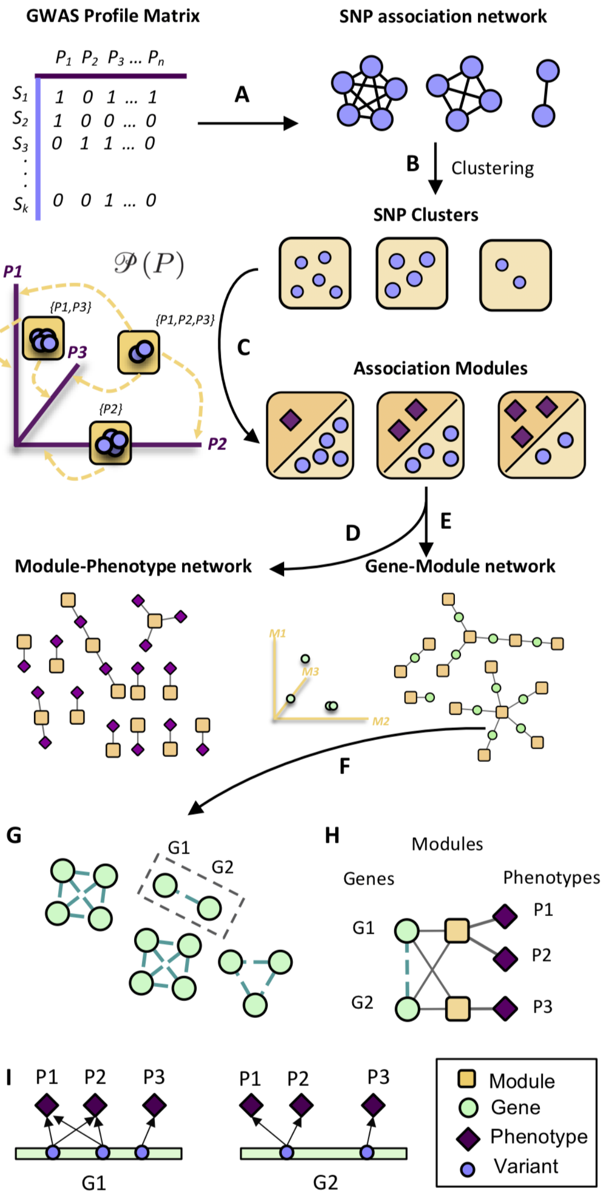


Figure 3.2: **Overview of MPA Decomposition and Signature Clustering.** (A) Construction of GWAS matrix and calculation of Proportional Similarity between all pairs of SNPs. (B) Clustering of the SNP association similarity network into groups of SNPs identical phenotype associations. (C) Association modules constructed as elements of the powerset of phenotypes observed in the SNP clusters. (D) Module-phenotype network links phenotypes to modules if phenotype is associated with all SNPs in module. (E) The gene-module network is constructed by mapping genes to association modules if the module contains a SNP that resides within that gene. (F) Signature clustering groups genes with the same module associations. (G,H) Clustering genes in powerset space results in groups of genes with the same pattern of MPA signatures with the same set of phenotypes. Modules exist as a layer between genes and phenotypes. (I) Example: Both G1 and G2 contain SNP(s) associating with both P1 and P2, as well as a SNP associating with only P3.

followed by the addition
N-methyl-N-trimethylsilyltrifluoroacetamide with 1% trimethylchlorosilane. Samples were heated for 1 hour at 70 to generate trimethylsilyl (TMS) derivatives. Samples were injected in an inert XL gas chromatograph-mass spectrometer (Agilent Technologies Inc., Santa Clara, CA, U.S.A.), fitted with an Rtx-5MS with Integra-Guard (5% diphenyl/95% dimethyl polysiloxane) capillary column (30 m by 250 μm by 0.25 μm film thickness) (Restek, Bellefonte, PA, U.S.A.). A standard quadrupole GC-MS was operated in the electron impact (70 eV) ionization mode, targeting 2.5 full-spectrum (50-650 Da) scans per second, as described previously [22]. A large user-created database (>2400 spectra) of mass spectral electron impact ionization fragmentation patterns of TMS-derivatized compounds, as well as the Wiley Registry 10th Edition with the NIST 2014 mass spectral database, were used to identify the metabolites of interest. Metabolites were quantified by extracting a key, characteristic mass-to-charge (m/z) for each known and unidentified metabolite using an automated data extraction program. Preprocessing of the resulting raw GC-MS data included alignment using XCMS [23] and normalization for amount of leaf sample analyzed, fraction of extracted sample analyzed, and internal standard recovered.

3.2.4 Outlier Analysis

We performed outlier detection on each of the respective phenotypes, to account for measurement variability and technical/experimental error, using R [24]. This determines which, if any, metabolite intensities, for a given genotype, are very different from those observed for other genotypes. We applied a variant of the method discussed in Leys *et al.* (2013) [25], using the median absolute deviation (MAD) from the median. Our approach differs in that it takes into account the asymmetry of the distribution of intensity values, as lower intensities are more frequent. We thus calculated the MAD for the upper and lower tails of the distribution separately. By investigating the distribution of intensities and the MAD distance from the median, for a random sample of metabolites, we determined that a MAD distance of 5 is appropriate for outlier detection, this was done using the ggplot2 package in R [26]. Any intensity value of a metabolite for a given genotype that was more than 5 MADs from the median was removed from the analysis. Also, to mitigate potential biases from under-represented metabolites, we excluded any metabolite that had less than 100 non-zero, non-outlier values.

3.2.5 GWAS

The EMMAX software [27] was used to statistically associate measured phenotypes with SNPs in *Populus trichocarpa*. Covariates were included to account for population structure by estimating a kinship matrix using the default parameters for Balding-Nichols method implemented in the emmax-kin program [28]. This was run in a parallel fashion using a customized Python script which made use of the NumPy [29], SciPY (http://www.scipy.org/) [30], pandas [31] and mpi4py [32, 33, 34] modules. The Benjamini-Hochberg stepwise procedure [35] was used to control the false discovery rate (FDR), and associations passing a FDR of 0.1 were considered significant associations. A total of 413 phenotypes had at least one significant SNP association, and 131,282 SNPs had at least one significant phenotype association.

3.2.6 Profile Matrix Construction

A GWAS profile matrix M was constructed in which each row represented a SNP that resides within a gene region, each column represented a phenotype and each entry M_{ij} was defined as:

$$M_{ij} = \begin{cases} 1 & \text{if SNP } i \text{ is associated with phenotype } j \\ 0 & \text{otherwise} \end{cases} \tag{3.1}$$

Each row of the matrix M represents the GWAS profile of a particular SNP. SNPs were mapped to their respective genes using the *P. trichocarpa* version 3 genome annotation [17] available on Phytozome [36] through the genome portal of the Department of Energy

Joint Genome Institute [37, 38]. A gene was considered to consist of its coding sequences as well as regulatory elements such as 5' and 3' UTRs.

3.2.7 Gene-Phenotype (GP) Matrix Construction

The gene-phenotype matrix GP was constructed from the GWAS results by mapping the SNPs in the GWAS profile matrix to genes. This would result in a matrix in which entry ij would be defined as 1 if gene i contained a SNP significantly associated with phenotype j, and zero otherwise, and then pruning the resulting matrix to include only genes with more than one phenotype association. Equivalently, this can be seen as constructing a network by creating an edge between a gene and phenotype if that particular gene contained a SNP significantly associated with that phenotype.

3.2.8 Association Module Construction

The procedure for the construction of association modules is shown in Figure 3.2, steps A though C. The GWAS profiles of all pairs of SNPs in the GWAS profile matrix M were compared by calculating the Proportional Similarity Index between all pairs of rows of M. The Proportional Similarity Index between two vectors X and Y is defined as [39]:

$$PS(X,Y) = \frac{2\sum_i \min(x_i, y_i)}{\sum_i (x_i + y_i)} \tag{3.2}$$

where X and Y are the GWAS profiles of two SNPs (i.e. two rows of the matrix M), x_i is the ith entry in row X and y_i is the ith entry in row Y. This was performed in parallel using a customized Perl script which made use of the Parallel::MPI::Simple Perl module, developed by Alex Gough and available on The Comprehensive Perl Archive Network (CPAN) at www.cpan.org. This all-vs-all comparison results in a complete, unpruned SNP association network in which nodes represent SNPs and edges represent the similarity between the phenotype associations of SNPs.

We extracted association modules from the SNP association network as follows: First we identify SNPs that reside within genes with multiple phenotype associations (MPA genes). We extracted SNPs within MPA genes and the edges between these SNPs, and then pruned the network to only include edges between SNPs which have identical phenotype associations. This was achieved by applying a Proportional Similarity threshold of 1 (Supplementary Text S3.1, S3.2, S3.3, Figure S3.1). Nodes of the resulting subnetwork were then clustered into groups using MCL [40, 41] available from http://micans.org/mcl/. Each resulting cluster represents a group of SNPs with the same phenotype associations, i.e. a group of SNPs driven together by a particular set of phenotypes, or, an element of the powerset of phenotypes. These *modules* of phenotypes form the axes of the powerset space.

3.2.9 Decomposition Matrix Construction

The procedure for decomposition matrix construction is shown in Figure 3.2, steps D and E. The GM matrix was constructed by mapping modules to genes which contained SNPs within that module. Thus, the GM matrix was constructed such that each entry ij was defined as 1 if module j contained a SNP that resides within gene i, and zero otherwise. This can also be seen as constructing a network by connecting gene nodes to module nodes which contain SNPs that reside within that gene region.

The MP matrix was constructed by mapping modules to phenotypes which drive the association between SNPs within the module. Thus, the MP matrix was constructed such that each entry ij was defined as 1 if phenotype j had a significant GWAS association with all SNPs in module i. This could alternatively be seen as creating a network by connecting phenotype nodes to module nodes if that phenotype has a GWAS association with all SNPs in that module.

The GP, MP and GM matrices represent MPA matrices, and the GM and MP matrices are referred to as the decomposition matrices (Supplementary Text S3.4, S3.5, Figure S3.2).

These matrices were visualized as bipartite networks in Cytoscape [42] using an Allegro layout.

3.2.10 Signature Clustering

Signature clustering was performed by calculating the similarity between all pairs of rows (genes) of the GM matrix using the proportional similarity metric (described above), applying a threshold of 1, and clustering the resulting similarity network using MCL [40, 41].

3.2.11 Annotation and Functional Enrichment

P. trichocarpa gene boundaries as defined in the Ptrichocarpa_210_v3.0.gene.gff3 annotation file obtained from version 3 genome annotation [17] available on Phytozome was used. Functional annotations of *P. trichocarpa* genes were obtained from version 3 genome annotation [17] available on Phytozome [36] through the genome portal of the Department of Energy Joint Genome Institute [37, 38].

Mapman annotations of *P. trichocarpa* were obtained by splitting the protein translations of *P. trichocarpa* genes into three sets and using the Meractor tool [43] to assign Mapman terms to each gene. The BINGO Cytoscape plugin [44] was used to determine enriched Gene Ontology (GO) terms in the set of type 1 and type 2 MPA genes.

3.2.12 Co-expression Network

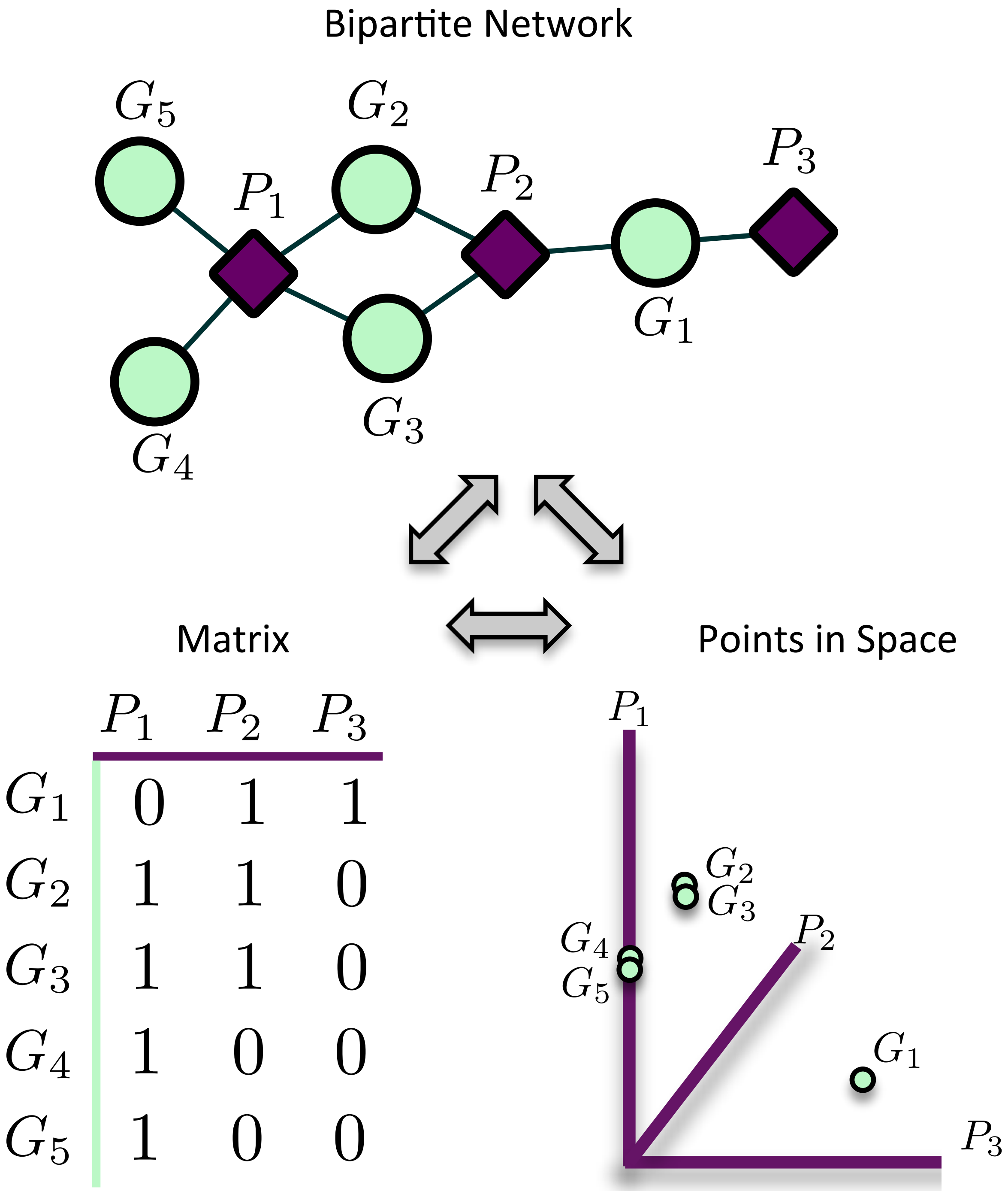
A *P. trichocarpa* gene co-expression network was constructed as described in Weighill *et al.* (2018) [19] making use of the *P. trichocarpa* (Nisqually-1) RNA-seq data derived from JGI Plant Gene Atlas project (Sreedasyam et al., unpublished), consisting of samples for various tissues (leaf, stem, root and bud tissue) and libraries generated from nitrogen source study. A list of sample descriptions was accessed from Phytozome at https://phytozome.jgi.doe.gov/phytomine/aspect.do?name=Expression.

3.3 Results and Discussion

3.3.1 MPA Decomposition: Construction of a New Space

Genotyping of 882 *P. trichocarpa* genotypes and metabolic profiling of 585 of these genotypes, followed by GWAS analysis provided a network of associations between SNPs and metabolic phenotypes (see *Methods and Materials*). Mapping of these SNPs to the genes in which they reside resulted in gene-to-phenotype associations. These gene-phenotype associations can be represented as multiple different data structures. Genes can be represented as points in multi-dimensional phenotype space, indicating their respective phenotype associations (Figure 3.3). The closer genes are to each other in phenotype space, the more shared phenotype associations they have. Alternatively, these associations can be represented as a bipartite network, linking a gene g_i to phenotype p_k if g_i contained a SNP significantly associated with p_k (Figure 3.3). Bipartite networks are useful for the visualization and investigation of points in high dimensional space, as well as for the representation of complex relationships between multiple objects. Thus, bipartite networks were used throughout MPA decomposition as the mathematical foundation as well as a visualization tool.

GWAS results represented as a bipartite network of SNPs connected to their associated phenotypes (Figure 3.4a) do not give any indication of MPA signatures as there is no obvious information about which SNPs belong to which genes. Thus, bipartite SNP-phenotype networks give no indication of how many phenotype associations a given gene has. GWAS results represented as a bipartite network of genes connected to their associated phenotypes (Figure 3.4b) can give an indication as to whether or not a gene has multiple phenotype associations in that it is associated with more than one phenotype, but cannot give any indication as to the type of MPA signature (type 1 or type 2) exhibited



$$\begin{array}{c|ccc} & P_1 & P_2 & P_3 \\ \hline G_1 & 0 & 1 & 1 \\ G_2 & 1 & 1 & 0 \\ G_3 & 1 & 1 & 0 \\ G_4 & 1 & 0 & 0 \\ G_5 & 1 & 0 & 0 \end{array}$$

Figure 3.3: **Representation of matrices as spaces and bipartite networks.** Matrices of GWAS results can easily be represented as points in high dimensional space, with rows representing points and columns representing variables/axes. Equivalently, matrices can be represented as bipartite networks, connecting row objects (genes) with column variables if the corresponding entry is non-zero. This provides a useful way to visualize high dimensional spaces as bipartite networks.

by the gene. Mapping the SNPs in the SNP-phenotype network to the genes in which they are present results in a gene-SNP-phenotype network (Figure 3.4c). From this network, it is possible to deduce the type of MPA signature exhibited by a gene through some amount of visual inspection, for example, looking at the SNPs within a gene and what their associated phenotypes are. However, the structure of this network does not allow the MPA signature of a gene to be readily extracted using simple node properties such as degree. For example, one cannot simply calculate the connectivity (degree) of each gene node in Figure 3.4c in order to determine the type of MPA signature exhibited, since one can have multiple SNPs within the same gene associating with the same set of phenotypes. In addition, it is not easy to determine which genes exhibit the same MPA signatures. The process of MPA decomposition allows one to maintain the topology of SNP associations within a gene while still being able to determine the type of MPA signature using simple network measures such as degree.

The first phase of MPA decomposition involved the construction of the powerset space, a new multi-dimensional space in which each dimension/axis represents a particular subset of phenotypes. Construction of this space is described in *Methods and Materials*. Briefly, we calculated the similarity between all pairs of SNP vectors, based on their phenotype associations. Applying a threshold and clustering the resulting SNP associations resulted in modules of SNPs with the same phenotype associations (Figure 3.2A-C). While representing non-overlapping sets of SNPs, these modules also represented overlapping sets of phenotypes. In particular, each module represented the set of phenotypes which were associated with all SNPs within the module. Thus, each module also represented an element of the powerset of phenotypes $P(P)$ observed in the SNP-phenotype GWAS associations. These observed elements of the powerset were used to construct the powerset space, with each element/module representing a different dimension of this space (Figures 3.2C, 3.2D). Genes were subsequently mapped from phenotype space into powerset space (Figure 3.2E). Thus, SNP elements were used to generate the modules of powerset space, but genes where the elements mapped in the powerset space. Represented as bipartite networks, the module-phenotype bipartite network defined the axes of powerset space, and the gene-module bipartite network mapped the genes into powerset space. While

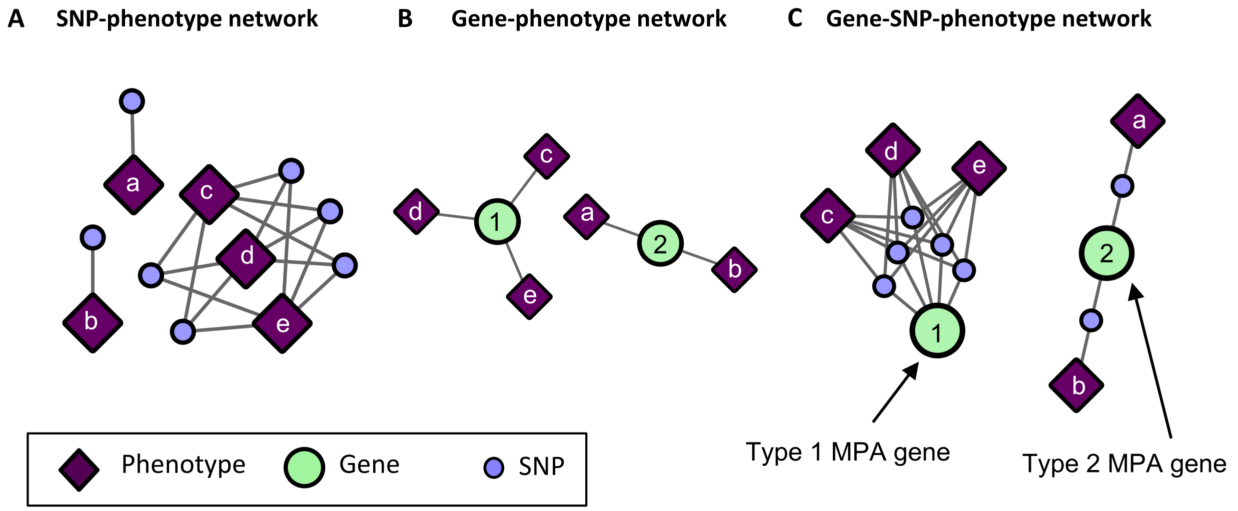


Figure 3.4: **Example of SNP-phenotype, gene-phenotype networks and gene-SNP-phenotype networks.** (a) SNP-phenotype bipartite networks simply connect SNPs to phenotypes with which they have a significant association, and do not provide information regarding MPA signatures within genes. (b) Gene-phenotype networks contain connections between genes and phenotypes. An edge will be drawn between a gene and a phenotype if that gene contains a SNP associated with that phenotype. Gene-phenotype networks do not provide information as to which type of MPA signature is exhibited. (c) Gene-SNP-phenotype networks are SNP-phenotype networks with the SNPs connected to genes in which they reside. These networks are more complicated, and MPA signatures can be deduced from their structure through further analysis, however, the network is not in a form in which MPA signatures can be extracted easily using standard network topology measures such as degree.

phenotype space provided information as to the individual phenotype associations of genes, powerset space indicated a gene's associations with sets of phenotypes at the SNP level, providing a detailed MPA signature. The mapping from phenotype space to powerset space was defined in terms of three bipartite networks, linked by a decomposition relationship (Figure 3.5, Supplementary Text S3.6), namely a gene-phenotype (GP), a gene-module (GM) and a module-phenotype (MP) network. In the GP network (Figure 3.6), nodes represented either genes or phenotypes, and an edge was defined between gene G_i and phenotype P_j if gene G_i contained a SNP which was statistically associated with phenotype P_j in the GWAS analysis. Nodes in the GM network (Figure 3.7) represented either genes or modules, and an edge was defined between gene G_i and module M_j if M_j contained a SNP that resided within gene G_i. Nodes in the MP network (Figure 3.8) represented either association modules or phenotypes, and an edge was defined between module M_i and phenotype P_j if the correlation of SNPs within M_i is driven by phenotype P_j.

3.3.2 Powerset Space Unravels Multi-Phenotype Association Signatures

The GP network (Figure 3.6) represents genes in phenotype space, and provides information regarding which genes are associated with which phenotypes, and can thus indicate which genes are have multiple phenotype associations and are potentially pleiotropic. Of the 41,335 genes in *P. trichocarpa*, 2,964 genes had GWAS hits with more than 1 metabolite phenotype each, and are thus considered MPA genes with respect to the metabolic phenotypes.

The GM network (Figure 3.7) represents genes in powerset space, which in turn is defined by the MP network (Figure 3.8). The GM network unravels the MPA signatures of genes, representing their associations with sets of phenotypes. Genes that are connected to one module exhibit a Type 1 MPA signature because they contain SNPs which are associating with the same set of phenotypes, whereas genes connected to more than one module

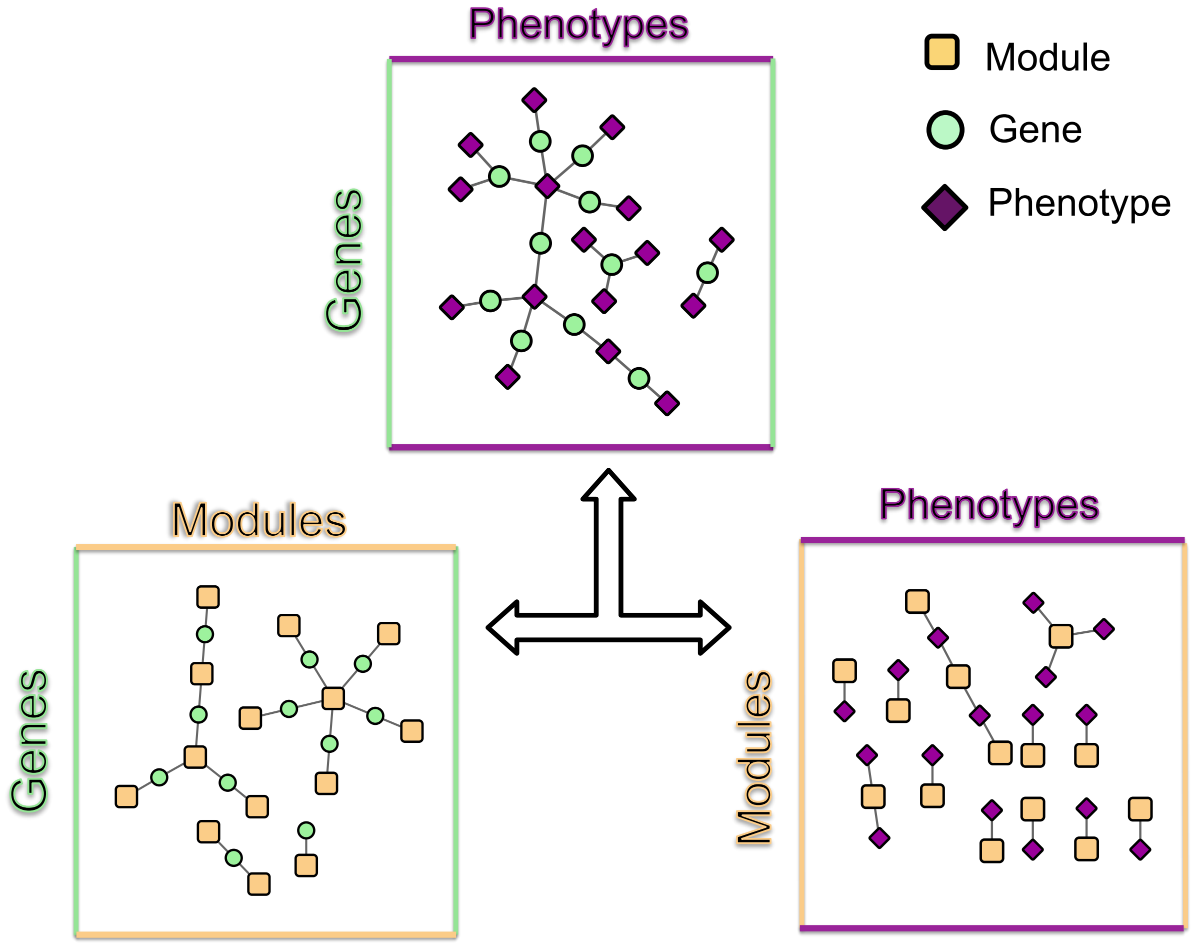


Figure 3.5: **MPA Decomposition.** The gene-phenotype matrix is decomposed into two matrices, a gene-module (GM) matrix and a module-phenotype (MP) matrix (Supplementary Text S3.4 and S3.5). The GM matrix represents genes in powerset space. *Association modules* (elements of the powerset of phenotypes) form the basic units of MPAs and are considered latent variables. Signature clustering is performed on genes in module space (GM matrix).

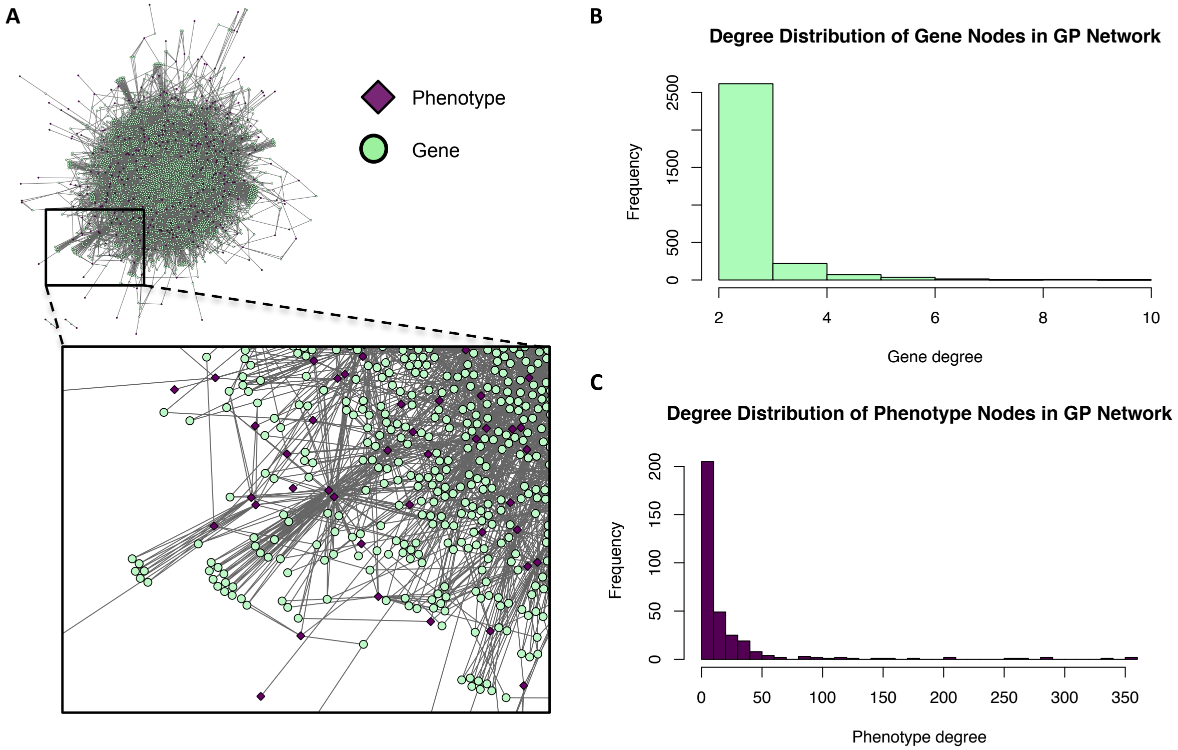


Figure 3.6: **Gene-Phenotype (GP) Network.** (a) The GP network. Green nodes represent MPA genes, pink diamonds represent metabolites (phenotypes). An edge connects a gene to a phenotype if that gene contains a SNP associated with that phenotype. (b) Degree distribution of the gene (green) nodes in the GP network. (c) Degree distribution of the phenotype (pink) nodes in the GP network.

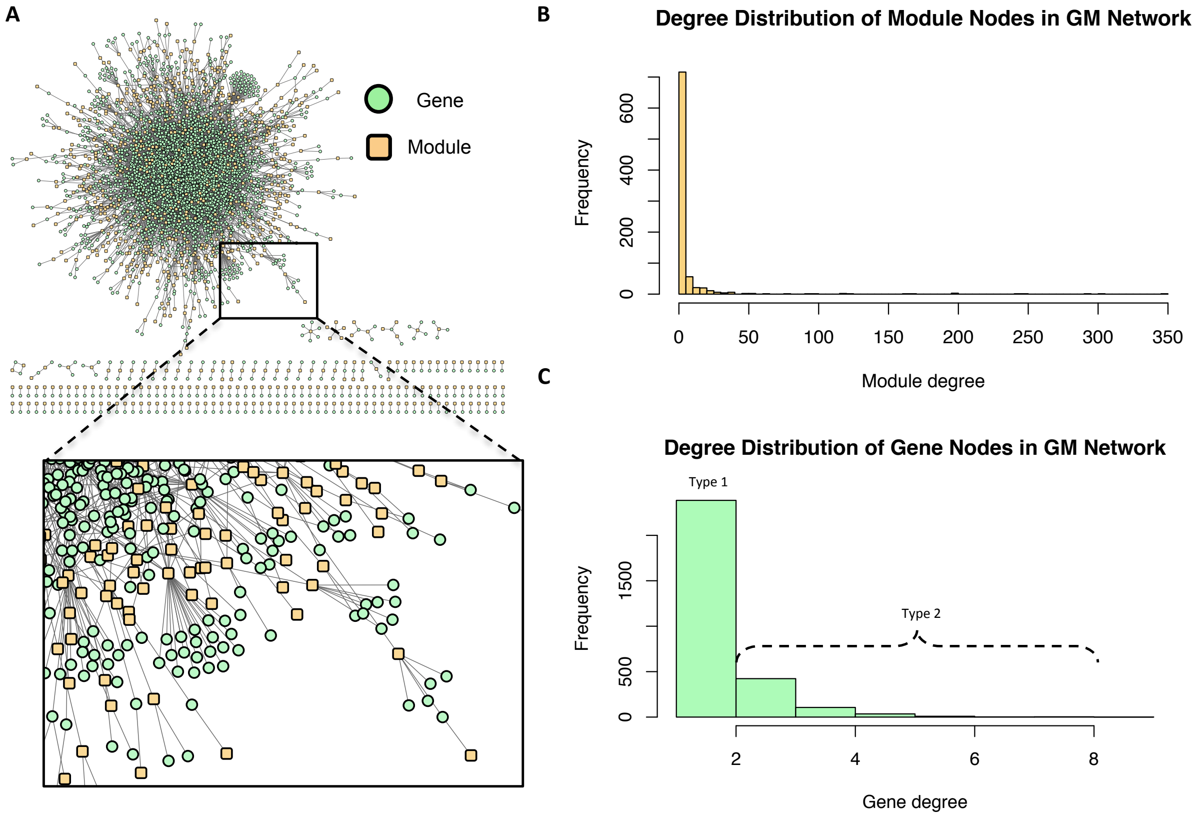


Figure 3.7: **Gene-Module (GM) Network.** (a) The GM network. Green nodes represent MPA genes and yellow nodes represent association modules. A gene node is connected to a module node if the module contains a SNP which resides within that gene. (b) Degree distribution of the module (yellow) nodes in the GM network. (c) Degree distribution of the gene (green) nodes in the GM network.

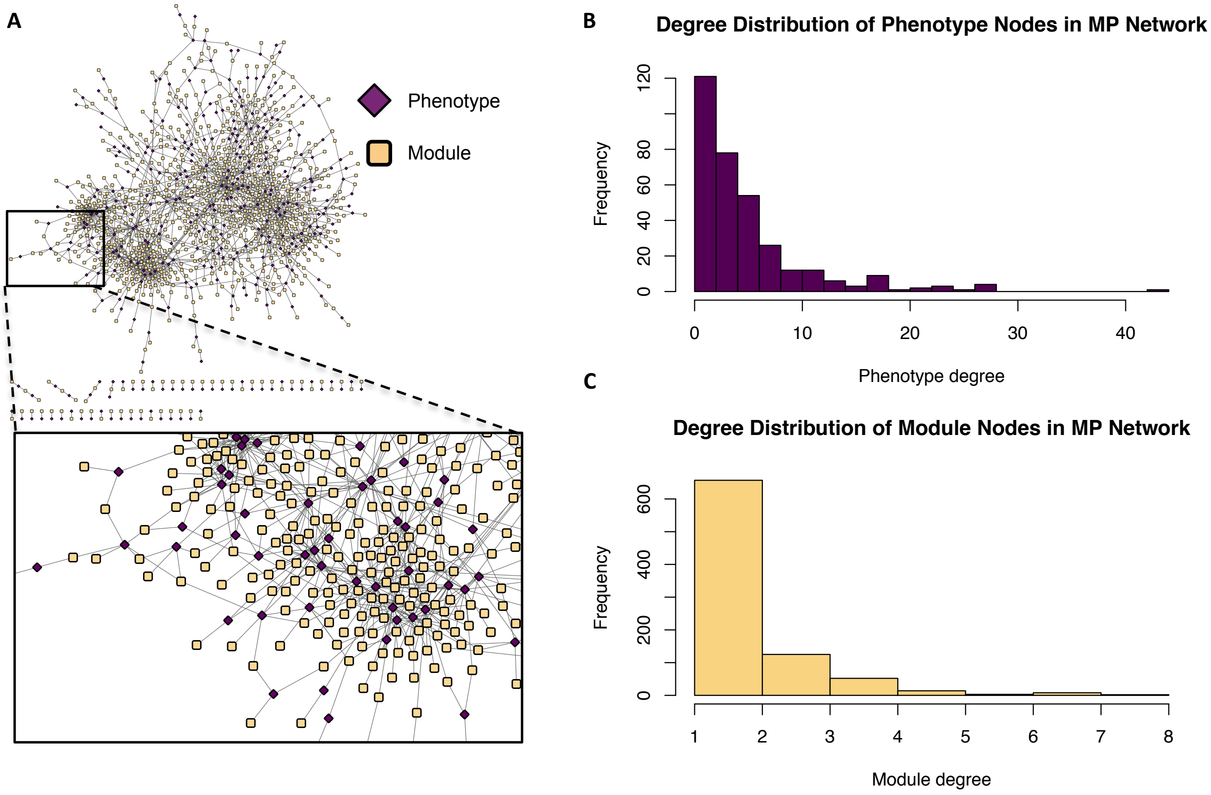


Figure 3.8: **Module-Phenotype (MP) Network.** (a) The MP network. Yellow nodes represent association modules and pink nodes represent phenotypes. A module node is connected to a phenotype node if the phenotype is associated with all SNPs within the module and is thus considered a driving phenotype of the module. (b) Degree distribution of the phenotype (pink) nodes in the MP network. (c) Degree distribution of the module (yellow) nodes in the MP network.

exhibit a Type 2 MPA signature because they contain SNPs which associate with different sets of phenotypes. Mapping of genes to module space thus reveals the Type 1 and Type 2 MPA patterns, as well as complex combinations of Type 1/Type 2 patterns that exist within genes (Figure 3.9). Phenotype associations of genes cannot be distinguished as Type 1 or Type 2 in phenotype space, whereas module space clearly indicates the MPA signature exhibited by a gene, revealing precisely which sets of phenotypes each individual SNP in a gene is associated with (Figure 3.9). Module space also goes beyond classifying genes as exhibiting Type 1 or Type 2 MPA signatures, but characterizes each unique topology of variant-phenotype associations within a gene separately. The high density of SNPs in this population and the rapid decay of LD allows for the high resolution of MPA signatures. Figure S3.3A shows the variation in LD in the region including 5kb upstream and downstream of Potri.001G419800, the type 2 MPA gene in Figure 3.9F. One can see that both associating variants in this gene are in a region of low LD. Figure S3.3B shows a pairwise LD heatmap of 100 variants in this region including the two associating variants in Potri.001G419800. One can see that these two associating variants exist within two separate LD blocks.

The beta value derived from each SNP-phenotype association gives an indication of the effect that the SNP has on the value of the phenotype. One can look at the beta values from the GWAS analysis to see if the minor allele of a given SNP has statistically a positive or negative affect on the phenotype value. This will inform the researcher of the potential functional affect of each SNP. Overall, positive and negative beta values are present in associations in the set of type 1 MPA genes, type 2 MPA genes and single phenotype association (SPA) genes, although negative beta values are far more prevalent across all categories (Figure S4S3.4 indicating that most minor alleles have negative effects on the phenotype (metabolite) values.

Of the 10,566 genes that had at least one phenotype hit, 2,964 exhibited a MPA signature by associating with more than one phenotype (Figure S3.5A). Of those MPA genes, type 2 MPA signatures were far more abundant, with 2,468 genes exhibiting a type 2 MPA signature and 496 genes exhibiting a type 1 MPA signature (Supplementary

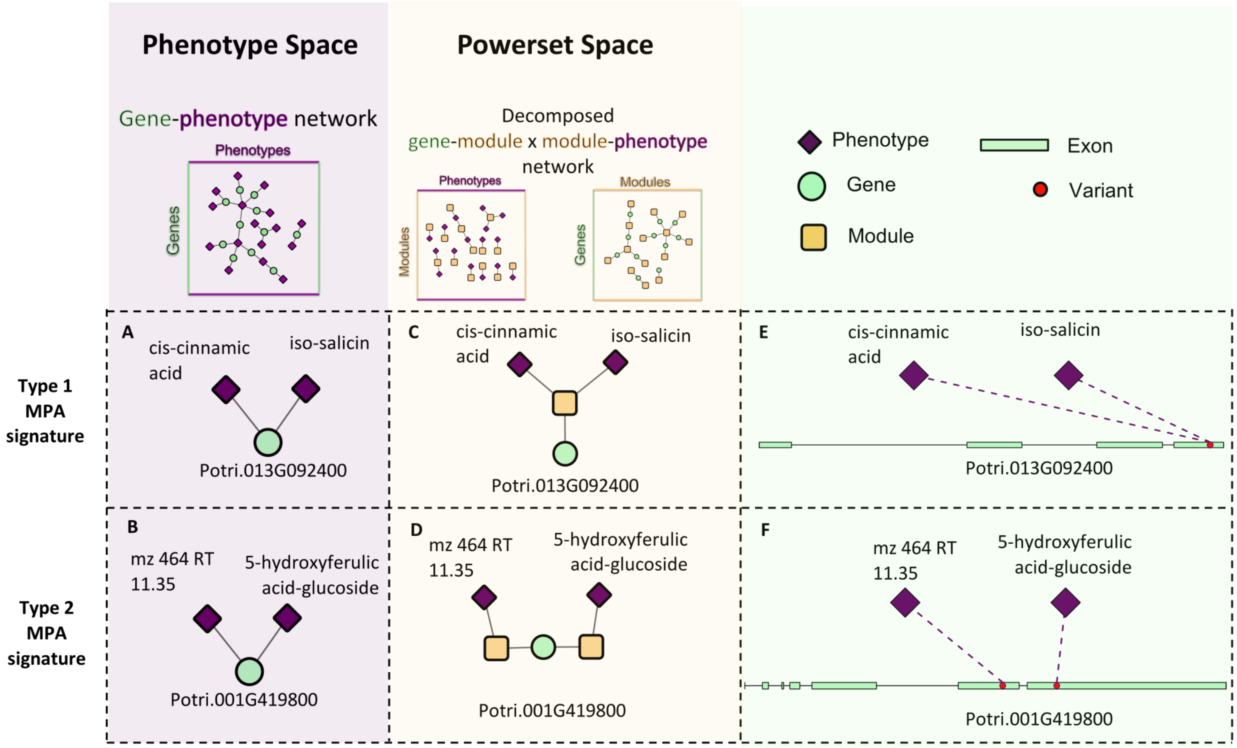


Figure 3.9: **Signature Decomposition Example.** Two genes, Potri.013G092400 (A) and Potri.001G419800 (B) have the same surrounding network topology in the GP network in that they are both connected to two phenotypes. Projecting the genes into powerset space though MPA decomposition of the GP network indicates that they exhibit different MPA signatures in that Potri.013G092400 exhibits a type 1 MPA signature (C), containing a SNP associating with two phenotypes (E) and Potri.001G419800 exhibits a type 2 MPA signature. (D) containing two SNPs, each with a different phenotype association (F).

Table S3.1, Figure S3.5B). MPA genes represented a broad range of functions (Figure S3.6). No functional enrichment was found in the set of type 1 MPA genes. However, various GO terms were found to be enriched in the set of type 2 MPA genes, including developmental functions such as root development, shoot development, leaf development, fruit development, symbiosis, encompassing mutualism through parasitism, various regulatory functions such as RNA gene silencing function and response to stress and DNA repair (see Supplementary Figures S3.7, S3.8, S3.9, Supplementary Table S3.2, Supplementary File S3.1 for complete enrichment results).

Chaperones are classic examples of pleiotropic genes, assisting in the folding of various proteins [45, 46, 47]. Querying the MPA networks for potential pleiotropic chaperones, we uncovered 14 potential chaperones based on there best *Arabidopsis* hit annotation, that contain MPA signatures (Supplementary Table S3.1), 12 of which contain type 2 MPA signatures. It is encouraging to see these classic pleiotropic genes appearing in the MPA networks, and interesting that they mostly exhibit type 2 MPA signatures.

3.3.3 Signature Clustering in Powerset Space

Clustering of genes in phenotype space produces groups of genes with the same overall set of phenotype associations. However, it does not provide any information as to the topology of Type 1/Type 2 associations of SNPs within the gene. Powerset space is defined by sets of phenotypes, and thus, clustering genes in this space groups genes based on the topology of Type 1/Type 2 associations of SNPs within the gene. After mapping genes to the newly constructed powerset space, genes were clustered (Figure 3.2F, *Methods and Materials*) resulting in groups of genes containing the same MPA signature. Members of a given cluster represented genes harboring identical MPA signatures. This means that genes within the same signature cluster have associations with the same modules. For example, the signature cluster driven by two modules, one involving associations with cis-3-O-caffeoyl-quinate and the other involving associations with gentisic acid-2-O-glucoside contains two genes, Potri.016G125500.v3.0 (homolog of *Arabidopsis thaliana* TRICHOME

Table 3.1: IDs, *Arabidopsis thaliana* best hits and corresponding descriptions of genes in the gentisic acid/cis-3-caffeoyl-quinate signature cluster (Figure 3.10).

Gene ID	***A. thaliana* Best Hit**	**Description**
Potri.012G132600	AT2G45650	AGAMOUS-like 6
Potri.016G125500	AT2G38320	TRICHOME BIREFRINGENCE-LIKE 34

BIREFRINGENCE-LIKE 34) and Potri.012G132600.v3.0 (homolog of *Arabidopsis thaliana* AGAMOUS-like 6). These genes have associations with both cis-3-O-caffeoyl-quinate and gentisic acid-2-O-glucoside, however a given SNP within these genes is associated with either caffeoyl-quinate *or* gentisic acid-2-O-glucoside, but not both (Figure 3.10). This exemplifies what MPA decomposition and signature clustering accomplishes - the extraction of detailed multi-phenotype association signatures within genes, and the grouping of genes based on these detailed MPA signatures.

MPA signature clusters varied in size and complexity, ranging from large sets of genes having simple MPA signatures (Figures 3.11A, 3.11B) to single gene clusters harboring very complex MPA signatures (Figures 3.11C, 3.11D). An inverse relationship existed between the cluster size, and the number of associated phenotypes, with a minimum gene cluster size of one and a maximum gene cluster size of 42 (Figure S3.10). Complex MPA signatures are possible in this population partly because of the rapid rate with which Linkage Disequilibrium (LD) decays, dropping below 0.2 within 100bp (Figure S3.11).

These signature clusters are easily combined with other data types in a "lines of evidence" fashion, as introduced in Weighill *et al.* (2018) [19]. Signature clusters such as those in Figure 3.10 can be merged with their neighbors in a co-expression network, providing additional insights into the functioning of these genes. Potri.016G125500 (TBL34) and Potri.012G132600 (AGL6) appeared in the same signature cluster, and are associated with many cell-wall related genes/phenotypes. TBL34 and AGL6 both associated with gentisic acid-2-O-glucoside and cis-3-O-caffeoyl-quinate, and both co-expressed with the same two transcription factors (Figure 3.12). An interesting regulatory circuit is potentially revealed, in that AGL6 potentially activates two transcription factors (positive co-expression edges)

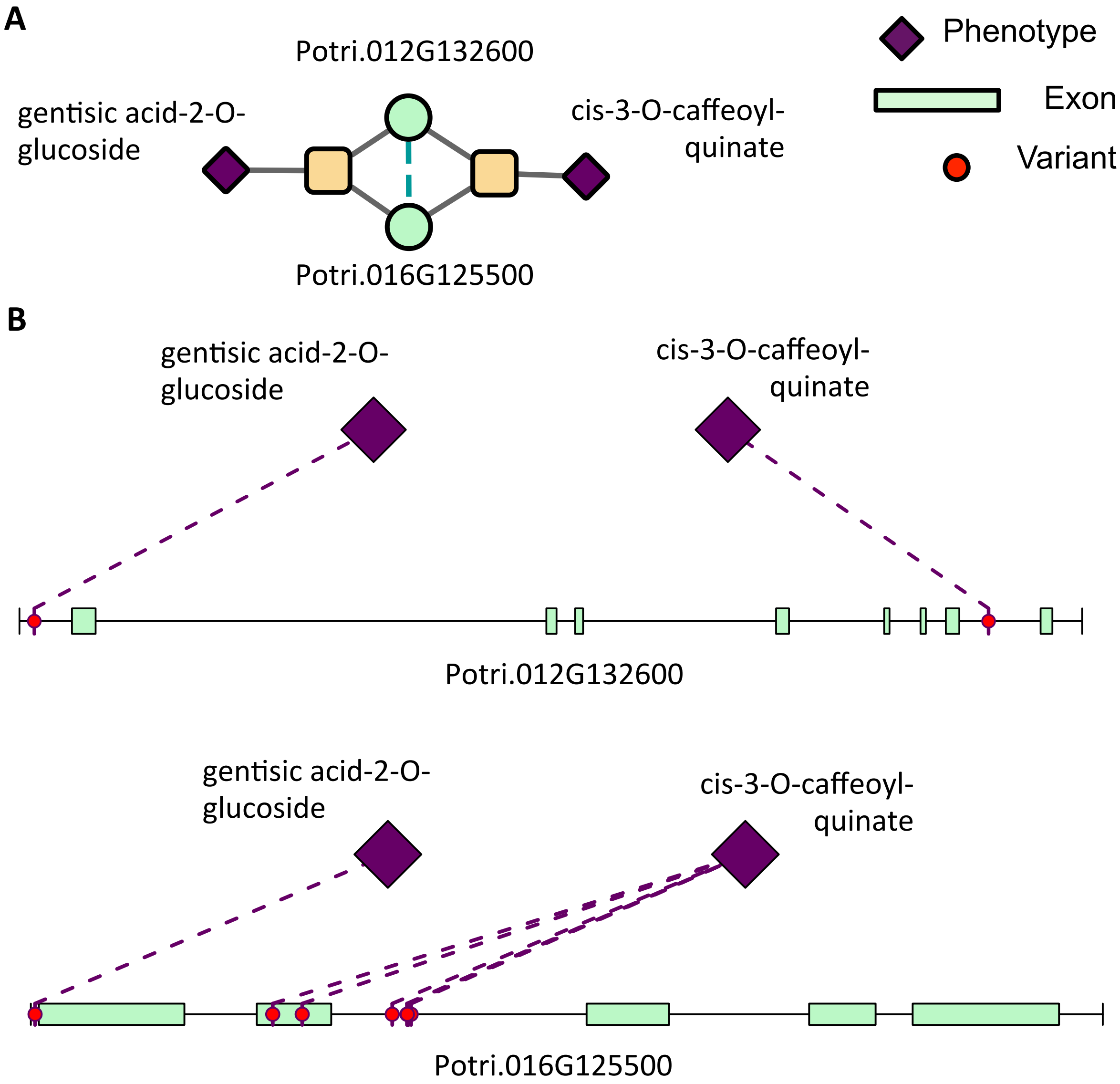


Figure 3.10: **Type 2 Signature Cluster.** (A) Signature cluster defined by a Type 2 association with gentisic acid-2-O-glucoside and cis-3-O-caffeoyl-quinate. (B) Associating SNP positions within genes in this signature cluster. These SNP associations have negative effect sizes (beta values) on the phenotype values. See Table 3.1 for gene information.

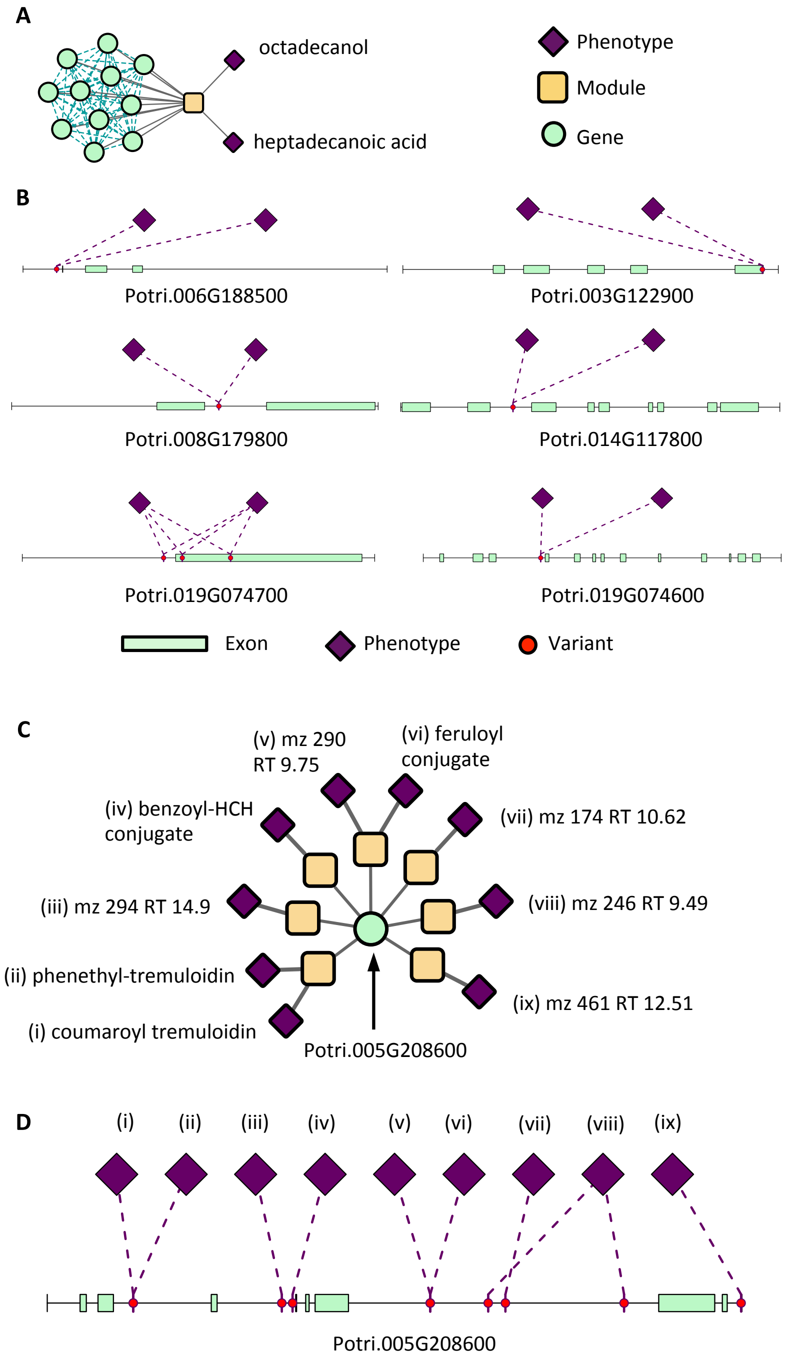


Figure 3.11: **Simple and Complex MPA Signatures.** (A) Signature cluster defined by a type one SNP association with octadecanol and heptadecanoic acid. See Table 3.2 for gene information. (B) Associating SNP positions within a selection of the genes in this signature cluster. These SNP associations have negative effect sizes (beta values) on the phenotype values. (C) Single-gene cluster of Potri.005G208600, bearing a unique, complex MPA signature consisting of 7 modules and 9 phenotypes. (D) Associating SNP positions of Potri.005G208600. These SNP associations have negative effect sizes (beta values) on the phenotype values.

Table 3.2: IDs, *Arabidopsis thaliana* best hits and corresponding descriptions of genes in the fatty acid signature cluster (Figure 3.11).

Gene ID	***A. thaliana* Best Hit**	**Description**
Potri.003G122900	AT1G63120	RHOMBOID-like 2
Potri.006G188500	AT4G31985	Ribosomal protein L39 family protein
Potri.008G179800	AT3G26000	Ribonuclease inhibitor
Potri.014G117800	AT2G47230	DOMAIN OF UNKNOWN FUNCTION 724 6
Potri.019G074600	AT4G10030	alpha/beta-Hydrolases superfamily protein
Potri.019G074700	AT1G71490	Tetratricopeptide repeat (TPR)-like superfamily protein
Potri.019G075000	AT3G44540	fatty acid reductase 4
Potri.019G075200	AT3G44540	fatty acid reductase 4
Potri.019G075300	AT4G33790, AT3G44540	fatty acid reductase 4, Jojoba acyl CoA reductase-related male sterility protein
Potri.019G075400	AT1G71460	Pentatricopeptide repeat (PPR-like) superfamily protein
Potri.019G087100	AT4G12600	Ribosomal protein L7Ae/L30e/S12e/Gadd45 family protein

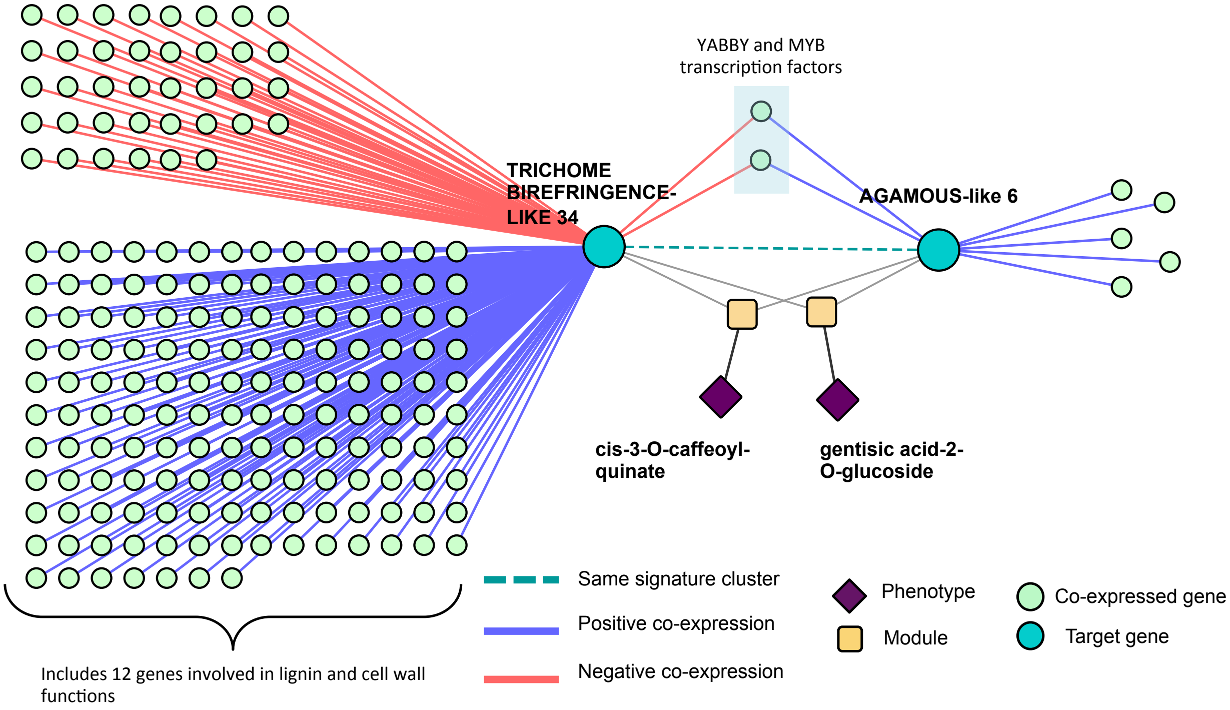

YABBY and MYB transcription factors
TRICHOME BIREFRINGENCE-LIKE 34
AGAMOUS-like 6
cis-3-O-caffeoyl-quinate
gentisic acid-2-O-glucoside
Includes 12 genes involved in lignin and cell wall functions
Same signature cluster
Positive co-expression
Negative co-expression
Phenotype
Module
Co-expressed gene
Target gene

Figure 3.12: **Co-expression lines of evidence.** Co-expression relationships of the signature cluster consisting of TBL34 and AGL6 from Figure 3.10.

which, in turn potentially repress TBL34 (negative co-expression edges). TBL34 is also positively co-expressed with 12 genes involved in cell wall and lignin biosynthesis functions (Figure 3.12). TBL genes are known to o-acetylate xylose [48], a function which has been found to be essential for resistance to certain pathogens [49]. Gentisic acid and its conjugate is a pathogen-induced signalling molecule [50] which itself has been found to induce pathogen resistance in plants [51] and induce expression of pathogenesis-related proteins [50]. Various AGL genes are also cell-wall related in that they impact lignin content [52, 53, 54]. This could be a regulatory circuit of biotic-stress-related cell wall remodeling, in which AGL6 potentially regulates xylose o-acetylation via TBL34.

3.3.4 Extensions to Pleiotropy

Several definitions of pleiotropy involve a gene associating with multiple, apparently disparate, unrelated phenotypes (see for example Stearns *et al.* (2010) [55]), and not all MPAs can be interpreted as pleiotropic signatures. However, if the two phenotypes are disparate enough, one can start to hypothesize about potential pleiotropic functioning of the gene in question. In this particular study, we demonstrated our method on a collection of molecular phenotypes of metabolite concentrations. If two metabolites in a MPA exist within separate pathways, one could consider it a potentially pleiotropic interaction.

A particular example of this phenomenon found in our analysis is Potri.002G178400. This gene has a type 2 MPA association with shikimic acid and raffinose (Figure 3.13). Based on existing knowledge found in PlantCyc on the Plant Metabolic Network (PMN) online resource [56], these two metabolites are found in different pathways. Shikimic acid is involved in reactions in pathways “chlorogenic acid biosynthesis I”, “chlorogenic acid biosynthesis II”, “phaselate biosynthesis”, “phenylpropanoid biosynthesis”, “simple coumarins biosynthesis”, and “chorismate biosynthesis from 3-dehydroquinate” whereas raffinose is involved in reactions in pathways “lychnose and isolychnose biosynthesis”, “stellariose and mediose biosynthesis”, “ajugose biosynthesis II (galactinol-independent)”, “stachyose degradation” and“stachyose biosynthesis”. Supplementary File S3.2 contains a high resolution PDF showing the positions of raffinose (red boxes) and shikimic acid (blue box) in the *P. trichocarpa* Cellular Overview metabolic map generated on the Plant Metabolic Network online resource. Potri.002G178400 contains two Pfam domains, namely pfam01565 (FAD binding domain) and pfam04030 (D-arabinono-1,4-lactone oxidase). This is an interesting example of a potentially pleiotropic gene, which affects two different metabolic phenotypes. A possible explanation for the mechanism of this pleiotropic interaction is through competition for carbon, with shikimic acid committing carbon to secondary metabolism and raffinose being the product of storage for primary carbon metabolism.

3.3.5 Future Prospects and Implications

P. trichocarpa was an ideal species for the demonstration of the MPA decomposition for several reasons. Firstly, a large collection of 1,100 *P. trichocarpa* accessions have

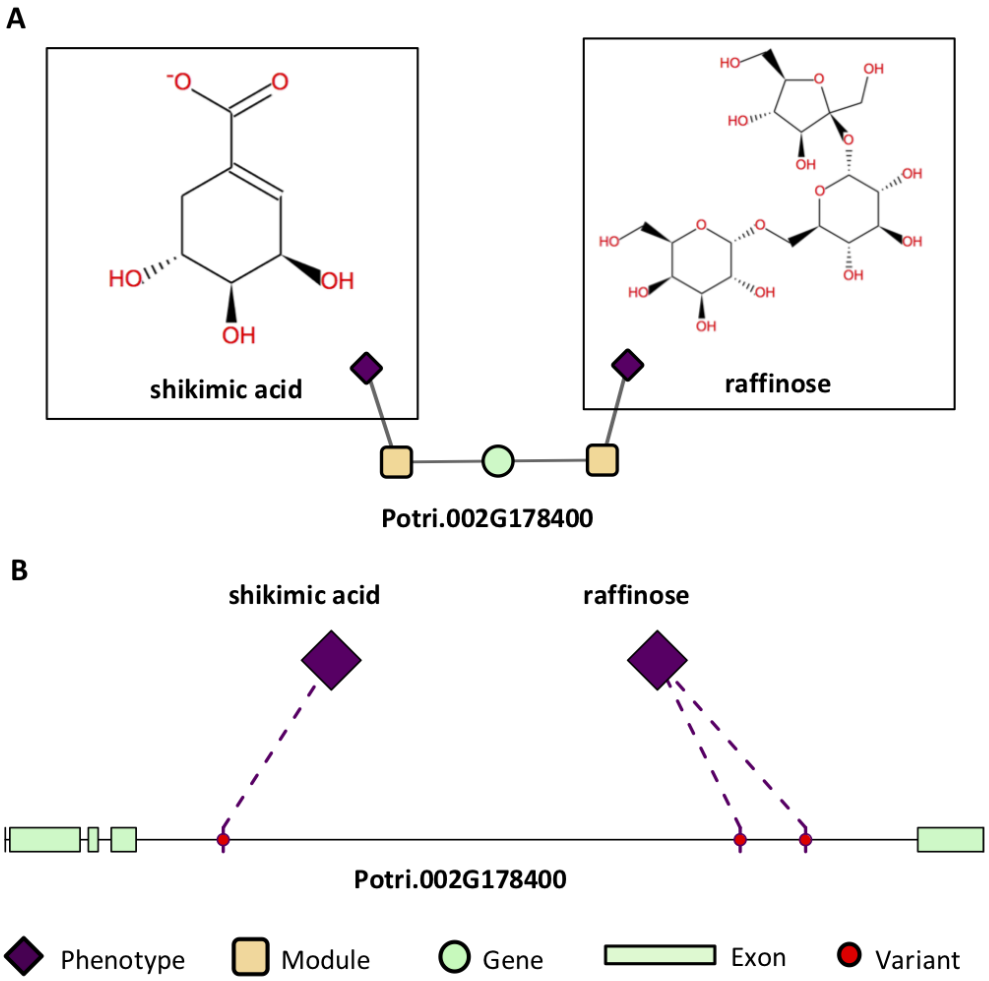


Figure 3.13: **Pleiotropic Signature.** (A) An example of a potentially pleiotropic signature of Potri.002G178400, involving a type 2 MPA with two metabolites in different pathways. (B) Associating SNP positions within Potri.002G178400.

been clonally propagated in common gardens, resequenced and genotyped, [17, 57, 58] providing a dense set of ∼28 million variants which are publicly available (DOI 10.13139/OLCF/1411410). Secondly, linkage disequilibrium (LD) decays very rapidly within this population of *P. trichocarpa* (Figure S3.11). This, in combination with the dense SNP genotyping, allowed for very fine-scale MPA signatures to be resolved. Thirdly, many other different ’omics datasets exist for *P. trichocarpa* including genome scale methylation data across 10 different tissues [59] as well as a gene expression atlas are available on Phytozome [36]. This provides extra data layers which can be integrated with the MPA networks in order to provide further interpretation and context to the GWAS associations seen in the MPA signatures, in a Lines of Evidence approach [19]. Lastly, Poplar is an important bioenergy crop [60] and is the target of extensive research. Thus, this method should be highly valuable to researchers aiming to attempt to genetically modify *P. trichocarpa* in order to impact phenotypes important to bioenergy.

The ease with which these MPA networks can be integrated with other network layers such as co-expression, co-methylation and SNP co-evolution networks provides a powerful strategy for furthering understanding and knowledge about the components of the system, which could aid in the annotation of genes/metabolites of previously unknown function.

MPA decomposition produces signature clusters from GWAS results which can easily be merged with other data types for further interpretation. It is intended that this method will be a valuable tool in the planning of future genetic modification experiments. The resolution of the MPA signatures revealed by this method provides a useful tool to use alongside new CRISPR-based gene editing technologies to achieve high precision genome editing. This method thus provides an informed strategy for increasing the precision of future synthetic biology efforts. Researchers aiming to modify a specific gene in order to impact a particular phenotype can select genes from the signature cluster best suited to the functions they want to modify. The module decomposition also provides information as to which variants/parts of genes are associating with one phenotype or more than one phenotype, and thus can inform the researcher whether the modification of a particular location within a gene will affect more than one phenotype.

MPA decomposition will also be particularly useful in the processing and interpretation of large GWAS datasets such as eQTN studies, involving associations between millions of variants and tens of thousands of phenotypes. Future application of this method to the expanding pool of phenotypic data available will allow for the generation of comprehensive signature clusters representing the global pleiotropic potential of a given organism, and inform the planning and precision of future synthetic biology efforts to impact a wide variety and scale of phenotypes. As such, this approach should have broad impacts by developing high resolution models of MPA/pleiotropy prediction that will form the foundation of future bioengineering design efforts.

3.4 Funding

Funding 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.

This research was also supported by the Plant-Microbe Interfaces Scientific Focus Area (http://pmi.ornl.gov) in the Genomic Science Program, the Office of Biological and Environmental Research (BER) in the U.S. Department of Energy Office of Science, and by the Department of Energy, Laboratory Directed Research and Development funding (7758), at the Oak Ridge National Laboratory. Oak Ridge National Laboratory is managed by UT-Battelle, LLC, for the US DOE under contract DE-AC05-00OR22725.

An award of computer time was provided by the INCITE program. This research used resources of the Oak Ridge Leadership Computing Facility (OLCF) and 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.

Support for DOI 10.13139/OLCF/1411410 dataset is provided by the U.S. Department of Energy, project BIF102 under Contract DE-AC05-00OR22725. Project BIF102 used resources of the Oak Ridge Leadership Computing Facility 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 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.

This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan)

3.5 Supplementary Material

3.5.1 Text S3.1: Proportional Similarity Threshold

A proportional similarity threshold of 1 was chosen when calculating the similarity between the SNP vectors in phenotype space. While this might seem overly stringent, we particularly want to extract groups of SNPs with *identical* phenotype associations in order to form the modules. This is what allows us to easily define modules as equivalence classes. Otherwise, our modules would not represent elements of the powerset of phenotypes observed in SNP-phenotype associations. These modules (elements of the powerset of phenotypes) allow us to rigorously and precisely characterize the MPA signature of genes, which subsequently allows us to cluster genes based on their MPA signatures. For other purposes, for example, if one wants to simply cluster SNPs/genes to obtain groups of genes with similar phenotype associations for functional analysis, one could adjust this threshold. However, for our purposes, in order to characterize *exact* MPA signatures to aid in the planning of genetic modification experiments, we chose a threshold of 1.

3.5.2 Text S3.2: Proportional Similarity Distributions

The Proportional Similarity was calculated between all pairs of SNP GWAS profile vectors (see *Methods and Materials*). The distribution of Proportional Similarity values can be seen in Figure S3.1A. Of the pairs of SNPs which have non-zero Proportional Similarity values (i.e. those pairs of SNPs which shared at least one phenotype association), many had a proportional similarity value of 1. This is explained by the degree distributions of the SNPs in the original SNP-phenotype GWAS network (Figure S3.1B). The degree distribution of a network indicates the probability (or, in this case, frequency) at which a node can be found to have a certain number of edges connected to it [61]. Therefore, the distribution in Figure S3.1B indicates that, of the SNPs which had significant phenotype associations, most of them had precisely one phenotype association. This could skew the Proportional Similarity distribution since any pairs of these “1-phenotype-hit” SNPs which are associated with the same phenotype will have a Proportional Similarity index of 1. However, it is important to keep in mind that these “1-phenotype-hit” SNPs can still contribute to MPA signatures within genes, as two “1-phenotype-hit” SNPs within the same gene that have different associations is precisely what we define as Type 2 MPA signatures.

3.5.3 Text S3.3: Formal Definition of Association Modules

Let the SNP-phenotype GWAS network G be defined as $G = (U, V, E)$, where U is the set of all SNPs within genes with at least two phenotype associations, V is the set of all phenotypes with at least one SNP association and E is the set of edges defined as:

$$E = \{\{u, v\} | u \in U \wedge v \in V \wedge u \text{ is significantly associated with } v\} \tag{3.3}$$

We will define association modules as equivalence classes of U under the relation R. Notation as in [62] will be used.

First we define the binary relation R for any $x, y \in U$ as:

$$xRy \iff PS(x, y) = 1 \tag{3.4}$$

where $PS(x, y)$ is the Proportional Similarity between x and y (see *Methods and Materials*). Since it is true that:

$$PS(x, x) = 1 \tag{3.5}$$

$$PS(x, y) = PS(y, x) \tag{3.6}$$

$$PS(x, y) = 1 \wedge PS(y, z) = 1 \implies PS(x, z) = 1 \tag{3.7}$$

we have that reflexivity, symmetry and transitivity hold:

$$xRx \text{ is true} \tag{3.8}$$

$$xRy \iff yRx \tag{3.9}$$

$$xRy \wedge yRz \implies xRz \tag{3.10}$$

Thus we have that R is an equivalence relation. We define the equivalence class of any element $x \in U$ as:

$$P_R x = \{y | y \in U \wedge yRx\} \tag{3.11}$$

Each association module M_i is defined as an equivalence class of U under the relation R.

3.5.4 Text S3.4: Formal Definitions of MPA Matrices

Below we provide mathematical definitions for the construction of the MPA matrices.

Recall that the SNP-phenotype network G is defined as $G = (U, V, E)$, where U is the set of all SNPs with at least one phenotype hit, V is the set of all phenotypes with at least one SNP hit and E is the set of edges defined as:

$$E = \{\{u, v\} | u \in U \wedge v \in V \wedge u \text{ is significantly associated with } v\} \tag{3.12}$$

We define S_{G_i} to represent the set of SNPs which reside within gene G_i. The gene-phenotype matrix GP is constructed such that each row $G_i \in \{G_1 ... G_m\}$ represents a gene, and each column $P_i \in \{P_1 ... P_l\}$ represents a phenotype. We define each entry GP_{ij} of the gene-phenotype matrix as:

$$GP_{ij} = \begin{cases} 1 & \text{if } \exists s \in U | s \in S_{G_i} \wedge \{s, P_j\} \in E \\ 0 & \text{otherwise} \end{cases} \tag{3.13}$$

Intuitively, this means that entry GP_{ij} will be 1 if there exists a SNP within a MPA gene G_i that is associated with phenotype P_j, and 0 otherwise.

The gene-module matrix GM is constructed such that each row $G_i \in \{G_1 ... G_m\}$

represents a gene and each column $M_i \in \{M_1...M_n\}$ represents a association module. Each entry GM_{ij} is then defined as:

$$GM_{ij} = \begin{cases} 1 & \text{if } \exists s \in U | s \in S_{G_i} \wedge s \in M_j \\ 0 & \text{otherwise} \end{cases} \tag{3.14}$$

Intuitively, this means that entry GM_{ij} will be a 1 if module M_j contains a SNP that resides within gene G_i, and zero otherwise.

The module-phenotype matrix MP was constructed such that each row $M_i \in \{M_1...M_n\}$ represents a association module, and each column $P_i \in \{P_1...P_l\}$ represents a phenotype. We define Q_{M_i} to be the set of phenotypes driving the correlation between the SNPs within module M_i, i.e:

$$Q_{M_i} = \{P_i \in V | \forall s \in M_i, \{s, P_i\} \in E\} \tag{3.15}$$

We then define each entry of the module-phenotype matrix MP_{ij} to be:

$$MP_{ij} = \begin{cases} 1 & \text{if } P_j \in Q_{M_i} \\ 0 & \text{otherwise} \end{cases} \tag{3.16}$$

We refer to the gene-module and module-phenotype matrices as the *decomposition matrices*, and refer collectively to the set of all three matrices (gene-phenotype GP, gene-module GM and module-phenotype MP matrices) as the *MPA matrices*.

3.5.5 Text S3.5: MPA Cube

The three MPA matrices can be seen as different sides of a *MPA cube C* as shown in Figure S3.2A. We define the first dimension of the cube to be genes, the second dimension to be association modules, and the third dimension to be phenotypes. We define each entry C_{ijk} to be:

$$C_{ijk} = \begin{cases} 1 & \text{if } (\exists s \in M_j | s \in S_{G_i}) \wedge (\forall s \in M_j : \{s, P_k\} \in E) \\ 0 & \text{otherwise} \end{cases} \tag{3.17}$$

One can retrieve the individual MPA matrices from the MPA cube simply by "viewing" the cube from different angles, as illustrated in Figure S3.1B. Imagine a transparent box in the dimensions of the MPA cube being filled by $1 \times 1 \times 1$ small cubes. Each small cube is colored black if the corresponding entry in the MPA cube is 1, and transparent if the corresponding entry in the MPA cube is 0. Viewing the transparent box from different sides will reveal a pattern of black and transparent squares, representing the binary values in one of the three MPA matrices, depending on which side you are viewing the cube from. For example, in Figure S3.1B, viewing the cube from the top will reveal the MP matrix, while viewing the cube from the front will reveal the GP matrix and viewing the cube's right side will reveal the GM matrix.

3.5.6 Text S3.6: Composition and Decomposition Relationships

The three MPA matrices satisfy the following equation:

$$GP = \text{bin}(GM \cdot MP) \tag{3.18}$$

where bin() is a binarizing function, setting all entries in a matrix which are greater than one to the value one, and $\cdot$ is normal matrix multiplication. This is a decomposition-like relationship, in that the GP matrix is, with the exception of the binarizing function, decomposed (or factorized) into matrices with an intervening latent variable, namely the association module variable.

The bipartite MPA networks can be seen to have a composition relationship as outlined in Figure 3.5 in the main text. If a gene G_i and a module M_j are connected in the GM network, and that module M_j is connected to phenotype P_k in the MP network, then gene G_i will be connected to phenotype P_k in the GP network.

3.5.7 Supplementary Files

File S3.1. Cytoscape session: Cytoscape session containing interactive networks and p-values for the BINGO [44] results for Type 2 MPA genes.

File S3.2. Metabolic pathway: Positions of raffinose (red) and shikimate (blue) in the PlantCyc metabolic pathway map for *P. trichocarpa* on the Plant Metabolic Network (PMN) online resource [56].

3.5.8 Supplementary Tables

See attached excel files.

Table S3.1: Gene IDs, SNP IDs, beta values and annotation information for MPA genes. Annotation information was derived from the version 3 genome annotation on Phytozome [36].

Table S3.2: GO-terms and their associated adjusted p-values from the GO enrichment analysis, sorted by adjusted p-value. Interactive networks of the GO enrichment analysis per GO hierarchy (Biological Process, Molecular Function and Cellular Component) as well as the associated p-values can be found in the Cytoscape session in Supplementary File S3.1.

Table S3.3: Annotation information for the primary transcripts of the 14 chaperone-related genes identified as MPA genes. Functional information shown was obtained from the version 3.0 gene annotation of *P. trichocarpa* on Phytozome [36] and includes PFAM domains, as well as the ID, name and description of the best *Arabidopsis thaliana* hit. The type of MPA signature exhibited (type 1 or type 2) is also shown.

3.5.9 Supplementary Figures

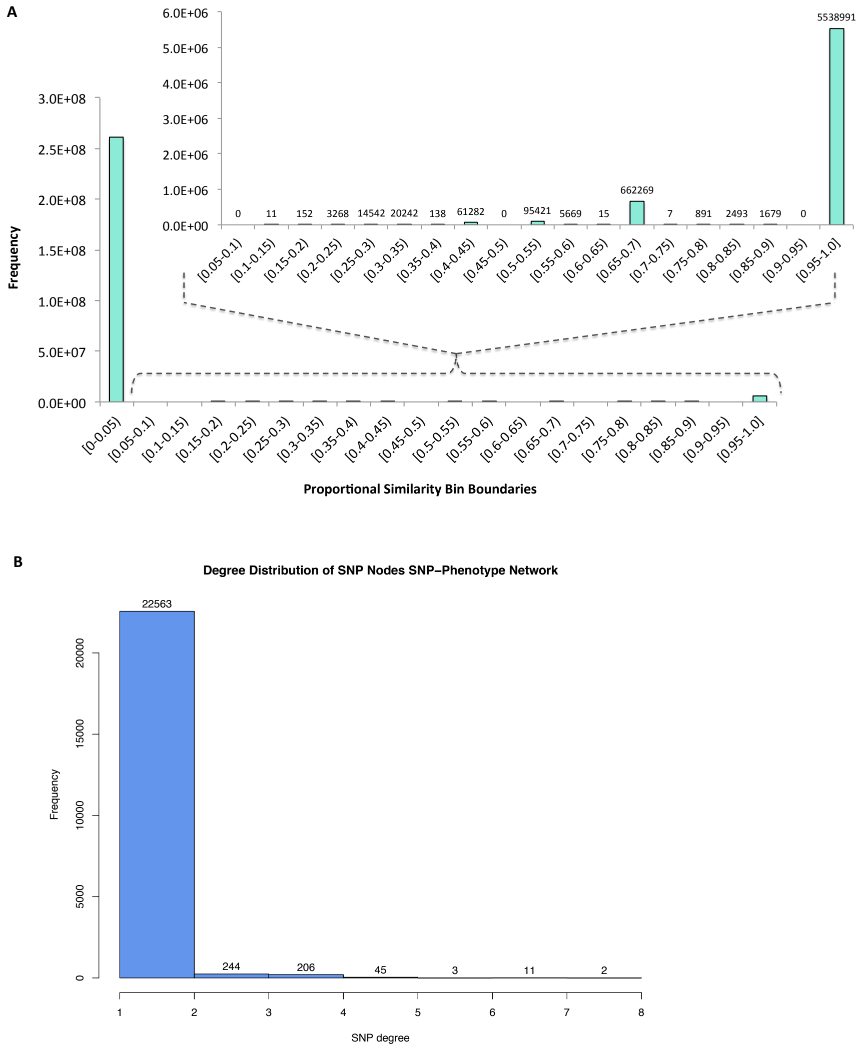
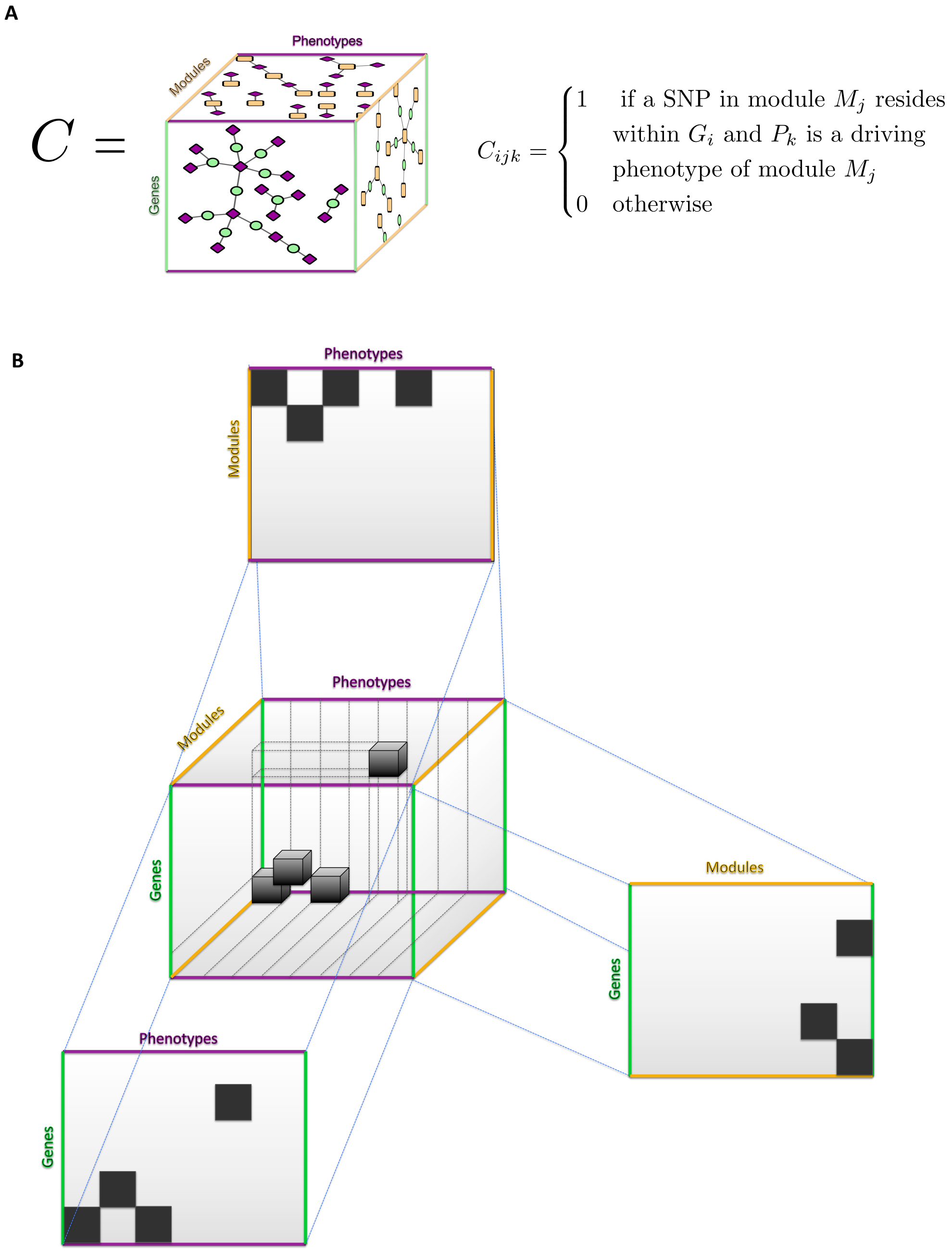


Figure S3.1: **Distributions:** (A) Distribution of the Proportional Similarity edge weights in the SNP association network. (B) Degree distribution of SNP nodes in the SNP-phenotype GWAS bipartite network.



$$C = \quad C_{ijk} = \begin{cases} 1 & \text{if a SNP in module } M_j \text{ resides} \\ & \text{within } G_i \text{ and } P_k \text{ is a driving} \\ & \text{phenotype of module } M_j \\ 0 & \text{otherwise} \end{cases}$$

Figure S3.2: **MPA Cube.** (a) Definition of the MPA cube. (b) Projection onto a particular side of the cube results in one of the MPA matrices.

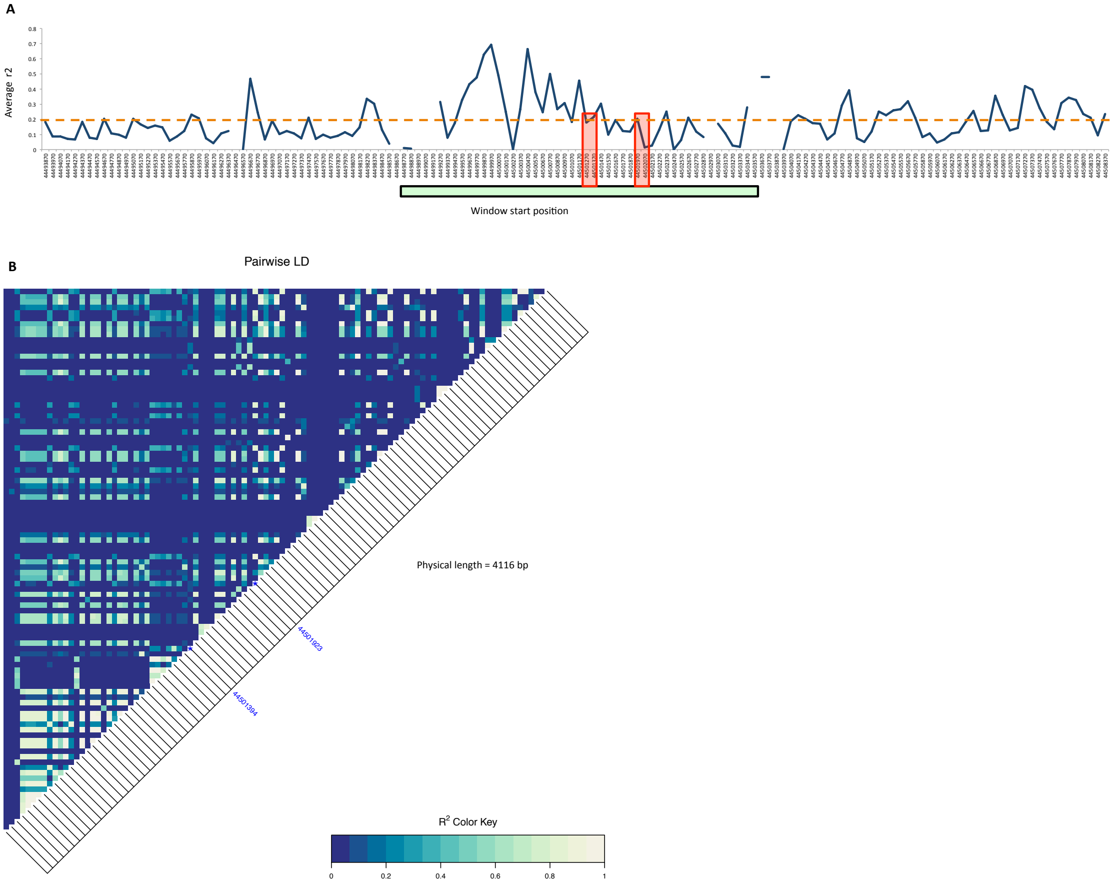


Figure S3.3: **Local LD example.** (A) Variation in LD in the region including 5kb upstream and downstream of Potri.001G419800. The green bar denotes the gene region, and then red bars hhighlight the overlapping bins containing the associating variants within the gene. LD r^2 values were calculated for pairs SNPs within 200bp windows across this region, overlapping by 100bp using PLINK [63]. (B) Pairwise LD heatmap of 100 variants in this region shown in (A) including the two associating variants in Potri.001G419800. LD values were calculated using PLINK [63] and plotted using LDheatmap [64].

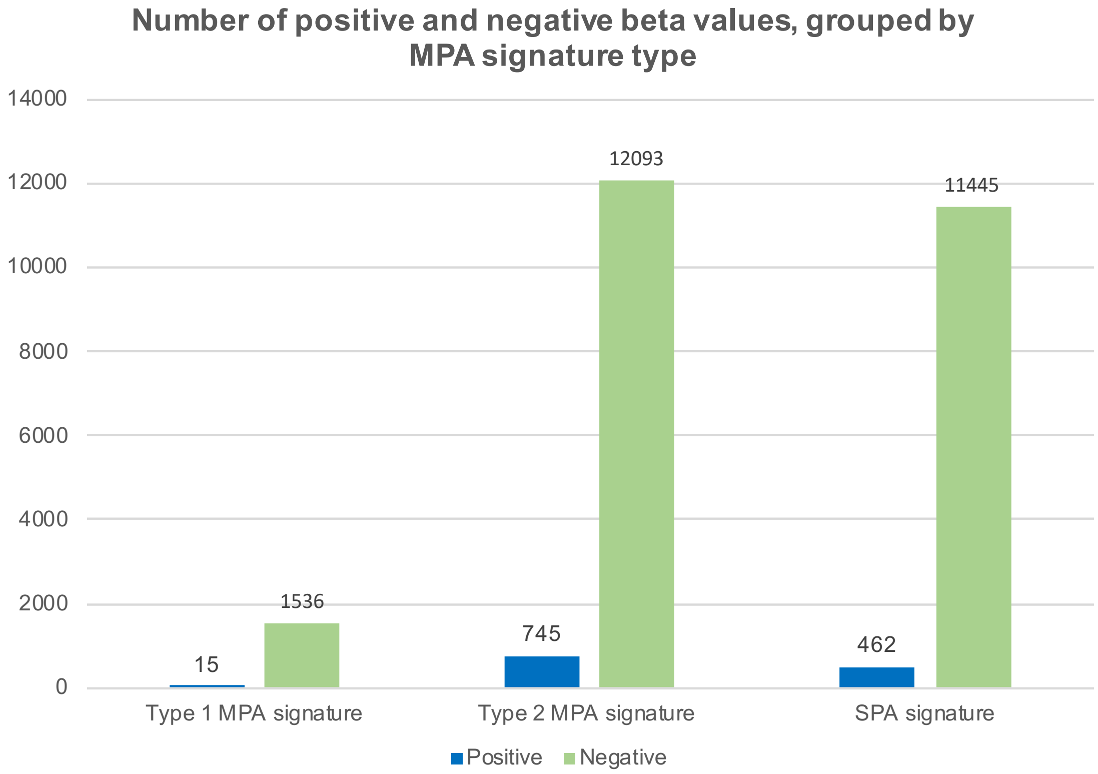


Figure S3.4: **Beta values.** Number of positive and negative significant beta values by MPA signature type.

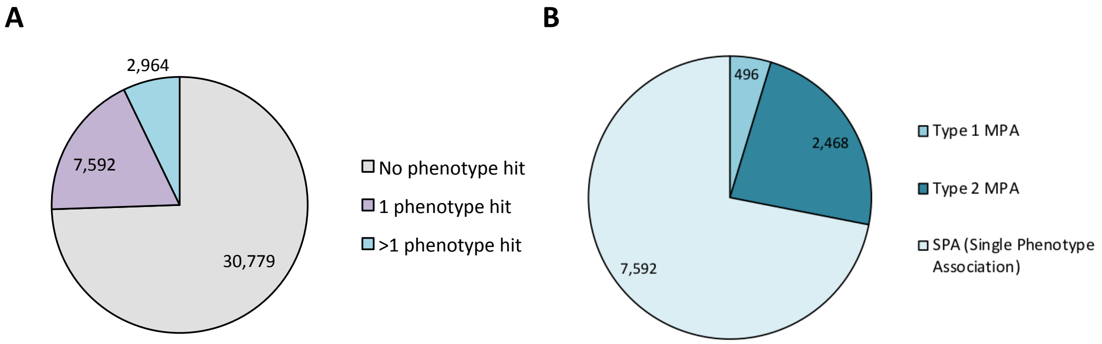


Figure S3.5: **Gene GWAS Association Counts.** Summary of the number of genes (A) with different numbers of GWAS phenotype hits and (B) different MPA signatures.

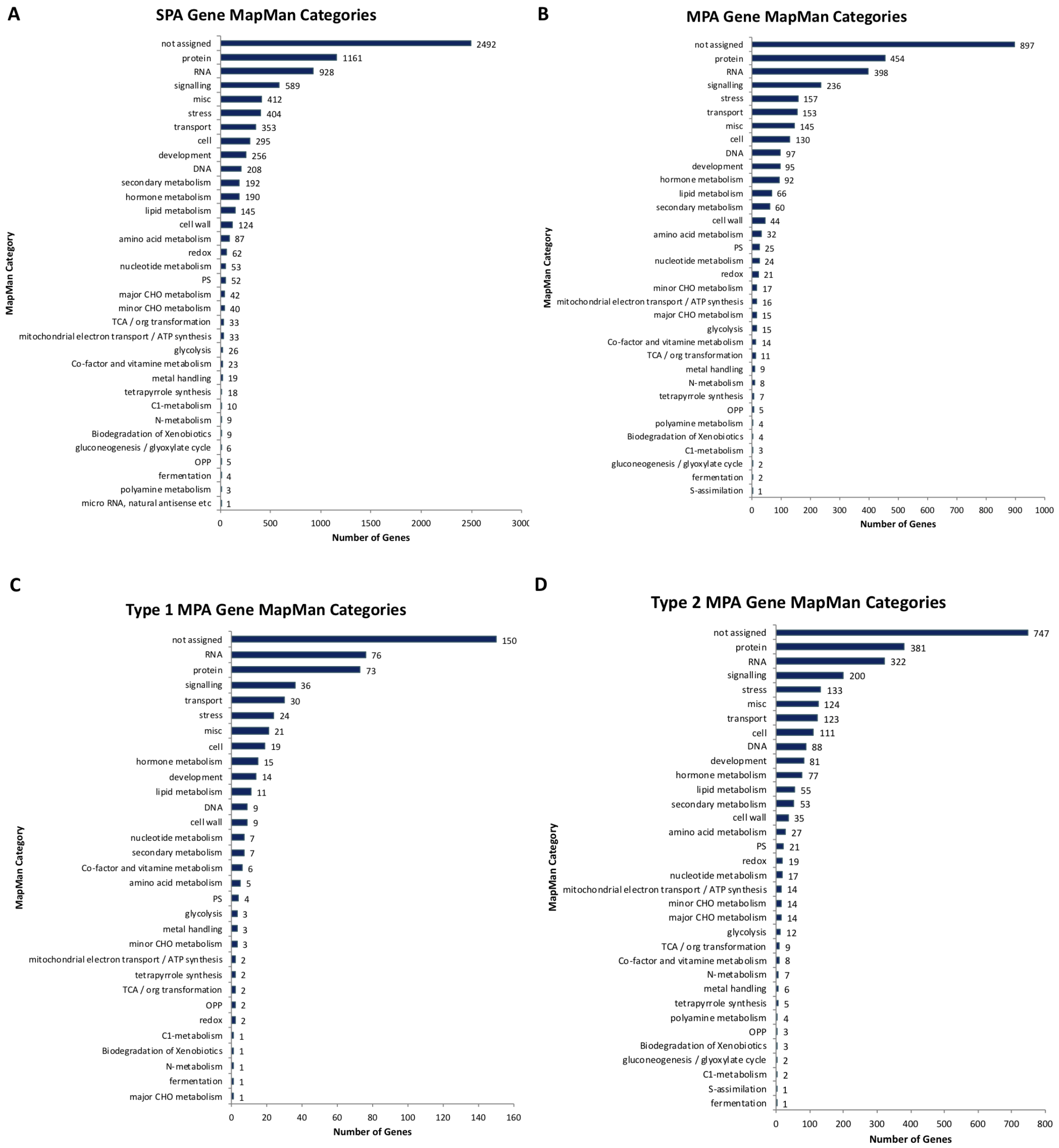


Figure S3.6: **Functional Annotations.** Number of genes annotated with different high-level MapMan categories for (A) non-MPA genes, (B) all MPA genes, (C) type 1 MPA genes and (D) type 2 MPA genes.

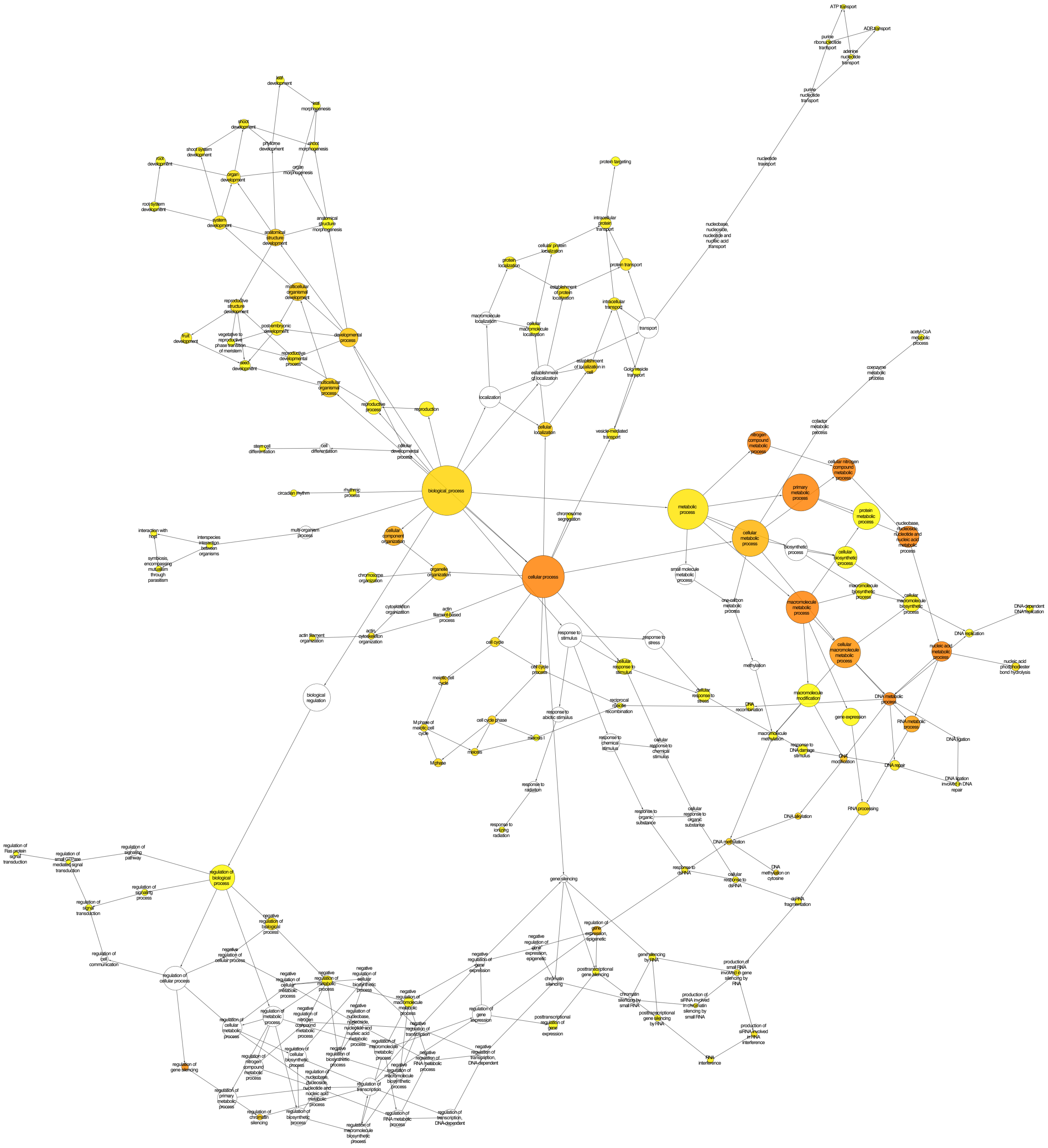

ATP transport
ADP transport
purine ribonucleotide transport
adenine nucleotide transport
purine nucleotide transport
nucleotide transport
nucleobase, nucleoside, nucleotide and nucleic acid transport
leaf development
leaf morphogenesis
shoot development
phyllome development
shoot morphogenesis
shoot system development
root development
organ development
organ morphogenesis
root system development
system development
anatomical structure development
anatomical structure morphogenesis
protein targeting
intracellular protein transport
cellular protein localization
protein localization
protein transport
establishment of protein localization
intracellular transport
macromolecule localization
cellular macromolecule localization
transport
multicellular organismal development
reproductive structure development
post-embryonic development
developmental process
fruit development
vegetative to reproductive phase transition of meristem
reproductive developmental process
seed development
multicellular organismal process
establishment of localization
establishment of localization in cell
Golgi vesicle transport
acetyl-CoA metabolic process
coenzyme metabolic process
localization
reproductive process
reproduction
cellular localization
vesicle-mediated transport
cofactor metabolic process
nitrogen compound metabolic process
cellular nitrogen compound metabolic process
stem cell differentiation
cell differentiation
cellular developmental process
primary metabolic process
biological_process
circadian rhythm
rhythmic process
chromosome segregation
metabolic process
protein metabolic process
nucleobase, nucleoside, nucleotide and nucleic acid metabolic process
interaction with host
interspecies interaction between organisms
multi-organism process
cellular component organization
cellular metabolic process
biosynthetic process
cellular biosynthetic process
symbiosis, encompassing mutualism through parasitism
small molecule metabolic process
chromosome organization
organelle organization
cellular process
macromolecule biosynthetic process
one-carbon metabolic process
cellular macromolecule biosynthetic process
macromolecule metabolic process
DNA-dependent DNA replication
cytoskeleton organization
actin filament-based process
DNA replication
actin filament organization
actin cytoskeleton organization
response to stimulus
response to stress
cell cycle
cellular response to stimulus
cellular macromolecule metabolic process
nucleic acid metabolic process
nucleic acid phosphodiester bond hydrolysis
methylation
meiotic cell cycle
cell cycle process
biological regulation
macromolecule modification
DNA metabolic process
reciprocal meiotic recombination
cellular response to stress
DNA recombination
M phase of meiotic cell cycle
cell cycle phase
response to abiotic stimulus
gene expression
RNA metabolic process
meiosis I
meiosis
macromolecule methylation
DNA ligation
response to chemical stimulus
cellular response to chemical stimulus
response to DNA damage stimulus
DNA modification
DNA repair
M phase
DNA ligation involved in DNA repair
response to radiation
RNA processing
response to organic substance
cellular response to organic substance
DNA alkylation
response to ionizing radiation
DNA methylation
regulation of Ras protein signal transduction
regulation of signaling pathway
regulation of small GTPase mediated signal transduction
regulation of biological process
response to dsRNA
cellular response to dsRNA
DNA methylation on cytosine
gene silencing
regulation of signaling process
regulation of signal transduction
dsRNA fragmentation
negative regulation of biological process
regulation of gene expression, epigenetic
negative regulation of gene expression, epigenetic
regulation of cell communication
negative regulation of cellular process
negative regulation of gene expression
gene silencing by RNA
production of small RNA involved in gene silencing by RNA
regulation of cellular process
negative regulation of cellular metabolic process
negative regulation of metabolic process
negative regulation of cellular biosynthetic process
chromatin silencing
posttranscriptional gene silencing
negative regulation of macromolecule metabolic process
negative regulation of nitrogen compound metabolic process
regulation of metabolic process
negative regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process
chromatin silencing by small RNA
production of siRNA involved in chromatin silencing by small RNA
regulation of gene expression
posttranscriptional gene silencing by RNA
regulation of cellular metabolic process
negative regulation of transcription
posttranscriptional regulation of gene expression
production of siRNA involved in RNA interference
regulation of cellular biosynthetic process
negative regulation of biosynthetic process
regulation of macromolecule metabolic process
negative regulation of transcription, DNA-dependent
regulation of nitrogen compound metabolic process
negative regulation of RNA metabolic process
regulation of gene silencing
RNA interference
negative regulation of macromolecule biosynthetic process
regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process
regulation of transcription
regulation of primary metabolic process
regulation of chromatin silencing
regulation of biosynthetic process
regulation of RNA metabolic process
regulation of transcription, DNA-dependent
regulation of macromolecule biosynthetic process

Figure S3.7: **Biological Process Enrichment.** Biological process GO terms enriched in the set of type 2 MPA genes. Enrichment was calculated using the BINGO Cytoscape plugin [44]. Yellow/orange nodes represent significantly over-represented GO terms. The more intense the orange color, the more significant the p-value. White nodes represent GO terms that are not significantly over-represented, but are parents of over-represented terms in the GO hierarchy. Node size corresponds to the number of genes in that particular category in the set tested for enrichment. Interactive networks can be seen and zoomed in the Cytoscape [42] session in Supplementary File S3.1.

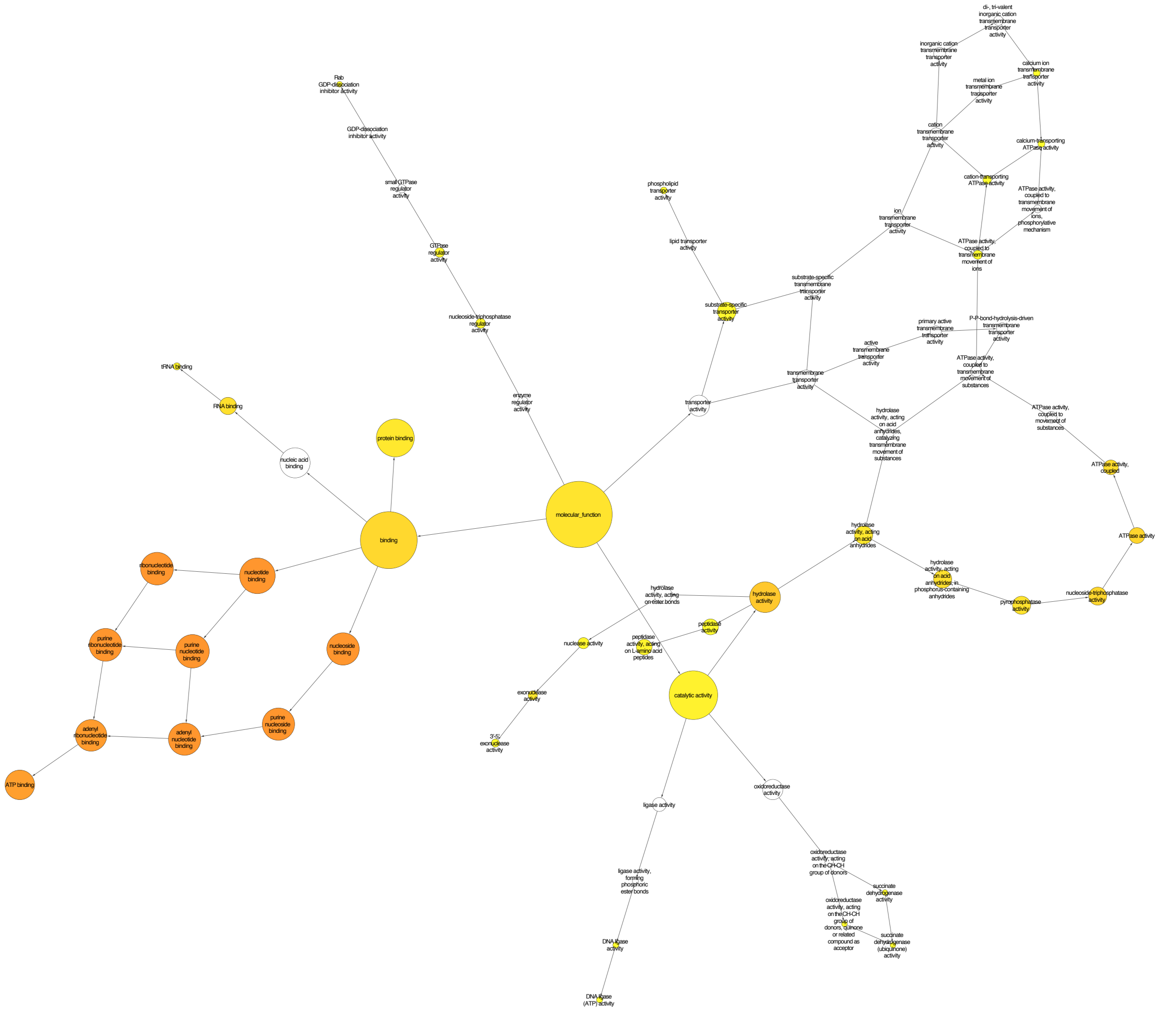

Rab GDP-dissociation inhibitor activity
GDP-dissociation inhibitor activity
small GTPase regulator activity
GTPase regulator activity
nucleoside-triphosphatase regulator activity
enzyme regulator activity
phospholipid transporter activity
lipid transporter activity
substrate-specific transporter activity
transporter activity
tRNA binding
RNA binding
nucleic acid binding
protein binding
binding
molecular_function
ribonucleotide binding
nucleotide binding
purine ribonucleotide binding
purine nucleotide binding
nucleoside binding
adenyl ribonucleotide binding
adenyl nucleotide binding
purine nucleoside binding
ATP binding
di-, tri-valent inorganic cation transmembrane transporter activity
inorganic cation transmembrane transporter activity
calcium ion transmembrane transporter activity
metal ion transmembrane transporter activity
cation transmembrane transporter activity
calcium-transporting ATPase activity
cation-transporting ATPase activity
ATPase activity, coupled to transmembrane movement of ions, phosphorylative mechanism
ion transmembrane transporter activity
ATPase activity, coupled to transmembrane movement of ions
substrate-specific transmembrane transporter activity
primary active transmembrane transporter activity
P-P-bond-hydrolysis-driven transmembrane transporter activity
active transmembrane transporter activity
transmembrane transporter activity
ATPase activity, coupled to transmembrane movement of substances
ATPase activity, coupled to movement of substances
hydrolase activity, acting on acid anhydrides, catalyzing transmembrane movement of substances
ATPase activity, coupled
ATPase activity
hydrolase activity, acting on acid anhydrides
hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides
pyrophosphatase activity
nucleoside-triphosphatase activity
hydrolase activity, acting on ester bonds
hydrolase activity
peptidase activity
nuclease activity
peptidase activity, acting on L-amino acid peptides
exonuclease activity
3'-5' exonuclease activity
catalytic activity
ligase activity
oxidoreductase activity
ligase activity, forming phosphoric ester bonds
oxidoreductase activity, acting on the CH-CH group of donors
succinate dehydrogenase activity
oxidoreductase activity, acting on the CH-CH group of donors, quinone or related compound as acceptor
succinate dehydrogenase (ubiquinone) activity
DNA ligase activity
DNA ligase (ATP) activity

Figure S3.8: **Molecular Function Enrichment.** Molecular function GO terms enriched in the set of type 2 MPA genes. Enrichment was calculated using the BINGO Cytoscape plugin [44]. Yellow/orange nodes represent significantly over-represented GO terms. The more intense the orange color, the more significant the p-value. White nodes represent GO terms that are not significantly over-represented, but are parents of over-represented terms in the GO hierarchy. Node size corresponds to the number of genes in that particular category in the set tested for enrichment. Interactive networks can be seen and zoomed in the Cytoscape [42] session in Supplementary File S3.1.

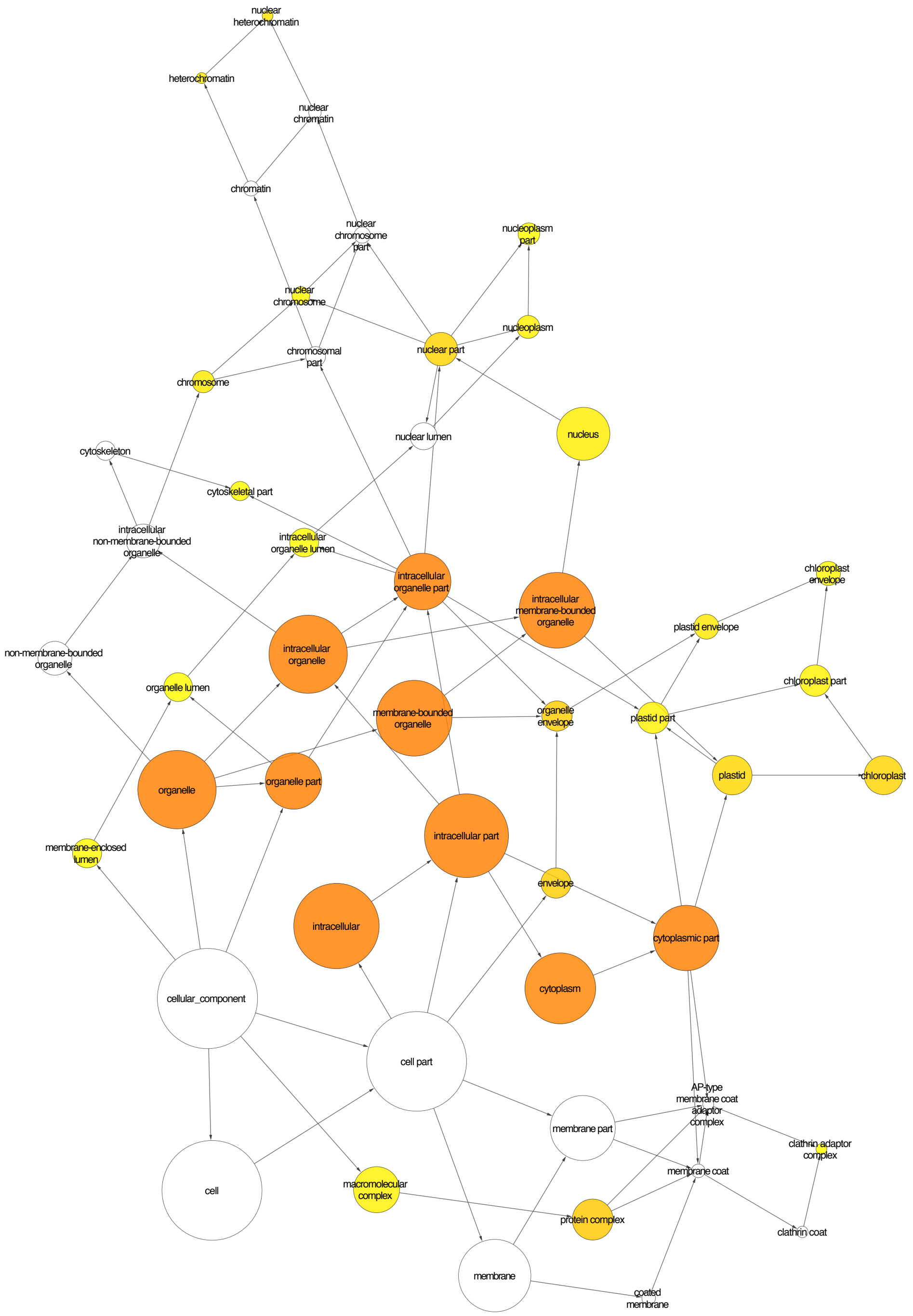


Figure S3.9: **Cellular Component Enrichment.** Cellular component GO terms enriched in the set of type 2 MPA genes. Enrichment was calculated using the BINGO Cytoscape plugin [44]. Yellow/orange nodes represent significantly over-represented GO terms. The more intense the orange color, the more significant the p-value. White nodes represent GO terms that are not significantly over-represented, but are parents of over-represented terms in the GO hierarchy. Node size corresponds to the number of genes in that particular category in the set tested for enrichment. Interactive networks can be seen and zoomed in the Cytoscape [42] session in Supplementary File S3.1.

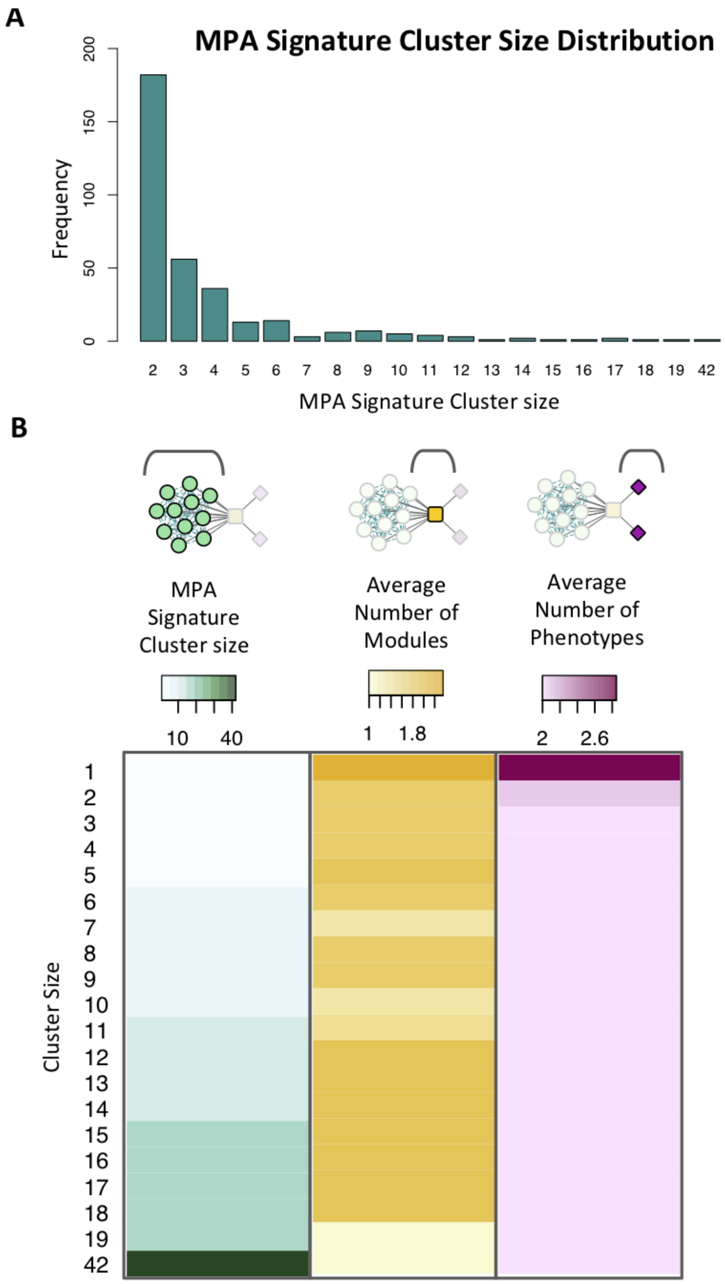


Figure S3.10: **Signature Clusters in Powerset Space.** (A) Cluster size distribution for signature clusters containing ≥ 2 genes. (B) Heatmap showing cluster size (green), average number of modules associated with genes of a given cluster size (yellow) and average number of phenotypes associated with genes in clusters of a given size (pink).

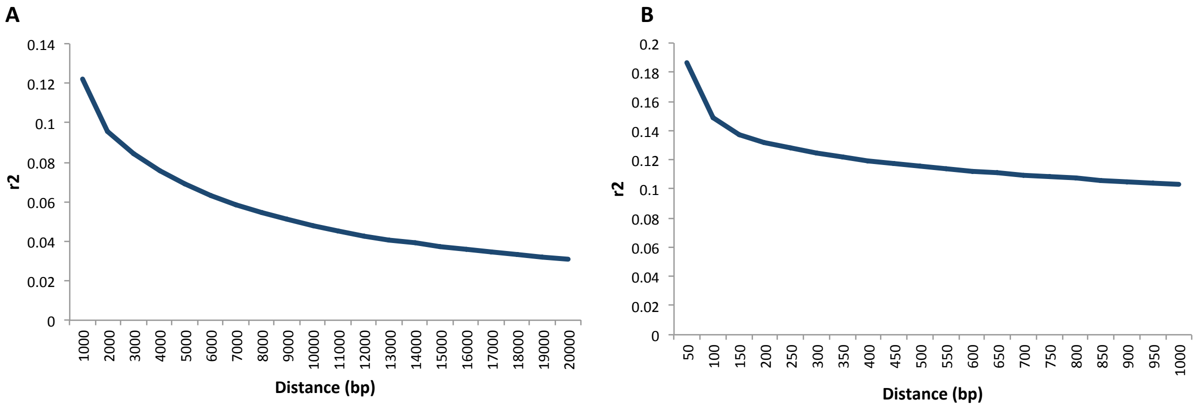


Figure S3.11: **Decay of Linkage Disequilibrium.** The decay of LD r^2 values plotted as the average r^2 value (y-axis) for SNPs within a given distance from each other (x-axis), for a length of (A) 20kb in 1kb windows and (B) 1kb in 50bp windows. LD values were calculated using PLINK [63].

Bibliography

[1] David Houle, Diddahally R Govindaraju, and Stig Omholt. Phenomics: the next challenge. *Nature reviews genetics*, 11(12):855–866, 2010. 145

[2] Sarah A Pendergrass, Anurag Verma, Anna Okula, Molly A Hall, Dana C Crawford, and Marylyn D Ritchie. Phenome-Wide Association Studies: Embracing Complexity for Discovery. *Human heredity*, 79(3-4):111–123, 2015. 145

[3] Molly A Hall, Anurag Verma, Kristin D Brown-Gentry, Robert Goodloe, Jonathan Boston, Sarah Wilson, Bob McClellan, Cara Sutcliffe, Holly H Dilks, Nila B Gillani, et al. Detection of Pleiotropy through a Phenome-Wide Association Study (PheWAS) of Epidemiologic Data as Part of the Environmental Architecture for Genes Linked to Environment (EAGLE) study. *PLoS genetics*, 10(12):e1004678, 2014. 145

[4] SA Pendergrass, K Brown-Gentry, SM Dudek, ES Torstenson, JL Ambite, CL Avery, S Buyske, C Cai, MD Fesinmeyer, C Haiman, et al. The use of phenome-wide association studies (PheWAS) for exploration of novel genotype-phenotype relationships and pleiotropy discovery. *Genetic epidemiology*, 35(5):410–422, 2011. 145

[5] Sarah A Pendergrass, Kristin Brown-Gentry, Scott Dudek, Alex Frase, Eric S Torstenson, Robert Goodloe, Jose Luis Ambite, Christy L Avery, Steve Buyske, Petra Bůžková, et al. Phenome-wide association study (phewas) for detection of pleiotropy within the population architecture using genomics and epidemiology (page) network. *PLoS genetics*, 9(1):e1003087, 2013. 145

[6] Anna L Tyler, Folkert W Asselbergs, Scott M Williams, and Jason H Moore. Shadows of complexity: what biological networks reveal about epistasis and pleiotropy. *BioEssays*, 31(2):220–227, 2009. 145

[7] Nadia Solovieff, Chris Cotsapas, Phil H Lee, Shaun M Purcell, and Jordan W Smoller. Pleiotropy in complex traits: challenges and strategies. *Nature Reviews Genetics*, 14(7):483–495, 2013. xix, 145, 146, 147

[8] Sophie Hackinger and Eleftheria Zeggini. Statistical methods to detect pleiotropy in human complex traits. *Open biology*, 7(11):170125, 2017. 146

[9] Shanya Sivakumaran, Felix Agakov, Evropi Theodoratou, James G Prendergast, Lina Zgaga, Teri Manolio, Igor Rudan, Paul McKeigue, James F Wilson, and Harry Campbell. Abundant Pleiotropy in Human Complex Diseases and Traits. *The American Journal of Human Genetics*, 89(5):607–618, 2011. 146

[10] Morgan E Levine, Peter Langfelder, and Steve Horvath. A Weighted SNP Correlation Network Method for Estimating Polygenic Risk Scores. In *Biological Networks and Pathway Analysis*, pages 277–290. Springer, 2017. 146

[11] Mayetri Gupta, Ching-Lung Cheung, Yi-Hsiang Hsu, Serkalem Demissie, L Adrienne Cupples, Douglas P Kiel, and David Karasik. Identification of Homogeneous Genetic Architecture of Multiple Genetically Correlated Traits by Block Clustering of Genome-Wide Associations. *Journal of Bone and Mineral Research*, 26(6):1261–1271, 2011. 146

[12] Kwang-Il Goh and In-Geol Choi. Exploring the human diseasome: the human disease network. *Briefings in functional genomics*, 11(6):533–542, 2012. 146

[13] Heather F Porter and Paul F O'Reilly. Multivariate simulation framework reveals performance of multi-trait gwas methods. *Scientific reports*, 7:38837, 2017. 148

[14] Manus PM Thoen, Nelson H Davila Olivas, Karen J Kloth, Silvia Coolen, Ping-Ping Huang, Mark GM Aarts, Johanna A Bac-Molenaar, Jaap Bakker, Harro J

Bouwmeester, Colette Broekgaarden, et al. Genetic architecture of plant stress resistance: multi-trait genome-wide association mapping. *New Phytologist*, 213(3):1346–1362, 2017. 148

[15] Nicholas A Furlotte and Eleazar Eskin. Efficient multiple trait association and estimation of genetic correlation using the matrix-variate linear mixed-model. *Genetics*, pages genetics–114, 2015. 148

[16] Matthew Stephens. A unified framework for association analysis with multiple related phenotypes. *PloS one*, 8(7):e65245, 2013. 148

[17] Gerald A Tuskan, S Difazio, Stefan Jansson, J Bohlmann, I Grigoriev, U Hellsten, N Putnam, S Ralph, Stephane Rombauts, A Salamov, et al. The Genome of Black Cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science*, 313(5793):1596–1604, 2006. 149, 152, 155, 175

[18] Gerald Tuskan, Gancho Slavov, Steve DiFazio, Wellington Muchero, Ranjan Pryia, Wendy Schackwitz, Joel Martin, Daniel Rokhsar, Robert Sykes, Mark Davis, et al. *Populus* resequencing: towards genome-wide association studies. In *BMC Proceedings*, volume 5, page I21. BioMed Central Ltd, 2011. 149

[19] Deborah A Weighill, Piet Jones, Manesh Shah, Priya Ranjan, Wellington Muchero, Jeremy Schmutz, Avinash Sreedasyam, David Macaya-Sanz, Robert Sykes, Nan Zhao, et al. Pleiotropic and epistatic network-based discovery: Integrated networks for target gene discovery. *Frontiers in Energy Research*, 2018. 149, 155, 168, 175

[20] Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A Albers, Eric Banks, Mark A DePristo, Robert E Handsaker, Gerton Lunter, Gabor T Marth, Stephen T Sherry, et al. The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158, 2011. 149

[21] Timothy J Tschaplinski, Jonathan M Plett, Nancy L Engle, Aurelie Deveau, Katherine C Cushman, Madhavi Z Martin, Mitchel J Doktycz, Gerald A Tuskan, Annick Brun, Annegret Kohler, et al. *Populus trichocarpa* and *Populus deltoides* Exhibit

Different Metabolomic Responses to Colonization by the Symbiotic Fungus *Laccaria bicolor*. *Molecular Plant-Microbe Interactions*, 27(6):546–556, 2014. 149

[22] Timothy J Tschaplinski, Robert F Standaert, Nancy L Engle, Madhavi Z Martin, Amandeep K Sangha, Jerry M Parks, Jeremy C Smith, Reichel Samuel, Nan Jiang, Yunqiao Pu, Arthur J Ragauskas, Choo Y Hamilton, Chunxiang Fu, Zeng-Yu Wang, Brian H Davidson, Richard A Dixon, and Jonathan R Mielenz. Down-regulation of the caffeic acid *O*-methyltransferase gene in switchgrass reveals a novel monolignol analog. *Biotechnology for Biofuels*, 5(1):1, 2012. 151

[23] Colin A Smith, Elizabeth J Want, Grace O'Maille, Ruben Abagyan, and Gary Siuzdak. XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Monlinear Peak Alignment, Matching, and Identification. *Analytical Chemistry*, 78(3):779–787, 2006. 151

[24] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. 151

[25] Christophe Leys, Christophe Ley, Olivier Klein, Philippe Bernard, and Laurent Licata. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, 49(4):764–766, 2013. 151

[26] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009. 151

[27] Hyun Min Kang, Jae Hoon Sul, Susan K Service, Noah A Zaitlen, Sit-yee Kong, Nelson B Freimer, Chiara Sabatti, Eleazar Eskin, et al. Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, 42(4):348–354, 2010. 152

[28] David J Balding and Richard A Nichols. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica*, 96(1-2):3–12, 1995. 152

[29] Stéfan van der Walt, S. Chris Colbert, and Gaël Varoquaux. The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science & Engineering*, 13(2):22–30, 2011. 152

[30] Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source scientific tools for Python, 2001. [Online; accessed 2016-08-16]. 152

[31] Wes McKinney. Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 51 – 56, 2010. 152

[32] Lisandro D. Dalcin, Rodrigo R. Paz, Pablo A. Kler, and Alejandro Cosimo. Parallel distributed computing using Python. *Advances in Water Resources*, 34(9):1124 – 1139, 2011. New Computational Methods and Software Tools. 152

[33] Lisandro Dalcín, Rodrigo Paz, Mario Storti, and Jorge D'Elía. MPI for Python: Performance improvements and MPI-2 extensions. *Journal of Parallel and Distributed Computing*, 68(5):655 – 662, 2008. 152

[34] Lisandro Dalcín, Rodrigo Paz, and Mario Storti. MPI for Python. *Journal of Parallel and Distributed Computing*, 65(9):1108 – 1115, 2005. 152

[35] Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995. 152

[36] David M Goodstein, Shengqiang Shu, Russell Howson, Rochak Neupane, Richard D Hayes, Joni Fazo, Therese Mitros, William Dirks, Uffe Hellsten, Nicholas Putnam, et al. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Research*, 40(D1):D1178–D1186, 2012. xiv, 152, 155, 175, 183

[37] Igor V Grigoriev, Henrik Nordberg, Igor Shabalov, Andrea Aerts, Mike Cantor, David Goodstein, Alan Kuo, Simon Minovitsky, Roman Nikitin, Robin A Ohm, et al. The Genome Portal of the Department of Energy Joint Genome Institute. *Nucleic Acids Research*, pages 1–7, 2011. 153, 155

[38] Henrik Nordberg, Michael Cantor, Serge Dusheyko, Susan Hua, Alexander Poliakov, Igor Shabalov, Tatyana Smirnova, Igor V Grigoriev, and Inna Dubchak. The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. *Nucleic Acids Research*, 42(D1):D26–D31, 2014. 153, 155

[39] Stephen A Bloom. Similarity Indices in Community Studies: Potential Pitfalls. *Mar. Ecol. Prog. Ser*, 5(2):125–128, 1981. 153

[40] Stijn Van Dongen. Graph clustering via a discrete uncoupling process. *SIAM Journal on Matrix Analysis and Applications*, 30(1):121–141, 2008. 154, 155

[41] Stijn Marinus Van Dongen. Graph clustering by flow simulation. 2001. 154, 155

[42] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, 13(11):2498–2504, 2003. xxiii, xxiv, 155, 189, 190, 191

[43] Marc Lohse, Axel Nagel, Thomas Herter, Patrick May, Michael Schroda, Rita Zrenner, Takayuki Tohge, Alisdair R Fernie, Mark Stitt, and Björn Usadel. Mercator: a fast and simple web server for genome scale functional annotation of plant sequence data. *Plant, Cell & Environment*, 37(5):1250–1258, 2014. 155

[44] Steven Maere, Karel Heymans, and Martin Kuiper. Bingo: a cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, 21(16):3448–3449, 2005. xxiii, xxiv, 155, 183, 189, 190, 191

[45] Todd A Sangster, Susan Lindquist, and Christine Queitsch. Under cover: causes, effects and implications of hsp90-mediated genetic capacitance. *Bioessays*, 26(4):348–362, 2004. 167

[46] Dong Yul Sung and Charles L Guy. Physiological and molecular assessment of altered expression of hsc70-1 in arabidopsis. evidence for pleiotropic consequences. *Plant Physiology*, 132(2):979–987, 2003. 167

[47] Wei J Gong and Kent G Golic. Loss of hsp70 in drosophila is pleiotropic, with effects on thermotolerance, recovery from heat shock and neurodegeneration. *Genetics*, 172(1):275–286, 2006. 167

[48] Sascha Gille, Amancio de Souza, Guangyan Xiong, Monique Benz, Kun Cheng, Alex Schultink, Ida-Barbara Reca, and Markus Pauly. O-acetylation of arabidopsis hemicellulose xyloglucan requires axy4 or axy4l, proteins with a tbl and duf231 domain. *The Plant Cell*, 23(11):4041–4053, 2011. 172

[49] Yaping Gao, Congwu He, Dongmei Zhang, Xiangling Liu, Zuopeng Xu, Yanbao Tian, Xue-Hui Liu, Shanshan Zang, Markus Pauly, Yihua Zhou, et al. Two trichome birefringence-like proteins mediate xylan acetylation, which is essential for leaf blight resistance in rice. *Plant physiology*, 173(1):470–481, 2017. 172

[50] José María Bellés, Rafael Garro, Joaquín Fayos, Pilar Navarro, Jaime Primo, and Vicente Conejero. Gentisic acid as a pathogen-inducible signal, additional to salicylic acid for activation of plant defenses in tomato. *Molecular plant-microbe interactions*, 12(3):227–235, 1999. 172

[51] Laura Campos, Pablo Granell, Susana Tárraga, Pilar López-Gresa, Vicente Conejero, José María Bellés, Ismael Rodrigo, and Purificación Lisón. Salicylic acid and gentisic acid induce rna silencing-related genes and plant resistance to rna pathogens. *Plant physiology and biochemistry*, 77:35–43, 2014. 172

[52] Cristina Ferrándiz, Sarah J Liljegren, and Martin F Yanofsky. Negative regulation of the shatterproof genes by fruitfull during arabidopsis fruit development. *Science*, 289(5478):436–438, 2000. 172

[53] Claudia Cosio, Philippe Ranocha, Edith Francoz, Vincent Burlat, Yumei Zheng, Sharyn E Perry, Juan-Jose Ripoll, Martin Yanofsky, and Christophe Dunand. 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(1):250–263, 2017. 172

[54] Estela Giménez, Benito Pineda, Juan Capel, María Teresa Antón, Alejandro Atarés, Fernando Pérez-Martín, Begoña García-Sogo, Trinidad Angosto, Vicente Moreno, and Rafael Lozano. Functional analysis of the arlequin mutant corroborates the essential role of the arlequin/tagl1 gene during reproductive development of tomato. *PLoS One*, 5(12):e14427, 2010. 172

[55] Frank W Stearns. One hundred years of pleiotropy: a retrospective. *Genetics*, 186(3):767–773, 2010. 172

[56] Pascal Schlapfer, Peifen Zhang, Chuan Wang, Taehyong Kim, Michael Banf, Lee Chae, Kate Dreher, Arvind K Chavali, Ricardo Nilo-Poyanco, Thomas Bernard, et al. Genome-wide prediction of metabolic enzymes, pathways and gene clusters in plants. *Plant physiology*, pages pp–01942, 2017. 173, 183

[57] Gancho T Slavov, Stephen P DiFazio, Joel Martin, Wendy Schackwitz, Wellington Muchero, Eli Rodgers-Melnick, Mindie F Lipphardt, Christa P Pennacchio, Uffe Hellsten, Len A Pennacchio, et al. Genome resequencing reveals multiscale geographic structure and extensive linkage disequilibrium in the forest tree *Populus trichocarpa*. *New Phytologist*, 196(3):713–725, 2012. 175

[58] Luke M Evans, Gancho T Slavov, Eli Rodgers-Melnick, Joel Martin, Priya Ranjan, Wellington Muchero, Amy M Brunner, Wendy Schackwitz, Lee Gunter, Jin-Gui Chen, et al. Population genomics of *Populus trichocarpa* identifies signatures of selection and adaptive trait associations. *Nature Genetics*, 46(10):1089–1096, 2014. 175

[59] Kelly J Vining, Kyle R Pomraning, Larry J Wilhelm, Henry D Priest, Matteo Pellegrini, Todd C Mockler, Michael Freitag, and Steven H Strauss. Dynamic DNA cytosine methylation in the *Populus trichocarpa* genome: tissue-level variation and relationship to gene expression. *BMC Genomics*, 13(1):1, 2012. 175

[60] Poulomi Sannigrahi, Arthur J Ragauskas, and Gerald A Tuskan. Poplar as a feedstock for biofuels: a review of compositional characteristics. *Biofuels, Bioproducts and Biorefining*, 4(2):209–226, 2010. 175

[61] Albert-László Barabási and Zoltán N Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113, 2004. 178

[62] Saunders MacLane and Garrett Birkhoff. *Algebra*. AMS Chelsea Publishing, third edition, 1988. 179

[63] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575, 2007. xxii, xxiv, 186, 193

[64] J.-H. Shin, S. Blay, B. McNeney, and J. Graham. Ldheatmap: An r function for graphical display of pairwise linkage disequilibria between single nucleotide polymorphisms. *J Stat Soft*, 16:Code Snippet 3, 2006. xxii, 186

Chapter 4

Centromere Wavelet Signatures and Co-evolution with CENH3 in *Populus trichocarpa*

This chapter has been submitted for publication in *Frontiers in Genetics* and is currently under review. This chapter contains contributions from other authors:

Deborah Weighill, David Macaya-Sanz, Stephen DiFazio, Wayne Joubert, Manesh Shah, Jeremy Schmutz, Avinash Sreedasyam, Gerald Tuskan and Daniel Jacobson

Author Contributions: DW and DJ conceived of and designed the study. DW performed the analysis, developed the wavelet centromere identification strategy, interpreted the results and wrote the manuscript, GT lead the sequencing of *Populus* genotypes, SD and DMS performed the SNP calling, JS and AS contributed the genome sequence and transcriptome expression analysis, MS mapped gene expression atlas reads and calculated gene expression TPM values, WJ developed the parallel GPU CCC metric code. DW, DJ, GT, DMS and AS edited the manuscript. The Department of Energy Joint Genome Institute (JGI) performed the sequencing of *P. trichocarpa* genomes, and Ryan McCormick, Sandra Truong and Anna Furches provided useful discussions about the manuscript.

Abstract

Various ‘omics data types have been generated for *Populus trichocarpa*, each providing a layer of information which can be represented as a density signal across a chromosome. We make use of genome sequence data, variants data across a population as well as methylation data across 10 different tissues, combined with wavelet-based signal processing to perform a comprehensive analysis of the signature of the centromere in these different data signals, and successfully identify putative centromeric regions in *P. trichocarpa* from these signals. Furthermore, using SNP (single nucleotide polymorphism) correlations across a natural population of *P. trichocarpa*, we find evidence for the co-evolution of the centromeric histone CENH3 with the sequence of the newly identified centromeric regions, and identify a new CENH3 candidate in *P. trichocarpa*.

4.1 Introduction

Integrating data from multiple different sources is a task which is becoming more prevalent with the increased availability of systems biology data from high-throughput 'omics technologies and phenotyping strategies [1]. Developing statistical and mathematical approaches to integrate this data in order to provide an increased understanding of the biological system is thus an important endeavor. For the bioenergy feedstock crop *Populus trichocarpa*, several heterogenous datasets have been generated. The full genome sequence is available and is currently in its third version [2]. A large collection of ∼ 28,000,000 Single Nucleotide Polymorphisms (SNPs) called across 882 genotypes are publicly available [https://doi.ccs.ornl.434gov/ui/doi/55], which were derived from the resequenced genomes of ∼1,000 *P. trichocarpa* genotypes propagated in common gardens [3, 4, 5]. Methyl-DNA immunoprecipitation (MEDIP)-seq DNA methylation data is also available for 10 different *P. trichocarpa* tissues [6]. A gene expression atlas for *P. trichocarpa* is also available on Phytozome [7].

Integration of multiple heterogeneous data types requires coercing them into mathematical structures that allow them to be compared/merged/layered. For example, each of the data types mentioned above provides feature(s) which can be represented as vectors of numbers, with each vector representing a signal which varies across a chromosome, for example, the gene density across a chromosome, or the methylation profile of a chromosome. Once represented as a signal, these data types are amenable to signal processing techniques. This study aims to make use of signal processing techniques of these multiple data types in order to attempt to identify chromosome structural features in *P. trichocarpa*.

The centromere is an important chromosomal structure which controls the segregation of chromosomes during cell division, and is the location for the assembly for the kinetochore protein complex [8, 9]. Centromeric chromatin contains a histone H3 variant specific to the centromere (CENH3), which has been found in many organsisms, including plants

[10]. Studies by [11, 12] have suggested that CENH3 is co-evolving with the sequence of the centromere.

Centromeric regions can vary in size, and can be small regions consisting of only one nucleosome such as in *Saccharomyces cerevisiae* [13, 9], while plant centromeric regions are large (Mb scale), and consist of repetitive sequences [14, 9]. Centromeres also have epigenetic characteristics in that plant centromeric regions have been found to be relatively highly methylated [15, 6].

Previously, putative centromere positions were identified in *P. trichocarpa* as chromosomal regions of low gene density and high methylation, presented visually, but coordinates were not reported [6]. Putative centromere positions have also been identified based on recombination rates along chromosomes through visual inspection of profiles of $4N_ec$ [4]. [16] identified putative centromeric repeats of *P. trichocarpa* which identified putative centromere positions on some of the *P. trichocarpa* chromosomes in a previous assembly of the genome. In [17], putative centromeres were identified as regions as the 250kb window on each chromosome with the lowest gene density. However, to our knowledge, there has not been a comprehensive study of *P. trichocarpa* centromeres integrating various available data types and multiple lines of evidence.

The large collection of data available for *P. trichocarpa* provides a source of multiple features which can be represented as density signals across each chromosome. Certain features, such as gene density and SNP density, can be readily constructed from the data available. Other lines of evidence, such as SNP correlation/co-segregation need to be calculated from the data before the chromosome signals can be constructed.

Such chromosome signals contain variation on multiple scales, including high frequency (narrow) peaks and low-frequency (broad) peaks. These different scales of peaks contain different information. Thus, techniques to analyse these signals at different scales are valuable (see [18, 19]). The Wavelet Transform, a signal processing technique, can be used to unpack the information in different scales of a signal, such as a density profile across a chromosome [18]. In general, the wavelet transform involves expressing a function (signal) as a linear combination of functions called wavelets. These functions are scaled

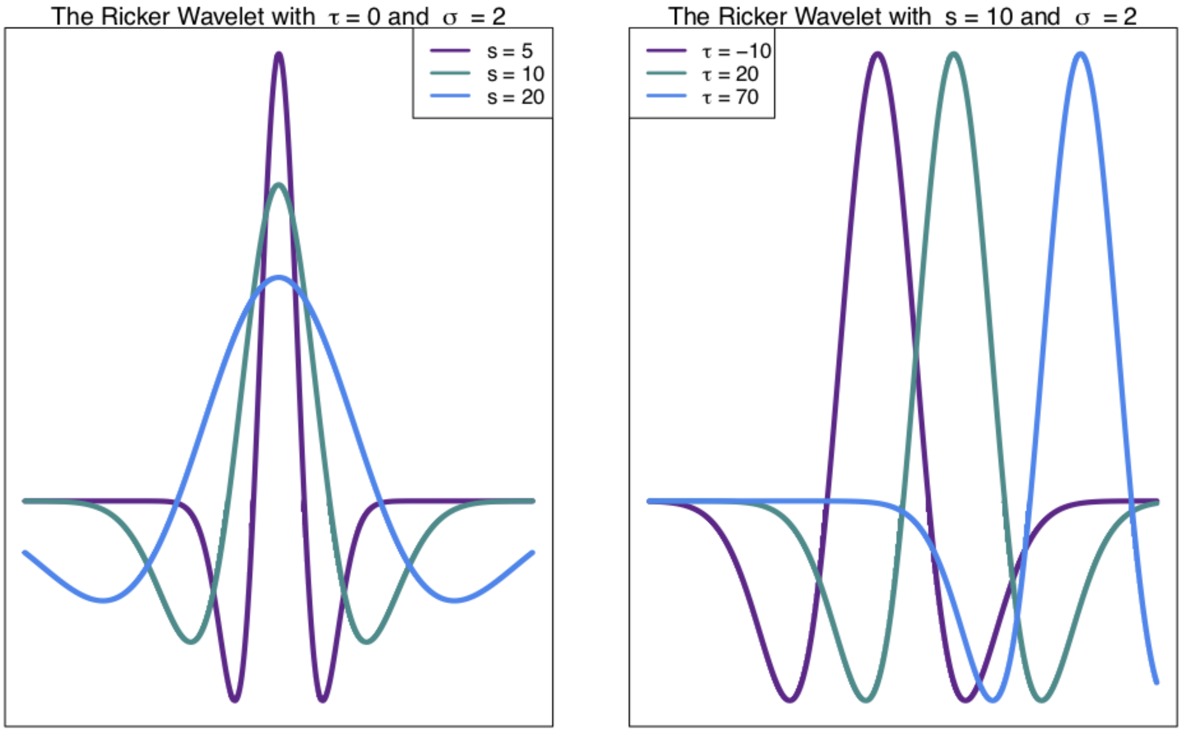


Figure 4.1: **Ricker Wavelet.** The Ricker wavelet shown for different values of scale s and translation τ in Equation 5.1 [21, 20].

translations of a mother wavelet, such as the Ricker Wavelet (Figure 4.1). What results from a wavelet transform is a wavelet coefficient $W(s, \tau)$ (Equation 5.1), for every scale s and translation (shift along the x-axis) τ [20].

$$W(s,\tau) = \frac{1}{\sqrt{s}} \int f(t)\psi^* \left(\frac{t-\tau}{s} \right) dt \tag{4.1}$$

Given the peak-like shape of the wavelet, a wavelet coefficient will indicate "how much of a peak" is present at a particular scale and at a particular position of the signal. Thus, the wavelet transform allows us to investigate the peaks of a signal at different scales and locations.

This study makes use of the Continuous Wavelet Transform (CWT) in characterizing chromosomal gene density, SNP density and methylation density signals in *P. trichocarpa*. We use the resulting CWT coefficient landscapes to identify the putative centromere locations and illustrate the wavelet signature of a centromere. We also investigate potential co-evolution signatures between the centromeric histone CENH3 and the newly identified centromeric regions through the calculation of SNP correlations across the population, and find evidence supporting the hypothesis of the co-evolution of putative *P. trichocarpa* CENH3 genes with the centromere sequences in *P. trichocarpa*. While wavelets have previously been used in chromosome classification [22], and the the discrete wavelet transform has been used in the analysis of feature profiles across a chromosome in human [18], to our knowledge this work presents the first use of the continuous wavelet transform in the identification of centromere positions from SNP and methylation density profiles. This study provides an example of how signal processing of multiple data types can be used to generate hypotheses surrounding the structure of chromosomes.

4.2 Methods and Materials

4.2.1 Variant Data and SNP Correlations

P. trichocarpa [2] variant data (DOI 10.13139/OLCF/1411410) was obtained from https://doi.ccs.ornl.434gov/ui/doi/55. This dataset consists of SNP 28,342,758 SNPs called across 882 *P. trichocarpa* genotypes and is derived the whole genome resequencing of a Genome Wide Association Study (GWAS) population clonally replicated in common gardens [3].

The most reliable SNPs within the dataset were selected, consisting of the 90% tranche (the tranche recovering 90% of the “true” SNPs). VCFtools [23] was used to extract the desired Tranche of SNPs from the VCF file and reformat it into .tfam and .tped files. Plink ([24], http://pngu.mgh.harvard.edu/purcell/plink/) was used to determine the minor allele frequency (MAF) and the call rate (fraction alleles observed) for each SNP, and removed all SNPs with MAF ≤ 0.01 and call rate ≤ 0.5.

Correlations between all pairs of SNPs were calculated using the Custom Correlation Coefficient (CCC) [25, 26]. This was performed on both the filtered set of SNPs as well as the entire 90% tranche, using a new, GPU implementation of the CCC metric for the calculation of SNP correlations [27] as well as the original software[25, 26], respectively. Calculation of the CCC between all pairs of SNPs using the original software was performed in parallel, as described in [28]. Briefly, the CCC between allele x at location i and allele y and location j is defined as:

$$CCC_{i_x j_y} = \frac{9}{2} R_{i_x j_y} \left(1 - \frac{1}{f_{i_x}}\right)\left(1 - \frac{1}{f_{j_y}}\right) \tag{4.2}$$

where $R_{i_x j_y}$ is the relative co-occurrence of allele x at location i and allele y at location j, f_{i_x} is the frequency of allele x at location i and f_{j_y} is the frequency of allele y at location j.

This was performed in a parallel fashion by constructing a Perl wrapper around the ccc binary, making use of the Parallel::MPI::Simple Perl module, developed by Alex Gough and available on The Comprehensive Perl Archive Network (CPAN) at www.cpan.org. “The set of ∼10 million SNPs was divided into 20 different blocks, and the CCC was calculated for each within-block and cross-block comparison in separate jobs, to a total of 210 MPI jobs ... A threshold of 0.7 was then applied.” (Quotation from [28].)

4.2.2 Chromosome Feature Profile Construction

SNP Density Profiles

A SNP density profile was created for each chromosome using the filtered set of SNPs by counting the number of these SNPs in non-overlapping 10kb windows across the chromosome.

Methylation Profiles

Methylation (MeDIP-seq) data from 10 *P. trichocarpa* tissues generated from the study by [6] re-aligned to the version 3 assembly of *P. trichocarpa* was downloaded from Phytozome [7]. This data consists of MeDIP-seq reads from tissues including bud, callus, female catkin, internode explant, leaf, make catkin, phloem, regenerated internode, root and xylem tissue.

Samtools [29] was used to view the data and BamTools stats [30] was used to investigate statistics of the reads in the bam files. BEDTools [31] was used to count the number of reads mapped to 10kb windows across the genome. This will allow us to construct a "mappped read density" distribution for each tissue and each chromosome, showing the number of reads which mapped to different regions of the genome, and thus indicating methylation hotsplots. The BEDOPs [32] software was used to convert .gtf files of the 10kb windows per chromosome into .bed files. GNU-Parallel [33] was used to run the BEDTools jobs in parallel.

Gene Density Profiles

Gene density profiles were constructed for each chromosome. Gene density for a given window was defined as the number of nucleotide positions within that window that reside within genes. Gene boundaries were determined from the Ptrichocarpa_210_v3.0.gene.gff3 annotation file obtained from the *P. trichocarpa* version 3 genome annotation [2] available on Phytozome [7] through the genome portal of the Department of Energy Joint Genome Institute [34, 35].

Genome Gap Density Profiles

Genome gap density profiles were constructed for each chromosome, similar to the approach for constructing SNP density profiles. For each non-overlapping 10kb window on a chromosome, the number of "N" positions were counted in the genome assembly file ptrichocarpa_210_v3.0.fa obtained from the version 3 genome assembly [2] available on Phytozome [7].

4.2.3 Continuous Wavelet Transform of Chromosome Feature Profiles

The CWT was performed on chromosome feature density profiles using the wmtsa R wavelet package [36, 37], the R programming language [38], RStudio [39] and various R packages and resources [40]. The CWT results in sets of wavelet coefficients at different scales. These were plotted as a heatmap/coefficient landscape, showing the numerical values of the different wavelet coefficients across the signal, at different scales. Plots were generated using custom R scripts and R packages [38, 41].

4.2.4 Centromere Position Identification

Putative centromeres were located for each chromosome by computationally identifying the "tooth-X-ray" signature in the wavelet landscapes. Let the matrix M represent the methylation wavelet landscape and let S represent the SNP wavelet landscape for a given chromosome. We identified the maximum wavelet coefficient in the upper third of the methylation wavelet landscape (internode explant tissue), and identified the scale p (row of M) at which this maximum coefficient was found. This identified the general pericentromeric scale. The borders of the approximate pericentromeric regions b_1 and b_2 were identified as the zeroes of the methylation wavelet coefficient vector at scale p (Supplementary Text S4.1, Figure S4.1). The minimum wavelet coefficient in the lower two thirds of S between the borders b_1 and b_2 was then identified, and the scale c (row of S) at which this minimum occurs was considered the centromeric scale. The methylation pericentromeric scale vector M_p (row p in matrix M) and the SNP centromeric scale vector S_c (row c of matrix S) were extracted, and scaled to have mean 0 and standard deviation 1. The approximate centromere locations were then identified as the position x at which the maximum

$$\max(M^*_{p,x} - S^*_{c,x}) \tag{4.3}$$

is obtained, where $M^*_{p,x}$ and $S^*_{c,x}$ represent the xth entry in the scaled vectors of M_p and S_c, respectively.

See Supplementary Text S4.1 and Figure S4.1 for further details.

4.2.5 Centromere Repeat Sequence Profiles

Plant centromere repeat sequences were downloaded from the PGSB Repeat Database [42] at http://pgsb.helmholtz-muenchen.de/plant/recat/index.jsp. The repeat sequences were then BLASTed [43] against the *P. trichocarpa* version 3 genome on Phytozome [7], using an E-value threshold of 10^{-5} and other default parameters. A density profile of BLAST hits was then constructed for each chromosome. The BLAST hit density for a given 10kb window was defined as the number of positions within the window that lay within a BLAST hit (E-value $\leq 10^{-5}$) with a plant centromeric repeat sequence. We obtained putative *P. trichocarpa* centromeric repeat sequences from [16], and constructed a BLAST hit density profile for these repeat sequences in a similar manner. These centromere repeat density profiles were visualized alongside of the predicted putative centromere positions.

4.2.6 Co-expression Network

Gene co-expression relationships were queried on PhytoMine though Phytozome [7, 44]. A custom co-expression network was also created as described in [28] using the *P. trichocarpa* (Nisqually-1) RNA-seq dataset from JGI Plant Gene Atlas project (Sreedasyam et al., unpublished). This dataset consists of samples for standard tissues (leaf, stem, root and bud tissue) and libraries generated from nitrogen source study. A list of sample descriptions was accessed from Phytozome at https://phytozome.jgi.doe.gov/phytomine/aspect.do?name=Expression. Networks were visualized in Cytoscape [45].

4.2.7 Co-Evolution of putative CENH3 genes

The genomic sequence of the *Arabidopsis thaliana* CENH3 gene (AT1G01370) was obtained from Phytozome [7] and BLASTed against the *P. trichocarpa* version 3 genome [2] on Phytozome using default parameters. Two BLAST hits were obtained, one gene on chromosome 14 (Potri.014G096400) and one on chromosome 2 (Potri.002G169000). While Potri.014G096400 contains functional annotations on Phytozome, including Panther PTHR11426:SF46 ("Histone H3-like centromeric protein A) and Pfam PF00125 ("Core histone H2A/H2B/H3/H4SNPs"), Potri.002G169000 contains no functional annotations, likely because of sequencing/assembly issues. There are various exons predicted in the gene which are not considered to be translated. However, when searching for domains in the genome sequence of Potri.002G169000 using CD-search at NCBI [46, 47, 48], Pfam PF00125 ("Core histone H2A/H2B/H3/H4SNPs") is identified in the sequence. Thus, we have two valid CENH3 candidates. SNPs which correlated with SNPs within these genes (CCC ≥ 0.7) were extracted from the SNP correlations. Density profiles of these SNPs were then constructed for all chromosomes in non-overlapping 10kb bins, similar to the profile construction described above.

4.3 Results and Discussion

4.3.1 Chromosome Feature Profiles and CWT Coefficient Landscapes

Chromosomal features including SNPs, genes, genome gaps and DNA methylation plotted as density signals across a chromosome result in signals that vary along the length of the chromosome (Figures 4.2, S4.2-S4.20). These profiles show the frequency of a particular feature in 10kb bins across each chromosome. These profiles vary on different scales, in that they contain peaks and valleys of different frequencies/broadness. Each of these signals has fine variation in the form of narrow, high frequency peaks, as well as broad, low-frequency peaks, as illustrated in the feature density profiles of chromosome 2 (Figure

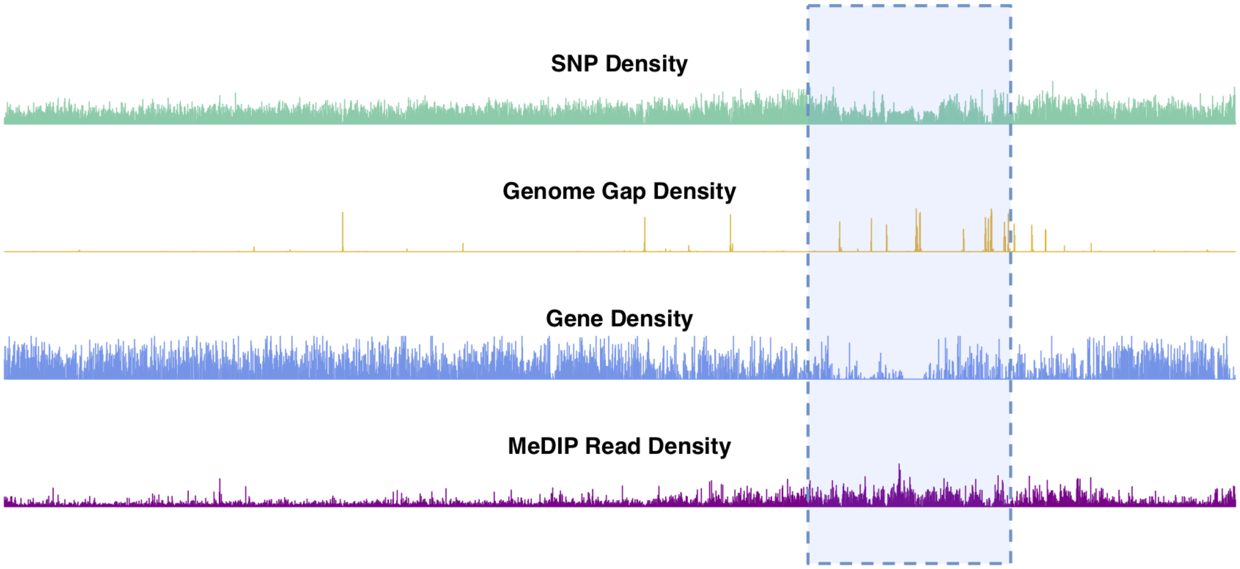


Figure 4.2: **Chromosome 2 feature density signals.** Feature density signals for SNP, gene, MeDIP read (internode explant tissue) and genome gap density in 10kb windows across *P. trichocarpa* chromosome 2.

4.2). The highlighted region in Figure 4.2 indicates the most prominent broad-scale feature, consisting of a large-scale valley in the SNP and gene density profiles, and a large-scale peak in the methylation (MeDIP-Seq read density) profile.

These large-scale peak-valley combinations of SNP, gene and methylation density profiles are observed easily on all chromosomes (Figure 4.3). One can see a large-scale peak in the methylation profile coinciding with valleys in the gene density and SNP density signals on each chromosome. The locations of these large-scale peak-valley combinations seem to agree with the putative *P. trichocarpa* centromere positions proposed by [6] on the basis of high methylation read coverage, high repeat-to-gene ratios and recombination valleys, and also agrees with some of the putative centromere positions identified through repeat elements [16].

The wavelet transform was used to characterize these signals at different scales, identifying peaks of different sizes. Applying the continuous wavelet transform (CWT) to such density signals results in a coefficient landscape for each signal, represented as a heatmap [18] (Figure 4.4). The x-axis of a coefficient landscape represents the position along the chromosome signal and the y-axis represents the scale, with small scales (high frequency

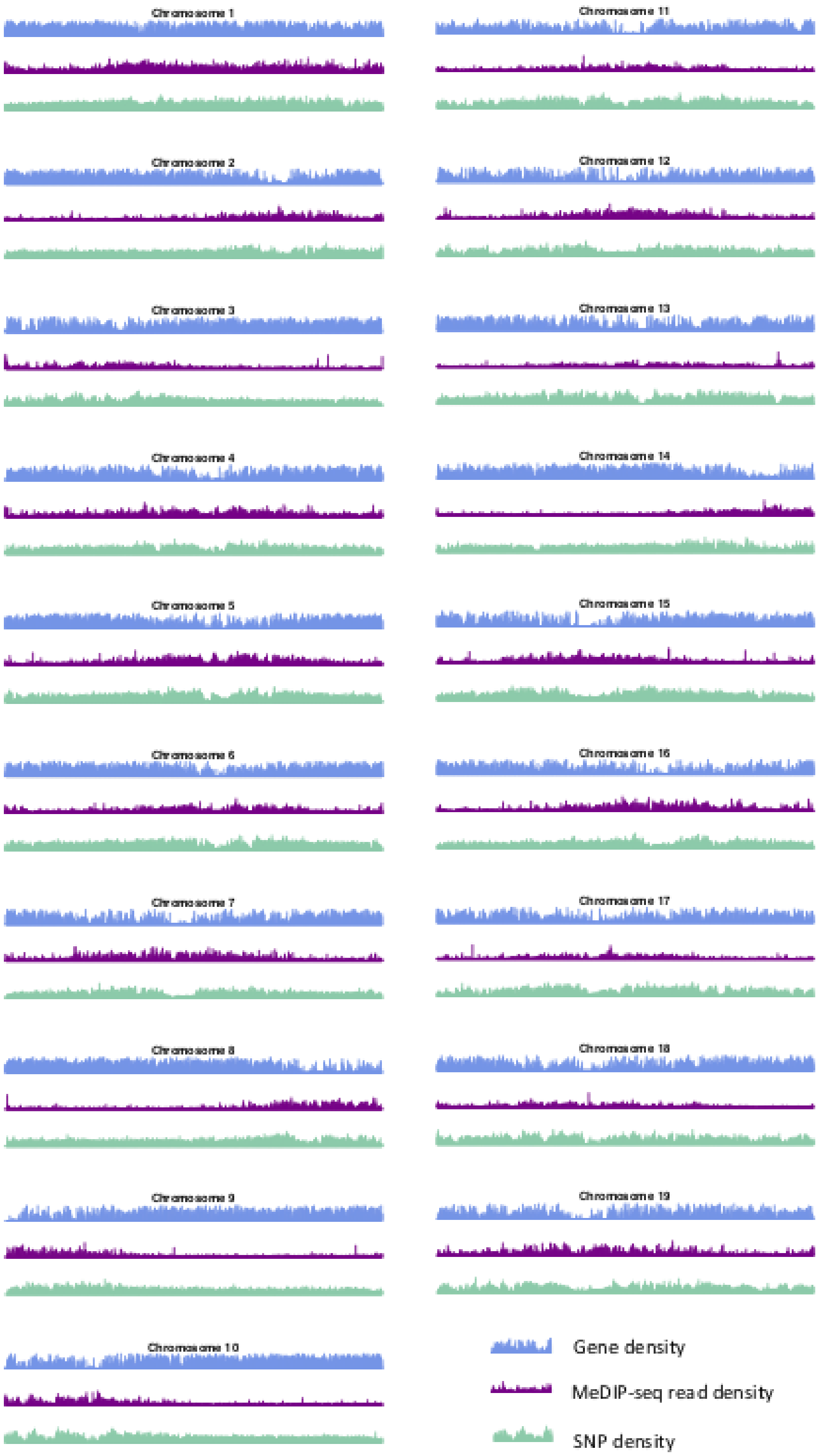


Figure 4.3: **Methylation, SNP and Gene Density.** SNP, gene and methylation (internode explant tissue) density profiles for all chromosomes of *P. trichocarpa*.

peaks) at the bottom and large scales (low frequency peaks) at the top. A wavelet coefficient is calculated for each signal position and each scale, thus resulting in a landscape. The wavelet coefficient landscapes clearly illustrate the detection of the large scale peaks (blue regions) and large scale valleys (red regions) in the upper half of the landscapes, corresponding to the visible large peaks and valleys of the signals. Plotting the wavelet coefficients at a particular scale shows the smoothed peaks and troughs of the signal at that scale (Figure 4.5A).

4.3.2 Wavelet Coefficient Landscape Signature of the Centromere

Identification of approximate centromere locations from gene density, SNP density and methylation wavelet landscapes requires knowledge of what patterns to look for. From the literature, we know that studies in *Arabidopsis* have found high methylation in the centromeric/pericentromeric regions [15], and found centromeric regions to be gene-sparse [49]. Similar conditions were found in *P. trichocarpa* [6, 50]. Though centromeric/pericentromeric regions as a whole are highly methylated, it has been found in Maize that the active centromere consists of repeats associated with CENH3 (the modified histone found in the active centromere) and is usually less methylated when compared to the pericentromeric regions [51]. A similar pattern can be observed in *Arabidopsis* [15]. Figure 4.6 shows the methylation CWT coefficient landscapes for each chromosome in internode explant tissue. One can clearly see the large-scale peaks in each chromosome indicated by the blue regions near the top of each profile, which correspond to the broad centromeric/pericentromeric regions. In 15 of the 19 chromosomes (chromosomes 1, 2, 4, 5, 6, 7, 8, 9, 10, 13, 14, 16, 17, 18, 19) we see evidence for the lowered methylation in the actual centromere when compared to the pericentromeric regions. In the coefficient landscapes, this is indicated by a medium-scale valley (red area) within and below the center of the large-scale peak, creating a "tooth-X-ray" like pattern (Figure 4.7). These centromeric wavelet coefficient signatures can also be seen in the methylation profiles of callus, female catkin, male catkin, leaf, phloem,

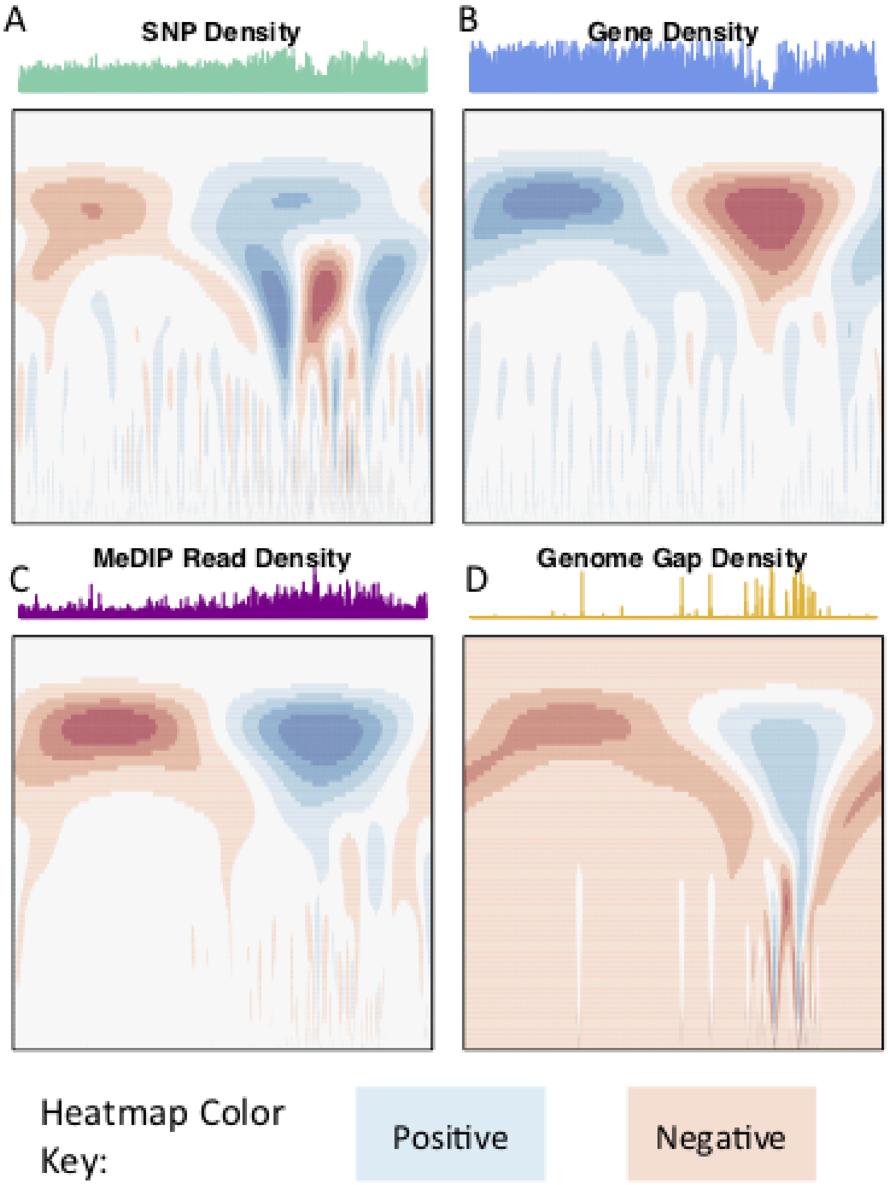


Figure 4.4: **Chromosome 2 CWT Landscapes.** CWT Coefficient landscapes of chromosome 2 for (A) SNP density, (B) gene density, (C) methylation (MeDIP-Seq read density, internode explant tissue) and (D) genome gap density. X-axes represent the bp dimension of the signals, Y-axes represent scales (*s* in Equation 5.1). Blue regions indicate positive coefficients and red regions indicate negative coefficients.

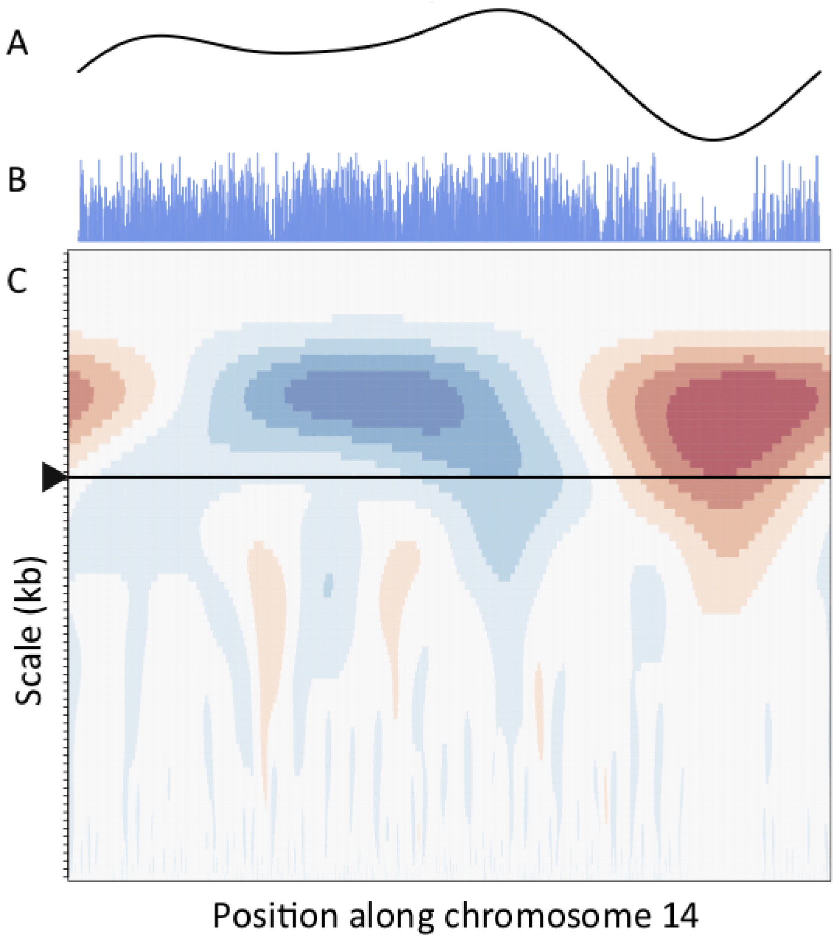


Figure 4.5: **CWT and Smooth Peaks.** CWT landscape of the gene density profile of chromosome 14. (B) is the original gene density signal, (C) is the CWT coefficient landscape of the signal and (A) shows the vector of wavelet coefficients of the scale corresponding to the large scale valley, as shown by the arrow in C.

regenerated internode, root and xylem tissues (Supplementary Figures S4.26-S4.34), but are mostly not visible in bud tissue (Supplementary Figure S4.25).

SNP density has been found to be higher in the pericentromere in *Arabidopsis* [52] and lower SNP density has been found in centromere regions in sorghum [53]. The SNP wavelet landscapes for all chromosomes all contain the "tooth-X-ray" like shape, indicating a medium-scale valley in SNP density within a large-scale peak (Figure S4.23). The location of this signature coincides with the large-scale peak in methylation (Figures S4.26-S4.34) and valley in gene density (Figure S4.22), known to be characteristic of centromeric locations. As with the methylation density, this "tooth-X-ray" shape could be indicating the pericentromeric and centromeric regions of the chromosome.

It is important to consider gaps in the assembled genome when interpreting chromosome density signals, because valleys in a density signal, such as SNP density, could be a meaningful biological signature (such as the centromere), or could be an artifact arising from a gap in the genome. Observing the density signals for all chromosomes (Figures S4.2-S4.20) and their wavelet landscapes (Figures S4.22-S4.34) one can see that in a few chromosomes, (for example, chromosome 18) the largest genome gap co-locates with the largest valley in SNP density. However, this is not true for all chromosomes. The locations of highest genome gap density do not always coincide with the largest valley in SNP density, for example, in chromosome 12 (Figure S4.21), and the largest genome gaps do not always correspond to approximate centromere locations. Thus, the tooth-X-ray shape cannot be purely driven by genome gaps, and, as such, does not appear to be an artifact.

4.3.3 Prediction of Centromere Position from Wavelet Coefficients

Based on the knowledge of centromere signatures in the literature, and the CWT landscapes of gene, SNP and methylation profiles, we attempted to locate the position of the centromere on each *P. trichocarpa* chromosome by computationally identifying the characteristic tooth-X-ray shape in the CWT landscapes. Briefly, for each chromosome, we

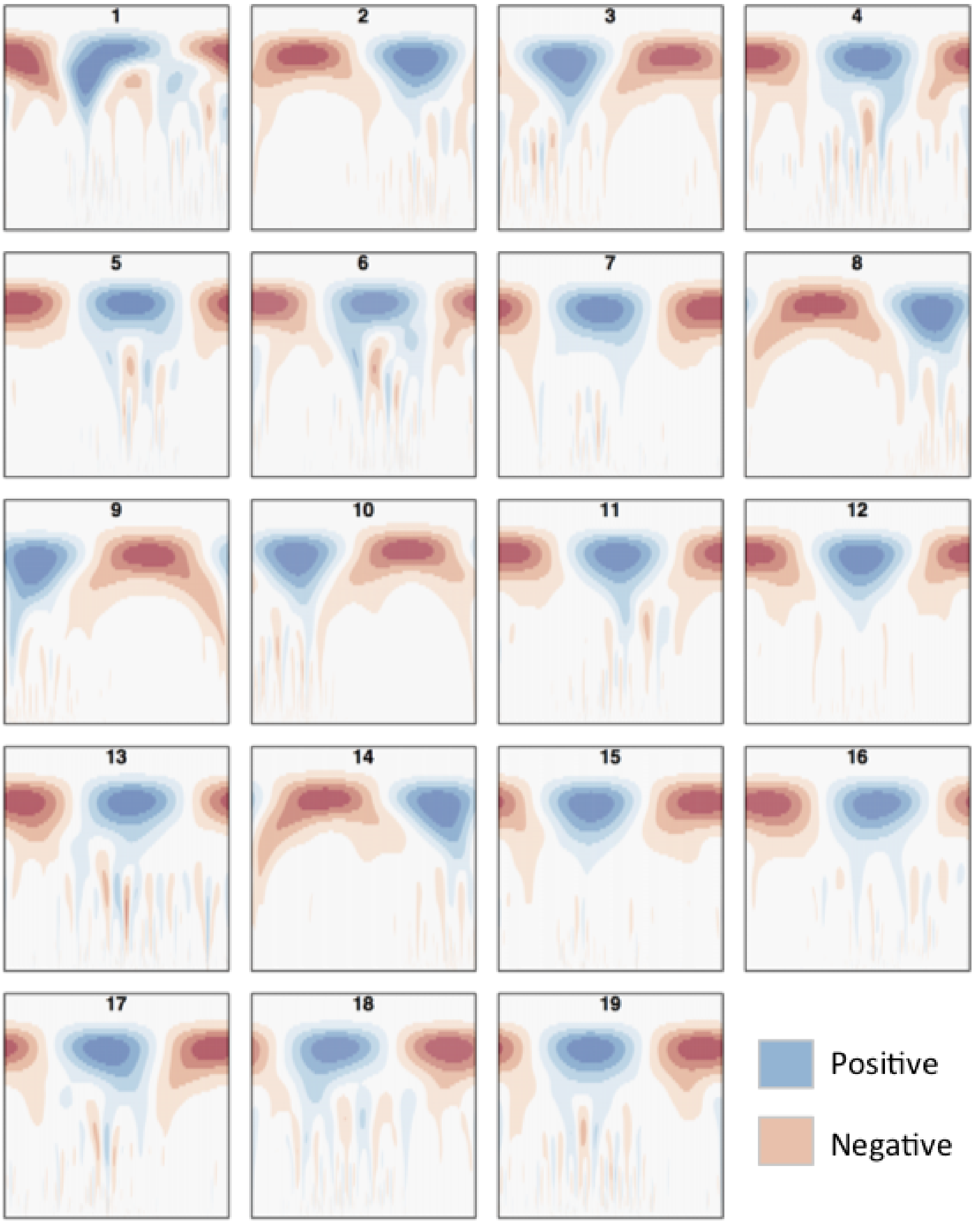


Figure 4.6: **Methylation (Internode Explant) CWT.** Methylation (internode explant tissue) CWT landscapes of each *P. trichocarpa* chromosome. For each heatmap, the x-axis represents position along the chromosome density signal (τ), the y-axis represents scale (s) and each entry represents the wavelet coefficient $W(s, \tau)$. Positive coefficients are colored blue and indicate peaks, negative coefficients are colored red and indicate valleys. The "tooth-x-ray" centromeric signature is evident in many chromosomes, consisting of a broad-scale peak encompassing the centromeric/pericentromeric regions, and the lower scale valley within the large peak indicating the centromeric region.

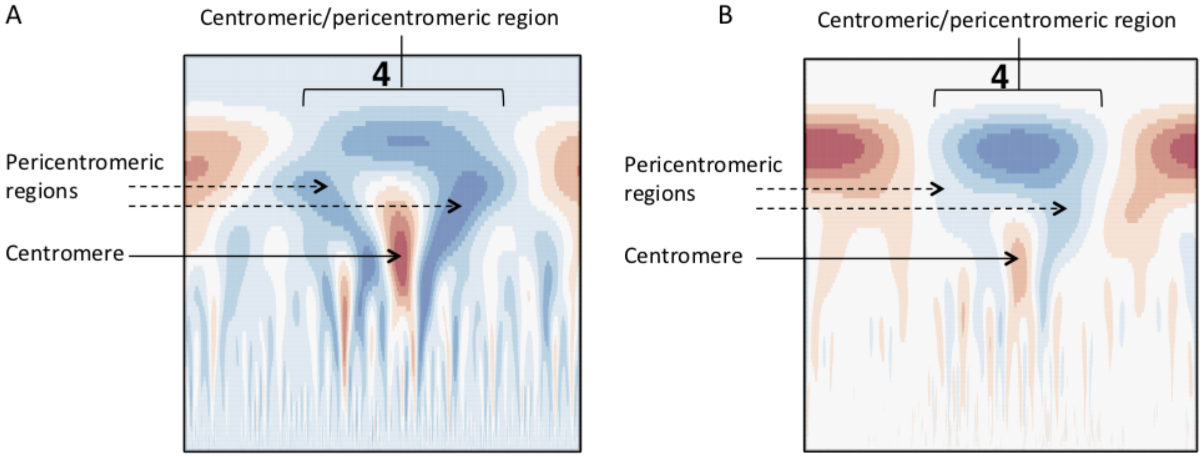


Figure 4.7: **Methylation Wavelet Signature of Centromere.** "Tooth-x-ray" centromeric signature for (A) SNP density and (B) methylation density, consisting of a broad-scale peak encompassing the centromeric/pericentromeric regions, and the lower scale valley within the large peak indicating the centromeric region.

calculate the CWT of the scaled SNP density and methylation profiles, resulting in two coefficient landscapes. We identify the pericentromeric scale as the scale at which we find the maximum wavelet coefficient in the upper third of the methylation landscape, and identify the borders of the pericentromeric region as the zeroes of the wavelet vector on either side of the maximum coefficient. We then identify the centromeric scale as the minimum wavelet coefficient in the SNP wavelet landscape within the borders of the pericentromeric region, and then consider the approximate center of the centromere location to be the point of maximum difference between the methylation wavelet coefficients at the pericentromere scale and SNP wavelet coefficients at the centromere scale (Figures 4.8, Figure S4.1, Supplementary Text S4.1), and the general centromeric region borders as the points of intersection between the these two vectors on either side of the center (Table S4.1, yellow bars in Figure 4.8).

Mapping centromere repeats from various plants from the PGSB Repeat Element Catalog [42] as well as repeat sequences which were found to identify centromeres on certain *P. trichocarpa* chromosomes in a previous assembly [16] using BLAST were consistent with the locations of centromeres identified using wavelet coefficients. Predicted centromere positions aligned well with the density profiles of repeat sequence BLAST hits, indicating

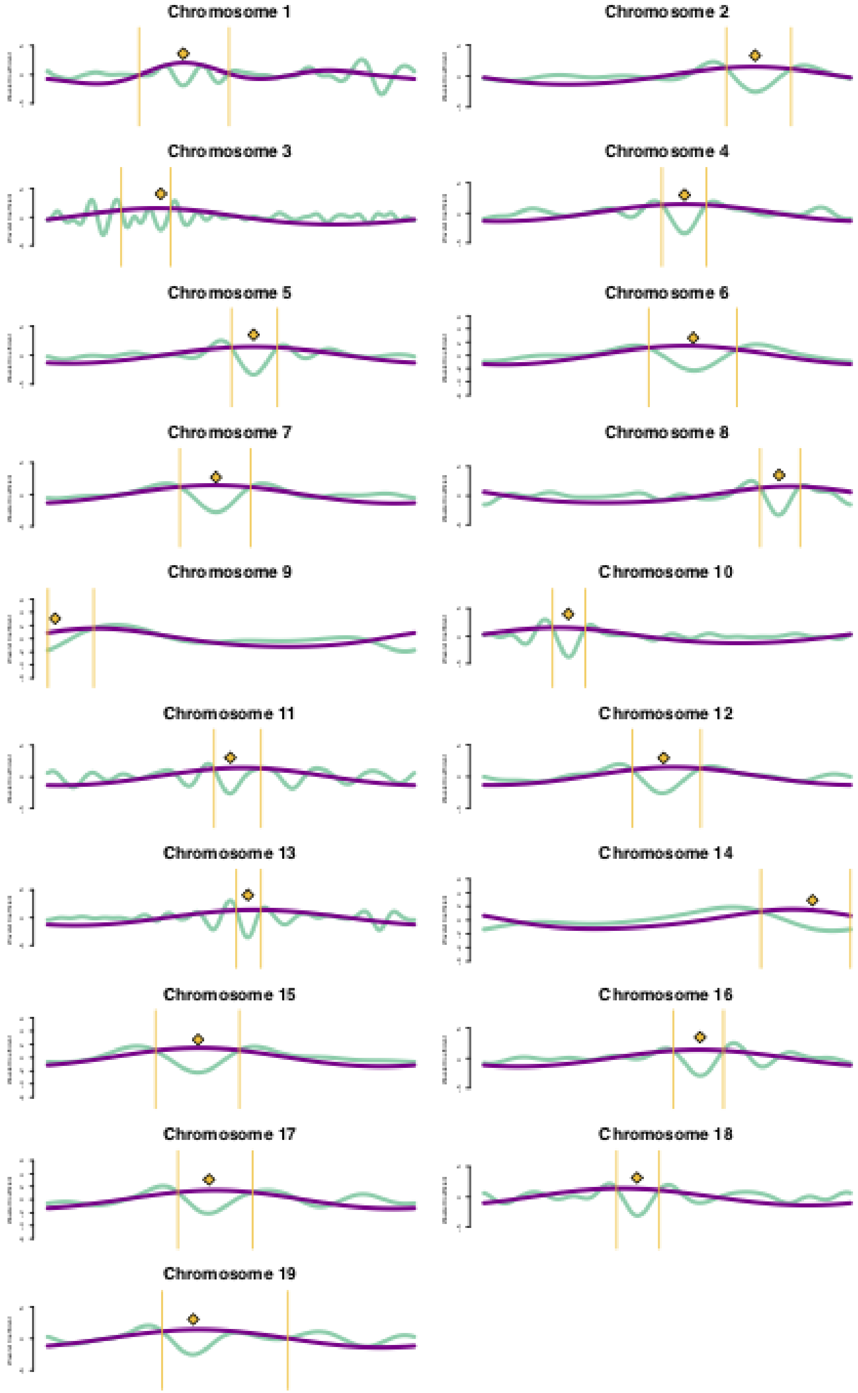


Figure 4.8: **Centromere Positions.** Line plots for each chromosome of methylation wavelet coefficients (internode explant tissue) at pericentromeric scale (purple lines) and SNP density wavelet coefficients at centromeric scale (green lines). Yellow diamonds represent the putative centromeric location, calculated as the point of maximum difference between the wavelet coefficients at these two scales. Yellow bars indicate the general centromeric region as the points of intersection between the two curves on either side of the centromeric region. See methods for further details.

that our centromere prediction strategy is likely identifying valid centromere positions (Figure 4.9). The wavelet-based centromere identification through the use of multiple lines of evidence allows us to be more certain of centromeric regions, and also allows more specific locations to be identified than can be done by simply looking at repeat density, which map to broad regions of the genome. Layering multiple data types allows for the identification of putative centromere positions based on multiple lines of evidence, and thus, allows one to be more certain of their location.

P. trichocarpa chromosomes contain homologous genome blocks, presumed to be derived from the salicoid genome duplication [2]. Looking at the positions of predicted centromeres in Figures 4.8 and 4.9, some paralogous chromosomes (see [2]) appear to have similar centromeric positions (for example, chromosomes 8 and 10, and chromosomes 12 and 15). This suggests that the current centromere positions potentially predate the salicoid duplication event.

4.3.4 Co-evolution of Putative CENH3 with Centromeric Sequences

The histone CENH3 epigenetically defines centromere position, and replaces normal histone H3 in the nucleosomes at the centromere [54]. Silencing of this gene in *Arabidopsis* has been found to cause dwarfism, reduced mitotic divisions and sterility [55]. CENH3 has been found to be adaptively evolving in *Arabidopsis* [10]. Analysis of CENH3 in various *Brassicaceae* showed that it is evolving adaptively at various sites which are potentially in contact with the centromeric DNA [11]. There is thus the hypothesis that CENH3 is co-evolving with the sequence of the centromere [11, 12]. In a study involving a *A. thaliana* CENH3-null mutant expressing a *Zea mays* CENH3, it was found that while the *Zea mays* CENH3 localized to the same locations as endogenous *A. thaliana* CENH3, the *Z. mays* CENH3 centromeres were weaker, and resulted in genome elimination in crosses with wild-type *A. thaliana* [56]. Thus, the sequence of CENH3 could potentially have an impact on the strength of the centromere.

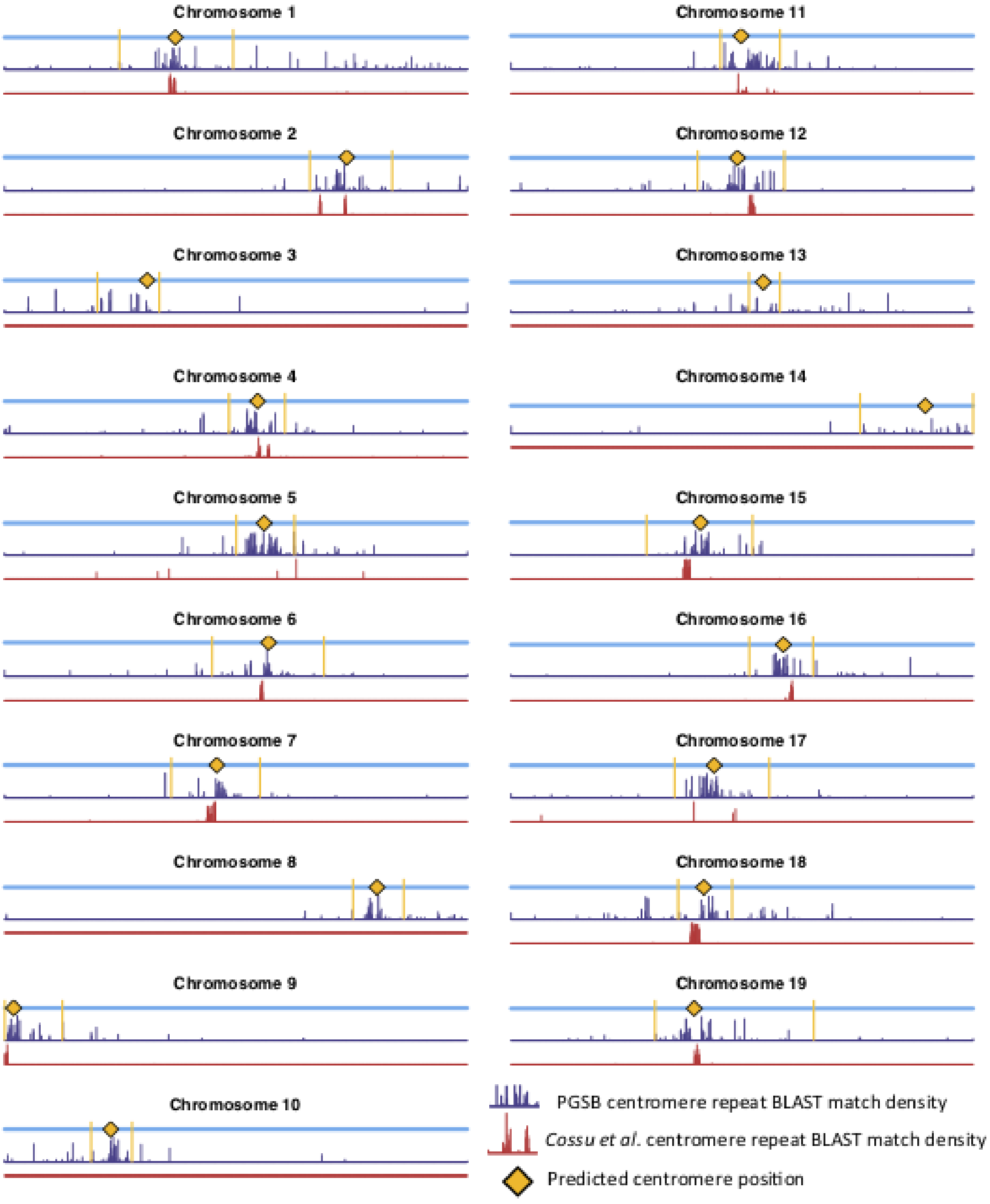


Figure 4.9: **Centromere positions and Centromere repeats.** Putative centromere positions (yellow diamonds) identified as in Figure 4.8 using methylation and SNP wavelet coefficients, as well as the density of BLAST matches of plant centromere repeat sequences (navy bars) and putative *P. trichocarpa* centromere repeat sequences [16] (red bars).

If the hypothesis of co-evolution between the CENH3 and centromeric sequences is true, one would expect to see correlations between Single Nucleotide Polymorphisms (SNPs) in *P. trichocarpa* CENH3 and *P. trichocarpa* centromeric regions. CENH3 is mostly a single copy in diploids such as *Arabidopsis* [54] but there are some species that contain more than one copy. Wheat has two distinct copies of CENH3, and they seem to be evolutionarily divergent. They have different expression patterns, and one of them shows positive selection [57]. We identified two putative CENH3 genes in *P. trichocarpa* (Potri.014G096400 on chromosome 14 and Potri.002G169000 on chromosome 2) as BLAST [43] matches of *A. thaliana* CENH3 (AT1G01370). It is interesting to note that chromosomes 2 and 14 are salicoid duplication paralogs. Of these two genes, Potri.014G096400 was annotated as being similar to a CENH3 gene, whereas Potri.002G169000 had no functional annotations. RNA-seq and EST information on Phytozome [7] confirmed that both of these genes are expressed (Figure S4.35). Expression information of these genes in the *P. trichocarpa* gene atlas on PhytoMine [44, 7] showed that the expression of these two genes varies across tissues, however, they are not co-expressed with one another (Figure 4.10). Both Potri.014G096400 and Potri.002G169000 genome sequences had multiple hits with CENH3 genes when BLASTed on NCBI.

We determined correlations between all pairs of $\sim$10,000,000 high confidence SNPs in a population of 882 *P. trichocarpa* genotypes using the CCC metric [25, 26, 27] and extracted SNPs within Potri.014G096400 and Potri.002G169000 that had correlations with SNPs elsewhere in the genome. When using a call rate constraint minimum of a 100 called alleles ($\sim 5\%$), a minimum overlap of 100 non-missing alleles in SNP correlations and requiring a minor allele frequency (MAF) $\geq$ 0.01, we find concentrations of SNPs in the centromeric region of various chromosomes which are correlated with SNPs in Potri.002G169000 (Figure 4.11, Figure S4.37). We thus find strong evidence for the co-evolution for CENH3 with the centromeric sequences.

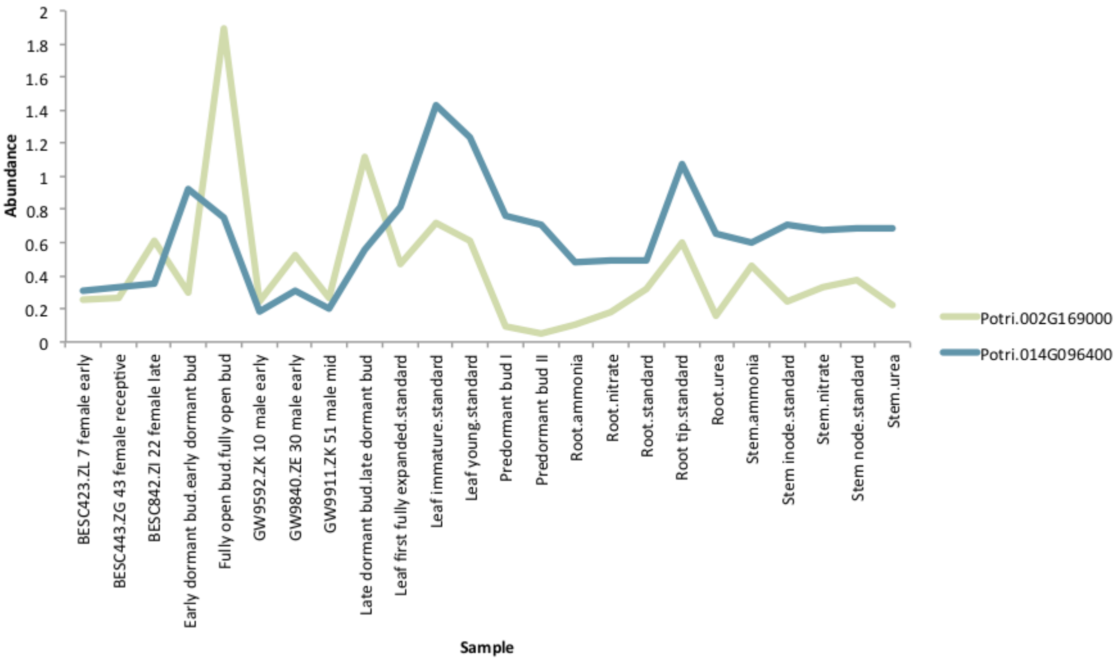


Figure 4.10: **CENH3 expression.** Expression levels of putative *P. trichocarpa* CENH3 genes, Potri.002G169000 and Potri.014G096400. Expression data obtained from PhytoMine on Phytozome [7].

In particular, it appears that Potri.002G169000 seems to have a co-evolution signature with the centromere, much more so than Potri.014G096400, in that Potri.002G169000 contained SNPs correlating with 13 out of 19 centromeric regions, whereas Potri.014G096400 contained SNPs correlating with 5 out of 19 centromeric regions (centromeric regions in Table S4.1). While both Potri.002G169000 and Potri.014G096400 on average have more mutations than other *P. trichocarpa* histones (an expected phenomemon as CENH3 histones accumulate mutations faster than normal histones, as mentioned in [58]), Potri.002G169000 contains more mutations than Potri.014G096400 (Figure S4.36, Table S4.2). Potri.014G096400 is also co-expressed with various other non-CENH3 histones, as well as a histone deacetylase and a histone methyltransferase on PhytoMine [44, 7] (Table S4.3), and the correlation neighbourhood of Potri.002G169000 and Potri.014G096400 do not overlap at all (Figure 4.12).

This seems to suggest that these two genes are functionally divergent. Given the facts that Potri.002G169000 has strong co-evolution signatures with the centromere (Figure 4.11) and Potri.014G096400 is co-expressing with non-CENH3 histones, one could hypothesize

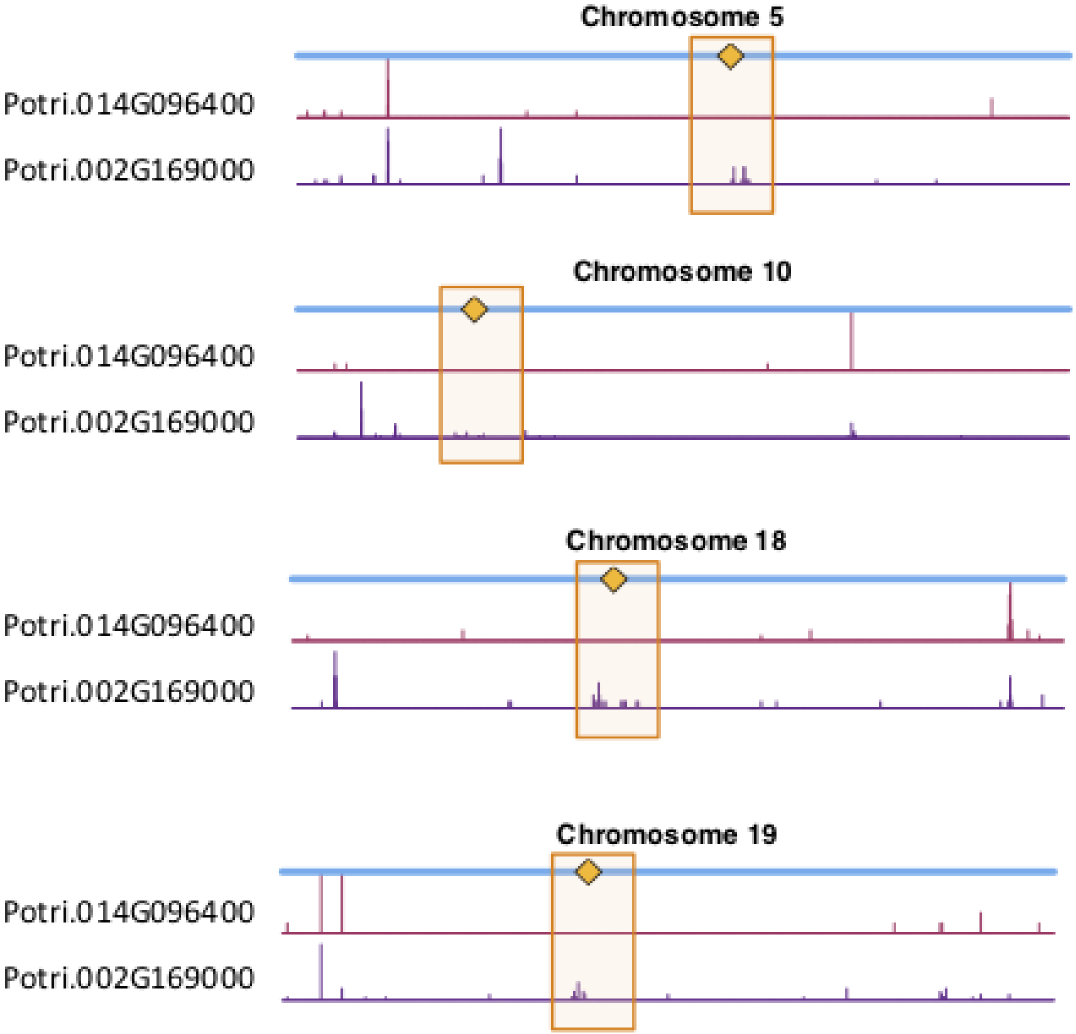


Figure 4.11: **CENH3 co-evolution.** SNP density profiles across a selection of chromosomes involving SNPs which correlate with SNPs in putative CENH3 genes, Potri.002G169000 and Potri.014G096400 across a population of *P. trichocarpa* genotypes. One can clearly see the clusters of SNPs in the centromeric regions which are correlating with SNPs within these CENH3 genes.

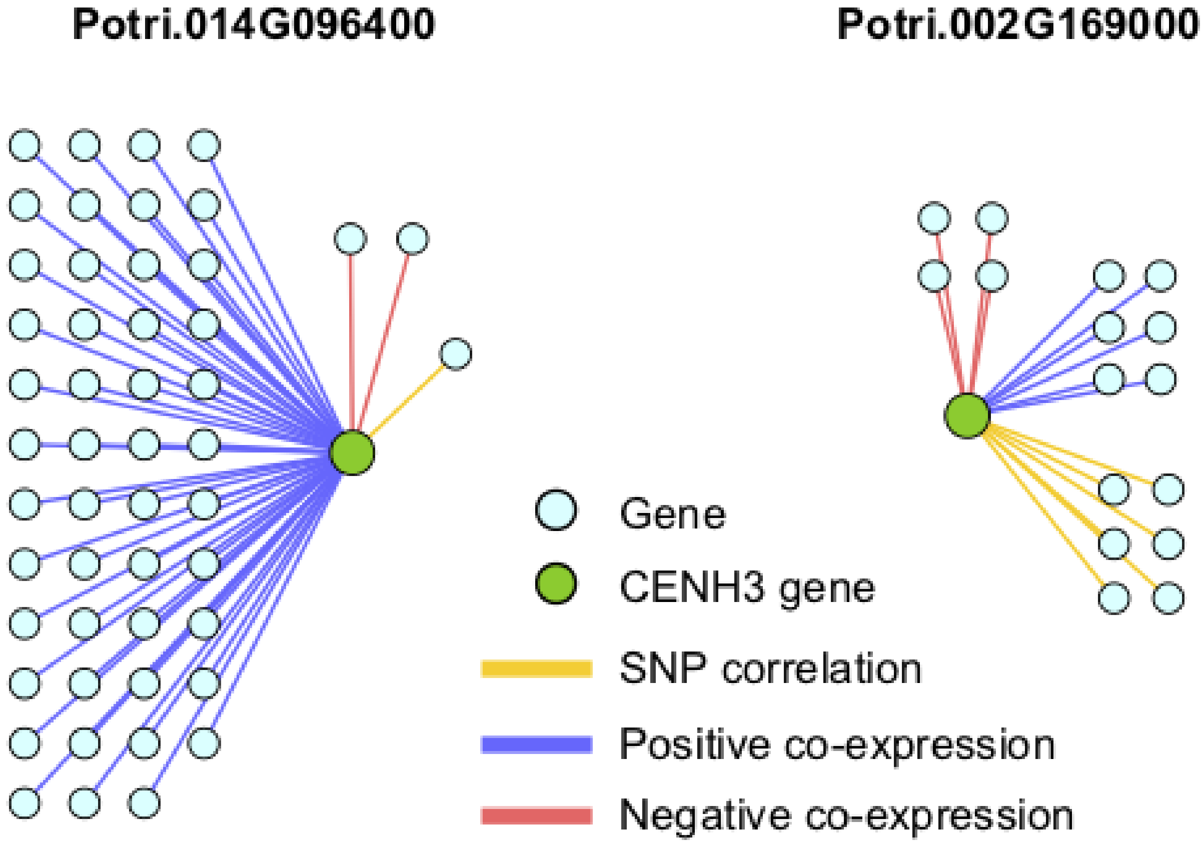


Figure 4.12: **CENH3 gene correlations.** Correlations of *P. trichocarpa* CENH3 genes (green circles) with other genes (aqua circles), including positive co-expression (blue), negative co-expression (red) and SNP correlations (yellow).

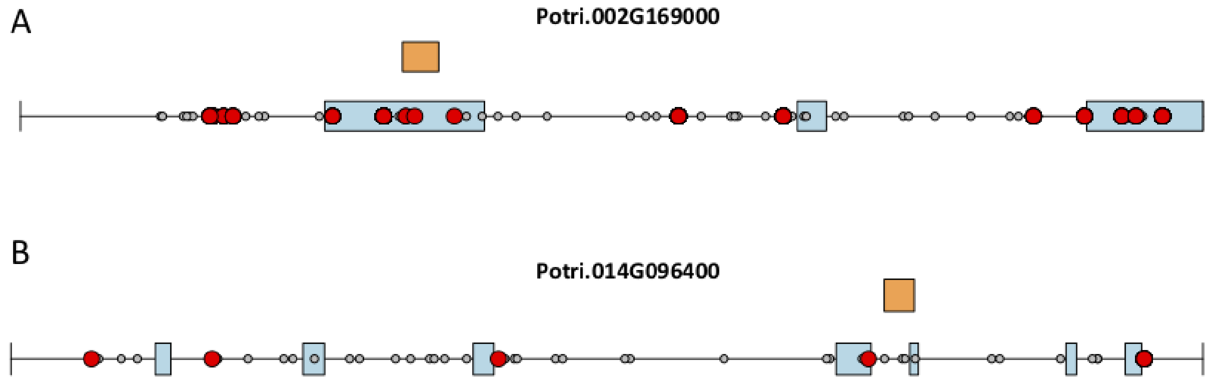


Figure 4.13: **CENH3 mutations.** SNPs in putative *P. trichocarpa* CENH3 genes (A) Potri.002G169000 and (B) Potri.014G096400. Exons (blue boxes) for Potri.014G096400 were determined from the v3.0 genome annotation on Phytozome [7], and from mapped ESTs on Phytozome JBrowse [59, 7] for Potri.002G169000. Grey circles represent SNPs, red circles represent SNPs that correlate with SNPs in centromeric regions. Orange rectangles indicate the location of the histone domain as determined using NCBI CDScan.

that Potri.002G169000 (a previously unannotated gene) is the primary functioning CENH3 in *P. trichocarpa* while Potri.014G096400 could be functioning more like a normal histone. If one looks at the position of SNPs within Potri.002G169000 and Potri.014G096400, it is evident that Potri.002G169000 contains more SNPs in transcribed regions of the gene that correlate with centromeric regions (Figure 4.13). In addition, Potri.002G169000 contains more SNPs in/near the histone domain that correlate with the centromere, when compared to Potri.014G096400. Potri.002G169000 also has more of the expected structure for a CENH3 gene, containing the histone domain in the C terminal, and having a variable N terminal domain [54].

Based on these various lines of evidence, we suggest that the previously unnannotated Potri.002G169000 is the primary functioning CENH3 gene in *P. trichocarpa*.

4.3.5 Concluding Remarks

In this study we performed wavelet-based signal processing of multiple, heterogeneous data types to identify centromere positions and properties in *P. trichocarpa*. We found centromeres to be in gene-sparse regions, and found centromeric/pericentromeric regions to hypermethylated relative to the rest of the chromosomes, and found centromeric DNA to be hypomethylated relative to pericentromeric regions in many chromosomes across various tissues. The "tooth-X-ray" wavelet signature was identified as a characteristic signature of the centromere in the wavelet landscapes of SNP density profiles.

The use of wavelet coefficients allowed us to identify the approximate centromeric locations. These locations were supported by mapping of repeat sequences, and could be further validated through experimental techniques such as ChIP (chromatin immunoprecipitation)-Seq. We also found evidence for the co-evolution of the sequence of the centromere-specific histone CENH3 with the sequences of the centromere on many chromosomes. In particular, we found that the previously unannotated gene Potri.002G169000 is the most likely candidate for an active, centromere-co-evolving CENH3 gene in *P. trichocarpa* and not the currently annotated CENH3 gene, Potri.014G096400.

This study illustrated the utility of wavelet-based signal processing of genomic signals to identify structural characteristics of chromosomes. While this study made primary use of the larger-scale wavelet coefficients, we would recommend the use of the smaller scale wavelet coefficients to investigate smaller-scale structural characteristics, such as nucleosome occupancy.

Funding

Funding 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.

This research was also supported by the Plant-Microbe Interfaces Scientific Focus Area (http://pmi.ornl.gov) in the Genomic Science Program, the Office of Biological and Environmental Research (BER) in the U.S. Department of Energy Office of Science, and by the Department of Energy, Laboratory Directed Research and Development funding

(7758), at the Oak Ridge National Laboratory. Oak Ridge National Laboratory is managed by UT-Battelle, LLC, for the US DOE under contract DE-AC05-00OR22725.

An award of computer time was provided by the INCITE program. This research also 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 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.

4.4 Supplementary Material

4.4.1 Text S4.1: Wavelet-based Centromere Identification

For the demonstrated centromere identification, we used the methylation density signal from internode explant tissue. Centromere identification was performed independently for each chromosome. SNP density and methylation density signals were mean centered (mean = 0) scaled to have standard deviation = 1. The continuous wavelet transform (CWT) was then performed on the scaled signal vectors. The methylation wavelet landscape was then used to identify the "general" centromeric/pericentromeric region of the chromosome. This gives us a general region of the chromosome to start looking for the "tooth-X-ray" shape, since identifying the centromere "valley" in the SNP wavelet landscapes by looking for minimum wavelet coefficients could instead identify valleys in other parts of the chromosome (see for example in chromosome 1, Figure S4.23, the minimum wavelet coefficient does not appear in the middle of the "tooth-X-ray" shape.)

The procedure for identifying centromere positions is as follows:

1. Identify the position of the maximum wavelet coefficient in the upper third of the methylation landscape. We call the scale at which this maximum occurs the "pericentromere scale".

2. Find the putative pericentromere borders as the zeros on their side of this maximum. If the centromere is near the end of the chromosome, the one "pericentromere border" might be the edge of the chromosome.

3. Identify the the minimum coefficient in the lower two thirds of the SNP wavelet landscape, between the approximate pericentromere borders. We call the scale at which this minimum occurs the "centromere scale".

4. Extract the SNP wavelet coefficient vector at centromere scale and the methylation wavelet coefficient vector at pericentromere scale.

5. Mean center (mean = 0) and scale these vectors to have standard deviation 1, and find the approximate centromere location as the position of maximum difference between these two vectors.

4.4.2 Supplementary Figures

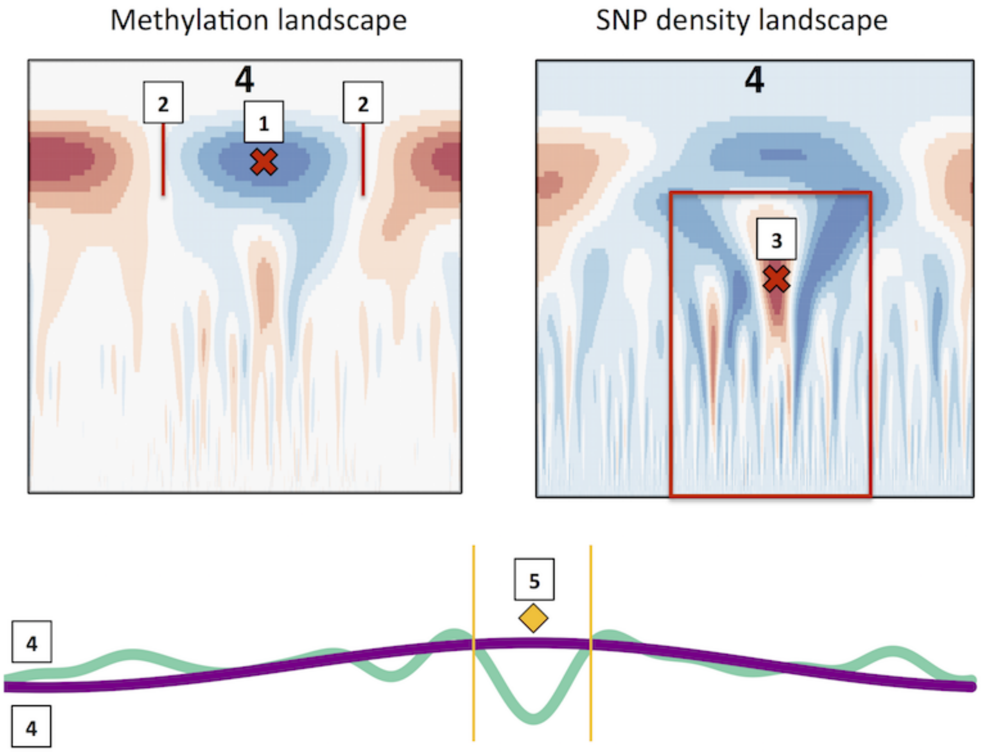


Figure S4.1: **Centromere identification.** Approach for identifying centromere positions from wavelet coefficient landscapes.

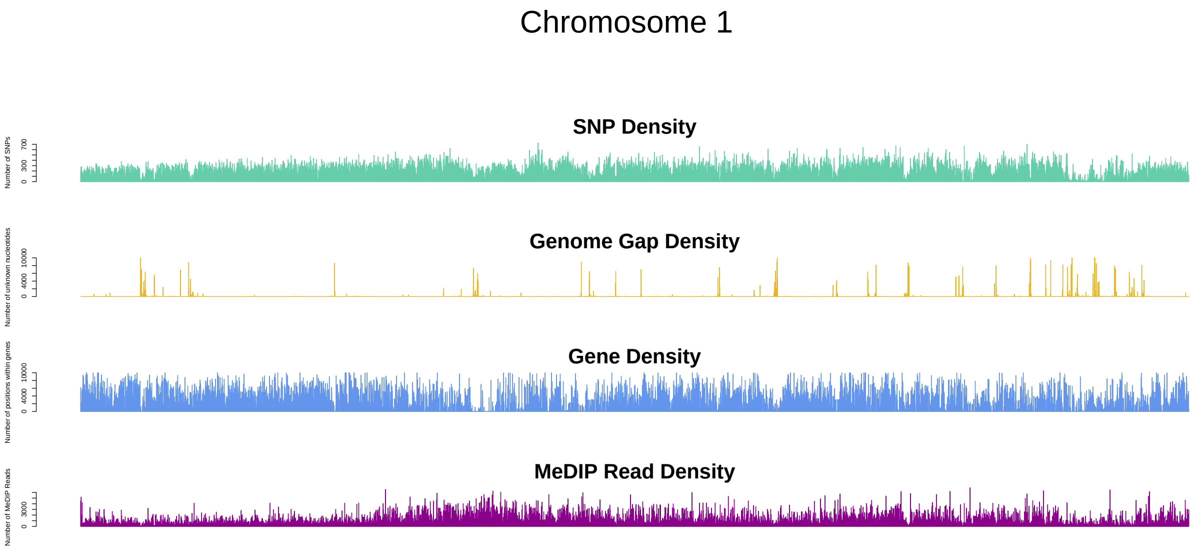


Figure S4.2: **Density profiles for chromosome 1.** Profile of SNP positions, genome gap positions, gene positions and MeDIP read positions in 10kb bins across *P. trichocarpa* chromosome 1.

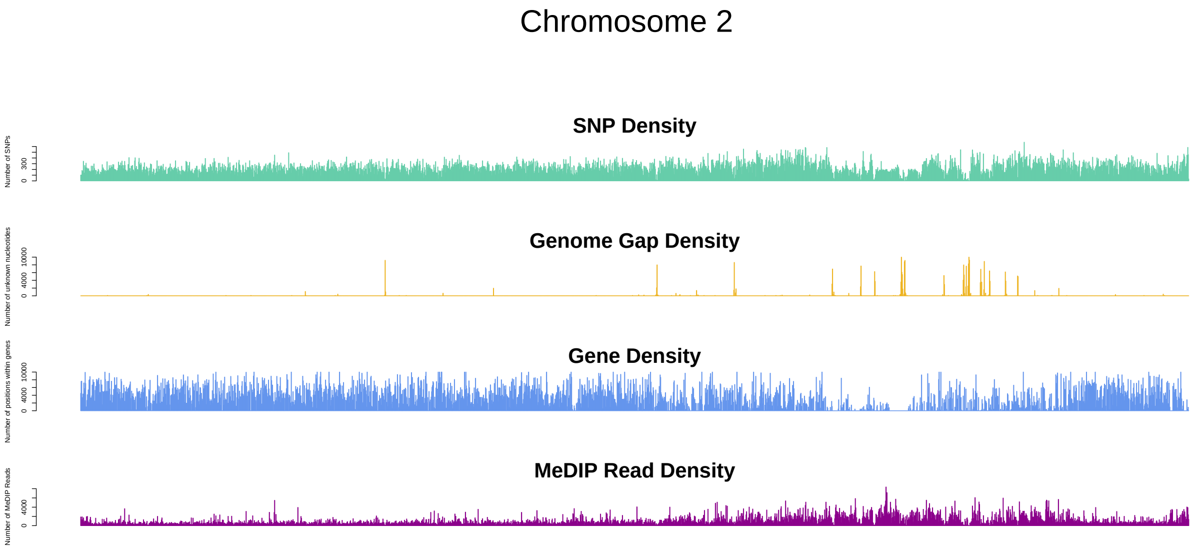


Figure S4.3: **Density profiles for chromosome 2.** Profile of SNP positions, genome gap positions, gene positions and MeDIP read positions in 10kb bins across *P. trichocarpa* chromosome 2.

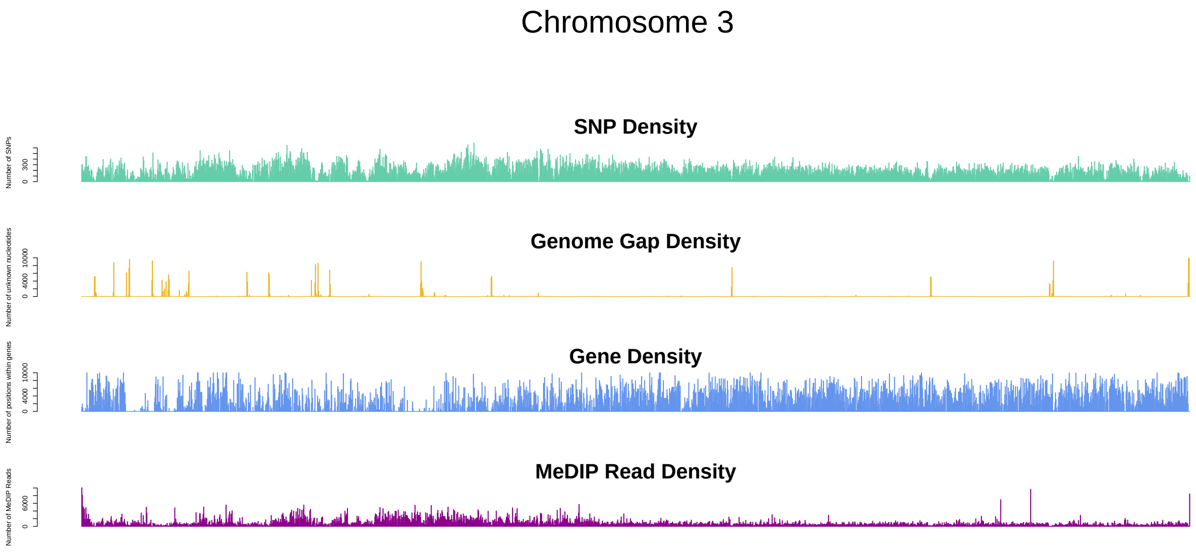


Figure S4.4: **Density profiles for chromosome 3.** Profile of SNP positions, genome gap positions, gene positions and MeDIP read positions in 10kb bins across *P. trichocarpa* chromosome 3.

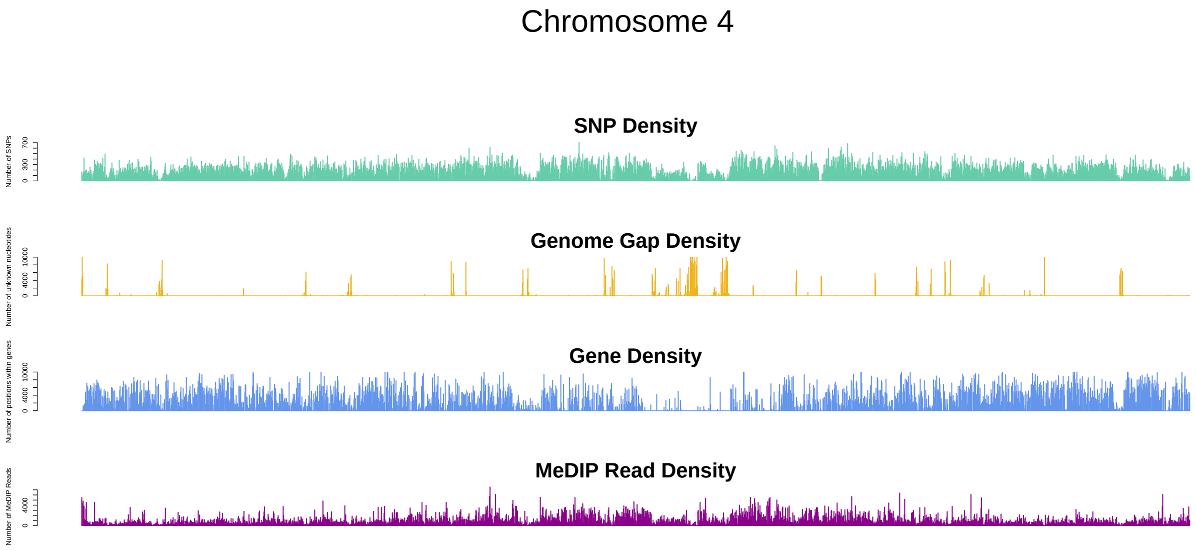


Figure S4.5: **Density profiles for chromosome 4.** Profile of SNP positions, genome gap positions, gene positions and MeDIP read positions in 10kb bins across *P. trichocarpa* chromosome 4.

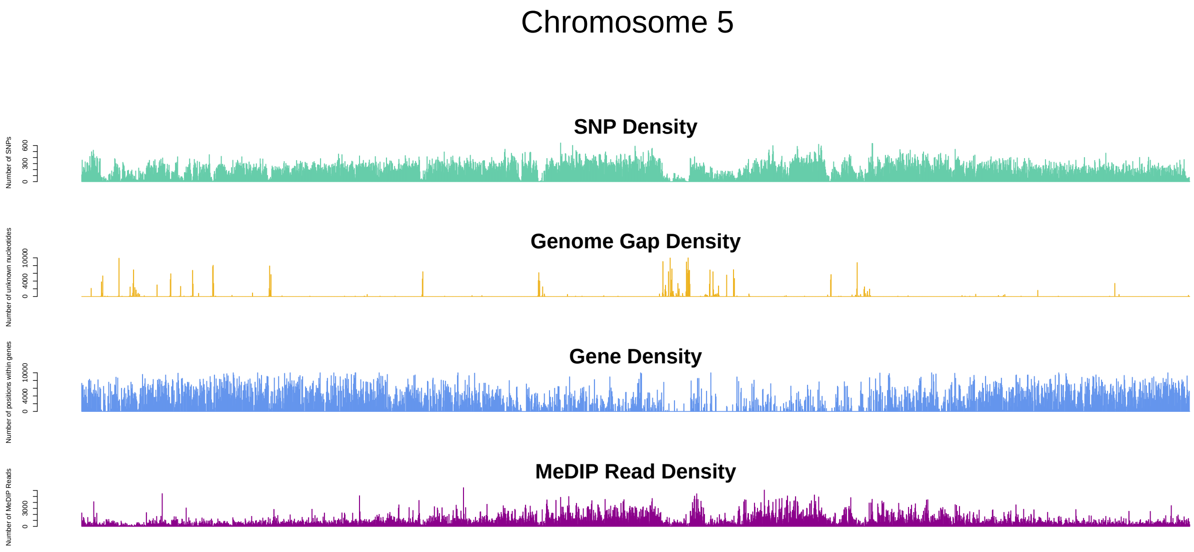


Figure S4.6: **Density profiles for chromosome 5.** Profile of SNP positions, genome gap positions, gene positions and MeDIP read positions in 10kb bins across *P. trichocarpa* chromosome 5.

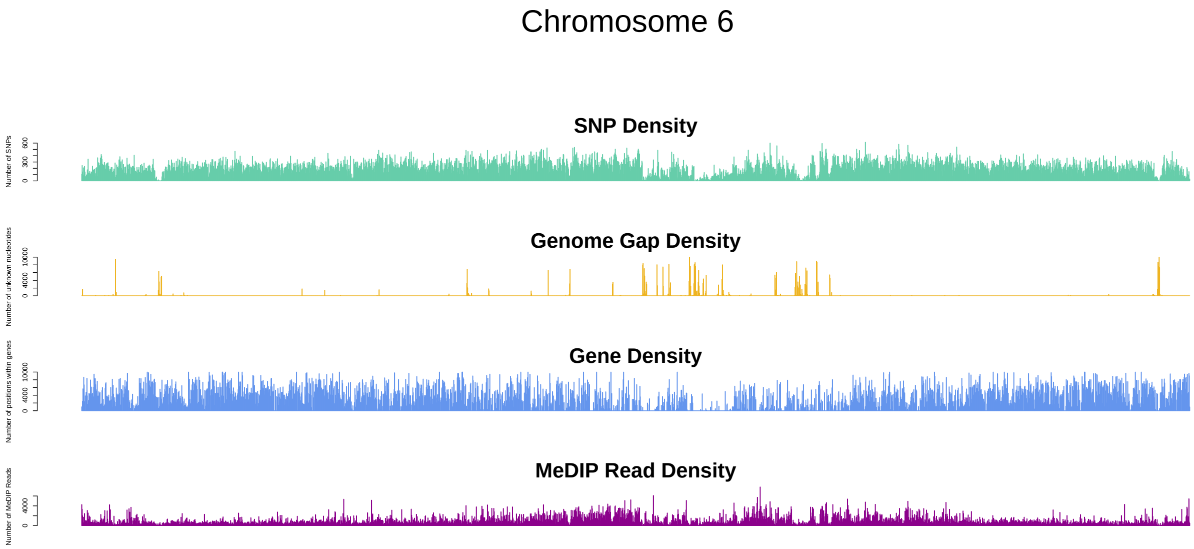


Figure S4.7: **Density profiles for chromosome 6.** Profile of SNP positions, genome gap positions, gene positions and MeDIP read positions in 10kb bins across *P. trichocarpa* chromosome 6.

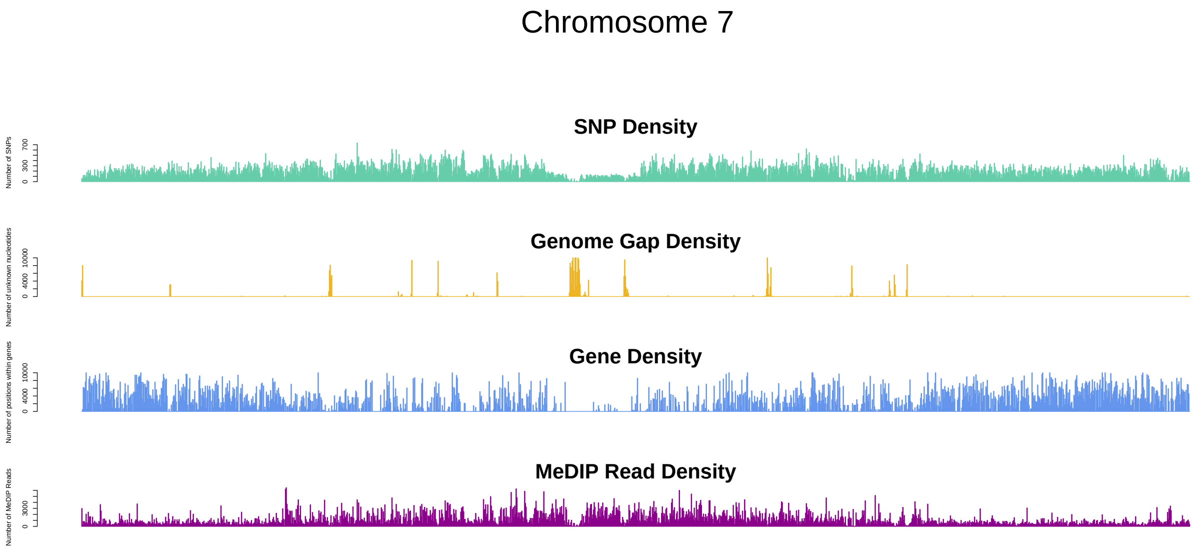


Figure S4.8: **Density profiles for chromosome 7.** Profile of SNP positions, genome gap positions, gene positions and MeDIP read positions in 10kb bins across *P. trichocarpa* chromosome 7.

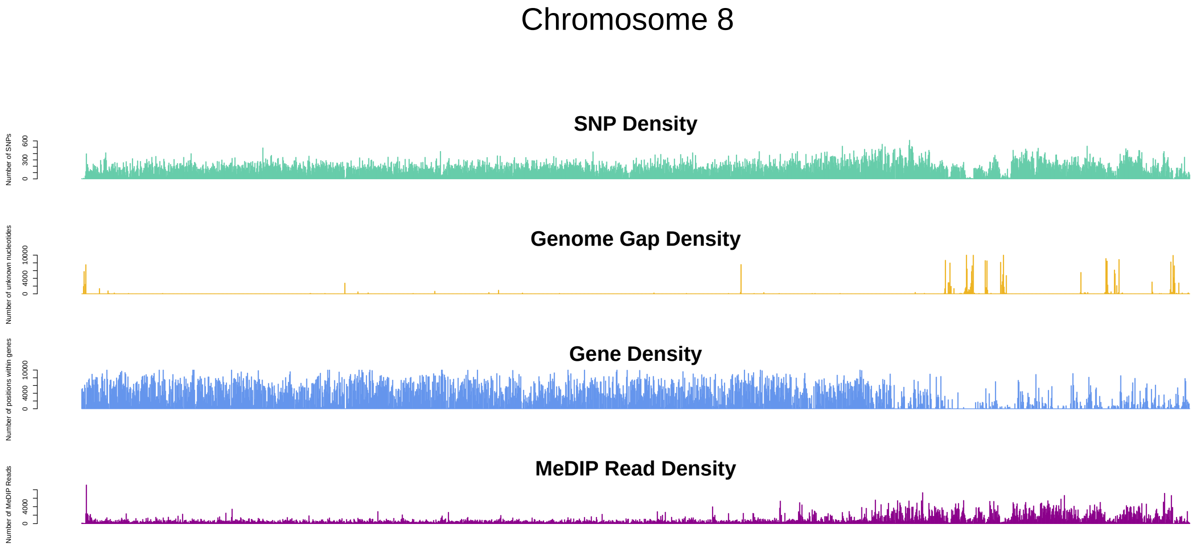


Figure S4.9: **Density profiles for chromosome 8.** Profile of SNP positions, genome gap positions, gene positions and MeDIP read positions in 10kb bins across *P. trichocarpa* chromosome 8.

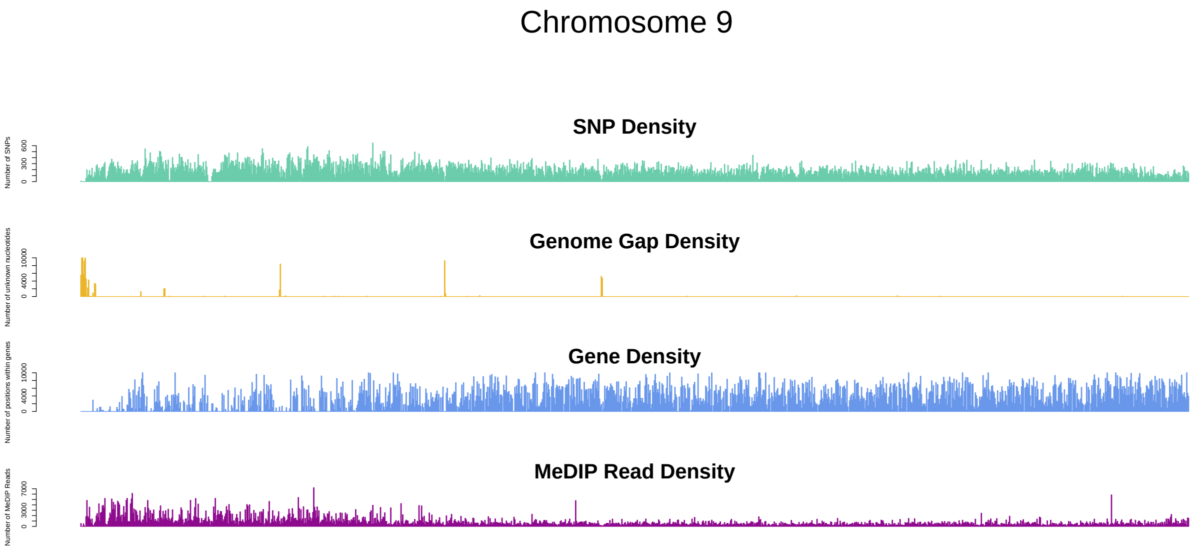


Figure S4.10: **Density profiles for chromosome 9.** Profile of SNP positions, genome gap positions, gene positions and MeDIP read positions in 10kb bins across *P. trichocarpa* chromosome 9.

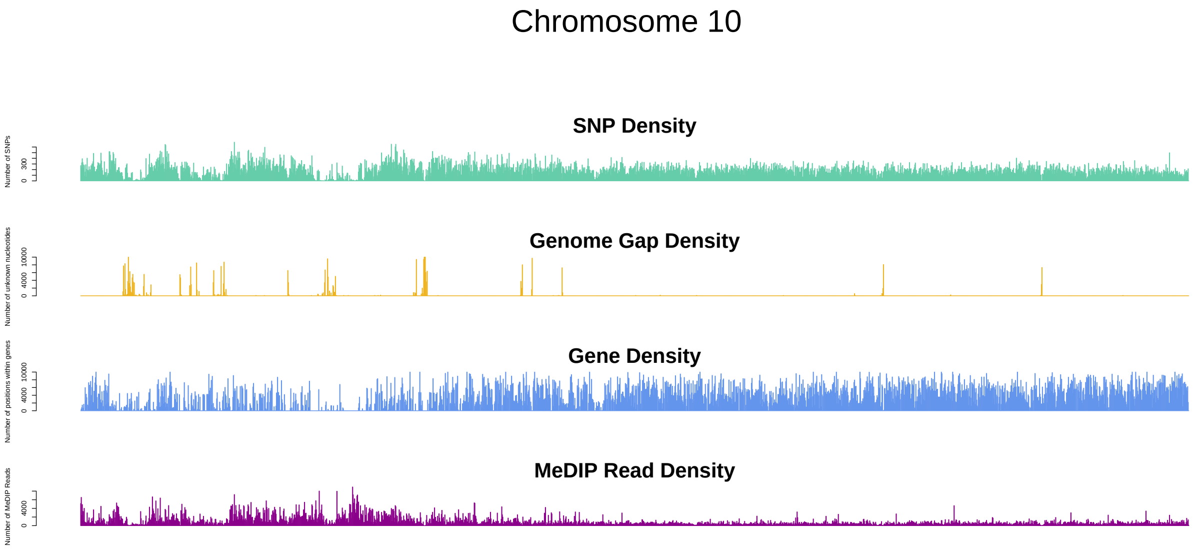


Figure S4.11: **Density profiles for chromosome 10.** Profile of SNP positions, genome gap positions, gene positions and MeDIP read positions in 10kb bins across *P. trichocarpa* chromosome 10.

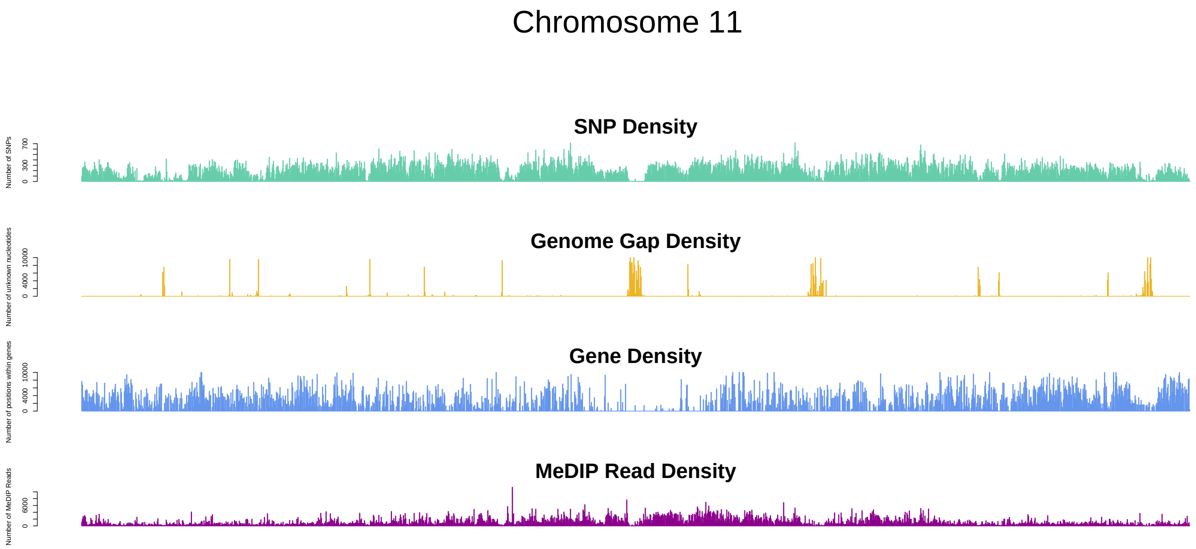


Figure S4.12: **Density profiles for chromosome 11.** Profile of SNP positions, genome gap positions, gene positions and MeDIP read positions in 10kb bins across *P. trichocarpa* chromosome 11.

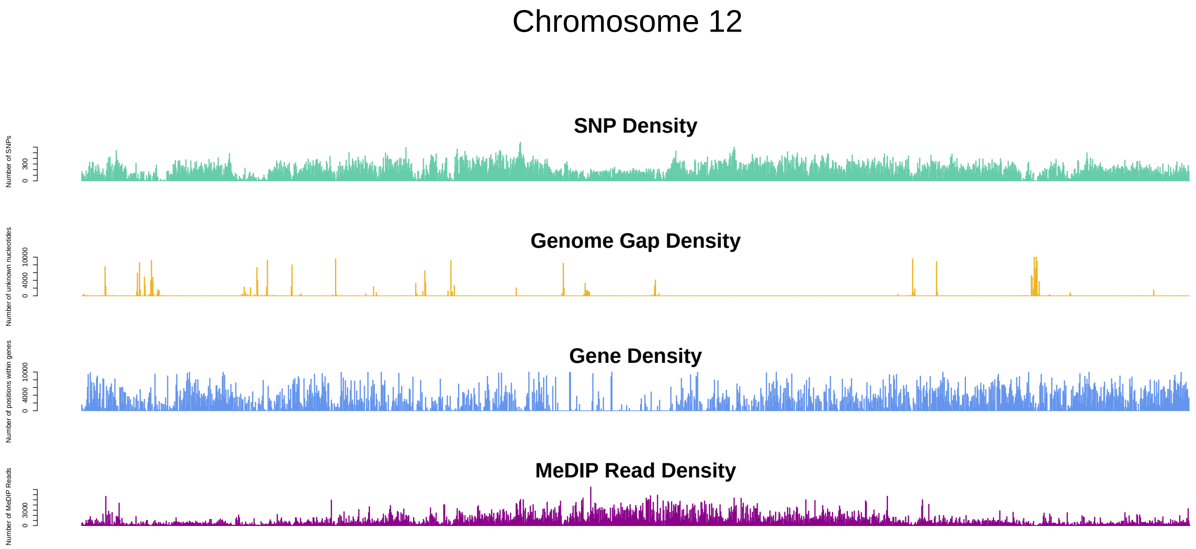


Figure S4.13: **Density profiles for chromosome 12.** Profile of SNP positions, genome gap positions, gene positions and MeDIP read positions in 10kb bins across *P. trichocarpa* chromosome 12.

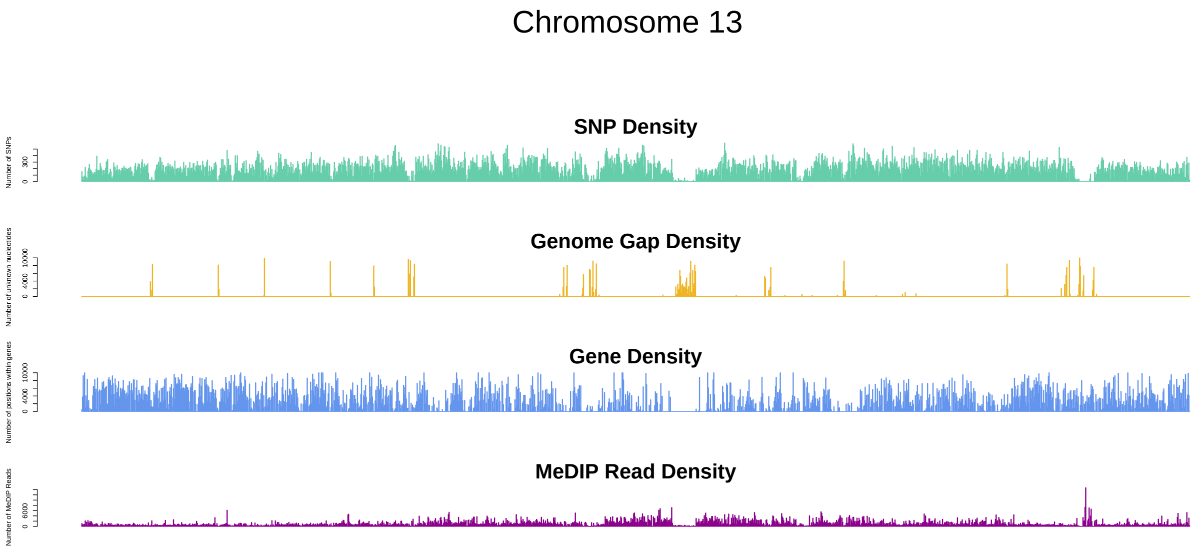


Figure S4.14: **Density profiles for chromosome 13.** Profile of SNP positions, genome gap positions, gene positions and MeDIP read positions in 10kb bins across *P. trichocarpa* chromosome 13.

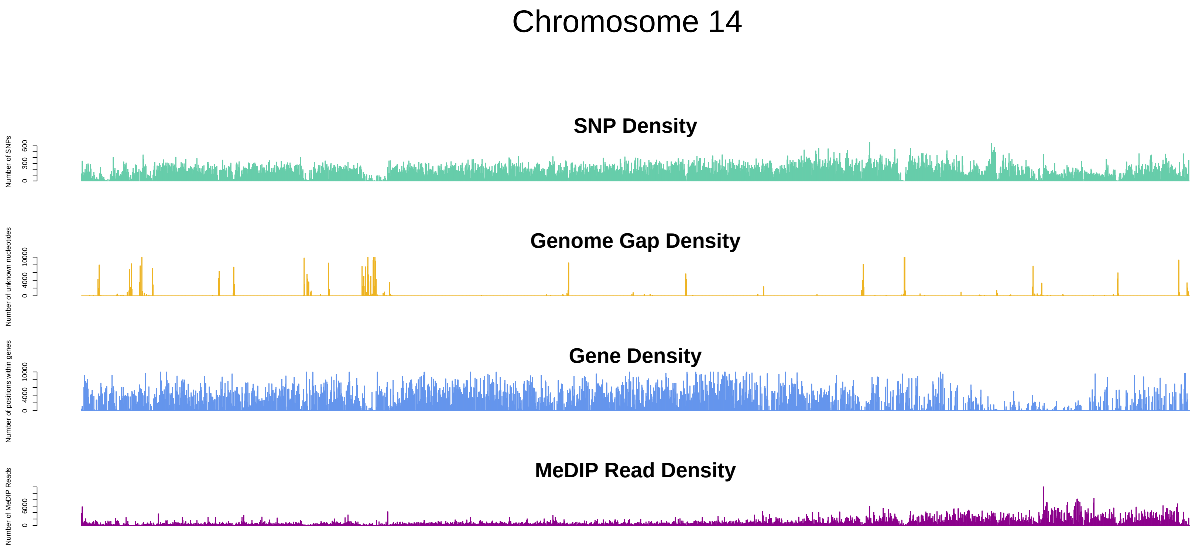


Figure S4.15: **Density profiles for chromosome 14.** Profile of SNP positions, genome gap positions, gene positions and MeDIP read positions in 10kb bins across *P. trichocarpa* chromosome 14.

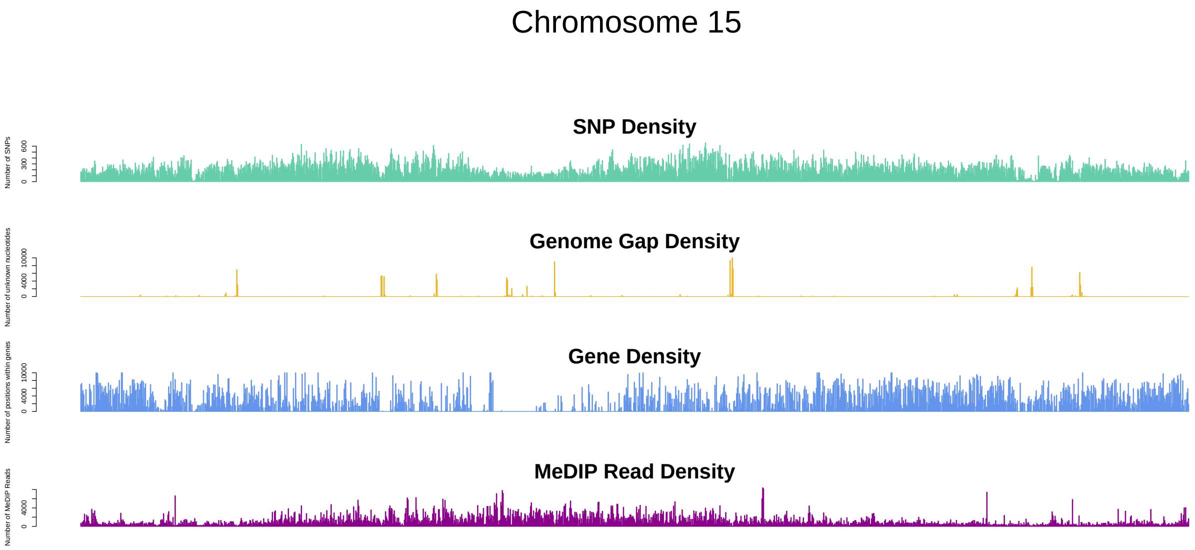


Figure S4.16: **Density profiles for chromosome 15.** Profile of SNP positions, genome gap positions, gene positions and MeDIP read positions in 10kb bins across *P. trichocarpa* chromosome 15.

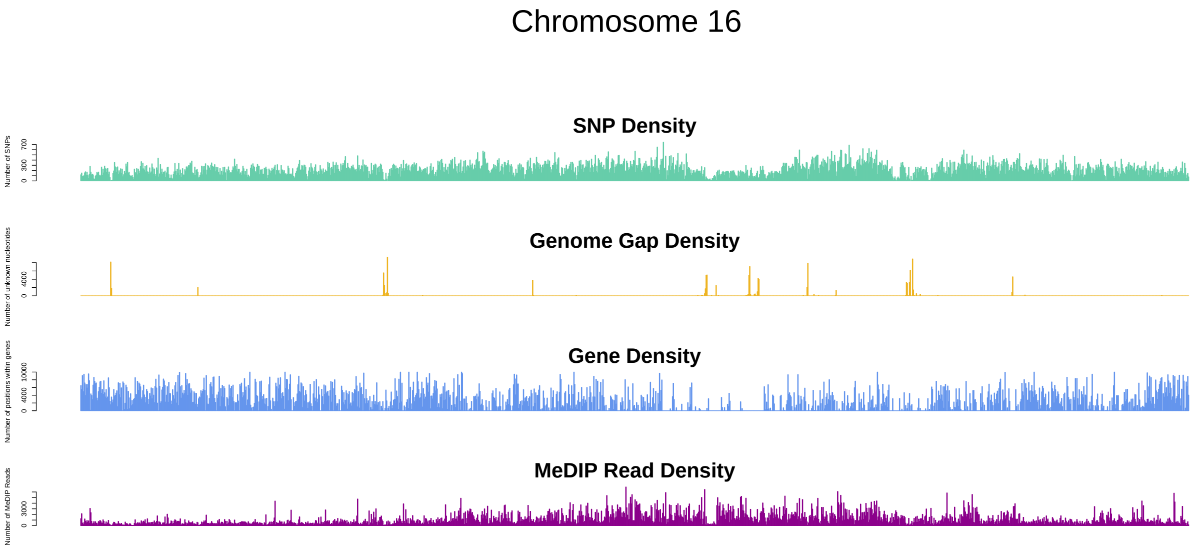


Figure S4.17: **Density profiles for chromosome 16.** Profile of SNP positions, genome gap positions, gene positions and MeDIP read positions in 10kb bins across *P. trichocarpa* chromosome 16.

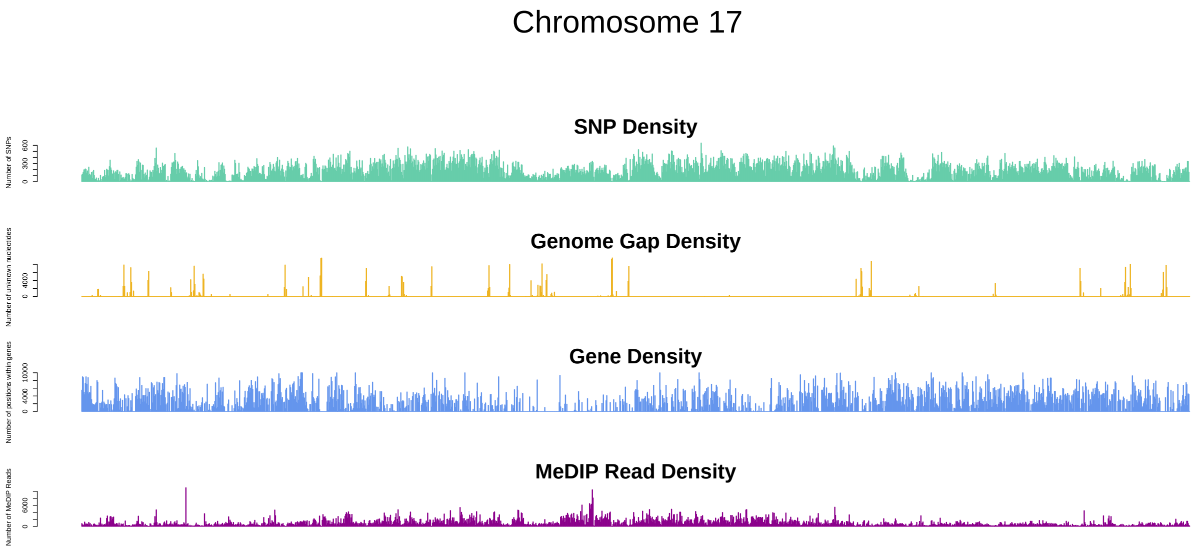


Figure S4.18: **Density profiles for chromosome 17.** Profile of SNP positions, genome gap positions, gene positions and MeDIP read positions in 10kb bins across *P. trichocarpa* chromosome 17.

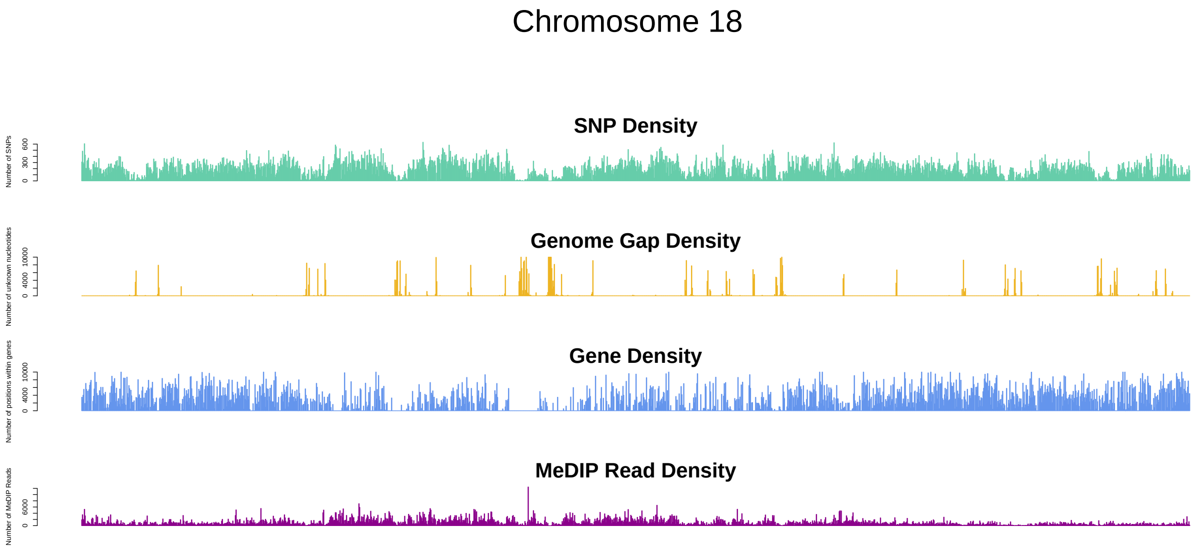


Figure S4.19: **Density profiles for chromosome 18.** Profile of SNP positions, genome gap positions, gene positions and MeDIP read positions in 10kb bins across *P. trichocarpa* chromosome 18.

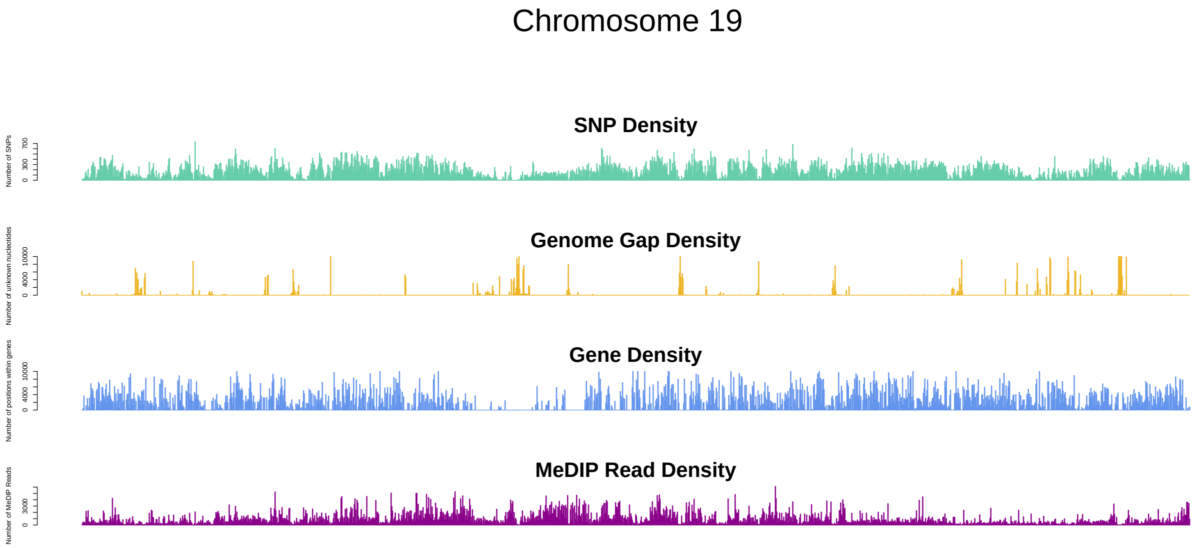


Figure S4.20: **Density profiles for chromosome 19.** Profile of SNP positions, genome gap positions, gene positions and MeDIP read positions in 10kb bins across *P. trichocarpa* chromosome 19.

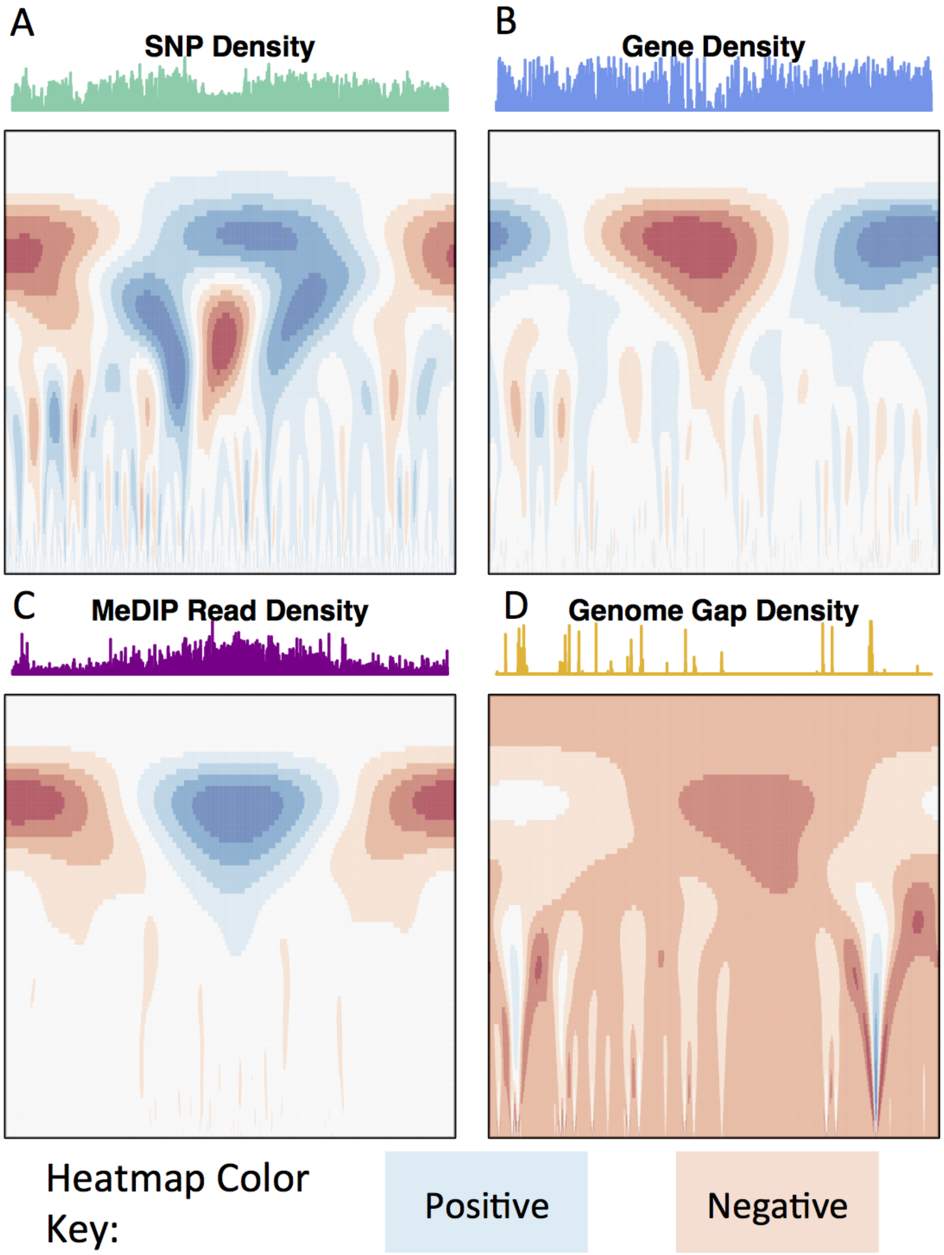


Figure S4.21: **CWT Coefficient landscapes of chromosome 12.** Coefficient landscapes for (A) SNP density, (B) gene density, (C) methylation (MeDIP-Seq read density, internode explant tissue) and (D) genome gap density on chromosome 12.

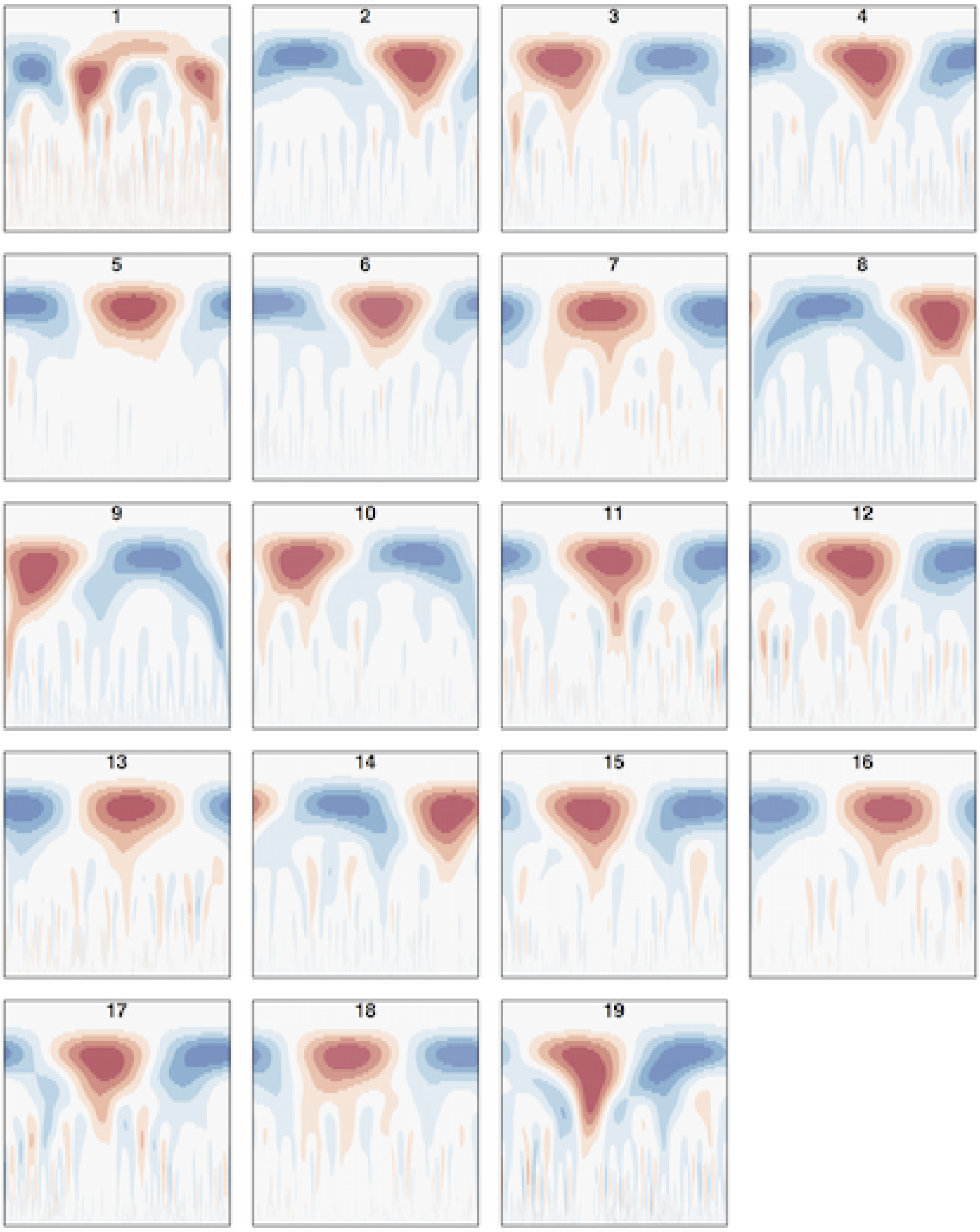


Figure S4.22: **Gene CWT.** Wavelet coefficient landscape for gene density.

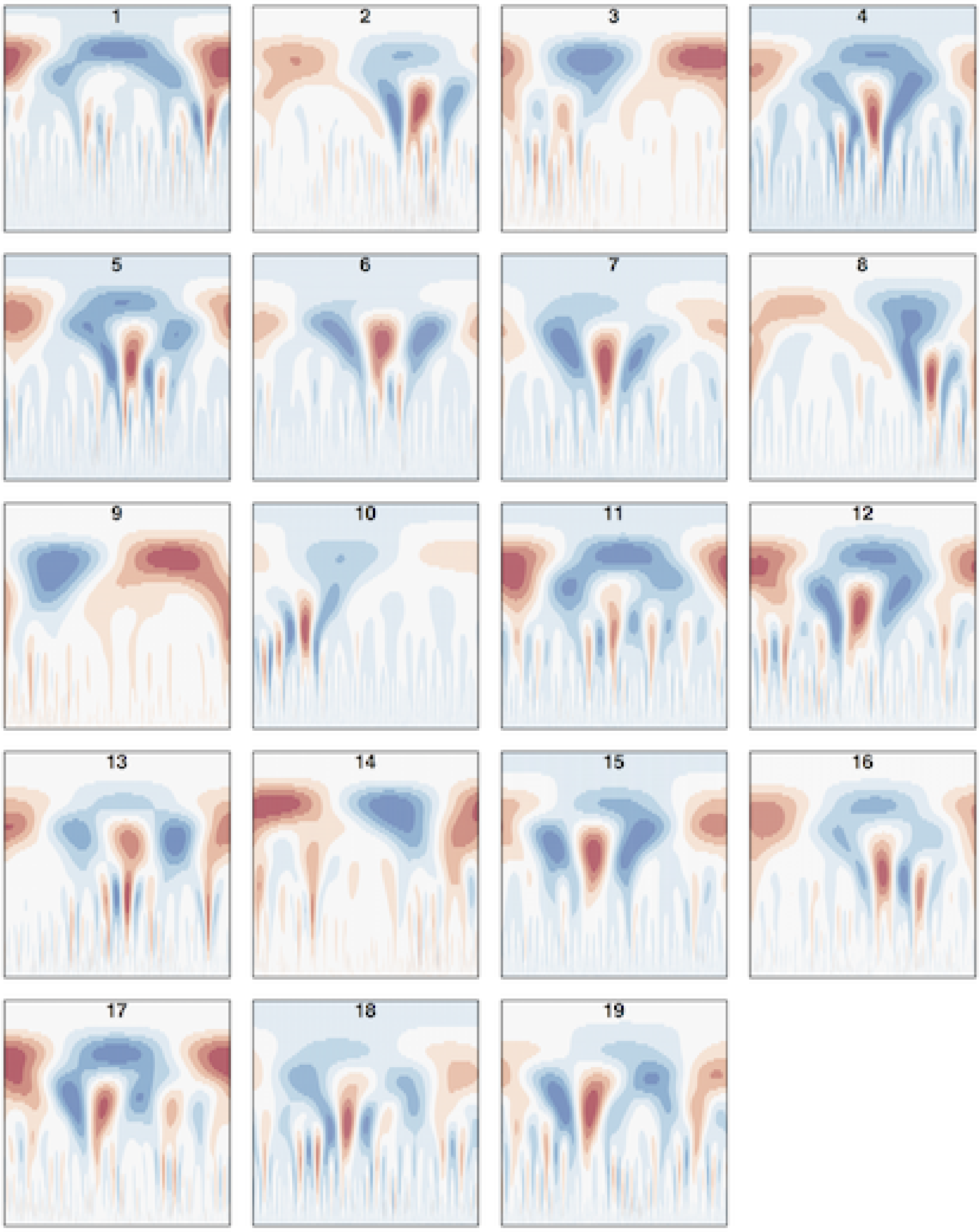


Figure S4.23: **SNP CWT.** Wavelet coefficient landscape for SNP density.

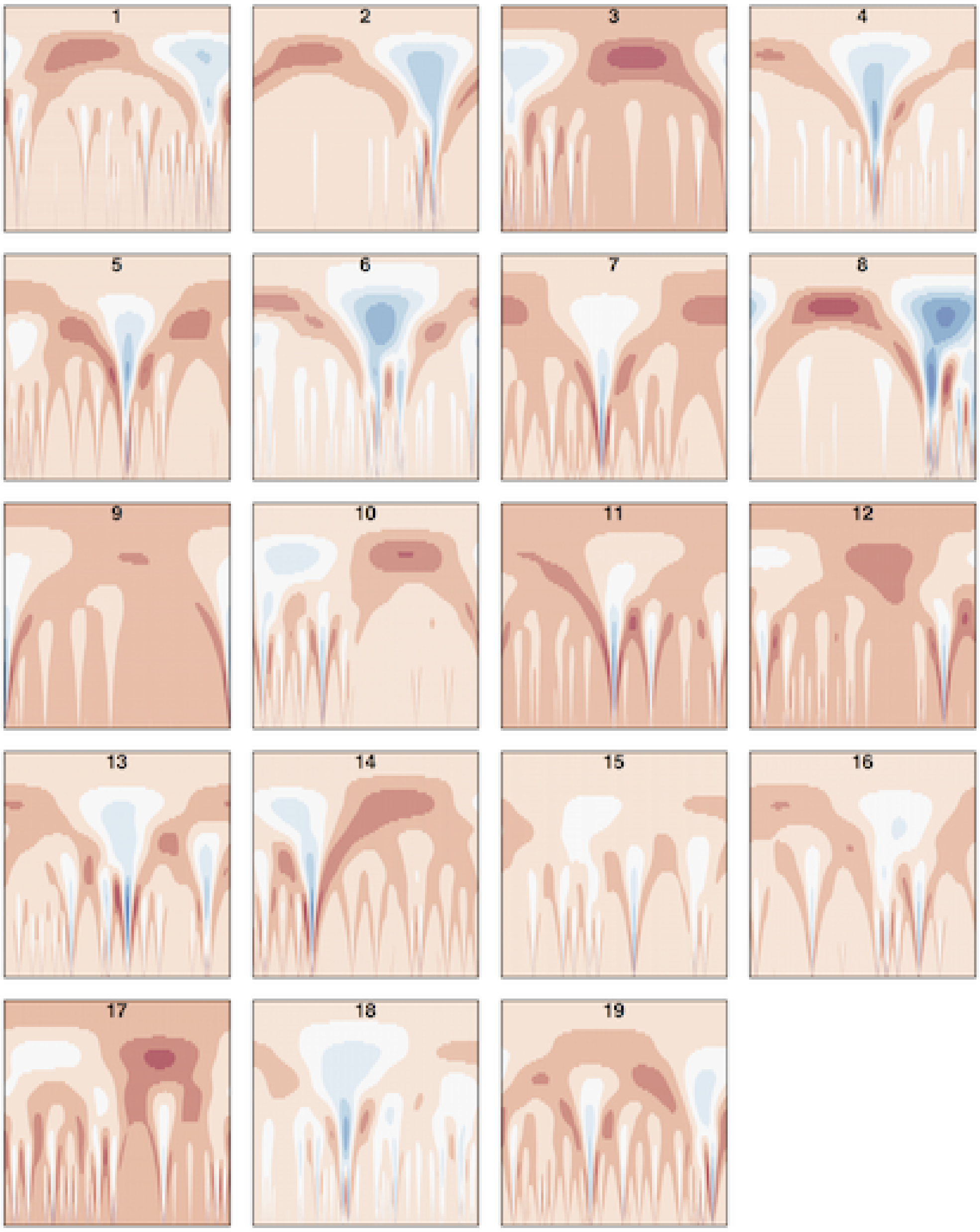


Figure S4.24: **Genome gap CWT.** Wavelet coefficient landscape for genome gap density.

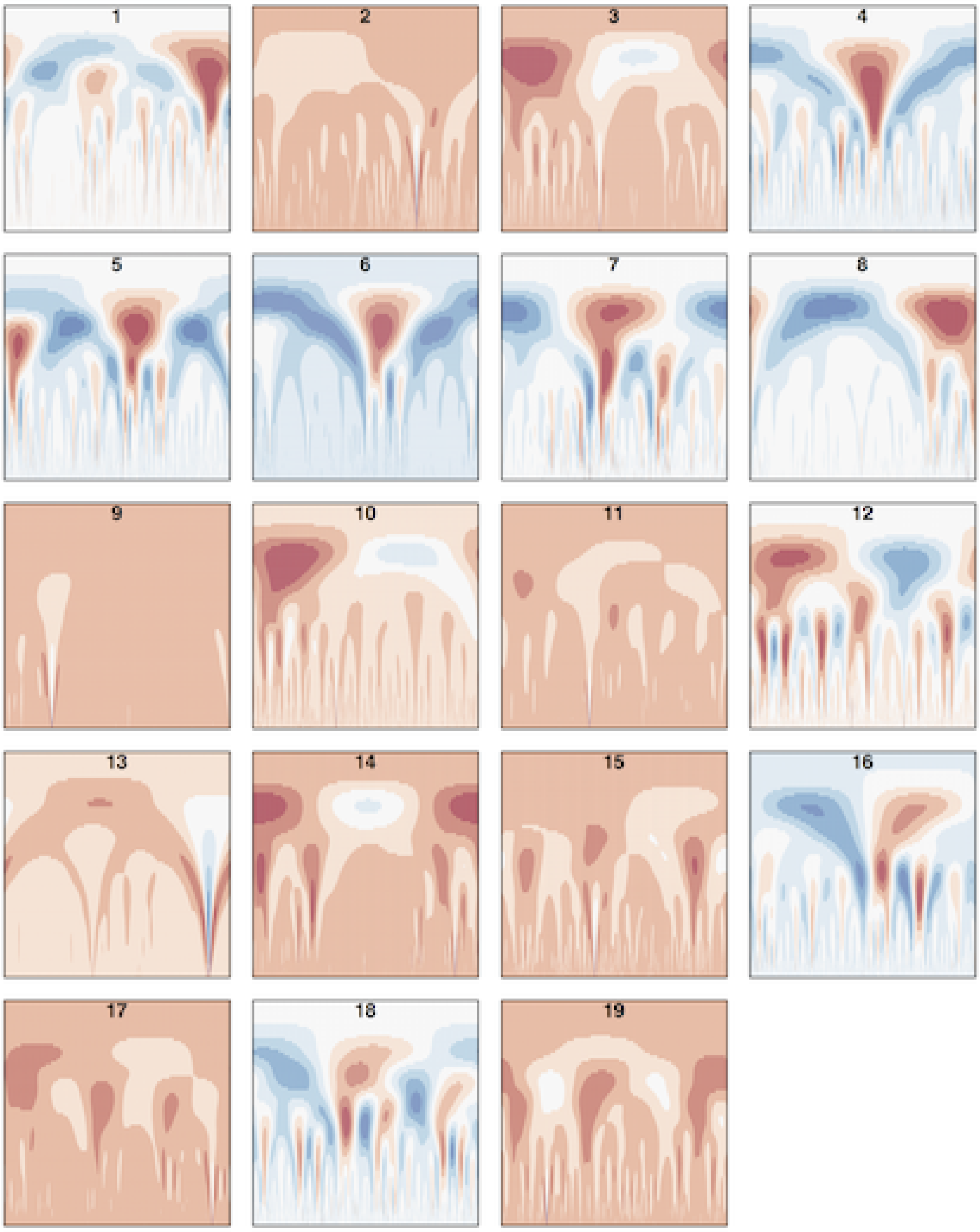


Figure S4.25: **Bud methylation CWT.** Wavelet coefficient methylation landscape for bud tissue.

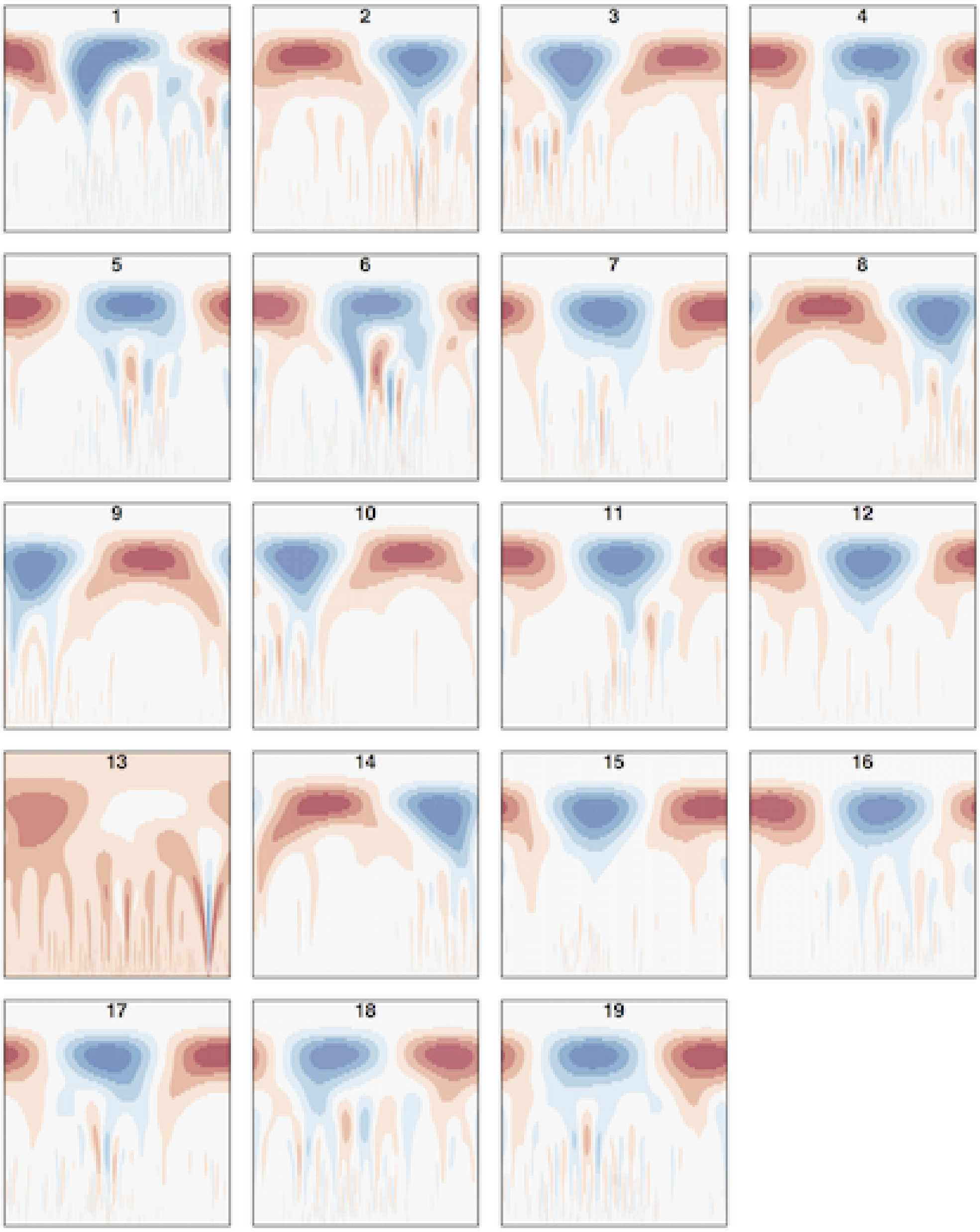


Figure S4.26: **Callus methylation CWT.** Wavelet coefficient methylation landscape for callus tissue.

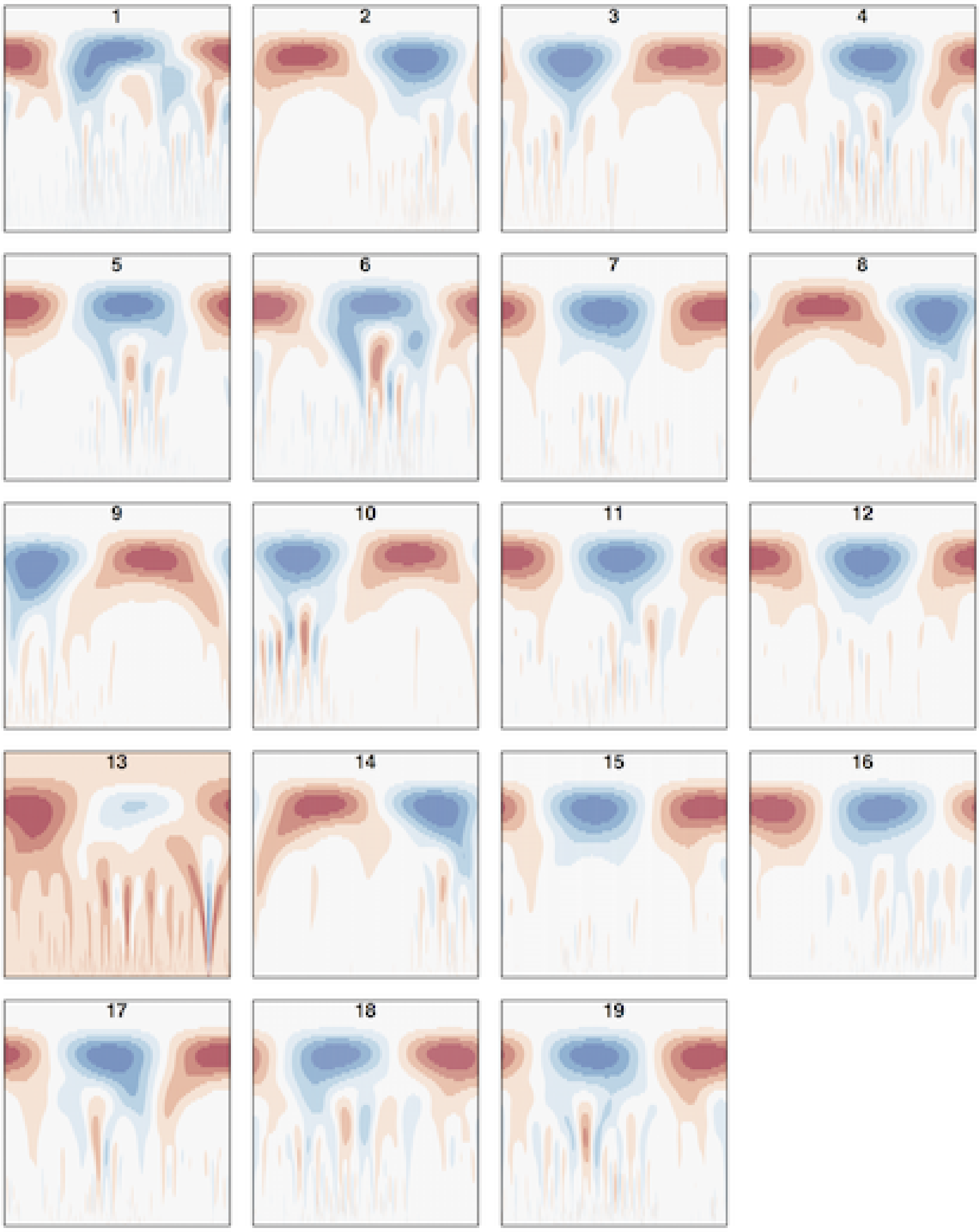


Figure S4.27: **Female catkin methylation CWT.** Wavelet coefficient methylation landscape for female catkin tissue.

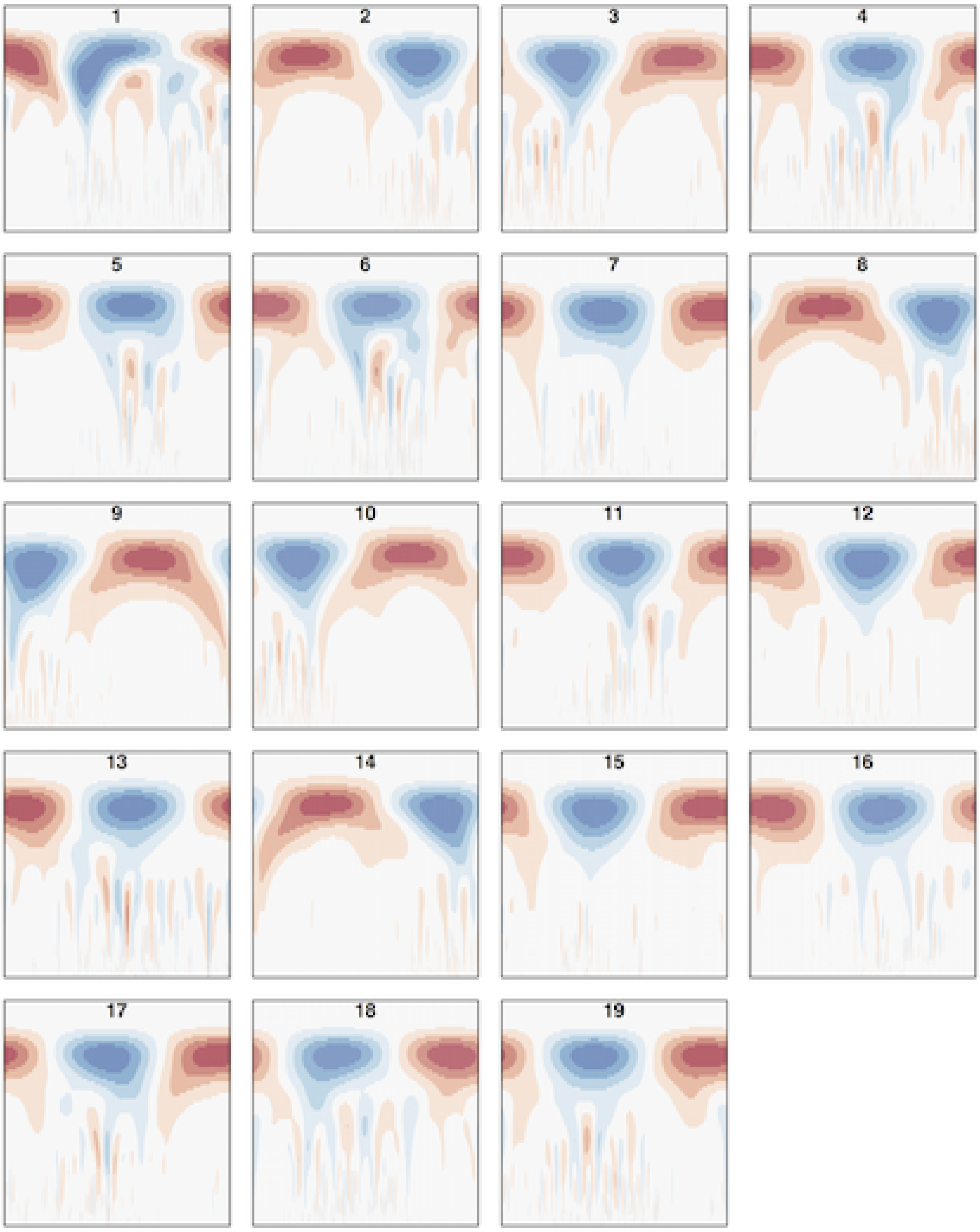


Figure S4.28: **internode explant methylation CWT.** Wavelet coefficient methylation landscape for internode explant tissue.

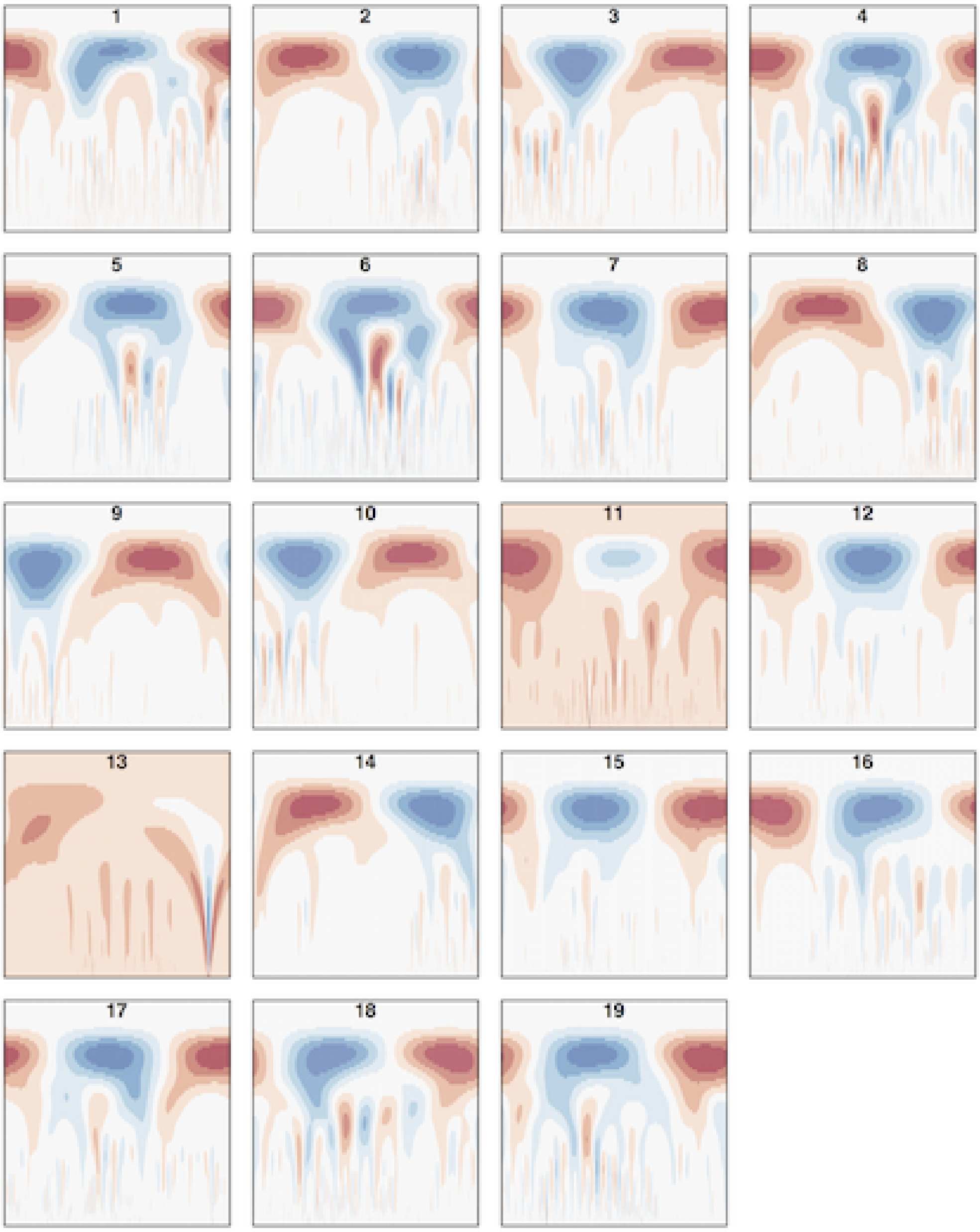


Figure S4.29: **Leaf methylation CWT.** Wavelet coefficient methylation landscape for leaf tissue.

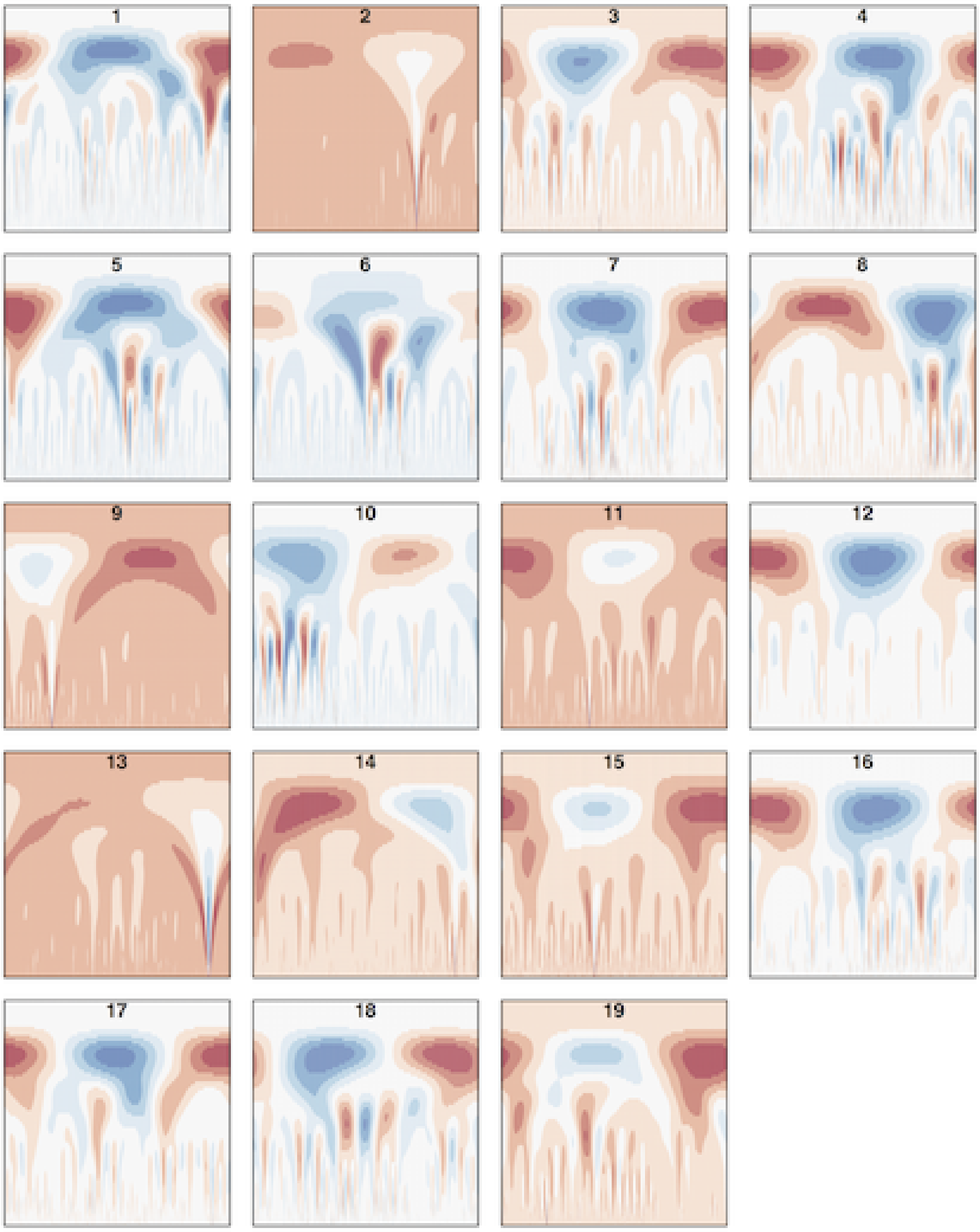


Figure S4.30: **Male catkin methylation CWT.** Wavelet coefficient methylation landscape for male catkin tissue.

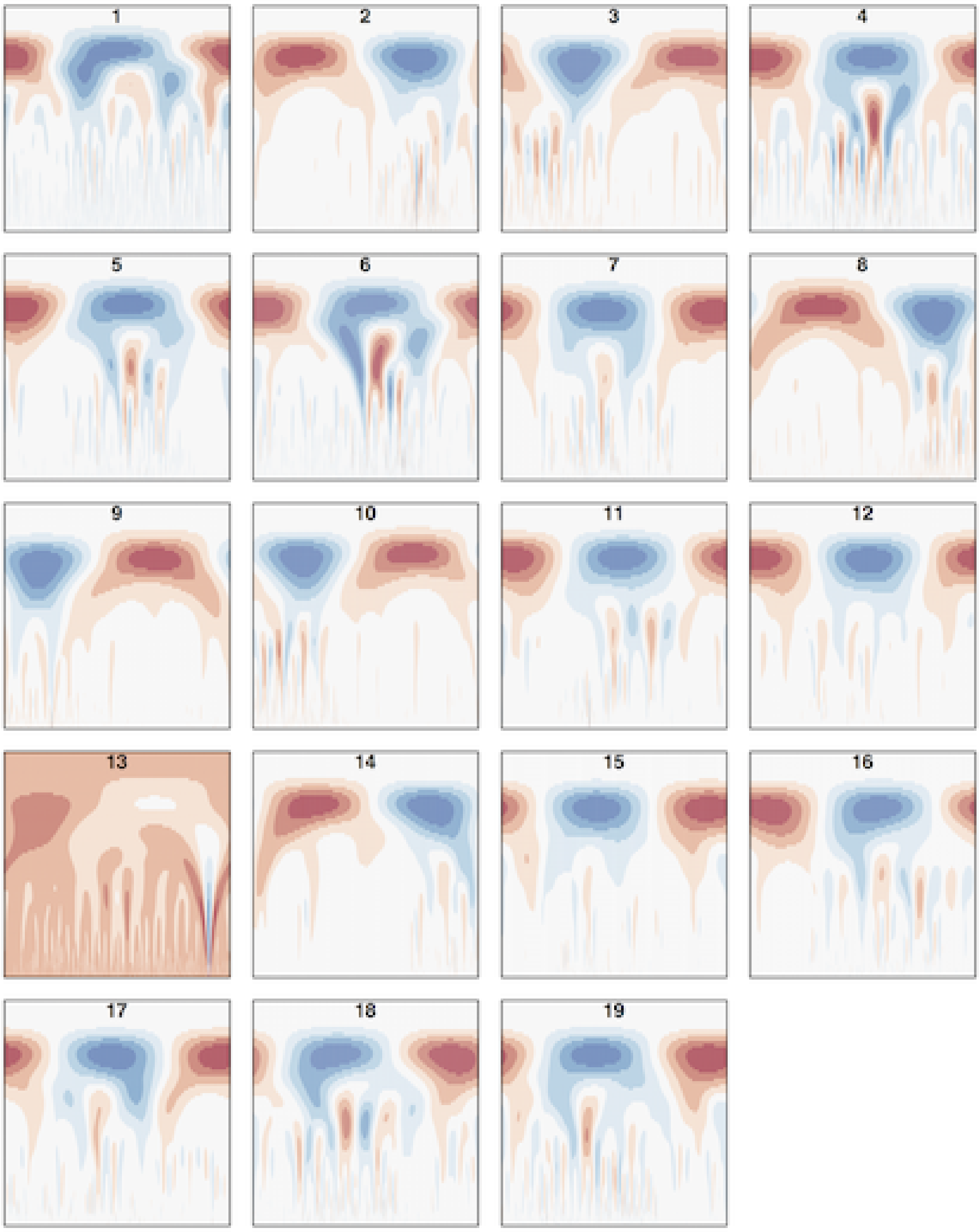


Figure S4.31: **Phloem methylation CWT.** Wavelet coefficient methylation landscape for phloem tissue.

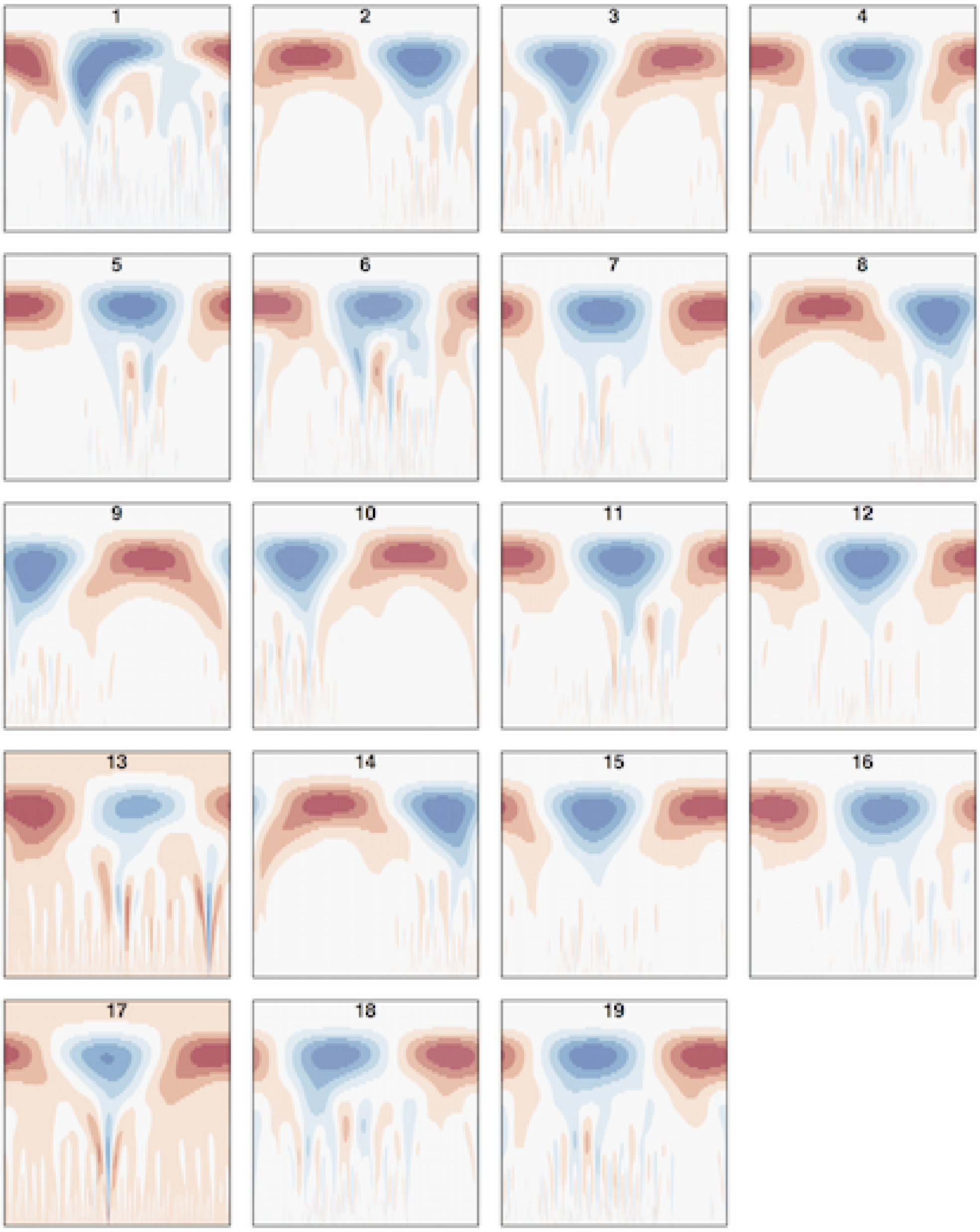


Figure S4.32: **Regenerated internode methylation CWT.** Wavelet coefficient methylation landscape for regenerated internode tissue.

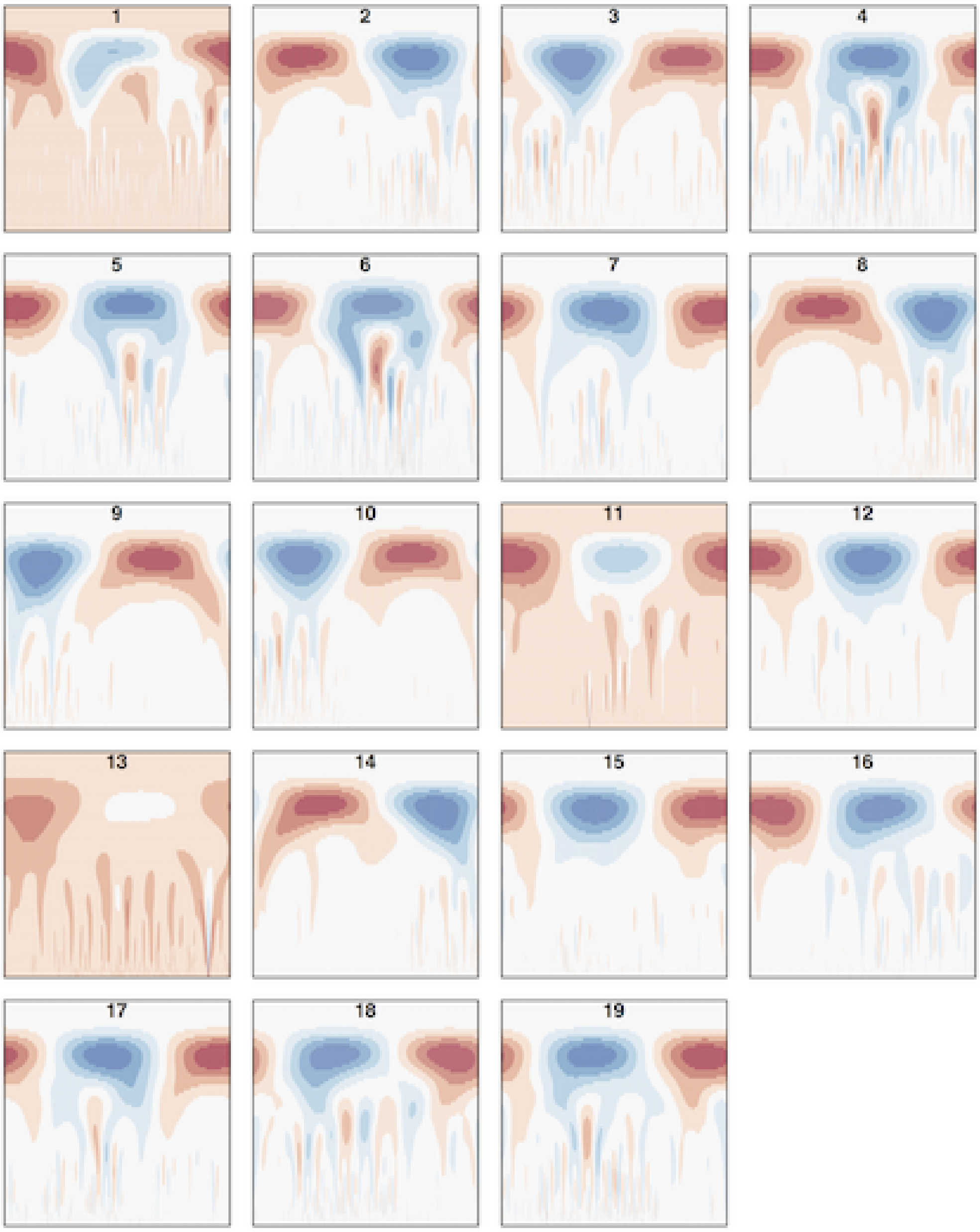


Figure S4.33: **Root methylation CWT.** Wavelet coefficient methylation landscape for root tissue.

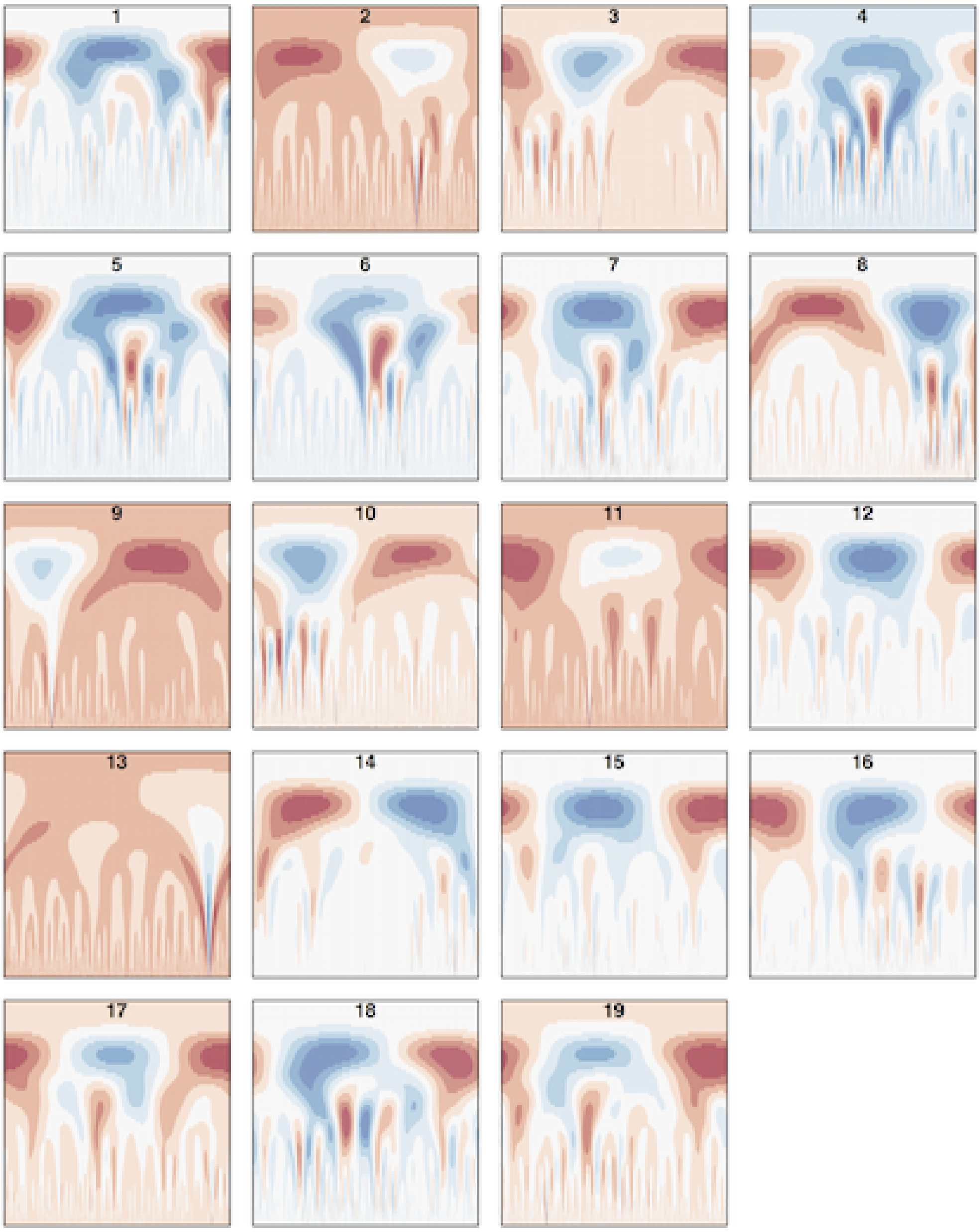


Figure S4.34: **Xylem methylation CWT.** Wavelet coefficient methylation landscape for xylem tissue.

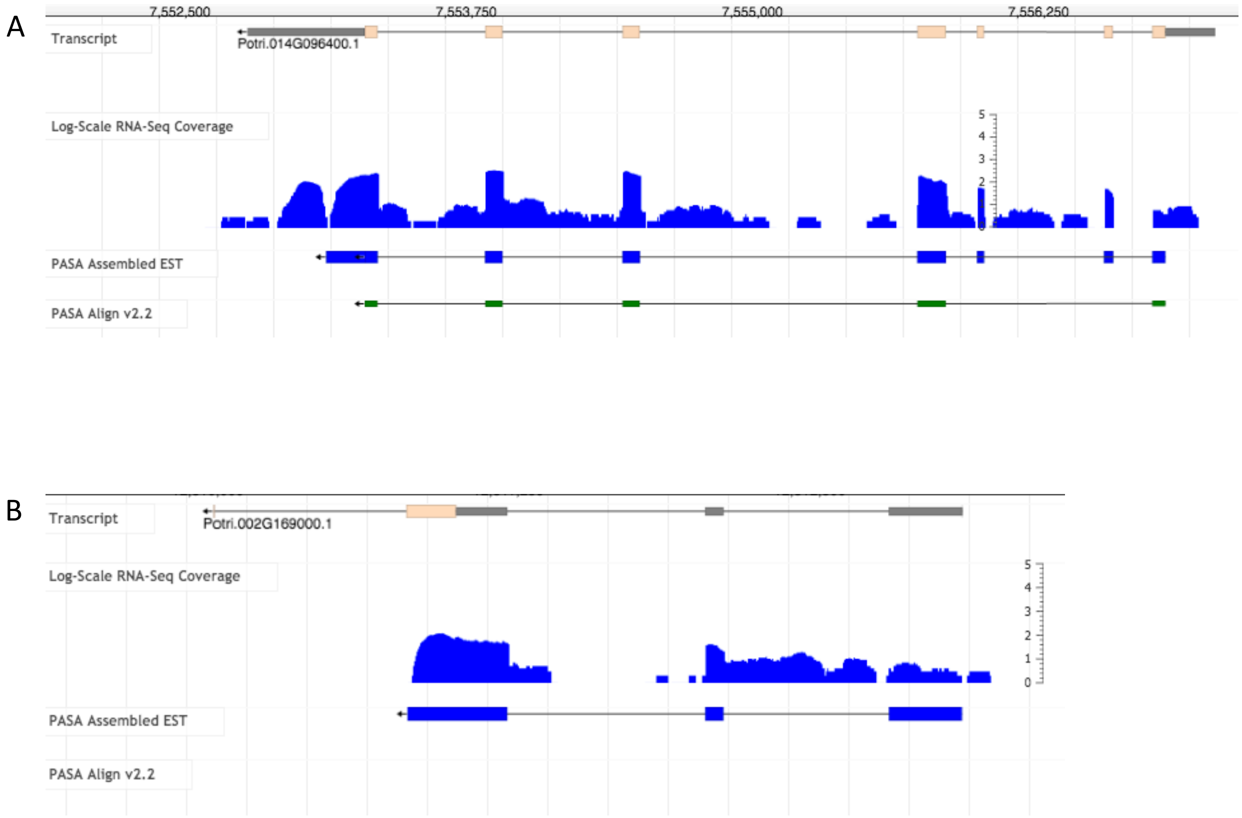


Figure S4.35: **CENH3 candidate genes expression evidence.** Two CENH3 candidates in *P. trichocarpa* (A) Potri.014G096400 and (B) Potri.002G169000 both show evidence of expression in the RNA-seq and EST coverage. Figure obtained using the Jbrowse plugin [59] on Phytozome [7].

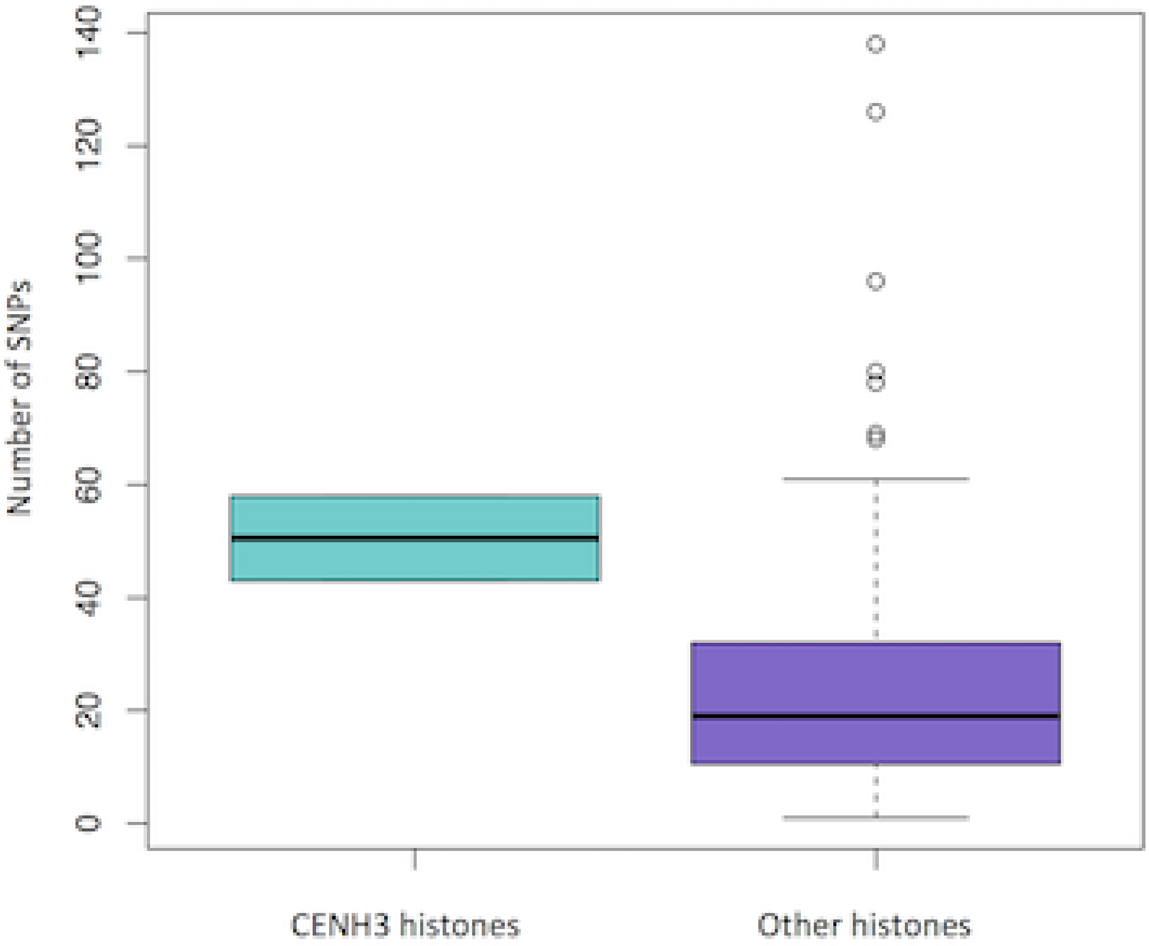


Figure S4.36: **SNPs in *P. trichocarpa* histones.** Boxplot showing the number of SNPs in the two candidate CENH3 histones versus other histones in *P. trichocarpa*.

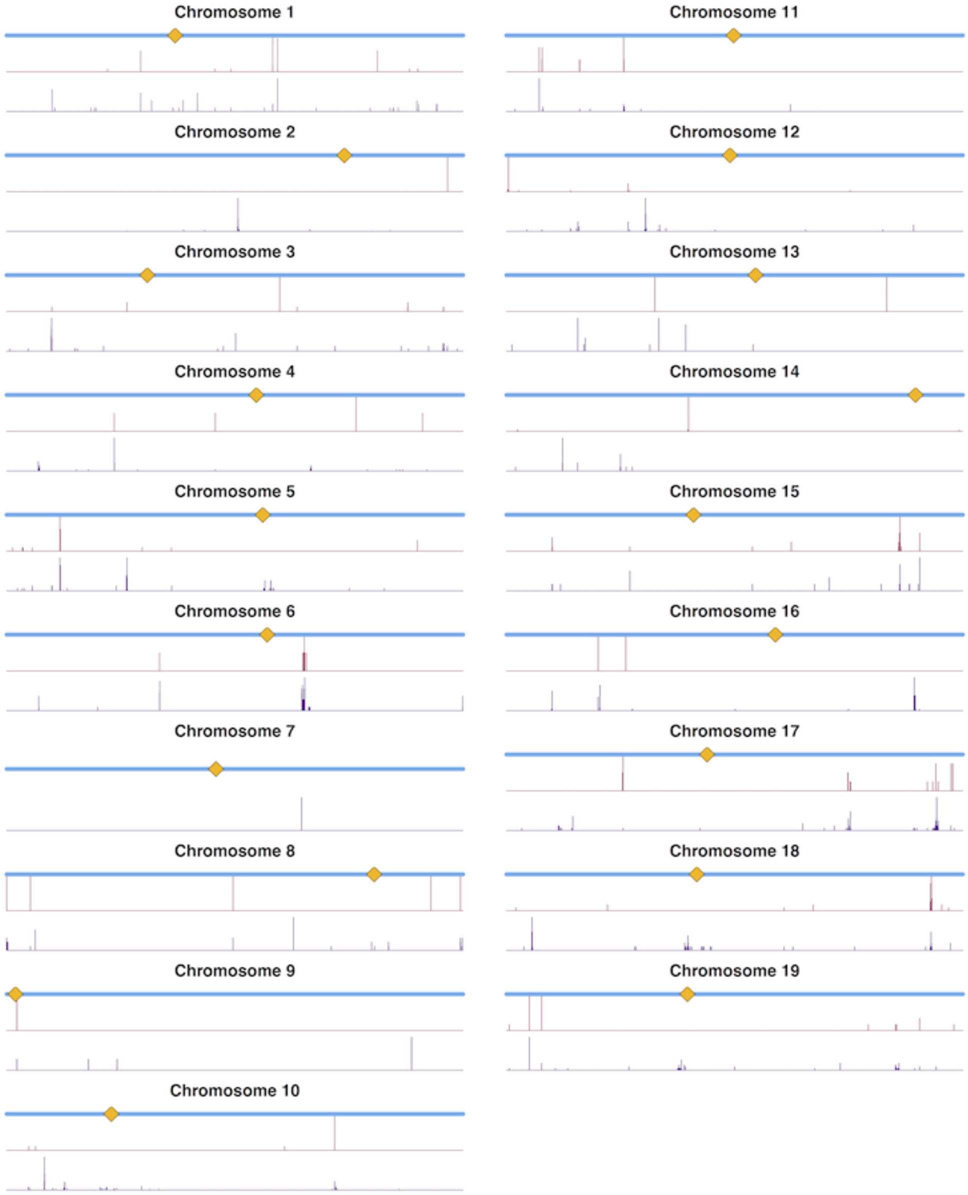


Figure S4.37: **CENH3 gene correlations.** Centromere positions (yellow diamonds) determined from wavelet coefficients and the density of SNPs correlating with SNPs in *P. trichocarpa* CENH3 genes. Red tracks are SNPs which correlate with SNPs in Potri.014G096400, and purple tracks are SNPs which correlate with SNPs in Potri.002G169000.

4.4.3 Supplementary Tables

Table S4.1: Positions in density signals of the approximate centromere locations, as well as the left and right borders indicated in Figure 8. Bp ranges indicate the borders of the particular bin in the density signal.

Chrom	Left border bin (bp)	Center bin (bp)	Right border bin (bp)
1	12,550,001-12,560,000	18,640,001-18,650,000	24,930,001-24,940,000
2	16,680,001-16,690,000	18,690,001-18,700,000	21,170,001-21,180,000
3	4,380,001-4,390,000	6,720,001-6,730,000	7,300,001-7,310,000
4	11,770,001-11,780,000	13,270,001-13,280,000	14,710,001-14,720,000
5	12,960,001-12,970,000	14,530,001-14,540,000	16,230,001-16,240,000
6	12,510,001-12,520,000	15,930,001-15,940,000	19,270,001-19,280,000
7	5,620,001-5,630,000	7,160,001-7,170,000	8,620,001-8,630,000
8	14,670,001-14,680,000	15,670,001-15,680,000	16,790,001-16,800,000
9	1-10,000	250,001-260,000	1,610,001-1,620,000
10	4,220,001-4,230,000	5,180,001-5,190,000	6,220,001-6,230,000
11	8,390,001-8,400,000	9,210,001-9,220,000	10,770,001-10,780,000
12	6,370,001-6,380,000	7,720,001-7,730,000	9,320,001-9,330,000
13	8,400,001-8,410,000	8,910,001-8,920,000	9,500,001-9,510,000
14	14,290,001-14,300,000	16,960,001-16,970,000	18,920,001-18,930,000
15	4,480,001-4,490,000	6,260,001-6,270,000	7,980,001-7,990,000
16	7,470,001-7,480,000	8,540,001-8,550,000	9,470,001-9,480,000
17	5,700,001-5,710,000	7,070,001-7,080,000	8,980,001-8,990,000
18	6,130,001-6,140,000	7,070,001-7,080,000	8,110,001-8,120,000
19	4,970,001-4,980,000	6,320,001-6,330,000	10,440,001-10,450,000

Table S4.2: Number of SNPs in *P. trichocarpa* histone genes.

See attached excel sheet

Table S4.3: Genes co-expressing with Potri.014G096400 on PhytoMine [44] from Phytozome [7]. Blue highlighted genes are histone-related.

See attached excel sheet

Bibliography

[1] David Gomez-Cabrero, Imad Abugessaisa, Dieter Maier, Andrew Teschendorff, Matthias Merkenschlager, Andreas Gisel, Esteban Ballestar, Erik Bongcam-Rudloff, Ana Conesa, and Jesper Tegnér. Data integration in the era of omics: current and future challenges. *BMC Systems Biology*, 8(2):I1, Mar 2014. 205

[2] Gerald A Tuskan, S Difazio, Stefan Jansson, J Bohlmann, I Grigoriev, U Hellsten, N Putnam, S Ralph, Stephane Rombauts, A Salamov, et al. The Genome of Black Cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science*, 313(5793):1596–1604, 2006. 205, 208, 210, 213, 223

[3] Gerald Tuskan, Gancho Slavov, Steve DiFazio, Wellington Muchero, Ranjan Pryia, Wendy Schackwitz, Joel Martin, Daniel Rokhsar, Robert Sykes, Mark Davis, et al. *Populus* resequencing: towards genome-wide association studies. In *BMC Proceedings*, volume 5, page I21. BioMed Central Ltd, 2011. 205, 208

[4] Gancho T Slavov, Stephen P DiFazio, Joel Martin, Wendy Schackwitz, Wellington Muchero, Eli Rodgers-Melnick, Mindie F Lipphardt, Christa P Pennacchio, Uffe Hellsten, Len A Pennacchio, et al. Genome resequencing reveals multiscale geographic structure and extensive linkage disequilibrium in the forest tree *Populus trichocarpa*. *New Phytologist*, 196(3):713–725, 2012. 205, 206

[5] Luke M Evans, Gancho T Slavov, Eli Rodgers-Melnick, Joel Martin, Priya Ranjan, Wellington Muchero, Amy M Brunner, Wendy Schackwitz, Lee Gunter, Jin-Gui Chen, et al. Population genomics of *Populus trichocarpa* identifies signatures of selection and adaptive trait associations. *Nature Genetics*, 46(10):1089–1096, 2014. 205

[6] Kelly J Vining, Kyle R Pomraning, Larry J Wilhelm, Henry D Priest, Matteo Pellegrini, Todd C Mockler, Michael Freitag, and Steven H Strauss. Dynamic DNA cytosine methylation in the *Populus trichocarpa* genome: tissue-level variation and relationship to gene expression. *BMC Genomics*, 13(1):1, 2012. 205, 206, 209, 214, 216

[7] David M Goodstein, Shengqiang Shu, Russell Howson, Rochak Neupane, Richard D Hayes, Joni Fazo, Therese Mitros, William Dirks, Uffe Hellsten, Nicholas Putnam, et al. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Research*, 40(D1):D1178–D1186, 2012. xiv, xxvi, xxvii, xxx, 205, 209, 210, 212, 213, 225, 226, 229, 258, 262

[8] C O, Connor. Chromosome segregation in mitosis: the role of centromeres. *Nature Education*, 1(1):28, 2008. 205

[9] Chao Feng, YaLin Liu, HanDong Su, HeFei Wang, James Birchler, and FangPu Han. Recent advances in plant centromere biology. *Science China Life Sciences*, 58(3):240–245, 2015. 205, 206

[10] Paul B Talbert, Ricardo Masuelli, Anand P Tyagi, Luca Comai, and Steven Henikoff. Centromeric localization and adaptive evolution of an arabidopsis histone h3 variant. *The Plant Cell*, 14(5):1053–1066, 2002. 206, 223

[11] Jennifer L Cooper and Steven Henikoff. Adaptive evolution of the histone fold domain in centromeric histones. *Molecular biology and evolution*, 21(9):1712–1718, 2004. 206, 223

[12] Steven Henikoff, Kami Ahmad, and Harmit S Malik. The centromere paradox: stable inheritance with rapidly evolving dna. *Science*, 293(5532):1098–1102, 2001. 206, 223

[13] Suzanne Furuyama and Sue Biggins. Centromere identity is specified by a single centromeric nucleosome in budding yeast. *Proceedings of the National Academy of Sciences*, 104(37):14706–14711, 2007. 206

[14] Shweta Mehrotra and Vinod Goyal. Repetitive sequences in plant nuclear dna: types, distribution, evolution and function. *Genomics, proteomics & bioinformatics*, 12(4):164–171, 2014. 206

[15] Xiaoyu Zhang, Junshi Yazaki, Ambika Sundaresan, Shawn Cokus, Simon W-L Chan, Huaming Chen, Ian R Henderson, Paul Shinn, Matteo Pellegrini, Steve E Jacobsen, et al. Genome-wide high-resolution mapping and functional analysis of dna methylation in arabidopsis. *Cell*, 126(6):1189–1201, 2006. 206, 216

[16] Rosa Maria Cossu, Matteo Buti, Tommaso Giordani, Lucia Natali, and Andrea Cavallini. A computational study of the dynamics of ltr retrotransposons in the populus trichocarpa genome. *Tree Genetics & Genomes*, 8(1):61–75, 2012. xxvi, 206, 212, 214, 221, 224

[17] Sara Pinosio, Stefania Giacomello, Patricia Faivre-Rampant, Gail Taylor, Veronique Jorge, Marie Christine Le Paslier, Giusi Zaina, Catherine Bastien, Federica Cattonaro, Fabio Marroni, et al. Characterization of the poplar pan-genome by genome-wide identification of structural variation. *Molecular biology and evolution*, 33(10):2706–2719, 2016. 206

[18] Chris CA Spencer, Panos Deloukas, Sarah Hunt, Jim Mullikin, Simon Myers, Bernard Silverman, Peter Donnelly, David Bentley, and Gil McVean. The Influence of Recombination on Human Genetic Diversity. *PLoS Genet*, 2(9):e148, 2006. 206, 208, 214

[19] Ryan F McCormick, Sandra K Truong, Avinash Sreedasyam, Jerry Jenkins, Shengqiang Shu, David Sims, Megan Kennedy, Mojgan Amirebrahimi, Brock Weers, Brian McKinley, et al. The sorghum bicolor reference genome: improved assembly and annotations, a transcriptome atlas, and signatures of genome organization. *bioRxiv*, page 110593, 2017. 206

[20] CM Leavey, MN James, J Summerscales, and R Sutton. An introduction to wavelet transforms: a tutorial approach. *Insight-Non-Destructive Testing and Condition Monitoring*, 45(5):344–353, 2003. xxiv, 207

[21] JA Tenreiro Machado, António C Costa, and Maria Dulce Quelhas. Wavelet analysis of human dna. *Genomics*, 98(3):155–163, 2011. xxiv, 207

[22] Qiang Wu and Kenneth R Castleman. Automated chromosome classification using wavelet-based band pattern descriptors. In *Computer-Based Medical Systems, 2000. CBMS 2000. Proceedings. 13th IEEE Symposium on*, pages 189–194. IEEE, 2000. 208

[23] Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A Albers, Eric Banks, Mark A DePristo, Robert E Handsaker, Gerton Lunter, Gabor T Marth, Stephen T Sherry, et al. The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158, 2011. 208

[24] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575, 2007. 208

[25] Sharlee Climer, Wei Yang, Lisa Fuentes, Victor G Dávila-Román, and C Charles Gu. A Custom Correlation Coefficient (CCC) Approach for Fast Identification of Multi-SNP Association Patterns in Genome-Wide SNPs Data. *Genetic Epidemiology*, 38(7):610–621, 2014. 208, 209, 225

[26] Sharlee Climer, Alan R Templeton, and Weixiong Zhang. Allele-Specific Network Reveals Combinatorial Interaction that Transcends Small Effects in Psoriasis GWAS. *PLoS Comput Biol*, 10(9):e1003766, 2014. 208, 209, 225

[27] Wayne Joubert, James Nance, Sharlee Climer, Deborah Weighill, and Daniel Jacobson. Parallel accelerated custom correlation coefficient calculations for genomics applications. *arXiv preprint arXiv:1705.08213*, 2017. 209, 225

[28] Deborah A Weighill, Piet Jones, Manesh Shah, Priya Ranjan, Wellington Muchero, Jeremy Schmutz, Avinash Sreedasyam, David Macaya-Sanz, Robert Sykes, Nan Zhao, et al. Pleiotropic and epistatic network-based discovery: Integrated networks for target gene discovery. *Frontiers in Energy Research*, 2018. 209, 212

[29] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, et al. The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009. 210

[30] Derek W Barnett, Erik K Garrison, Aaron R Quinlan, Michael P Strömberg, and Gabor T Marth. BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics*, 27(12):1691–1692, 2011. 210

[31] Aaron R Quinlan. Bedtools: the swiss-army tool for genome feature analysis. *Current protocols in bioinformatics*, pages 11–12, 2014. 210

[32] Shane Neph, M Scott Kuehn, Alex P Reynolds, Eric Haugen, Robert E Thurman, Audra K Johnson, Eric Rynes, Matthew T Maurano, Jeff Vierstra, Sean Thomas, et al. Bedops: high-performance genomic feature operations. *Bioinformatics*, 28(14):1919–1920, 2012. 210

[33] Ole Tange et al. Gnu parallel-the command-line power tool. *The USENIX Magazine*, 36(1):42–47, 2011. 210

[34] Igor V Grigoriev, Henrik Nordberg, Igor Shabalov, Andrea Aerts, Mike Cantor, David Goodstein, Alan Kuo, Simon Minovitsky, Roman Nikitin, Robin A Ohm, et al. The Genome Portal of the Department of Energy Joint Genome Institute. *Nucleic Acids Research*, pages 1–7, 2011. 210

[35] Henrik Nordberg, Michael Cantor, Serge Dusheyko, Susan Hua, Alexander Poliakov, Igor Shabalov, Tatyana Smirnova, Igor V Grigoriev, and Inna Dubchak. The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. *Nucleic Acids Research*, 42(D1):D26–D31, 2014. 210

[36] William Constantine and Donald Percival. *wmtsa: Wavelet Methods for Time Series Analysis*, 2016. R package version 2.0-2. 211

[37] Donald B Percival and Andrew T Walden. *Wavelet methods for time series analysis*, volume 4. Cambridge university press, 2006. 211

[38] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. 211

[39] RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA, 2016. 211

[40] William Constantine, Tim Hesterberg, Knut Wittkowski, Tingting Song, and Stephen Kaluzny. *splus2R: Supplemental S-PLUS Functionality in R*, 2016. R package version 1.2-2. 211

[41] Erich Neuwirth. *RColorBrewer: ColorBrewer Palettes*, 2014. R package version 1.1-2. 211

[42] Thomas Nussbaumer, Mihaela M Martis, Stephan K Roessner, Matthias Pfeifer, Kai C Bader, Sapna Sharma, Heidrun Gundlach, and Manuel Spannagl. Mips plantsdb: a database framework for comparative plant genome research. *Nucleic acids research*, 41(D1):D1144–D1151, 2012. 212, 221

[43] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990. 212, 225

[44] Alex Kalderimis, Rachel Lyne, Daniela Butano, Sergio Contrino, Mike Lyne, Joshua Heimbach, Fengyuan Hu, Richard Smith, Radek Štěpán, Julie Sullivan, et al. Intermine: extensive web services for modern biology. *Nucleic acids research*, 42(W1):W468–W472, 2014. xiv, 212, 225, 226, 262

[45] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, 13(11):2498–2504, 2003. 212

[46] Aron Marchler-Bauer and Stephen H Bryant. Cd-search: protein domain annotations on the fly. *Nucleic acids research*, 32(suppl_2):W327–W331, 2004. 213

[47] Aron Marchler-Bauer, Shennan Lu, John B Anderson, Farideh Chitsaz, Myra K Derbyshire, Carol DeWeese-Scott, Jessica H Fong, Lewis Y Geer, Renata C Geer, Noreen R Gonzales, et al. Cdd: a conserved domain database for the functional annotation of proteins. *Nucleic acids research*, 39(suppl_1):D225–D229, 2010. 213

[48] Aron Marchler-Bauer, Myra K Derbyshire, Noreen R Gonzales, Shennan Lu, Farideh Chitsaz, Lewis Y Geer, Renata C Geer, Jane He, Marc Gwadz, David I Hurwitz, et al. Cdd: Ncbi's conserved domain database. *Nucleic acids research*, 43(D1):D222–D226, 2014. 213

[49] Gregory P Copenhaver, Kathryn Nickel, Takashi Kuromori, Maria-Ines Benito, Samir Kaul, Xiaoying Lin, Michael Bevan, George Murphy, Barbara Harris, Laurence D Parnell, et al. Genetic definition and sequence analysis of arabidopsis centromeres. *Science*, 286(5449):2468–2474, 1999. 216

[50] Dan Liang, Zhoujia Zhang, Honglong Wu, Chunyu Huang, Peng Shuai, Chu-Yu Ye, Sha Tang, Yunjie Wang, Ling Yang, Jun Wang, et al. Single-base-resolution methylomes of populus trichocarpa reveal the association between dna methylation and drought stress. *BMC genetics*, 15(1):S9, 2014. 216

[51] Wenli Zhang, Hye-Ran Lee, Dal-Hoe Koo, and Jiming Jiang. Epigenetic modification of centromeric chromatin: hypomethylation of dna sequences in the cenh3-associated chromatin in arabidopsis thaliana and maize. *The Plant Cell*, 20(1):25–34, 2008. 216

[52] Stephan Ossowski, Korbinian Schneeberger, José Ignacio Lucas-Lledó, Norman Warthmann, Richard M Clark, Ruth G Shaw, Detlef Weigel, and Michael Lynch. The rate and molecular spectrum of spontaneous mutations in arabidopsis thaliana. *science*, 327(5961):92–94, 2010. 219

[53] Wubishet A Bekele, Silke Wieckhorst, Wolfgang Friedt, and Rod J Snowdon. High-throughput genomics in sorghum: from whole-genome resequencing to a snp screening array. *Plant biotechnology journal*, 11(9):1112–1125, 2013. 219

[54] Anshul Watts, Vajinder Kumar, and Shripad Ramachandra Bhat. Centromeric histone h3 protein: from basic study to plant breeding applications. *Journal of Plant Biochemistry and Biotechnology*, 25(4):339–348, Oct 2016. 223, 225, 229

[55] Inna Lermontova, Olga Koroleva, Twan Rutten, Joerg Fuchs, Veit Schubert, Izabel Moraes, David Koszegi, and Ingo Schubert. Knockdown of cenh3 in arabidopsis reduces mitotic divisions and causes sterility by disturbed meiotic chromosome segregation. *The Plant Journal*, 68(1):40–50, 2011. 223

[56] Shamoni Maheshwari, Takayoshi Ishii, C Titus Brown, Andreas Houben, and Luca Comai. Centromere location in arabidopsis is unaltered by extreme divergence in cenh3 protein sequence. *Genome Research*, 27(3):471–478, 2017. 223

[57] Jing Yuan, Xiang Guo, Jing Hu, Zhenling Lv, and Fangpu Han. Characterization of two cenh3 genes and their roles in wheat evolution. *New Phytologist*, 206(2):839–851, 2015. 225

[58] Shamoni Maheshwari, Ek Han Tan, Allan West, F Chris H Franklin, Luca Comai, and Simon WL Chan. Naturally occurring differences in cenh3 affect chromosome segregation in zygotic mitosis of hybrids. *PLoS genetics*, 11(1):e1004970, 2015. 226

[59] Mitchell E Skinner, Andrew V Uzilov, Lincoln D Stein, Christopher J Mungall, and Ian H Holmes. JBrowse: A next-generation genome browser. *Genome Research*, 19(9):1630–1638, 2009. xxvii, xxx, 229, 258

Chapter 5

Synteny, Ancestral Centromeres and Repeats: Further Insights into Genome Organization and Evolutionary History

This chapter is currently in preparation for publication, and contains contributions from other authors:

Deborah Weighill, David Macaya-Sanz, Angelica M. Walker, Stephen DiFazio, Gerald Tuskan and Daniel Jacobson

Author Contributions: DW and DJ conceived of and designed the study. DW performed the analysis, interpreted the results and wrote the manuscript, GT lead the sequencing of *Populus* genotypes, SD and DMS performed the SNP calling. AW ran RepeatMasker. The Department of Energy Joint Genome Institute (JGI) performed the sequencing of *P. trichocarpa* genomes, and Anna Furches provided useful discussions regarding synteny and transposable elements.

Abstract

The *Populus trichocarpa* genome has an interesting and complex evolutionary history which involves several genome duplication events and rearrangements. The genome also consists of various genomic/epigenomic elements, including single nucleotide polymorphisms (SNPs), DNA methylation, transposable elements (TEs) and repeat sequences. We constructed density profiles of syntenic blocks, centromere positions, different classes of TEs, SNPs, genes and methylation across the genome, and make use of the discrete wavelet transform to to unpack the information in different scales of these signals. Correlation analysis identified various scale-specific relationships between genomic features, and provided resources for a method useful for the interrogation and comparison of multiple data types in *P. trichocarpa*.

5.1 Introduction

The *Populus trichocarpa* genome has an interesting and complex evolutionary history which involves several genome duplication events and rearrangements [1]. As reported by Tuskan *et al.* (2006), the structure of the current *P. trichocarpa* genome contains large homologous chromosome blocks, or syntenic blocks. These originated from a whole genome duplication event called the Salicoid duplication, and subsequent genome rearrangement [1].

In the previous chapter, we investigated chromosomal structure and centromere positions in *P. trichocarpa* though signal processing of genomic features such as SNP density, gene density and methylation density. In this chapter, we include two new genomic features in order to investigate the evolutionary history leading up to the current *P. trichocarpa* genome. We wish to repeat similar analyses as performed by [1] in order to determine which of the hypotheses are still valid in the new, improved genome assembly, version 3.0. We also include a new data type, repeat sequences and transposable elements (TEs), which were masked from the original genome analysis.

TEs are mobile DNA elements falling into two main classes. Class I TEs (retroelements) duplicate via an RNA intermediate in a "copy-and-paste" type mechanism, whereas Class II TEs move via a "cut and paste" mechanism where the DNA element is excised and integrated elsewhere in the genome [2, 3]. Different TEs have very different distributions throughout the genome, and are known to have a significant impact on various aspects of the genome including genome size, genome arrangement and centromere function.

We wish to interrogate the similarities and differences between the distributions of syntenic blocks, centromere positions, different classes of TEs, SNPs, genes and methylation across the genome, in order to generate hypotheses surrounding the evolutionary history of the structure of the genome. Such comparisons between signals requires them to be compared at multiple scales, as different driving forces can operate at different scales. For this reason, we make use of the wavelet transform, a signal processing technique which can be used to unpack the information in different scales of a signal, such as a density profile across a chromosome [4]. In general, the continuous wavelet transform (CWT) involves expressing a function (signal) as a linear combination of basis functions called wavelets. These basis functions are scaled translations of a mother wavelet, such as the Ricker Wavelet. What results from a wavelet transform is a set of wavelet coefficients $W(s, \tau)$ (Equation 5.1), a coefficient for every scale s and translation τ [5].

$$W(s, \tau) = \frac{1}{\sqrt{s}} \int f(t)\psi^* \left(\frac{t - \tau}{s}\right) dt \tag{5.1}$$

The Discrete Wavelet Transform (DWT) is a sampled version of the CWT, and involves sampling of the bp and scale dimensions [5]. The DWT produces a series of sets of coefficients with one set of coefficients for each scale computed (Figure 5.1).

While the CWT is ideal when one wants to view the entire coefficient landscape as we did in the previous chapter, it does contain redundant information. When one wants to calculate the correlation between two features at different scales, one does not need every scale. A sampling of the scales is more convenient. Thus, the DWT is ideal for this application.

The study by Slavov *et al.* (2012) [6] involved correlating various chromosome features such as recombination rate and methylation across chromosomes, and the study by McCormick *et. al* (2017) [7] involved analyzing chromosome features at different scales using the Fourier transform. Spencer *et. al* (2006) [4] introduced using the DWT to calculate the correlations between various genomic signals at different scales, and performed this analysis on human chromosome 20. However, the use of the DWT to analyze such signals at different scales is novel in the large scale, extensive *P. trichocarpa* dataset.

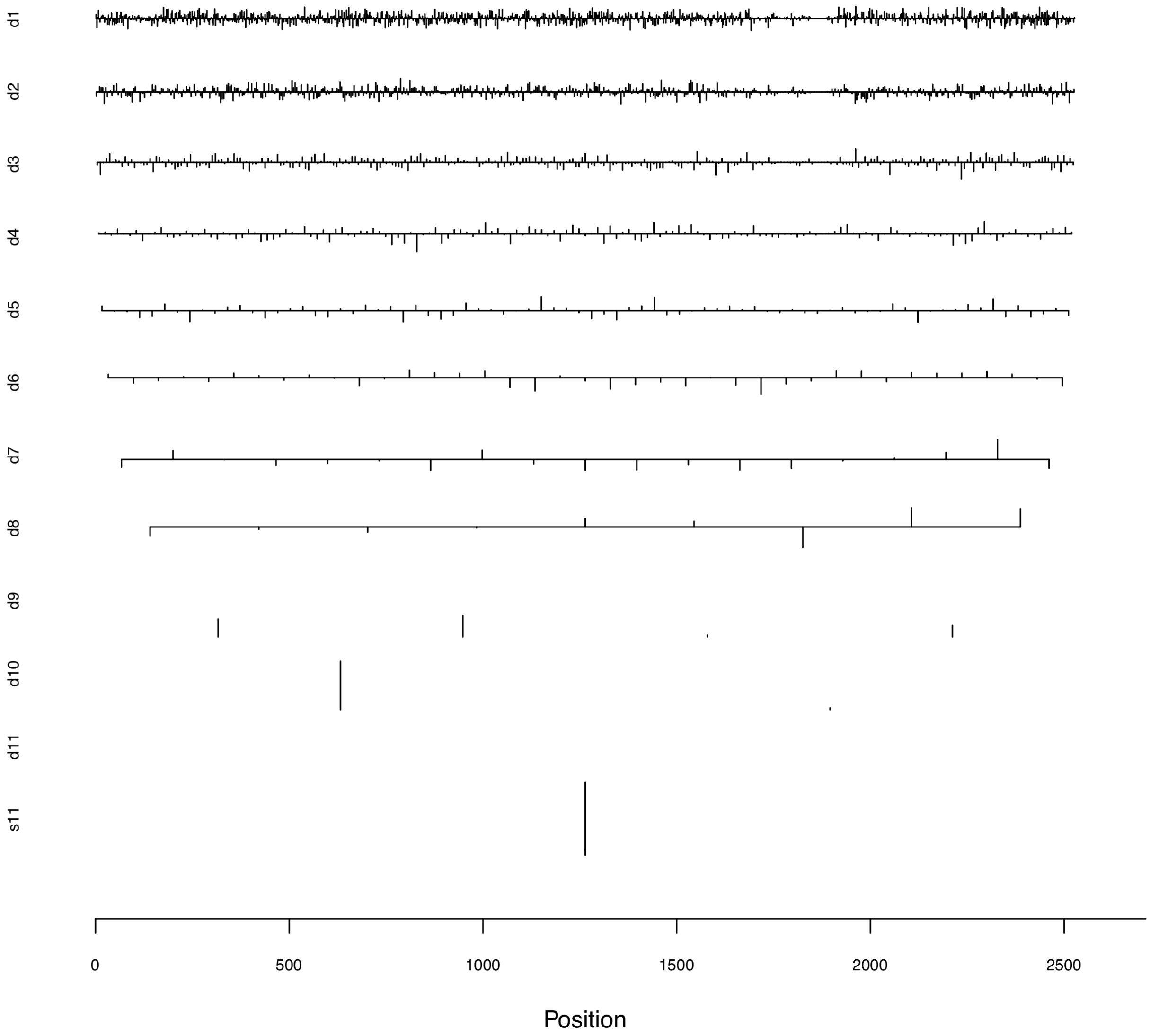


Figure 5.1: **Discrete Wavelet Transform (DWT).** Example of wavelet coefficients produced from applying the DWT to the gene density signal of *P. trichocarpa* chromosome 2. The x-axis represents position along the signal, and each line (d1 through d11) plots the wavelet coefficients of that particular scale across the signal, with d1 representing the smallest scale and d11 representing the largest scale.

5.2 Methods

5.2.1 Syntenic Block Construction

Syntenic blocks within the *P. trichocarpa* version 3.0 genome were constructed using CoGe SynMap [8, 9]. Syntenic segments were computed based on gene order, within a maximum of 10 non-matching genes between matching genes, and a minimum of 5 aligned genes per segment, similar to the parameters used in the syntenic block analysis of the original genome [1]. Synonymous substitution rates (Ks) were also calculated. All settings used can be seen in Supplementary Figure S5.1. Syntenic blocks were visualized using Circos [10].

5.2.2 Putative Ancestral Centromeres

For each chromosome, syntenic blocks which overlapped with putative centromere locations (Table S4.1 in Chapter 4) on a chromosome were extracted. The wavelet transform was used to smooth over smaller syntenic blocks that were close together in order to identify putative ancestral centromere boundaries. For each pair of chromosomes (i, j), nucleotide positions within syntenic blocks on chromosome i that overlapped with centromeric regions on chromosome j were extracted. These represented nucleotide positions potentially within ancestral centromeric regions. These positions were transformed into a density profile, counting the number of such positions in 10kb bins across chromosome i. These density profiles were constructed for each pair of chromosomes. Peaks in these density profiles indicate potential ancestral centromere locations (Figure 5.2B). The Continuous Wavelet Transform (CWT) was used to smooth over these density profiles to get ancestral centromere boundaries from ancestral centromeres potentially made up of many smaller syntenic blocks (Figure 5.2A, B and C).

The CWT was calculated for each density profile using the wmtsa R package [11, 12] using the “gaussian2” wavelet (Ricker wavelet) and a variance of 1. The scale s and position p

at which the maximum wavelet coefficient was obtained were identified, and the vector of coefficients at scale s was extracted. Syntenic block boundaries were identified as the x-intercepts on either side of the maximum peak.

5.2.3 Density Distribution Construction

Density profiles of various features across each chromosome were constructed. The construction of some of these density profiles was described in detail in Chapter 4 and are thus described briefly below.

SNP Density

Variant data used in the previous chapter, consisting of $\sim$10 million SNPs across 882 *P. trichocarpa* individuals corresponding to the 90% "PASS" tranche of the full set of 28,342,758 SNPs (DOI 10.13139/OLCF/1411410) were used. A density profile was calculated for each chromosome, counting the number of SNPs residing in non-overlapping 10kb bins across each chromosome.

Gene Density

Gene density profiles were constructed for each chromosome by counting the number of nucleotide positions within non-overlapping 10kb windows that reside within genes. Gene boundaries were determined from the annotation file obtained from the *P. trichocarpa* version 3 genome annotation available on Phytozome [13].

Methylation Density

Methylation (MeDIP-seq) reads from ten *P. trichocarpa* tissues (bud, callus, female catkin, internode explant, leaf, make catkin, phloem, regenerated internode, root and xylem) [14]

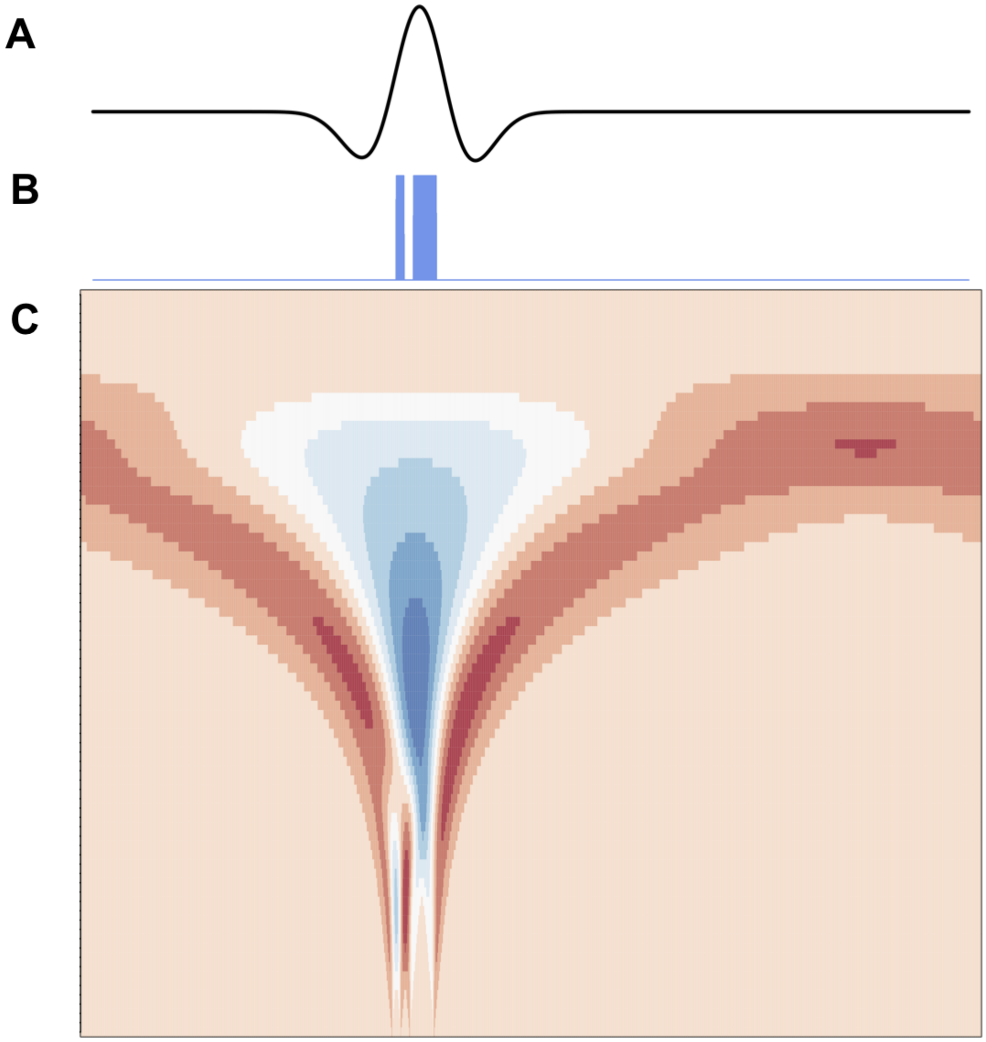


Figure 5.2: **Determining ancestral centromere boundaries.** For a given chromosome, the density profile of syntenic blocks which originate from a centromeric region on another chromosome was constructed (B). The continuous wavelet transform was calculated for this signal resulting in a coefficient landscape (C). The x-axis of a coefficient landscape represents the position along the chromosome, and the y-axis represents the scale, with small scales (high frequency peaks) at the bottom of the heatmap and large scales (low frequency peaks) at the top. A wavelet coefficient is calculated for each signal position and each scale, thus resulting in a landscape. The scale at which the maximum coefficient occurs was identified, and the wavelet coefficient vector at that scale was extracted (A). X-axis intercepts of this vector were considered putative ancestral centromere boundaries.

re-aligned to the version 3 assembly of *P. trichocarpa* were downloaded from Phytozome [13].

BEDTools [15] and GNU-Parallel [16] were used to count the number of reads mapped to 10kb windows across the genome, and a “mapped read density” distribution was constructed for each tissue and each chromosome, showing the number of reads which mapped to different regions of the genome, and thus indicating methylation hot spots.

Repeat Element Density

RepeatMasker (Smit, AFA, Hubley, R & Green, P. RepeatMasker Open-4.0. 2013-2015 [http://www.repeatmasker.org]) was used to identify repeat elements in the chromosomes of *P. trichocarpa* version 3. A density profile for each different family of repeats was constructed for each chromosome by counting the number of nucleotide positions that resided within an instance of that family of repeats in non-overlapping 10kb bins across the genome. These repeat densities were visualized using the IdeoViz R package [17].

Syntenic Block Boundary Density

A density vector was constructed for each chromosome indicating the number of syntenic block boundaries in non-overlapping 10kb bins across the genome.

Centromere/Ancestral Centromere Vectors

A centromere “density” vector was constructed for each chromosome using the same 10kb bins described in the density distributions above. Each element of the centromere vector was assigned the value “1” if the 10kb bin overlapped with the centromeric region and “0” otherwise. This was performed in a similar manner for ancestral centromeres. These were constructed to allow the locations of the centromere/ancestral centromeres to be correlated with other features such as gene density/TE density.

5.2.4 Discrete Wavelet Transform and Correlations

For each feature in Table 5.1, the Discrete Wavelet Transform (DWT) was calculated, producing a series of sets of coefficients with one set of coefficients for each scale computed. This was performed separately for each chromosome using the `wmtsa` R package [11, 12, 18, 19]. For each chromosome, the Spearman correlation between the wavelet coefficients of each level was calculated for each pair of features, as well as the associated p-value using the R `cor.test` function. A significance threshold of 0.05 was set. Significant correlations were visualized in Cytoscape [20].

5.3 Results and Discussion

5.3.1 Syntenic Blocks, Historic Genome Duplication Events and Ancestral Centromeres

Syntenic blocks within the *P. trichocarpa* genome as well as synonymous substitution rates (Ks) for homologous segements were constructed using CoGe [8]. A dotplot of the homologous chromosome regions can be seen in Figure 5.3A and a distribution of the Ks values in Figure 5.3B. A first look at the dotplot reveals some long regions of homologous segments, and some smaller, sparse regions. The Ks distribution (Figure 5.3B) as three major peaks. It is interesting to note that the long homologous chromosomal regions tend to fall within the blue peak, whereas the smaller, sparse homologous regions tend to fall within the green and orange peaks (Figures 5.3C and 5.3D). This distribution of Ks values as well as the major syntenic blocks on the dotplot agree with findings in the original genome study for *P. trichocarpa* [1], in which Tuskan *et al.* describe three genome duplications in the history of *Populus*, namely, the Salicoid duplication, the Eurosid duplication and an Ancient duplication. The major syntenic blocks presented by Tuskan *et al.* (2006) are thought to have arisen from genome rearrangements following the most recent whole-genome duplication event, the Salicoid duplication. Our distribution of Ks

Table 5.1: Genomic Features for which density profiles were constructed. "Feature name" shows the full name of the feature, "Source" shows the method/data where the feature arose from and "Code" shows a shortened name for the feature used in the networks (Figure 5.9).

Feature name	Source	Code
LTR/Caulimovirus density	RepeatMasker	Caulimovirus
LTR/Copia density	RepeatMasker	Copia
LTR/Gypsy density	RepeatMasker	Gypsy
DNA/hAT-Ac density	RepeatMasker	hAT-Ac
snRNA density	RepeatMasker	snRNA
rRNA density	RepeatMasker	rRNA
DNA/PIF-Harbinger density	RepeatMasker	PIF-Harbinger
tRNA density	RepeatMasker	tRNA
LINE/L1 density	RepeatMasker	LINE
DNA/CMC-EnSpm density	RepeatMasker	CMC-EnSpm
Low complexity repeat density	RepeatMasker	Low_complexity
Simple repeat density	RepeatMasker	Simple_repeat
Satellite density	RepeatMasker	Satellite
DNA/hAT-Tag1 density	RepeatMasker	hAT-Tag1
RC/Helitron density	RepeatMasker	Helitron
DNA/hAT-Tip100 density	RepeatMasker	hAT-Tip100
Gene density	version 3.0 gene annotation	Gene
SNP density	PASS tranche of *P. trichocarpa variants* (DOI 10.13139/OLCF/1411410)	SNP
Callus methylation density	MeDIP-Seq reads from callus tissue [14]	callus

Continued on next page.

Table 5.1: (continued)

Feature name	**Source**	**Code**
Root methylation density	MeDIP-Seq reads from root tissue [14]	root
Leaf methylation density	MeDIP-Seq reads from leaf tissue [14]	leaf
Xylem methylation density	MeDIP-Seq reads from xylem tissue [14]	xylem
Bud methylation density	MeDIP-Seq reads from bud tissue [14]	bud
Female catkin methylation density	MeDIP-Seq reads from female catkin tissue [14]	female_catkin
Male catkin methylation density	MeDIP-Seq reads from male catkin tissue [14]	male_catkin
Internode explant methylation density	MeDIP-Seq reads from internode explant tissue [14]	internode_explant
Phloem methylation density	MeDIP-Seq reads from phloem tissue [14]	phloem
Regenerated internode methylation density	MeDIP-Seq reads from regenerated internode tissue [14]	regenerated_internode
Centromere	centromere positions determined in Chapter 4	cen
Ancestral centromere	Ancestral centromere positions determined from wavelet transform of syntenic blocks	ancestral_centromere

Table 5.2: Syntenic blocks recovered in this analysis of *P. trichocarpa* version 3.0 compared to that of the original genome study [1]. Columns 1 and 2 show the source and target chromosomes of syntenic blocks identified in the original genome study. Column 3 indicates which of those syntenic blocks were not identified in the current analysis.

Source Chromosome	**Syntenic blocks with Source Chromosome in [1]**	**Missing in this analysis**
1	3, 9, 11, 17	
2	4, 5, 14	4
3	1, 5	5
4	2, 9, 11, 17	2
5	2, 3, 7	3
6	16, 18	
7	5, 14, 17	
8	10	
9	1, 4	
10	8	
11	1, 4, 13	
12	15	
13	11, 19	
14	2	
15	12	
16	6	
17	1, 4, 7	
18	6	
19	13	

values (as an estimate of divergence/evolutionary age) recovers the three peaks mostly likely representing the three duplication events described previously. Given that the blue peak seems to represent the longer syntenic blocks which match those found in the original genome study, we can tentatively conclude that these are the same syntenic blocks arising from the Salicoid duplication, represented by the blue peak (Figure 5.3C and 5.3D).

Table 5.2 shows the pairs of homologous chromosome regions found in Tuskan *et al.* (2006), as well as those which were not found in this analysis. It is expected that there might be some differences, primarily because the genome assembly has changed significantly since the genome's first release. However, the fact that all but two syntenic blocks were recovered indicates that the large scale structure of the genome has not changed much between version 1 and version 3.

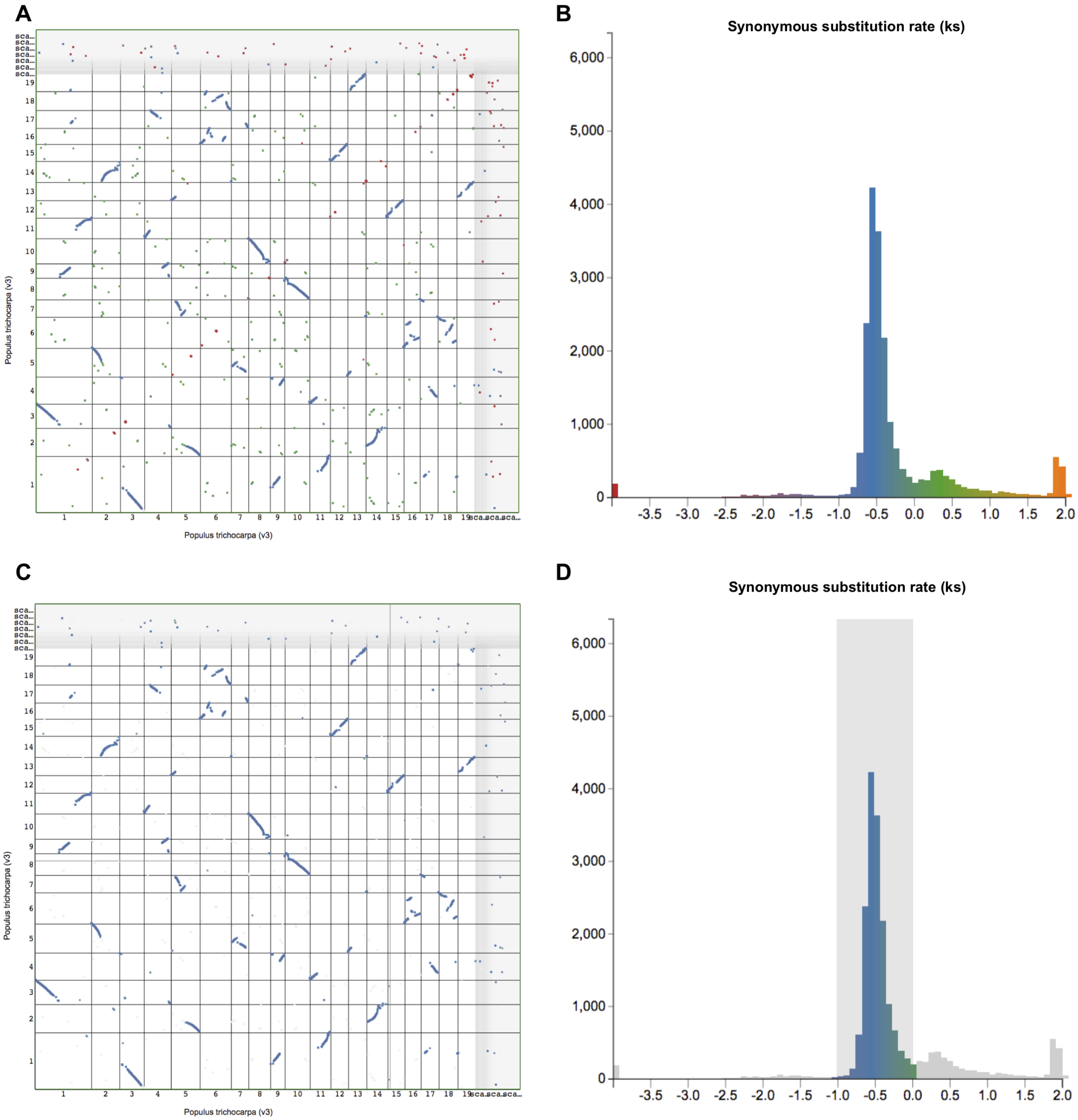


Figure 5.3: **Synonymous substitution rates.** (A) Syntenic dot plot and (B) Distribution of -$\log_{10}$ Ks values for *P. trichocarpa*, generated using CoGe [8, 9] (C) Syntenic blocks with ks values in the range of the selected blue peak in (D).

Table 5.3: Putative ancestral centromeres identified from syntenic blocks and centromere locations.

Ancestral centromere on	From chromosome
1	17
2	5
2	14
4	9
5	7
6	18
6	16
14	2

Figure S5.2 shows circos plots of all the syntenic blocks/homologous chromosome regions, centered around each chromosome separately, with each circos plot representing the syntenic blocks involving a particular chromosome. Centromeric regions predicted in Chapter 4 are shown as highlights on the chromosome ideogram. The homologous chromosome pairs are clearly visible, as well as the chromosome rearrangements which occurred. Visualizing only the syntenic blocks which overlap with centromeric regions provides information on the fate of active centromeres/centromeric DNA post rearrangement. One can also see evidence for cases where the active centromere of a given chromosome segment was maintained after the chromosome rearrangement. For example, one can see from Figure 5.4A that chromosome 1 and chromosome 3 share a very large syntenic block which Ks values suggests is from the salicoid duplication event. Figure 5.4B shows the syntenic blocks on chromosome 1 originating from centromeric regions, and indicates that the active centromere of the chromosome 1-3 duplication was possibly maintained post rearrangement. It also points to a section of DNA on chromosome 1 that originated from the centromeric region of chromosome 17, and is thus a potential ancestral centromere. More examples of putative ancestral centromeres can be found in Table 5.3.

Chromosome 11 is an interesting example which potentially could have acquired a new centromere post-rearrangement (Figure 5.4C). Chromosome 11 appears to be composed of segments of chromosome 1 and chromosome 4, however neither of those segments

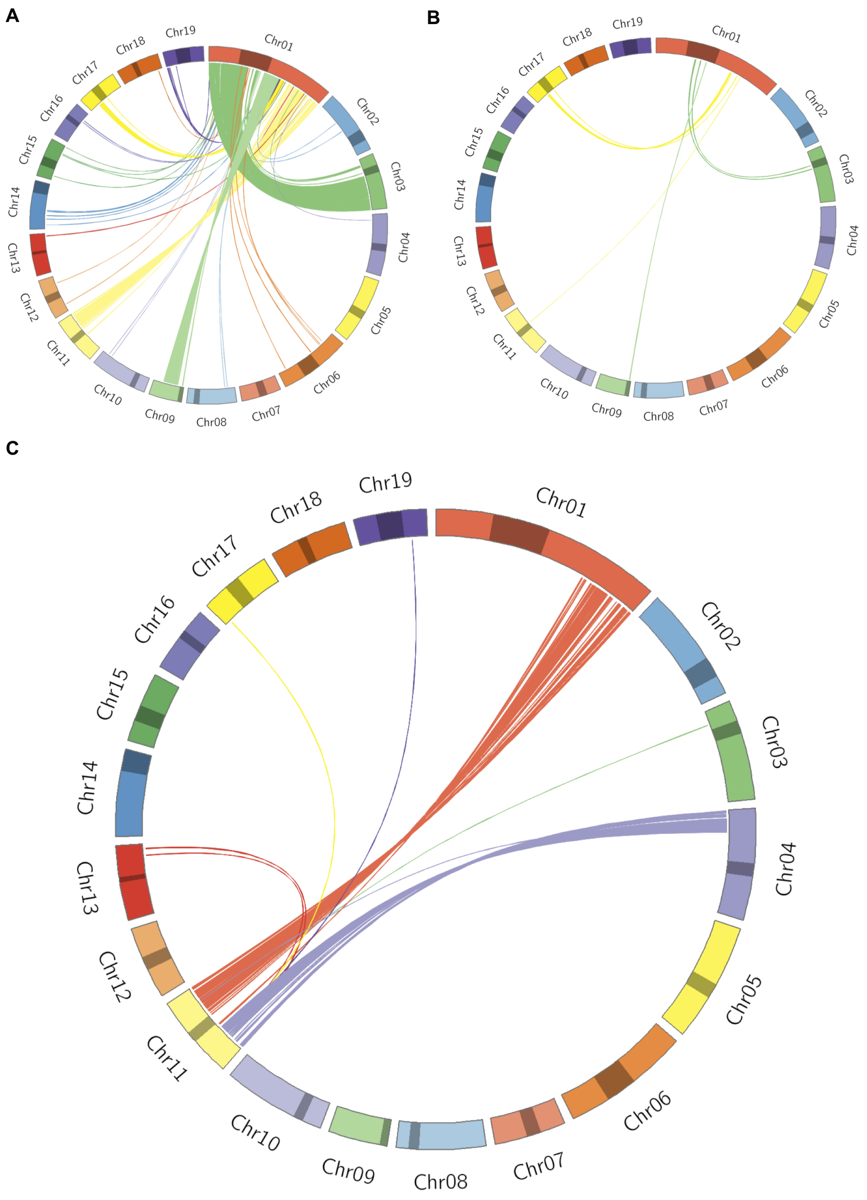


Figure 5.4: **Syntenic blocks and centromeres.** (A) Syntenic blocks on chromosome 1. (B) Syntenic blocks on chromosome 1 that originate in centromeric regions on other chromosomes. (C) Syntenic blocks for chromosome 11.

contained an active centromere. One can hypothesize that Chromosome 11 formed a new centromere post rearrangement.

5.3.2 Repeat and Transposable Element Density

Active plant centromeres are defined as DNA regions binding to the centromere-specific histone, CENH3 [21]. Centromeric DNA tends to consist of repeat sequences. While retrotransposons in general occur less frequently in active centromere regions [22, 23], certain retroelement groups are enriched in centromeric regions, such as gypsy repeats [22].

Various repeat sequences and retroelements/TEs were identified in the *P. trichocarpa* genome. These repeats belonged to many different families (Table 5.1) and had very different distributions across the genome. Three classes of repeats showed particular enrichment in the centromeric regions, including gypsy repeats derived from gypsy group retrotransposon, satellite repeats and Long Interspersed Nuclear Elements (LINE) repeats (Figures 5.5, 5.6 and 5.7). Gypsy repeats are known to be found in centromeric regions [24, 25, 26, 27]. We also see some putative peaks in the gypsy repeat density in some of the putative ancestral centromeric regions. It is expected that we see enrichments of satellite repeats in the centromeric regions as these are known components of centromeric DNA [28, 22, 26]. It is interesting that we also see satellite repeats in some of the potential ancestral centromeric regions. LINE transposable elements are known to be found in the pericentromeric regions [29], and there is some new evidence of LINE repeats occuring in active centromeres in sunflowers and bats [30, 31].

5.3.3 Discrete Wavelet Transform and Scale-Specific Correlation

Biological signals, such as the density profiles in Table 5.1, can have different relationships with each other depending on the scale at which one is looking [4]. While two features my be correlated at certain scales, they may be anti-correlated at others. Correlating

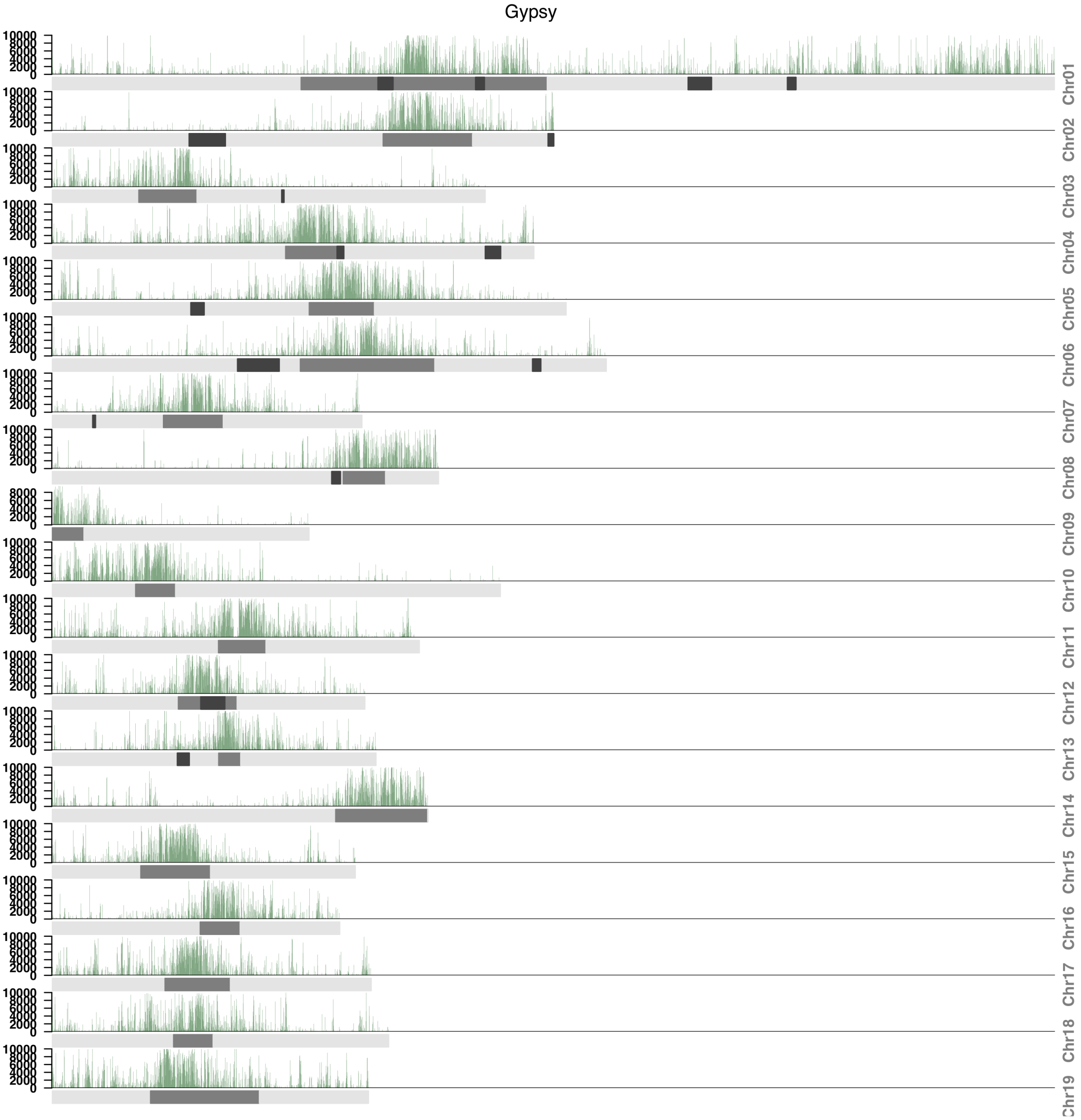


Figure 5.5: **Gypsy TEs.** Density of Gypsy TE repeats across all chromosomes in *P. trichocarpa*.

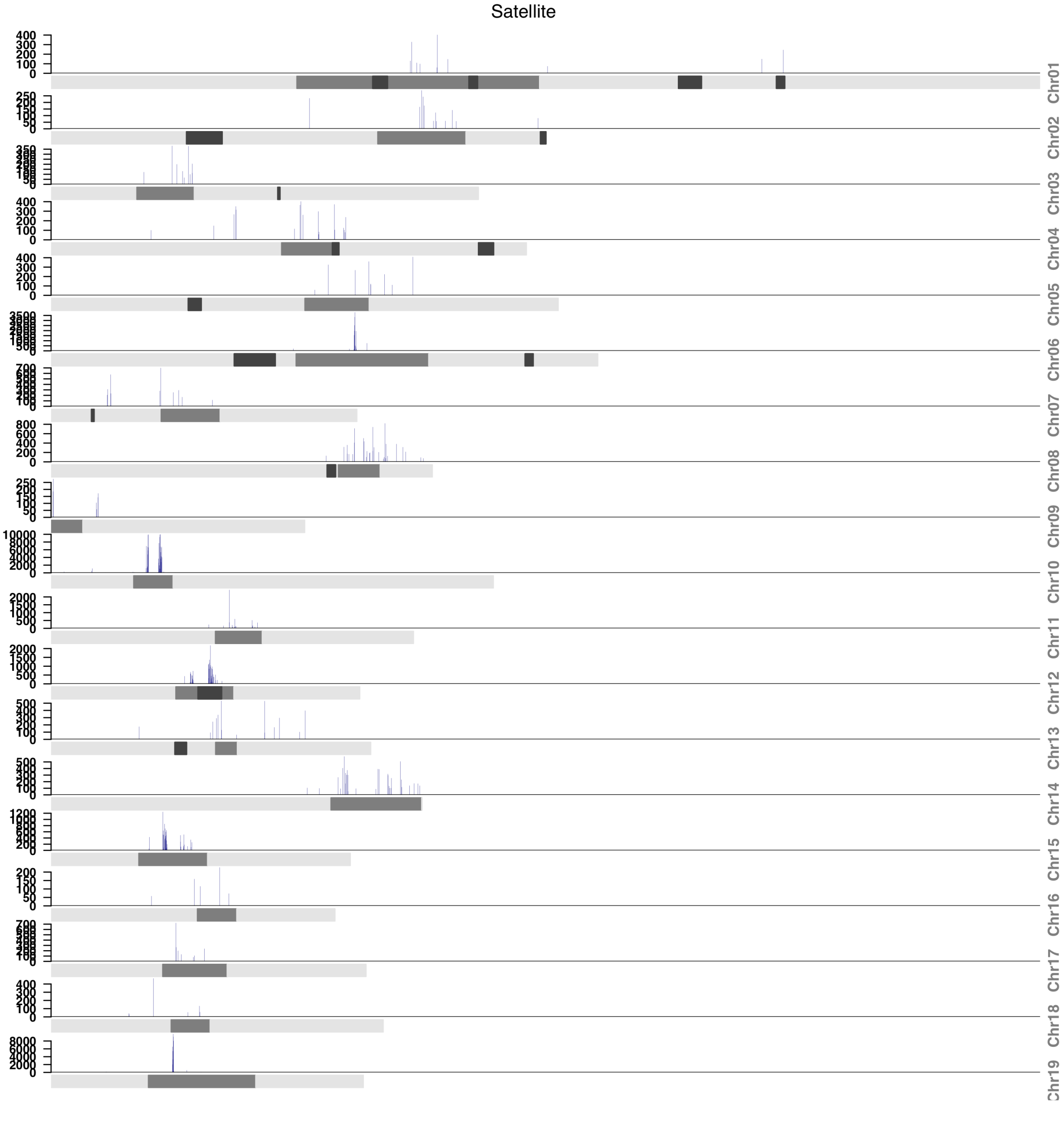


Figure 5.6: **Satellite repeats.** Density of Satellite repeats across all chromosomes in *P. trichocarpa*.

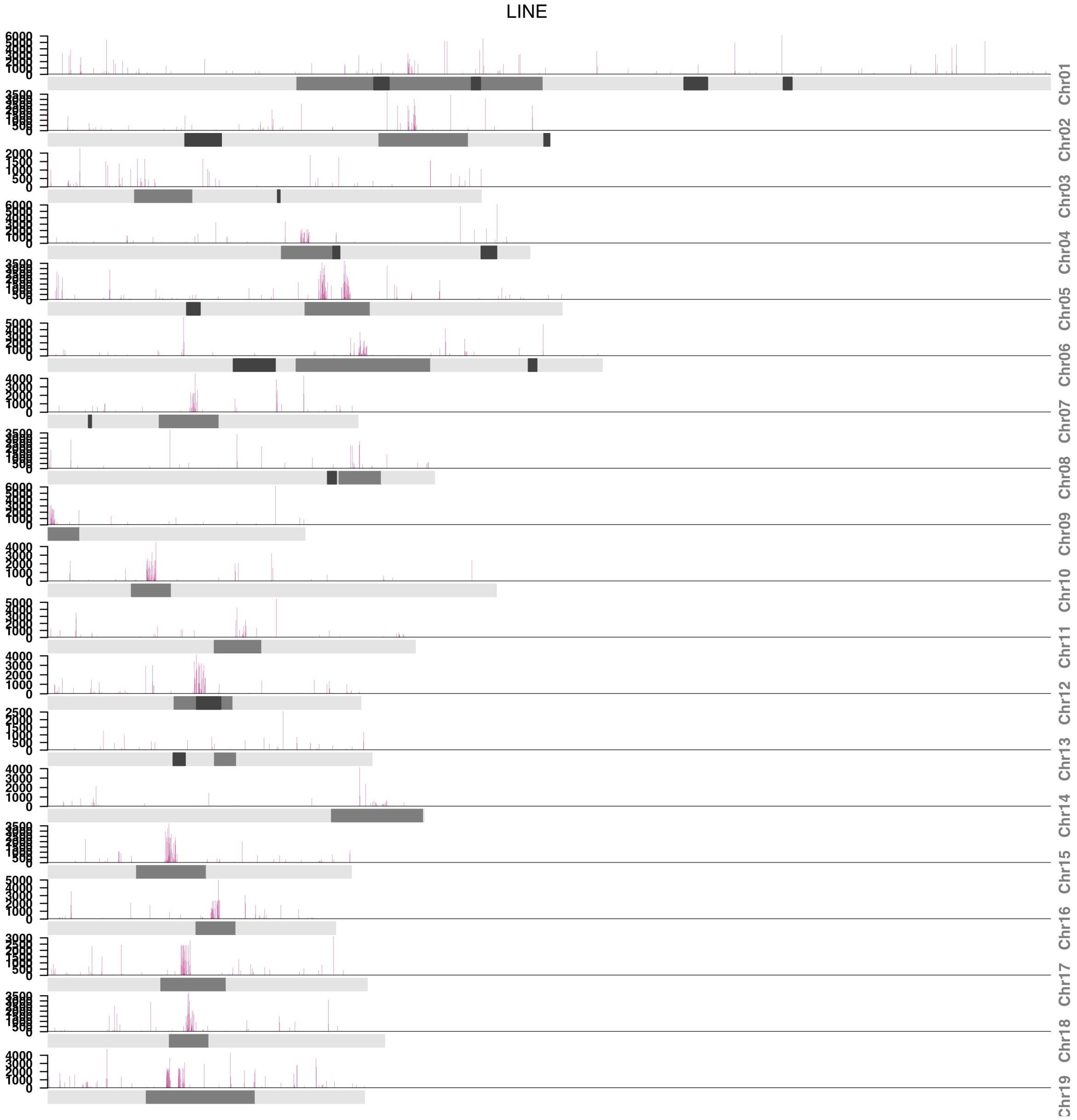


Figure 5.7: **LINE TEs.** Density of LINE TE repeats across all chromosomes in *P. trichocarpa*.

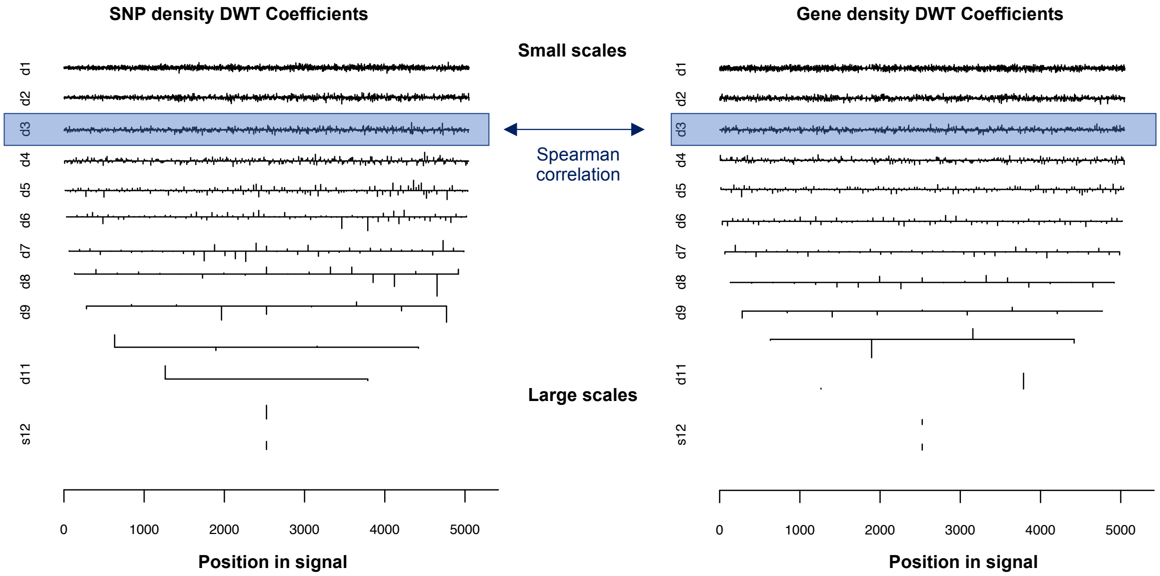


Figure 5.8: **DWT Coefficient Correlation Method.** Scale specific correlations are calculated by discrete wavelet transforming the signals and then calculating the correlation between the coefficients of matching scales of the features.

raw signals does not allow one to analyze these features at different scales. The Discrete Wavelet Transform (DWT) is a sampled version of the CWT, and involves the sampling of the base pair (x-axis) and scale (y-axis) dimensions [5] and produces a series of sets of coefficients, with one set of coefficients for each scale computed. Calculating regression or correlation between wavelet coefficients of two different signals allows for scale-specific inferences to be determined [32]. The Spearman correlation coefficient was calculated between the wavelet coefficients of each scale for each pair of features and each chromosome (Figure 5.8). Significant (p < 0.05) correlations were represented as networks and investigated for interesting scale-specific relationships.

Figure 5.9 shows the significant (p < 0.05) correlation between features at different scales/levels for chromosome 13, represented as networks. Each network (A-H) in Figure 5.9 represents the correlations of a particular scale, with Level 1 representing small scale features and Level 8 representing large-scale features. Therefore, the network in Figure 5.9A represents the significant Spearman correlations between the level 1 wavelet coefficients of features, thus representing correlations between fine-scale features of the signals, whereas Figure 5.9H represents the significant Spearman correlations between the level 8 wavelet coefficients of features, thus representing correlations between large-scale features of the signals. An explanatory example of scale-specific correlations can be seen when looking at the relationship between SNP density (blue nodes in Figure 5.9) and gene density (green nodes in Figure 5.9). At small scales (Level 1, Level 2, Level 3) one can see that SNP density and gene density are significantly negatively correlated. This small scale feature can be interpreted as the fact that SNPs tend to occur in intergenic space as opposed to within genes. However, at medium-large scales (Level 6) SNP density and gene density are positively correlated, likely driven by the low density of genes and SNPs found in centromeric regions. Similar patterns are seen on other chromosomes (see Figures S5.4 - S5.21 for DWT correlation networks on the other 18 chromosomes.)

This relationship between SNP density and gene density for a given chromosome can also be visualized as line plot of wavelet coefficient correlations (Figure 5.10). The x-axis represents the scale (level), and the y-axis represents the Spearman correlation coefficient between wavelet coefficients of the at a particular scale. Red points represent negative correlations, blue points represent positive correlations, solid circles represent significant ($p \leq 0.05$) correlations and empty circles represent non-significant correlations. One can see that the correlation between SNP density and gene density starts as negative at small scales, and becomes positive at larger scales, and the correlation drops off at the extremely large scales. This general pattern can be seen across all chromosomes (Figure 5.11). A hypothesis as to why the correlation drops off at very high scales is related to the difference in densities between gene and SNP density in the pericentromere. Gene density drops off at the pericentromeres and is at its lowest in the centromeric regions, whereas SNP density peaks in the pericentromeric regions and drops off rapidly in the centromeric regions (Figure 4.3 Chapter 4).

SNP density appears to be correlated with methylation density across small-medium scales, indicating that areas of high mutation tend to be highly methylated. A possible cause of this correlation is the fact that increased mutation rates are seen in methylated cytosines, in that they are easily deaminated to thymine [33, 34, 35]. One can see that in many cases, features correlate very differently at large scales than at small scales. These are details that

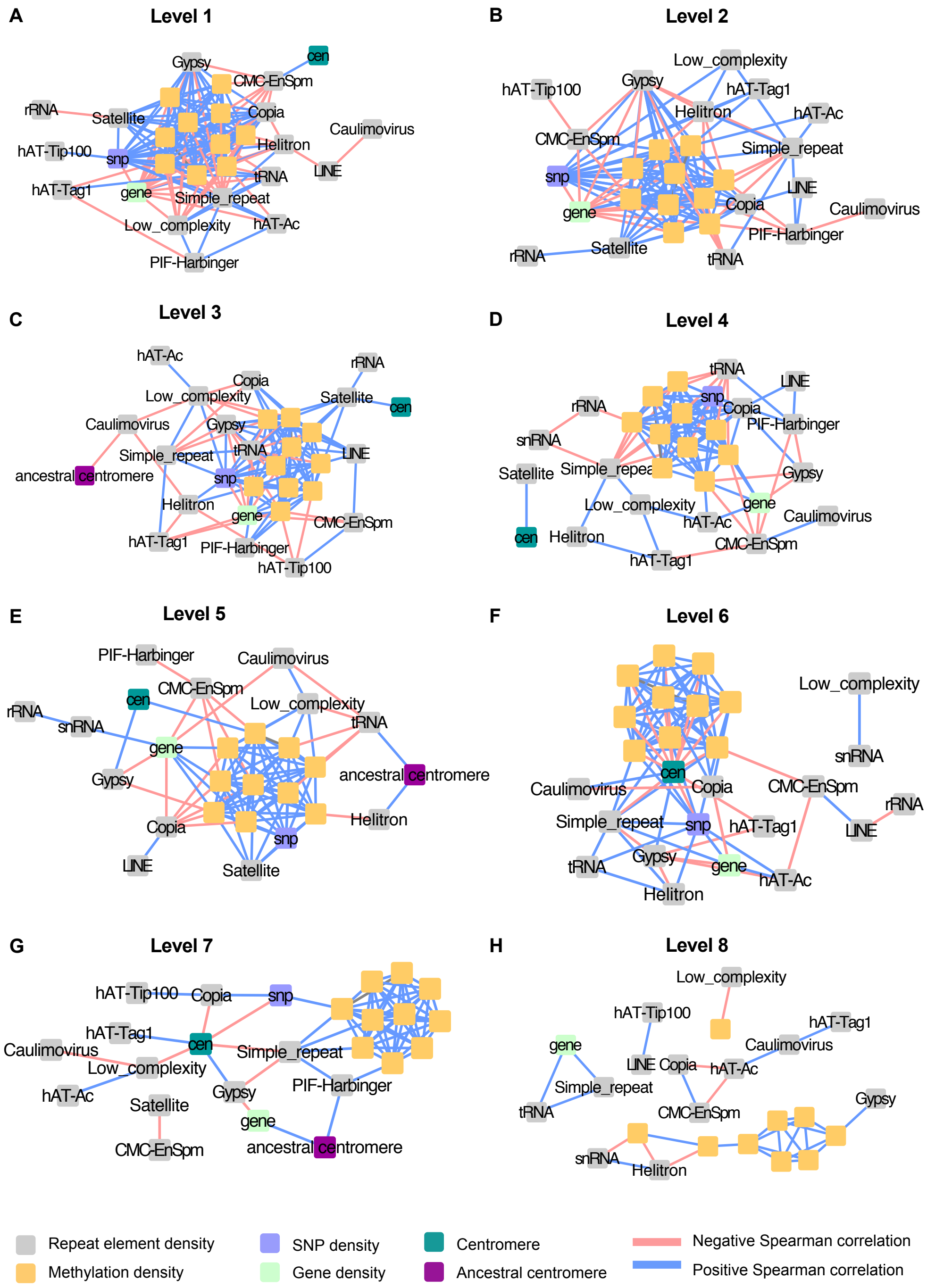


Figure 5.9: **DWT correlation networks for chromosome 13.** Discrete wavelet transform correlation networks representing the significant ($p < 0.05$) Spearman correlations between wavelet coefficients of different features, at each scale on chromosome 13. Each of (A) through (F) represents correlations of a specific scale, with (A) representing the smallest scale and (H) representing the largest scale. Each node represents a genomic feature, and each edge represents a significant Spearman correlation between two genomic features at a particular scale. Different types of features are indicated by different colors. Up to ten different methylation nodes exist at each scale, one for each tissue type methylation data was derived from (Table 5.1)

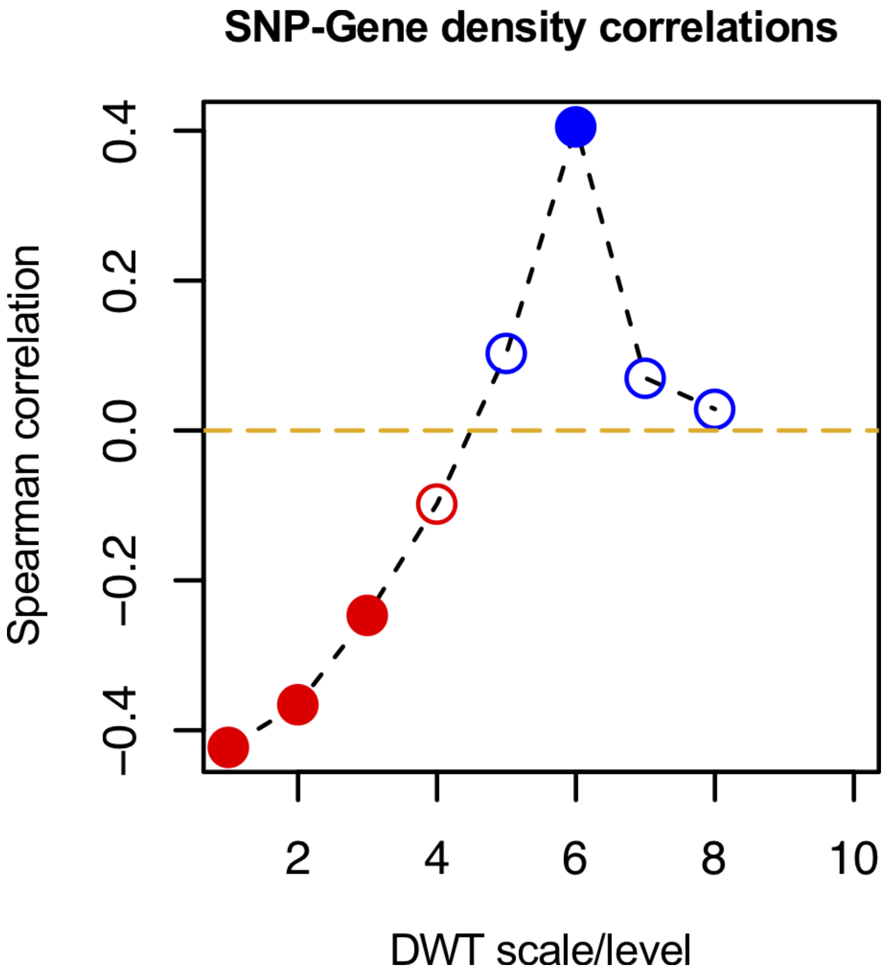


Figure 5.10: **SNP-gene DWT coefficient correlation for chromosome 13.** Each point represents the Spearman correlation coefficients between SNP density and gene density on chromosome 13 at a particular scale. The x-axis represents the scale, the y-axis represents the correlation coefficient value, solid circles represent significant correlations, empty circles represent non-significant correlations, blue circles represent positive correlations and red circles represent negative correlations. One can see that at small scales, SNP density and gene density are negatively correlated, whereas at medium-large scales they are positively correlated.

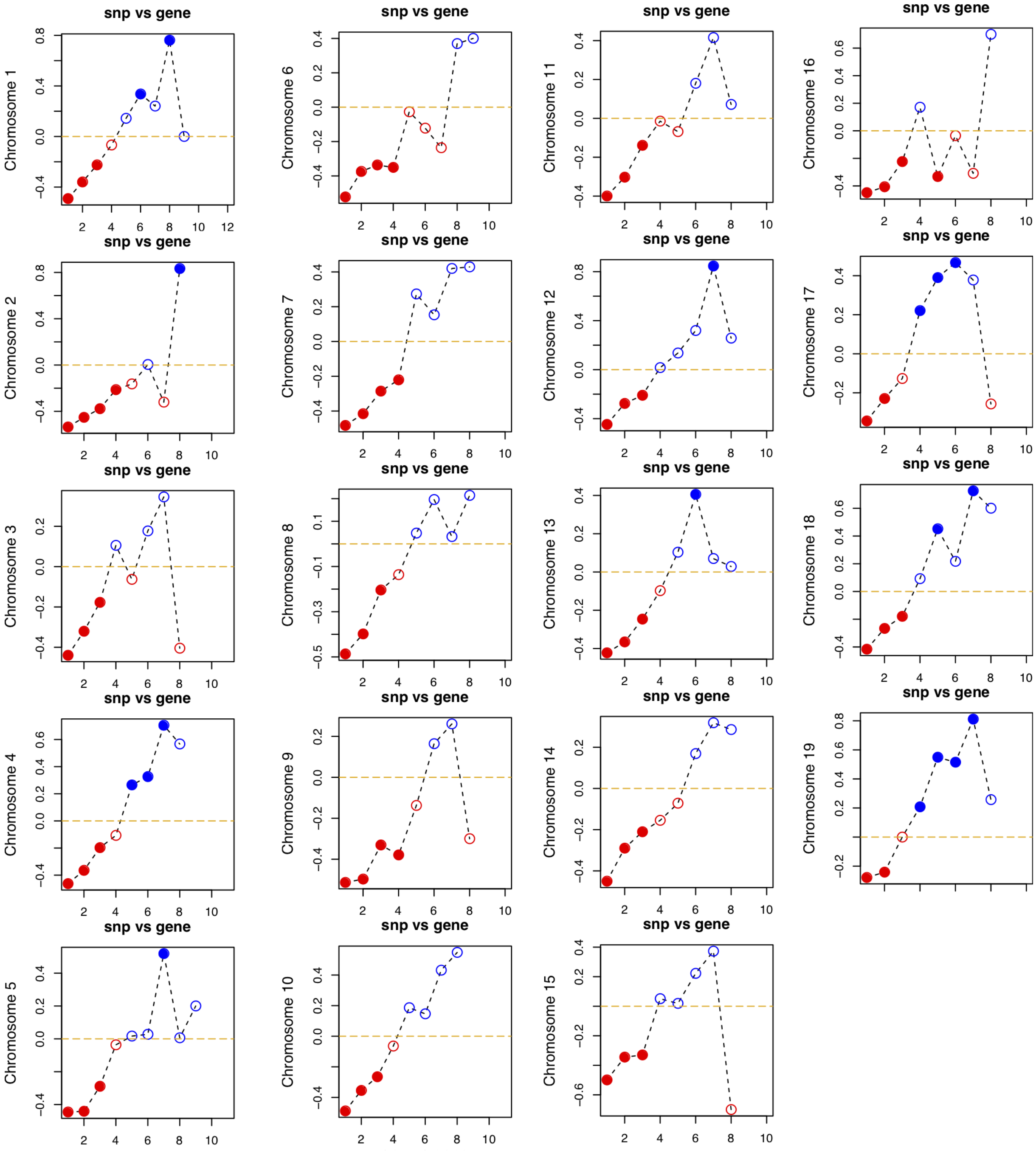


Figure 5.11: **SNP-gene DWT coefficient correlation for all chromosomes.** Similar to Figure 5.10, Spearman correlation coefficients between SNP density and gene density are plotted as a line plot across scales. A separate plot is shown for each chromosome. For each plot, the x-axis represents the scale, the y-axis represents the correlation coefficient value, solid circles represent significant correlations, empty circles represent non-significant correlations, blue circles represent positive correlations and red circles represent negative correlations.

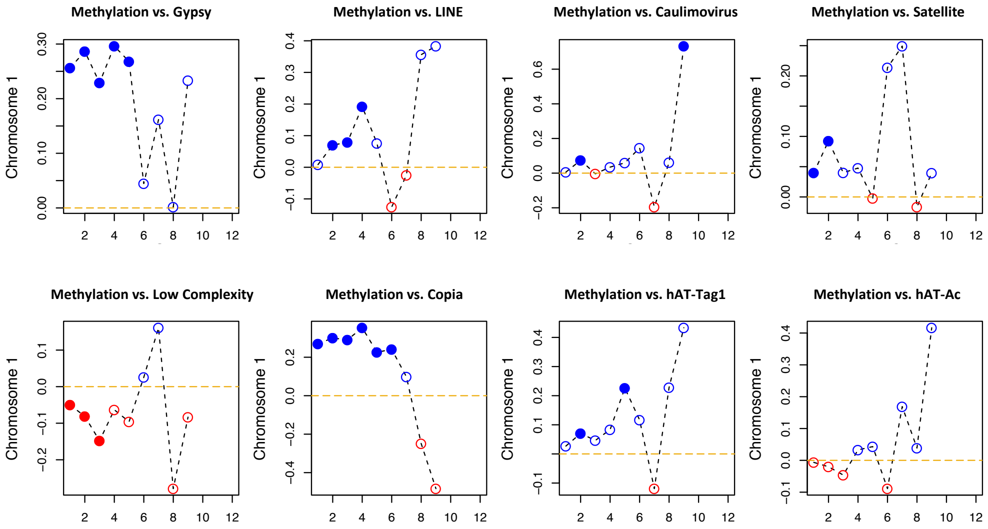


Figure 5.12: **DWT correlation between repeats and methylation.** Spearman correlation coefficients between different repeat class density and leaf tissue methylation density on chromosome 1 are plotted as a line plot across scales. For each plot, the x-axis represents the scale, the y-axis represents the correlation coefficient value, solid circles represent significant correlations, empty circles represent non-significant correlations, blue circles represent positive correlations and red circles represent negative correlations.

would be missed by analyzing the signal as a whole. Thus, the use of the wavelet transform to unravel the different scales of features can provide insights into chromosome structure and association between features, and suggest where multiple forces may be at work.

5.3.4 Correlates of Repeats and TEs

Many transposable element and repeat sequences are correlated with DNA methylation across various scales in various tissues (Figure 5.9). These include satellite repeats as well as copia, gypsy and LINE TEs. DNA methylation is known to be a defense mechanism against TEs as a silencing mechanism [28, 2, 36] and it has been noted that repetitive DNA sequences in plants are methylated in all (or most) tissues [37]. One can see in Figure 5.12 that Gypsy, LINE, Copia and Satellite repeats are significantly positively correlated with methylation in leaf tissue at several scales.

Interestingly, low complexity repeat sequences are significantly negatively correlated with methylation across several scales, especially low scales. This suggests that low complexity repeat sequences are generally not methylated. These correlation patterns are generally observable across most chromosomes. No significant correlations were found between hAT-Ac transposable elements and methylation at any scale on chromosome 1. This is true for most chromosomes, and there is no easily discernible pattern of the correlation coefficients. This indicates that there is no specific relationship between hAT-Ac transposable elements and methylation. One can see the expected correlation between the densities of satellite repeats and gypsy TEs with the position of the centromere (Figure 5.9C-G).

Transposable elements and repeat sequences are thought to contribute to genome instability in various ways. Inverted repeats can cause hairpin bends which can contribute to double strand breaks in the DNA [28]. Simple repeats (microsatellites) are difficult to replicate because of a process known as slippage [38]. DNA strands can dissociate during replication and then mis-align because because of the repeated sequence [38]. Repeat sequences can also cause unequal homologous recombination which can result in chromosome rearrangement and DNA loss [3, 39, 28]. The syntenic blocks and syntenic block boundaries provide a representation of the chromosome rearrangements which have happened in the evolutionary history of the *P. trichocarpa* genome. Investigating the correlations between the density of syntenic block boundaries and the density of repeat sequences at different scales revealed that syntenic block boundaries are significantly correlated with simple repeat sequences across multiple scales and multiple chromosomes (Table 5.4).

This provides a hypothesis as to the cause of the chromosomal rearrangements that occurred in the *P. trichocarpa* genome.

Table 5.4: Correlation of syntenic block boundary density with simple repeats.

Chromosome	Scale	Spearman correlation	p-value
1	2	-0.0578312274425647	0.0399643752344869
3	4	0.184028260110646	0.0319798347524092
5	4	0.171515348694062	0.0295944897603272
7	7	0.578950932303566	0.0485584019884882
8	2	0.128312766893837	0.00460927612992136
8	3	0.138092860354941	0.031407192590853
11	6	0.453705145990907	0.0153086010055195
11	7	0.546256831729449	0.0432786009618237
12	3	0.157904734168371	0.026683532852477
13	1	0.0986868825001212	0.00477802485347983
13	3	0.154932652102235	0.0269210183189224
13	6	0.413837837184921	0.0397351917575558
17	4	0.264324775120754	0.00787328447282763
17	6	0.412042857690601	0.0406893321529381

5.4 Conclusions and Future Directions

This study made use of multiple data types and wavelet-based signal processing to investigate the evolutionary history of the structure of the *P. trichocarpa* genome. The major syntenic blocks found in the original genome paper [1] were recovered in the new version 3.0 assembly, as well as the evidence for the Salicoid duplication event, the Eurosid duplication event and the ancient dupication event, in the form of the Ks distribution.

Expected distributions of TEs and repeat sequences were found, and these provide an opportunity to refine the centromere locations identified in Chapter 4. DWT-correlation analysis revealed that many TEs are likely methylated in *P. trichocarpa*, potentially a genome defense mechanism. In addition, correlation between the locations of syntenic block boundaries and simple repeat sequences suggest a hypotheses as to the reason for the exact rearrangement points in the genome.

This work has provided resources for a method useful for the interrogation and comparison of multiple data types in *P. trichocarpa*. Future work should dig deeper into the correlations found in the discrete wavelet transform. Correlations such as those between syntenic block boundaries and simple repeats should be investigated at the single base pair level. The distance and length of simple repeat sequences from every syntenic block boundary should be determined. This could provide more detailed evidence to support the hypothesis that the genome rearrangements were driven by repeat sequences. This should be similarly applied for every family of TEs and repeats.

The use of different similarity metrics should also be explored in the DWT-based scale-specific correlation. Different similarity metrics could extract different patterns and relationships between the wavelet coefficients of signals and provide different perspectives. The DUO metric will be a particularly interesting addition. The structure of the DUO similarity metric is similar to the SNP (Single Nucleotide Polymorphism) correlation metric, CCC, also developed by Sharlee Climer *et. al* [40, 41]. It categorizes values in a matrix into high, medium and low values, and then for each pair of rows, it calculates a scaled co-occurrence of all 4 possible combinations of high values and low values. Thus it will be particularly interesting in investigating if/how it isolates and compares the peaks in a wavelet coefficient vector.

In addition, network topology comparisons should be performed across DWT correlation networks from different scales and chromosomes, in order to determine which patterns are constant across the genome and which are chromosome specific. Topology comparison methods such as those discussed by Weighill *et al.* (2016) [42] can be used to perform such comparisons.

The correlation between different families of TEs and different gene functions should also be investigated. This can be achieved by performing similar DWT-correlation analysis, but this time separating the gene density signal into different components, with each component representing a different overall gene function (e.g. “kinase” or “glycosyl transferase”).

The rich datasets of *P. trichocarpa* combined with the power of scale-specific correlation analysis should prove incredibly useful in investigating the structure and evolutionary history of the *P. trichocarpa* genome.

Funding

Funding provided by The BioEnergy Science (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.

This research was also supported by the Department of Energy Laboratory Directed Research and Development funding (7758), at the Oak Ridge National Laboratory. Oak Ridge National Laboratory is managed by UT-Battelle, LLC, for the US DOE under contract DE-AC05-00OR22725.

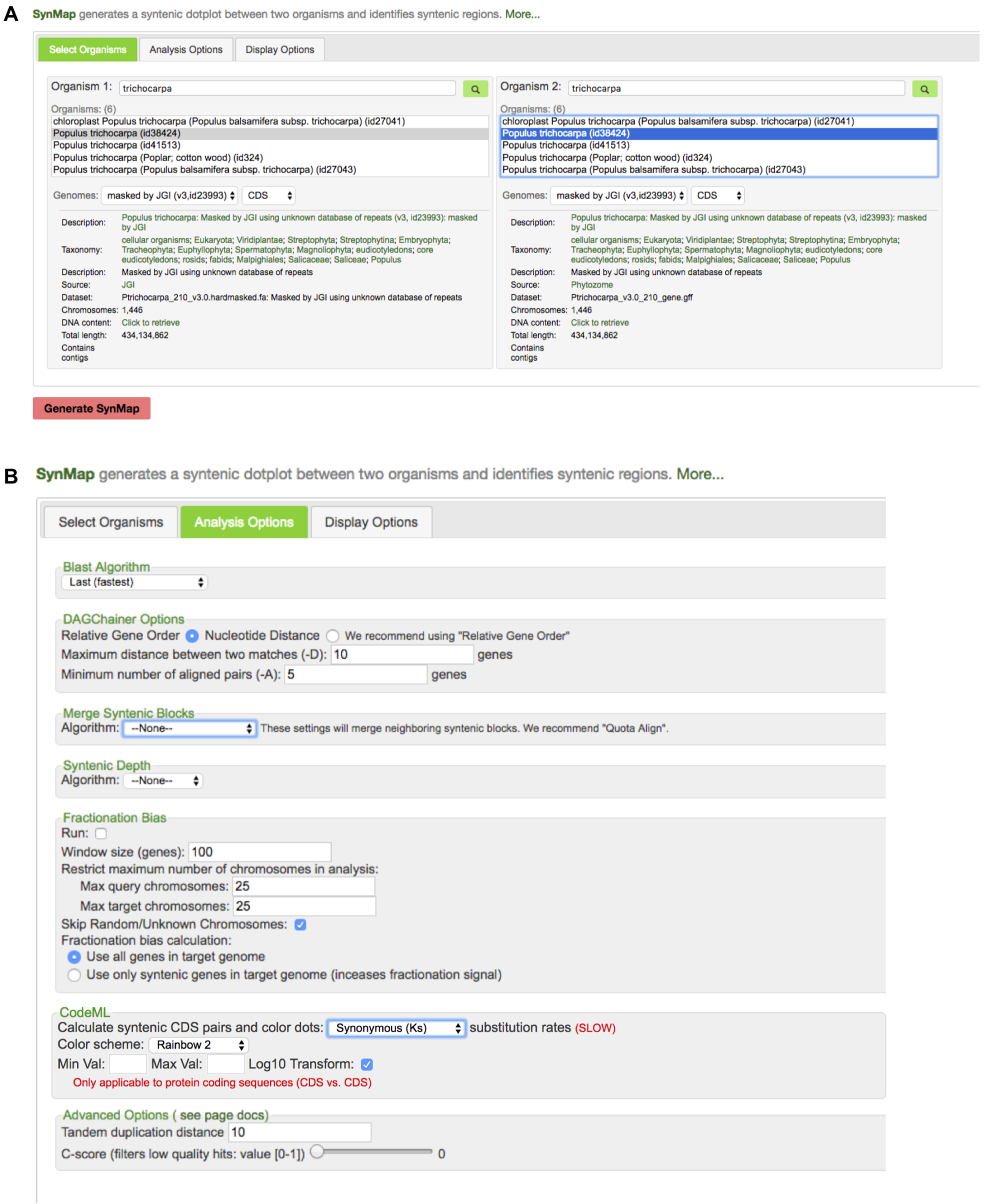
This research used resources of the Oak Ridge Leadership Computing Facility (OLCF) and the Compute and Data Environment for Science (CADES) at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE- AC05-00OR22725.

Support for the Poplar GWAS dataset is provided by the U.S. Department of Energy, Office of Science Biological and Environmental Research (BER) via the Bioenergy Science Center (BESC) under Contract No. DE-PS02-06ER64304. 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. Full Gene Atlas data sets are available at http://phytozome.jgi.doe.gov.

5.5 Supplementary Material

5.5.1 Supplementary Figures



SynMap generates a syntenic dotplot between two organisms and identifies syntenic regions. More...

Organism 1: trichocarpa

Organisms: (6)

chloroplast Populus trichocarpa (Populus balsamifera subsp. trichocarpa) (id27041)
Populus trichocarpa (id38424)
Populus trichocarpa (id41513)
Populus trichocarpa (Poplar; cotton wood) (id324)
Populus trichocarpa (Populus balsamifera subsp. trichocarpa) (id27043)

Genomes: masked by JGI (v3,id23993) CDS

Description:	Populus trichocarpa: Masked by JGI using unknown database of repeats (v3, id23993): masked by JGI
Taxonomy:	cellular organisms; Eukaryota; Viridiplantae; Streptophyta; Streptophytina; Embryophyta; Tracheophyta; Euphyllophyta; Spermatophyta; Magnoliophyta; eudicotyledons; core eudicotyledons; rosids; fabids; Malpighiales; Salicaceae; Saliceae; Populus
Description:	Masked by JGI using unknown database of repeats
Source:	JGI
Dataset:	Ptrichocarpa_210_v3.0.hardmasked.fa: Masked by JGI using unknown database of repeats
Chromosomes:	1,446
DNA content:	Click to retrieve
Total length:	434,134,862
Contains contigs	

Organism 2: trichocarpa

Organisms: (6)

chloroplast Populus trichocarpa (Populus balsamifera subsp. trichocarpa) (id27041)
Populus trichocarpa (id38424)
Populus trichocarpa (id41513)
Populus trichocarpa (Poplar; cotton wood) (id324)
Populus trichocarpa (Populus balsamifera subsp. trichocarpa) (id27043)

Genomes: masked by JGI (v3,id23993) CDS

Description:	Populus trichocarpa: Masked by JGI using unknown database of repeats (v3, id23993): masked by JGI
Taxonomy:	cellular organisms; Eukaryota; Viridiplantae; Streptophyta; Streptophytina; Embryophyta; Tracheophyta; Euphyllophyta; Spermatophyta; Magnoliophyta; eudicotyledons; core eudicotyledons; rosids; fabids; Malpighiales; Salicaceae; Saliceae; Populus
Description:	Masked by JGI using unknown database of repeats
Source:	Phytozome
Dataset:	Ptrichocarpa_v3.0_210_gene.gff
Chromosomes:	1,446
DNA content:	Click to retrieve
Total length:	434,134,862
Contains contigs	

SynMap generates a syntenic dotplot between two organisms and identifies syntenic regions. More...

Blast Algorithm
Last (fastest)

DAGChainer Options
Relative Gene Order ● Nucleotide Distance ○ We recommend using "Relative Gene Order"
Maximum distance between two matches (-D): 10 genes
Minimum number of aligned pairs (-A): 5 genes

Merge Syntenic Blocks
Algorithm: --None-- These settings will merge neighboring syntenic blocks. We recommend "Quota Align".

Syntenic Depth
Algorithm: --None--

Fractionation Bias
Run: □
Window size (genes): 100
Restrict maximum number of chromosomes in analysis:
Max query chromosomes: 25
Max target chromosomes: 25
Skip Random/Unknown Chromosomes: ☑
Fractionation bias calculation:
● Use all genes in target genome
○ Use only syntenic genes in target genome (inceases fractionation signal)

CodeML
Calculate syntenic CDS pairs and color dots: Synonymous (Ks) substitution rates (SLOW)
Color scheme: Rainbow 2
Min Val: Max Val: Log10 Transform: ☑
Only applicable to protein coding sequences (CDS vs. CDS)

Advanced Options (see page docs)
Tandem duplication distance 10
C-score (filters low quality hits: value [0-1]) 0

Figure S5.1: **Settings for Coge.** Options used for running CoGe to generate syntenic blocks.

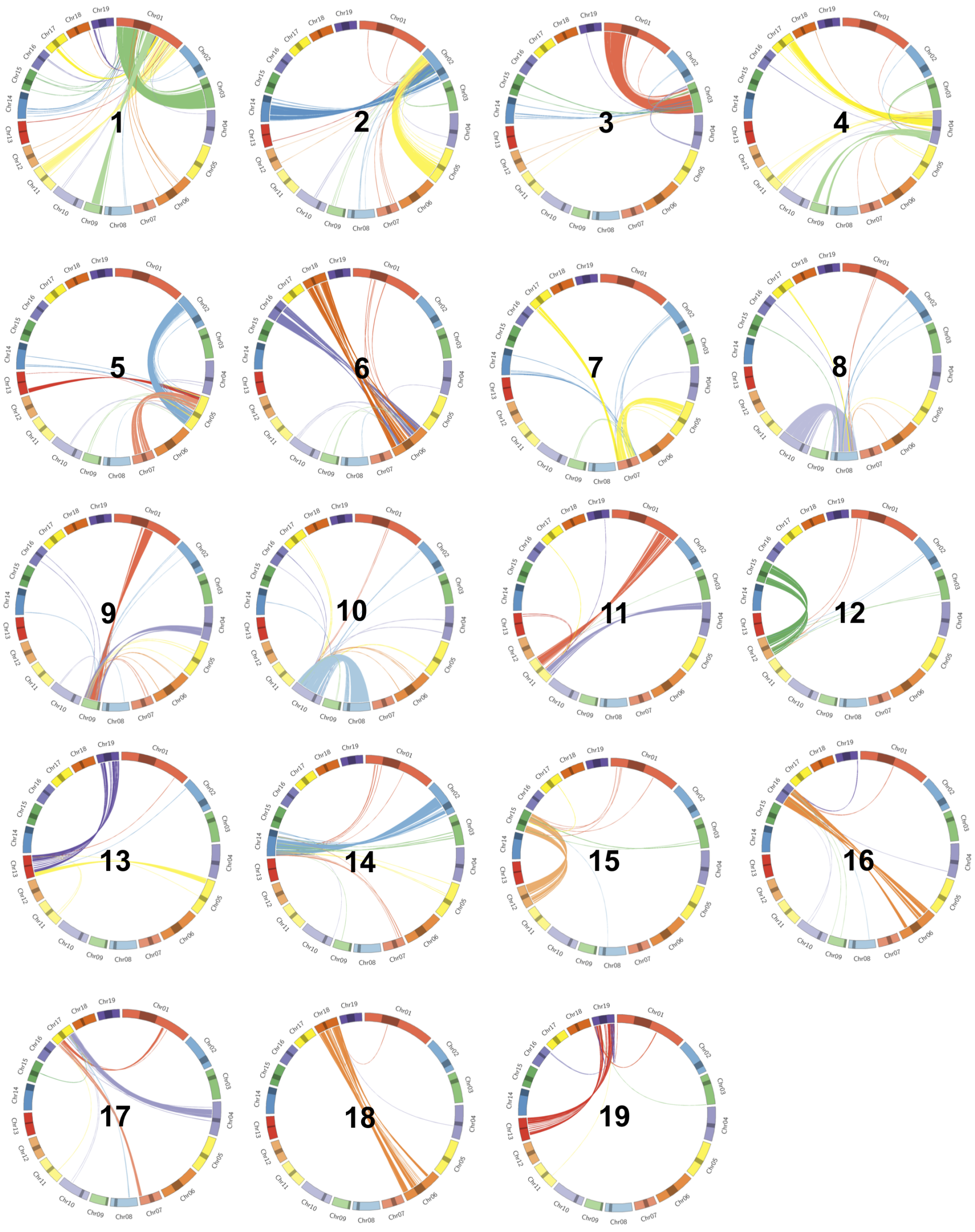


Figure S5.2: **Syntenic blocks.** Syntenic blocks for each chromosome.

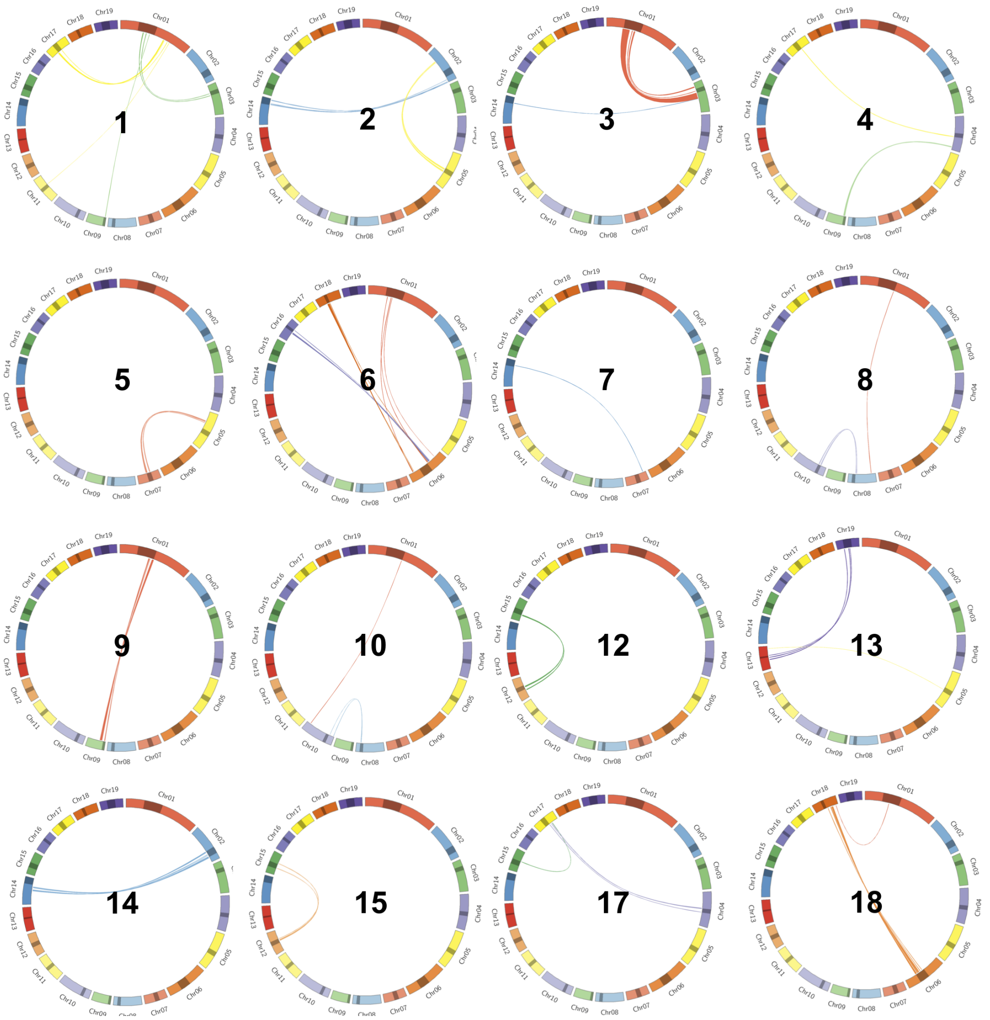


Figure S5.3: **Syntenic blocks overlapping centromeric regions.** Syntenic blocks for each chromosome that arise from a centromeric region on the source chromosome.

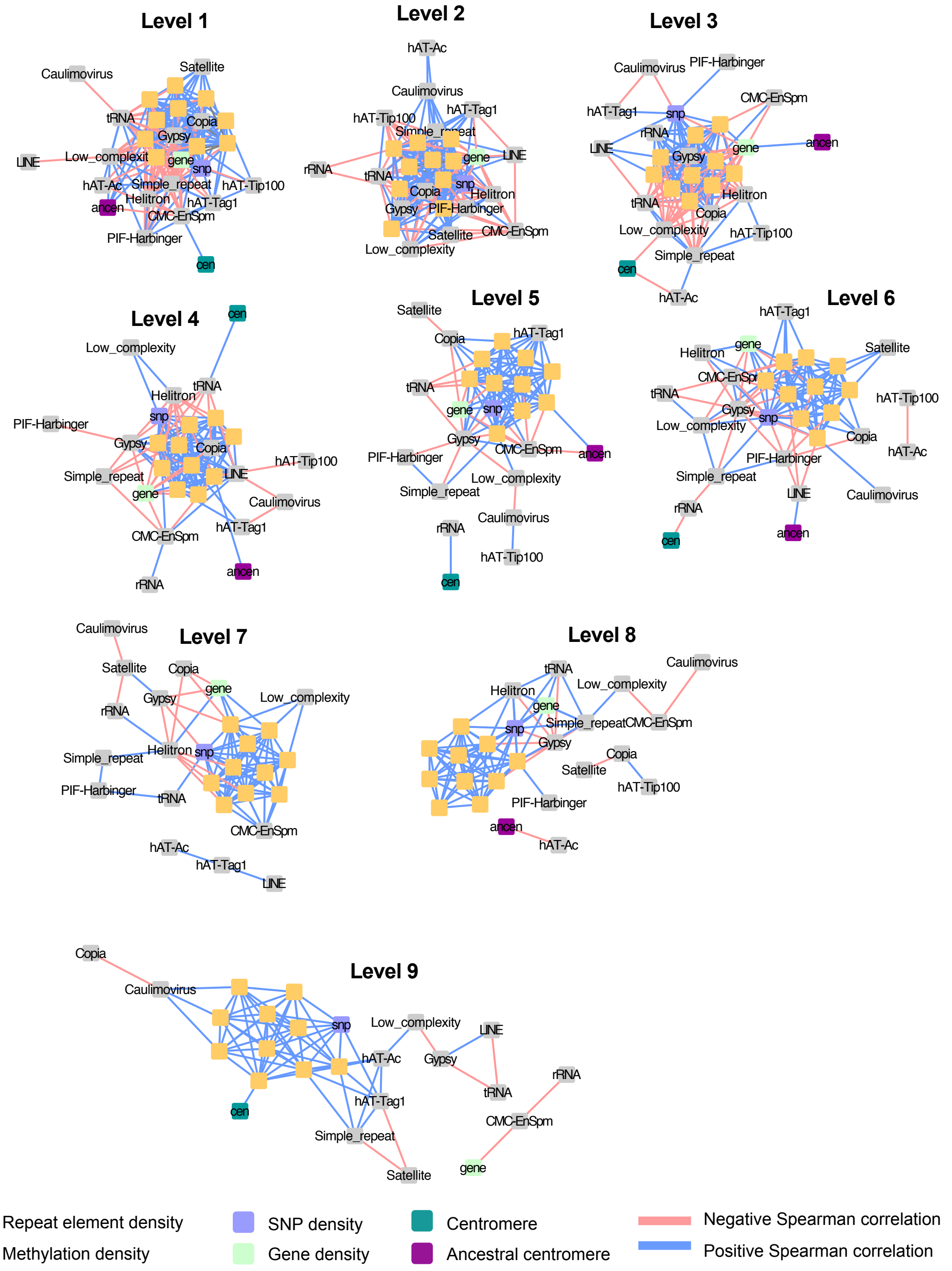


Figure S5.4: **DWT correlation networks for chromosome 1.** Discrete wavelet transform correlation networks representing the significant ($p < 0.05$) Spearman correlations between wavelet coefficients of different features, at each scale on chromosome 1. Each of (A) through (F) represents correlations of a specific scale, with (A) representing the smallest scale and (H) representing the largest scale. Each node represents a genomic feature, and each edge represents a significant Spearman correlation between two genomic features at a particular scale.

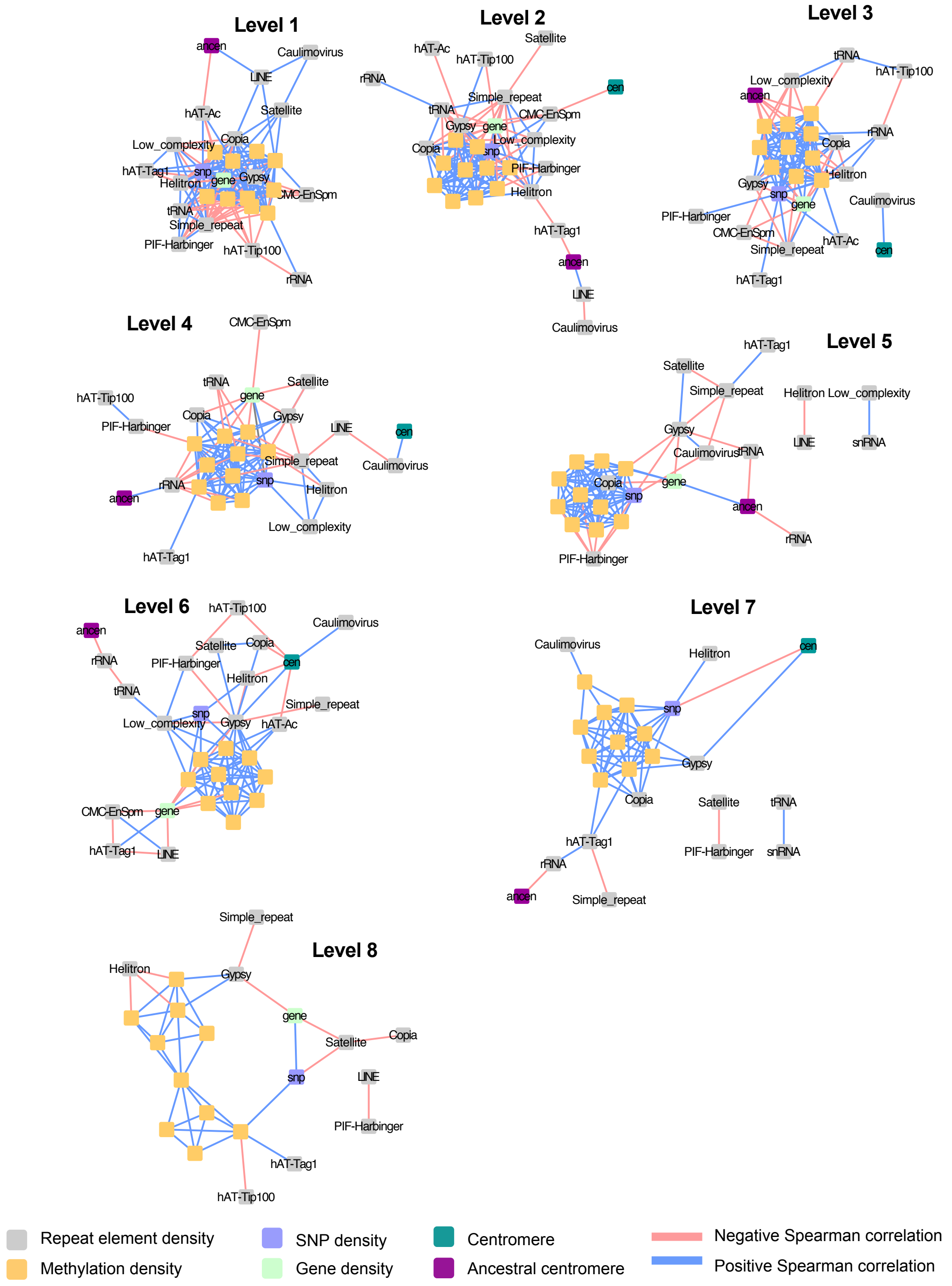


Figure S5.5: **DWT correlation networks for chromosome 2.** Discrete wavelet transform correlation networks representing the significant ($p < 0.05$) Spearman correlations between wavelet coefficients of different features, at each scale on chromosome 2. Each of (A) through (F) represents correlations of a specific scale, with (A) representing the smallest scale and (H) representing the largest scale. Each node represents a genomic feature, and each edge represents a significant Spearman correlation between two genomic features at a particular scale.

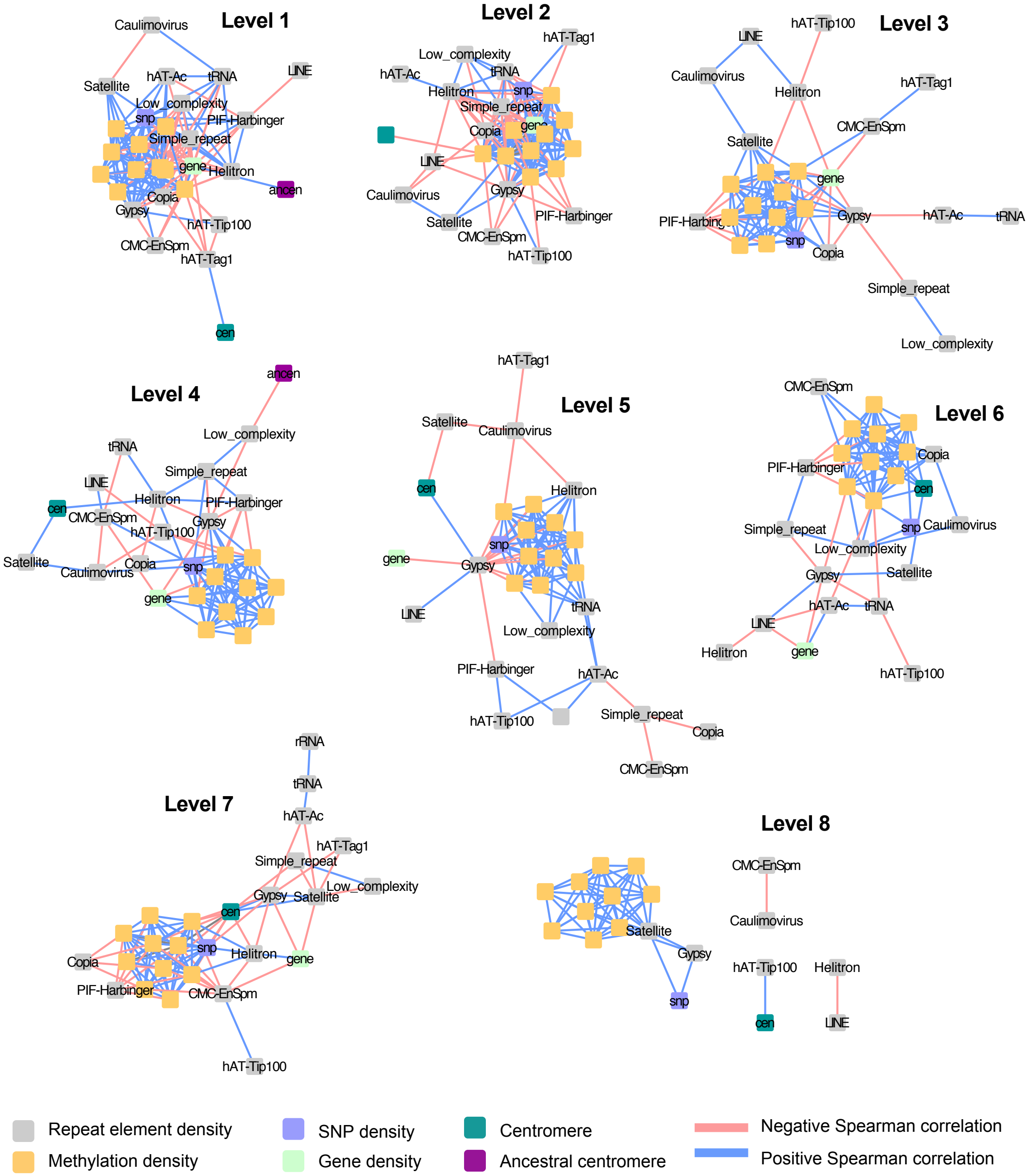


Figure S5.6: **DWT correlation networks for chromosome 3.** Discrete wavelet transform correlation networks representing the significant ($p < 0.05$) Spearman correlations between wavelet coefficients of different features, at each scale on chromosome 3. Each of (A) through (F) represents correlations of a specific scale, with (A) representing the smallest scale and (H) representing the largest scale. Each node represents a genomic feature, and each edge represents a significant Spearman correlation between two genomic features at a particular scale.

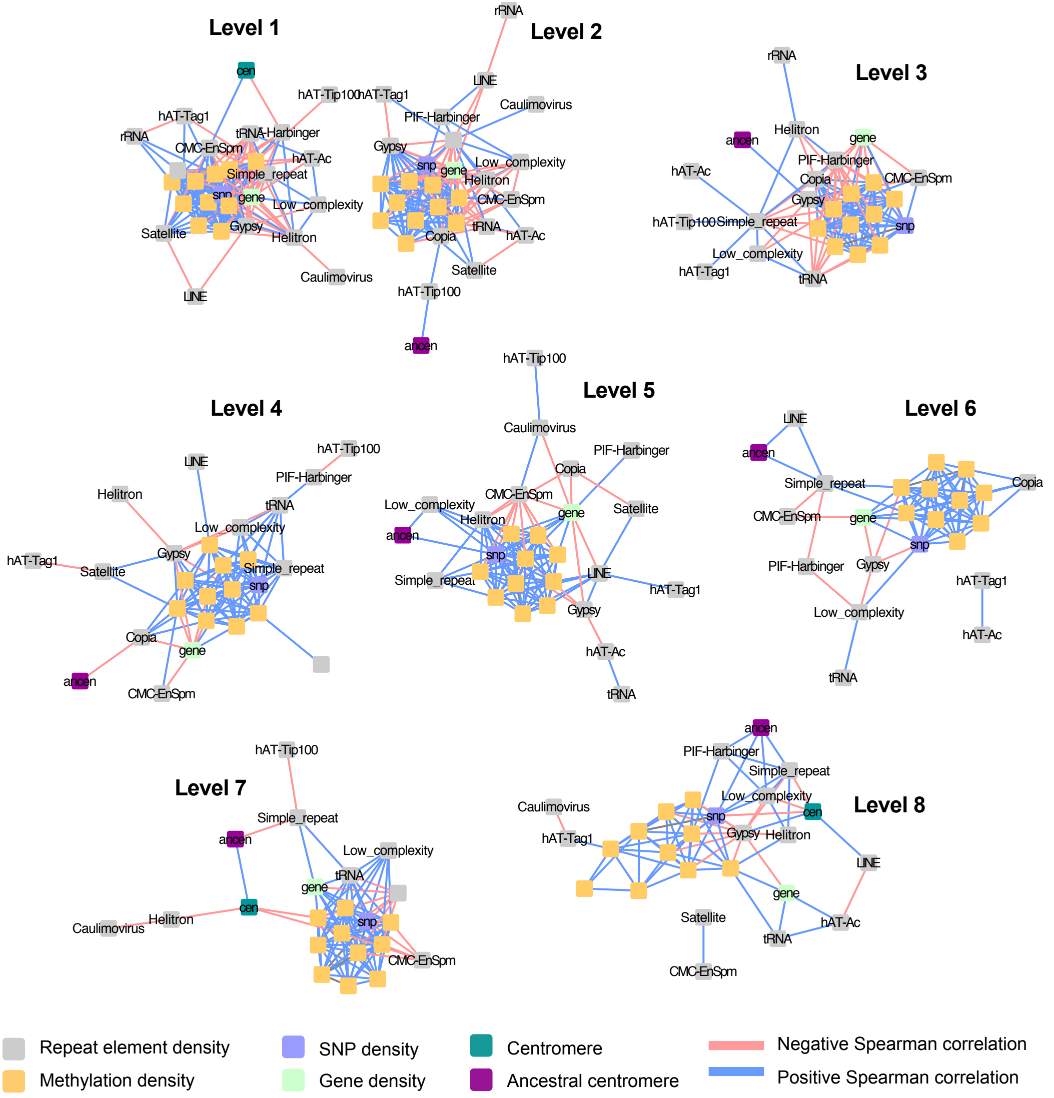


Figure S5.7: **DWT correlation networks for chromosome 4.** Discrete wavelet transform correlation networks representing the significant ($p < 0.05$) Spearman correlations between wavelet coefficients of different features, at each scale on chromosome 4. Each of (A) through (F) represents correlations of a specific scale, with (A) representing the smallest scale and (H) representing the largest scale. Each node represents a genomic feature, and each edge represents a significant Spearman correlation between two genomic features at a particular scale.

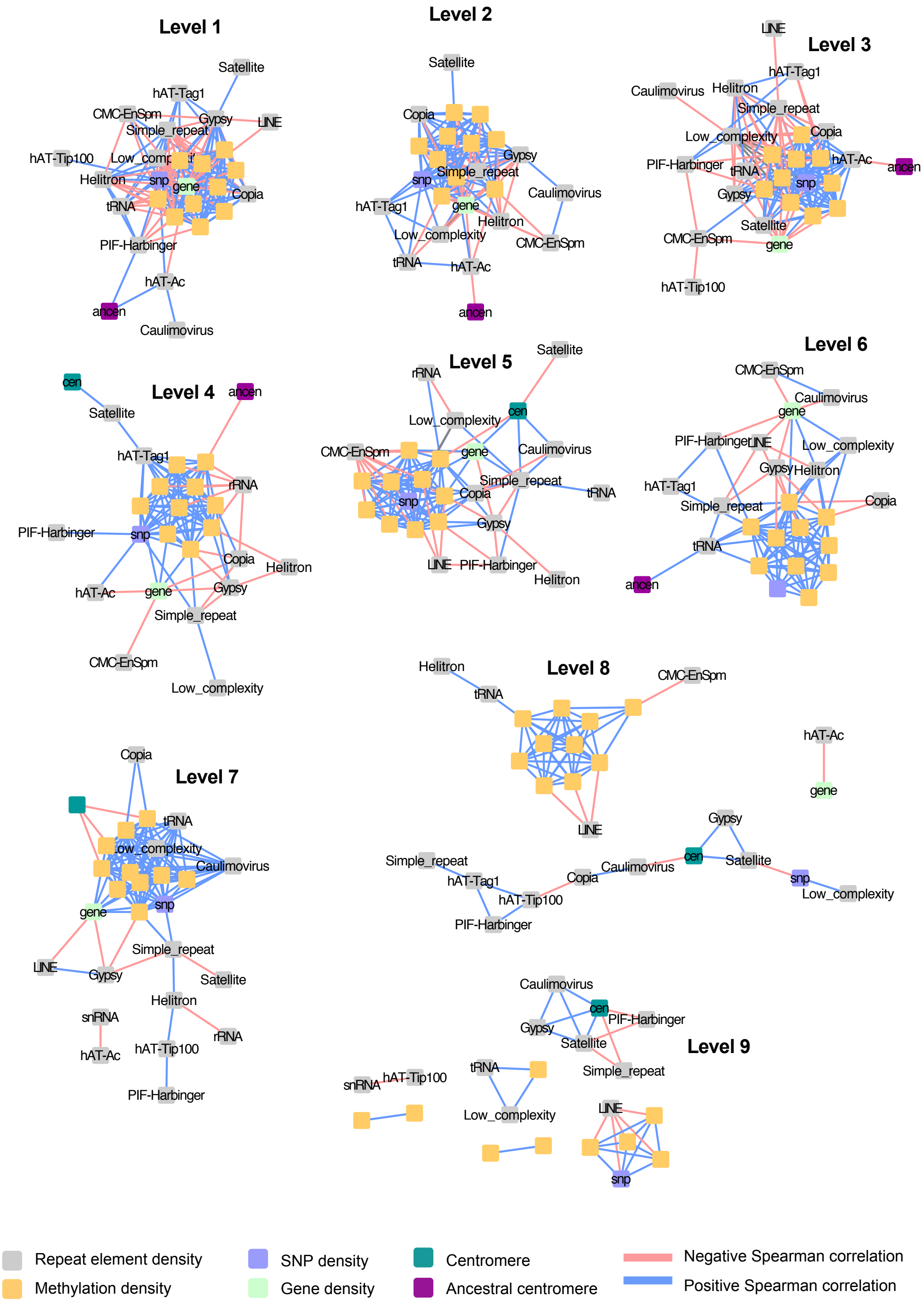


Figure S5.8: **DWT correlation networks for chromosome 5.** Discrete wavelet transform correlation networks representing the significant ($p < 0.05$) Spearman correlations between wavelet coefficients of different features, at each scale on chromosome 5. Each of (A) through (F) represents correlations of a specific scale, with (A) representing the smallest scale and (H) representing the largest scale. Each node represents a genomic feature, and each edge represents a significant Spearman correlation between two genomic features at a particular scale.

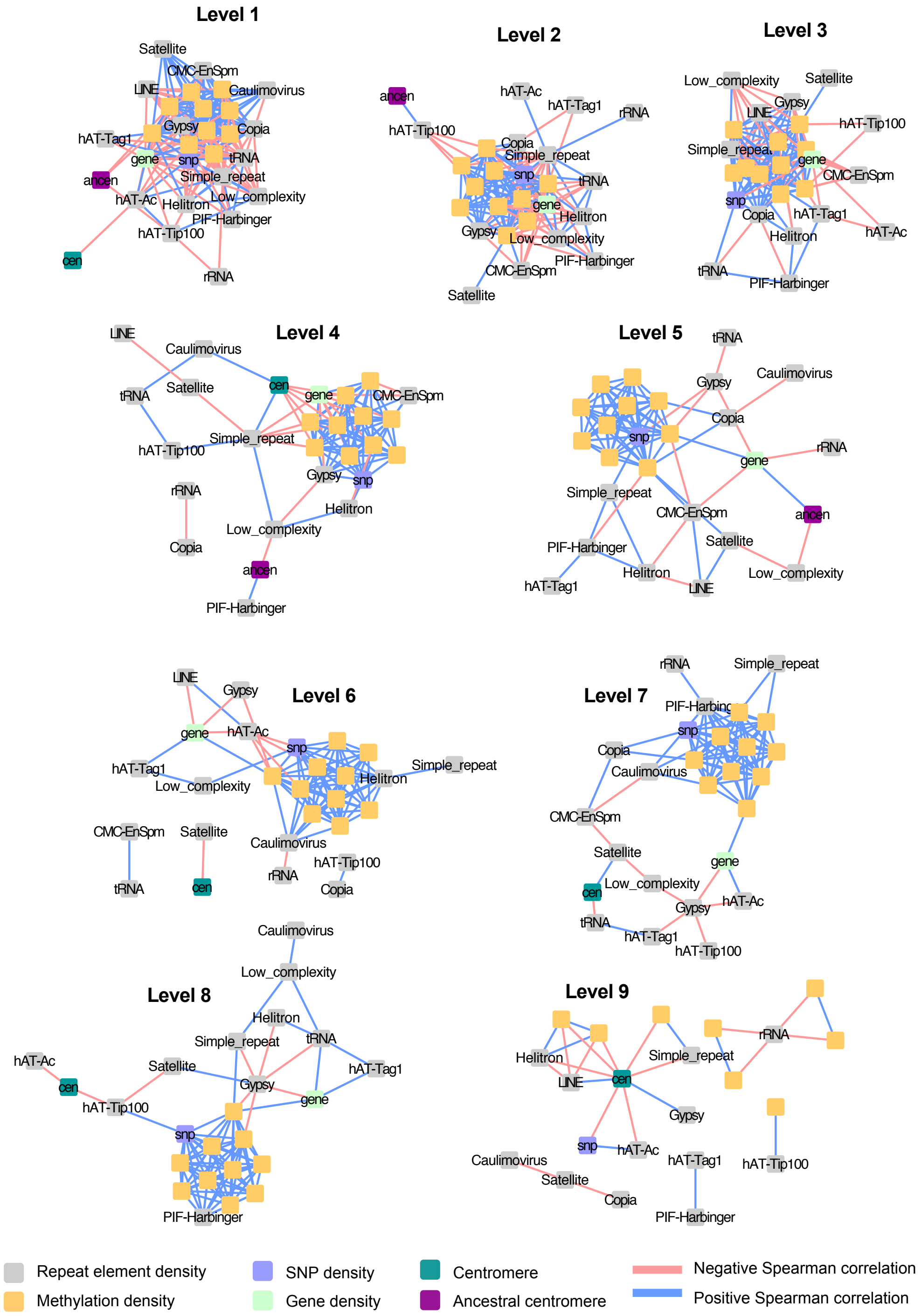


Figure S5.9: **DWT correlation networks for chromosome 6.** Discrete wavelet transform correlation networks representing the significant ($p < 0.05$) Spearman correlations between wavelet coefficients of different features, at each scale on chromosome 6. Each of (A) through (F) represents correlations of a specific scale, with (A) representing the smallest scale and (H) representing the largest scale. Each node represents a genomic feature, and each edge represents a significant Spearman correlation between two genomic features at a particular scale.

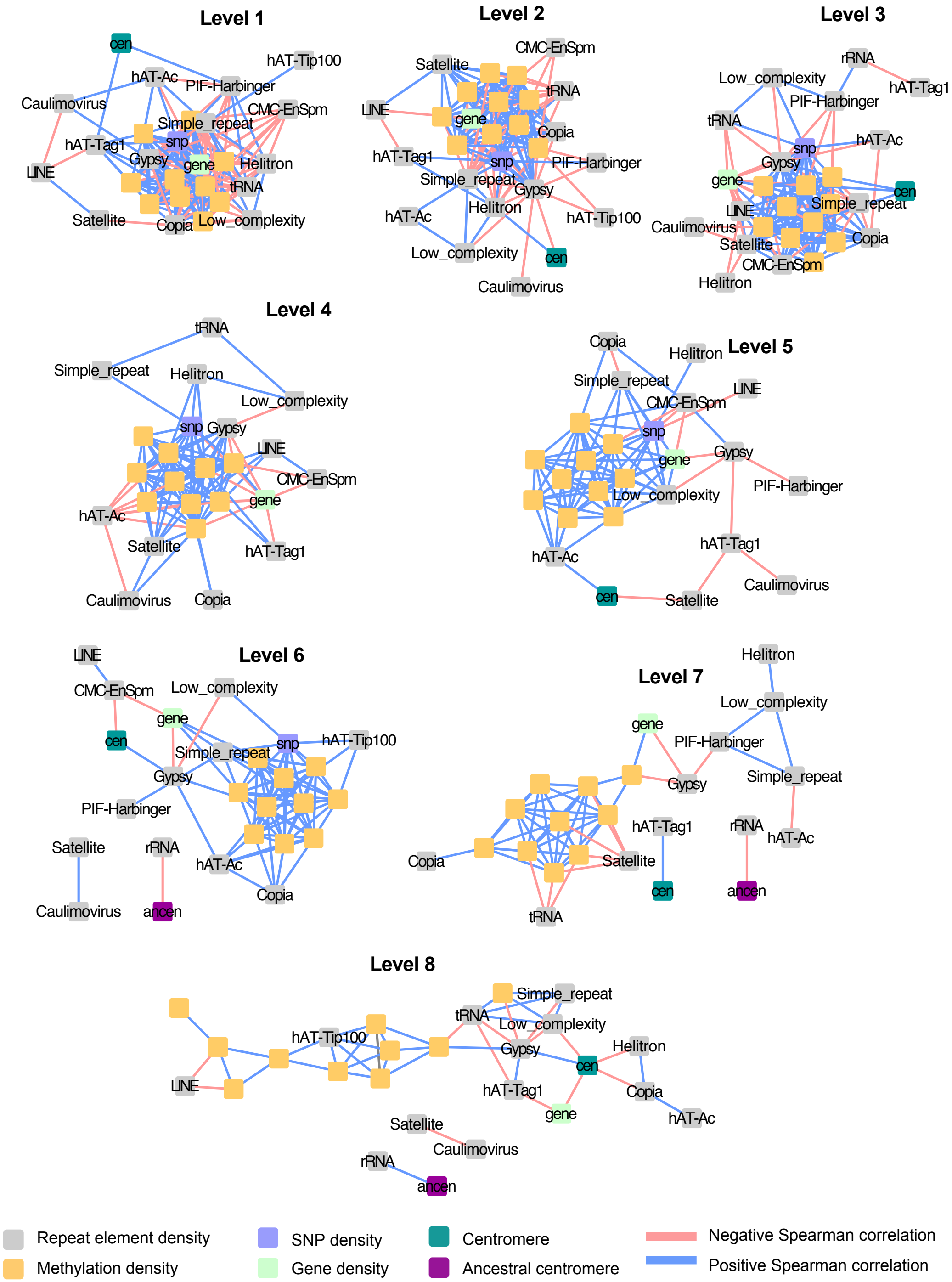


Figure S5.10: **DWT correlation networks for chromosome 7.** Discrete wavelet transform correlation networks representing the significant ($p < 0.05$) Spearman correlations between wavelet coefficients of different features, at each scale on chromosome 7. Each of (A) through (F) represents correlations of a specific scale, with (A) representing the smallest scale and (H) representing the largest scale. Each node represents a genomic feature, and each edge represents a significant Spearman correlation between two genomic features at a particular scale.

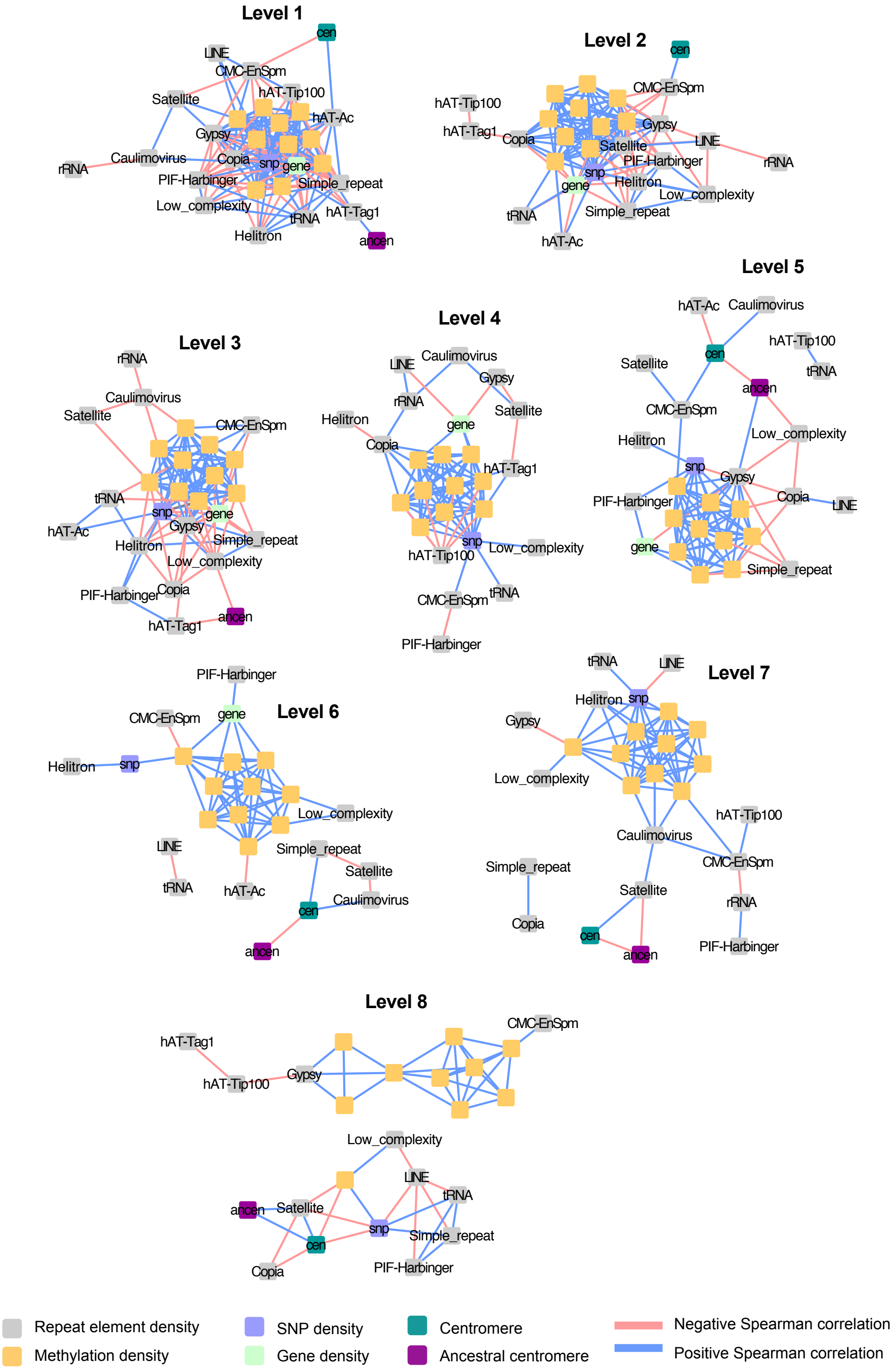


Figure S5.11: **DWT correlation networks for chromosome 8.** Discrete wavelet transform correlation networks representing the significant ($p < 0.05$) Spearman correlations between wavelet coefficients of different features, at each scale on chromosome 8. Each of (A) through (F) represents correlations of a specific scale, with (A) representing the smallest scale and (H) representing the largest scale. Each node represents a genomic feature, and each edge represents a significant Spearman correlation between two genomic features at a particular scale.

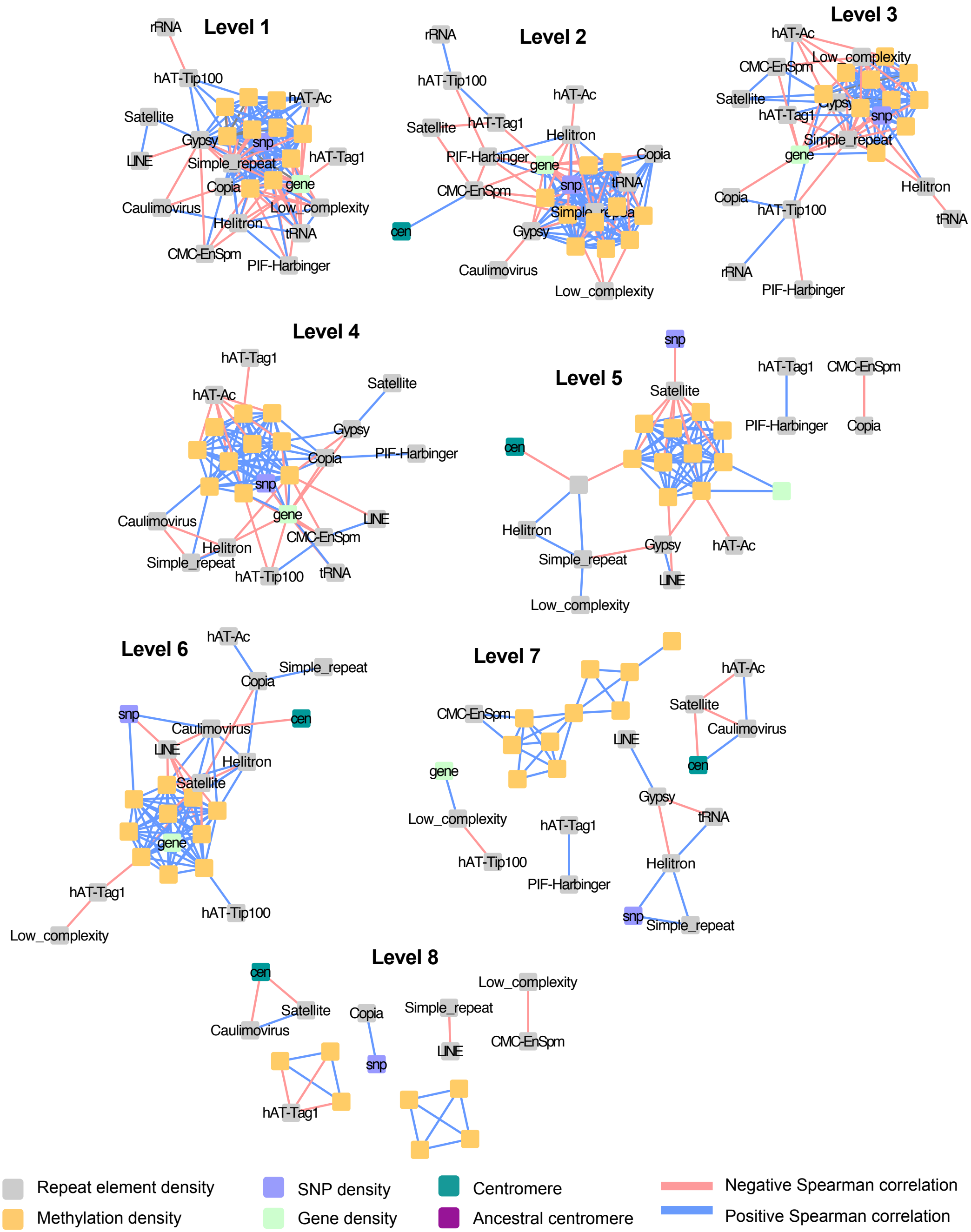


Figure S5.12: **DWT correlation networks for chromosome 9.** Discrete wavelet transform correlation networks representing the significant ($p < 0.05$) Spearman correlations between wavelet coefficients of different features, at each scale on chromosome 9. Each of (A) through (F) represents correlations of a specific scale, with (A) representing the smallest scale and (H) representing the largest scale. Each node represents a genomic feature, and each edge represents a significant Spearman correlation between two genomic features at a particular scale.

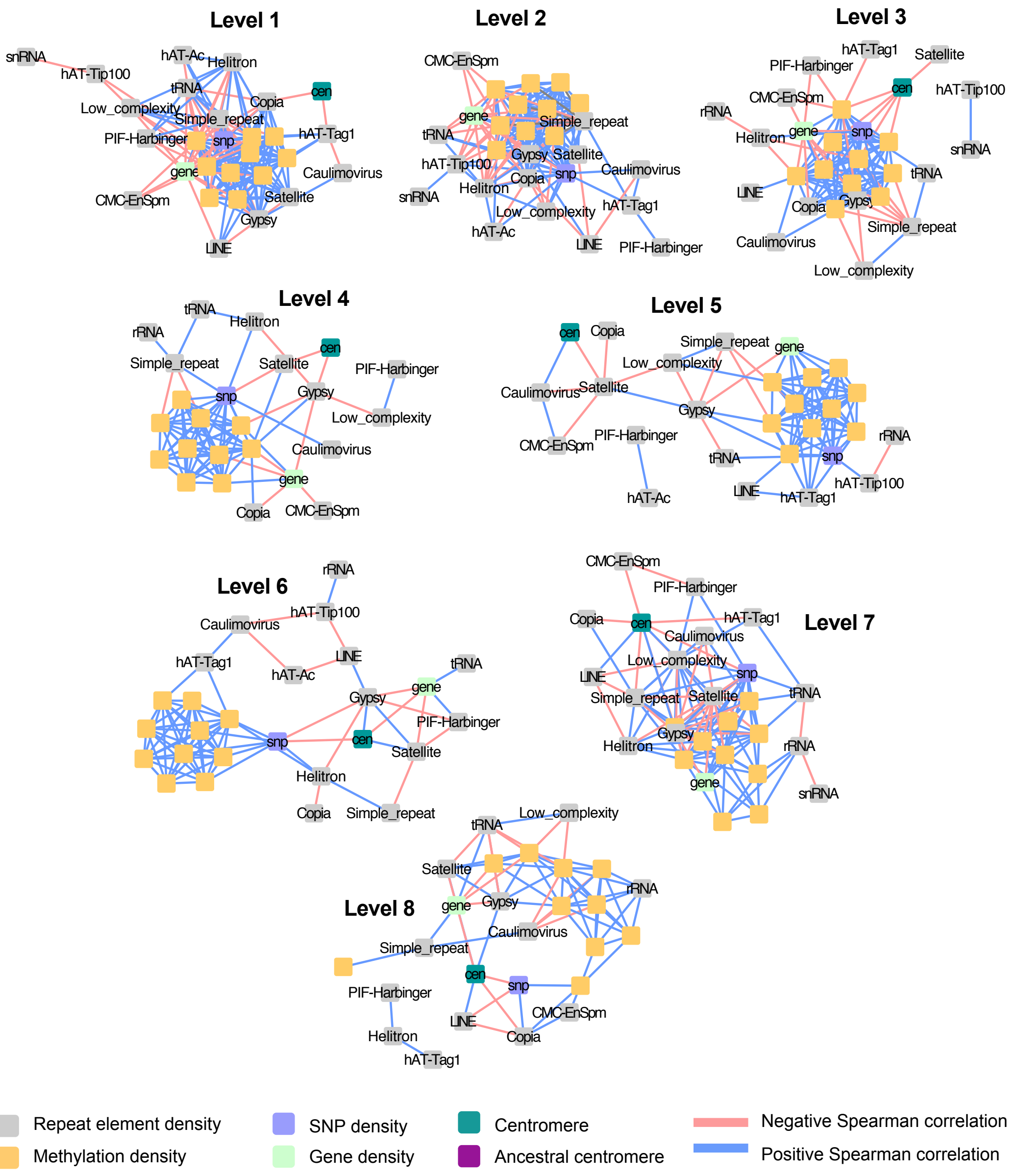


Figure S5.13: **DWT correlation networks for chromosome 10.** Discrete wavelet transform correlation networks representing the significant ($p < 0.05$) Spearman correlations between wavelet coefficients of different features, at each scale on chromosome 10. Each of (A) through (F) represents correlations of a specific scale, with (A) representing the smallest scale and (H) representing the largest scale. Each node represents a genomic feature, and each edge represents a significant Spearman correlation between two genomic features at a particular scale.

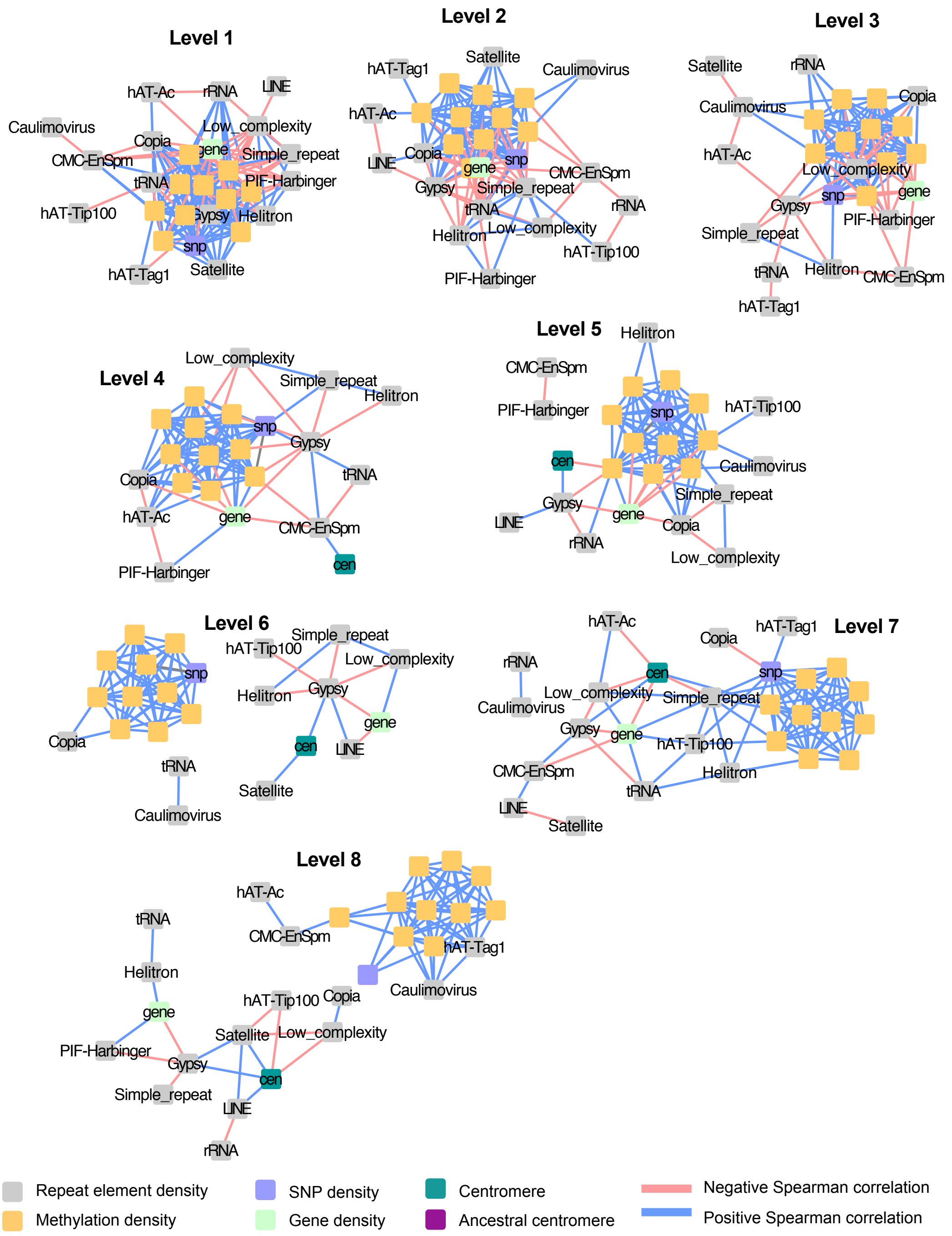


Figure S5.14: **DWT correlation networks for chromosome 11.** Discrete wavelet transform correlation networks representing the significant ($p < 0.05$) Spearman correlations between wavelet coefficients of different features, at each scale on chromosome 11. Each of (A) through (F) represents correlations of a specific scale, with (A) representing the smallest scale and (H) representing the largest scale. Each node represents a genomic feature, and each edge represents a significant Spearman correlation between two genomic features at a particular scale.

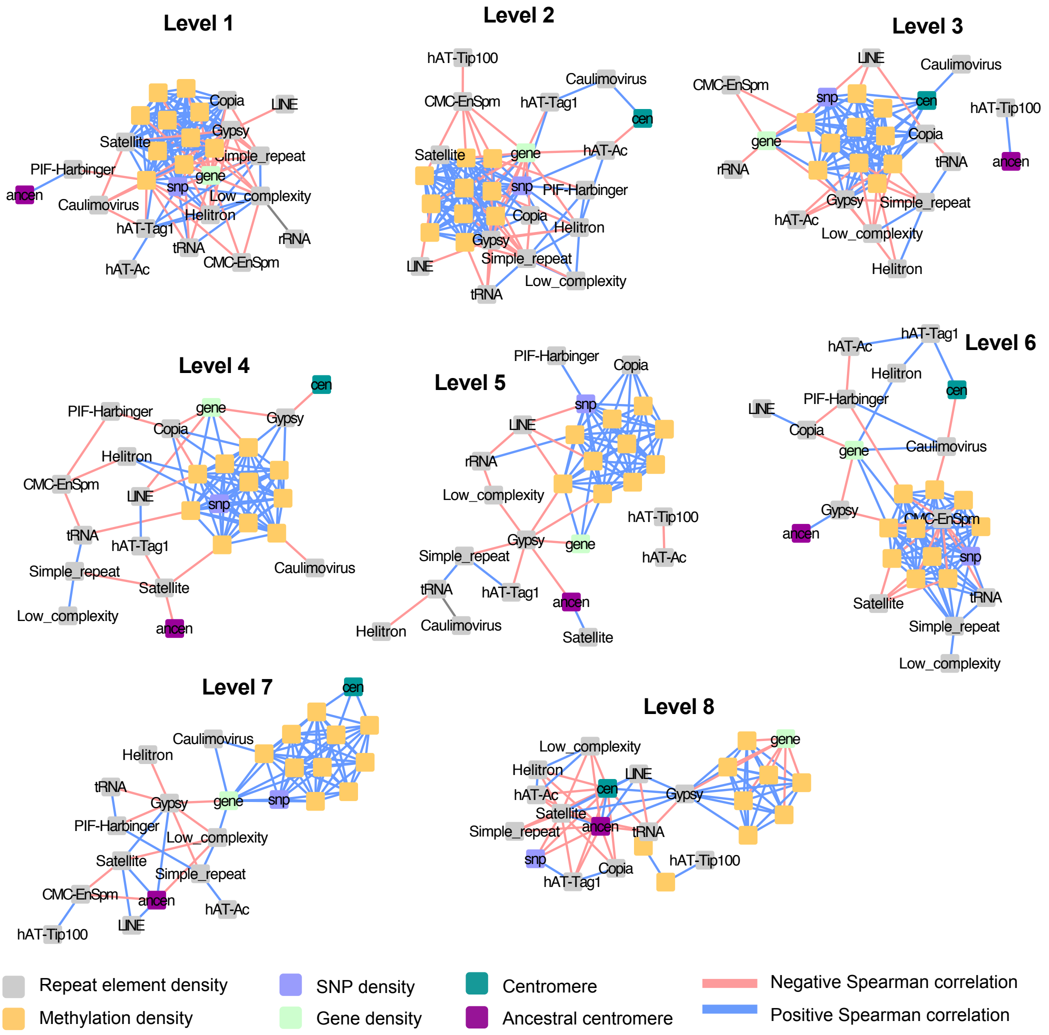


Figure S5.15: **DWT correlation networks for chromosome 12.** Discrete wavelet transform correlation networks representing the significant ($p < 0.05$) Spearman correlations between wavelet coefficients of different features, at each scale on chromosome 12. Each of (A) through (F) represents correlations of a specific scale, with (A) representing the smallest scale and (H) representing the largest scale. Each node represents a genomic feature, and each edge represents a significant Spearman correlation between two genomic features at a particular scale.

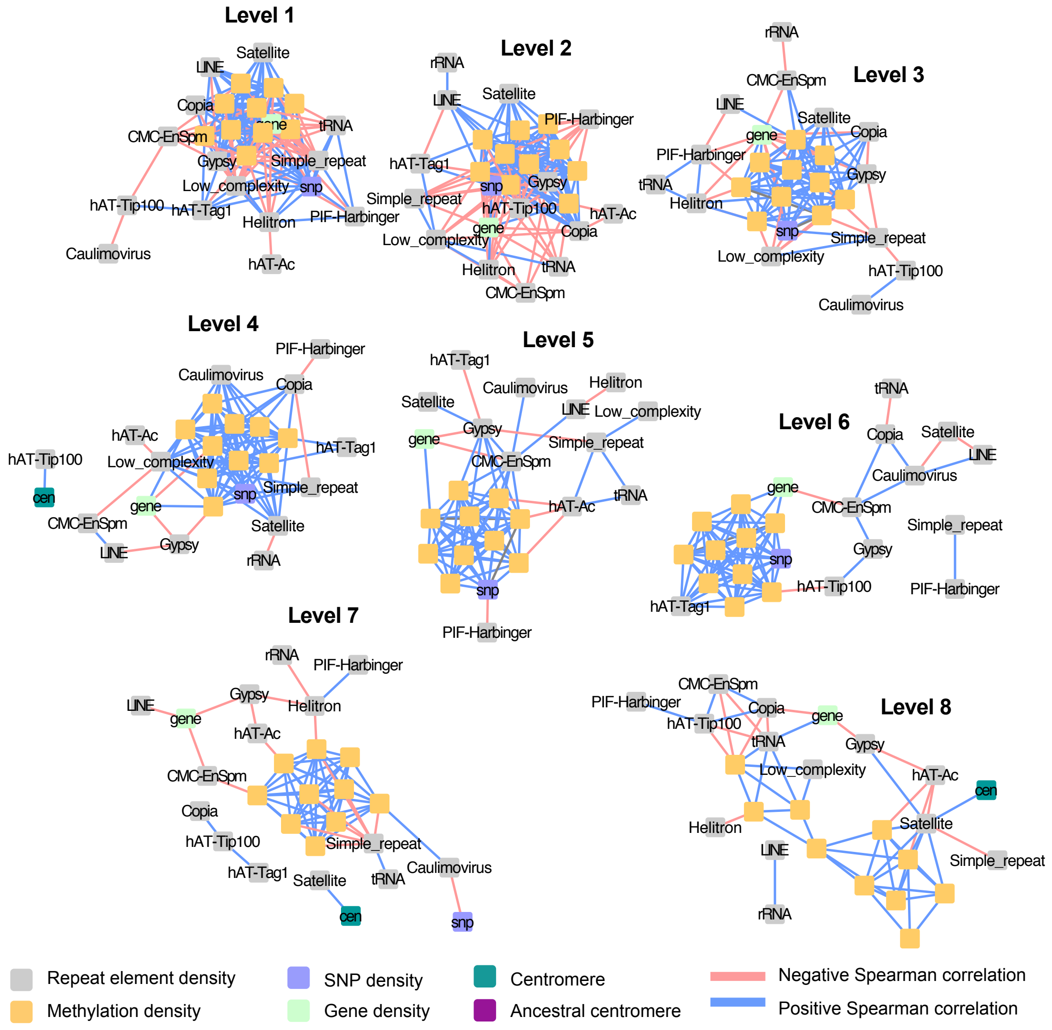


Figure S5.16: **DWT correlation networks for chromosome 14.** Discrete wavelet transform correlation networks representing the significant ($p < 0.05$) Spearman correlations between wavelet coefficients of different features, at each scale on chromosome 14. Each of (A) through (F) represents correlations of a specific scale, with (A) representing the smallest scale and (H) representing the largest scale. Each node represents a genomic feature, and each edge represents a significant Spearman correlation between two genomic features at a particular scale.

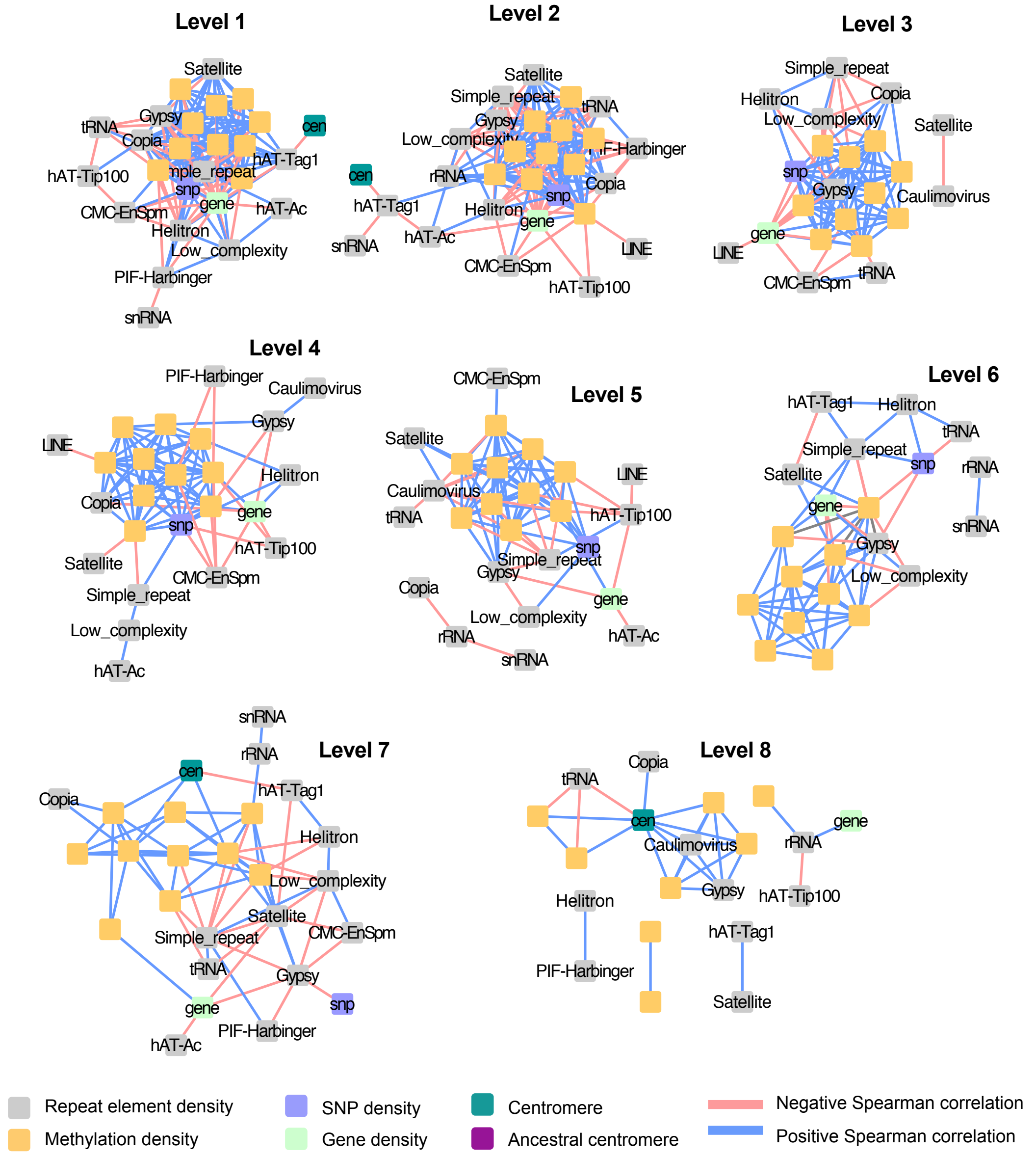


Figure S5.17: **DWT correlation networks for chromosome 15.** Discrete wavelet transform correlation networks representing the significant ($p < 0.05$) Spearman correlations between wavelet coefficients of different features, at each scale on chromosome 15. Each of (A) through (F) represents correlations of a specific scale, with (A) representing the smallest scale and (H) representing the largest scale. Each node represents a genomic feature, and each edge represents a significant Spearman correlation between two genomic features at a particular scale.

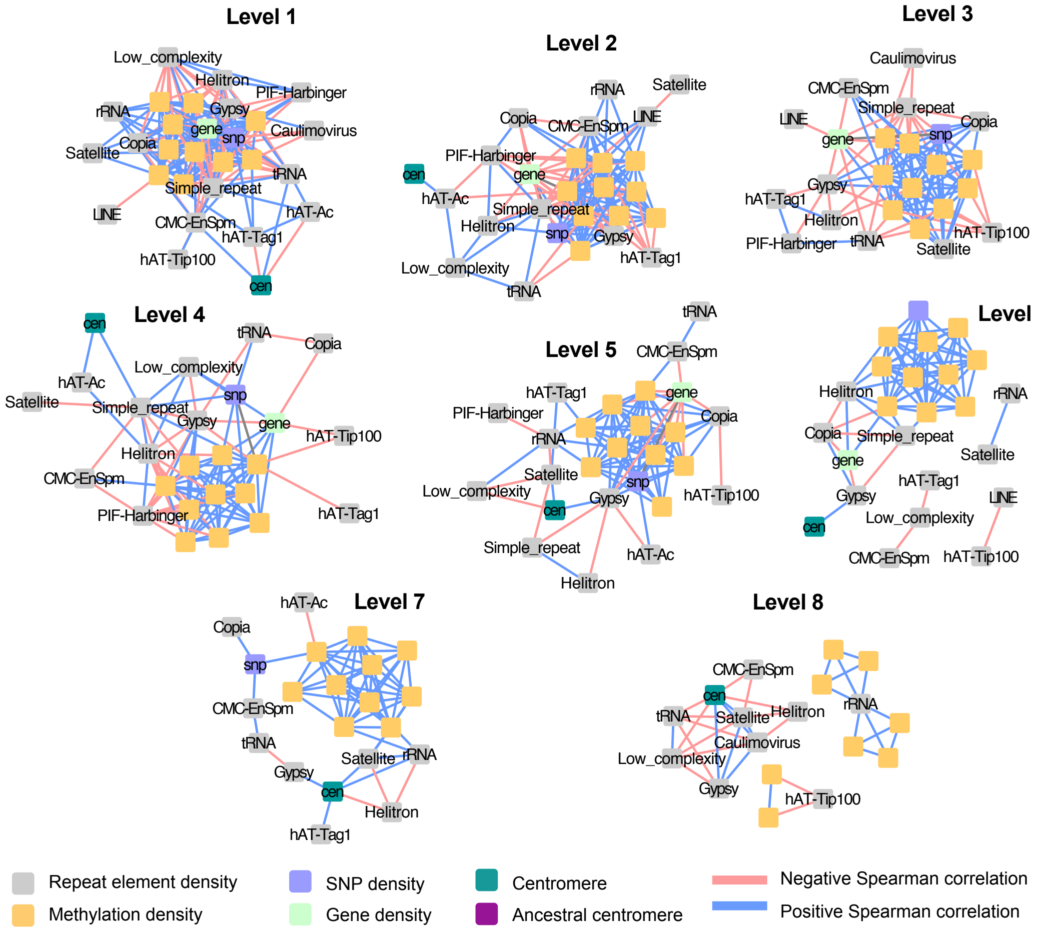


Figure S5.18: **DWT correlation networks for chromosome 16.** Discrete wavelet transform correlation networks representing the significant ($p < 0.05$) Spearman correlations between wavelet coefficients of different features, at each scale on chromosome 16. Each of (A) through (F) represents correlations of a specific scale, with (A) representing the smallest scale and (H) representing the largest scale. Each node represents a genomic feature, and each edge represents a significant Spearman correlation between two genomic features at a particular scale.

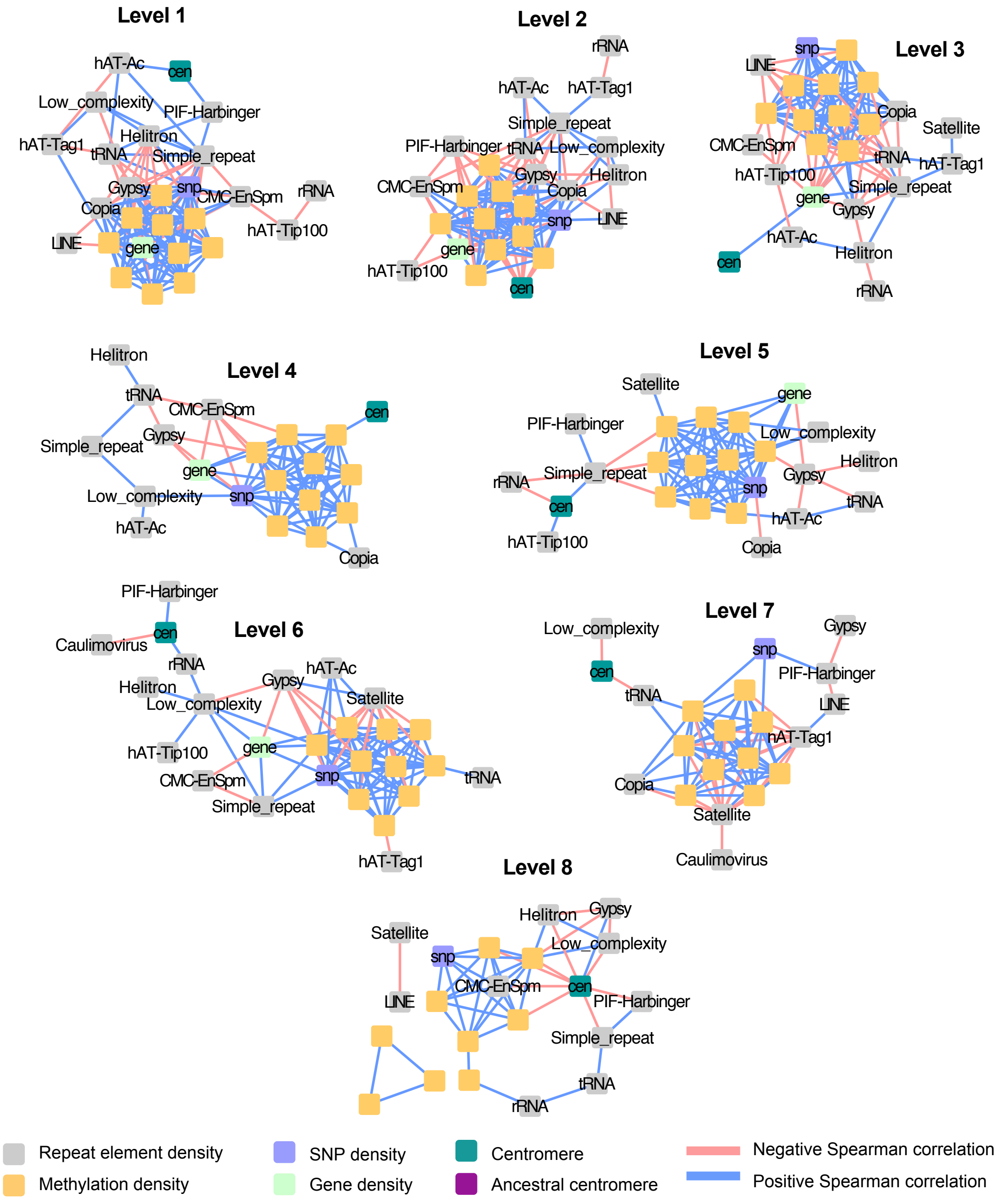


Figure S5.19: **DWT correlation networks for chromosome 17.** Discrete wavelet transform correlation networks representing the significant (p < 0.05) Spearman correlations between wavelet coefficients of different features, at each scale on chromosome 17. Each of (A) through (F) represents correlations of a specific scale, with (A) representing the smallest scale and (H) representing the largest scale. Each node represents a genomic feature, and each edge represents a significant Spearman correlation between two genomic features at a particular scale.

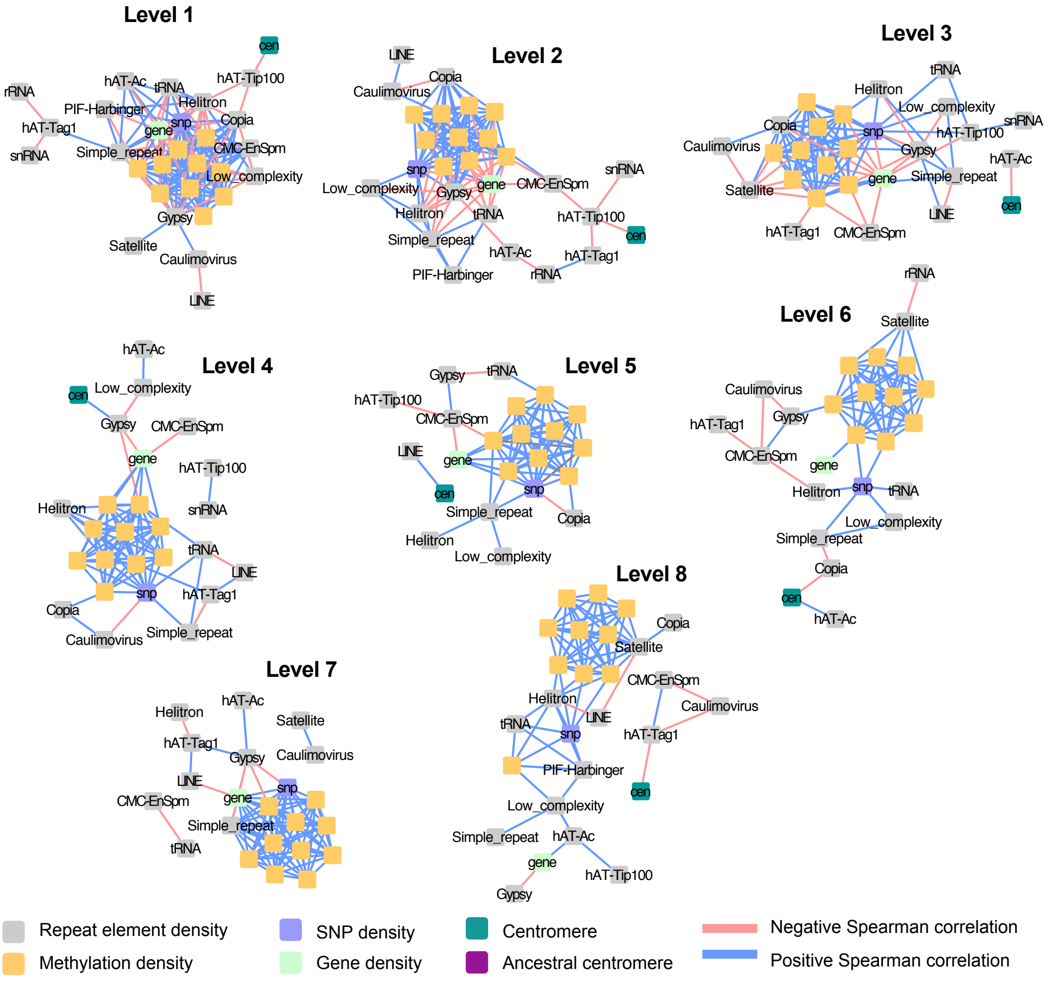


Figure S5.20: **DWT correlation networks for chromosome 18.** Discrete wavelet transform correlation networks representing the significant ($p < 0.05$) Spearman correlations between wavelet coefficients of different features, at each scale on chromosome 18. Each of (A) through (F) represents correlations of a specific scale, with (A) representing the smallest scale and (H) representing the largest scale. Each node represents a genomic feature, and each edge represents a significant Spearman correlation between two genomic features at a particular scale.

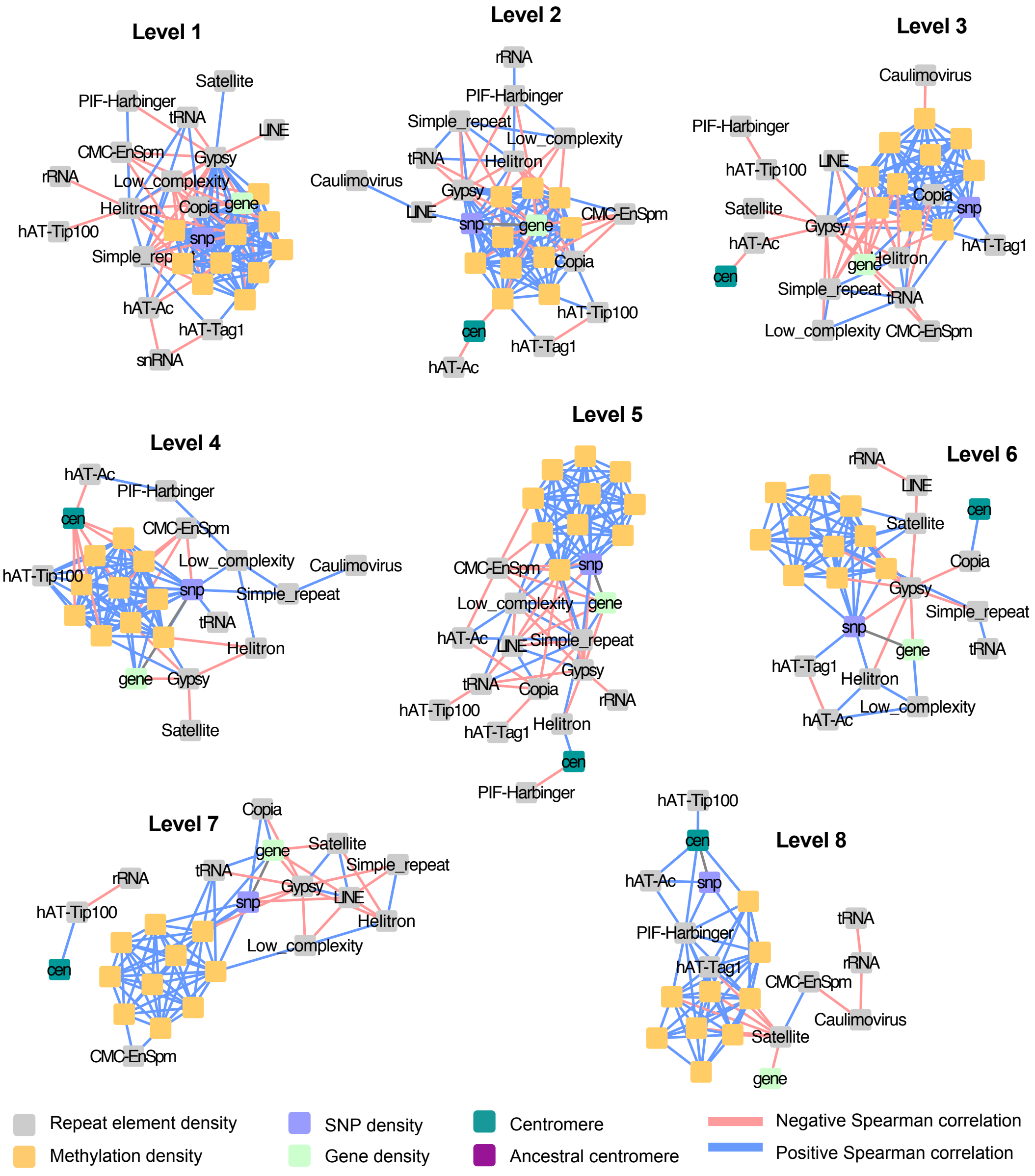


Figure S5.21: **DWT correlation networks for chromosome 19.** Discrete wavelet transform correlation networks representing the significant ($p < 0.05$) Spearman correlations between wavelet coefficients of different features, at each scale on chromosome 19. Each of (A) through (F) represents correlations of a specific scale, with (A) representing the smallest scale and (H) representing the largest scale. Each node represents a genomic feature, and each edge represents a significant Spearman correlation between two genomic features at a particular scale.

Bibliography

[1] Gerald A Tuskan, S Difazio, Stefan Jansson, J Bohlmann, I Grigoriev, U Hellsten, N Putnam, S Ralph, Stephane Rombauts, A Salamov, et al. The Genome of Black Cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science*, 313(5793):1596–1604, 2006. xv, 273, 276, 280, 283, 298

[2] Damon Lisch. Epigenetic regulation of transposable elements in plants. *Annual review of plant biology*, 60:43–66, 2009. 273, 296

[3] Jeffrey L Bennetzen and Hao Wang. The contributions of transposable elements to the structure, function, and evolution of plant genomes. *Annual review of plant biology*, 65:505–530, 2014. 273, 297

[4] Chris CA Spencer, Panos Deloukas, Sarah Hunt, Jim Mullikin, Simon Myers, Bernard Silverman, Peter Donnelly, David Bentley, and Gil McVean. The Influence of Recombination on Human Genetic Diversity. *PLoS Genet*, 2(9):e148, 2006. 274, 287

[5] CM Leavey, MN James, J Summerscales, and R Sutton. An introduction to wavelet transforms: a tutorial approach. *Insight-Non-Destructive Testing and Condition Monitoring*, 45(5):344–353, 2003. 274, 291

[6] Gancho T Slavov, Stephen P DiFazio, Joel Martin, Wendy Schackwitz, Wellington Muchero, Eli Rodgers-Melnick, Mindie F Lipphardt, Christa P Pennacchio, Uffe Hellsten, Len A Pennacchio, et al. Genome resequencing reveals multiscale

geographic structure and extensive linkage disequilibrium in the forest tree *Populus trichocarpa*. *New Phytologist*, 196(3):713–725, 2012. 274

[7] Ryan F McCormick, Sandra K Truong, Avinash Sreedasyam, Jerry Jenkins, Shengqiang Shu, David Sims, Megan Kennedy, Mojgan Amirebrahimi, Brock Weers, Brian McKinley, et al. The *Sorghum bicolor* reference genome: improved assembly and annotations, a transcriptome atlas, and signatures of genome organization. *bioRxiv*, page 110593, 2017. 274

[8] Asher Haug-Baltzell, Sean A Stephens, Sean Davey, Carlos E Scheidegger, and Eric Lyons. SynMap2 and SynMap3D: web-based whole-genome synteny browsers. *Bioinformatics*, page btx144, 2017. xxxi, 276, 280, 284

[9] Eric Lyons, Brent Pedersen, Josh Kane, and Michael Freeling. The value of nonmodel genomes and an example using SynMap within CoGe to dissect the hexaploidy that predates the rosids. *Tropical Plant Biology*, 1(3-4):181–190, 2008. xxxi, 276, 284

[10] Martin I Krzywinski, Jacqueline E Schein, Inanc Birol, Joseph Connors, Randy Gascoyne, Doug Horsman, Steven J Jones, and Marco A Marra. Circos: an information aesthetic for comparative genomics. *Genome research*, 2009. 276

[11] William Constantine and Donald Percival. *wmtsa: Wavelet Methods for Time Series Analysis*, 2016. R package version 2.0-2. 276, 280

[12] Donald B Percival and Andrew T Walden. *Wavelet methods for time series analysis*, volume 4. Cambridge university press, 2006. 276, 280

[13] David M Goodstein, Shengqiang Shu, Russell Howson, Rochak Neupane, Richard D Hayes, Joni Fazo, Therese Mitros, William Dirks, Uffe Hellsten, Nicholas Putnam, et al. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Research*, 40(D1):D1178–D1186, 2012. 277, 279

[14] Kelly J Vining, Kyle R Pomraning, Larry J Wilhelm, Henry D Priest, Matteo Pellegrini, Todd C Mockler, Michael Freitag, and Steven H Strauss. Dynamic DNA

cytosine methylation in the *Populus trichocarpa* genome: tissue-level variation and relationship to gene expression. *BMC Genomics*, 13(1):1, 2012. 277, 281, 282

[15] Aaron R Quinlan. BEDTools: the Swiss-army tool for genome feature analysis. *Current protocols in bioinformatics*, pages 11–12, 2014. 279

[16] Ole Tange et al. GNU Parallel-The Command-Line Power Tool. *The USENIX Magazine*, 36(1):42–47, 2011. 279

[17] Shraddha Pai and Jingliang Ren. *IdeoViz: Plots data (continuous/discrete) along chromosomal ideogram*, 2017. R package version 1.10.0. 279

[18] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. 280

[19] RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA, 2016. 280

[20] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, 13(11):2498–2504, 2003. 280

[21] Luca Comai, Shamoni Maheshwari, and Mohan PA Marimuthu. Plant centromeres. *Current opinion in plant biology*, 36:158–167, 2017. 287

[22] Jiming Jiang, James A Birchler, Wayne A Parrott, and R Kelly Dawe. A molecular view of plant centromeres. *Trends in plant science*, 8(12):570–575, 2003. 287

[23] Rebecca J Mroczek and R Kelly Dawe. Distribution of retroelements in centromeres and neocentromeres of maize. *Genetics*, 165(2):809–819, 2003. 287

[24] Amar Kumar and Jeffrey L Bennetzen. Plant retrotransposons. *Annual review of genetics*, 33(1):479–532, 1999. 287

[25] Tim Langdon, Charlotte Seago, Michael Mende, Michael Leggett, Huw Thomas, John W Forster, Howard Thomas, R Neil Jones, and Glyn Jenkins. Retrotransposon evolution in diverse plant genomes. *Genetics*, 156(1):313–325, 2000. 287

[26] Hiroshi Mizuno, Takashi Matsumoto, and Jianzhong Wu. Composition and structure of rice centromeres and telomeres. In *Rice Genomics, Genetics and Breeding*, pages 37–52. Springer, 2018. 287

[27] Gernot G Presting, Ludmilla Malysheva, Jörg Fuchs, and Ingo Schubert. A ty3/gypsy retrotransposon-like sequence localizes to the centromeric regions of cereal chromosomes. *The Plant Journal*, 16(6):721–728, 1998. 287

[28] Savannah J Klein and Rachel J O'Neill. Transposable elements: genome innovation, chromosome diversity, and centromere conflict. *Chromosome Research*, pages 1–19, 2018. 287, 296, 297

[29] R Keith Slotkin and Robert Martienssen. Transposable elements and the epigenetic regulation of the genome. *Nature reviews genetics*, 8(4):272, 2007. 287

[30] Kiyotaka Nagaki, Keisuke Tanaka, Naoki Yamaji, Hisato Kobayashi, and Minoru Murata. Sunflower centromeres consist of a centromere-specific line and a chromosome-specific tandem repeat. *Frontiers in plant science*, 6:912, 2015. 287

[31] Cibele Gomes de Sotero-Caio, Diogo Cavalcanti Cabral-de Mello, Merilane da Silva Calixto, Guilherme Targino Valente, Cesar Martins, Vilma Loreto, Maria José de Souza, and Neide Santos. Centromeric enrichment of line-1 retrotransposons and its significance for the chromosome evolution of phyllostomid bats. *Chromosome Research*, 25(3-4):313–325, 2017. 287

[32] Timothy H Keitt and Dean L Urban. Scale-specific inference using wavelets. *Ecology*, 86(9):2497–2504, 2005. 291

[33] Stephan Ossowski, Korbinian Schneeberger, José Ignacio Lucas-Lledó, Norman Warthmann, Richard M Clark, Ruth G Shaw, Detlef Weigel, and Michael Lynch.

The rate and molecular spectrum of spontaneous mutations in arabidopsis thaliana. *science*, 327(5961):92–94, 2010. 292

[34] Jun Cao, Korbinian Schneeberger, Stephan Ossowski, Torsten Günther, Sebastian Bender, Joffrey Fitz, Daniel Koenig, Christa Lanz, Oliver Stegle, Christoph Lippert, et al. Whole-genome sequencing of multiple arabidopsis thaliana populations. *Nature genetics*, 43(10):956, 2011. 292

[35] Mao-Lun Weng, Claude Becker, Julia Hildebrandt, Matthew T Rutter, Ruth G Shaw, Detlef Weigel, and Charles B Fenster. Fine-grained analysis of spontaneous mutation spectrum and frequency in arabidopsis thaliana. *Genetics*, pages genetics–301721, 2018. 292

[36] Tu N Le, Yuji Miyazaki, Shohei Takuno, and Hidetoshi Saze. Epigenetic regulation of intragenic transposable elements impacts gene transcription in arabidopsis thaliana. *Nucleic acids research*, 43(8):3911–3921, 2015. 296

[37] Jeffrey L Bennetzen. Transposable element contributions to plant gene and genome evolution. *Plant molecular biology*, 42(1):251–269, 2000. 296

[38] Hans Ellegren. Microsatellites: simple sequences with complex evolution. *Nature reviews genetics*, 5(6):435, 2004. 297

[39] Pamela S Soltis, D Blaine Marchant, Yves Van de Peer, and Douglas E Soltis. Polyploidy and genome evolution in plants. *Current opinion in genetics & development*, 35:119–125, 2015. 297

[40] Sharlee Climer, Wei Yang, Lisa Fuentes, Victor G Dávila-Román, and C Charles Gu. A Custom Correlation Coefficient (CCC) Approach for Fast Identification of Multi-SNP Association Patterns in Genome-Wide SNPs Data. *Genetic Epidemiology*, 38(7):610–621, 2014. 299

[41] Sharlee Climer, Alan R Templeton, and Weixiong Zhang. Allele-Specific Network Reveals Combinatorial Interaction that Transcends Small Effects in Psoriasis GWAS. *PLoS Comput Biol*, 10(9):e1003766, 2014. 299

[42] Deborah A Weighill and Daniel Jacobson. Network metamodeling: Effect of correlation metric choice on phylogenomic and transcriptomic network topology. In *Network Biology*, pages 143–183. Springer, 2016. 299

Chapter 6

Conclusion

6.1 Concluding Remarks and Future Work

The work presented in this dissertation has made substantial contributions to data analysis capabilities in biofuels research. Techniques were developed for the integration of different multiple data types, combining the information in multiple large ’omics datasets. The methods developed and applied have provided strategies to gain insight into fundamental knowledge about the relationships between different data types in *P. trichocarpa*, as well as strategies to combine the information in these different data types to identify target genes involved in bioenergy-related phenotypes.

The Pleiotropy Decomposition technique presented in Chapter 2 describes a post-GWAS method which characterizes the pleiotropic signatures of genes using data from a multi-phenotype GWAS analysis. The representation of pleiotropic signatures as combinations of “pleiotropy modules” allows genes to be clustered not only by the phenotypes which they are associated with, but also the topology of SNP-phenotype associations within genes. This detailed characterization and clustering will allow for candidate genes to be identified which have pleiotropic signatures which are theoretically favorable for modification to impact certain phenotypes. This method will be applied to more phenotypes as they become available, including gene expression phenotypes. This will ultimately provide a global view of the pleiotropic interactions in *P. trichocarpa*.

The Lines of Evidence “LOE” data layering method presented in Chapter 3 provides an approach to integrate multiple ’omics datasets to identify novel candidate genes involved in biofuels-related phenotypes. Association network layers are constructed for each data type, and subsequently, LOE scores are calculated for each gene in the genome based on their connectivity to nodes already known to be involved in a particular function of interest. This method produces a useful ranking of genes, allowing for an evidence-based selection of candidate genes for testing. The network representation also provides useful context for the interpretation of the position of candidate genes in the entire system, and allows for hypotheses generation surrounding the mechanism of a gene’s involvement in the target function of interest. Future work will involve investigating optimal weights for different data layers in LOE scores by using a cross-validation analysis.

The pleiotropic signature clusters in Chapter 3 are easily integrated with LOE networks in Chapter 2, and a combined pleiotropy decomposition/LOE approach could also yield information on interesting gene regulatory circuits of genes which have similar pleiotropic signatures.

The wavelet-based signal processing in Chapter 4 allowed for the identification of the approximate centromere positions in *P. trichocarpa*. These positions had not been adequately reported in previous scientific literature. The wavelet analysis also provided a method to investigate scale-specific relationships between the different ’omics data layers, providing insights into the evolution of the *P. trichocarpa* genome, as well as potential driving forces of genome structure at different scales, as detailed in Chapter 5. This work provides a foundation for future scale-specific analysis. In particular, future work could include investigating scale-specific relationships between various genomic elements and measured phenotypes.

The collection of methods and approaches presented in this dissertation provide tools to aid in the design of plants optimized for biofuels production. The LOE methods will allow one to construct layered networks involving GWAS networks of relevant phenotypes, and the resulting scores will inform the researcher about new candidate genes involved in a particular biofuels-related phenotype of interest. These candidate genes can then be considered as targets for genetic modification or to potentially inform genomic selection procedures. Pleiotropy decomposition of the GWAS networks in the LOE analysis will unravel the multi-phenotype associations of genes, allowing for the investigation of the pleiotropic signatures of new candidate genes. This will assist in determining if modification of these candidate genes will cause unintended consequences by affecting other phenotypes. Signature clustering will allow the identification of genes with similar pleiotropic signatures as candidate genes, potentially identifying regulatory circuits or genes of shared function. The wavelet transform analysis presented provides a first step in attempting to identify scale-specific associations between genomic/epigenomic features and biofuels phenotypes, which could lead to new insights into scale-specific impacts on phenotype.

In summary, this dissertation has provided useful approaches for data integration and target gene identification, and provided strategies for the investigation of fundamental structural and functional genomic information about *P. trichocarpa*.

Vita

Deborah Ann Weighill, née Grobler, was born to parents Gregory Grobler and Lynette Grobler in Paarl, South Africa. She attended Stellebosch University in South Africa and completed a Bachelor of Science in Mathematical Sciences (Cum Laude) in 2011, a Bachelor of Science Honours in Grapevine and Wine Biotechnology (Cum Laude) in 2012, as well as a Master of Science in Grapevine and Wine Biotechnology (Cum Laude) in 2014. She then moved to Oak Ridge, Tennessee and joined the Bredesen Center for Interdisciplinary Research and Graduate Education, obtaining a PhD in May of 2019 with a focus in Computational Biology.