

94. A Systems Biology, Whole-Genome Association Analysis of The Molecular Regulation of Biomass Growth and Composition in *Populus deltoides*

Annette Fahrenkrog,¹ Leandro Neves,¹ Juan Acosta,² Brad Barbazuk,^{1,3,4} and Matias Kirst^{1,2,3*}
(mkirst@ufl.edu)

¹Plant Molecular and Cellular Biology Program, University of Florida, Gainesville; ²School of Forest Resources and Conservation, University of Florida, Gainesville; ³Genetics Institute, University of Florida, Gainesville; ⁴Biology Department, University of Florida, Gainesville.

Project Goals: The overall goal of this project is to identify genetic polymorphisms that regulate biomass productivity and composition in poplar trees. The aims proposed to identify the genetic variants of interest are: 1) optimization of sequence-capture for unbiased, high-throughput and low-cost recovery of target coding and regulatory sequences in *P. deltoides*; 2) “whole-genome” genotyping of a *P. deltoides* unstructured population for association mapping; and 3) identification of significant SNP-trait associations with biomass growth and carbon partitioning to define genes and alleles that regulate trait variation. The genes identified with this approach will be used for the implementation of efficient molecular breeding and germplasm selection strategies, accelerating poplar breeding.

Poplars are fast growing trees with high potential to become a major feedstock for the biofuel industry. For these trees to be widely adopted for biofuel production, genotypes with specialized wood quality and growth traits are required. A better understanding of the genetic regulation of biomass growth and wood composition is needed to accelerate the development of new poplar clones through genomics-assisted breeding. Despite their importance, little is known about the genes that regulate these economically important traits. We are using a genome-wide association genetics approach to discover these genes in eastern cottonwood (*P. deltoides*), a poplar species with a wide distribution in North America and high levels of genetic and phenotypic variation.

In order to genotype a *P. deltoides* association population composed of 579 individuals, we used the *P. trichocarpa* version 2.2 reference genome to design probes for sequence capture. A total of 227,943 RNA-based 120 nucleotide long probes were designed to capture 23,835 intergenic regions (one probe every 15 Kb) and 18,153 genes (exons and part of 5'- and 3'-UTRs) previously found to be expressed in vegetative tissues in poplar. The targeted region corresponds to 27.35 Mb (5.7%) of the poplar genome. Sequence capture was performed on pools of 12 samples, and the captured DNA was sequenced using the Illumina HiSeq 2000 platform. We obtained an average of 27.3 million reads per genotype, and mean and median on-target depths of 26.1X and 25.6X, respectively. The percentage of bases sequenced at a depth above 15X (considered sufficient for identification of heterozygous variants) was 59.5%. Capture efficiency was high, with 100% of the genes and 96.3% of the intergenic regions captured.

Three different software (GATK HaplotypeCaller, FREEBAYES and SAMTOOLS) were used to call SNPs in the poplar association population, identifying 1.32 M, 1.31 M and 0.74 M SNPs respectively. In order to generate a high confidence SNP set, 529,628 SNPs identified by all three programs (with matching position and alternative allele calls) were selected. This consensus SNP set contains a high proportion (57.9%) of low frequency variants (MAF < 0.05). Rare variants (MAF < 0.005) constitute 23.3% of the total SNPs identified.

To avoid detection of spurious associations in the genome-wide association study (GWAS) we are carrying out to find significant marker-trait correlations, we assessed the presence of relatedness and population structure in our samples. The relatedness analysis performed with the software KING identified duplicated as well as closely related individuals that were excluded from further analyses. One

possibly hybrid individual was also removed. The remaining 430 unrelated samples were analyzed with the software STRUCTURE and the ΔK methodology, identifying two subpopulations that follow an east-west pattern. This structure needs to be accounted for when performing GWAS.

The association population was grown in a greenhouse and phenotyped for height, diameter, biomass production (separating leaves, stem and roots), and wood composition (lignin percentage, syringyl to guaiacyl lignin ratio and five-carbon and six-carbon sugar content). We are currently carrying out GWAS applying novel methods to detect marker-trait associations for rare variants, as well as traditional methods for common variants.

In summary, the multiplexed exome capture and resequencing protocol optimized in this project's first aim efficiently captured genes and intergenic regions and allowed genotyping the *P. deltoides* association population for a high number of SNP markers at reduced sequencing costs. This genotyping method is suited for the identification of rare variants. The consensus SNP set obtained after identifying variants with three different programs in the second aim of this project is expected to be of high quality. This approach has been used successfully in other studies to increase variant detection accuracy. With the necessary data already generated (genotypes, phenotypes and assessment of relatedness and structure in the population), the search for significant marker-trait associations underway in the project's third aim is currently under way and will be demonstrated in the investigator meeting.

This research was supported by the DOE Office of Science, Office of Biological and Environmental Research (BER), grant no. DE-SC0003893.