FINAL TECHNICAL REPORT (non-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 5

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 detect genetic polymorphisms that can be applied to accelerate breeding for bioenergy we pursued the following goal:

(1) Identifying significant SNP-trait associations with biomass growth and carbon partitioning to define genes and alleles that regulate trait variation.

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

MATERIALS AND METHODS

Tests for association between markers and phenotypes. Two GWAS strategies were followed to detect associations between SNP markers and traits in the 391 unrelated individuals with phenotypic data: i) Single-variant association tests, with higher power to detect marker-trait associations with common variants, were carried out with PLINK version 1.9 (Purcell *et al.*, 2007); and ii) Multiple-variant association test with higher power to detect associations with rare variants were performed with the R package SKAT version 1.0.1 (Wu *et al.*, 2011) (referred to as "SKAT package"). Only genic SNPs were included in all association analyses performed (PLINK and SKAT package). Association tests with PLINK used a linear model for analysis of quantitative traits and were done for the consensus SNPs with minor allele frequency (MAF) > 0.003835 (corresponding to SNPs present in at least two samples in the population), consensus SNPs with MAF ≥ 0.05 (referred to as "common SNPs" hereafter), and functional SNPs with MAF > 0.003835. Analysis with SKAT was performed using all consensus SNPs and functional SNPs, grouping variants by gene. Seven different association tests available in the SKAT package for R (Wu *et al.*, 2011; Lee *et al.*, 2012; Ionita-Laza *et al.*, 2013) were used: (i) sequence kernel association test (SKAT), (ii) burden test, (iii) optimal unified SKAT test (SKAT-O), (iv) combined sum test with SKAT test for common and low-frequency variants (SKAT-C), (v) combined sum test with burden test for common and low-frequency variants (Burden-C), (vi) adaptive sum test with SKAT test for common and low-frequency variants (SKAT-A), and (vii) adaptive sum test with burden test for common and low-frequency variants (Burden-A). Original SKAT and Burden tests are designed to identify association of rare variants with phenotype. SKAT is a variance component test that assigns a higher weight to rare variants in the statistical model compared with the weight assigned to common variants (Ionita-Laza *et al.*, 2013). The Burden test collapses all rare variants into one genetic score and tests it for association with the trait using a linear model (Lee *et al.*, 2014a). SKAT-O is an omnibus test that combines SKAT and Burden tests. It behaves like either one of the individual tests when a single test is more powerful (Lee *et al.*, 2012). The combined sum tests SKAT-C and Burden-C analyze common and rare variants separately and then combine the results into an overall test statistic in a way that both variant classes contribute to the test statistic equally. The adaptive sum tests SKAT-A and Burden-A differ from the combined sum tests in that they test for association between markers and traits using different weight parameters and then report the result with the lower p-value (Ionita-Laza *et al.*, 2013). Combined and adaptive sum tests define rare variants using a threshold that depends on sample size, considering variants with minor allele frequency (MAF) ≤ 1/√2n to be rare (1/√2n = 1/√(2 x 391) = 0.036 in this study). Detailed descriptions of the seven tests applied using the SKAT package are available in the articles published by Wu *et al*. (2011;

SKAT), Lee *et al*. (2012; SKAT-O) and Ionita-Laza *et al*. (2013; adaptive and combined sum tests). For a review see Lee *et al*. (2014).

All GWAS models (PLINK and SKAT package) included STRUCTURE ancestry as a covariate to correct for population structure, assuming four subpopulations. For comparison, six principal components were also used to correct for population structure. The results of this analysis are available in the Supporting Information. A 5% false discovery rate (FDR) threshold was used to correct for multiple testing. For all genes found to be significantly associated with a trait, their function and gene name matching the *P. trichocarpa* version 3.0 (v3.0) genome annotation were obtained from the PopGenIE (Sjödin *et al.*, 2009) database. For a reduced number of v2.2 gene names not found in PopGenIE, the corresponding v3.0 gene was identified aligning the gene sequence to the *P. trichocarpa* v3.0 genome using BLAST 2.2.29 (blastn, e-value ≤ $1x10^{-5}$). Linkage disequilibrium (LD) between SNPs significantly associated with a trait was assessed with PLINK 1.9 (Purcell *et al.*, 2007) and LD in genes or regions of interest was plotted using the R package snp.plotter (Luna & Nicodemus, 2007). Linkage disequilibrium decay with physical distance within genes that contained two or more consensus SNPs in the total population was estimated for markers with MAF ≥ 0.10. Pairwise LD between markers was calculated with PLINK 1.9 and LD decay was obtained following Marroni *et al.* (2011). Manhattan and q-q plots were generated with the R package qqman (Turner, 2014).

Detection of signatures of selection in genes associated with lignin percentage. Only genic SNPs identified with GATK were used for this analysis. This marker set was filtered removing SNPs with a quality score below 50, mapping quality below 30, strand bias (p-value ≤ 0.00001), and end distance bias (p-value ≤ 0.00001). Also, genotypes with a quality score below 20 and depth below eight were recoded as missing and SNPs with more than 25% missing data were excluded. Tajima's D analysis was performed by gene with the PopGenome package for R (Pfeifer *et al.*, 2014).

RESULTS

Identification of putative regulators of complex traits in *P. deltoides*. Genome-wide association analysis in crop plants has been done typically by sequentially testing the correlation between alleles at individual loci and traits of interest (Huang & Han, 2014). This approach is suitable for analysis of common variants, but has limited power to detect associations with low-frequency variants, unless a large population is available or the variant has a significant impact on the phenotype (Li & Leal, 2008; Lee *et al.*, 2014a). Thus, low-frequency variants are commonly excluded from GWAS (Lipka *et al.*, 2015). To address these limitations, methods that group low-frequency variants by genomic region can be used to increase the power to detect trait associations (Lee *et al.*, 2012). Because approximately half of the consensus SNPs obtained for the population under study are common SNPs and the other half are low frequency SNPs, both approaches described above were applied.

Single-variant and multiple-variant associations were tested with eight phenotypes measured in the association population. For each trait, the phenotype range (and heritability estimate) obtained for the 15 week old plants used in this study was: i) height: 10.6 - 113.9 cm (0.71); ii) diameter: 3.8 - 8.3 mm (0.51); iii) leaf biomass: 1.6 - 15.9 g (0.53); iv) stem biomass: 0.4 - 9.2 g (0.61); v) lignin percentage: 17.8 - 28.3 % (0.64); vi) lignin S:G ratio: 1.0 - 1.8 (0.63); vii) 5-carbon sugars: 20.6 - 29.3 mass to charge ratio (m z^{-1}) sum of peak intensities associated with this trait obtained with pyrolysis-molecular beam mass spectrometry (Py-MBMS) (0.41); and viii) 6-carbon sugars: 22.9 - 38.0 m z^{-1} peak intensity sum (0.51). Although trees were phenotyped at a very young age, a positive correlation (r = 0.3) was observed for height and diameter between the plants grown in the glasshouse and three year old clones of these plants currently growing in the field. Additionally, lignin percentage and lignin S:G ratio measured in this population are similar to the range of values reported for two year old *Populus nigra* trees (lignin percentage: 19.5 - 26.5 %; lignin S:G ratio: 1.3 - 2.1) grown in field conditions (Guerra *et al.*, 2013). This indicates a degree of overlap in the genetic control of these traits throughout developmental stages.

Single-marker association tests carried out with PLINK 1.9 (Purcell *et al.*, 2007) used three different SNP sets: consensus SNPs (including 334,679 SNPs with MAF > 0.003835, threshold that removed alleles detected less than three times in the population), common SNPs (185,526 consensus SNPs with MAF ≥ 0.05), and functional SNPs (76,804 consensus SNPs predicted by SNPEFF 4.0 (Cingolani *et al.*, 2012) to cause missense and nonsense mutations). The latter set was selected because of the higher probability of these SNPs to affect protein function and, potentially, plant phenotypes.

Analysis of common SNPs, representing loci that would be normally included in association studies, identified 22 genes (23 SNPs) associated with a phenotype at a 5% FDR significance level (Fig. 1, Table 1). A much larger number of genes were identified as associated with a trait when including low-frequency variants. A set of 240 genes (294 SNPs) was significantly associated with a trait when using the consensus SNPs in single-variant association tests, at a 5% FDR significance level. Out of these, 17 genes were also identified when analyzing common SNPs. Most of the identified genes, 211 out of 240, were associated with lignin percentage. Low-frequency polymorphisms caused 191 of these associations with lignin percentage and all of them were negatively correlated with the trait. The majority of associations detected with height (six out of seven total associations) and stem biomass (seven out of eight total associations) were also caused by low-frequency SNPs. To assess if there is evidence of selection acting on the genes containing low-frequency variants associated with lower lignin percentage, we estimated their Tajima's D and compared it to the mean value estimated for all genes characterized in this study. On average, Tajima's D was significantly lower (*p-value* = 0.0004) among the genes associated with lower lignin content (-1.242) relative to all the genes characterized in this study (-1.065).

Association tests carried out using the functional SNPs identified 83 genes (94 SNPs) significantly associated with a trait at a 5% FDR significance level. Among these, 78 genes were associated with lignin percentage, including 10 genes not identified when analyzing the consensus SNPs or only common SNPs. The proportion of significant associations detected out of the total number of SNPs tested was 19.6% higher in this set compared to the consensus SNPs (87 SNPs out of 76,804 SNPs tested for functional SNPs v/s 317 SNPs out of 334,679 SNPs tested for all consensus SNPs) at a significance threshold of 4.8×10^{-5}, corresponding to the highest significant p-value at 5% FDR for the consensus SNPs. These associations are of special interest because they have a higher probability of representing the causative polymorphism responsible for part of the phenotypic variation.

Sequence kernel association and burden tests identify associations missed by individual SNP analysis. Due to the presence of a high number of low frequency SNPs in *P. deltoides*, we applied association tests designed to assess the combined effect of these variants on traits. Variants were grouped by gene and seven association methods (Wu *et al.*, 2011) were utilized (see Materials and Methods section). SKAT, Burden and SKAT-O are designed to detect trait associations with low-frequency variants, while the combined and adaptive sum tests also have high power to detect associations with common variants (Ionita-Laza *et al.*, 2013). The power of the seven tests to detect associations with low-frequency SNPs depends on the genetic architecture of the trait and each test has particular scenarios where it performs better. Their respective advantages and disadvantages have been summarized elsewhere (Ionita-Laza *et al.*, 2013; Lee *et al.*, 2014a). Analysis of the consensus and functional SNPs with the SKAT package identified 62 and 60 associations, respectively, at a 5% FDR significance level. The vast majority of genes identified with the SKAT package were associated with lignin percentage (consensus SNPs: 57/62 associations, functional SNPs: 60/60 associations) and 22 associations were shared by both SNP sets. Also, the use of seven different multiple-marker association tests allowed the discovery of a large number of genes contributing to biomass previously not detected by the analysis of individual variants. Among the 100 total associations identified with the SKAT package, 51 were not identified in the marker-by-marker association analysis conducted with PLINK.

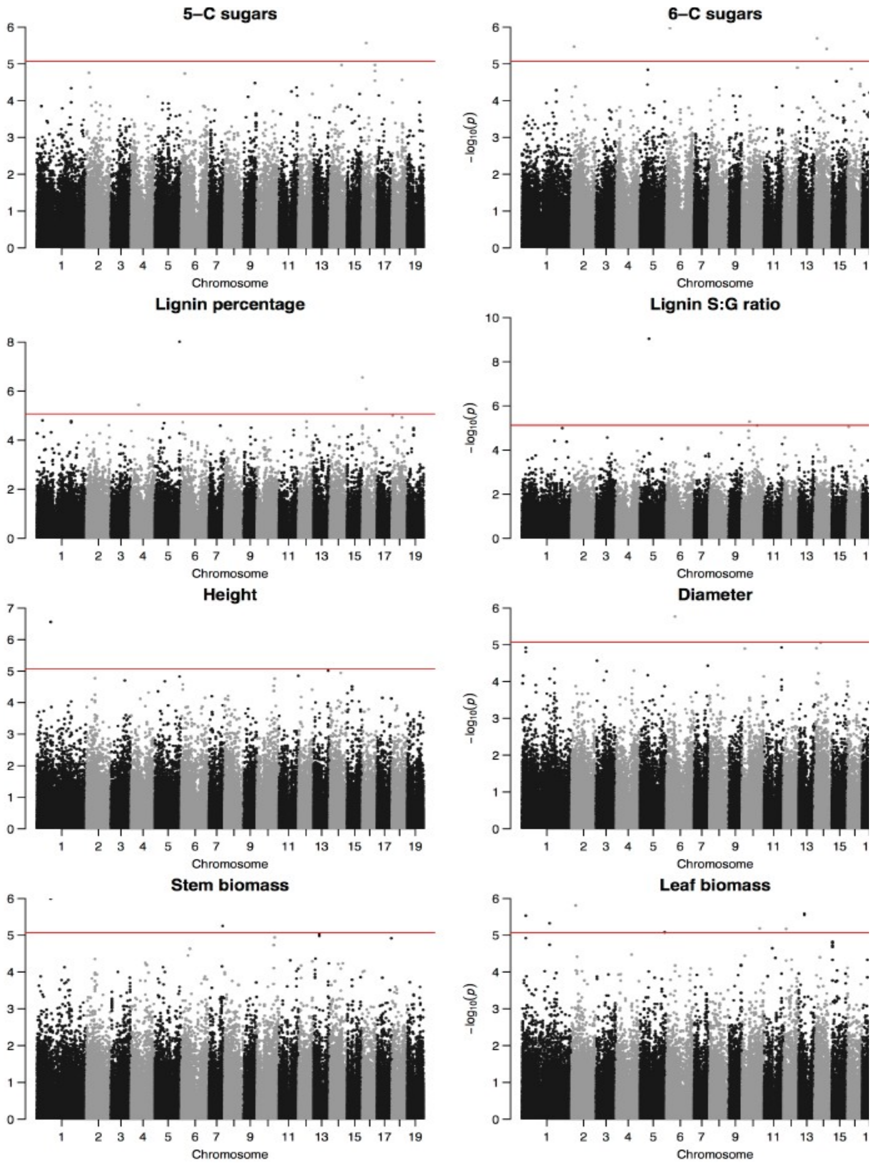

Figure 1: Manhattan plots for single-marker association tests using common SNPs, performed with eight traits in *Populus deltoides*. The red line in the Manhattan plots indicates a 5% FDR significance threshold.

Table 1: Significant trait associations with common SNPs identified by single-marker tests in *Populus deltoides*.

Trait	Gene v2.2 (v3.0)	SNP	Allele	Frequency	P-value	Q-value	Beta	Annotation
5-C sugars	POPTR_0016s05750 (Potri.016G057100)	scaffold_16:3693589	C	0.057	2.71E-06	0.020	-0.040	Transcription initiation factor
6-C sugars	POPTR_0002s03910 (Potri.002G038000)	scaffold_2:2450309	G	0.172	3.40E-06	0.023	0.034	Intracellular protein transport
6-C sugars	POPTR_0006s04850 (Potri.006G050000)	scaffold_6:3373462	A	0.102	1.06E-06	0.008	-0.056	Nodulin-like protein, nutrient transport, plant-microbe interactions
6-C sugars	POPTR_0014s02780 (Potri.014G027900)	scaffold_14:2312090	C	0.052	2.03E-06	0.014	0.062	Remorin, hormone and pathogen response
6-C sugars	POPTR_0014s15490 (Potri.014G156700)	scaffold_14:11541215	C	0.146	3.92E-06	0.025	0.041	Unknown
Diameter	POPTR_0006s10910 (Potri.006G108600)	scaffold_6:8278651	C	0.189	1.70E-06	0.012	0.368	Serine/threonine protein kinase
Height Stem biomass	POPTR_0001s16600 (Potri.001G165800)	scaffold_1:13379115	C	0.096	2.77E-07 1.01E-06	0.002 0.007	23.080 1.645	*Spt20*, chromatin remodelling
Leaf biomass	POPTR_0001s04530 (Potri.001G008000)	scaffold_1:3555495	G	0.079	2.95E-06	0.019	-1.327	Unknown
Leaf biomass	POPTR_0001s27920 (Potri.001G272300)	scaffold_1:26764684	A	0.076	4.74E-06	0.030	1.566	Unknown
Leaf biomass	POPTR_0002s06000 (Potri.002G059100)	scaffold_2:3984967	T	0.469	1.55E-06	0.011	1.031	Transcription factor
Leaf biomass Lignin percentage	POPTR_0005s25790 (Potri.005G236600)	scaffold_5:23971775	T	0.053	8.33E-06 9.67E-09	0.047 0.000	-1.852 -2.020	Methyltransferase
Leaf biomass	POPTR_0010s19410 (Potri.010G186800)	scaffold_10:17325446	A	0.051	6.55E-06	0.039	3.018	Unknown
Leaf biomass	POPTR_0012s03370 (Potri.012G036700)	scaffold_12:2666211	T	0.305	6.76E-06	0.039	-0.810	Protein degradation
Leaf biomass	POPTR_0013s06840 (Potri.013G071000)	scaffold_13:5530235	A	0.499	2.82E-06	0.019	-9.799	Unknown
		scaffold_13:5530250	G	0.499	2.59E-06	0.018	-9.817	
Lignin percentage	POPTR_0004s08740 (Potri.004G088700)	scaffold_4:7324296	C	0.051	3.66E-06	0.025	-1.374	Protein degradation
Lignin percentage	POPTR_0016s00260 (Potri.016G000600)	scaffold_16:29160	A	0.079	2.77E-07	0.002	-1.530	Protein degradation
Lignin percentage	POPTR_0016s05850 (Potri.016G058100)	scaffold_16:3775933	T	0.078	5.33E-06	0.034	-0.996	Unknown
Stem biomass	POPTR_0007s13210 (Potri.007G021800)	scaffold_7:13329788	A	0.232	5.60E-06	0.039	1.474	RNA helicase
Lignin S:G ratio	POPTR_0005s11950 (Potri.005G117500)	scaffold_5:8723671	T	0.174	9.10E-10	0.000	-0.079	*F5H3*, lignin biosynthesis
Lignin S:G ratio	POPTR_0010s05920 (Potri.010G049400)	scaffold_10:7348539	T	0.229	5.20E-06	0.036	0.058	Unknown

PERSONEL TRAININED IN THIS PROJECT

Graduate students:

Leandro Gomide Neves (PhD Plant Molecular and Cellular Biology, University of Florida)

Círtia Leite Ribeiro (PhD Plant Molecular and Cellular Biology, University of Florida)

Annette Fahrenkrog (PhD Plant Molecular and Cellular Biology, University of Florida)

Laboratory Technicians:

Madelyn Thiele

Gregory Brown

Christopher Dervinis

PUBLICATIONS RESULTING FROM THIS PROJECT

Fahrenkrog AM, Neves LG, Resende MF Jr, Vazquez AI, de Los Campos G, Dervinis C, Sykes R, Davis M, Davenport R, Barbazuk WB, Kirst M. Genome-wide association study reveals putative regulators of bioenergy traits in Populus deltoides. New Phytol. 2016 Sep 6. doi: 10.1111/nph.14154.

Ribeiro CL, Silva CM, Drost DR, Novaes E, Novaes CR, Dervinis C, Kirst M. Integration of genetic, genomic and transcriptomic information identifies putative regulators of adventitious root formation in Populus. BMC Plant Biol. 2016 Mar 16;16:66. doi: 10.1186/s12870-016-0753-0.

Drost DR, Puranik S, Novaes E, Novaes CR, Dervinis C, Gailing O, Kirst M. Genetical genomics of *Populus* leaf shape variation. BMC Plant Biol. 2015 Jun 30;15:166. doi: 10.1186/s12870-015-0557-7.

Harfouche A, Meilan R, Kirst M, Morgante M, Boerjan W, Sabatti M, Scarascia Mugnozza G. Accelerating the domestication of forest trees in a changing world. Trends Plant Sci. 2012 Feb;17(2):64-72. doi: 10.1016/j.tplants.2011.11.005.

CONFERENCE PROCEEDINGS RESULTING FROM THIS PROJECT

Fahrenkrog, A.M., L.G. Neves, J.J. Acosta, B. Barbazuk and M. Kirst. 2014. Poster presentation: Genotyping Populus deltoides by exome resequencing. Plant and Animal Genome Conference XXII, January 11-15, San Diego, USA.

Ribeiro, C.L., C. Dervinis, G.F. Peter, T. Martin and M. Kirst. 2014. Oral presentation: "Unknown Function" gene HC1 regulates growth and hydraulic conductivity in poplar trees. Plant and Animal Genome Conference XXII, January 11-15, San Diego, USA.

Ribeiro, C.L., B. Miles, T. Martin, G.F. Peter and M. Kirst. 2013. Oral presentation: HC1 regulates growth and hydraulic conductivity in poplar. Third International Plant Vascular Biology, July 26-30, Helsinki, Finland.

Fahrenkrog, A.M., L.G. Neves, J.J. Acosta, B. Barbazuk and M. Kirst. 2013. Poster presentation: Exome resequencing and population structure analysis in Populus deltoides. IUFRO Tree Biotechnology 2013 Conference, May 26-June 1, Asheville, USA.

Ribeiro, C.L., B. Miles, T. Martin, G.F. Peter and M. Kirst. 2013. Oral presentation: HC1 regulates growth and hydraulic conductivity in poplar. IUFRO Tree Biotechnology 2013 Conference, May 26-June 1, Asheville, USA.

Fahrenkrog, A.M., L.G. Neves, B. Barbazuk and M. Kirst. 2013. Poster presentation: Exome resequencing in a Populus deltoides association population. Plant and Animal Genome Conference XXI, January 12-16, San Diego, USA.

Ribeiro, C.L., B. Miles, T. Martin, G.F. Peter and M. Kirst. 2013. Oral presentation: HC1 regulates growth and hydraulic conductivity in poplar. Plant and Animal Genome Conference XXI, January 12-16, San Diego, USA.

Ribeiro, C.L., E. Novaes, C. Dervinis and M. Kirst. 2012. Poster: Functional analysis of a candidate gene involved in the regulation of biomass growth and carbon partitioning in Populus. Plant and Animal Genome Conference XIX, January 14-18, San Diego, USA.

INTELECTUAL PROPERTY RESULTING FROM THIS PROJECT

Kirst, M. Material and methods to increase plant growth and yield (Publication number: US 20140201867 A1)

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.