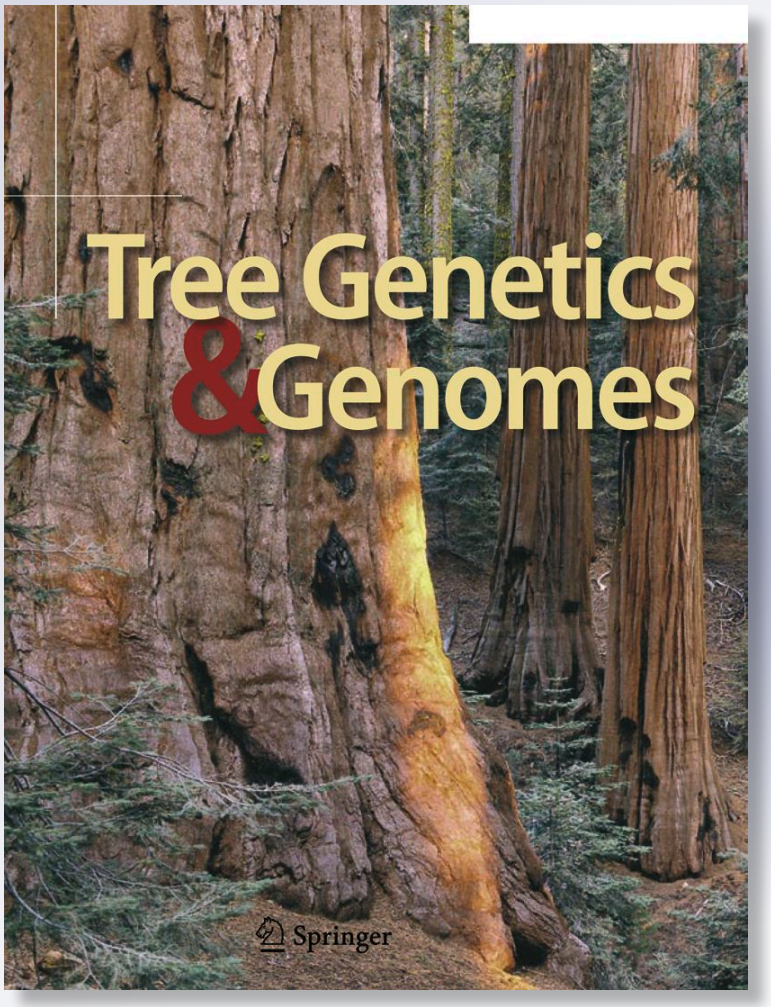
SNP markers trace familial linkages in a cloned population of Pinus taeda*— prospects for genomic selection*

Jaime Zapata-Valenzuela, Fikret Isik, Christian Maltecca, Jill Wegrzyn, David Neale, Steve McKeand & Ross Whetten

Tree Genetics & Genomes

ISSN 1614-2942

Tree Genetics & Genomes
DOI 10.1007/s11295-012-0516-5



Springer

Tree Genetics & Genomes
DOI 10.1007/s11295-012-0516-5

ORIGINAL PAPER

SNP markers trace familial linkages in a cloned population of *Pinus taeda*—prospects for genomic selection

Jaime Zapata-Valenzuela · Fikret Isik · Christian Maltecca · Jill Wegrzyn · David Neale · Steve McKeand · Ross Whetten

Received: 24 September 2011 / Revised: 2 April 2012 / Accepted: 17 April 2012

Abstract Advances in DNA sequencing technology have made possible the genotyping of thousands of single-nucleotide polymorphism (SNP) markers, and new methods of statistical analysis are emerging to apply these advances in plant breeding programs. We report the utility of markers for prediction of breeding values in a forest tree species using empirical genotype data (3,406 polymorphic SNP loci). A total of 526 *Pinus taeda* L. clones tested widely in field trials were phenotyped at age 5 years. Only 149 clones from 13 full-sib crosses were genotyped. Markers were fit simultaneously to predict marker additive and dominance effects. Subsets of the 149 genotyped clones were used to train a model using all markers. Cross-validation strategies were followed for the remaining subset of genotyped individuals. The accuracy of genomic estimated breeding values ranged from 0.61 to 0.83 for wood lignin and cellulose content, and from 0.30 to 0.68 for height and volume traits. The accuracies of predictions based on markers were comparable with the accuracies based on pedigree. Because of the small number of SNP markers used and the relatively small population size, we suggest that observed accuracies in this study trace familial linkage rather than historical linkage disequilibrium with trait loci. Prediction accuracies of models that use only a subset of markers were generally comparable with the accuracies of the models using all markers, regardless of whether markers are associated with the phenotype. The results suggest that using SNP loci for selection instead of phenotype is efficient under different relative lengths of the breeding cycle, which would allow cost-effective applications in tree breeding programs. Prospects for applications of genomic selection to *P. taeda* breeding are discussed.

Keywords Loblolly pine · Marker-aided selection · Quantitative genetics · Genomic selection · Marker–trait association

Communicated by D. Grattapaglia

Jaime Zapata-Valenzuela, Fikret Isik, and Ross Whetten contributed equally to this work.

J. Zapata-Valenzuela · F. Isik (✉) · S. McKeand · R. Whetten
Department of Forestry and Environmental Resources,
North Carolina State University,
Raleigh, NC, USA
e-mail: fisik@ncsu.edu

C. Maltecca
Department of Animal Science, North Carolina State University,
Raleigh, NC, USA

J. Wegrzyn · D. Neale
Department of Plant Science, University of California-Davis,
Davis, CA, USA

Introduction

Selection based on phenotypes and resemblance among relatives has been successful in forest tree breeding programs during the last five decades (McKeand et al. 2006). However, the current breeding strategy for pines is logistically complex, expensive, and time-consuming, because of their long breeding cycles and large physical sizes. Many traits of interests are measured indirectly, due to the long life and slow growth of trees, and these traits generally have low (<0.2) narrow-sense heritabilities (White et al. 2007). Early reports discussed the rationale for marker-aided selection in conifers as another form of indirect selection, with the primary objective of reducing the length of the breeding cycle (Neale and Williams 1991).

Initial uses of molecular markers in conifers focused on mapping, dissection, and understanding of gene action of

Published online: 31 May 2012

loci controlling quantitative traits in forest trees (Knott et al. 1997; Kaya et al. 1999). Extensive quantitative trait locus (QTL) studies in two independent *Pinus taeda* populations reported that QTLs explained between 5.3% and 15.7% of the phenotypic variance observed within the population for several wood property traits, and those populations in turn contained only a fraction of the genetic variation present in the whole loblolly pine population (Neale et al. 2002; Brown et al. 2003). These results are consistent with the expectation that individual pedigrees do not capture all the genetic variation in the population for quantitative traits and that many genes of relatively small effect are likely to be involved in controlling complex phenotypes in conifer populations. These concerns were identified early in the development of research programs on marker-assisted selection in conifers (Neale and Williams 1991; Strauss et al. 1992), but the experimental verification of these factors led to a shift in emphasis toward association genetics approaches as a tool for dissecting the genetic basis of complex phenotypes (Neale et al. 2002).

Conifers are well-suited for association genetics studies to identify specific genes associated with complex phenotypes, because of the relatively low levels of linkage disequilibrium (LD) in the genome (Neale and Savolainen 2004). For example, in *P. taeda* L. (commonly known as loblolly pine) and in *Pinus sylvestris* L., LD decays to mean r^2 values of less than 0.25 within 2,000 bp (Brown et al. 2004; Pyhäjärvi et al. 2007). Significant associations between single-nucleotide polymorphism (SNP) markers and wood property or carbon-isotope discrimination traits have been identified in *P. taeda* (González-Martínez et al. 2007, 2008). Cumbie et al. (2011) reported associations of SNP markers with height growth and carbon isotope discrimination in *P. taeda*. Quesada et al. (2010) reported associations of SNP markers with resistance to a fungal pathogen, *Fusarium circinatum* in *P. taeda*. In all these reports, the fraction of phenotypic variation explained by individual markers was relatively low, consistent with the model of polygenic inheritance of most complex traits in conifers.

An ongoing challenge of marker-assisted selection (MAS) in trees is the complex inheritance of quantitative traits. The proportion of the total phenotypic variance explained by markers was considered a key determinant of the efficiency of MAS in breeding (Lande and Thompson 1990). However, Goddard and Hayes (2009) suggested that markers individually do not have to explain considerable variation in phenotype in order to be collectively good predictors of breeding values. The results of both QTL and SNP association studies reviewed above are consistent with the classic infinitesimal model that many genes of small effect contribute to genetic variation in many conifer phenotypes (Neale 2007), and experimental studies on laboratory model species and in agricultural breeding populations have shown that this model is accurate for many phenotypic traits in many other species as well (Mackay et al. 2009). A key challenge for approaches that seek to identify individual genes associated with specific phenotypes is the need for a stringent statistical threshold for declaring a significant effect to avoid high levels of false positive results. This can lead to high levels of false-negative results or failure to detect loci that fall below the threshold of statistical significance, which in turn leads to an inability to account for more than a fraction of the genetic variation in the phenotype of interest (Yang et al. 2010).

A new technology called genomic selection (GS) is revolutionizing dairy cattle breeding by reducing the length of the breeding and selection cycle (Hayes et al. 2009; VanRaden et al. 2009) and has been proposed to be useful for breeding annual crops and trees as well (Piepho 2009; Grattapaglia and Resende 2010; Jannink et al. 2010). GS is based on estimating breeding values as the sum of effects of individual marker alleles (or haplotypes of those alleles) across the entire genome, without imposing a specific threshold of statistical significance for the association of any particular marker with the trait of interest. With sufficiently dense marker coverage, all genes affecting traits of interest are assumed to be in LD with at least one marker (Meuwissen et al. 2001). Marker effects are first estimated in a large training population, for which both phenotypic and genotypic data are available, and the resulting statistical model of allele effects is then used to predict breeding values in prediction populations for which genotypes, but no phenotypic data are available. This approach is focused on prediction of breeding values, rather than identification of specific genes related to a particular trait, and is therefore less limited by the polygenic inheritance of many phenotypes in agricultural and forest species (Jannink et al. 2010). The term "polygenic inheritance" refers only to the fact that many genes, each of relatively small effect, are involved in controlling phenotypic variation. This should not be confused with the term "polygenic effect," used in some cases to refer to genetic variation that is not accounted for by markers (Meuwissen and Goddard 2010).

Utility of GS in forest trees has been tested in few studies. Resende et al. (2011) reported comparable accuracy of GS to the accuracy of selection based on phenotype at two sites. They reported a selection efficiency of 53% to 112% for growth. The highest-density marker dataset for pines reported to date has been that described by Eckert et al. (2010), also used in the studies of Quesada et al. (2010), Cumbie et al. (2011), Resende et al (2011), and this study. While the 3,000 to 4,000 SNP used in these studies do provide extensive coverage of the pine linkage map (Eckert et al. 2010), they are unlikely to provide markers in LD with all (or even a substantial proportion of) the loci controlling phenotypes of commercial interest. In the absence of high-density SNP

genotype datasets, Grattapaglia and Resende (2010) used deterministic simulation studies to explore the potential value of GS in tree breeding, assessing the effects of the level of LD, the effective population size, the size of the training population, and trait heritability. Those authors suggested that the obstacle of low levels of LD in most forest tree species could be overcome by working in populations of related individuals with relatively small effective population size, such as elite populations used in some tree breeding programs. This is consistent with a previous simulation study, which also commented that making predictions in a population descended from the individuals used for training the GS model would require lower marker density and smaller training population size than making predictions in a population of individuals with no known relationship to the training population (Meuwissen 2009).

In this study, we tested the accuracy of GS methods in predicting breeding values using a dataset of 3,406 SNP loci in a population of 149 clonally replicated *P. taeda* genotypes derived from a structured mating design. Our objectives were to (1) evaluate the accuracy of genomic estimated breeding values (GEBV) compared with breeding values estimated by a traditional polygenic model in *P. taeda*, (2) compare different cross-validation methods as a means of evaluating statistical models, (3) compare the accuracy of GEBV predicted using the full set of markers against values predicted using marker subsets of varying sizes, and (4) explore the efficiency of genomic selection compared with selection based on pedigree alone.

Methods

Experimental population and phenotypes

Eighteen *P. taeda* parents were used as females and males to generate 13 full-sib families. The majority of families were obtained by single-pair mating. Several families were genetically related by a common male or female parent (see pedigree in Fig. 1). The number of offspring per family ranged from 1 to 34, and they were cloned via somatic embryogenesis (Bettinger et al. 2009). A total of 526 clones were field-tested on 16 sites planted between 2000 and 2002. Out of 526 clones phenotyped, only 149 clones were genotyped. An alpha-lattice incomplete block design with single tree plots was used as the field layout. Every cloned progeny was not represented in each incomplete block or at each site. A clone was represented by 16 to 50 ramets across all sites. The test sites were located across varying edaphic conditions and productivity classes in the coastal plain of South Carolina, Georgia, and in the gulf coast of Mississippi in the southern US.

A total of 21,974 pine trees were measured at 5 years after planting. The median number of observation per clone was 40. Height, diameter at 1.4 m above ground, and stem forking (as binary variable) were assessed and used to calculate volume of individual trees according to Goebel and Warner (1966). Wood cores (12 mm thick) were collected from trees from three sites to determine lignin and cellulose contents using near-infrared spectroscopy models as described by Hodge and Woodbridge (2010). The models for lignin and cellulose had a correlation of 0.97 and 0.82, respectively.

Genotyping

We had resources to genotype only 149 pine clones out of 526 in the experimental population, and they were selected randomly. Needles were sampled from one or multiple copies of each of 149 cloned offspring from one of three field progeny test locations during the 2008 growing season. Approximately 23,000 SNP markers, of which we chose 5,379 for genotyping, were identified through the resequencing of 7,535 uniquely expressed sequence tag contigs in 18 *P. taeda* haploid megagametophytes, as described by Eckert et al (2010). Selection of SNP for genotyping was based largely on quality scores derived from the original sequence data and not on functional or site annotations. This ensured thorough coverage of the available sequence

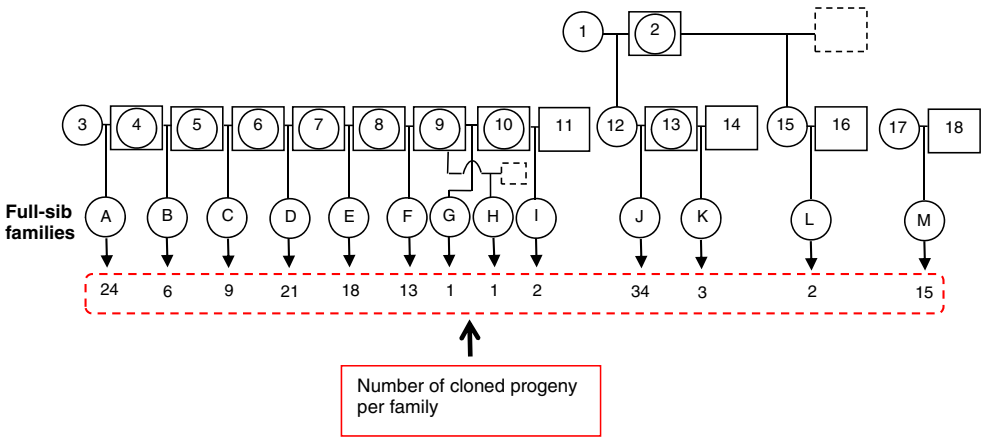


Fig. 1 Pedigree structure of the 149 clones used for GS predictions. Parents 2, 4, 5, 6, 7, 8, 9, 10, and 13 were used either as female or male in a partial diallel mating design fashion. Parent 9 crossed with 10 gave G, but crossed with unknown male gave H. *Dashed squares* were unknown male contribution (wind-pollinated). *Letters* indicated 13 full-sib families. Number of cloned progeny produced for each full-sib family is given *below arrows*

resource for *P. taeda* (http://dendrome.ucdavis.edu/adept2/). Genotyping of SNP utilizing the Infinium platform (Illumina, San Diego) was carried out at Illumina. Arrays were imaged on a Bead Array reader (Illumina), and genotype calling was performed using BeadStudio v. 3.1.3.0 (Illumina). Sample phenotypes and SNP genotype data are available for all samples through the TreeGenes database at http://dendrome.ucdavis.edu/DiversiTree (Wegrzyn et al. 2008).

We carried out exploratory data analysis on SNP markers using the ALLELE procedure of SAS GENETICS software (SAS Institute Inc. 2010). Out of 5,379 SNP markers analyzed, 758 SNP markers failed in genotyping or contained missing genotypes for more than 15% of the clones, and 1,215 markers were monomorphic (homozygous in all individuals). Thus, a total of 3,406 SNP markers were informative and used for prediction of breeding values. Missing genotypes for each of the remaining 3,406 marker loci were imputed using the genotype frequencies for loci across all 149 clones in the dataset by randomly sampling a genotype for each missing data point based on the frequencies of genotypes at that locus. Three independent imputations were used to sample the stochastic effects of the imputed genotypes, using a stochastic method that imputes a categorical genotype (0, 1, or 2 copies of the minor allele) based on the frequency of genotypes observed at the same locus across all families.

Statistical analyses

The following linear mixed model was fit to predict breeding values of 526 clones for height, volume, lignin, and cellulose content using measurements from field progeny tests:

$$y = Xb + Zu + e \tag{1}$$

where y is the vector of phenotypic observations, b is the vector of fixed intercept and design variables (site effects, incomplete block effects within sites); u is the vector of random tree (clone) effect and clone by site interaction effect; X and Z are incidence matrices, and e is the vector of residuals. The variance of y is var(y)=V=ZGZ'+R, where G and R are variance–covariance matrices of random effects and residuals, respectively. The G matrix was substituted by a numerator relationship matrix (A) derived from the pedigree of all 526 clones and scaled by a ratio of residual variance and the variance of clones (Mrode 2005). Using the variance components, clone mean repeatability values were calculated for traits as $H_c^2=\sigma_c^2/(\sigma_c^2+\sigma_{cs}^2/_s+\sigma_e^2/_{sr})$, where σ_c^2 is the variance explained by the clone effect, σ_{cs}^2 is the variance due to clone by site interaction, s is the number of sites (16 for growth traits, three for wood traits), σ_e^2 is the residual variance, and r is the mean number of ramets per clone per site for each trait. ASReml software (Gilmour et al. 2009) was used to estimate variances and solve mixed-model equations. The predicted breeding values (EBV) of 149 clones from this analysis were used as 'pseudo-phenotypes' for further marker-based prediction models. We also ran the same model without pedigree to obtain EBV of the 149 cloned offspring. Analysis without pedigree eliminates the contribution of parents to offspring breeding values, and the resulting estimated breeding values can be compared with those obtained from a model that includes pedigree information to determine if deregression of breeding values is needed.

Prediction of breeding values based on additive and dominant marker effects

The following linear mixed model was used to predict GEBV of clones using markers as independent variables.

$$\hat{u} = \mu 1_n + Ta + Wd + e \tag{2}$$

Where $\hat{u}$ is the vector of pseudo-phenotypes or predicted breeding values (EBV) of 149 clones obtained from phenotype (Eq. 1); μ is the overall mean; 1_n is the vector of 1 s with n dimension (number of trees); a is the random additive marker effects with expectations $\hat{a} \sim N\left(0, I\sigma_a^2\right)$; T is the design matrix which relates observations to additive marker effect; d is the random dominant marker effects with expectations $\hat{d} \sim N\left(0, I\sigma_d^2\right)$; W is the design matrix which relates observations to dominant marker effect; e is the vector of random residual effects with expectations $\hat{e} \sim N\left(0, I\sigma_e^2\right)$, and I is the identity matrix (Mrode 2005). The GEBV of the individual j across all loci was obtained by solving the following equation (Legarra and Misztal 2008).

$$\text{GEBV} = \sum\nolimits_i T_{ij}\hat{a}_i + \sum\nolimits_i W_{ij}\hat{d}_i \tag{3}$$

Where T and W are the conditional design matrices allocating each j^{th} EBV to the marker effect at the i^{th} locus, determined as $+a_i$ or $-a_i$ (coded as 11 or 22) for the additive effects, and d_i (coded as 12) for the heterozygote or dominant effect. In matrix form, the mixed-model equations solved to predict the additive and dominant marker effects were:

$$\begin{bmatrix} X'X & X'T & X'W \\ T'X & T'T + A^{-1}\lambda_1 & T'W \\ W'X & W'Z & W'W + D^{-1}\lambda_2 \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{a} \\ \hat{d} \end{bmatrix} = \begin{bmatrix} X'\hat{u} \\ T'\hat{u} \\ W'\hat{u} \end{bmatrix} \tag{4}$$

Where $\lambda_1 = \sigma_e^2/\sigma_a^2$ and $\lambda_2 = \sigma_e^2/\sigma_d^2$ are the shrinkage factors (lambda) for additive and dominant marker effects (Mrode 2005), A^{-1} is the inverse of marker additive genetic variance, D^{-1} is the inverse of marker dominance genetic

variances; $\widehat{b}$, $\widehat{a}$, and $\widehat{d}$ are the vectors of solutions (predictions) for fixed effect (mean), marker additive, and dominance genetic effects, respectively. Other terms were explained before. This method assumes that all the markers have the same additive and dominant variances, a similar assumption of classical BLUP analysis (Legarra et al. 2008). We also specified an extra term (*Zu*) for "polygenic effects" to disentangle markers in LD with the trait loci and the markers that track relationship in the reference population (Liu et al. 2011).

We used the GS3 software described by Legarra et al. (2008) to solve the large number of mixed-model equations for markers. The software implements the Gauss-Seidel method with residual updating, as an iterative method to solve linear systems of equations and obtain estimates of best linear unbiased estimates and BLUP (Legarra and Misztal 2008). The Gauss-Seidel algorithm is converted to a Gibbs sampler to estimate the variance components (VCE option in GS3 software). A prior inverted chi-squared distribution was used for variance component estimation, using estimated additive, dominant and residual variances reported by Isik et al. (2003, 2005) and Sykes et al. (2006) for *P. taeda*. The Gibbs sampler was run for 10,000 iterations for lignin and cellulose contents and 5 million iterations for height and volume. The first 2,000 cycles were discarded as burn-in. We tested the convergence of the sample variances produced by the MCMC iterations using the coda R package (Plummer et al. 2006). We also examined the effect of various shrinkage factors on accuracy of predictions by fixing the ratio of error and genetic variances and running a BLUP option in the GS3 software. This option allowed fixing the variance components for different levels of ratios expressed as logarithmic value in base 10.

Accuracy and cross-validation of predictions

The cross-validation methods were designed as follows: The 149 clones were divided into a training data set and a validation data set. In the first scenario, about 90% of the clones (136) were sampled for the training set, either within each of the 13 families or at random from the whole population without family consideration. The remaining clones were used for the validation (13 clones). In the second scenario, about 50% of clones (75) were sampled either within family or randomly from the whole population for training, and the remaining clones were used for validation (74 clones). The model parameters estimated in the training set were used to predict GEBV in the validation set. For each scenario, three independent samplings were carried out. For training and validation sets, we used all 3,406 markers.

We used breeding values of trees based on clonally replicated phenotype measurements (EBV) as the basis (true BV) to compare with the breeding values based on the markers (GEBV). The accuracies of the GEBV were estimated as the correlation (r) between the true breeding value EBV (g) based on the phenotype (pedigree) and ($\widehat{g}$) the GEBV based on markers.

$$r_{(g,\widehat{g})} = \mathrm{Cov}(g,\widehat{g})/\sqrt{\sigma_g^2\sigma_{\widehat{g}}^2} = \sqrt{\sigma_g^2/\sigma_{\widehat{g}}^2} \quad (5)$$

where $\sigma_{\widehat{g}}^2$ is the variance of GEBV, and σ_g^2 is the variance of EBV. Theoretically, the covariance is $\mathrm{Cov}(g,\widehat{g}) = \sigma_{\widehat{g}}^2$. Mean accuracy values ($r_{\mathrm{m}}$ for marker prediction accuracy) are reported for each scenario. Each accuracy value is the mean of nine replications, e.g., three independently sampled sets of training and validation clones for each of three imputed data sets of genotypes.

We permuted the genotypes and phenotypes of individuals to break up any association in the data to test a hypothesis of no marker–trait association. Permutation was repeated three times, and the resulting datasets were fit using Eq. 2 to estimate the accuracy values between markers and traits in the absence of any true association between genotypes and phenotypes.

Using a subset of marker loci for GEBV

Using a smaller set of markers (based on association testing) to predict the genetic merit of individuals could be an attractive strategy. If successful, this might reduce the cost of genotyping and also might simplify analytical approaches for routine genetic data analysis. We explored the performance in GS for a reduced set of markers using SNPs selected by association tests carried out in TASSEL (Bradbury et al. 2007), as well as SNP selected disregarding the association with phenotype. The mixed linear model employed included markers as fixed effects and clones as random effects, plus a residual term. The phenotypes used for association were BLUP values of clones from each training population. We did not include a population stratification Q-matrix because population structure is unlikely to be important in a set of progeny from a structured mating design. The average relationship between clones was estimated by a kinship K-matrix from all markers to help correct for spurious associations. Independent association tests were carried out on different training populations composed of randomly selected clones and gave variable results in terms of the number of significant markers.

Once the association tests were run, different numbers of markers were selected according the following three threshold probability values—$Pr \leq 0.1$, $Pr \leq 0.01$, and $Pr \leq 0.001$, without probability adjustment for multiple testing. The number of significant markers detected in different training groups varied, so the range of number of markers identified at each probability threshold is reported. Finally, we ran our

marker model (Eq. 2) on the training/validation clones to obtain accuracy values (r_m) between EBV and GEBV obtained from the three described subsets of markers. As a control, random sets of markers were also used in separate training/validation groups for direct comparison of the accuracy of random markers with the above subsets selected based on association testing.

Efficiency of genomic selection

We estimated the efficiency of GS based on the correlated response of the target unknown trait (EBV) and genetic parameters from marker data as $CR_c = i_m\ H_m\ r_m\ \sigma_c$, where i_m is the selection intensity using markers, H_m is the square root of the proportion of trait heritability explained by markers (Lande and Thompson 1990), r_m is the accuracy of the marker-only model as a measure of the genetic correlation between the markers and the phenotypes obtained from Eq.5, and σ_c is the square root of the variance explained by clone effect from progeny test data. Additionally, the direct response to selection from progeny test data is given by $R_c = i_c\ H_c\ \sigma_c$, where i_c is the selection intensity based on classical BLUP, H_c is the square root of the clone mean repeatability, i.e., the genetic variance explained by the pedigree effect over the phenotypic variance. The relative efficiency per year (E/year) of indirect selection on the direct response to selection is given by the ratio of CR_c/R_c (Falconer and Mackay 1996).

$$\mathrm{E/year} = r_m[(H_m/T_m)/(H_c/T_c)] \quad (6)$$

It is assumed that the selection intensity of clones based on classical BLUP (i_c) and based on GS (i_m) is the same, and these terms were dropped from Eq. 6. The generation intervals in scenarios with and without GS are variable according to the number of years required to accomplish each breeding phase. We used five different ratios of T_m/T_c, ranging from 0.1 to 0.5 to calculate efficiency per year using markers versus progeny testing. We assumed that a cycle of clonal progeny testing for *P. taeda* takes 15 years. Different time ratios (7.5, 6, 4.5, 3, and 1.5 years) for using GS were modeled. Selection efficiency was calculated for two scenarios, using all the markers and using a subset of markers associated with the phenotypes under a probability $Pr \leq 0.001$, for lignin and cellulose.

Results

Variance explained by marker effects

Additive marker effects explained considerable fractions of the variance for wood chemical traits lignin (18.2%) and cellulose (22.9%), whereas for height and volume traits, these variance components could not be estimated (Table 1). We found that marker variances converged for lignin and cellulose with low levels of autocorrelation in samples from the Markov chain produced by the GS3 software, but none of the additive, dominant, or residual variances reached convergence for height or volume, even after increasing the number of iterations from 30,000 to 5 million. Similar results were obtained for dominant marker effects, where the percentage of phenotypic variance explained by the markers was considerable for lignin (16.4%) and cellulose (5.7%), but dominant variances for height and volume traits were also undetermined.

The clone term accounted for 22.9% (cellulose) to 38% (tree height) of total phenotypic variance when a mixed model was fitted to progeny test data using all 526 clones (Table 1). Clone mean repeatability values based on variance components from progeny test data for height and volume were higher than the estimates for lignin and cellulose.

Relationships among clones do not affect the estimated breeding values

In cattle breeding, the contribution of parents to an individual bull's breeding values is high (Moser et al. 2010), and a deregression approach is recommended to remove parent average effects before using the breeding values as pseudo-phenotypes for genomic estimated breeding values (Garrick et al. 2009). This approach may not be crucial for cloned progeny tests in forest trees. We estimated breeding values of clones using pedigree first and then excluded pedigree and obtained a correlation of 0.997 between the two predictions. Also, repeatabilities of clone means for

Table 1 Clone mean repeatability (H_c^2) values with standard errors (SE) within parentheses, percent of phenotypic variances explained by clone effect (% σ_c^2) based on progeny test data, and percent of additive (%σ_a^2) and dominant (%σ_d^2) genetic variances explained by markers for lignin, cellulose, height, and volume traits in a cloned population of *P. taeda*

	Pedigree only model		Markers model	
Trait	H_c^2 (SE)	% σ_c^2	% σ_a^2	% σ_d^2
Lignin	0.76 (0.09)	32.9	18.2	16.4
Cellulose	0.69 (0.13)	22.9	22.9	5.7
Height	0.95 (0.01)	38.0	na[a]	na[a]
Volume	0.94 (0.01)	31.5	na[a]	na[a]

[a] na=variances could not be determined for height and volume, because the Gibbs sampler failed to converge, even after 5 million iterations using GS3 software

traits were high, ranging from 0.69 to 0.95 (Table 1), showing a greater weight given to information coming from clones rather than from relatives when BLUP of clones are estimated. We conclude that the accuracy of individual-genotype breeding values in our study is based almost entirely on the extensive replication of individuals (16 to 50 copies of each cloned genotype were measured), rather than on information derived from relatives, so deregression to remove shrinkage effects is not necessary. The estimated breeding values (EBV) obtained from the model including pedigree were used as pseudo-phenotypes for genomic estimated breeding value.

The accuracy of genomic estimated breeding values (GEBV)

The accuracies of predictions based on marker additive and dominance effects for traits using different cross-validation scenarios were generally higher for lignin and cellulose than height and volume (Table 2). The range of accuracies for the GEBV for lignin and cellulose was between 0.61 and 0.83, whereas the range for height and volume was between 0.30 and 0.68 across validation methods. One important finding for this population was that markers could predict clone genetic values as accurately as a model based on pedigree. The accuracy values were particularly comparable for lignin and cellulose traits.

We broke up the association of markers with the phenotypes three times, and we found near-zero accuracy values after running our marker effect model (data not shown). Secondly, including or excluding clones used in validation as part of the pedigree-only model to obtain EBV did not make a difference in terms of accuracy of GEBV of those clones. The correlation coefficient between the two cases was almost perfect (0.99).

Table 2 Accuracies of genomic estimated breeding values from markers (model 2) for two different validation populations, with 10% and 50% of clones sampled for validation

Trait	10% sampled		50% sampled	
	Within	Random	Within	Random
Lignin	0.70	0.66	0.76	0.69
Cellulose	0.83	0.61	0.79	0.76
Height	0.47	0.68	0.55	0.52
Volume	0.30	0.56	0.40	0.36

Sampling was carried out either as 10% of the clones within each cross (within) or random across all crosses. The accuracies are the average of nine sampling replications for each scenario

The standard error of accuracy values varied from 0.01 to 0.09. Number of markers used for training and validation was 3406

The effect of fixed error and additive genetic variances on accuracy of GEBV

The effect of different levels of fixed shrinkage values ($\lambda_1=\sigma_e^2/\sigma_a^2$) used in BLUP on the accuracy of GEBV for lignin and volume are presented in Fig. 2. The shrinkage factor is the inverse of heritability used in the BLUP and can be interpreted as the effect of heritability on accuracy of GEBV. The curves suggest that fixing variance components while using the BLUP option in GS3 software versus letting the software calculate variances from data using MCMC sampling had a small effect. For example, the accuracy values for lignin ranged from 0.61 to 0.70 for different ratios of error and additive genetic variances. The effect of shrinkage factor on predicted volume accuracy values was negligible.

GEBV based on subset of markers

The accuracies of predictions for traits using subsets of markers selected based on association testing are presented in Table 3. As expected, the accuracy values decreased as the number of markers was reduced. When we used a randomly sampled subset of markers of a comparable size to the subset selected based on association testing, they yielded similar results. For example, when we selected a subset of markers (ranged from 410 to 717) based on *F*-test probability values of $Pr \leq 0.1$, the accuracy of GEBV for traits ranged from 0.34 (volume) to 0.75 (cellulose). When a random sample of 800 SNP markers (without association testing) was used in estimation of genomic breeding values, the accuracy values were comparable, ranging from 0.37 (volume) to 0.72 (cellulose).

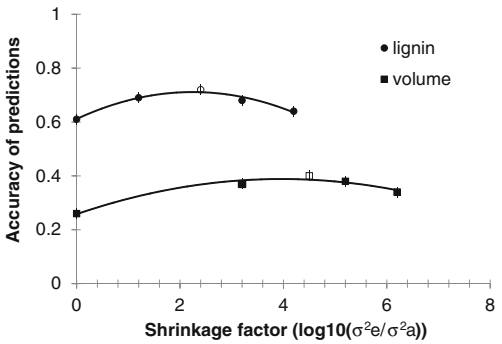


Fig. 2 The effect of shrinkage factor used in GS3 software with the BLUP option on the accuracy of predictions using SNP markers for lignin and volume. The shrinkage factor in BLUP is a ratio of error and genetic variance ($\lambda=\sigma_e^2/\sigma_a^2$). With the BLUP option in GS3, variances are not estimated from data but fixed by the analyst. The accuracies are based on cross-validation method 50/50% of clones sampled randomly within family. The *filled symbols* represent the variances fixed by the analyst. The *unfilled symbols* (one for lignin and one for volume) represent the shrinkage factor estimated by MCMC algorithm in GS3 software. Standard errors of each mean accuracy values were plotted

Table 3 Mean accuracies (r_m) of predictions for subsets of markers selected from association testing and random sampling without association testing

Probability value used	# of markers selected	Lignin	Cellulose	Height	Volume
All markers	3406	0.76	0.79	0.55	0.40
$Pr \leq 0.1$	563	0.72	0.75	0.50	0.34
Random	563	0.60	0.68	0.48	0.32
$Pr \leq 0.01$	132	0.69	0.68	0.48	0.36
Random	132	0.55	0.68	0.51	0.31
$Pr \leq 0.001$	18	0.67	0.56	0.39	0.25
Random	18	0.50	0.45	0.39	0.20

Training and validation for subsets of markers were based on random sampling of 50% of the clones within family. Randomly selected markers (rows starting with Random) produced comparable accuracy values to markers based on association tests (rows with *Pr* values of 1, ≤0.1, ≤0.01, and ≤0.001)

Standard errors of mean accuracy values for GEBV varied from 0.01 to 0.13

Efficiency of using SNP markers in selection compared with progeny testing

Selection efficiency (E/year) obtained using all markers was higher than that obtained using a subset of markers (Fig. 3). Using markers to predict genetic merit of trees for lignin was superior (E/year>1) compared with progeny testing under different time scenarios. The efficiency of GEBV over progeny testing was higher when we assumed that selection based on markers takes up to 50% of the time needed for progeny testing. A GEBV approach using either all markers or a subset of markers was very similar for lignin and cellulose under the assumptions compared, so we only plotted the results for lignin. Selection efficiencies for growth traits were not estimated since variances were not available from marker models.

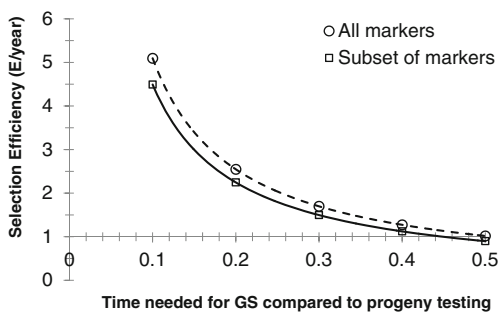


Fig. 3 Selection efficiency (E/year) of GS using all (3,406) or subsets of markers (4 to 32) compared with progeny testing for lignin content. We assumed that the time needed to complete GS is a fraction (0.1 to 0.5) of time required to complete progeny testing. GS is superior to progeny testing when efficiency is greater than 1

Discussion

Using empirical genomic and phenotypic data, we demonstrated the potential of SNP markers to predict genetic merit of cloned offspring in a small population of *P. taeda*. Markers were efficient in predictions of genomic estimated breeding values in this population, particularly for lignin and cellulose content.

We found considerable amounts of additive and dominant genetic variance explained by the markers for lignin and cellulose. On the other hand, models did not converge for growth traits, and thus estimates of variance components for those traits were not available. We speculate that increasing the training population size (number of unrelated clones) and number of SNP markers used in analysis of growth traits would help to converge the models and estimate variances. The difference between wood chemical traits (lignin and cellulose) and growth traits (height and volume) suggests that the genetic architecture of those traits differs, or that the SNP markers used are biased toward detecting variation in genes that control variation in wood properties. Lignin and cellulose contents may be controlled by genes that have alleles of larger effect or which are in stronger LD with the markers and therefore have greater power to explain phenotypic variation. Conversely, growth traits may be controlled by many genes of small, additive effect, or which have weak, if any, LD with the SNP markers. With the small population size, either explanation could lead to convergence issues in our markers model. Previous QTL and association studies for wood property and growth traits support these hypotheses (Kaya et al. 1999; Brown et al. 2003; González-Martínez et al. 2007).

Accuracies of genomic estimated breeding values

The accuracy of predictions based on markers was comparable to that of a traditional pedigree-only method, suggesting that SNP markers can be utilized in *P. taeda* breeding programs for selection of superior individuals and thus substantially increase genetic gains. When marker–phenotype relationships were permuted in the dataset, the accuracy of the predictions dropped to near zero, suggesting that markers are in LD with trait loci and thus are good predictors.

Effective population sizes (Ne) and marker density have previously been noted as the most significant factors in determining prediction accuracy (Hayes et al. 2009; Grattapaglia and Resende 2010). For height and volume, little, if any, variation was explained by SNP markers, and the prediction accuracies for these traits were lower. In order to capture historical LD between trait loci and markers, high marker density and large (>1,000 individuals) training populations will be needed to implement genomic selection for these traits, as reported by Solberg et al. (2008). In the near future, when a

reference sequence for the pine genome is achieved, increased marker density and discovery of additional genes will help provide resources useful in accounting for more phenotypic variation with genotypic information.

Genomic selection can provide higher accuracy than pedigree-based BLUP, if markers are in LD with loci that control genetic variation in the trait (Calus and Veerkamp 2007). In the absence of such LD, comparable accuracies from a marker model indicate that the predictive accuracy is largely from reconstruction of the relationship matrix by marker data (Legarra et al. 2008). There can still be important advantages of using markers for predictions, even when the accuracy is comparable to that from a pedigree-based analysis. The ability to obtain GEBV of genetic entries without a known pedigree could reduce the cost of breeding by eliminating the requirement for making multiple controlled crosses per seed parent, which is an expensive part of the tree breeding process. Construction of a realized genomic relationship matrix can be carried out in the absence of a known pedigree, allowing tree breeders to use a single pollen mix within an elite subline and still retain the potential to predict offspring performance using genomic relationships and BLUP methods.

The effect of variances on the accuracy of GEBV

The choice of shrinkage factor ($\lambda_1=\sigma_e^2/\sigma_a^2$) in BLUP approach had a small effect on the accuracy of GEBV, as shown for lignin and volume (Fig. 2). The highest accuracy value was reached when the variance components estimated from data using MCMC sampling, rather than any of the variances provided by the analyst. As the shrinkage factor increased (i.e., as heritability decreased), the power of the prediction decreased slightly. Using a deterministic simulation approach, Grattapaglia and Resende (2010) suggested that heritability of the trait had relatively a small impact on the accuracy of predictions. The findings shown in Fig. 2 are consistent with their suggestion, although we tested only the effects of changing the shrinkage factor and not changes in the true underlying heritability of the traits. The approach of estimating variance components from the data avoids the need to guess the adequate shrinkage factor value, as was done in an early report of a ridge regression method (Meuwissen et al. 2001). Also, the prediction is not sensitive for a trait where variance components are difficult to estimate, as was found for volume.

Cross-validation methods

Cross-validation of genome-wide predictions is important, because breeding programs are designed to make decisions about the likely future performance of the individuals in the selection population. One interesting aspect of the validation is that fewer individuals (75) in the training population (based on 50% sampling) seemed to be enough to reach a comparable prediction level versus using 136 clones in the training groups (based on 10% sampling). In general, a larger training population size (>1,000) is considered an important factor to achieve higher accuracies, especially for low-heritability traits where detection of markers in LD with loci controlling phenotypic variation is an important contributor to prediction accuracy (Hayes et al. 2009).

Including genotyped clones in the estimation of EBV, used as a pseudo-phenotype for later analyses, could potentially bias GEBV. In our study of 526 clonally replicated progeny, only 149 had genotypes available, and we hypothesized that the contribution of these 149 clones to the estimates of EBV would be relatively small. In order to test this hypothesis, we first left out the 149 genotyped clones in the BLUP model used to estimate EBV and obtained GEBVs. Then, we included those genotyped clones in estimation of EBV to obtain GEBVs. The correlation coefficient between two different GEBV of those clones was 0.99. This suggests that, in this study, using all the trees to obtain EBV as new phenotypes to train marker models does not introduce bias in the predictive potential of the markers.

A third model that combined the polygenic effect plus the additive and dominant marker effects was also tested, and the prediction accuracies for all four traits were compared with those of the marker-only and pedigree-only models. No improvements of the accuracies from this full model were found. We suggest that this lack of increase in predictive power of the combined model could be due to collinearity effects between independent variables, where the marker information is highly correlated with the pedigree information, as previously reported (Legarra et al. 2008; Jannink et al. 2010), and we conclude that the markers are simply providing the same information as the pedigree.

In the context of a forest tree breeding program, there are additional validation methods to explore that might be relevant to specific types of applications. For example, the training population could include phenotypes and genotypes from a subset of sites in a set of multi-environment progeny tests, and the validation population could be drawn from the remaining sites. The results would assess the efficiency of GS for a particular growing environment and address the question of the magnitude of genotype by environment interaction, which was proposed to limit the application of molecular markers in forest tree breeding due to the need of specific predictions for any particular environment (Strauss et al. 1992). In a recent study on loblolly pine, Resende et al. (2011) observed decreasing accuracies as measurements at younger ages were used for validation. In the same study, they reported a reduction in accuracy of GEBV across sites, using data from one site for training the model and data from another site as validation population, with greater reduction in accuracy for sites that are geographically more distant.

Association testing and GS

The decision to use fewer markers in prediction of breeding values has an important implication in the cost of genotyping. SNP genotyping is currently a major expense in a marker-assisted breeding program, and our results suggest that using subsets of markers could be as efficient as using all available markers to estimate GEBV, which also would impact GS costs if implemented in operational breeding programs. The cost was not included in our analysis, but there are tools to economically assess the efficiency for selection using MAS or GS approaches as described by Strauss et al. (1992).

Interestingly, we found that using smaller subset of markers, whether selected based on association testing with phenotype through independent *F*-tests or chosen at random, did not decrease accuracy values considerably. This could be due to the closed highly pedigreed population we used for validation. A more diverse population would be a better scenario to test the relationship between fewer markers and their predictive ability.

Efficiency of GS

The results of this study suggest that efficiency of GS per year would be greater than that of traditional progeny testing of selection for lignin and cellulose content. The cost-effectiveness of marker applications in tree breeding increases as the length of the breeding cycle is reduced. Previous reports (Legarra et al. 2008; Cumbie 2010) used the breeder's equation in combination with the correlated response predicted for genetic values, based on estimates obtained either from markers or phenotypes. We also used this approach and included time as a factor in the comparison, assuming a clonal testing cycle of 15 years without the use of markers. This time was estimated based on starting with seed from elite families, and assuming it would take 9 years to produce rooted cuttings, plant and evaluate clonal field trials, and allowing 1 year for lignin/cellulose determinations and 5 years for breeding of selected individuals and collection of the next generation of seeds. Under GS, different approaches could be used to reduce time required for a breeding cycle. One realistic scenario would involve a generation interval that starts with collecting seed from elite crosses, followed by two and half years to grow the seedlings, determine SNP genotypes of selection candidates, and select clones to be topgrafted to stimulate flowering, and 5 years to breed and obtain seeds, for a total 7.5 years or a ratio of 0.5 for a reduced cycle. We explored more reduced time scenarios, but they would become more difficult to achieve unless new breeding techniques are developed and integrated into GS system (e.g., a combination of DNA extraction at somatic embryogenesis stage and accelerated flowering and cone production in the topgrafting). The efficiency per year of slightly less than one achieved with smaller subsets of markers for selection on lignin and cellulose is still promising in terms of the relative efficiency of GS over traditional breeding and clonal testing. This would make genotyping more cost-effective and reduce the challenges of managing and interpreting large amounts of SNP data in operational tree breeding programs.

Conclusions

This study explores the utility of GS methods in a forest tree species using empirical marker genotype and phenotype data. Cross-validation methods demonstrated the utility of genomic selection in this small pine breeding population. We speculate that the observed accuracies in this study are mainly the results of markers tracing the LD due to limited effective population size in this small cloned pine population. Prediction accuracies of models that use only a subset of markers were generally comparable with the accuracies of the models using all markers, and the subset of markers used do not need to be associated with the phenotype to be effective. For lignin and cellulose content, a GS scenario is efficient under different relative lengths of the breeding cycle, which would allow cost-effective applications of GS in a tree breeding program. The results in this study should be interpreted cautiously, because of the small size of the genotyped training population, and should be verified with larger training populations. We envision that GS implementation in forest tree breeding during the next generations will yield accurate breeding values for candidate selections, overcoming the unpredictable mapping and positioning of QTL experiments. In the near future, GS will likely improve the efficiency of forest tree breeding by reducing the need for expensive field testing, shortening the breeding cycle, and achieving more realized genetic gain per unit time and cost.

Acknowledgments The authors thank Plum Creek Timber Company, Inc., and CellFor, Inc., for the field measurement data; David Barker and Josh Steiger (NCSU Cooperative Tree Improvement Program) for collection of wood cores, preparation of samples, and collection of NIR spectra; and Dr. Gary Hodge (Camcore) for prediction of lignin and cellulose content. This work was supported by the Conifer Translational Genomics Network Coordinated Agricultural Project (USDA NRI #2007-02781 and AFRI #2009-01879), the NC State University Cooperative Tree Improvement Program, and the Department of Forestry and Environmental Resources at NCSU.

References

Bettinger P, Clutter M, Siry J, Kane M, Pait J (2009) Broad implications of southern United States pine clonal forestry on planning and management of forests. Int For Rev 11(3):331–345

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

Brown GR, Bassoni DL, Gill GP, Fontana JR, Wheeler NC, Megraw RA, Davis MF, Sewell MM, Tuskan GA, Neale DB (2003) Identification of quantitative trait loci influencing wood property traits in loblolly pine (*Pinus taeda* L.). III. QTL verification and candidate gene mapping. Genetics 164:1537–1546

Brown GR, Gill GP, Kuntz RJ, Langley CH, Neale DB (2004) Nucleotide diversity and linkage disequilibrium in loblolly pine. PNAS 101:15255–15260

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. J Anim Breed Genet 124:362–368

Cumbie WP (2010) Association genetics for growth, carbon isotope discrimination, and stem quality in loblolly pine. Dissertation, North Carolina State University

Cumbie WP, Eckert A, Wegrzyn J, Whetten R, Neale D, Goldfarb B (2011) Association genetics of carbon isotope discrimination, height and foliar nitrogen in a natural population of *Pinus taeda* L. Heredity. doi:10.1038/hdy.2010.168

Eckert AJ, van Heerwaarden J, Wegrzyn JL, Nelson CD, Ross-Ibarra J, González-Martínez SC, Neale DB (2010) Patterns of population structure and environmental associations to aridity across the range of loblolly pine (*Pinus taeda* L., Pinaceae). Genetics 185 (3):969–982

Falconer DS, Mackay TFC (1996) Introduction to quantitative genetics. Fourth edition. Longman Group, Ltd, Essex, p 464

Garrick DJ, Taylor JF, Fernando RL (2009) Deregressing estimated breeding values and weighting information for genomic regression analyses. Genet Sel Evol 41:55. doi:10.1186/1297-9686-41-55

Gilmour AR, Gogel BJ, Cullis BR, Thompson R (2009) ASReml User. Guide release 3.0. VSN International Ltd., Hemel Hempstead, HP1 1ES, United Kingdom

Goddard ME, Hayes BJ (2009) Mapping genes for complex traits in domesticated animals and their use inbreeding programs. Nat Rev Genet 10:381–391

Goebel NB, Warner JR (1966) Total and bark volume tables for small diameter Loblolly, Shortleaf, and Virginia Pine in the upper South Carolina Piedmont. Forest Research Series No. 9, Clemson University

González-Martínez SC, Wheeler NC, Ersoz E, Nelson CD, Neale DB (2007) Association genetics in *Pinus taeda* L. I. Wood property traits. Genetics 175:399–499

González-Martínez SC, Huber D, Ersoz E, Davis JM, Neale DB (2008) Association genetics in *Pinus taeda* L. II. Carbon isotope discrimination. Heredity 101(1):19–26

Grattapaglia D, Resende MDV (2010) Genomic selection in forest tree breeding. Tree Genet Genome 7(2):241–255

Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME (2009) Invited review: genomic selection in dairy cattle: progress and challenges. J Dairy Sci 92(2):433–443

Hodge GR, Woodbridge WC (2010) Global near infrared models to predict lignin and cellulose content of pine wood. J Near Infrared Spectrosc 18:367–380

Isik F, Li B, Frampton J (2003) Estimates of additive, dominance and epistatic genetic variances from a clonally replicated test of loblolly pine. For Sci 49(1):77–88

Isik F, Boos DD, Li B (2005) The distribution of genetic parameter estimates and confidence intervals from small disconnected diallels. Theor Appl Genet 110:1236–1243

Jannink JL, Lorenz AJ, Iwata H (2010) Genomic selection in plant breeding: from theory to practice. Brief Funct Genom 9(2):166–177

Kaya Z, Neale DB, Sewell MM (1999) Identification of quantitative trait loci influencing annual height- and diameter-increment growth in loblolly pine (*Pinus taeda* L.). Theor Appl Genet 98 (3/4):586–592

Knott SA, Neale DB, Sewell MM, Haley CS (1997) Multiple marker mapping of quantitative trait loci in an outbred pedigree of loblolly pine. Theor Appl Genet 94(6–7):810–820

Lande R, Thompson R (1990) Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics 124:743–756

Legarra A, Misztal I (2008) Technical note: computing strategies in genome-wide selection. J Dairy Sci 91:360–366

Legarra A, Robert-Granie C, Manfredi E, Elsen JM (2008) Performance of genomic selection in mice. Genetics 180:611–618

Liu Z, Seefried FR, Reinhardt F, Rensing S, Thaller G, Reents R (2011) Impacts of both reference population size and inclusion of a residual polygenic effect on the accuracy of genomic prediction. Genet Sel Evol 43:19

Mackay TFC, Stone EA, Ayroles JF (2009) The genetics of quantitative traits: challenges and prospects. Nat Rev Genet 10:565–577

McKeand SE, Jokela EJ, Huber DA, Byram TD, Allen HL, Li B, Mullin TJ (2006) Performance of improved genotypes of loblolly pine across different soils, climates, and silvicultural inputs. For Ecol Manag 227:178–184

Meuwissen THE (2009) Accuracy of breeding values of unrelated individuals predicted by dense SNP genotyping. Genet Sel Evol 41:35

Meuwissen THE, Goddard ME (2010) Accurate prediction of genetic values for complex traits by whole-genome resequencing. Genetics 185:623–631

Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome wide dense marker maps. Genetics 157:1819–1829

Moser G, Khatkar MS, Hayes BJ, Raadsma HW (2010) Accuracy of direct genomic values in Holstein bulls and cows using subsets of SNP markers. Genet Sel Evol 42:37–52

Mrode RA (2005) Linear models for the prediction of animal breeding values. CAB International, Wallingford

Neale DB (2007) Genomics to tree breeding and forest health. Curr Opin Genet Dev 17:539–544

Neale DB, Savolainen O (2004) Association genetics of complex traits in conifers. Trends Plant Sci 9:325–330

Neale DB, Williams CG (1991) Restriction fragment length polymorphism mapping in conifers and applications to forest genetics and tree improvement. Can J For Res 21:545–554

Neale DB, Sewell MM, Brown GR (2002) Molecular dissection of the inheritance of wood property traits in loblolly pine. Ann For Sci 59:595–605

Piepho HP (2009) Ridge regression and extensions for genomewide selection in maize. Crop Sci 49:1165–1176

Plummer M, Best N, Cowles K, Vines K (2006) CODA: convergence diagnosis and output analysis for MCMC. R News 6(1):7–11

Pyhäjärvi T, García-Gil MR, Knürr T, Mikkonen M, Wachowiak W, Savolainen O (2007) Demographic history has influenced nucleotide diversity in European *Pinus sylvestris* populations. Genetics 177:1713–1724

Quesada T, Gopal V, Cumbie WP, Eckert AJ, Wegrzyn JL, Neale DB, Goldfarb B, Huber DA, Casella G, Davis JM (2010) Association mapping of quantitative disease resistance in a natural population of loblolly pine (*Pinus taeda* L.). Genetics 186(2):677–686

Resende MFR, Muñoz P, Acosta JJ, Peter GF, Davis JM, Grattapaglia D, Resende MDV, Kirst M (2011) Accelerating the domestication of trees using genomic selection: accuracy of prediction models across ages and environments. New Phytol. doi:10.1111/j.1469-8137.2011.03895.x

SAS Institute Inc (2010) SAS online doc 9.2. SAS Institute Inc, Cary, pp 2002–2005

Solberg TR, Sonesson AK, Woolliams J, Meuwissen THE (2008) Genomic selection using different marker types and densities. J Anim Sci 86:2447–2454

Strauss SH, Lande R, Namkoong G (1992) Limitations of molecular-marker-aided selection in forest tree breeding. Can J For Res 22:1050–1061

Sykes R, Li B, Isik F, Kadla J, Chang HM (2006) Genetic variation and genotype by environment interactions of juvenile wood chemical properties in *Pinus taeda* L. Ann For Sci 63:897–904

VanRaden PM, Van Tassell CP, Wiggans GR, Sonstegard TS, Schnabel RD, Taylor JF, Schenkel F (2009) Invited review: reliability of genomic predictions for North American Holstein bulls. J Dairy Sci 92:16–24

Wegrzyn JL, Lee JM, Tearse BR, Neale DB (2008) TreeGenes: a forest tree genome database. Int J Plant Genom. doi:10.1155/2008/412875

White TL, Adams WT, Neale DB (2007) Forest genetics. CABI Publishing CAB International, Cambridge

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, Goddard ME, Visscher PM (2010) Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42 (7):565–569