ASSOCIATION MAPPING OF ADAPTIVE AND BREEDING TRAITS IN EAST TEXAS LOBLOLLY PINE (*PINUS TAEDA* L.) BREEDING POPULATIONS USING HIGHT-DENSITY SNP GENOTYPING

Vikram E. Chhatre,[1] Thomas D. Byram, David B. Neale, Jill L. Wegrzyn, and Konstantin V. Krutovsky

[1]Department of Ecosystem Science and Management, Texas A&M University, College Station, TX

Loblolly pine (*Pinus taeda* L.) is the most commercially and ecologically important tree in the Southeastern US and is the main species in the Western Gulf Forest Tree Improvement Program (WGFTIP), one of the largest tree improvement programs in the US. Recent availability of genomic markers through the Conifer Translational Genomics Network (CTGN) has enabled a genome wide survey of population parameters in the WGFTIP loblolly pine breeding populations reported here.

Materials and Methods

The study included first- and second-generation selections from the WGFTIP East Texas breeding population. The first-generation selections were from natural stands and plantations originating at the western limit of the natural distribution of loblolly pine. The first-generation selections were subsequently partitioned into sublines and subjected to breeding and controlled pollination. Their progeny contributed the second-generation selections. Genome wide variation, population substructure and adaptive trait associations were investigated in both first- and second-generation populations using single nucleotide polymorphism (SNP) markers developed through the CTGN.

Genetic variation and its partitioning within the breeding populations were analyzed in 1,706 trees using 4,264 SNPs. These SNPs are based on amplicons representing partial sequences of ~3,000 expressed genes and were originally discovered in a small range-wide population set in the NSF funded ADEPT2 resequencing project. The tree samples were subdivided into 14 (first-generation) and 8 (second-generation) populations based on their geographical origin and 30 breeding groups based on their pedigree and the WGFTIP breeding strategy. Population structure was analyzed using Bayesian analysis as implemented in software *STRUCTURE* (Pritchard et al. 2000) and the ΔK parameter of Evanno et al. (2005). Individual computer runs for each putative cluster were permuted using LargeKGreedy algorithm implemented in the *CLUMPP* software (Jakobsson & Rosenberg 2007) and then visualized using the *DISTRUCT* (Rosenberg 2004). F_{ST} outlier method was used to detect candidate markers with alleles that were putatively affected by natural selection. The blast homology search was done to assess their functional significance. Significant associations between markers and adaptive traits were also studied using *TASSEL* (Bradbury et al. 2007). Haplotypes were reconstructed using the *fastPHASE* program (Scheet & Stephens 2006), and linkage disequilibrium (LD) between SNPs in all 12 linkage groups was estimated using HAPLOVIEW program (Barrett et al. 2005).

Results and Discussion

Population Substructure

Population structure appears to be weak as indicated by *STRUCTURE* analysis. The log probability of data suggests the number of clusters to be between 3 and 6, and ΔK suggests the number of clusters to be no more than 6 (Figure 1). Given the uniformity of the environment across the region and the limited area sampled, population structure appears to be mostly subtle if not an artifact of sampling.

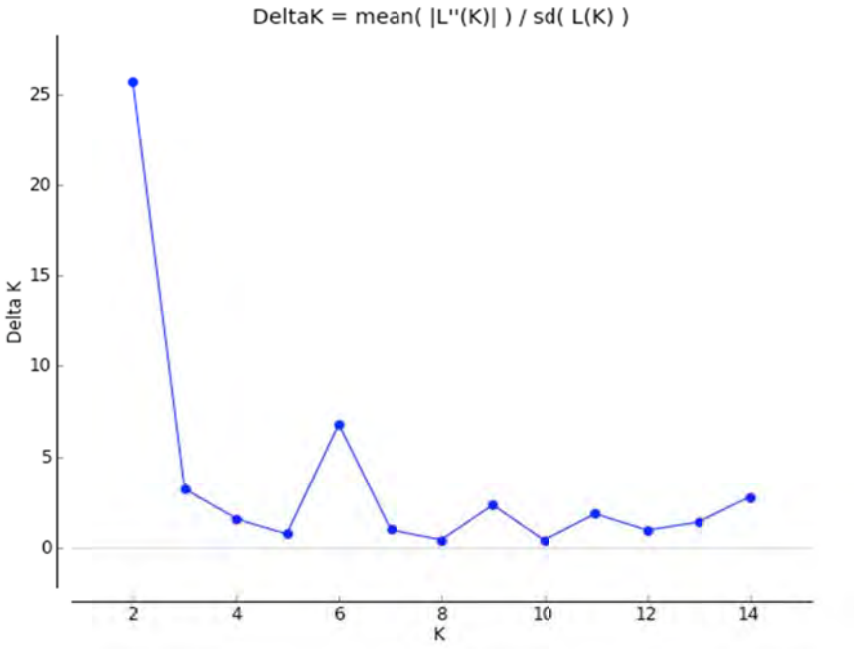

Figure 1. ΔK Estimator of population structure (Evanno et al. 2005).

Signatures of Natural Selection

F_{ST} outlier analysis for all 4,264 marker genotypes in the first generation samples revealed several markers that contributed to extremely high or extremely low F_{ST} estimates. Allelic variation in these markers demonstrates signatures of possible balancing or diversifying selection. Detailed functional annotation has been done for these markers.

Genome-Wide Linkage Disequilibrium

With approximately 100 SNPs per linkage group mapped using a relatively small segregating population (Eckert et al. 2009), the map resolution was insufficient to observe the rate of LD decay (Fig. 2). Most LDs were observed between closely linked SNPs, but there were a few significant LDs observed between markers separated by considerable distances. We hypothesize

that this could be due to several reasons including unaccounted family structure, population substructure, mapping errors and epistasis.

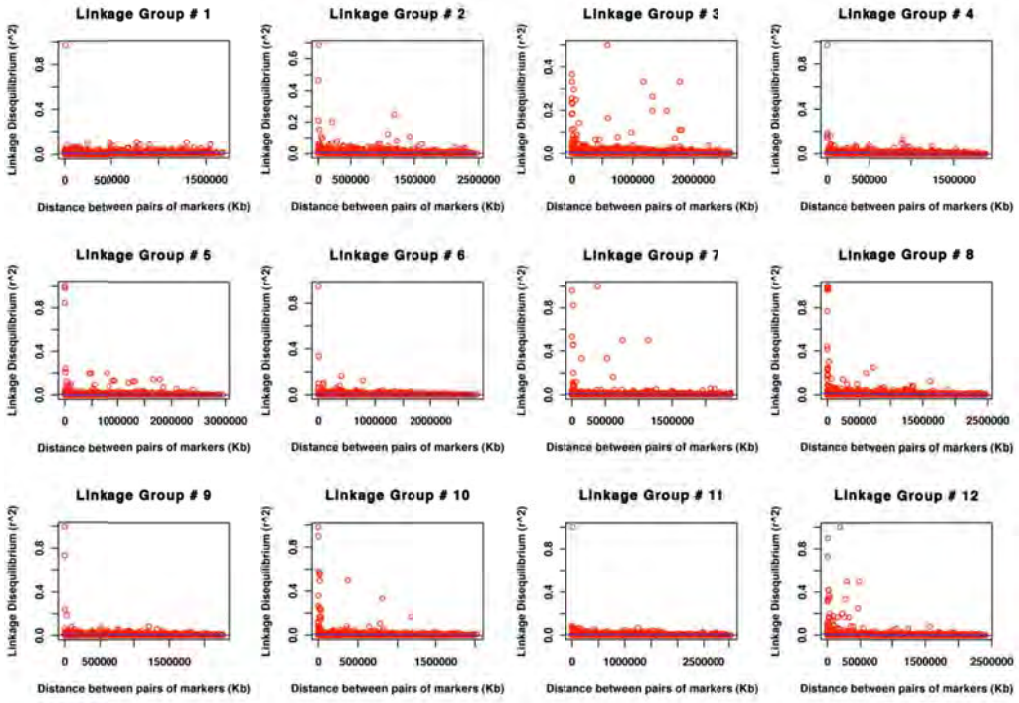

Figure 2. Pairwise LDs for SNPs mapped in 12 loblolly pine linkage groups.

Association Mapping Analysis

Significant associations of SNPs with several phenotypic traits such as height, diameter, survival on droughty sites, stem forking, wood specific gravity, etc. were detected. Sequences with SNPs that resulted in such associations were further annotated using their homology with functional genes in related and model species. Examples of associated genes based on BlastX functional annotation include decarboxylases, reductases, RNA polymerases, stress proteins, beta tubulins, chlorophyll binding proteins, metallothionein-like proteins, CDC2 protein kinases, arabinofuranosidases, Acyl CoA synthetase, sodium symporter, serine-rich proteins, phosphoglyceride transport proteins etc.

Conclusions

The SNP diversity is relatively high in the studied populations. Inbreeding is low, and many populations have excess of heterozygotes, especially in second-generation selection populations. Population differentiation is low in natural stands but higher among second- generation populations and breeding groups, attributable to their relatedness imposed due to the breeding strategy. Population substructure is relatively weak, but there could be up to 6 subpopulations. SNPs contributing to extremely high or low F_{ST} were detected and may exhibit signatures of selection. Numerous associations were detected between SNPs and adaptive phenotypic traits,

but most of them failed the false positive rate test. There are no long-distance LD blocks in the current population, but current SNP density is insufficient for tracing the rate of LD decay. The relatively sparse SNP resolution suggests insufficient power for detecting most associations between current SNP markers and QTLs.

Acknowledgements

The authors acknowledge the support from Texas A&M University Genetics graduate program, the USDA, the NSF, the Texas Forest Service & Western Gulf Forest Tree Improvement Program and the SFTIC early career travel grant to Vikram Chhatre.

References Cited

Eckert AJ, B Pande, ES Ersoz, MH Wright, VK Rashbrook, CM Nicolet and DB Neale. 2009. High-throughput genotyping and mapping of single nucleotide polymorphisms in loblolly pine (*Pinus taeda* L.). *Tree Genet Genomes* 5: 225–234.

Pritchard JK, M Stephens, P Donnelly. 2000. Inference of population structure using multilocus genotypic data. *Genetics* 155(2): 945-959.

Evanno G, S Regnaut, J Goudet. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol Ecol* 14(8): 2611-2620.

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

Rosenberg NA. 2004. *Distruct*: a program for the graphical display of population structure. *Molecular Ecology Notes* 4: 137-138.

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y and Bucker ES. 2007. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* 23(19): 2633-2635.

Scheet P and M Stephens. 2006. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet* 78(4): 629-644.

Barrett JC, B Fry, J Maller and MJ Daly. 2005. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21(2): 263-265.