FINAL TECHNICAL REPORT (ARRA funding)

Title: A systems biology, whole-genome association analysis of the molecular regulation of biomass growth and composition in *Populus deltoides*

Project Number: DE-SC0003893

Period: Year 1-4

INTRODUCTION

Poplars trees are well suited for biofuel production due to their fast growing habit, favorable wood composition and adaptation to a broad range of environments. The availability of a reference genome sequence, ease of vegetative propagation and availability of transformation methods also make poplar an ideal model for the study of wood formation and biomass growth in woody, perennial plants. The objective of this project was to conduct a genome-wide association genetics study to identify genes that regulate bioenergy traits in *Populus deltoides* (eastern cottonwood). Association mapping was pursued by combining sequence-capture followed by high-throughput sequencing to genotype coding and regulatory sequences in the whole-genome. To identify genetic polymorphisms to be evaluated for the genetic regulation of biomass productivity and carbon partitioning we pursued the following goals:

(1) Optimize sequence-capture for unbiased, high-throughput and low cost recovery of target coding and regulatory sequences in P. deltoides

(2) Carry out "whole-genome" genotyping of a P. deltoides unstructured population for association mapping

This report represents research outcomes derived from PROJECT DE-SC0003893, from years 1-4. Results derived from PROJECT DE-SC0003893, year 5, are described in a separate report.

MATERIALS AND METHODS

Plant material and DNA extraction. The *Populus deltoides* Bartr. ex Marsh. (eastern cottonwood) population used in this study is composed of 579 individuals sampled in 15 states in Central, Southern and Eastern US and maintained at the University of Florida. Leaves were collected from each genotype, lyophilized in a freeze-drier (Freezezone 18L Bulk Tray Dryer, Labconco, Kansas City, MO, USA) and ground in a Geno/Grinder tissue homogenizer (SPEX SamplePrep, Metuchen, NJ, USA). Genomic DNA (gDNA) was extracted from the lyophilized leaf tissue using the DNeasy Plant Mini Kit (Qiagen, Valencia, CA, USA).

Probe design for sequence capture. A set of 227,943 RNA-based 120 nucleotide long probes was designed to capture 23,835 intergenic regions (with 23,835 probes) and 18,153 genes (with 204,108 probes), previously identified as expressed in poplar vegetative tissues (Quesada *et al.*, 2008). A hybrid *P. trichocarpa/P. deltoides* reference genome was generated modifying the *P. trichocarpa* version 2.2 (v2.2) reference based on a collection of random shotgun sequences of a single *P. deltoides* individual. Non-overlapping probes targeting exons, 5' and 3' UTR's and putative regulatory regions spanning 500 bp upstream of transcription start sites were designed based on this reference. In instances where exons were smaller than 120 bp, probe sequences expanded into introns. Each gene was targeted with an average of 11.2 probes. Besides the target genes, a set of 22,864 probes was designed to capture intergenic regions spaced evenly along the poplar genome in 15 Kbp windows. Finally, 971 intergenic probes were obtained from a capture set designed for *P. trichocarpa* (Zhou & Holliday, 2012; Zhou *et al.*, 2014). The total targeted region corresponds to 27.35 Mb (5.7%) of the poplar genome.

Library preparation and sequencing. Sequencing libraries were prepared following Bentley *et al*. (2008), with modifications. Genomic DNA was sheared to an average size of 200 bp using a Covaris E210 machine (Covaris, Woburn, MA, USA). The ends of the sheared DNA molecules were repaired using the End-It DNA End-Repair kit (Epicentre Biotechnologies, Madison, WI, USA), and tailed with a single A using 15U Klenow fragment (3′ to 5′ exo-; New England Biolabs Inc., Ipswich, MA, USA) and 0.2 mM dATP (Promega, Madison, WI, USA). The A-tailed DNA was ligated to custom adapters that included a barcode sequence at the 5' end, using the Fast-Link DNA Ligation kit (Epicentre Biotechnologies, Madison, WI, USA). PCR enrichment was performed for 10 cycles using Illumina paired-end primers and Phusion High-

Fidelity DNA Polymerase (New England Biolabs Inc., Ipswich, MA, USA). The samples were purified using the Agencourt AMPure XP bead purification kit (Beckman Coulter, Brea, CA, USA) after each step of the protocol. After PCR enrichment, the libraries were quantified with the Quant-iT PicoGreen dsDNA Assay Kit (Invitrogen, Waltham, MA, USA). A total of 48 pools of gDNA libraries were prepared before sequence capture – 45 pools contained 12 samples and three pools contained 13 samples. The barcoded libraries were mixed in equimolar amounts to obtain a total of 750 ng of DNA in each pool. Multiplexed sequence capture was performed using the SureSelect Target Enrichment kit (Agilent Technologies, Santa Clara, CA, USA), following the manufacturer's protocol. The captured DNA from each pool was sequenced in paired-end mode for 100 cycles in a separate lane of an Illumina HiSeq 2000 Sequencing System (Illumina, San Diego, CA, USA). Sequences are available in the National Center for Biotechnology Information Short Read Archive (www.ncbi.nlm.nih.gov; Accession number: SRP066162).

Sequencing data processing. The FASTX-toolkit 0.0.13 (http://hannonlab.cshl.edu/fastx_toolkit) was used to split the sequencing reads from each lane according to their barcode sequence, generating separate files for each sample present in the pool. The barcode sequence was removed from the reads and, if present, bases with quality below a PHRED score of 20 were trimmed from the 3' end of the read. Only reads longer than 50 nucleotides were kept and were further filtered by quality, retaining only reads with 90% of their sequence with PHRED scores above 20. The filtered reads were aligned to the hybrid *P. trichocarpa*/*P. deltoides* reference genome using MOSAIK 2.2 (Lee *et al.*, 2014b), using the following parameters: maximum percentage of mismatches (-mmp) 0.2, local alignment search for paired-end reads enabled in a 500 bp radius (-ls 500), and hash size (–hs) 15. Reads that lost a mate during quality filtering were aligned as single-end reads. Finally, the coverage tool available in BEDTOOLS 2.22.1 (http://bedtools.readthedocs.org/) was used to assess on-target depth. The data has been deposited in the Sequence Read Archive (SRA, Accession Number SRP066162).

Variant identification. The data pre-processing steps recommended in the GATK best practices workflow (DePristo *et al.*, 2011) were performed prior to variant identification. PCR duplicates

were marked with PICARD 1.115 (http://broadinstitute.github.io/picard) and indel realignment (IR) and base quality score recalibration (BQSR) were performed with GATK 3.1 (McKenna *et al.*, 2010; DePristo *et al.*, 2011; Van der Auwera *et al.*, 2013). Sets of known indels and SNPs were developed for this population to be used for IR and BQSR, respectively. SNPs and indels were identified in multi-sample calling mode with SAMTOOLS 1.1 (Li, 2011) and FREEBAYES 0.9.15 (Garrison & Marth, 2012) (parameters --theta 0.004 --min-alternate-count 2, --min-alternate-fraction 0.3, --min-coverage 6). Variants were also identified with GATK 3.1 using the HaplotypeCaller tool in single-sample calling mode, followed by joint genotyping of the population with the GenotypeGVCFs tool. Indels were removed and, in order to generate a high-confidence SNP set, the SNPs overlapping all callers were selected using the vcf-isec tool from VCFTOOLS 0.1.12a (Danecek *et al.*, 2011). To be able to intersect the SNP set obtained with FREEBAYES with the SNP sets identified with the other two callers, FREEBAYES SNPs had to be processed with the vcfallelicprimitives and vcfcreatemulti tools from VCFLIB (https://github.com/ekg/vcflib) to decompose the complex variant representation used by FREEBAYES into standard SNP and indel format. Among the SNPs overlapping between callers, SNPs with discordant alternative base calls were removed using custom scripts. VCFTOOLS was used to filter the remaining SNPs, using SNP quality and mapping quality thresholds of 50 and 30, respectively, and recoding genotypes with a quality score below 20 and depth below eight as missing. This final SNP set is referred to "consensus SNPs" hereafter. The software SNPEFF 4.0 (Cingolani *et al.*, 2012) was used to annotate SNPs within exons for their predicted effect. Polymorphisms predicted to cause missense and nonsense mutations are referred to as "functional SNPs".

Relatedness and population structure. We investigated the presence of duplicated and related samples in our population with the software KING 1.4 (Manichaikul *et al.*, 2010, 2012), identifying a group of 500 putatively unrelated individuals. After removing samples with more than 15% missing data from this set and SNPs with more than 5% missing data, the genetic relationship matrix (GRM) was computed with the R package BGData (https://github.com/QuantGen/BGData) and additional samples with a kinship coefficient above 0.35 (on a scale from 0 - 1, with one representing clones) were removed resulting in a final population of 425 unrelated individuals. The genetic structure of this population was determined

with the software STRUCTURE 2.3.4 (Pritchard *et al.*, 2000) using 8,664 intergenic SNPs pruned to remove SNPs in linkage disequilibrium (r^2 threshold of 0.2). Subpopulation numbers (K) ranging from 1 to 10 were evaluated, repeating each analysis 10 times. Modeling decisions included the use of an admixture model and correlated allele frequencies. A burn-in of 20,000 iterations and 100,000 iterations of the Markov Chain Monte Carlo (MCMC) were used. The most likely number of subpopulations was inferred with the delta K method (Evanno *et al.*, 2005) implemented in the STRUCTURE HARVESTER program (Earl & VonHoldt, 2011). The CLUMPP program (Jakobsson & Rosenberg, 2007) was used to generate a consensus ancestry estimate from the 10 independent runs for each K and ancestry bar plots were generated with the POPHELPER package for R (R Core Team., 2015; Francis, 2016). For comparison, structure was also assessed with principal component analysis (PCA) using the SNPRelate package for R (Zheng *et al.*, 2012; R Core Team., 2015).

Phenotyping for growth and wood composition. Out of the 425 unrelated individuals composing the *P. deltoides* association population, a set of 391 *P. deltoides* genotypes was successfully phenotyped. Plants were clonally propagated from apical green cuttings in a glasshouse at the University of Florida. Plants were distributed in the glasshouse following a randomized complete block design with four blocks, each one containing one biological replicate per genotype. The plants were grown for 15 weeks with nutrient solution (Peters 20-20-20, at a rate of 4mM nitrogen, Scotts Co, Marysville, OH, USA) in a flood irrigation system of ebb-and-flow benches. Benches were flooded twice a day for 30 min. Height was recorded at weeks one to nine, 11, 13 and 15. At week 15, stem diameter was measured at 5 cm above the soil and the plants were harvested placing leaves and stem in separate barcoded paper envelopes to measure biomass. A 10 cm fraction of the stem was collected from the base of the plant for wood composition analysis, pealed to remove bark and phloem, and the xylem was dried in a freeze-drier (Freezezone 18L Bulk Tray Dryer, Labconco, Kansas City, MO, USA). Stem and leaf samples collected for biomass analysis were dried in a 65˚C drying room. Once dry, tissues were conditioned in the laboratory for more than one week before being weighed with high-precision balances (Ohaus Corp, Parsipanny, NJ, USA). The dry xylem tissue collected for wood composition analysis was coarsely ground in a coffee grinder and stored at room temperature in barcoded 20 ml HDPE scintillation vials (Thermo Fisher Scientific, Chicago, IL, USA). For the

391 phenotyped samples, material from three biological replicates was ground to a fine powder in a Geno/Grinder tissue homogenizer (SPEX SamplePrep, Metuchen, NJ, USA) for wood composition analysis. Ground samples were completely randomized and two subsamples (technical replicates) of approximately 4 mg from each biological replicate were scooped into 80 µl stainless steel Eco-cups (Frontier Laboratories Ltd., Fukushima, Japan) and covered with glass fiber paper type A/D (Pall Corporation, Ann Arbor, MI, USA). Wood composition was measured using pyrolysis-molecular beam mass spectrometry (py-MBMS) at the National Renewable Energy Laboratory (NREL, Golden, CO, USA). Analysis procedures followed the methods described previously (Novaes *et al.*, 2009). A spectrum ranging from 30 to 450 mass-to-charge (m z^{-1}) ratio was obtained for each sample analyzed. Peak intensity was mean normalized and technical replicates were averaged for each individual. Peaks associated with each wood component of interest were summed to estimate the proportion of these components in the wood sample. The following m z^{-1} peak intensities were summed per trait: i) total lignin: 120, 124, 137, 138, 150, 152, 154, 164, 167, 168, 178, 180, 181, 182, 194, 208 and 210; ii) syringyl lignin: 154, 167, 168, 182, 194, 208 and 210; iii) guaiacyl lignin: 124, 137, 138, 150, 164 and 178; iv) five-carbon sugars (C5): 57, 73, 85, 96 and 114; and v) six-carbon sugars (C6): 57, 60, 73, 98, 126 and 114. The syringyl (S) lignin peak sum was divided by the guaiacyl (G) lignin peak sum to obtain an estimate of lignin S:G ratio. For each sample analyzed, NREL also reported the estimated lignin percentage calculated using the known lignin percentage of standard samples (previously characterized with wet chemistry) included in each run.

Statistical analysis of phenotypic data. Phenotypic data (height at week 15; diameter at week 15; leaf biomass; stem biomass; lignin percentage; lignin S:G ratio; C5 m z^{-1} peak intensity sum and C6 m z^{-1} peak intensity sum) was analyzed using ASREML version 3.0 (Gilmour *et al.*, 2006). A log transformation was applied to C5 peak intensity sum and C6 peak intensity sum because of the non-normal distribution of residuals. Data points where the residuals deviated more than three standard deviations from the mean were considered outliers and excluded from downstream analyses. The following linear mixed model was used to obtain an accurate trait estimate for each sample, while accounting for experimental effects:

$$y_{ijkl} = \mu + R_i + b_{j(i)} + g_k + e_{ijkl}$$

where y_{ijkl} corresponds to the measure of the trait in the *l*th plant of the *k*th genotype in the *j*th bench within the *i*th block, μ is the population mean, R_i is the fixed effect of block, $b_{j(i)}$ is the random effect of bench within block, g_k is the random effect of genotype assuming $g_k \sim N(0, I\sigma_g^2)$ and e_{ijkl} is the random error effect within the experiment. Heritability of each trait was obtained from the variance components estimated by ASREML, using the following equation:

$$h^2 = \frac{\sigma_g^2}{\left(\sigma_b^2 + \sigma_g^2 + \sigma_e^2\right)}$$

where σ^2_g, σ^2_b, and σ^2_e are the variance components corresponding to genotype, bench within block, and residual effects.

RESULTS

Efficient capture and sequencing of the gene space in a natural population of *P. deltoides*. We genotyped a *P. deltoides* population composed of 579 individuals by targeted resequencing, capturing 18,153 genes and 23,835 intergenic regions. The captured DNA was sequenced with the Illumina HiSeq 2000 platform, generating an average of 27.3 million reads per individual. Reads were aligned to a hybrid *P. trichocarpa*/*P. deltoides* reference genome using MOSAIK 2.2 (Lee *et al.*, 2014b), obtaining mean and median on-target coverage of 30.3× and 28.9×, respectively. Among the on-target bases sequenced, 76.3% reached a mean sequencing depth greater than 13×, shown to be sufficient to detect heterozygous SNPs with 95% probability in exome sequencing studies (Meynert *et al.*, 2013). Overall efficiency of the enrichment strategy was high, capturing 100% of the targeted genes and 96.3% of the intergenic regions.

Low-frequency variants represent a high proportion of polymorphic loci in *P. deltoides*. Genetic polymorphisms were identified in the 579 samples using the HaplotypeCaller tool from GATK 3.1 (McKenna *et al.*, 2010; DePristo *et al.*, 2011; Van der Auwera *et al.*, 2013),

FREEBAYES 0.9.15 (Garrison & Marth, 2012) and SAMTOOLS 1.1 (Li, 2011), detecting 1,385,938 (198,139); 1,483,939 (147,995); and 746,338 (219,817) SNPs (indels), respectively. A high-confidence SNP set was obtained by excluding indels and intergenic SNPs, and retaining only those DNA variants identified by all three methods. A consensus SNP set increases variant detection accuracy, with low false-positive rates (Zook *et al.*, 2014; Baes *et al.*, 2014; Trubetskoy *et al.*, 2015). While this approach will discard true SNPs that were not identified by all three SNP callers, it minimizes the evaluation of associations with false SNPs. These SNPs were used to identify unrelated samples, generating a final dataset for GWAS consisting in an association population composed of 391 unrelated individuals with phenotypic information and a consensus SNP set containing 358,153 biallelic SNPs segregating in this population. A high proportion (48.2%) of these SNPs are low-frequency variants (MAF < 0.05). The sequence capture approach was highly successful in retrieving DNA variant information for the vast majority of the genomic regions of interest, with 17,972 out of 18,153 genes (99.0%) containing at least one consensus SNP.

Evidence of population structure in *P. deltoides*. Genetic structure of the association population was assessed with STRUCTURE (Pritchard *et al.*, 2000) using 8,664 intergenic consensus SNPs. According to the delta K method (Evanno *et al.*, 2005), the most likely number of subpopulations (K) was two, followed by a smaller delta K peak at K=6. Nevertheless, visual inspection of the ancestry coefficient bar plots for K ranging from two to six (Fig. 1a) and comparison of the proposed subpopulations with geographic origin of the samples where this information was available (168 individuals) revealed that the population is probably divided into four subpopulations (Fig. 1c). This finding was confirmed by population structure analysis with PCA, where the first two principal components (respectively explaining 1.99% and 0.90% of the genetic variance in the population, Fig. S2) clearly separated the four subpopulations (Fig. 1b). The structure pattern, with subpopulations identified along a longitudinal gradient, is consistent with a scenario of isolation by distance.

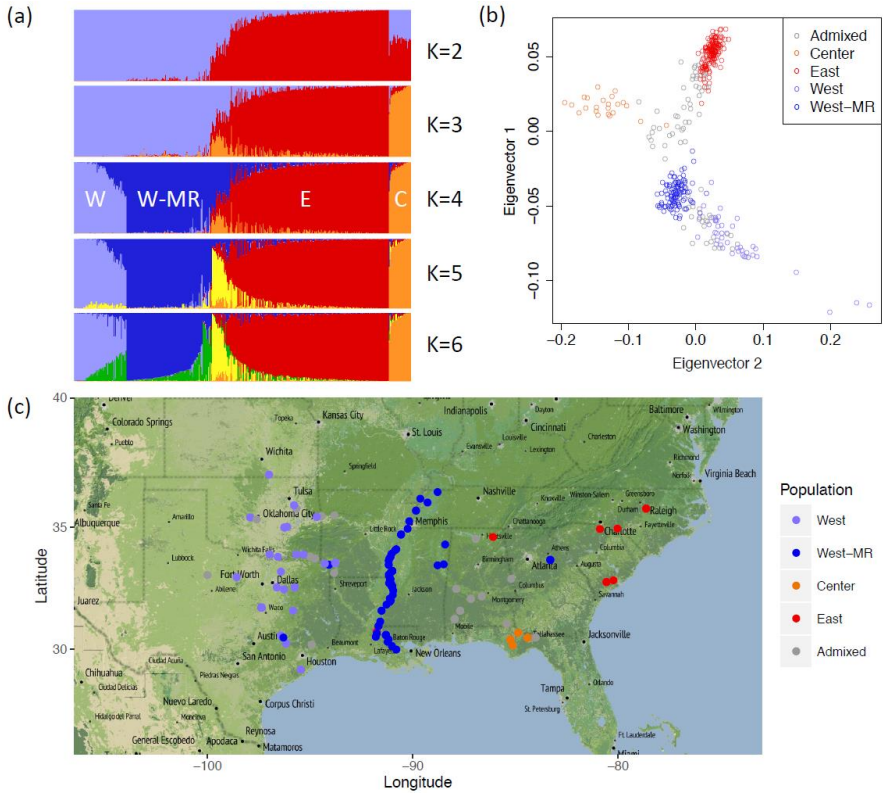

Figure 1: *Populus deltoides* population structure. **(a)** Ancestry coefficient bar plots obtained with STRUCTURE for an assumed number of subpopulations (K) from two to six. The four subpopulations considered to correct for population structure in the association analyses are labeled in the bar plot for K = 4. E: East subpopulation, C: Center subpopulation, W: West subpopulation, W-MR: West - Mississippi River subpopulation. **(b)** Population structure revealed by principal component analysis and **(c)** map showing 168 individuals with known sampling location. Samples in **(b)** and **(c)** are colored according to the subpopulation they were assigned to by STRUCTURE at K = 4.

REFERENCES

Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Angel G, Levy-moonshine A, Shakir K, Roazen D, Thibault J, Banks E, *et al.* 2013. From FastQ Data to High-Confidence Variant Calls : The Genome Analysis Toolkit Best Practices Pipeline. *Current Protocols in Bioinformatics* **43**: 11.10.1–11.10.33.

Baes CF, Dolezal M a, Koltes JE, Bapst B, Fritz-Waters E, Jansen S, Flury C, Signer-Hasler H, Stricker C, Fernando R, *et al.* 2014. Evaluation of variant identification methods for whole genome sequencing data in dairy cattle. *BMC Genomics* **15**: 948.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, *et al.* 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**: 53–59.

Betts MJ, Russell RB. **2003**. Amino-Acid Properties and Consequences of Substitutions. In: Barnes MR, Gray IC, eds. Bioinformatics for Geneticists: A Bioinformatics Primer for the Analysis of Genetic Data. Chichester, UK: John Wiley & Sons, Ltd., 311–342.

Bouquet A, Juga J. **2013**. Integrating genomic selection into dairy cattle breeding programmes: a review. *Animal : An International Journal of Animal Bioscience* **7**: 705–713.

Brachi B, Morris GP, Borevitz JO. **2011**. Genome-wide association studies in plants: the missing heritability is in the field. *Genome Biology* **12**: 232.

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. **2012**. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w 1118; iso-2; iso-3. *Fly* **6**: 80–92.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, *et al.* 2011. The variant call format and VCFtools. *Bioinformatics (Oxford, England)* **27**: 2156–2158.

DePristo MA, Banks E, Poplin R, Garimella K V, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, *et al.* 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* **43**: 491–498.

Earl DA, VonHoldt BM. **2011**. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources* **4**: 359–361.

Edwards MD, Symbor-Nagrabska A, Dollard L, Gifford DK, Fink GR. **2014**. Interactions between chromosomal and nonchromosomal elements reveal missing heritability. *Proceedings of the National Academy of Sciences of the United States of America* **111**: 7719–7722.

Evanno G, Regnaut S, Goudet J. **2005**. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology* **14**: 2611–2620.

Evans LM, Slavov GT, Rodgers-Melnick E, Martin J, Ranjan P, Muchero W, Brunner AM, Schackwitz W, Gunter L, Chen J-G, *et al.* 2014. Population genomics of *Populus trichocarpa* identifies signatures of selection and adaptive trait associations. *Nature Genetics* **46**: 1089–1096.

Francis RM. **2016**. POPHELPER: An R package and web app to analyse and visualise population structure. *Molecular Ecology Resources*: doi: 10.1111/1755–0998.12509.

Garrison E, Marth G. **2012**. Haplotype-based variant detection from short-read sequencing. *arXiv preprint*: arXiv:1207.3907.

Gibson G. **2012**. Rare and common variants: twenty arguments. *Nature Reviews. Genetics* **13**: 135–145.

Gilmour AR, Gogel BJ, Cullis BR, Thompson R. **2006**. *ASReml User Guide Release 2.0*. VSN International Ltd, Hemel Hempstead, HP1 1ES, UK.

Guerra FP, Wegrzyn JL, Sykes R, Davis MF, Stanton BJ, Neale DB. **2013**. Association genetics of chemical wood properties in black poplar (*Populus nigra*). *New Phytologist* **197**: 162–176.

Hefer CA, Mizrachi E, Myburg AA, Douglas CJ, Mansfield SD. **2015**. Comparative interrogation of the developing xylem transcriptomes of two wood-forming species : *Populus trichocarpa* and *Eucalyptus grandis*. *New Phytologist* **206**: 1391–1405.

Huang X, Han B. **2014**. Natural variations and genome-wide association studies in crop plants. *Annual Review of Plant Biology* **65**: 531–551.

Ionita-Laza I, Lee S, Makarov V, Buxbaum JD, Lin X. **2013**. Sequence kernel association tests for the combined effect of rare and common variants. *American Journal of Human Genetics* **92**: 841–853.

Jakobsson M, Rosenberg NA. **2007**. CLUMPP: A cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* **23**: 1801–1806.

Kelley LA, Sternberg MJE. **2009**. Protein structure prediction on the Web: a case study using the Phyre server. *Nature Protocols* **4**: 363–371.

Lee S, Abecasis GR, Boehnke M, Lin X. **2014a**. Rare-Variant Association Analysis: Study Designs and Statistical Tests. *The American Journal of Human Genetics* **95**: 5–23.

Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, Christiani DC, Wurfel MM, Lin X. **2012**. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *American Journal of Human Genetics* **91**: 224–37.

Lee W-P, Stromberg MP, Ward A, Stewart C, Garrison EP, Marth GT. **2014b**. MOSAIK: a hash-based algorithm for accurate next-generation sequencing short-read mapping. *PloS one* **9**: e90581.

Li H. **2011**. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics (Oxford, England)* **27**: 2987–2993.

Li B, Leal SM. **2008**. Methods for Detecting Associations with Rare Variants for Common Diseases : Application to Analysis of Sequence Data. *The American Journal of Human Genetics* **83**: 311–321.

Lin Z, Hayes BJ, Daetwyler HD. **2014**. Genomic selection in crops , trees and forages : a review. *Crop and Pasture Science* **65**: 1177–1191.

Lipka AE, Kandianis CB, Hudson ME, Yu J, Drnevich J, Bradbury PJ, Gore MA. **2015**. From association to prediction: statistical methods for the dissection and selection of complex traits in plants. *Current Opinion in Plant Biology* **24**: 110–118.

Luna A, Nicodemus KK. **2007**. snp.plotter: An R-based SNP/haplotype association and linkage

disequilibrium plotting package. *Bioinformatics* **23**: 774–776.

MacKay JJ, O'Malley DM, Presnell T, Booker FL, Campbell MM, Whetten RW, Sederoff RR. **1997**. Inheritance, gene expression, and lignin characterization in a mutant pine deficient in cinnamyl alcohol dehydrogenase. *Proceedings of the National Academy of Sciences of the United States of America* **94**: 8255–8260.

Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen W-M. **2010**. Robust relationship inference in genome-wide association studies. *Bioinformatics (Oxford, England)* **26**: 2867–2873.

Manichaikul A, Palmas W, Rodriguez CJ, Peralta CA, Divers J, Guo X, Chen WM, Wong Q, Williams K, Kerr KF, *et al.* 2012. Population structure of hispanics in the United States: The multi-Ethnic study of Atherosclerosis. *PLoS Genetics* **8**: e1002640.

Marroni F, Pinosio S, Zaina G, Fogolari F, Felice N, Cattonaro F, Morgante M. **2011**. Nucleotide diversity and linkage disequilibrium in *Populus nigra* cinnamyl alcohol dehydrogenase (CAD4) gene. *Tree Genetics and Genomes* **7**: 1011–1023.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, *et al.* 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**: 1297–1303.

Mckown AD, Guy RD, Klápště J, Geraldes A, Friedmann M, Cronk QCB, El-Kassaby YA, Mansfield SD, Douglas CJ. **2014**. Geographical and environmental gradients shape phenotypic trait variation and genetic structure in *Populus trichocarpa*. *New Phytologist* **201**: 1263–1276.

McKown AD, Klápště J, Guy RD, Geraldes A, Porth I, Hannemann J, Friedmann M, Muchero W, Tuskan GA, Ehlting J, *et al.* 2014. Genome-wide association implicates numerous genes underlying ecological trait variation in natural populations of *Populus trichocarpa*. *New Phytologist* **203**: 535–553.

Meynert AM, Bicknell LS, Hurles ME, Jackson AP, Taylor MS. **2013**. Quantifying single nucleotide variant detection sensitivity in exome sequencing. *BMC Bioinformatics* **14**: 195.

Muchero W, Guo J, DiFazio SP, Chen J-G, Ranjan P, Slavov GT, Gunter LE, Jawdy S, Bryan AC, Sykes R, *et al.* 2015. High-resolution genetic mapping of allelic variants associated

with cell wall chemistry in *Populus*. *BMC Genomics* **16**: 1–14.

Novaes E, Osorio L, Drost DR, Miles BL, Boaventura-Novaes CRD, Benedict C, Dervinis C, Yu Q, Sykes R, Davis M, *et al.* 2009. Quantitative genetic analysis of biomass and wood chemistry of *Populus* under different nitrogen levels. *New Phytologist* **182**: 878–890.

Petit RJ, Hampe A. **2006**. Some Evolutionary Consequences of Being a Tree. *Annual Review of Ecology, Evolution, and Systematics* **37**: 187–214.

Pfeifer B, Wittelsburger U, Ramos-Onsins SE, Lercher MJ. **2014**. PopGenome: An Efficient Swiss Army Knife for Population Genomic Analyses in R. *Molecular Biology and Evolution* **31**: 1929–1936.

Porth I, El-Kassaby YA. **2015**. Using *Populus* as a lignocellulosic feedstock for bioethanol. *Biotechnology Journal* **10**: 510–524.

Porth I, Klápště J, Skyba O, Friedmann MC, Hannemann J, Ehlting J, El-Kassaby YA, Mansfield SD, Douglas CJ. **2013a**. Network analysis reveals the relationship among wood properties, gene expression levels and genotypes of natural *Populus trichocarpa* accessions. *New Phytologist* **200**: 727–742.

Porth I, Klapšte J, Skyba O, Hannemann J, McKown AD, Guy RD, Difazio SP, Muchero W, Ranjan P, Tuskan GA, *et al.* 2013b. Genome-wide association mapping for wood characteristics in *Populus* identifies an array of candidate single nucleotide polymorphisms. *New Phytologist* **200**: 710–726.

Pritchard JK, Stephens M, Donnelly P. **2000**. Inference of population structure using multilocus genotype data. *Genetics* **155**: 945–959.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira M a R, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, *et al.* 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* **81**: 559–575.

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

R Core Team. 2015. R: A Language and Environment for Statistical Computing. R Foundation

for Statistical Computing. Vienna, Austria.

Robinson MR, Wray NR, Visscher PM. **2014**. Explaining additional genetic variation in complex traits. *Trends in Genetics* **30**: 124–132.

Sannigrahi P, Ragauskas AJ, Tuskan GA. **2010**. Poplar as a feedstock for biofuels : A review of compositional characteristics. *Biofuels, Bioproducts and Biorefining* **4**: 209–226.

Sjödin A, Street NR, Sandberg G, Gustafsson P, Jansson S. **2009**. The *Populus* Genome Integrative Explorer (PopGenIE): A new resource for exploring the *Populus* genome. *New Phytologist* **182**: 1013–1025.

Stanton BJ, Neale DB, Li S. **2010**. *Populus* Breeding: From the Classical to the Genomic Approach. In: Jansson S, Bhalerao RP, Groover AT, eds. Genetics and Genomics of *Populus*. New York, NY, USA: Springer New York, 309–348.

Trubetskoy V, Rodriguez A, Dave U, Campbell N, Crawford EL, Cook EH, Sutcliffe JS, Foster I, Madduri R, Cox NJ, *et al.* 2015. Consensus Genotyper for Exome Sequencing (CGES): improving the quality of exome variant genotypes. *Bioinformatics (Oxford, England)* **31**: 187–193.

Turner SD. **2014**. qqman: an R package for visualizing GWAS results using QQ and manhattan plots. *bioRxiv*: 5165.

Vanholme B, Cesarino I, Goeminne G, Kim H, Marroni F, Van Acker R, Vanholme R, Morreel K, Ivens B, Pinosio S, *et al.* 2013. Breeding with rare defective alleles (BRDA): a natural *Populus nigra* HCT mutant with modified lignin as a case study. *New Phytologist* **198**: 765–76.

Wegrzyn JL, Eckert AJ, Choi M, Lee JM, Stanton BJ, Sykes R, Davis MF, Tsai C-J, Neale DB. **2010**. Association genetics of traits controlling lignin and cellulose biosynthesis in black cottonwood (*Populus trichocarpa, Salicaceae*) secondary xylem. *New Phytologist* **188**: 515–532.

Weng J-K, Chapple C. **2010**. The origin and evolution of lignin biosynthesis. *New Phytologist* **187**: 273–285.

Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. **2011**. Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. *The American Journal of Human Genetics* **89**: 82–93.

Xu S. **2003**. Theoretical Basis of the Beavis Effect. *Genetics* **165**: 2259–2268.

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW*, et al.* 2010. Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics* **42**: 565–569.

Zasada JC, David AJ, Gilmore DW, Landhausser SM. **2001**. Ecology and silviculture of natural stands of *Populus* species. In: Dickman DI, Isebrands JG, Eckenwalder JE, Richardson J, eds. Poplar Culture in North America. Ottawa, Canada: NRC Research Press, 119–151.

Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. **2012**. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**: 3326–3328.

Zhou L, Bawa R, Holliday JA. **2014**. Exome resequencing reveals signatures of demographic and adaptive processes across the genome and range of black cottonwood (Populus trichocarpa). *Molecular Ecology* **23**: 2486–99.

Zhou L, Holliday JA. **2012**. Targeted enrichment of the black cottonwood (Populus trichocarpa) gene space using sequence capture. *BMC Genomics* **13**: 703.

Zook JM, Chapman B, Wang J, Mittelman D, Hofmann O, Hide W, Salit M. **2014**. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nature Biotechnology* **32**: 246–251.