National Research Council Short Term Mobility grants call 2013

Next generation genome sequencing and adaptive polymorphism analysis of poplar drought sensitive and tolerant genotypes

Beneficiary: Dr. Giovanni Emiliani, Tree and Timber Institute (IVaLSA)

INTRODUCTION

Currently, there is a growing interest in the fast production of biomass as feedstock to displace fossil fuels and to reduce greenhouse gas emissions. Considerable amount of energy are locked-up in lignocellulose, the main component of plant cell walls (Möller et al. 2007). Lignocellulose can be used as raw material for the manufacture of various bio-based products, for example bioethanol (Schubert 2006). Poplar is a good candidate as biorefinery feedstock (together with Salix, Miscanthus, and Triticum; Möller et al. 2007) due to its rapid juvenile growth (Bradshaw et al. 2000), good coppicing ability, and completely sequenced genome (Boerjan 2005; Tuskan et al. 2006). However, the variability in quantitative and qualitative biomass production is huge in the *Populus* genus and among the vast number of interspecific hybrids resulting from crosses among the 30 species belonging to the genus (Cervera et al. 2005).

Genome Wide Association Studies (GWAS) have emerged as a powerful approach for identifying genes underlying complex diseases at an unprecedented rate (International HapMap Consortium, 2007; Altshuler et al., 2008). However, despite their promise, GWAS have largely not been applied to the dissection of complex traits in crop plant (Nordborg, M. & Weigel, 2008; Atwell et al., 2010; Gore et al., 2009). This is due mainly to the lack of effective genotyping techniques for plants and the limited resources for developing high-density haplotype maps like those seen in other well-developed systems, such as the human genome HapMap project (International HapMap Consortium, 2007). Nevertheless the availability of high density genotyping array for poplar (Geraldes et al., 2013) enables genome wide association studies of relevant trait associated with biomass production, with a special emphasis on drought stress response, and biofuel transformation potential.

The WATBIO (Development of improved perennial non-food biomass and bioproduct crops for water stressed environments) project is funded under the Seventh Framework Programme (FP7). Led by the University of Southampton, the WATBIO consortium is investigating the productivity of crops in a future climate. Particular focus is being given to increasingly more episodes of drought and water shortage. Crop productivity, for instance, during the 2003 drought across Europe dropped

by 30 %. The primary aim is to characterise the vast amount of DNA variation in poplar and harness this to produce better crops, through the sequencing of the genome and the analysis of polymorphism of association population trees, sampled from across contrasting sites in Europe, including droughted southern sites. The identification of polymorphism in the sequence of genes might give a clue to survival in stressful environments. These DNA variants can then be used in breeding programmes, enabling to harness the power of molecular biology without the necessity of GM crops.

MATERIAL AND METHODS

Association population and phenotypic data. The *populus nigra* association population was constructed sampling individuals across major European fluvial systems (Figure 1), for a total of 800 individuals. For these preliminary analyses three traits were considered: biomass (expressed as stem volume index) for year 2012, tree height (of the principal stem) and stem number obtained in a common short rotation coppice test field in Northington (UK).

Single Nucleotide Polymorphisms (SNPs) analysis. The 800 poplar genotypes were genotyped with the 34 k poplar SNPs array (Geraldes et al., 2013). After quality control, rare SNPs pruning, etc., 8250 common SNPs were selected for association analyses.

Genome-wide association analysis. Association analyses were conducted using the simple model and the compressed MLM. The genotype data set for were generated after imputation of missing genotypes, with a total of 8250 common SNP sites (minor allele frequency > 0.05).

For the simple model analysis, we used the following equation:

$$y = X\alpha + e$$

For the compressed MLM analysis, we used the equation:

$$y = X + Q\beta + K\mu + e$$

In these equations, y represents phenotype, X represents genotype, Q is the covariance and K is the relative kinship matrix. $X\alpha$ and $P\beta$ represent fixed effects, and $K\mu$ and e represent random effects. The top five principal components were used to build up the P matrix for population-structure correction. The matrix of simple matching coefficients was used to build up the K matrix, and this step was followed by compression. The analyses were performed using GAPIT (Lipka et al., 2012).

Population genetic analyses. Covariate analysis was done using the software Admixture (Alexander et al., 2009) after calculation of the best K number inside the same software (Figure 2). PCA analysis were performed with GAPIT. To minimize the contribution from regions of extensive

strong LD, if a pair of SNPs within the 50-kb region had r^2 greater than 0.8, we removed one of them. The first two principal components were plotted against each other for the *indica* population and the *japonica* population, respectively. LD was calculated using the software TASSEL (Bradbury et al., 2007) with default settings (Figure 3 and 4). Pairwise r^2 was calculated for all SNPs in a 50-kb window and averaged across the whole genome. Sequence diversity (π) was calculated in a 100-kb window as the average number of pairwise difference per site for all pairs of total sampled genotypes. The population-differentiation statistics (*F*ST) were computed as described, using a 100-kb window, between the, among the metapopulation as resulted by the K-means clustering.

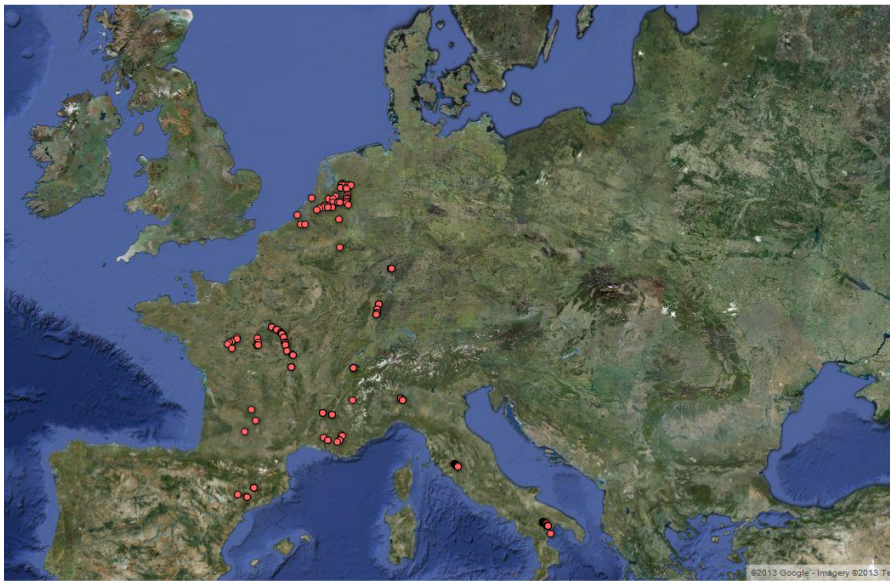

Figure 1: Geographic distribution of sampled individuals for the association population constitution

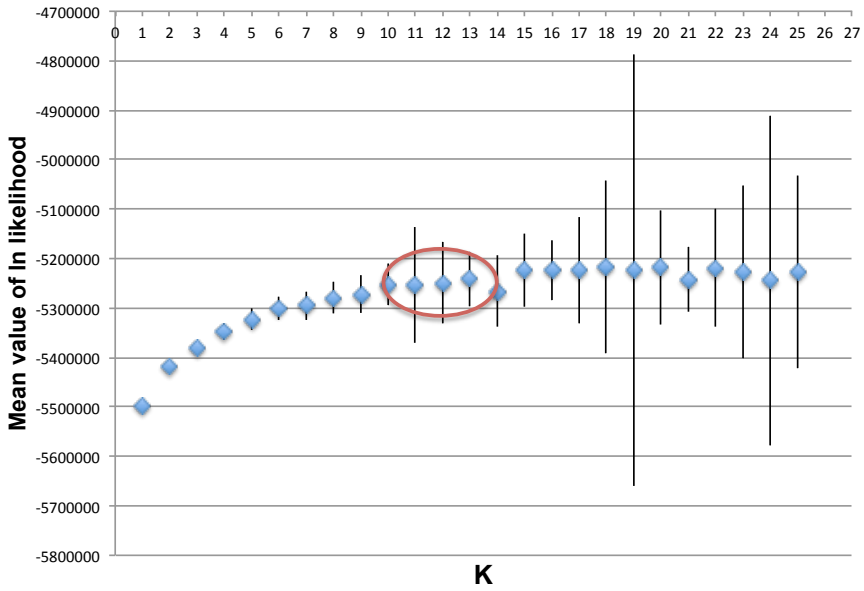

Figure 2: Individuation of the best K value for association analysis.

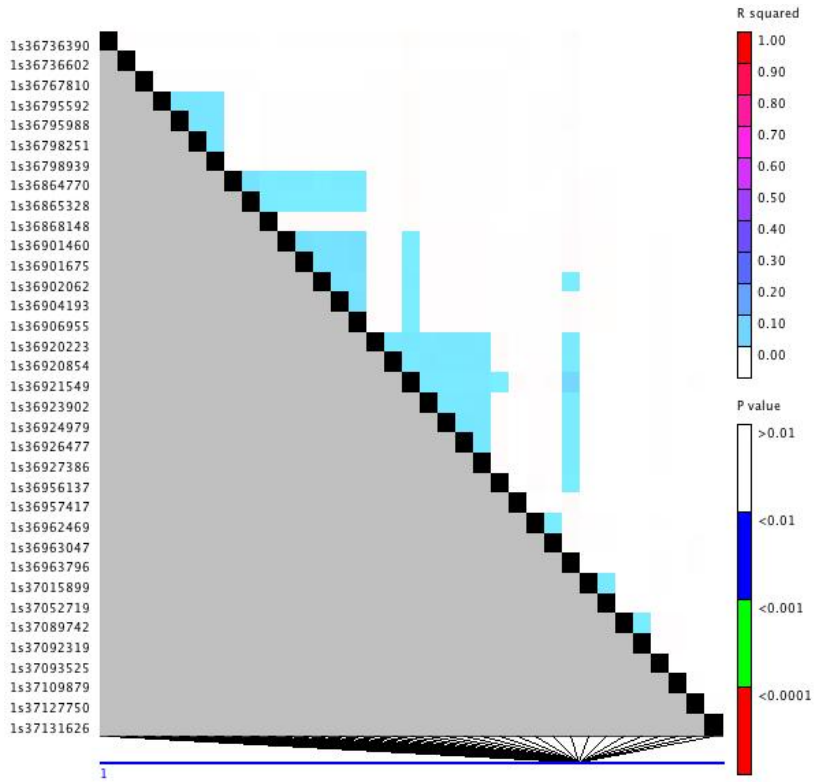

Figure 3: Linkage disequilibrium (r^2) between adjacent SNPs for chromosome 1

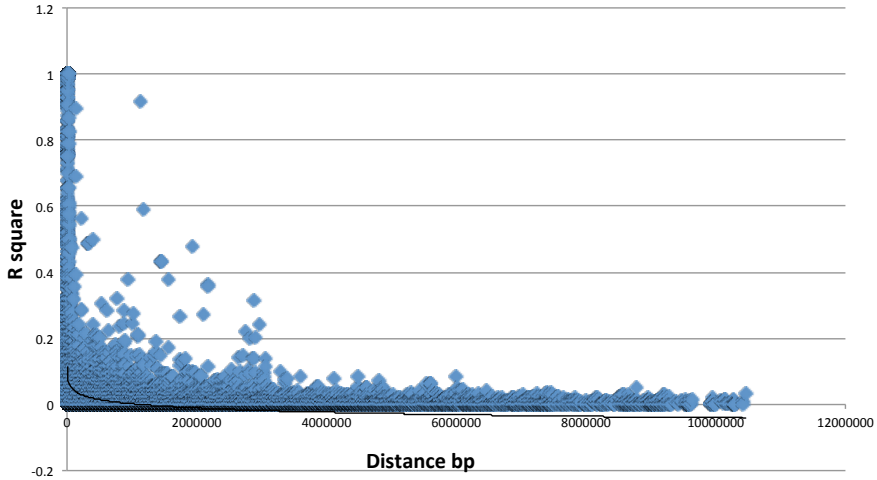

Figure 4: Linkage disequilibrium (r^2) decay for chromosome 1

RESULTS

The analyses of population structure highlighted the presence of population differentiation as shown by the plot of the first 2 principal components calculated with GAPIT (Figure 5). This findings highlights the importance of an adequate analysis of population structure prior to GLM or MLM GWAS. Calculation of the best K (figure 2) resulted in the individuation of 9 clusters to be considered in the subsequent analyses. As expected, Linkage Disequilibrium analyses showed, a low level of association (figure 3) and a rapid decay of chromosome-wide association between the loci used to GWAS analyses.

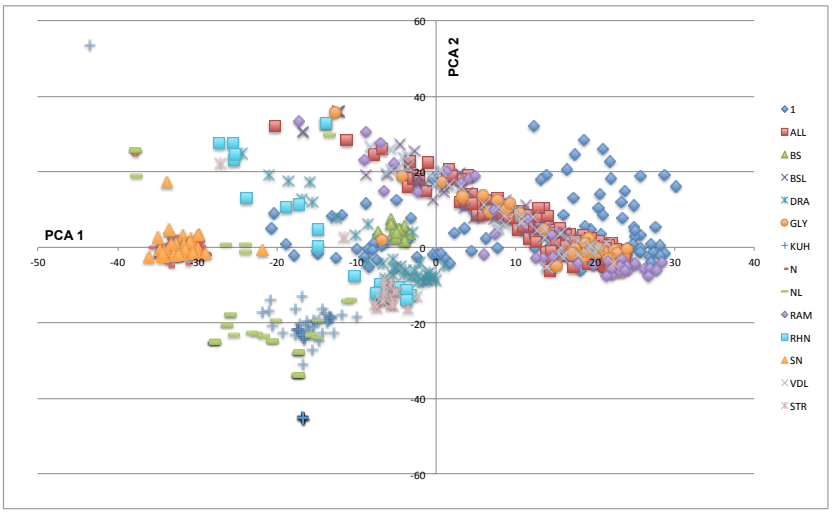

Figure 5: Principal components analyses based on SNPs data of poplar individuals belonging to the association population.

Figure 6 shows the SNPs associated to biomass yield for 2012 plotted on their chromosomal location (chromosome 1-19). Only two SNPs resulted significantly ($-\log_{10}(p) < 4.5$, genome wide significance threshold), associated to biomass, as shown also by the Quartile-Quartile plot of figure 7.

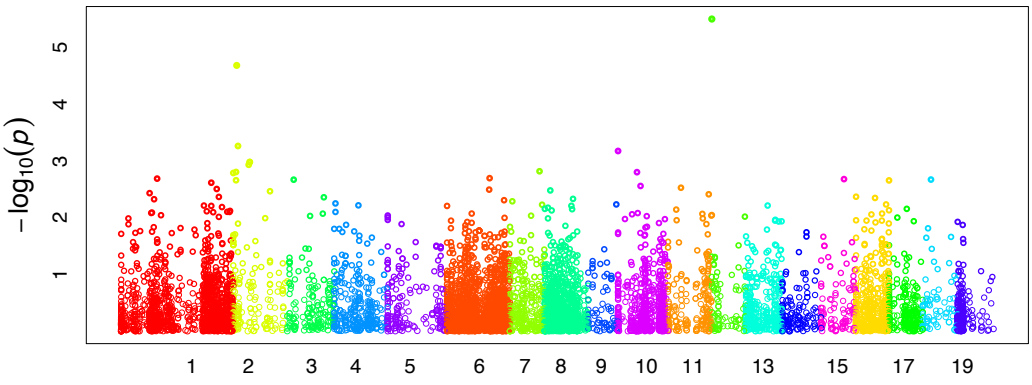

Figure 6: The Manhattan plot showing the SNPs associated to biomass

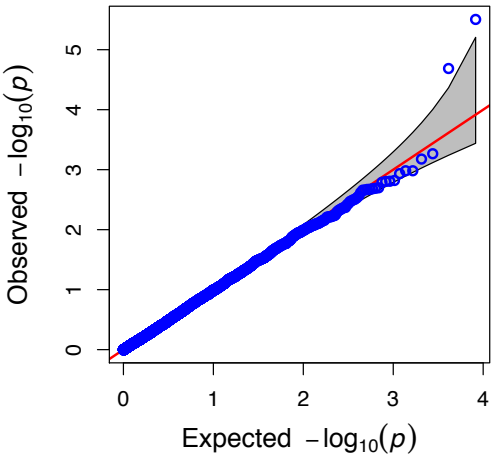

Figure 7: Quartile-Quartile plot showing the expected and observed SNPs associated to biomass

Figure 8 shows the SNPs associated to tree height plotted on their chromosomal location (chromosome 1-19). For this trait no SNPs resulted significantly ($-\log_{10}(p) < 4.5$, genome wide significance threshold), associated to tree height, as shown also by the Quartile-Quartile plot of figure 9.

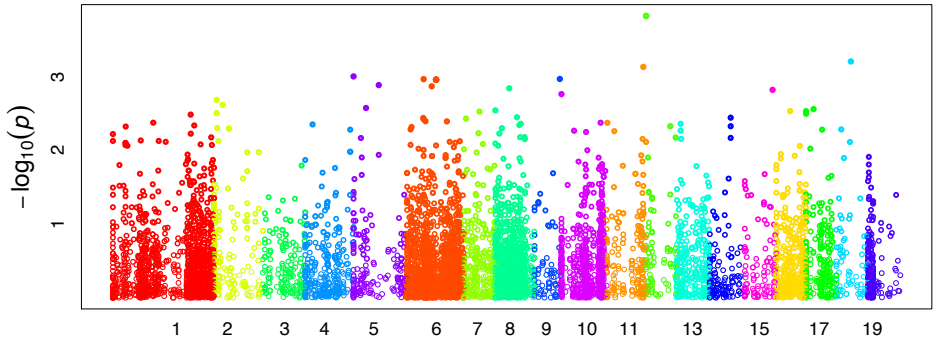

Figure 8: The Manhattan plot showing the SNPs associated to tree height

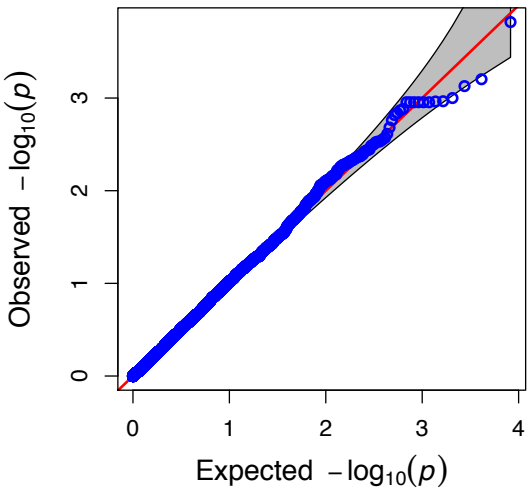

Figure 9: Quartile-Quartile plot showing the expected and observed SNPs associated to tree height

Figure 10 shows the SNPs associated to the average stem plotted on their chromosomal location (chromosome 1-19). For this trait 4 SNPs resulted at the boundary of statistical significance ($-\log_{10}(p) < 4.5$, genome wide significance threshold), associated to stem number, as shown also by the Quartile-Quartile plot of figure 11.

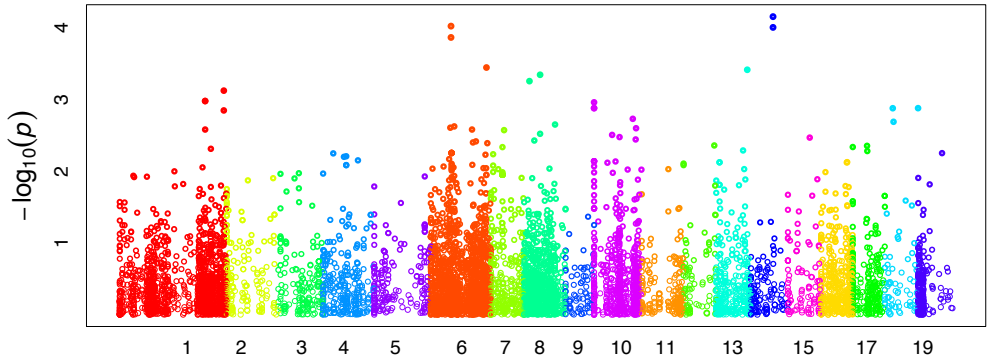

Figure 10: The Manhattan plot showing the SNPs associated to stem number

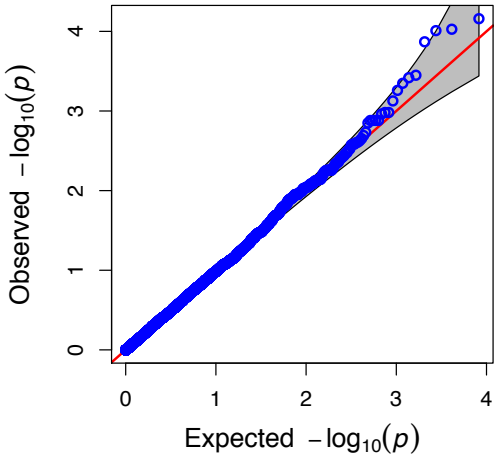

Figure 11: Quartile-Quartile plot showing the expected and observed SNPs associated to stem number

PRELIMINARY CONCLUSION AND PERSPECTIVES

The GWAS approach developed here is able to discover genetic variants among poplar individuals and population associated to complex traits like biomass production, paving the way for genetic improvement. The work is going ahead performing association analyses for trait related to biofuel transformation (saccharification potential) and drought stress response (wood anatomical propeprties).

CITED BIBLIOGRAPHY

Alexander D.H., J. Novembre, and K. Lange. 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Research, 19:1655:1664

Bradbury, P.J. et al. 2007. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23, 2633-2635

Alexander E. Lipka, Feng Tian, Qishan Wang, Jason Peiffer, Meng Li, Peter J Bradbury, Michael Gore, Edward S Buckler and Zhiwu Zhang 2012 GAPIT: Genome Association and Prediction Integrated Tool. Bioinformatics doi: 10.1093/bioinformatics/bts44

The International HapMap Consortium. 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature 449, 851–861

Altshuler, D., Daly, M.J. & Lander, E.S. 2008. Genetic mapping in human disease. Science 322, 881–888.

Nordborg, M. & Weigel, D. 2008. Next-generation genetics in plants. Nature 456, 720–723.

Atwell, S. et al. 2010. Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature 465, 627–631

Gore, M.A. et al. 2009. A first-generation haplotype map of maize. Science 326, 1115–1117

Möller R, Toonen M, van Beilen J, Salentijn E, Clayton D. 2007. Crop platforms for cell wall biorefining: lignocellulose feedstocks. EPOBIO project. CNAP, University of York, UK, pp 162

Schubert C (2006) Can biofuels finally take center stage? NatureBiotechnol 24:777–784

Bradshaw HD Jr, Ceulemans R, Davis J, Stettler R (2000) Emerging model systems in plant biology: poplar (*Populus*) as a model forest tree. J Plant Growth Regul 19:306–313

Boerjan W (2005) Biotechnology and the domestication of forest trees. Curr Opin Biotechnol 16:159–166

Cervera MT, Storme V, Soto A, Ivens B, Van Montagu M, Rajora OP, Intraspecific and interspecific genetic and phylogenetic relationships in the genus Populus based on AFLP markers. Theor Appl Genet 111:1440–1456

Tuskan GA, et al. 2006. The genome of blackcottonwood, Populus trichocarpa (Torr. & Gray). Science313:1596–1604

Boerjan W 2005. Biotechnology and the domestication of forest trees. Curr Opin Biotechnol 16:159–166

Geraldes ASP, Difazio P, Slalov GT, Ranjan P, Muchero W, Hannemann J, Gunter LE, Wymore AM, Grassa CJ, Farzaneh N, Porth I, Mckown AD, Skyba O, Li E, Fujita, J. Klapste J, J. Martin J, Schackwitz W, Pennacchio C, Rokhsar D, Friedmann MC, Wasteneys GO, Guy RD, Y. A. EL-Kassaby YA, Mansfield Q, Cronk CB, Ehlting J, Douglas CJ, Tuskan GA. 2013. A 34K SNP genotyping array for Populus trichocarpa: Design, application to the study of natural populations and transferability to other *Populus* species. Molecular Ecology Resources

Firenze 1 Luglio 2013 Dr Giovanni Emiliani