Dissection of complex traits in forest trees – opportunities for marker-assisted selection

Saravanan Thavamanikumar\textsuperscript{1,2}, Simon G. Southerton\textsuperscript{3}, Gerd Bossinger\textsuperscript{1,2}, and Bala R. Thumma\textsuperscript{3,*}

\textsuperscript{1}Department of Forest and Ecosystem Science, The University of Melbourne, Creswick, Victoria 3363, Australia

\textsuperscript{2}Co-operative Research Centre for Forestry, Private Bag 12, Hobart, Tasmania 7001 Australia

\textsuperscript{3}CSIRO Plant Industry, GPO Box 1600, Acton, ACT 2601, Australia

\textsuperscript{*}Corresponding author: Bala R. Thumma, CSIRO Plant Industry, GPO Box 1600, Acton, ACT 2601, Australia

Email: Reddy.Thumma@csiro.au

Phone: +61-02-62464856
ABSTRACT

Due to their long reproductive cycles and the time to expression of mature traits, marker-assisted selection is particularly attractive for tree breeding. In this review we discuss different approaches used for developing markers and propose a method for application of markers in low linkage disequilibrium (LD) populations. Identification of useful markers for application in tree breeding is mainly based on two approaches, quantitative trait locus mapping (QTL) and association genetic studies. While several studies have identified significant markers, effect of the individual markers is low making it difficult to utilize them in breeding programs. Recently genomic selection (GS) was proposed for overcoming some of these difficulties. In GS, high density markers are used for predicting phenotypes from genotypes. Currently small effective populations with high LD are being tested for GS in tree breeding. For wider application, GS needs to be applied in low LD populations which are found in many tree breeding programs. Here we propose an approach in which the significant markers from association studies may be used for developing prediction models in low LD populations using the same methods as in GS. Preliminary analyses indicate that a modest numbers of markers may be sufficient for developing prediction models in low LD populations. Genomic selection based on large numbers of random markers or small numbers of associated markers is poised to make marker-assisted selection a reality in forest tree breeding.

Key words: Genomic selection, genomic estimated breeding values, association mapping, QTL mapping, linkage disequilibrium.
Introduction

Traits that do not follow classical Mendelian monogenic inheritance are termed “complex” or “quantitative” traits (Lander and Schork 1994). Such traits are usually governed by variation in multiple genes and are typically influenced by the environment. One of the major challenges in the field of genetics is to account for all inherited phenotypic variation at the molecular level. In addressing this challenge it is often difficult to identify molecular markers that exhibit perfect co-segregation with complex traits because of their polygenic inheritance and other factors including epistasis, incomplete penetrance, phenocopy and locus heterogeneity (Lander and Schork 1994). However, it is essential to dissect complex traits into their genetic components in order to facilitate breeding through marker-assisted selection (MAS). In tree species, where most phenotypic traits are complex, the use of MAS for tree improvement is particularly attractive because conventional selection is impeded by long generation times and the long delay in the full expression of mature traits (Grattapaglia et al. 1996).

Attempts at identifying suitable markers for breeding have until recently focused on the mapping of linkage-based quantitative trait loci (QTL). Using this approach significant progress has been made in identifying genomic regions that underlie complex traits in numerous species (Goddard and Hayes 2009; Grattapaglia et al. 2009; Mauricio 2001). However, due to several limitations of the QTL approach (see below), the focus of marker discovery research in trees has increasingly shifted to association genetic mapping. In this paper, we (1) review the methods for dissection of complex traits such as QTL and association studies, (2) discuss the advances made using these studies and limitations of these approaches in applied breeding programs, (3) discuss the current method of genomic selection to overcome some of these limitations and (4) discuss how genomic selection models can be applied in low LD populations using markers identified from association studies.

Quantitative trait locus (QTL) mapping

QTL mapping is generally performed in pedigrees from controlled crosses and involves associating variation in regions of a genome with variation in measured phenotypic traits. Generally, only two alleles segregate at a locus in backcross and $F_2$ pedigrees constructed in
inbred crop species. This is because segregation in the mapping generation results from meiosis in a single parent. However, in outcrossing and highly heterozygous forest trees, up to four alleles can segregate at a single locus (Sewell and Neale 2000). This is because segregation results from separate meioses and crossovers in the two parents (Groover et al. 1994). Where a QTL segregates for more than two alleles, there is an increased likelihood that allelic interactions such as simple dominance, over-dominance or under-dominance will occur (Groover et al. 1994).

QTL mapping in forest trees

Until recently, QTL mapping was the principal means of dissecting the genetics of complex traits in forest tree species. QTL have been identified for growth and wood property traits (Brown et al. 2003; Devey et al. 2004; Markussen et al. 2003; Thumma et al. 2010b), vegetative propagation (Marques et al. 2002; Scotti-Saintagne et al. 2005; Thumma et al. 2010a), and disease resistance (Newcombe et al. 1996) in a range of forest trees. While several QTL have been identified, individual QTL however typically explain only a small proportion of phenotypic variation. For example, using an outcrossed F$_2$ family of _E. globulus_, a recent study identified a number of QTL influencing growth (diameter at breast height) and wood properties (wood density, cellulose content, pulp yield and lignin content) (Freeman et al. 2009). One to four QTL were identified for each of these traits, which jointly explained between 3.8% and 34.9% of the phenotypic variation in the same population. Because of the small number of individuals used most of these marker effects may have been over estimated (Beavis 1998).

Though QTL mapping helps determine the complexity of the genetic architecture underlying phenotypic traits, the mode of action of specific genes within QTL remains unclear as the genomic regions identified are usually large (Doerge 2002). Most QTL span hundreds, and occasionally thousands, of genes (Nadeau and Frankel 2000). For example, 95% confidence intervals for three separate QTL observed for wood basic density in a recent QTL study in _E. globulus_ (Freeman et al. 2009) covered 39, 40 and 28 centimorgans (cM) respectively. Similarly, in _E. nitens_, five separate QTL identified for wood basic density had 95% confidence intervals ranging from 38 to 50 cM (Thumma et al. 2010b). In eucalypts, a single cM, which represents the distance across which one recombination will occur in 1% of trees in a single generation, was estimated to be approximately 200 kb (Grattapaglia and Bradshaw
Assuming 5 to 10 expressed sequence tags (ESTs) per 200 kb\(^1\), this equates to between 135 and 400 ESTs per QTL. QTL mapping resolution cannot be increased by increasing the number of markers alone as the major limitation to map resolution is the low number of recombination events observed. Although the resolution could be increased by increasing the number of progeny along with more markers, the increased resolution would still likely to be insufficient for identification of candidate genes.

While the marker effects observed within species are small to moderate, marker effects observed in crosses between species are however larger. Shepherd et al (2006, 2008) identified large effect QTL explaining as much as 66% of variation in rooting ability by analysing the genetic architecture of adventitious root system in \textit{Pinus} and \textit{Corymbia}. In \textit{Pinus}, by comparing QTL detected in F\(_1\) and F\(_2\) populations they observed that the size of the QTL effects observed within species was lower compared to that of between species (Shepherd et al. 2006). Similarly using an inter-specific cross in \textit{Populus}, (Bradshaw and Stettler 1995) observed growth QTL explaining as much as 45% of genetic variation in stem volume. Some of the large effect QTL observed in these inter-specific crosses may be real. Transgressive segregation may explain the large effect QTL observed in the inter-specific crosses. Complimentary gene action of the recombined alleles in hybrids that were fixed with opposite sign in parents results in transgressive segregation (Stelkens and Seehausen 2009).

Because QTL studies are conducted in pedigrees from a small number of parents, only a fraction of the potential variation underlying a quantitative trait can be investigated. As the interval between the marker and the QTL is generally large, recombination between the marker and the QTL may make it difficult to observe the QTL in a different pedigree (Verhaegen et al. 1997). Since a QTL detected in one pedigree may not segregate in another pedigree, validating QTL results in different pedigrees (Lander and Schork 1994; Rocha et al. 2007) is challenging. For example, in a QTL study in \textit{E. globulus} conducted across multiple pedigrees and sites, only 19 of the 98 QTL, which affected wood properties and growth, were significant in more than one family (Freeman et al. 2011). This can limit the ability of QTL

\(^1\)estimated using the \textit{Eucalyptus} Genome Database browser (http://eucalyptusdb.bi.up.ac.za/)
studies to guide broader inferences on the genetic architecture of any given trait. In contrast to QTL mapping, association studies can be used to explore population wide variation.

**Association genetic mapping**

Detection of significant associations between allelic variation and phenotypic variation at a population level is variously termed as association genetics, association mapping or linkage disequilibrium (LD) mapping. Association studies are generally carried out in populations containing several hundreds of unrelated individuals. The resolution of marker-trait associations identified is much higher than in QTL studies as association studies exploit the low LD generally found in natural populations. However to exploit the advantage of high resolution due to low LD, high density markers are needed. Accumulation of numerous recombination events over many generations in natural populations breaks all long-range linkages. This results in short stretches of high LD between the loci and permits fine-scale mapping of marker-trait associations (Nordborg and Tavare 2002). Moreover, association genetics exploits trait variation existing in the whole population while QTL studies are restricted to variation present between the parents of the pedigree used in the study (Neale and Savolainen 2004). Therefore, as markers are identified in analyses of diverse populations, these markers may be used in most early generation tree breeding populations which are based on a large number of families.

Understanding of the nucleotide diversity (\(\pi\)) patterns in a population is important for association studies. Several studies have examined nucleotide diversity in tree species (Table 1). Nucleotide diversity is measured as average number of base pair (nucleotide) differences per base pair between orthologous (allelic) sequences of two trees drawn at random from a population. In forest trees the estimates of nucleotide diversities have ranged from 0.00156 (Cryptomeria japonica; Kado et al. 2008) to 0.01110 (Populus tremula; Ingvarsson 2005). While most of the studies have analysed only a small number of genes it appears that the available nucleotide diversity is large enough to be useful for association studies. Compared to studies in humans, association studies are new to forest trees, however, early results are promising and a number of loci associated with a range of traits have been identified (Table 2).
Linkage disequilibrium and its impact on the resolution of association mapping

In addition to nucleotide diversity, linkage disequilibrium (LD) has important implications in association mapping. Linkage disequilibrium is the non-random association of alleles at two or more loci, not necessarily on the same chromosome, in which some combinations of alleles occur more or less frequently in a population than would be expected if alleles were inherited independently (Neale et al. 2002; Nordborg and Tavare 2002). A number of evolutionary phenomena including selection, genetic drift, gene flow, mutation, and factors including an active mating system, population size and structure generate LD (Rafalski and Morgante 2004; Terwilliger and Weiss 1998). Recombination between alleles leads to breakdown in LD (Terwilliger and Weiss 1998).

In general, LD decays more quickly in outcrossing species than in inbreeding species. For instance, LD decays rapidly in outcrossing species like maize (decays in just 100 bp; Tenaillon et al. 2001), potato (decays within 1 kb; Simko et al. 2006), loblolly pine (decays within 800 bp; Gonzalez-Martinez et al. 2006), Scots pine (decays within 750 bp; Garcia-Gil et al. 2003), Douglas fir (decays within 2 kb; Krutovsky and Neale 2005), white spruce (decays within 65 bp; Pavy et al. 2012), Shining gum (decays within 850 bp; Thumma et al. 2005), and European aspen (decays within 500 bp; Ingvarsson 2005). In contrast, LD extends over longer distances in selfing species like soybean (extends over 50 kb; Zhu et al. 2003), Arabidopsis (extends to 250 kb; Hagenblad and Nordborg 2002) and rice (extends to 100 kb; Garris et al. 2003). LD patterns may vary between different populations of a species that show significant structure (Goddard et al. 2000; Kidd et al. 2004) and between different parts of the genome (Daly et al. 2001; Dawson et al. 2002). In structured populations, LD should be computed within populations as pooling population samples with different allele frequencies will produce LD between unlinked loci (Mangin et al. 2011). Although LD patterns in different populations were studied (Heuertz et al. 2006; Ingvarsson 2005; Pyhajarvi et al. 2007), LD in different regions of the genome is not yet characterized well in forest trees.

The extent of LD has two main implications for association studies. Rapid decay of LD enables high resolution mapping of associations between the marker and the trait (Grattapaglia and Kirst 2008; Thumma et al. 2005) and can even lead to the identification of causative polymorphisms (Gaut and Long 2003; Thumma et al. 2009). However, as high
High densities of markers are needed to attain an appropriate level of genomic coverage, low LD makes genome-wide association studies (GWAS), such as those carried out in humans (Petukhova et al. 2010) and Arabidopsis (Todesco et al. 2010), unfeasible in many forest tree species. Consequently, candidate gene-based association studies are currently the preferred approach in forest trees (Gonzalez-Martinez et al. 2007; Thumma et al. 2005). In candidate gene-based association studies markers from candidate genes controlling the relevant traits are tested for association while in GWAS random markers spread across the genome are tested. However, as the cost of next generation sequencing continues to fall, GWAS based on sequencing individual genomes may soon become possible at least for species with reference sequences such as Eucalyptus and Populus.

**Experimental design of association studies**

Two main designs based on whole populations and individual families are used for conducting association studies in forest trees. In the population-based approach, unrelated individuals or a few individuals from many different families are used while in the family-based approach multiple individuals from a smaller number of different families are used. However, the main drawback of population-based approaches is that they can suffer from demographic problems, such as population structure, which can lead to false positives (Lander and Schork 1994). These spurious associations may be avoided by performing association studies in homogenous populations rather than in structured populations (Lander and Schork 1994; Pritchard et al. 2000b). Alternatively, statistical models can be used to infer population structure (Pritchard et al. 2000a) or to account for multiple levels of relatedness (Yu et al. 2006). Accounting for population structure may sometimes result in false negatives when the polymorphism associated with the trait coincides with the population structure (Yu and Buckler 2006; Zhao et al. 2007).

Family-based designs, on the other hand, are robust against false positives associated with population structure. Family-based approaches such as quantitative transmission disequilibrium tests (QTDT) exploit information from linkage and linkage disequilibrium to find associations (Benyamin et al. 2009). Using this approach Gonzalez-Martinez et al. (2008) identified two SNPs from a dehydrin and a cell wall gene that were significantly associated with water use efficiency in Pinus taeda. While family-based association studies are robust against population structure, in general they have low power and resolution.
compared to population-based association studies. Another drawback of this method is the requirement for heterozygous parents in the families.

**Selection of candidate genes and polymorphisms for association studies**

Candidate genes for association studies are selected from expressed sequence tag (EST) libraries derived from tissues likely to influence the trait in question. For wood quality studies in eucalypts, a number of EST libraries have been generated from developing xylem tissue (Paux et al. 2004; Rengel et al. 2009). A large number of candidate genes have been identified based on their ‘differential expression’ in a range of forest tree species including eucalypts (Foucart et al. 2006; Paux et al. 2004; Qiu et al. 2008), loblolly pine (Allona et al. 1998; Whetten et al. 2001) and spruce (Friedmann et al. 2007). With the steady decrease in the cost of next generation sequencing, even larger numbers of candidate genes are expected to become available in the near future, even for species with currently limited genomic resources (Neale and Kremer 2011). Recent developments in genomics technology are also making it possible to identify functional SNPs controlling traits. For example, using next generation methods for transcriptome sequencing it is possible to detect not only candidate genes but also functional markers controlling the expression of these candidate genes. RNA sequencing of *Eucalyptus camaldulensis* seedlings subjected to water stress was recently used to identify several candidate genes along with several potential functional SNPs based on differential allelic expression (Thumma et al. 2012).

**SNP genotyping and association analysis**

Common SNPs are identified by sequencing candidate genes in several individuals (Thumma et al. 2009; Gonzalez-Martínez et al. 2006). To capture species-wide diversity, trees representing all the major sub-populations or races of a species are typically included (Thavamanikumar et al. 2011). Subsequently, common and haplotype tagging SNPs (Johnson et al. 2001) are selected for genotyping in the larger population. The frequency and position of polymorphisms, the presence of haplotypes and LD between the SNPs are some of the factors influencing the selection of SNPs for genotyping (Tabor et al. 2002). SNPs from coding regions, especially nonsynonymous substitutions (substitutions that will change the amino acid sequence), are generally targeted as these substitutions are more likely to impact phenotype than substitutions in coding regions that do not alter the amino acid sequence.
(Terwilliger and Weiss 1998). However, there is increasing evidence that other silent site mutations are similarly likely to confer phenotypic effects by a variety of mechanisms. For example, two intronic mutations in the *cinnamoyl CoA reductase* gene (*CCR*) were reported to associate with microfibril angle (Thumma et al. 2005) and a synonymous SNP in *EniCOBL4A* gene was reported to associate with cellulose content by affecting allelic expression in *E. nitens* (Thumma et al. 2009). Similarly, an intronic SNP in the *α-tubulin* gene of *Pinus taeda* was strongly associated with microfibril angle (Gonzalez-Martinez et al. 2007). In white spruce, one intronic SNP each from β-expansin and β-TIP (β-Tonoplast Intrinsic Protein) was associated with percentage of early wood and tracheid diameter in the latewood, respectively (Beaulieu et al. 2011). In *Drosophila melanogaster*, a number of intronic mutations from several loci have been found to associate with bristle number (Long et al. 1998; Lyman et al. 1999). Likewise, a number of promoter variants were found to associate with a variety of diseases in human association studies (Clement et al. 2009; Goyenechea et al. 2009; Lan et al. 2009; Marquet et al. 2008). These findings suggest that important markers may be missed if only coding non-synonymous SNPs are selected for genotyping.

**Validation of marker-trait associations**

False positives in association mapping studies can result from population stratification when sub-populations differ in their SNP allele frequency and mean trait values (Koller et al. 2004) or due to multiple testing errors. A range of methods are now available to account for population structure and to correct for multiple testing errors. Several statistical methods have been proposed to account for this structure when testing for association (Pritchard et al. 2000b; Yu et al. 2006). Validation of marker-trait associations in one or more independent populations is another way of identifying robust markers and reducing false positives. Associations have been successfully validated in independent populations in some studies (Dillon et al. 2012; Dillon et al. 2010; Thumma et al. 2009; Thumma et al. 2005). Functional studies may also help in validating associations. For example, in *E. nitens*, Thumma et al. (2009) have identified a synonymous SNP in the *EniCOBL4A* gene associated with cellulose content. Using allelic expression imbalance studies and nuclear protein binding studies, these authors showed that this synonymous SNP is a *cis*-acting regulatory polymorphism affecting allelic expression. Recently, Beaulieu et al. (2011) have identified 13 SNPs associated with
wood traits in white spruce. Expression of three genes was significantly correlated with SNP genotypes indicating the functional significance of the associated SNPs.

Low frequency polymorphisms have low power in detecting associations as large populations are required to test the effect of rare variants (Long and Langley 1999). For this reason, and because populations of a few hundred individuals are generally used in association studies in forest trees, common SNPs (> 0.10 MAF, minor allele frequency) are generally used. Recent genome-wide association studies in humans and a number of candidate-gene based association studies in forest trees have shown that the effect of common SNPs is typically modest (Table 2). In view of these results it has been suggested that rare SNPs with large effect on the traits may exist which are not detected by current methods (Goldstein 2009). In order to identify such rare SNPs large populations with thousands of individuals are required (Grattapaglia and Kirst 2008); inevitably there is a trade-off between the power to detect an association and the size of the effect a polymorphism has on phenotype.

While several studies have identified markers linked to QTL and SNP markers within the candidate genes affecting a trait, these markers are not currently being used in breeding programs. In addition to the difficulties of using markers from QTL studies as described earlier, the effect of markers identified from both QTL and association studies are small. Often individual markers explain less than 5% of total variation. Two SNPs that were independently associated with microfibril angle (MFA) in *E. nitens* explained between 3.4 and 5.9% of total variation in MFA (Thumma et al. 2005). In *Populus tremula*, two nonsynonymous SNPs were independently associated with bud set (Ingvarsson et al. 2008). After correcting for the possible upward bias in the effect size, these two SNPs explained 1.4 and 5.9% of the variation in bud set. In a large multilocus association mapping study involving 117 genes in *Pseudotsuga menziesii*, all SNPs independently explained less than 5% of the phenotypic variation in traits (Eckert et al. 2009a). In white spruce, 13 SNPs associated with wood quality explained between 3 to 5% of trait variation individually (Beaulieu et al. 2011). A large proportion of the trait variation may be captured by combining the effects of several markers (Yang et al. 2010). The recently developed genomic selection (GS) approach exploits the effects of several markers for predicting phenotypes for selection in breeding programs.
Genomic selection – a new approach for marker-assisted selection in tree breeding

Selecting individuals based on genomic estimated breeding values (GEBVs) is termed genomic selection (Hayes et al. 2009). Genomic selection was pioneered by Meuwissen et al. (2001) and has been quickly adopted by animal and plant breeders (Hayes et al. 2009; Heffner et al. 2009; Lorenzana and Bernardo 2009; Luan et al. 2009). Recently this approach has been applied to tree breeding (Resende et al. 2012a; Resende et al. 2012b; Resende et al. 2012c; Grattapaglia and Resende 2011). In GS, large numbers of progeny from a highly structured population with high LD are genotyped with thousands of markers. The effects of all the markers are used in developing prediction models using ‘training’ or ‘discovery’ populations. These models are then used to predict genomic breeding values of the individuals of the validation population using only the genotype data. The accuracies of the GEBVs are obtained by correlating GEBVs against traditional pedigree-based breeding values. Accuracies of GS models are also obtained by testing the ability of markers to predict phenotypes in one generation by using the model developed in an another generation (Habier et al. 2007). Simulation analyses have indicated that accuracies as high as 0.85 are possible with markers alone (Meuwissen et al. 2001). With high accuracies of GEBVs it is possible to double the genetic gain by shortening the breeding cycle (Hayes et al. 2009). It has been estimated that the application of GS in dairy cattle breeding could lead to cost savings of as much as 92% (Schaeffer 2006).

In GS, as currently applied, a high density of markers is necessary to ensure that at least some of the markers are in LD with the genomic regions affecting the traits. As the number of predictors are generally higher than the number of samples (p >> n), marker effects cannot be estimated using multiple regression by ordinary least squares (Jannink et al. 2010). To overcome multicollinearity and the reduced predictive ability associated with using large number of predictors compared to the samples, statistical models treating markers as random effects are used to develop the prediction models (Jannink et al. 2010). Ridge regression best linear unbiased prediction (RR-BLUP) and Bayesian regression (Meuwissen et al. 2001) are some of the statistical models used for developing prediction models. All of these methods are aimed at reducing the marker variance by shrinking the marker effects toward zero. RR-
BLUP assumes equal marker variances while Bayesian models allow each marker to have its own variance (Meuwissen et al. 2001).

Traditional best linear unbiased predictions (BLUP) using mixed model equations are now widely used in animal and tree breeding for predicting breeding values (White et al. 2007). In traditional BLUP, phenotypic information from individuals as well as their relatives is used to predict breeding values. Information from relatives in current and past generations is incorporated via the additive relationship matrix, which contains information on the proportion of the genome identical by descent (IBD) between pairs of individuals calculated from their known pedigree relationships (Lynch and Walsh 1998). It has been shown that the accuracy of predictions can be higher when the expected relationship matrix from the pedigree is replaced by the realized relationship matrix generated from markers (Goddard 2009; Hayes et al. 2009). As the markers are capable of capturing the random Mendelian sampling variations i.e., variations between the siblings, the accuracy of the predictions would be higher than using the pedigree-based expected relationship matrix (Daetwyler et al. 2007). The prediction of GEBVs using the realized relationship matrix is termed genomic best linear unbiased prediction (GBLUP). GBLUP assumes that a large number of genes with equal effects control the traits similar to RR-BLUP (Zhang et al. 2010). GBLUP is equivalent to RR-BLUP and the accuracies of GEBVs from the two methods are similar (Hayes et al. 2009). Accuracies of GEBVs from GBLUP or RR-BLUP are higher if a trait is controlled by a large number of QTL while the accuracies from Bayesian methods are higher if a trait is influenced by a small number of QTL (Daetwyler et al. 2010).

When the density of markers is high, most of the genetic effects will be captured by the markers. However, when the marker density is low inclusion of a polygenic term in the model may help in capturing genetic variance not captured by the markers (Haley and Visscher 1998). Simulation studies have shown that when marker density is high the accuracy of GEBVs did not improve with the inclusion of polygenic effect in the model. At lower marker densities however the accuracy of the GEBVs improved with the inclusion of the polygenic effect. However, the inclusion of polygenic effect reduced the bias between the estimated variance components and the simulated values even for high density markers suggesting increased bias in the estimated values without the polygenic effect at higher marker densities (Calus and Veerkamp 2007; Solberg et al. 2009).
Simulation studies have also shown that inclusion of polygenic effect will improve the persistence of marker effects over several generations (Solberg et al. 2009). At higher densities markers capture family relationships leading to spurious associations, i.e., association due to family relationships rather than markers linked to the causative QTL. Such spurious associations due to family relationships decay faster over time compared to associations due to markers in LD with the QTL. By including the polygenic effect in the model, associations due family relationships are captured by polygenic term leaving the QTL variation to be captured by the markers. This will lead to persistence of the accuracies as the predictions are based on the markers in LD with the QTL which decay more slowly over time (Solberg et al. 2009). In association studies similar approaches (marker based kinship estimates) are used to control for relatedness among the individuals (Yu et al. 2006).

GS models utilize both LD between markers and the QTL as well as the family relationships in estimating GEBVs. RR-BLUP mainly exploits the family relationships while Bayesian methods use both family relationships as well as LD between the markers and the QTL in estimating the GEBVs (Habier et al. 2007). As the contribution of genetic relationships towards the accuracy of GEBVs is halved in each descendant generation following the generation in which the model was developed, the ranking of the individuals may change in different generations for the methods such as RR-BLUP which mainly exploit the genetic relationships in estimating GEBVs. Bayesian methods on the other hand exploit LD between markers and the QTL better than RR-BLUP and therefore the accuracies of Bayesian methods will persist over several generations (Habier et al. 2007).

**Genomic selection in forest trees**

Simulation studies in forest trees have indicated that in populations with small effective size and high LD, GS can substantially improve the efficiency of conventional tree breeding (Grattapaglia and Resende 2011; Iwata et al. 2011). Only a few markers per genetic distance (two to 20 markers/cM depending on the effective population size) are needed to reach the accuracies of selection based on traditional BLUP estimates. Simulation studies have shown that selection efficiencies (gain with markers in comparison to phenotypic selection) of more than 100% could be obtained with GS (Grattapaglia and Resende 2011). In a proof of concept study, Resende et al. (2012c) tested the utility of GS for predicting growth traits in *Pinus taeda*. By using genotype information from 4825 markers in 800 individuals from a
structured population they developed prediction models for height and DBH. Accuracies of GS for these traits ranged from 0.64 to 0.74. Compared to phenotypic selection, a selection efficiency of between 1.53 and 2.00 was obtained with GS. Similarly, Resende et al. (2012a) have observed accuracies of between 0.55 and 0.88 for pulp yield and growth traits in two Eucalyptus breeding populations using more than 3000 markers. While the accuracies of the models were found to be high, the accuracy for predicting across different populations and sites were low in the two studies suggesting significant genotype by environment (G x E) interactions. Significant G x E interactions between populations may arise due to differences in allele frequencies and LD patterns between the populations. Results from these studies however clearly show that marker-assisted selection with GS has the potential to significantly improve the efficiency of traditional tree breeding. In practical tree breeding programs, GS effectively replaces the prolonged testing phase and thus drastically reducing the breeding cycles (Figure 1). In Eucalyptus breeding, GS would result in 50% reduction in breeding cycle resulting in economic return of 20 times on the investment (Resende et al. 2012a).

The main limitation of applying GS in forest tree breeding is the need to develop breeding populations with high LD using small effective populations. Most of the early generation tree breeding populations generally contain large numbers of open pollinated families (Eldridge et al. 1994). Linkage disequilibrium will typically be low in such populations. Other difficulties associated with GS in high LD populations are the potential breakdown in the accuracy of prediction models in advanced generations and the inability of the models to predict traits across sites and/or populations due to genotype X environment interactions and differences in LD between the populations (Hayes et al. 2009). As the populations used for GS are derived from a few parents, loss of genetic diversity and increase in the rates of inbreeding are of concern. However, as markers can capture random Mendelian segregation variations, application of GS within a given population results in reduced inbreeding compared to traditional BLUP-based selections (Hayes et al. 2009; Heffner et al. 2009; Jannink et al. 2010). Using traditional BLUP two siblings from a full-sib family may receive similar breeding values if their phenotypes are similar. Therefore both of them will be selected using traditional BLUP methods leading to increased selection of individuals with similar genetic makeup. However, with markers, the siblings can be separated by capturing random Mendelian variations among the siblings leading to lower inbreeding (Hayes et al. 2009).
Some of the limitations of using GS in high LD populations may be overcome by using markers identified from association studies. Once markers from several candidate genes explaining a significant proportion of a trait are identified, allele effects of all the markers can be used to generate marker-based breeding values as in GS. The pre-selected markers may be used in random effects models as in GS or in fixed effects models (fixed regression-least squares) to estimate the marker effects and to predict GEBVs. As the markers identified from association studies are often in high LD with traits the predictions should be valid over several generations. However, the identification of flip-flop effects (Lin et al. 2007; same marker showing significant association in different populations at different sites but with allelic effects swapped) in recent association studies (Southerton et al. 2010; Dillon et al. 2012) indicate that results of GS using markers from association studies may also be population specific unless stable markers with allelic effects in the same direction from across sites are used.

Lack of large number of markers is the main impediment for developing prediction models in low LD populations. Recent developments in sequencing and genotyping technologies should accelerate identification of markers associated with the traits. Transcriptome sequencing can be used to identify candidate genes and SNPs at genome wide scale (Thumma et al. 2012). Potential genes and markers showing differential expression can be used in association studies to test the effect of the SNPs in different populations.

In a preliminary study to test the efficacy of using markers from association studies for application of GS in low LD populations, we used a modest number of markers (90 SNPs) associated with several wood quality traits at a marginal significance level in *E. nitens* (Thumma et al. unpublished data) to develop prediction models. Prediction models were developed using RR-BLUP, Bayesian LASSO and fixed regression models. Using these prediction models we were able to predict several wood traits with accuracies ranging from 0.20 to 0.47 in three different populations. These predictive abilities are similar to those reported for several wood quality traits in a highly structured and high LD population of loblolly pine (Resende et al. 2012c) and the predictive ability of pulp yield (0.38) in *Eucalyptus* (Resende et al. 2012a). Results from this preliminary study indicate that GS models can be applied in low LD populations using a modest number of markers. Increasing the number of markers should potentially increase the accuracies of predictions.
Conclusions

QTL mapping has been used extensively in forest trees to identify genomic regions that contribute to variation in complex traits. However, the low resolution of QTL mapping limits its application to tree breeding. Association mapping has recently been used successfully to identify several SNP markers in forest tree species which could be used in breeding programs. Due to the moderate to high nucleotide diversity and low LD in most forest tree species candidate gene-based association studies are most likely to yield success. However, the low proportion of phenotypic variation explained by individual SNPs is consistent with earlier results from QTL studies. The small effect of individual QTL/SNPs makes it difficult to apply these markers in breeding programs. Genomic selection has been proposed to overcome the limitations of using markers in breeding programs. By using large numbers of markers in prediction models GS can significantly improve the selection efficiency in breeding programs. So far the application of GS has been restricted to highly structured populations with high LD. For wider application in forest tree breeding programs, GS needs to be tested in low LD populations. For this, markers previously identified from association studies can be used to develop prediction models. As the markers from association studies are in general in high LD with the causal variants, prediction models from such markers should persist over several generations. Preliminary results indicate a modest number of markers associated with several traits at nominal significance level may be enough for developing prediction models in low LD populations. However, identification of sufficient markers is important for developing prediction models with high accuracy in low LD populations. Application of GS in high LD advanced generation populations or candidate gene-based genomic selection in low LD early generation populations promises to have a major impact on forest tree breeding.
**Figure Legend**

Figure 1: Improving the efficiency of tree breeding with GS. Genomic selection effectively replaces expensive and time consuming progeny trials which may take between 8-20 years depending on the species and the breeding system thus shortening the breeding cycle. With the accuracies of genomic selection similar to traditional phenotype based selection, marker genotypes of seedlings can be used for selecting superior individuals for the next breeding cycle. The selected individuals can be used for field testing and clonal trials. The genotype and phenotype information from some of these individuals can be used to re-train and refine the model.
Thavamanikumar et al. Dissection of complex traits in forest trees

Table 1. Nucleotide diversity and linkage disequilibrium studies in forest tree species

<table>
<thead>
<tr>
<th>Common name</th>
<th>Scientific name</th>
<th>Function of genes studied</th>
<th>Number of genes studied</th>
<th>Mean total nucleotide diversity ($\theta$)</th>
<th>LD decay</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loblolly pine</td>
<td><em>Pinus taeda</em></td>
<td>Wood formation</td>
<td>19</td>
<td>0.00398</td>
<td>Within 2000 bp</td>
<td>Brown et al. (2004)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Drought response</td>
<td>18</td>
<td>0.00507</td>
<td>Within 800 bp</td>
<td>Gonzalez-Martinez et al. (2006)</td>
</tr>
<tr>
<td>Maritime pine</td>
<td><em>Pinus pinaster</em></td>
<td>Wood formation</td>
<td>8</td>
<td>0.00241</td>
<td>-</td>
<td>Eveno et al. (2008)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Drought response</td>
<td>11</td>
<td>0.00548</td>
<td>-</td>
<td>Pot et al. (2005)</td>
</tr>
<tr>
<td>Radiata pine</td>
<td><em>Pinus radiata</em></td>
<td>Wood formation</td>
<td>8</td>
<td>0.00186</td>
<td>-</td>
<td>Pot et al. (2005)</td>
</tr>
<tr>
<td>Scots pine</td>
<td><em>Pinus sylvestris</em></td>
<td>Cold-related</td>
<td>14</td>
<td>0.00600</td>
<td>Within 200 bp</td>
<td>Wachowiak et al. (2009)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Allozyme coding</td>
<td>6</td>
<td>0.01000$^a$</td>
<td>No appreciable decay</td>
<td>Pyhajarvi et al. (2011)</td>
</tr>
<tr>
<td>Douglas Fir</td>
<td><em>Pseudotsuga menziesii</em></td>
<td>Cold Hardiness and wood formation</td>
<td>18</td>
<td>0.00655</td>
<td>Within 1500 bp</td>
<td>Krutovsky and Neale (2005)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cold Hardiness</td>
<td>121</td>
<td>0.00435</td>
<td>Within 1000 bp</td>
<td>Eckert et al. (2009b)</td>
</tr>
<tr>
<td>Norway spruce</td>
<td><em>Picea abies</em></td>
<td>Growth cessation</td>
<td>22</td>
<td>0.00208</td>
<td>Within 100 bp</td>
<td>Heuertz et al. (2006)</td>
</tr>
<tr>
<td>White spruce</td>
<td><em>Picea glauca</em></td>
<td>Multiple functions</td>
<td>105</td>
<td>0.00430</td>
<td>Within 65 bp</td>
<td>Pavy et al. (2012)</td>
</tr>
<tr>
<td>European aspen</td>
<td><em>Populus tremula</em></td>
<td>Multiple functions</td>
<td>5</td>
<td>0.01110</td>
<td>Within 500 bp</td>
<td>Ingvarsson (2005)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Multiple functions</td>
<td>77</td>
<td>0.00420</td>
<td>Within 200 bp</td>
<td>Ingvarsson (2008)</td>
</tr>
<tr>
<td>Balsam poplar</td>
<td><em>Populus balsamifera</em></td>
<td>Multiple functions</td>
<td>590</td>
<td>0.00280</td>
<td>No appreciable decay</td>
<td>Olson et al. (2010)</td>
</tr>
<tr>
<td>Japanese cedar</td>
<td><em>Cryptomeria japonica</em></td>
<td>Multiple functions</td>
<td>7</td>
<td>0.00251</td>
<td>-</td>
<td>Kado et al. (2003)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Multiple functions</td>
<td>5</td>
<td>0.00213</td>
<td>-</td>
<td>Kado et al. (2006)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Multiple functions</td>
<td>10</td>
<td>0.00156</td>
<td>-</td>
<td>Kado et al. (2008)</td>
</tr>
</tbody>
</table>

$^a$ Watterson’s theta: $\theta_w$
Table 2. Association mapping studies in forest tree species

<table>
<thead>
<tr>
<th>Common name</th>
<th>Scientific name</th>
<th>Markers used</th>
<th>Traits studied</th>
<th>Marker-trait associations found</th>
<th>Annotation of the associated SNPs</th>
<th>Independent validation attempt</th>
<th>Percentage of phenotypic variation explained</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shining gum</td>
<td>Eucalyptus nitens</td>
<td>SNP (25), SSR (14)</td>
<td>Wood quality</td>
<td>2</td>
<td>I</td>
<td>Yes</td>
<td>4.6</td>
<td>Thumma et al. (2005)</td>
</tr>
<tr>
<td>Tasmanian blue gum</td>
<td>Eucalyptus globulus</td>
<td>SNP (11)</td>
<td>Wood quality</td>
<td>1</td>
<td>S</td>
<td>Yes</td>
<td>-</td>
<td>Thumma et al. (2009)</td>
</tr>
<tr>
<td>Spotted gum</td>
<td>Corymbia citriodora subsp. variegata</td>
<td>SNP (74)</td>
<td>Wood quality, growth, disease resistance</td>
<td>9</td>
<td>S, I</td>
<td>Yes</td>
<td>3.4 to 8</td>
<td>Dillon et al. (2012)</td>
</tr>
<tr>
<td>European aspen</td>
<td>Populus tremula</td>
<td>SNP (41), SSR (26), SNP (39)</td>
<td>Phenology</td>
<td>2</td>
<td>NS</td>
<td>No</td>
<td>1.5 to 5</td>
<td>Ingvarsson (2008)</td>
</tr>
<tr>
<td>Blackbutt</td>
<td>Eucalyptus pilularis</td>
<td>SNP (52), SSR (12)</td>
<td>Wood quality</td>
<td>2</td>
<td>NS</td>
<td>No</td>
<td>4.1 to 4.2</td>
<td>Sexton et al. (2010)</td>
</tr>
<tr>
<td>Loblolly pine</td>
<td>Pinus taeda</td>
<td>SNP (58), SSR (22)</td>
<td>Wood quality</td>
<td>4</td>
<td>I, NS</td>
<td>No</td>
<td>2.2 to 3.6</td>
<td>Gonzalez-Martinez et al. (2007)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SNP (46)</td>
<td>-</td>
<td>Carbon isotope discrimination</td>
<td>4</td>
<td>I, NS</td>
<td>No</td>
<td>0.5 to 3.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SNP (3938), SSR (23)</td>
<td>Disease resistance</td>
<td>10</td>
<td>S, NS, UTR</td>
<td>No</td>
<td>4.8 to 7.2</td>
<td>Quesada et al. (2010)</td>
</tr>
<tr>
<td>Radiata pine</td>
<td>Pinus radiata</td>
<td>SNP (149)</td>
<td>Wood quality</td>
<td>10</td>
<td>S, I</td>
<td>Yes</td>
<td>2.0 to 6.5</td>
<td>Dillon et al. (2010)</td>
</tr>
<tr>
<td>Maritime pine</td>
<td>Pinus pinaster</td>
<td>SNP (384)</td>
<td>Growth and wood quality</td>
<td>2</td>
<td>S</td>
<td>Yes</td>
<td>10.0 to 11.4</td>
<td>Lepoittevin et al. (2011)</td>
</tr>
<tr>
<td>Douglas Fir</td>
<td>Pseudotsuga menziesii</td>
<td>SNP (228), SSR (6)</td>
<td>Isozymes (25), Cold-Hardiness</td>
<td>30</td>
<td>I, S, NS</td>
<td>No</td>
<td>1.9 to 4.3</td>
<td>Eckert et al. (2009a)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SNP (113)</td>
<td>SSR (26)</td>
<td>Growth Cessation</td>
<td>6</td>
<td>S, NS</td>
<td>No</td>
<td>6.2 to 12.7</td>
</tr>
<tr>
<td>White Spruce</td>
<td>Picea glauca</td>
<td>SNP (944)</td>
<td>Wood quality</td>
<td>25</td>
<td>S, NS, UTR</td>
<td>No</td>
<td>2.6 to 5.4</td>
<td>Beaulieu et al. (2011)</td>
</tr>
<tr>
<td>Sika Spruce</td>
<td>Picea sitchensis</td>
<td>SNP (339), SNP (98)</td>
<td>Bud set timing and cold-hardiness</td>
<td>45</td>
<td>S, NS</td>
<td>No</td>
<td>0.7 to 5.4</td>
<td>Holliday et al. (2010)</td>
</tr>
</tbody>
</table>

*a Used for identification/accounting of population structure

*b I – Intron, S – Synonymous, NS – Nonsynonymous, UTR – Untranslated Regions
References


Calus MPL, Veerkamp RF (2007) Accuracy of breeding values when using and ignoring the polygenic effect in genomic breeding value estimation with a marker density of one SNP per cM. Journal of Animal Breeding and Genetics 124 (6):362-368


Thavamanikumar et al. Dissection of complex traits in forest trees


Kado T, Matsumoto A, Ujino-Ihara T, Tsumura Y (2008) Amounts and patterns of nucleotide variation within and between two Japanese conifers, sugi (Cryptomeria japonica) and hinoki (Chamaecyparis obtusa) (Cupressaceae sensu lato). Tree Genetics & Genomes 4 (1):133-141


Mangin B, Siberchicot A, Nicolas S, Doligez A, This P, Cierco-Ayrolles C (2011) Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. Heredity:1-7


resource dedicated to wood formation in Eucalyptus. Bmc Plant Biology 9:36


Thumma BR, Baltunis BS, Bell JC, Emebiri LC, Moran GF, Southerton SG (2010a) Quantitative trait locus (QTL) analysis of growth and vegetative propagation traits in Eucalyptus nitens full-sib families. Tree Genetics & Genomes 6 (6):877-889


Thumma BR, Southerton SG, Bell JC, Owen JV, Henery ML, Moran GF (2010b) Quantitative trait locus (QTL) analysis of wood quality traits in Eucalyptus nitens. Tree Genetics & Genomes 6 (2):305-317


Wachowiak W, Balk PA, Savolainen O (2009) Search for nucleotide diversity patterns of local adaptation in dehydrins and other cold-


Discovery population
develop the prediction model

Validation population
validate the model using genotype information only

Select best genotypes
genotype the seedlings and select best individuals using genomic breeding values

Field testing, clonal trials and deployment

Make crosses
Induce early flowering and cross the selected individuals

Re-train the model
Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:
Thavamanikumar, S; Southerton, SG; Bossinger, G; Thumma, BR

Title:
Dissection of complex traits in forest trees - opportunities for marker-assisted selection

Date:
2013-06-01

Citation:

Persistent Link:
http://hdl.handle.net/11343/283141