CBI Performance Metric for FY20: Report on genomic science-based advances and testing of new plant feedstocks for bioenergy purposes.

Q3 Metric: Report on bioenergy-relevant insights gained from analyses of the poplar genome.
June 2020

1. Introduction

Poplar (*Populus* spp.) is one of the two primary plant feedstocks that have been studied in the BioEnergy Science Center project (BESC, 2007-2017) and the Center for Bioenergy Innovation project (CBI, 2018-present). Over the past 12 years of our BESC and CBI research, we have sought and continue to accelerate domestication of poplar plants for bioenergy production. One main goal of BESC was to overcome lignocellulosic biomass recalcitrance for sugar release and CBI seeks to increase biomass yield, improve sustainable production of bioenergy crops, and enable valuable bio-based fuels and products from plant biomass (Fig. 1).

Poplar is a fast-growing perennial tree recognized for its economic potential in biofuels production, with various natural populations across broad environmental gradients in the US. Published in Science in 2006 [1], the genome of *Populus trichocarpa* was the 3rd plant genome sequenced, with more than 40,000 genes predicted in the genome. This poplar genome has been serving as a reference for functional genomics and population genomics research in poplar as well as comparative and evolutionary genomics research among diverse lineages of plant species. With the leadership of Oak Ridge National Laboratory and DOE Joint Genome Institute (JGI), rich genomics resources have been created for poplar, including an improved high-quality genome sequence assembly and gene annotation, genome-resequencing data for more than 1000 poplar genotypes, and poplar gene expression atlas for various tissue types and experimental conditions. Furthermore, the poplar genomics resources have been widely used by a large scientific community including the researchers in BESC and CBI for identification of genes associated with bioenergy traits.

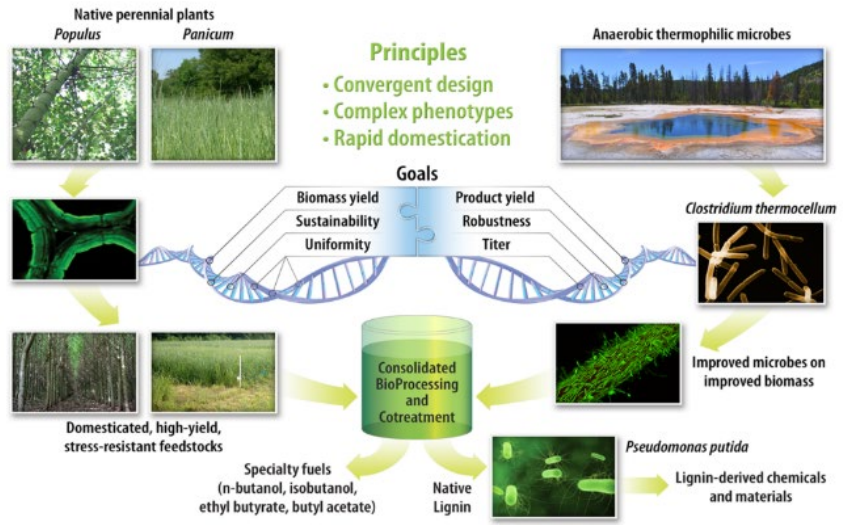
Native perennial plants
Populus
Panicum
Principles
• Convergent design
• Complex phenotypes
• Rapid domestication
Anaerobic thermophilic microbes
Goals
Biomass yield
Sustainability
Uniformity
Product yield
Robustness
Titer
Clostridium thermocellum
Consolidated BioProcessing and Cotreatment
Improved microbes on improved biomass
Domesticated, high-yield, stress-resistant feedstocks
Pseudomonas putida
Specialty fuels (n-butanol, isobutanol, ethyl butyrate, butyl acetate)
Native Lignin
Lignin-derived chemicals and materials

Figure 1. Research goals of the Center for Bioenergy Innovation

Here we summarize our BESC-/CBI-based research on the expansion of poplar genomics resources (pangenome and centromere), poplar genome-wide association study (GWAS), new approaches and algorithms for analysis of big genomics data, poplar functional genomics related to bioenergy, and elucidation of specific gene functions utilizing genomics and phenomics approaches.

Highlighted Results

2. Poplar Genome-Wide Association Study (GWAS) Resources

Populus trichocarpa GWAS population SNP dataset

Based on our initial success with identifying signatures of selection and genetic basis of phenotypic trait variation through GWAS analysis of approximately 500 *P. trichocarpa* accessions [2, 3], we created a new SNP dataset that includes genetic variations found in 882 poplar trees, and provides useful information to scientists studying plants as well as researchers more generally in the fields of biofuels, materials science, and secondary plant compounds. For nearly 12 years, researchers with DOE's BESC and CBI, multi-institutional organizations headquartered at ORNL, have studied the genome of *Populus* — a fast-growing perennial tree recognized for its economic potential in biofuels production. This GWAS dataset includes more than 28 million single nucleotide polymorphisms (SNPs) that have been derived from 17 trillion bases of sequence data generated from 882 undomesticated *Populus* genotypes. Each SNP represents a variant in a single DNA nucleotide that can act as a biological marker and/or causal allele within a protein sequence, helping scientists locate genes associated with certain characteristics, conditions or diseases. GWAS has great potential for revealing the molecular basis of lignin biosynthesis, composition and structure [4]. The applications of this resource have been used, among other things, to 1) seek genetic control of cell-wall recalcitrance — a natural characteristic of plant cell walls that prevents the release of sugars under microbial conversion and restricts biofuels production and 2) identify the molecular mechanisms controlling deposition of lignin in plant cell walls. Lignin is a polyphenolic polymer that strengthens plant cell walls and acts as a barrier to microbial access to cellulose during saccharification — the process of breaking cellulose down into simple sugars for fermentation. Although the resource's most immediate applications are in fundamental plant sciences, ORNL researchers plan to use the GWAS data to inform applied work in areas such as the production of 1) cleaner, sustainable transportation biofuels, 2) carbon fiber for lightweight vehicles, and 3) alternatives to conventional plastics and building insulation materials [5].

Populus association genetics

The BESC *Populus trichocarpa* association study population consists of 1,100 trees collected from across the range of the species from California to British Columbia. We have taken a population genomics approach to determine the distribution of neutral and adaptive genetic variation in this population. For example, the detection of reliable phenotype-genotype associations and molecular signatures of selection requires detailed knowledge about genome-wide patterns of allele frequency variation, linkage disequilibrium (LD) and recombination. Initially, we resequenced 16 *P. trichocarpa* genomes to an average depth of 39× and genotyped 120 trees from 10 *P. trichocarpa* subpopulations using 22,280 SNPs assayed with an Illumina Infinium BeadArray. Analyses of the population structure at multiple spatial scales revealed significant geographic differentiation that was consistent with models of Isolation by Distance (Fig. 2). Furthermore, latitudinal allele frequency gradients were strikingly common across the genome, affecting approximately 25% of SNPs. The decay of genome-wide LD relative to physical distance was much slower than expected compared with smaller-scale studies in *Populus* and other forest trees. LD in the GWAS population dropped below 0.2 within 6 kb. Consistent with this observation, estimates of effective population size from LD (Ne ≈ 4000-6000) were remarkably low relative to the large census sizes of *P. trichocarpa* stands. Fine-scale rates of recombination varied widely across the genome but were largely predictable based on DNA sequence and methylation. Taken together, these results have implications for our understanding of the evolutionary history of *P. trichocarpa* and the design of robust selection scans [4]. Importantly, the extensive LD suggests that genomic selection from association studies in undomesticated populations may be more feasible in *Populus* than previously assumed. Using this GWAS resource, we identified candidate genes related to many biofuels relevant tratis in *P. trichocarpa* [e.g., 6].

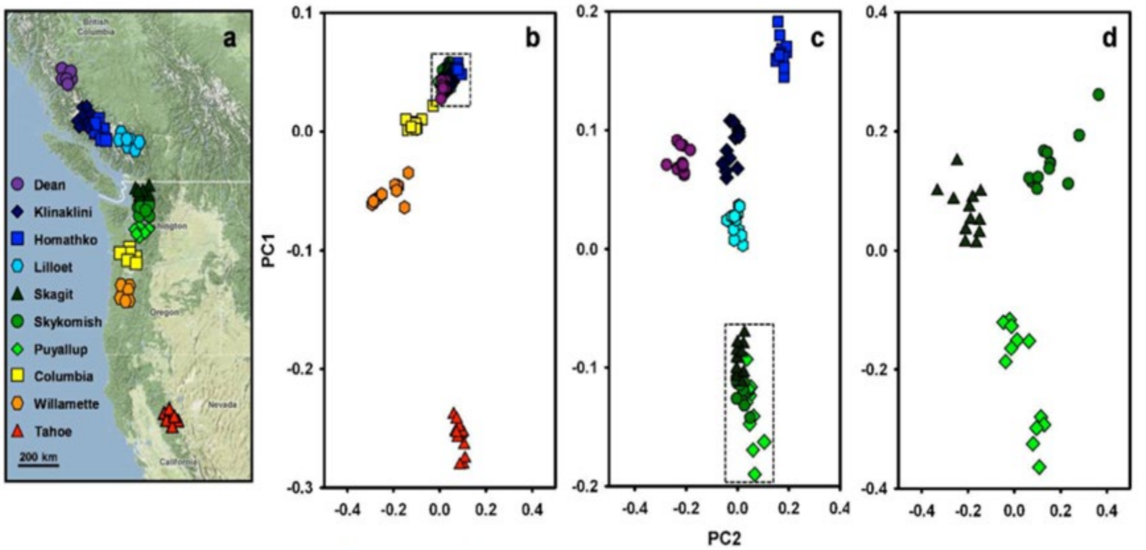

Figure 2. Population structure in *P. trichocarpa* based on Infinium SNP data. **(a)** Sampling locations of 120 *P. trichocarpa* trees from 10 subpopulations chosen to span the geographic distribution of the 16 resequenced trees **(b-d)** Major axes of variation based on principle components analysis of 22,280 SNPs on a range-wide scale **(b)**, after excluding trees from the clearly differentiated Tahoe, Willamette and Columbia subpopulations **(c)** and within the most homogeneous set of subpopulations in western Washington **(d)**.

Field trials for the GWAS population were established at four sites spanning most of the latitudinal range of the collection area: Placerville, CA; Corvallis, OR; Clatskanie, OR; and Agassiz, BC. Each trial was established with three clonal replicates of all 1,100 genotypes and survival ranges between ~70% in Placerville to ~90% in Clatskanie. These clones have been measured for a wide range of adaptive traits, including timing of bud flush and bud set, growth, crown architecture, leaf morphology, wood anatomy and disease/insect susceptibility. Genetically based phenotypic differentiation among groups of individuals may be estimated as QST (i.e., the degree of genetic differentiation among populations), which in turn is analogous to FST (i.e., the proportion of the total genetic variance contained in a subpopulation), or genetic differentiation among the groups at molecular markers. (Note, that a neutral trait will diverge due to demographic factors and thus the neutral expectation is that QST will be equivalent to FST.) We have found that QST was significantly greater than FST, suggesting that divergent selection has occurred for these traits. These data demonstrated that our network of plantations were able to accurate estimation of the quantitative components of phenotypic variation and ultimately the identification of the underlying loci controlling these traits. Such traits can be targeted by breeding programs to enhance the productivity of bioenergy plantations across a wide range of environments, thus ensuring a stable supply of feedstock for lignocellulosic biofuel refineries.

Summary: We have established genome-wide association study (GWAS) population consisting of more than 1000 undomesticated Populus trichocarpa trees and identified more than 20 million single nucleotide polymorphisms (SNPs) from this GWAS population. This fully sequenced GWAS resource has been used to link genetic variants to bioenergy-related traits, such as cell-wall recalcitrance and deposition of lignin in plant structures.

3. Poplar Genome Sequences

The addition of a thousand poplar genomes

As described above in the creation of the poplar GWASs population and common gardens, over a thousand addition *Populus trichocarpa* individuals were resequenced by JGI using DNA supplied by BESC and CBI researchers. These genomes were assembled and are available at the JGI website [https://phytozome.jgi.doe.gov/].

The *Populus trichocarpa* pangenome

The core and pan genomes of *P. trichocarpa* were constructed from 1,036 genotypes combined with the high-quality geneomic sequence reference (V3.1) to exploit the natural variation of this population for identifying potential biologically relevant gene targets. DNA reads from each genotype were quality/adapter-trimmed using Skewer and then assembled using Abyss. Skewer and Abyss are standard publicly available software packages. The assembled contigs were mapped back to the high-quality *P. trichocarpa* reference genome. Examining the mappings revealed near complete coverage of most of the reference genes. A core matrix was constructed showing presence/absence of reference genes in each genotype. The missing reference regions were quantified and GWAS was performed, treating these deleted regions as phenotypes, to search for patterns in subpopulations. Contigs from the unmapped assemblies for each genotype were mapped back to the pan-assembly; using these mappings, a presence/absence matrix was constructed for the newly predicted "pan"-genes (Fig. 3).

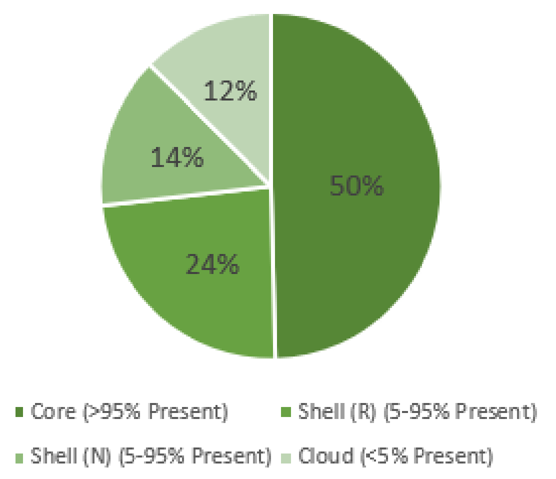

Figure 3. Gene-based pangenome of *Populus trichocarpa*. Annotated genes from the core and pan genomes of *P. trichocarpa* are characterized as core, cloud, and shell pan-genes in the high-confidence pangenome further divided into reference (R) and non-reference (N).

A recent promising development for the pangenome was the introduction of the new *P. trichocarpa* assembly V4.1 (courtesy of JGI, not yet publicly released). This high-quality PacBio assembly separates out the main genome and the alternate haplotype (300 MB of sequence). Using the new assembly, we filtered out ~25% of our pangenome contigs based on close similarities to the new sequences (either main or haplotype). This revealed that the old V3.1 assembly likely consisted of a combination of haplotype and main sequences, consistent with our previous pangenome assembly. The new assembly has enabled better resolution, facilitating classification of pan-contigs into reassembled primary references, reassembled haplotypes, or true pan contigs, significantly improving construction of the pangenome.

Centromere wavelet signatures and co-evolution with CENH3 in *P. trichocarpa*

Various 'omics data types have been generated for *P. trichocarpa*, each providing a layer of information that 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 (Fig. 4). 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* [7].

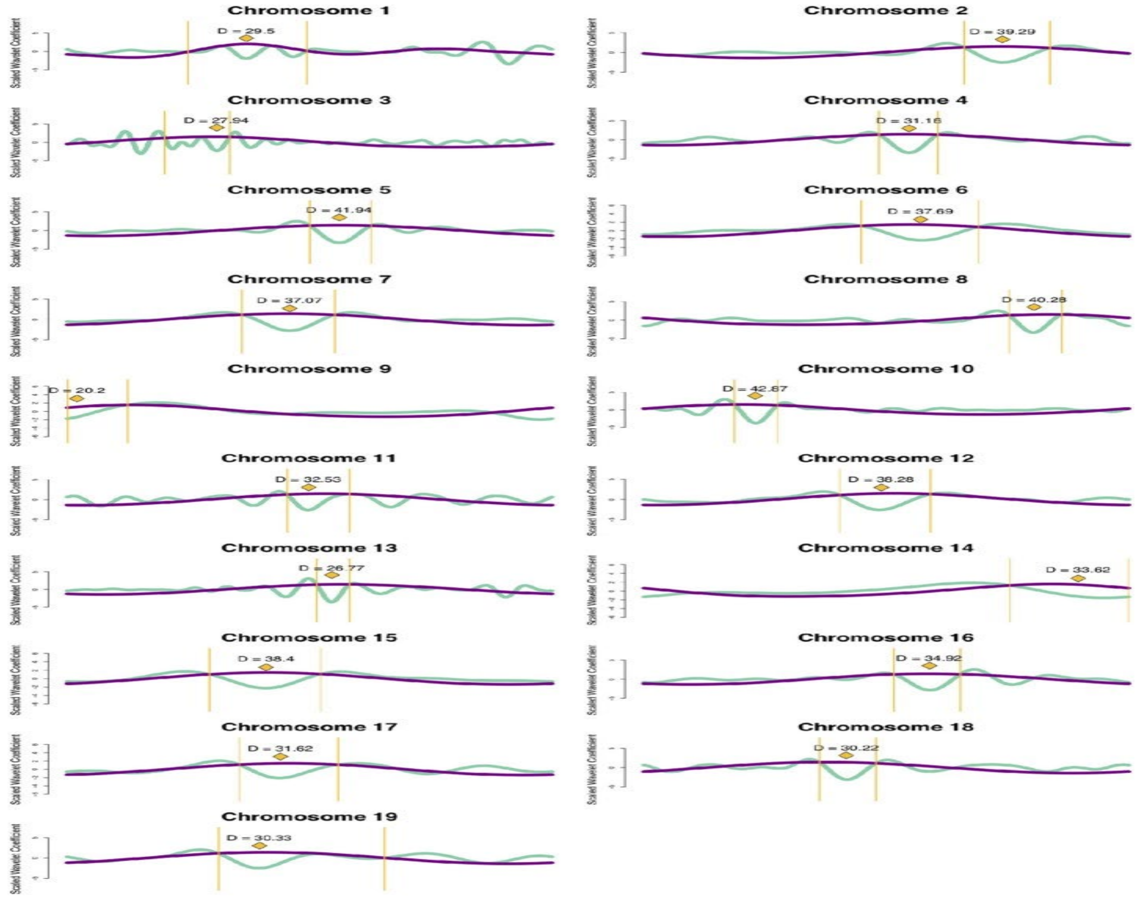

Figure 4. Centromere positions of *Populus trichocarpa*. Line plots for each chromosome of methylation wavelet coefficients (internode explant tissue) at pericentromeric scale (purple lines) and SNP density wavelet coefficients at centromeric scale (green lines). Yellow diamonds represent the putative 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. The maximum difference (D) between the unscaled wavelet coefficients used to determine the putative centromeric location (yellow diamonds) are shown for each chromosome.

Summary: We constructed the pan genome of P. trichocarpa to cover genomic diversity of more than 1000 poplar variants. Based on this pan genome, in combination with the P. trichocarpa reference genome sequence, we divided the poplar genes into core, cloud, and shell pan-genes. These data provide an new opportunity to discover gene-to-phenotype relationship that would be invisible to a reference genome approach. Also, through wavelet-based signal processing of genomic and epigenomic sequencing data, we predicted centromeric regions in P. trichocarpa and found evidence for the co-evolution of the centromeric histone CENH3 with the centromeric regions. The knowledge of where centromeres reside inform our ability to apply genomic selection algorithms to create new improved genotypes in P. trichocarpa.

4. Analysis and New Approaches and Algorithms

Multi-trait GWAS analysis

In conjunction with the poplar and computational biology teams, we developed multi-omics assessments using classical phenotypic measures to implicate specific genes for sustainability traits in *P. trichocarpa*. This approach increases the power to discern genes, gene networks and key metabolites involved in sustainability-related adaptive traits, such as leaf area and stomatal density (Fig. 5). For example, we have implicated mitochondrial transcription termination factor 1 (mTERF1) as central to controlling stomatal

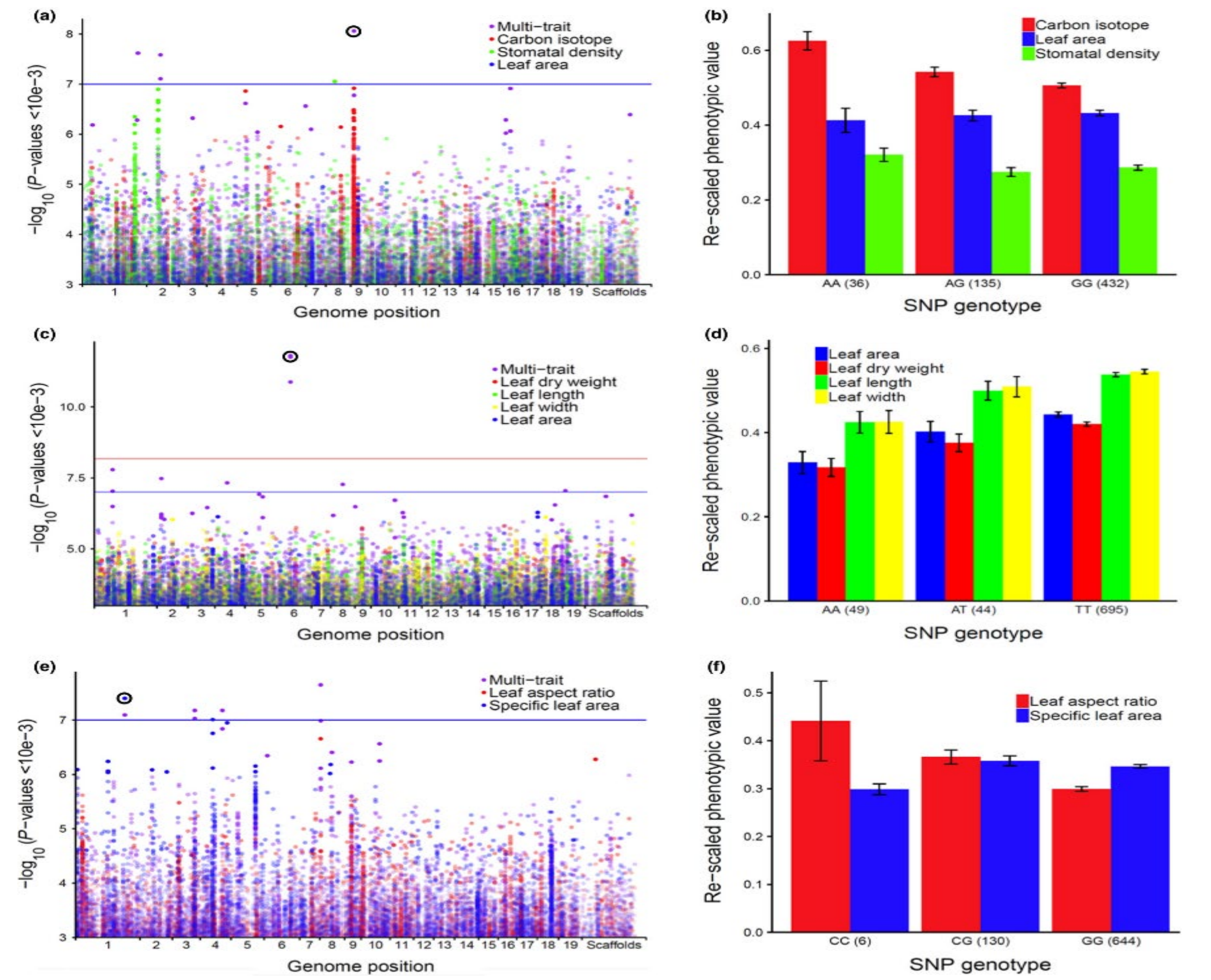

Figure 5. Manhattan plots comparing GEMMA univariate and multivariate GWAS in *Populus trichocarpa*. The colors of the dots correspond to single-trait or multi-trait associations. *P*-values are converted to $-\log_{10}$ (*P*-value). Single nucleotide polymorphisms (SNPs) above red lines passed Bonferroni correction test ($P \leq 7.37 \times 10^{-9}$), SNPs above blue lines are considered suggestive associations ($P \leq 1 \times 10^{-7}$). Only SNPs with $P \leq 1 \times 10^{-3}$) are plotted. (a) Carbon isotope, stomatal density, and leaf area. (b) Allelic effects of SNP near Potri.009G015500 (c) Leaf area, leaf dry weight, leaf length, and leaf width. (d) Allelic effects of SNP near gene *Potri.006G132500*. (e) Leaf aspect ratio and specific leaf area. (f) Allelic effects of SNP near gene *Potri.001G371800*. (b, d, f) SNPs depicted are circled in the corresponding Manhattan plots and the error bars represent ± SEs for rescaled phenotypic values.

density and leaf area. mTERFs are implicated in abiotic stresses that impair chloroplast function, including drought and it has been proposed that chloroplasts might act as sensors capable of perceiving stress and transmitting this information to the nucleus. At the same time, identification of these networks allows for further annotation of the poplar genome. Our approach identified additional genes and metabolites implicated in controlling plant response to abiotic stressors, such as YABBY, WUSCHEL and LEA genes, providing targets for future genomic selection efforts for stress resistance [8].

Coevolution network and genome-wide epistasis (2018 Gordon Bell Prize)

GWAS seeks to identify genetic variants that contribute to individual phenotypes. Alleles often function in complex networks that are often challenging to dissect with extant statistical procedures, particularly when the effects of multiple variants are nonadditive or epistatic. A custom correlation coefficient (CCC) approach was implemented on the Summit supercomputer. CCC can identify groups of SNPs that tend to co-occur in a population and consequently can be used to find combinations of SNPs that associate with specific phenotypes. CCC was used to create a co-evolutionary network in *P. trichocarpa* that is being used in generating sets of SNPs for genome-wide epistasis studies and as a network layer in our network mining algorithm. The calculations done with CCC on Summit reached 2.41 exaflops of performance, making it, at the time, the fastest scientific calculation ever performed. We were awarded the Gordon Bell Prize for our work with CCC on Summit.

iRF and epistatic GWAS

A full model of all higher-order interactions of both cellular and organismal components is one of the ultimate grand challenges of systems biology. The combinatorial space of all the possible interactions among biological objects in an organism is very large (10^{170} interactions), and there are considerable computing challenges in searching for these relationships. To explore this vast biological combinatorial space and to develop petascale applications that can create models of such systems, we have implemented scalable iRFs in R and C and have used the latter to perform epistatic GWAS on all the metabolomics phenotypes in the *P. trichocarpa* GWAS population, identifying up to ninth-order epistatic interactions between SNPs and metabolites. There is considerable evidence to suggest that a significant proportion of heritability is often missed by traditional single-SNP GWAS tests as they do not account for epistasis. Therefore, this new scalable code will allow us to detect associations between multiple genes and complex phenotypes relevant to CBI that would be missed by traditional GWAS methods. Yield, drought tolerance, microbiome association, and pathogen resistance are complex phenotypes that are very likely to have epistatic relationships among responsible genes.

Pleiotropy decomposition for unraveling complex gene-phenotype relationships

Pleiotropy is the phenomenon whereby a gene is involved in multiple functions. Different types of pleiotropic patterns exist and can be detected in the results of GWAS. Here, 882 *P. trichocarpa* genotypes were used to perform GWAS of untargeted metabolomics phenotypes. Pleiotropy decomposition, a new method that decomposes the results of a multi-phenotype GWAS study into three bipartite networks to unravel and regroup genes, was developed to identify pleiotropic connections (Fig. 6). This method allows us to find multiple patterns of pleiotropy within individual genes and to cluster genes based on pleiotropic patterns; it is proving valuable for the interpretation of large GWAS data sets. It will also aid in future synthetic biology efforts designed to optimize phenotypes of interest [9, 10]

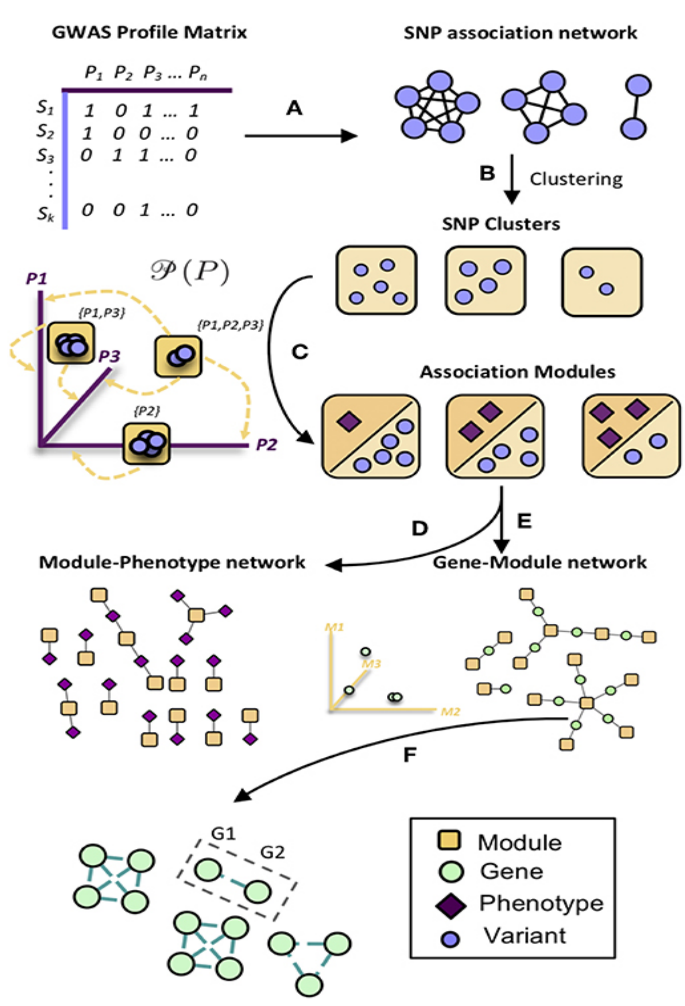

Figure 6. Overview of MPA decomposition and signature clustering. (A) The GWAS profile matrix M representing SNP-phenotype associations was constructed, and the Proportional Similarity between all pairs of SNPs (rows of M) was calculated. (B) Clustering of the SNP association similarity network results in clusters of SNPs with the same phenotype associations. (C) Association modules are constructed as elements of the powerset of phenotypes observed in the SNP clusters. Association modules can thus be seen as non-overlapping sets of SNPs, or overlapping sets of phenotypes. These modules form the axes of powerset space. (D) The module-phenotype network associates the phenotypes present in each element of the powerset observed in the SNP association clusters. (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 is performed in GM (powerset) space, grouping genes with the same module associations. Clustering genes in powerset space results in groups of genes with the same pattern of MPA signatures with the same set of phenotypes. For example, a signature cluster could involve G1 and G2 containing SNPs associating with both phenotypes P1 and P2, as well as a SNP associating with only P3.

Summary: To unleash the potential of poplar genome sequence and GWAS resource for bioenergy research, we developed new computational biology capabilities for analysis of large genomic datasets. Specifically, we used multivariate GWAS to identify the genes associated with multiple traits related to water-use efficiency, stomatal density, and leaf growth. We performed coevolution network analysis on the fastest supercomputer in the world to identify combinations of SNPs associated with specific phenotypes. We implemented the iterative random forest algorithm to perform epistatic GWAS and identified epistatic interactions between SNPs and metabolites, which are often missed by traditional single-SNP GWAS tests. We decomposed the results of a multi-phenotype GWAS and identified multiple patterns of pleiotropy within individual genes.

5. Poplar Functional Genomics Related to Bioenergy

Nature’s Knockout Association Test (NatKAT) identifies adaptive locus affecting poplar leaf cuticle

The pan-genome of *P. trichocarpa* contains thousands of structural variations or genes that are present in some of the population but not all. Some pan-genes are common, others are rare [as noted above]. This variation provides an assay of natural gene knock outs generated by mutation and natural selection over evolutionary time. We applied NatKAT to concentrations of leaf metabolites measured in the same population, potentially revealing which genes are associated with adaptive traits. One pan-gene (26802_g) strongly associated with C25 and C27 alkenes, which are the dominant components of leaf cuticular wax in *P. trichocarpa* (Fig. 7). The 26802_g gene is orthologous to ketoacyl-CoA-synthase genes involved in very long-chain fatty-acid elongation. The genotypes that possess 26802_g (20% of the population) show little to no C25 or C27 alkene content, whereas those without 26802_g accumulate the alkenes normally, indicating that the presence of 26802_g probably affects the ability to produce alkene precursors. Owing to the role of long alkenes in leaf cuticle formation, this pan-gene may confer an adaptation for aridity, as well as for foliar disease resistance.

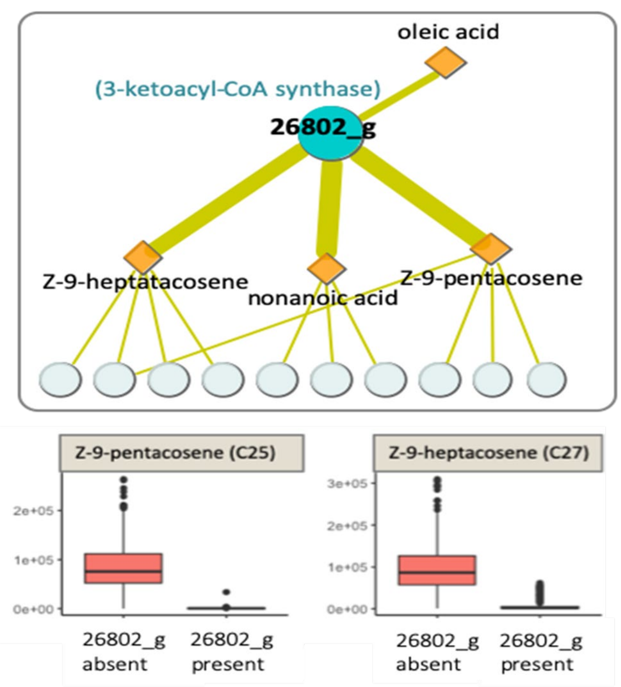

Figure 7. Network view of NatKAT result. Presence/absence of pan-genes (circles) associates with metabolites (diamonds). Presence of 26802_g results in failure to accumulate cuticular C25 and C27 alkenes.

Genome-wide prediction of *cis*- and *trans*-regulatory elements in poplar

We established an eQTN mapping pipeline to conduct global prediction of regulatory elements for >20,000 genes from both leaf and xylem transcriptomes. Of these, ~10,000 were shared between the two tissues and ~5,000 exhibited tissue-specific expression. We demonstrated numerous instances where regulatory elements mapped to the same genomic position across tissues, supporting our predictions of master regulatory loci, such as the novel transcriptional regulators on chromosome II and XIV of ten members of the Fasciclin gene family (Fig. 8). This gene family has been widely implicated in cell wall biosynthesis in numerous plant genera, but this is the first report of these specific transcriptional regulators targeting multiple members of the family. Additionally, three-layer transcriptional regulatory hierarchies were constructed to identify transcriptional master regulators that target more than 100 CBI-relevant genes. From this analysis, two candidate genes (*XBAT* and *GOLGIN*) were selected for cis-genic CRISPR-Cas 9 validation.

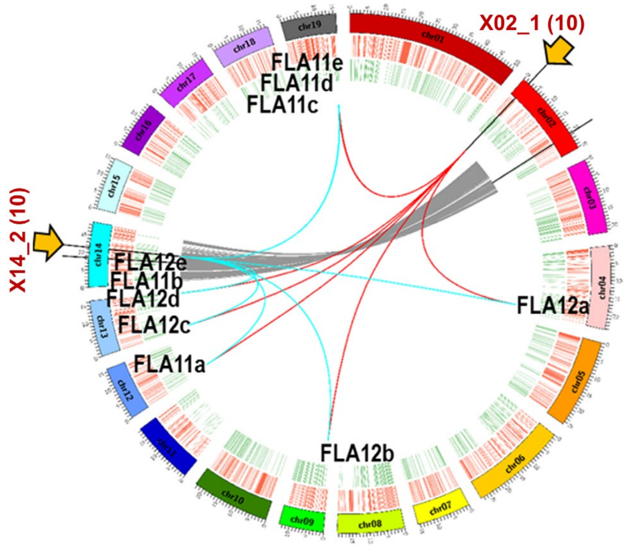

Figure 8. Mapping novel transcriptional regulators on chromosome II and XIV of ten members of the Fasciclin gene family in *Populus*.

Developing and applying conventional genomic selection algorithms in *Populus*

We evaluated the efficacy of using whole-genome marker data to guide selection of crosses to improve a suite of traits that are most relevant to production of biofuels and bioproducts. We first used the well-accepted Genomic Best Linear Unbiased Predictor (GBLUP) method to estimate single phenotypes measured in our *P. trichocarpa* common gardens. These phenotypes included height, diameter, lignin content and S:G (syringyl-guaiacyl) ratio, disease resistance and drought tolerance. Using a cross-validation procedure, we observed predictive abilities ranging from 0.11 to 0.50, meaning that up to 50% of the genetic variance in some traits could be captured using marker data alone. To facilitate simultaneous selection for all of these traits, we created a selection index based on a linear weighted combination of the traits, coupled with the variance-covariance structures of the traits. This enabled us to balance trade-offs due to negative genetic correlations with relative gains of the individual trait. We modeled three different selection scenarios that varied based on the relative importance of the lignin content and S:G ratio. On this basis, we selected 15 mating pairs with a wide range of expected genetic gains. These same breeding pairs were used for the prediction of progeny performance using machine learning algorithms developed by the CompBio Team. These crosses were performed by GreenWood Resources and 12 of the families are being propagated for field evaluations.

Targeted quantitative trait nucleotide (QTN) stacking and genomic selection

Genomic selection algorithms targeting productivity parameters (Jmax25, Red light25, Resistwp25, water-use efficiency (WUE), and diameter at breast height (DBH) were evaluated, suggesting that the Ridge Regression Best Linear Unbiased Prediction algorithm exhibited superior performance over Bayesian linear, Bayesian sparse linear mixed, and random forest models. We also established that GWAS-informed, SNP-phenotype associations, in which only 70 markers were used, achieved prediction accuracies with r≥0.80 for all traits, including DBH. In comparison, 10,000 randomly selected markers had prediction accuracies near ~r=0.90 and consumed significantly more computational time. For the second genomic selection task targeting cell wall chemistry, two field expeditions were conducted at Westport, OR, to collect wood, leaf, and xylem samples to evaluate phenotypic properties of the parental and progeny lines from the 7×7 cross. Owing to the unanticipated early termination of the Clatskanie, OR, site, scheduled for spring 2019, crosses to validate predicted progeny performance were made during winter 2018/2019 to meet CBI goals. To expedite selection of parental genotypes, outputs from the pleiotropic deconstruction analyses were being used to evaluate trait complementarity or tradeoffs.

Drought tolerance in poplar

In conjunction with the Poplar Team, the GWAS population was planted in Sedona, AZ, and Boardman, OR, in a drip-irrigated trial to determine the variability of drought tolerance in *P. trichocarpa*. The Boardman trial was measured for height and diameter after two growing seasons under reduced and full irrigation. Trees averaged 5.9 m in height under full irrigation and 3.6 m under reduced irrigation. Substantial variation in the relative response to drought was found, with some clones showing much less (designated "drought-tolerant") or much greater growth reduction ("drought sensitive") (Fig. 9). Drought-induced leaf senescence was scored at Boardman, OR, and GWAS analysis showed that a major locus involving a guard cell osmoregulator was significantly associated with the trait in two independent treatment blocks at the site, suggesting this trait is under strong genetic control with marginal to low environmental influence. Through integration of the *P. trichocarpa* genome sequence with the analysis of transcriptomics and proteomics data, we have identified poplar genes involved in drought stress response [11, 12]

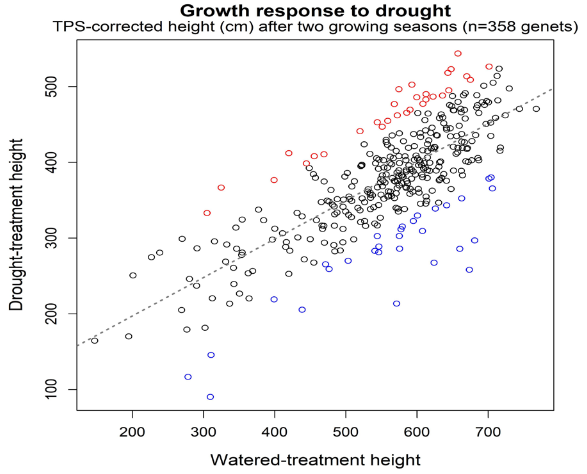

Figure 9. Height of the poplar GWAS population after 2 growing seasons under full (7 h) and reduced (4 h) irrigation. Drought-resistant lines have the highest positive residuals (red) relative to drought-sensitive lines (blue).

Summary: The genomic resources described above is accelerating our ability to identify genes and putative functions. We identified genes in the P. trichocarpa pan-genome, which affect poplar leaf cuticle leaf cuticular wax, using nature's knockout association test. Using genome-wide eQTN mapping, we predicted regulatory elements for more than 20,000 genes and identified transcriptional master regulators regulating more than 100 bioenergy-relevant genes. Our GWAS analysis identified a major genomic locus encoding a guard cell osmoregulator associated with drought tolerance. We applied conventional genomic selection algorithms to simultaneously select for multiple traits to balance trade-offs caused by negative genetic correlations with the relative gains of individual traits.

6. Examples of Specific Gene Functions Discovered Utilizing Genomic and Phenotypic Approaches

New cell wall regulatory genes in *Populus trichocarpa*

Understanding the regulatory network controlling cell wall biosynthesis is of great interest in *Populus trichocarpa*, both because of its status as a model woody perennial and its importance for lignocellulosic products. We searched for genes with putatively unknown roles in regulating cell wall biosynthesis using an extended network-based Lines-of-Evidence (LOE) pipeline to combine multiple omics data sets in *P. trichocarpa*, including gene coexpression, gene comethylation, population level pairwise SNP correlations, and two distinct SNP-metabolite Genome Wide Association Study (GWAS) layers (Fig. 10). By incorporating validation, ranking, and filtering approaches we produced a list of nine high priority gene candidates for involvement in the regulation of cell wall biosynthesis. We subsequently performed a detailed investigation of candidate gene GROWTH-REGULATING FACTOR 9 (*PtGRF9*). To investigate the role of *PtGRF9* in regulating cell wall biosynthesis, we assessed the genome-wide connections of *PtGRF9* and a paralog across data layers with functional enrichment analyses, predictive transcription factor binding site analysis, and an independent comparison to eQTN data. Our findings indicate that PtGRF9

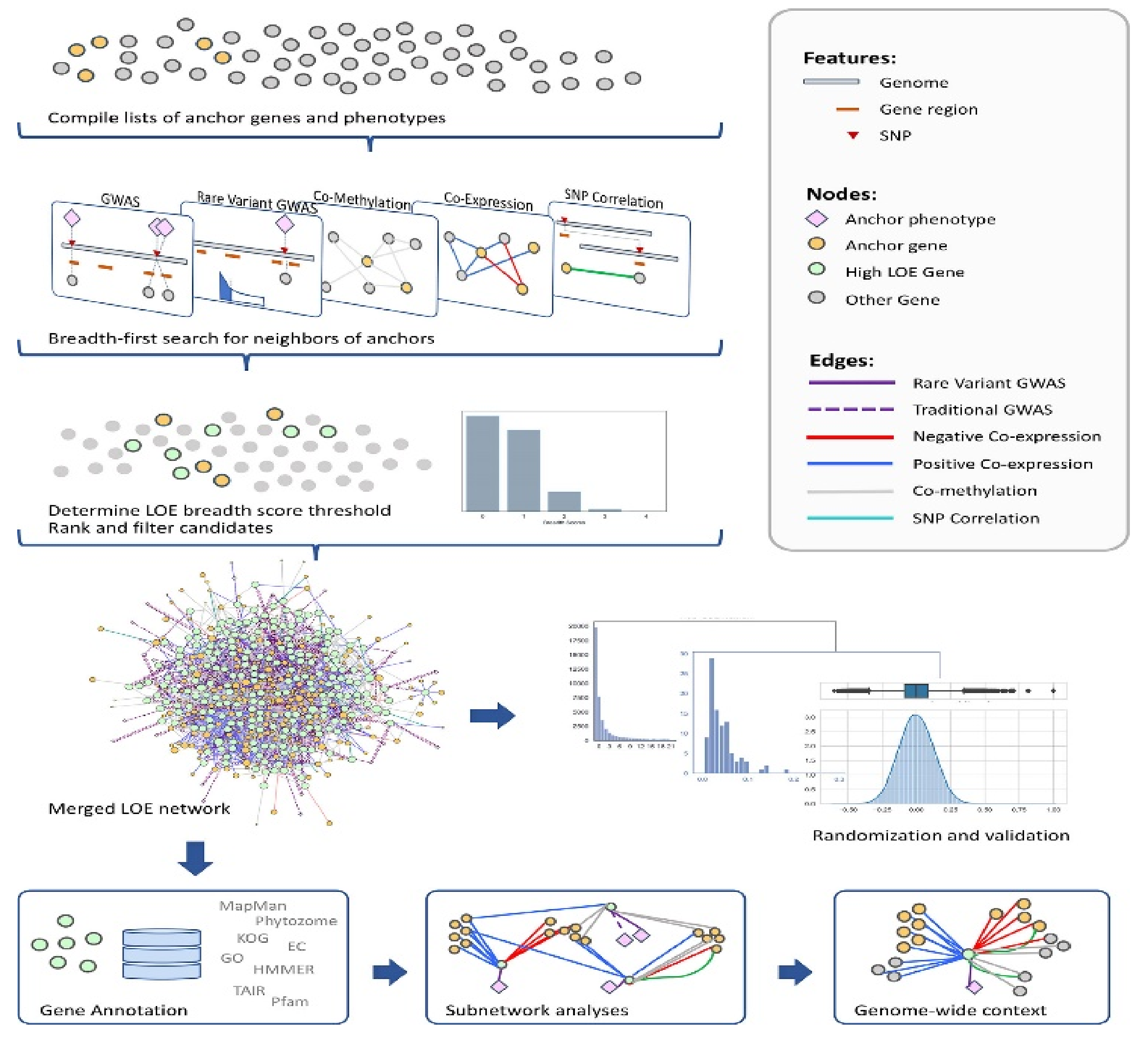

Figure 10. Overview of the method for identifying new candidate genes involved in cell wall regulation through data layering and calculation of Line of Evidence (LOE) scores.

likely affects the cell wall by directly repressing genes involved in cell wall biosynthesis, such as *PtCCoAOMT* and *PtMYB.41*, and indirectly by regulating homeobox genes. Furthermore, evidence suggests that *PtGRF9* paralogs may act as transcriptional co-regulators that direct the global energy usage of the plant. Using our extended pipeline, we show multiple lines of evidence implicating the involvement of these genes in cell wall regulatory functions and demonstrate the value of this method for prioritizing candidate genes for experimental validation [13].

Our GWAS analysis revealed that the 5-enolpyruvylshikimate 3-phosphate synthase gene (PtrEPSP) in *P. trichocarpa* was associated with lignin biosynthesis, and our molecular experiments demonstrated that PtrEPSP repressed the expression of PtrMYB021, which is a master regulator of the phenylpropanoid pathway and lignin biosynthesis [14]. Also, using the *Populus* genomics resources, we have identified additional poplar genes involved in cell wall biosynthesis [15-20], lignin content and composition [21-23], pectin biosynthesis [24], and the production of xylan and homogalacturonan [25, 26].

We performed association genetic analyses based on phenotypic data from the field trial in Corvallis, OR, and SNP genotypes for 28,733 genetic loci. Association genetics statistical approaches were used to identify specific SNP within explicit genetic loci that controlled sugar release and other relevant cell wall phenotypes. After controlling for the confounding effects of population structure and relatedness using three statistical approaches, we identified 473 associations (at $p \leq 0.001$) with traits related to biomass productivity, crown architecture, and phenology. Association analyses were conducted using MBMS data for hemicellulose, cellulose, lignin content, lignin S/G ratio, and 24 individual molecular beam mass spectrometry (MBMS) spectra from 774 genotypes as well as saccharification-based glucose, xylose, and combined glucose/xylose release (Fig. 11). The average increase in sugar release yield associated with each SNP is approximately 23–28% above wild-type depending upon the sugar analyzed. These genes have been accepted within our transformation pipeline. Approximately 42 genes containing SNP variants with strong associations were identified, and invention disclosures were submitted for a set of 16 with the most robust support from multiple phenotypes and analyses. We also identified 46 genes and their amino acid substitutions that are controlling the phenotypes measured in this population.

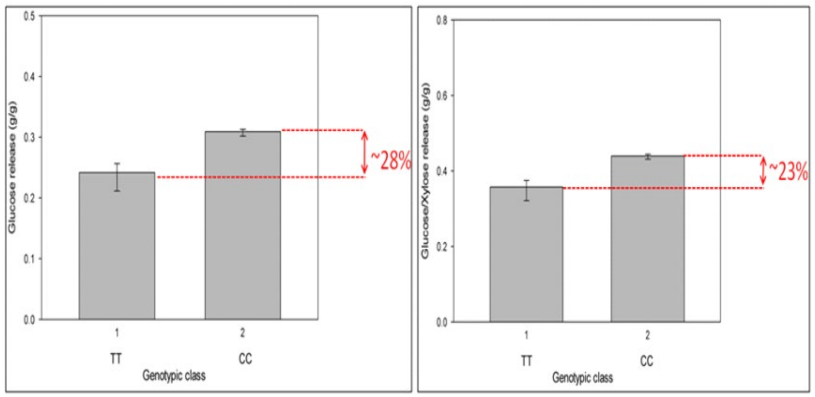

Figure 11. Evaluation of glucose (left) and glucose + xylose (right) enzymatic release for 774 genotypes (CC) showing less recalcitrant phenotypes compared to wildtype (TT).

Poplar genes mediating plant–pathogen interactions

Invasive microbes causing diseases such as Dutch elm disease negatively affect ecosystems and economies around the world. The deployment of resistant genotypes for combating introduced diseases typically relies on breeding programs that can take decades to complete. To demonstrate how this process can be accelerated, we employed a genome-wide association mapping of *ca*. 1,000 resequenced *Populus trichocarpa* trees individually challenged with *Sphaerulina musiva*, an invasive fungal pathogen. Among significant associations, three loci associated with resistance were identified and predicted to encode one putative membrane-bound L-type receptor-like kinase and two receptor-like proteins (Fig. 12). A susceptibility-associated locus was predicted to encode a putative G-type D-mannose–binding receptor-like kinase. Multiple lines of evidence, including allele analysis, transcriptomics, binding assays, and overexpression, support the hypothesized function of these candidate genes in the *P. trichocarpa* response to *S. musiva* [27]. Furthermore, we found that a G-type lectin receptor-like kinases (lecRLKs) in *P. trichocarpa* mediates the interaction between poplar and fungus *Laccaria bicolor* [28]. Also, our genome-wide comparative genomics analysis has provided new insights into the classification, domain architecture and expression of lecRLKs in the *Populus* [29].

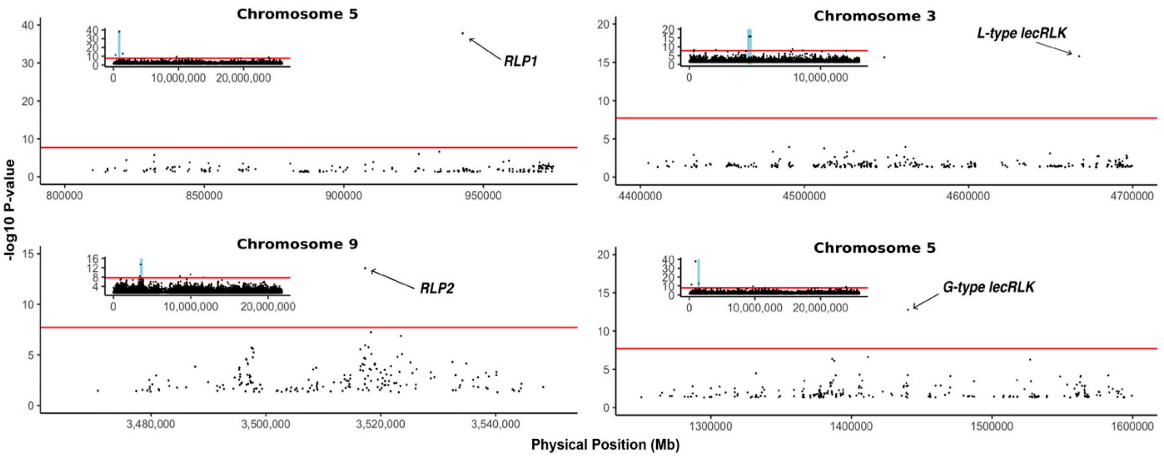

Figure 12. Summary of phenotyping and allele analysis identifying resistance- and susceptibility-associated loci in the *P. trichocarpa* association-mapping population inoculated with the fungal pathogen *S. musiva*. The Manhattan plot of *P. trichocarpa* chromosome 5 region depicts significant associations of receptor-like protein 1 (RLP1). The Inset depicts the entire length of *P. trichocarpa* chromosome 5. The Manhattan plot of *P. trichocarpa* chromosome 9 region depicts significant associations of receptor-like protein 2 (RLP2). The Inset depicts the entire length of *P. trichocarpa* chromosome 9. The Manhattan plot of *P. trichocarpa* chromosome 3 region depicts significant associations of L-type lectin receptor-like kinase (L-type lecRLK). The Inset depicts the entire length of *P. trichocarpa* chromosome 3. Manhattan plot of *P. trichocarpa* chromosome 5 depicting significant association of a G-type lectin receptor-like kinase (G-type lecRLK) with susceptibility. Each black dot on the Manhattan plots corresponds to a marker, its level of significance, and its physical position on the chromosome (in megabases). The red line represents the Bonferroni-corrected significance threshold, and the blue squares in the Insets represent the region of each chromosome immediately surrounding the significant associations that are visualized in the local Manhattan plot.

HCT2 in caffeoylquinic acid biosynthesis and its regulation by defense-responsive transcription factors in *Populus*

3-*O*-caffeoylquinic acid, also known as chlorogenic acid (CGA), functions as an intermediate in lignin biosynthesis in the phenylpropanoid pathway. It is widely distributed among numerous plant species and acts as an antioxidant in both plants and animals. Using GC-MS, we discovered consistent and extreme variation in CGA content across a population of 739 4-yr-old *Populus trichocarpa* accessions. We performed genome-wide association studies (GWAS) from 917 *P. trichocarpa* accessions and expression-based quantitative trait loci (eQTL) analyses to identify key regulators. The GWAS and eQTL analyses resolved an overlapped interval encompassing a hydroxycinnamoyl-CoA:shikimate hydroxycinnamoyl transferase 2 (*PtHCT2*) that was significantly associated with CGA and partially characterized metabolite abundances (Fig. 13). *PtHCT2* leaf expression was significantly correlated with CGA abundance and it was regulated by *cis*-eQTLs containing W-box for WRKY binding. Among all nine *PtHCT* homologs, *PtHCT2* is the only one that responds to infection by the fungal pathogen *Sphaerulina musiva* (a *Populus* pathogen). Validation using protoplast-based transient expression system suggests that *PtHCT2* is regulated by the defense-responsive WRKY. These results are consistent with reports of CGA functioning as an antioxidant in response to biotic stress. This study provides insights into data-driven and omics-based inference of gene function in woody species [30].

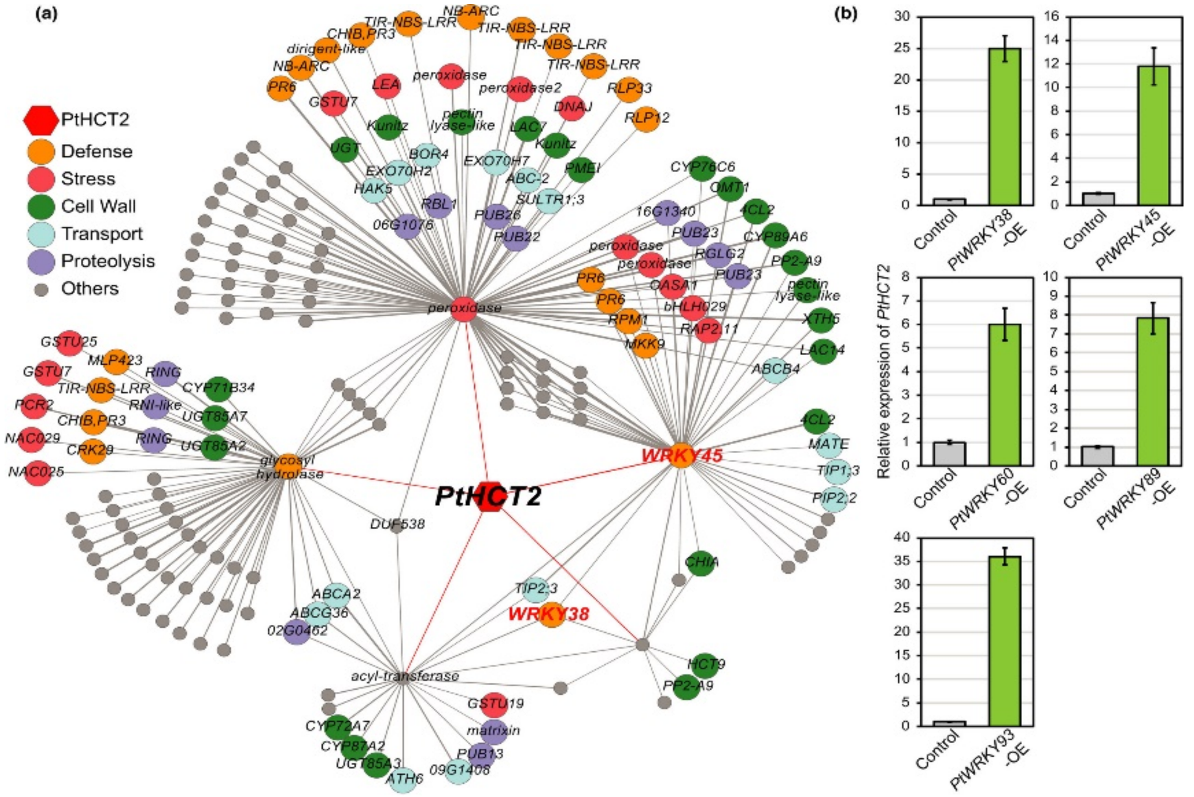

Figure 13. Co-expression network of *PtHCTs* in *Populus*. (a) Based on the functional annotation, the genes in the *PtHCT2* co-expression network were classified into the following groups: defense response (orange nodes), stress response (pink nodes), cell wall related (green nodes), transport (cyan nodes), proteolysis (purple nodes) and others (grey nodes). Two *WRKYs* (*WRKY38* and *WRKY45*) are among the *PtHCT2* co-expression network. (b) Regulation of *PtHCT2* by *Pt*WRKYs. Five *PtWRKYs* (*PtWRKY38*, *45*, *60*, *89* and *93*) were transiently overexpressed in *Populus* protoplasts. The expression levels of *PtHCT2* were analyzed by using qRT-PCR with three biological replicates. Error bars, ± SE.

PolyQ length modulates the function of *Populus* ANGUSTIFOLIA protein

Polyglutamine (polyQ) stretches have been reported to occur in proteins across many organisms including animals, fungi and plants. Expansion of these repeats has attracted much attention due their associations with numerous human diseases including Huntington's and other neurological maladies. This suggests that the relative length of polyQ stretches is an important modulator of their function. Here, we report the identification of a *Populus* C-terminus binding protein (CtBP) ANGUSTIFOLIA (*PtAN1*) which contains a polyQ stretch whose functional relevance had not been established. Analysis of 917 resequenced *Populus trichocarpa* genotypes revealed three allelic variants at this locus encoding 11-, 13- and 15-glutamine residues. Transient expression assays using *Populus* leaf mesophyll protoplasts revealed that the 11Q variant exhibited strong nuclear localization whereas the 15Q variant was only found in the cytosol, with the 13Q variant exhibiting localization in both subcellular compartments (Fig. 14). We assessed functional implications by evaluating expression changes of putative *PtAN1* targets in response to overexpression of the three allelic variants and observed allele-specific differences in expression levels of putative targets. Our results provide evidence that variation in polyQ length modulates *Pt*AN1 function by altering subcellular localization [31].

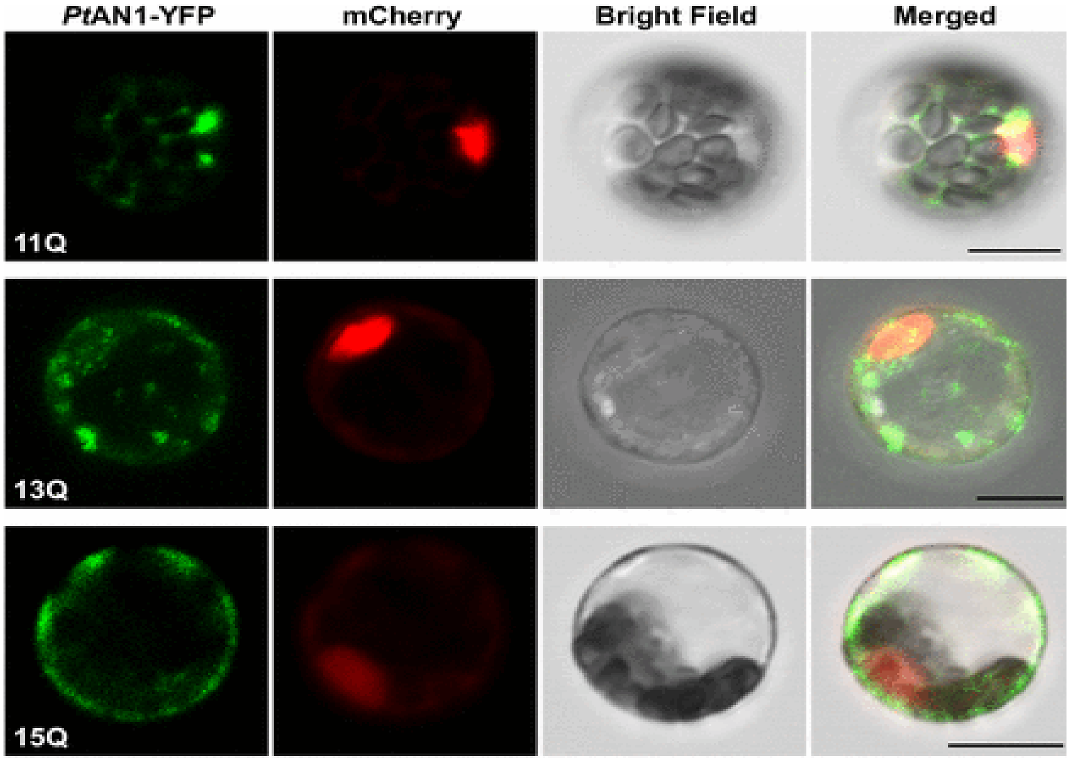

Figure 14. Subcellular localization of *Pt*AN1 in *Populus* protoplasts. YFP was fused with different alleles of *Pt*AN1. *35S:PtAN1-YFP* co-transfected with VirD2NLS-mCherry tagged nucleus marker into poplar mesophyll protoplasts. YFP signal is indicated as green color and mCherry signal is indicated as red color. Scale bar is 10 µm.

Defining the genetic components of callus formation

A characteristic feature of plant cells is the ability to form callus from parenchyma cells in response to biotic and abiotic stimuli. Tissue culture propagation of recalcitrant plant species and genetic engineering for desired phenotypes typically depends on efficient *in vitro* callus generation (Fig. 15). Callus formation is under genetic regulation, and consequently, a molecular understanding of this process underlies successful generation for propagation materials and/or introduction of genetic elements in experimental or industrial applications. Herein, we identified 11 genetic loci significantly associated with callus formation in *P. trichocarpa* using a genome-wide association study (GWAS) approach. Eight of the 11 significant gene associations were consistent across biological replications, exceeding a chromosome-wide–log10 (p)=4.46 [p = 3.47E−05] Bonferroni-adjusted significance threshold. These eight genes were used as hub genes in a high-resolution co-expression network analysis to gain insight into the genome-wide basis of callus formation. A network of positively and negatively co-expressed genes, including several transcription factors, was identified. As proof-of-principle, a transient protoplast assay confirmed the negative regulation of a Chloroplast Nucleoid DNA-binding-related gene (Potri.018G014800) by the LEC2 transcription factor. Many of the candidate genes and co-expressed genes were 1) linked to cell division and cell cycling in plants and 2) showed homology to tumor and cancer-related genes in humans. The GWAS approach based on a high-resolution marker set, and the ability to manipulate targets genes *in vitro*, provided a catalog of high-confidence genes linked to callus formation that can serve as an important resource for successful manipulation of model and non-model plant species, and likewise, suggests a robust method of discovering common homologous functions across organisms [32].

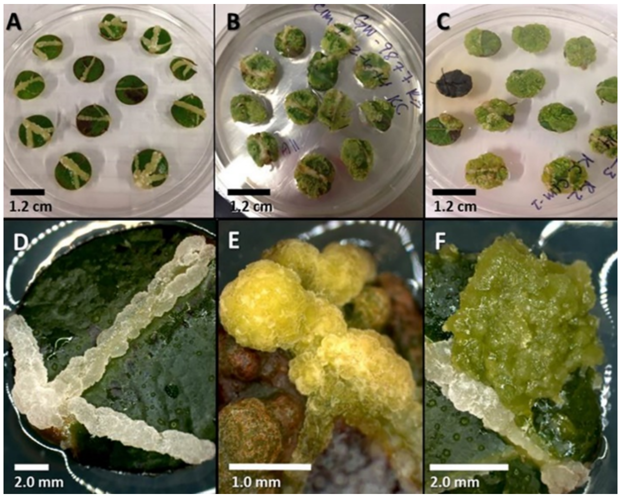

Figure 15. Callus formation on *Populus* leaf disc explants after 30 days on a callus induction medium. (**A**) 12 replicate leaf disk explants with callus along the midrib, (**B**) 12 replicate leaf disk explants with callus across the explant, (**C**) 12 replicate leaf disk explants with callus along the cut margin, (**D**) white friable callus along the midrib, (**E**) light green compact callus, and **(F**) green friable callus.

Summary: Our analysis of GWAS and multiple omics data identified nine high priority gene candidates regulating cell wall biosynthesis, receptor-like kinases responsive for disease (caused by Sphaerulina musiva) resistance and susceptibility in poplar. Our GWAS and eQTL analyses revealed a key defense-responsive gene (PtHCT2) involved in caffeoylquinic acid biosynthesis in poplar. Also, we identified a novel Populus gene (PtAN1), in which variation in polyQ length modulates PtAN1 function by altering its protein subcellular localization. Finally, our GWAS analysis identified 11 genetic loci associated with callus formation in Populus trichocarpa, which have potential application for improving genetic transformation in non-model plant species including bioenergy crops.

7. Concluding Remarks

The poplar genome sequence was published right before the launching of the DOE Bioenergy Research Centers (BRCs). Within BESC and CBI, we have significantly expanded the poplar genomics resources by resequencing more than 1000 poplar genotypes, creating poplar pangenome, and identifying centromeric regions overlooked by regular genome annotation. To facilitate the utilization of the large poplar genomics datasets, we have developed new computational biology capabilities for linking genomic variants to multiple plant traits, constructing coevolution network, performing epistatic genome-wide association study (GWAS), informing genomic selection via targeted quantitative trait nucleotide stacking, and unraveling complex gene-phenotype relationships using pleiotropy decomposition. By leveraging the rich poplar genomics resources, we have identified genes affecting poplar leaf cuticle and *cis-trans-*regulatory elements of more than 20,000 genes. Through integrative analysis of GWAS and gene expression data, we have elucidated the function of poplar genes regulating cell wall biosynthesis, mediating plant–pathogen interactions, and promoting callus formation. These achievements not only provide new insights into the molecular mechanisms underlying biomass recalcitrance, disease resistance and drought tolerance in bioenergy crop poplar, but also establish unique genomics resources and advanced capabilities for analysis of big genomics data, laying a solid foundation for future poplar genomics research to maximize the potential of poplar for bioenergy production.

References

1. Tuskan, G.A. et al. (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. &amp; Gray). Science 313 (5793), 1596-1604.
2. McKown, A.D. et al. (2014) Genome-wide association implicates numerous genes underlying ecological trait variation in natural populations of *Populus trichocarpa*. New Phytol 203 (2), 535-53.
3. Evans, L.M. et al. (2014) Population genomics of *Populus trichocarpa* identifies signatures of selection and adaptive trait associations. Nature Genetics 46 (10), 1089-1096.
4. Tuskan, G.A. et al. (2019) Population-level approaches reveal novel aspects of lignin biosynthesis, content, composition and structure. Current Opinion in Biotechnology 56, 250-257.
5. Tuskan, G. et al. (2016) *Populus trichocarpa* genome-wide association study (GWAS) population SNP dataset released. https://www.osti.gov/biblio/1411410, DOI: 10.13139/OLCF/1411410.
6. Bdeir, R. et al. (2019) Genome-wide association studies of bark texture in *Populus trichocarpa*. Tree Genetics & Genomes 15 (1), 14.
7. Weighill, D. et al. (2019) Wavelet-Based genomic signal processing for centromere identification and hypothesis generation. Frontiers in genetics 10, 487-487.
8. Chhetri, H.B. et al. (2019) Multitrait genome-wide association analysis of *Populus trichocarpa* identifies key polymorphisms controlling morphological and physiological traits. New Phytologist 223 (1), 293-309.
9. Weighill, D. et al. (2019) Multi-phenotype association decomposition: Unraveling complex gene-phenotype relationships. Frontiers in Genetics 10 (417), https://doi.org/10.3389/fgene.2019.00417.
10. Weighill, D. et al. (2018) Pleiotropic and epistatic network-based discovery: Integrated networks for target gene discovery. Frontiers in Energy Research 6 (30), https://doi.org/10.3389/fenrg.2018.00030.
11. Abraham, P.E. et al. (2018) Quantitative proteome profile of water deficit stress responses in eastern cottonwood (*Populus deltoides*) leaves. PLOS ONE 13 (2), e0190019.
12. Garcia, B.J. et al. (2018) Phytobiome and transcriptional adaptation of *Populus deltoides* to acute progressive drought and cyclic drought. Phytobiomes Journal 2 (4), 249-260.
13. Furches, A. et al. (2019) Finding new cell wall regulatory genes in *Populus trichocarpa* using multiple lines of evidence. Frontiers in Plant Science 10 (1249), https://doi.org/10.3389/fpls.2019.01249.

14. Xie, M. et al. (2018) A 5-enolpyruvylshikimate 3-phosphate synthase functions as a transcriptional repressor in *Populus*. The Plant Cell 30 (7), 1645-1660.
15. Yang, X. et al. (2011) Identification of candidate genes in *Arabidopsis* and *Populus* cell wall biosynthesis using text-mining, co-expression network analysis and comparative genomics. Plant Sci 181 (6), 675-87.
16. Ye, C.Y. et al. (2011) Comparative analysis of GT14/GT14-like gene family in *Arabidopsis*, *Oryza*, *Populus*, *Sorghum* and *Vitis*. Plant Sci 181 (6), 688-95.
17. Yang, Y. et al. (2017) Overexpression of a Domain of Unknown Function 266-containing protein results in high cellulose content, reduced recalcitrance, and enhanced plant growth in the bioenergy crop *Populus*. Biotechnology for Biofuels 10 (1), 74.
18. Yang, Y. et al. (2017) Overexpression of a Domain of Unknown Function 231-containing protein increases O-xylan acetylation and cellulose biosynthesis in *Populus*. Biotechnology for Biofuels 10 (1), 311.
19. Badmi, R. et al. (2018) A new calmodulin-binding protein expresses in the context of secondary cell wall biosynthesis and impacts biomass properties in *Populus*. Frontiers in Plant Science 9 (1669), https://doi.org/10.3389/fpls.2018.01669.
20. Zhang, J. et al. (2019) Overexpression of a serine hydroxymethyltransferase increases biomass production and reduces recalcitrance in the bioenergy crop *Populus*. Sustainable energy & fuels 3 (1), 195-207.
21. Bryan, A.C. et al. (2016) Knockdown of a laccase in *Populus deltoides* confers altered cell wall chemistry and increased sugar release. Plant Biotechnol J 14 (10), 2010-20.
22. Yang, Y. et al. (2019) PdWND3A, a wood-associated NAC domain-containing protein, affects lignin biosynthesis and composition in *Populus*. BMC Plant Biology 19 (1), 486.
23. Zhang, J. et al. (2020) Overexpression of a Prefoldin β subunit gene reduces biomass recalcitrance in the bioenergy crop *Populus*. Plant Biotechnology Journal 18 (3), 859-871.
24. Biswal, A.K. et al. (2018) Sugar release and growth of biofuel crops are improved by downregulation of pectin biosynthesis. Nature Biotechnology 36 (3), 249-257.
25. Biswal, A.K. et al. (2018) Working towards recalcitrance mechanisms: increased xylan and homogalacturonan production by overexpression of GAlactUronosylTransferase12 (GAUT12) causes increased recalcitrance and decreased growth in Populus. Biotechnology for Biofuels 11 (1), 9.
26. Biswal, A.K. et al. (2015) 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. Biotechnol Biofuels 8, 41.
27. Muchero, W. et al. (2018) Association mapping, transcriptomics, and transient expression identify candidate genes mediating plant–pathogen interactions in a tree. Proceedings of the National Academy of Sciences 115 (45), 11573-11578.
28. Labbé, J. et al. (2019) Mediation of plant–mycorrhizal interaction by a lectin receptor-like kinase. Nature Plants 5 (7), 676-680.
29. Yang, Y. et al. (2016) Genome-wide analysis of lectin receptor-like kinases in *Populus*. BMC Genomics 17 (1), 699.
30. Zhang, J. et al. (2018) Genome-wide association studies and expression-based quantitative trait loci analyses reveal roles of HCT2 in caffeoylquinic acid biosynthesis and its regulation by defense-responsive transcription factors in *Populus*. New Phytologist 220 (2), 502-516.
31. Bryan, A.C. et al. (2018) A variable polyglutamine repeat affects subcellular localization and regulatory activity of a *Populus* ANGUSTIFOLIA protein. G3: Genes|Genomes|Genetics 8 (8), 2631-2641.
32. Tuskan, G.A. et al. (2018) Defining the genetic components of callus formation: A GWAS approach. PLOS ONE 13 (8), e0202519.