MAXIMIZING THE USE OF MOLECULAR MARKERS IN PINE BREEDING IN THE CONTEXT OF GENOMIC SELECTION

By

PATRICIO R. MUNOZ DEL VALLE

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2012

© 2012 Patricio Munoz Del Valle

To my wife for all her support in this long and rough road

ACKNOWLEDGMENTS

I would like to thank my committee for all their support in these years. Special thanks to Dr. Gary Peter because he always supports me and found the time in his busy agenda to discuss important topics of my research as well as broader areas. Thanks to Dr. Matias Kirst for his support, time and constant push toward a better science. Thanks to Dr. John Davis for helping me improve considerably my understanding on genetics of plant diseases. Thanks to Dr. James Olmstead for his support and for give me a chance to interact with students from the other side of the classroom. Thanks to Dr. George Casella, who originally was in my committee, he helps me understanding the number and formulas when they were too entangled for me. I have long conversation and discussion of several important and daily-trivial topics with all of you that in different ways help me to improve my holistic view of research and science.

I would like to thank my family, especially to my wife (Claudia), my daughter (Antonia) and son (Ignacio) for their love, support in all times, understand me in difficult times and for disconnecting me from demanding research. Thanks to my parent, brother and sisters, for their love and for helping me to reach my goals from the beginning.

Thanks to my friends and colleagues, especially Thomas and Charlotte for their support and friendship that I am sure will last until the end of our days. Thanks to Dr. Salvador Gezan and Marcio Resende for their friendship, constant support, great discussions and helping me increase my knowledge in different areas of programming, statistics and more. To the Chilean community (families: Gonzalez-Costagliola, Riveros-Chavez, Garcia-Villaseñor, Forcael-Salgado, Rodriguez-Hananias and Gladys, Andres, Francisco, Fernando, Rodrigo) for their friendship and support. Thanks to my

colleagues Greg Powell, Eliana Kampf and Chris Dervinis, for their constant support and their important role on logistics in their respective areas.

Finally, I would like to thanks to all the members of the Forest Genomic Lab, the Plant Molecular and Cellular Biology program (PMCB), the Plant Molecular Breeding Initiative (PMBI), the Cooperative of Forest Genetic Research Program (CFGRP) and the Forest Biology Research Cooperative (FBRC) for their support in different aspect of my research.

TABLE OF CONTENTS

page

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AIC	Akaike information criteria
BA	Average branch angle
BD	Average branch diameter
BLC	Basal height of the live crown
BLUP-BV	Traditional best linear unbiased predicted breeding value
BV	Breeding values
CWAC	Crown width across the planting beds
CWAL	Crown width along the planting beds
DBH	Stem diameter at chest height
GS	Genomic Selection
GS-BV	Genomic selection predicted breeding value
HT	Total stem height
RRM	Realized relationship matrix
SNP	Single nucleotide polymorphism
SSR	Simple-sequence-repeat

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

MAXIMIZING THE USE OF MOLECULAR MARKERS IN PINE BREEDING IN THE CONTEXT OF GENOMIC SELECTION

By

Patricio Munoz Del Valle

December 2012

Chair: Gary F. Peter
Major: Plant Molecular and Cellular Biology

By 2030 demand for renewable energy, food and fiber is expected to double. To sustainably meet this increase in demand from the current land base, plant breeders need to develop higher yielding crops that require fewer inputs and better resist diseases and environmental change. Of particular importance is accelerating improvement in quantitative traits (QT), which show complex patterns of inheritance. Genomic selection (GS) provides an approach where molecular markers can be used directly in breeding programs regardless of the genetic architecture. While most GS studies have concentrated on prediction of breeding values, here this approach is extended to include non-additive variation and to maximize the use of molecular markers (SNPs) in pine breeding.

With a relatively dense panel of SNPs, a method to detect and correct errors in the pedigree information is presented, based on a marker-derived additive relationship matrix. The impact of pedigree errors on genetic parameter estimates and breeding value prediction is demonstrated. In addition, the performance of four published analytical methods for GS that differ in assumptions regarding the distribution of markers additive-effect is presented. Methods include: ridge regression–best linear

unbiased prediction (RR–BLUP), Bayes A, Bayes Cpi, and Bayesian LASSO. Furthermore, a modified RR–BLUP (RR-BLUP_B) that utilizes a selected subset of markers was developed and evaluated. All five methods for GS were evaluated for seventeen different traits of importance in pine breeding and with different predicted genetic architecture and heritabilities. While for QT no significant difference among methods was detected, for traits controlled by fewer genes, Bayes Cpi and RR-BLUP_B performed significantly better. Finally, the use of a dense panel of SNPs to partition the genetic variance into additive and non-additive components was evaluated. For tree height, use of the SNP-derived relationship matrices (additive and non-additive) in a statistical model including additive and non-additive effects performed best, not only to partition the genetic variances but also to improve considerably the breeding value prediction ability in trend, magnitude and top individual selection. This study indicates that markers can be used beyond prediction of additive effects, positively impacting the genetic gain of the breeding program.

CHAPTER 1
INTRODUCTION

Our rapidly expanding population and their desire for a better standard of living are quickly increasing demand for renewable energy, food and fiber (FAO 2002). At the same time, less land is available for production due to increased urbanization, climate variability and environmental change (IPCC 2007). To meet these increases in demand, with less land, plant breeders need to accelerate development of higher yielding crops that require fewer inputs and better resist diseases and environmental change (Collard and Mackill 2008). To help meet these challenges, molecular markers offered the promise of accelerating genetic improvement (Stuber *et al.* 1982, Soller 1978). Unfortunately marker assisted selection (MAS) only worked in a few cases where traits were controlled by a small number of genes (Lande and Thompsom 1990; Dekkers 2004; Jannink *et al.* 2010). However, most economically important traits in animal and plant breeding programs show complex inheritance and are quantitative traits (Buckler *et al.* 2009; Goddard 2009c; Hayes and Goddard 2010).

Quantitative traits are controlled by a large number of genes, each with a small effect, only detectable using large populations (Visscher 2008). For quantitative traits selecting for positive alleles in a few of the hundreds of genes affecting a trait has limited applicability in breeding (Bernardo 2008). To overcome this limitation, a new methodology known as genomic selection (GS) was proposed. GS uses all molecular markers simultaneously (as a random effect) to make predictions in the breeding context (Meuwissen *et al.* 2001). In GS, breeding values (BV) obtained from traditional Best Linear Unbiased Prediction (BLUP-BV) analysis of phenotypic data from a training population are used to develop GS prediction models based on genotypic (molecular

markers) information. GS models predict the genomic BVs (GS-BV) for a validation population, which are then correlated with known traditional BLUP-BVs to estimate the accuracy of the models (Goddard and Hayes 2009). The overall idea of GS is to select genotypes in advanced generations based only on the GS-BV predicted from the molecular marker information using models previously constructed and validated with the parental population (Hayes and Goddard 2010). From the statistical point of view, GS has advantages over other methods (e.g. least square means), because it accounts for the co-linearity of molecular markers since all co-variables (markers) come from the same individual. Also, in GS all markers are retained for prediction, enabling the capture of more variation due to small-effect quantitative trait loci (QTL), and undetected causal loci (Hamblin *et al.* 2011). GS can be used for any trait, independent of its genetic architecture, thus predictions using GS can be performed for quantitative traits while MAS is limited to traits with few genes of large effect (Jannink *et al.* 2010). From the point of view of breeding, GS provides higher accuracy for prediction and shorter rotations than current phenotypic selection, which translates into increased genetic gain per unit time (Resende *et al.* 2012a).

The greatest theoretical and practical advances in GS have occurred with cattle breeding (Hayes 2009b) aimed at decreasing the cost of testing that can reach as much as $50,000 per bull (Schaeffer 2006). Bulls are selected by their additive genetic value (General combining ability, GCA), ignoring non-additive effects. In this respect, GS methods and models developed for cattle are completely equivalent to traditional animal breeding as they both only consider additive effects. Consequently, commercial dairy cattle breeding programs currently use additive GS models operationally (Wiggans *et al.*

2011). However, plant-breeding programs differ substantially from cattle breeding. Plant breeders use mating designs with complex pedigrees, and often exploit non-additive effects through full-sib families or cloning of progeny for deployment (White *et al.* 2007).

In the case of conifers, breeding is a long, complex multi-step process (White and Carson 2004). Since the mid 1950's, only three cycles of breeding and selection have been completed with the most advanced programs only in their fourth cycle of improvement (Neale and Kremer 2011). During this time, a number of advances have been made to accelerate pine breeding and improve genetic gain, the most important being the use of top grafting of immature scion into sexually mature trees (Bramlett 1997), and early selection for growth traits (Lambeth 1980). Even with these improvements, pine breeding and selection cycles still span more than a decade. Thus, in tree breeding the general objective of implementing GS is to shorten the breeding cycle, thereby increasing the gain per unit time. Recently the potential for dramatic increases in genetic gain per cycle with GS have been shown (Wong and Bernardo 2008; Iwata *et al.* 2011; Resende *et al*. 2012a) considering only additive effects. The reality is that tree breeders exploit additive as well as non-additive effects (White *el at.* 2007). A single parent can be crossed to several other individuals (as male and female), full-sib progeny from these crosses are tested as seedlings and/or clonally propagated cuttings, and families are then selected by their deviation relative to the average of the two parents (Specific Combining Ability). In addition, these crosses are performed under mating designs that create complex relationships (pedigrees) in a few generations of improvement (White and Carson 2004).

Given the important differences between tree and cattle breeding, development of GS models that are completely equivalent to traditional tree improvement scenarios are a priority. Additionally, considering the potential for success of GS in plant breeding and the rapidly decreasing cost of molecular markers, especially single nucleotide polymorphism (SNP), the use of molecular markers and GS likely will soon become the norm in all breeding programs. This implies an important initial investment to obtain SNP markers in a significant number of individuals (Collard and Mackill 2008). Although, molecular markers (SNPs) as presented above are used exclusively for prediction of BV, they could be used in other beneficial ways for the breeding program. The objective of this research is to present how to maximize the use of the molecular marker in pine breeding programs by pedigree corrections, GS analytic method selection and by exploring the trait genetic architecture.

The length of the pine breeding cycle not only slows progress but also increases the chance for errors because numerous people are involved in the many steps that need to occur over the 12-15 years it takes to complete one breeding cycle. Errors in the population pedigree decrease the accuracy of genetic predictions, which reduces the genetic gains that can be achieved even with traditional phenotypic selection (Ericsson 1999; Banos *et al.* 2001; Sanders *et al.* 2006). Recently, we concluded that the breeding cycle could be shortened by using GS with additive effects (Resende *et al.* 2012a); however, errors in the pedigree will still reduce accuracies. If molecular markers are developed with the purpose of using genomic selection for additive effects, then it should be possible to use them to first correct the pedigree to obtain less biased BV and GS-BV predictions. Thus, the second chapter provides a method to detect and correct

errors in the pedigree information using a dense panel of markers that posteriorly will be used for GS. Although the pedigree can be corrected with a low density of markers, such as microsatellites (SSRs), if GS is planned, then using the dense panel of markers needed for GS to first correct the pedigree will improve prediction model construction that in the practice will benefit the breeding program.

Genomic selection is expected to be particularly valuable for traits that are costly to phenotype and expressed late in the life cycle of long-lived species, such as pines (Resende *et al.* 2012a). Alternative approaches to genomic selection prediction models may perform differently for traits with distinct genetic properties (de los Campos *et al.* 2009a; Habier *et al.* 2011; Meuwissen *et al.* 2001). In the third chapter the performance of four different original analytical methods of genomic selection: (i) ridge regression–best linear unbiased prediction (RR–BLUP), (ii) Bayes A, (iii) Bayes Cpi, and (iv) Bayesian LASSO, that differ with respect to assumptions regarding distribution of marker effects, are evaluated and compared. In addition, a modified RR–BLUP (RR–BLUP B) that utilizes a selected subset of markers was evaluated. Proof of concept was provided in Resende et al. (2012a), which evaluated GS for yield traits. This chapter extends the analysis to seventeen different traits of importance in pine breeding and with different predicted genetic architecture and heritabilities. The results show that correct selection of the methodology should be based on the genetic architecture of the trait.

In the Southeastern US, pine plantations exploit both additive and non-additive effects as parental crosses are selected for family or clonal deployment (White *et al.* 2007). Incorporating non-additive effects in GS has had less attention because it cannot

be exploited in cattle breeding where the most of the advances for GS have been made. Thus, for forest tree improvement it is important to explore how to incorporate non-additive effects in the GS models, understanding the contribution to genetic variance, and finally make GS fully comparable to the current methods of tree improvement. The fourth chapter uses a dense panel of molecular markers in a novel way to improve partitioning of the genetic variance into additive and non-additive components. Unbiased estimation of the variance components is essential for breeding, as bias affects BV prediction, which can be detrimental to the breeding program. Here we evaluated the potential of incorporating non-additive effects into GS, which is highly relevant for tree breeding programs that typically exploit some portion of non-additive effects by deployment of families or clonal material. In these cases, the use of GS with only additive effects does not fit with the necessities of the breeding program, by not exploiting non-additive variation to improve genetic gain.

CHAPTER 2
DENSE PANEL OF MARKERS FOR CORRECTING PEDIGREE ERRORS IN BREEDING POPULATIONS: IMPACT ON HERITABILITY, BREEDING VALUE AND GENOMIC SELECTION ACCURACY[1]

Background

A central goal of quantitative genetics is to estimate the level of genetic control and genetic correlation amongst complex traits. This information is used in breeding, for the selection of elite parents, families and individuals, as well as for subsequent generations of genetic improvement. Genetic tests are designed to provide phenotypic information for estimation of parameters such as variance components, heritability, genetic correlations and breeding values. Breeding values (BV) are typically estimated with best linear unbiased prediction (BLUP-BV), based on the theory of resemblance between relatives due to genetic factors (Lynch and Walsh 1998), commonly derived from the pedigree (Mrode 2005). Consequently, better estimates of genetic parameters are obtained when the pedigree information is accurate. Pedigree errors are common, averaging 10% in cattle and tree breeding populations (Banos *et al.* 2001; Visscher *et al.* 2002; Doerksen and Herbinger 2010). The presence of such errors can lead to incorrect estimates of the additive variance, causing a decrease in the BLUP-BV prediction accuracy (Ericsson 1999; Banos *et al.* 2001; Sanders *et al.* 2006). In traditional phenotypic selection, a decrease in BV accuracy has been shown to reduce genetic gains by 4.3% to 17% (Geldermann *et al.* 1986; Israel and Weller 2000).

Dense panels of molecular markers can be used to empirically estimate the actual relationships between relatives through the construction of a realized relationship matrix (RRM) (Powell *et al.* 2010) and provide precise estimates of the proportion of the

[1] Chapter submitted to Genomic Selection Evolution Journal

genome that is shared among individuals. Molecular markers have been used to correct errors in the pedigree using different strategies; most of them rely on parent-progeny genotype data (Bennewitz *et al.* 2002; Wiggans *et al.* 2010; Hayes 2011) or in the diagonal of the RRM matrix (Simeone *et al.* 2011). If a dense panel of markers is used in breeding populations with a complex pedigree, the RRM values among individuals are normally distributed around the expectation for a given class (i.e. expectation [unrelated]=0.0) (Yang *et al.* 2010; Simeone *et al.* 2011). Thus, the current progeny population RRM diagonal and off-diagonal elements can be used to correct pedigree errors. This corrected pedigree should increase the accuracy of the BLUP-BV predictions.

Increasing the accuracy of BLUP-BV not only improves gains from traditional phenotypic selection, but should also improve the accuracy of genomic selection models. Genomic selection (GS) models are developed to predict BV using only information from estimated marker effects (Meuwissen *et al.* 2001). Typically, the inputs for constructing GS prediction models are phenotypes deregressed from the BLUP-BV (Garrick *et al.* 2009). The models are then tested in a validation population to obtain GS predicted BVs (GS-BV) and estimate the accuracy of genomic prediction (Goddard *et al.* 2009). The utility of GS in plant and animal breeding depends on the accuracy of the GS models developed to predict BV (Goddard and Hayes 2009; Habier *et al.* 2010; Jannink *et al.* 2010; Grattapaglia and Resende 2011; Heffner *et al.* 2010). Recently, a number of analytical approaches (Gianola *et al.* 2006; de los Campos *et al.* 2009a; Habier *et al.* 2011; Legarra *et al.* 2011b) have been developed to study factors that contribute to GS accuracy (Habier *et al.* 2009; Habier *et al.* 2010; Iwata and Jannink

2011) and to increase GS accuracy relative to the original approaches proposed by Meuwissen *et al.* (2001). Higher accuracy and less bias in the estimated BLUP-BVs are expected to improve the accuracy of all GS models. However, the effect of correcting pedigree errors on BLUP-BVs used to develop GS-BV prediction models has not been assessed.

Here we report for a loblolly pine breeding population, the effect of pedigree correction based on construction of a realized relationship matrix from a dense panel of genetic markers. The original and corrected pedigrees were used to generate BLUP-BVs and posteriorly GS models using ridge regression BLUP. The accuracies of the uncorrected and corrected pedigrees on BLUP-BV and GS-BV were compared.

Methods

Data

Phenotypic and genotypic data were collected from one field test located in Nassau (Florida, USA) containing 956 clonally propagated loblolly pine trees (~8 ramets per genotype) of a genetic test design with 61 families derived from 32 parents crossed in a circular mating design (details in Baltunis *et al.* 2005). The field site was established using single-tree plots in eight replicates (one ramet in each replicate), utilizing a resolvable alpha incomplete block design (Williams *et al.* 2002). Two silvicultural treatments were applied; four replicates were grown under high intensity and four replicates under operational culture.

Phenotype measurements were taken for basal height of the live crown (BLC, cm), crown width across the planting beds (CWAC, cm), crown width along the planting beds (CWAL, cm), stem diameter (DBH, cm) and total stem height (HT, cm), as described in Baltunis *et al.* (2007a) and Resende *et al.*(2012). The traits branch angle average (BA,

degrees) and average branch diameter (BD, cm) were measured only in the high intensity silvicultural treatment. The age for each measurement is listed in Table 2-1, together with the trait-age combination used hereafter.

Genomic DNA was extracted from needle tissue using the QIAGEN DNeasy Plant Kit, and quantified with a NanoDrop microvolume spectrophotometer. One microgram of DNA from each clone was genotyped using an Illumina Infinium™ assay (Illumina, San Diego, CA) designed detect 7,216 SNPs that were identified through the resequencing of 7,535 uniquely expressed sequence tag (EST) contigs in 18 loblolly pine haploid megagametophytes (Eckert *et al.* 2010). After filtering for monomorphic markers a total of 4,825 SNPs were selected for analysis.

Realized Relationship Matrix and Pedigree Corrections

A total of 2,182 SNP markers, with a minor allele frequency greater than 0.12, were used to construct the RRM. The RRM was estimated by determining identity by state coefficients relative to the parents of the current population as the base population (Powell *et al.* 2010). Relatedness estimates were adjusted for sampling error and shrunk toward the expected values to lessen error as recommended by Yang *et al.* (2010). Using relationships estimated in the RRM, the pedigree was corrected based on the normality of the distribution of the relationship coefficients around their expected values (i.e. 0.5 for full-sib). First, the RRM matrix was paired with the A matrix. Second, duplicated individuals (different label, but same genotype) were identified, and the ones with the fewer missing values were kept. Third, the relationship coefficient limits for the full-sib and half-sib classes were defined based on the normal distribution using all relationships in each class. Fourth, individual or groups of individuals not matching the expected pattern were identified. Fifth, conflictive individuals were re-assigned to a new

pedigree by searching across all relationships in the dataset for the parent or family where these individuals match the expectation. In the last step, an individual was re-assigned to a new parent/family only if the conflictive individual matched the expectation, given by the defined boundaries, with all individuals from that parent/family. Once the new parent/family was identified, the individuals were re-labeled generating the corrected pedigree. This process was iterative, as every time the pedigree of an individual was corrected the relationship class distributions changed across the database and were recalculated.

BLUP/REML Analysis: Variance Component Estimation and Breeding Values Prediction

To investigate the effects of BLUP-BV predictions on GS, two alternative linear mixed models were fitted independently using ASReml v.3.0 (Gilmour *et al.* 2009) for each trait. Accuracy for all BLUP analyses was estimated based on the prediction error variance for each clone separately (Mrode 2005) and the average was reported.

Original pedigree BLUP (Ori-BLUP)

This model assumes no errors in the original pedigree

$$y = Xb + Z_1 i + Z_2 a + Z_3 f + Z_4 n + Z_5 d_1 + Z_6 d_2 + Z_7 d_3 + e \quad (1)$$

where *y* is the measure of the trait being analyzed (see above), *b* is a vector of fixed effects (i.e. culture type and replication within culture type), *i* is a vector of random incomplete block effect within replication ~N(0, $I\sigma^2_{iblk}$), *a* is a vector of random additive effects of clones ~N(0, $A\sigma^2_a$), *f* is a vector of random family effect ~N(0, $I\sigma^2_f$), *n* is a vector of random non-additive effects of clones ~N(0, $I\sigma^2_n$), d_1 is a vector of random additive by culture type interaction ~N(0, $DIAG\sigma^2_{d1}$), d_2 is a vector of random family by

culture type interaction ~N(0, DIAGσ^2_{d2}), *d*$_3$ is a vector of random non-additive by culture type interaction ~N(0, DIAGσ^2_{d3}), *e* is the random residual effect ~N(0, DIAGσ^2_e) as one specific error for each treatment was fitted, X and Z_1-Z_7 are incidence matrices and I, A and DIAG are the identity, numerator relationship and block diagonal matrices respectively.

Corrected pedigree BLUP (Corr-BLUP)

This model assumes that the original pedigree contains errors that were corrected using the relationships derived from the RRM and implemented for analysis in the corrected version of the pedigree. Therefore, in this analysis a corrected version of the A matrix will be used (A_{cor}). This analysis uses the same model described above (equation 1), although in this case *a* is a vector of random additive effects of clones ~N(0, $A_{cor}\sigma^2_a$).

Genomic Selection and Validation

The breeding value estimated in each of the above models was deregressed and the parental average of each family removed (Garrick *et al.* 2009) for genomic selection analysis. The deregressed phenotype derived by using the original and the corrected pedigree was used as input for a ridge-regression BLUP with the 4,825 markers used as covariates as described previously (Resende *et al.* 2012b). Each analysis was repeated 10 times in a cross validation scheme (Kohavi 1995). The predictive ability of each model was estimated as the correlation between the genomic selection predicted breeding values (GS-BVs) and the deregressed phenotype that were used as input in the generation of the GS-BVs.

Results

Pedigree Correction

The relationship coefficients derived from the molecular markers is expected to be a normal distribution centered at 0.5 and 0.25 for full- and half-sib families, respectively. The distribution with the original pedigree was bi-modal and asymmetrical for half-sibs, with a large frequency closer to the expected 0.25 value and a second peak close to zero (Figure 2-1 top-left panel). For the full-sib class a tri-modal asymmetrical distribution was observed, the highest peak (mode) around the corrected 0.5 expectation value, with the second and third peaks around the 0.25 and a zero relationship, respectively (Figure 2-1 bottom-left panel). In the original pedigree, before corrections, the most frequent relationship found in the dataset yielded biased average relationship coefficients (Table 2-2), with unrelated, half-sibs, and full-sibs individuals being underestimated, and the diagonal being slightly overestimated. The standard deviations for full-sib and half-sib individuals were the largest (Table 2-2). However, correcting the pedigree gave mean values that agreed with the expectations for the given classes, causing a 27 to 67% decrease in the standard deviation (Table 2-2, Figure 2-1 right panels).

By using the RRM, different types of pedigree errors were detected and corrected, including duplicated genotypes (clones) with different labels, from which only one was kept. Individuals with either one or both incorrect parents (sixty-nine in total), were reassigned to the correct parent using the coefficients from the RRM. Eleven new parents, one female and ten male, were added, as they did not exist in the pedigree records. Parents of four complete families and two grandparents were reassigned.

Finally, three individuals were removed because they yielded inconsistent relationships across the pedigree (Table 2-3).

Estimation of Breeding Values with Original and Corrected Pedigree Relationship Matrices

Both the original and corrected pedigrees were used to estimate BVs from a traditional BLUP analysis. Heritability estimates derived from the REML/BLUP analysis using the corrected pedigree decreased slightly for eight of the traits (maximum decrease of 5%), when compared with the REML/BLUP analysis using the original pedigree. In seven traits the heritability increased by a maximum of 21% for BD_6 (Table 2-4). Overall, with the corrected version of the pedigree, breeding value accuracy decreased slightly in only four traits, with a maximum reduction of 0.94% for BA_6. The BV accuracy increased for 11 traits (maximum increase of 5.8% for HT_6). Importantly, in all but one trait (BLC_6) the models with the corrected pedigree fit the data substantially better, measured by the Akaike Information Criteria (AIC, Table 2-4).

Accuracy of Genomic Selection Predictive Models with Original and Corrected Pedigrees

Breeding values obtained from BLUP analyses with the original and corrected pedigrees were posteriorly deregressed and used as response variables to generate genomic selection models for the 15 trait-age combinations. The predictive ability of the models increased for 13 of the 15 traits when the corrected version of the pedigree was used (Figure 2-2). The two traits that decreased with the corrected pedigree (BA_6 and CWAC_6) were reduced by 1.1 and 2.3%, respectively; whereas the predictive ability of the remaining 13 traits increased from 1 to 15% with an average of 7.2%.

Discussion

Pedigree Correction

Genetic improvement of trees is logistically complex, time-consuming and expensive. Over the last 40 years, forest tree breeders have decreased breeding cycle time and improved the estimates of heritability of most traits which led to greater gains per cycle (White *et al.* 2007). Most breeders calculate BLUP-BVs from phenotypic information obtained from field trials with progeny from pedigreed breeding populations, to rank parents and progeny for selection. Despite these advances, it is still vital to decrease breeding cycle time and increase gain per cycle.

The gain per cycle is affected by the accuracy of BLUP-BV. Errors in the pedigree can lead to biased BLUP-BV predictions, and have been estimated to average 10% (Banos *et al.* 2001; Visscher *et al.* 2002; Doerksen and Herbinger 2010), although these vary from program to program. Correcting pedigree errors should improve BLUP-BV predictions and improve heritability estimates. Pedigree errors have been corrected by genotyping (e.g. SSR fingerprinting) parents and progeny or from the diagonal of the realized relationship matrix to detect foreign populations (Simeone *et al.* 2011). Here we propose the use of the normality property of the different relationship classes to correct errors in the pedigree. Recent advances in genotyping methods enable the rapid development of dense panels of molecular markers that, as we show, can be used to correct historical errors carried in the pedigree. The use of a dense panel of markers has the advantage of being a byproduct of the genomic selection objective.

To correct errors, a realized relationship matrix (Powell *et al.* 2010) is constructed for the breeding population. In the relationship matrix, a normal symmetric and uni-modal distribution for each relationship class (i.e. unrelated, half-sib or full-sib) is

expected because of Mendelian sampling (Simeone *et al.* 2011). This has been observed with 294,831 SNPs markers on 3,925 human individuals with a standard deviation between 0.004 and 0.005 (Yang *et al*. 2010). As more markers are added more precise estimations of the Mendelian sampling will be obtained and, thus, smaller standard deviations are observed (Hayes *et al.* 2009a). In our case, we detected a bi-modal asymmetrical distribution for half-sibs, indicating problems in the recorded pedigree and showing a bias for the mean relationship (see Figure 2-1). The additional peak observed in the distribution was centered on zero, indicating that unrelated individuals were being misclassified as half-sibs. After re-assignment of individuals and correction of pedigree, the expected normal distribution was observed, as well as a considerable decrease in the standard deviation. This also was the case for the full-sib relationship class. Although we obtained a large decrease in the standard error, our estimations are still high compared with those obtained by Yang *et al.* (2010) or Simeone *et al.* (2011), probably due to the reduced number of SNPs markers (~2,300) genotyped on a smaller population (~860 individuals) with many different relationship classes derived from the circular mating design (i.e. unrelated, half-sibs, full-sibs, etc.). Better estimations are expected as more markers and individuals are added in future studies.

The extended length of a pine breeding cycle and their reproductive biology, contribute to a high likelihood of pedigree errors. Pines are wind pollinated and pollen from foreign genotypes is commonly present during controlled pollinations. Similarly, the length of the breeding cycle implies that record keeping is prone to include errors (White *et al.* 2007). Most errors can be corrected by re-assigning individuals, parents or

families present in the known pedigree, although the necessity of adding new parents indicates pollen contamination (Adams *et al.* 1988). In our case, three individuals were dropped from further analysis as they yielded inconsistent relationships. These inconsistent relationships of these three individuals were due to large amounts of missing SNP data, indicating genotyping problems.

Estimation of Breeding Value with Original and Corrected Relationship Matrices

Independently of the stage when the errors originated, our results show that pedigree errors decrease the accuracy of the BLUP-BV prediction, as previously reported in pines and dairy cattle (Ericsson 1999; Banos *et al.* 2001; Sanders *et al.* 2006). In addition to improved BLUP-BV accuracy, using the corrected pedigree, instead of the original, dramatically increased the fit of the data (Table 2-4). This indicated that with the original pedigree the heritability was slightly overestimated in eight traits and underestimated in seven (Table 2-4). The impact of correcting the pedigree on the BLUP analysis not only depends on the number of errors but also on how much difference existed between the phenotypic value of the individual, and the average of the family where the individual was incorrectly assigned. This happens because the traditional BLUP analysis shrinks the individual records towards the parental average of the family defined in the A matrix. When the phenotype of the mislabeled individual is similar to the family average in which this individual was misassigned, the estimated breeding value will be less biased than in a situation where the difference between the phenotypic value and the average of the family is large. However, even in those cases, there are some practical considerations regarding inbreeding and selection. If the best performing individuals are mislabeled, then related individuals may be selected inadvertently, or conversely, selection of superior unrelated

individuals may be avoided because they are labeled as the same family. Both cases will impact the potential genetic gain, the first through inbreeding depression and the second in the loss of opportunity of selection one of the best individuals.

Accuracy of Genomic Selection Predictive Models with Original and Corrected Pedigrees

Genomic selection offers the possibility to dramatically accelerate tree genetic improvement by eliminating, in some phases, the need of field tests to select superior individuals. Furthermore, selection of elite individuals can be more accurate compared to traditional phenotypic selection (Resende *et al.* 2012a). Many different methodologies have been proposed to construct GS prediction models with the aim of increasing their accuracy. However, for most quantitative traits there is not a clear advantage of any of the proposed prediction methods (Heslot *et al.* 2012; Resende *et al.* 2012b). Nonetheless, other opportunities exist for improvement of the accuracy of GS prediction models. One approach adopted in this study was the improvement of the BLUP-BV used as input for constructing the GS models by correcting errors in the pedigree.

When BVs derived from the corrected pedigree were deregressed and used to construct GS models, the accuracy of these models increased for 13 of 15 traits. This included seven out of the eight traits that previously had a decrease in heritability in the BLUP analysis. This indicates that GS models more efficiently capture associations between markers and QTLs, when the correct pedigree is used to estimate BLUP-BV. The traits BA_6 and CWAC_6 showed a smaller GS prediction ability with the new pedigree; however, these traits showed a slightly smaller or equal accuracy for the BLUP-BV prediction and a high increase in data fitting (AIC) indicating that the original pedigree was overestimating the GS predictive ability in these two cases.

Table 2-1. Age of trait measurement and code trait-age combination

Trait	Age measured	Code	Trait	Age measured	Code
BA	6	BA_6	DBH	3	DBH_3
BD	6	BD_6	DBH	4	DBH_4
BLC	4	BLC_4	DBH	6	DBH_6
BLC	6	BLC_6	HT	1	HT_1
CWAC	2	CWAC_2	HT	2	HT_2
CWAC	6	CWAC_6	HT	3	HT_3
CWAL	2	CWAL_2	HT	6	HT_6
CWAL	6	CWAL_6			

Table 2-2. Original and corrected pedigree mean and standard deviation for relationship classes in the population.

		Original Pedigree		Corrected Pedigree	
Relationship Class	Expected Relationship Coefficient	Mean	Standard Deviation	Mean	Standard Deviation
Unrelated	0.0000	-0.0382	0.044	-0.0005	0.015
Half Sibs	0.2500	0.1974	0.089	0.2500	0.042
Full Sibs	0.5000	0.4563	0.121	0.5001	0.061
Self	1.0000	1.0121	0.055	0.9997	0.040

Table 2-3. Number of individuals in each pedigree category in the original and new pedigree

Category	Original pedigree	Corrected Pedigree
Clones	956	940
Females	26	26
Males	27	37
Families	61	71

Table 2-4. Narrow-sense heritability (h^2_m), accuracy of breeding values (Acc(BV)) and fitting of models (log L = maximum of Log(likelihood), AIC= Akaike Information Criteria) by traditional BLUP analysis using a full genetic model with original pedigree or using a full genetic model with corrected pedigree (from eq. 1) on 15 trait-age combination.

	Full Original Pedigree BLUP				Full Corrected Pedigree BLUP			
Trait	h^2	Acc(BV)	Log L	AIC	h^2	Acc(BV)	Log L	AIC
BA_6	0.33(0.08)	0.82	-9056.1	18122.3	0.33(0.08)	0.81	-9015.3	18040.7
BD_6	0.12(0.04)	0.69	2004.1	-3998.2	0.15(0.05)	0.72	2003.5	-3996.9
BLC_4	0.19(0.06)	0.78	-8167.9	16353.8	0.22(0.02)	0.81	-8044.6	16107.1
BLC_6	0.31(0.08)	0.79	37.7	-65.4	0.35(0.03)	0.82	38.5	-67.0
CWAC_2	0.23(0.02)	0.82	-5355.0	10728.0	0.22(0.02)	0.82	-5251.8	10521.6
CWAC_6	0.43(0.10)	0.85	-4898.4	9806.8	0.45(0.02)	0.85	-4834.6	9679.1
CWAL_2	0.21(0.02)	0.82	-4779.5	9577.0	0.21(0.02)	0.82	-4673.5	9365.0
CWAL_6	0.27(0.08)	0.79	-3898.8	7807.6	0.27(0.03)	0.79	-3838.3	7686.7
DBH_3	0.27(0.02)	0.83	-4304.4	8626.9	0.26(0.02)	0.83	-4292.3	8602.6
DBH_4	0.28(0.02)	0.83	-6165.2	12348.5	0.27(0.02)	0.83	-6146.8	12311.6
DBH_6	0.32(0.02)	0.85	-7996.2	16010.3	0.31(0.02)	0.85	-7971.0	15959.9
HT_1	0.11(0.03)	0.75	-3727.4	7472.8	0.12(0.03)	0.77	-3622.3	7262.6
HT_2	0.27(0.02)	0.82	-29071.5	58160.9	0.27(0.02)	0.84	-28950.9	57919.7
HT_3	0.28(0.08)	0.83	-2593.0	5203.9	0.27(0.02)	0.84	-2456.6	4931.2
HT_6	0.26(0.07)	0.80	-5091.4	10194.8	0.31(0.02)	0.85	-4944.6	9901.1

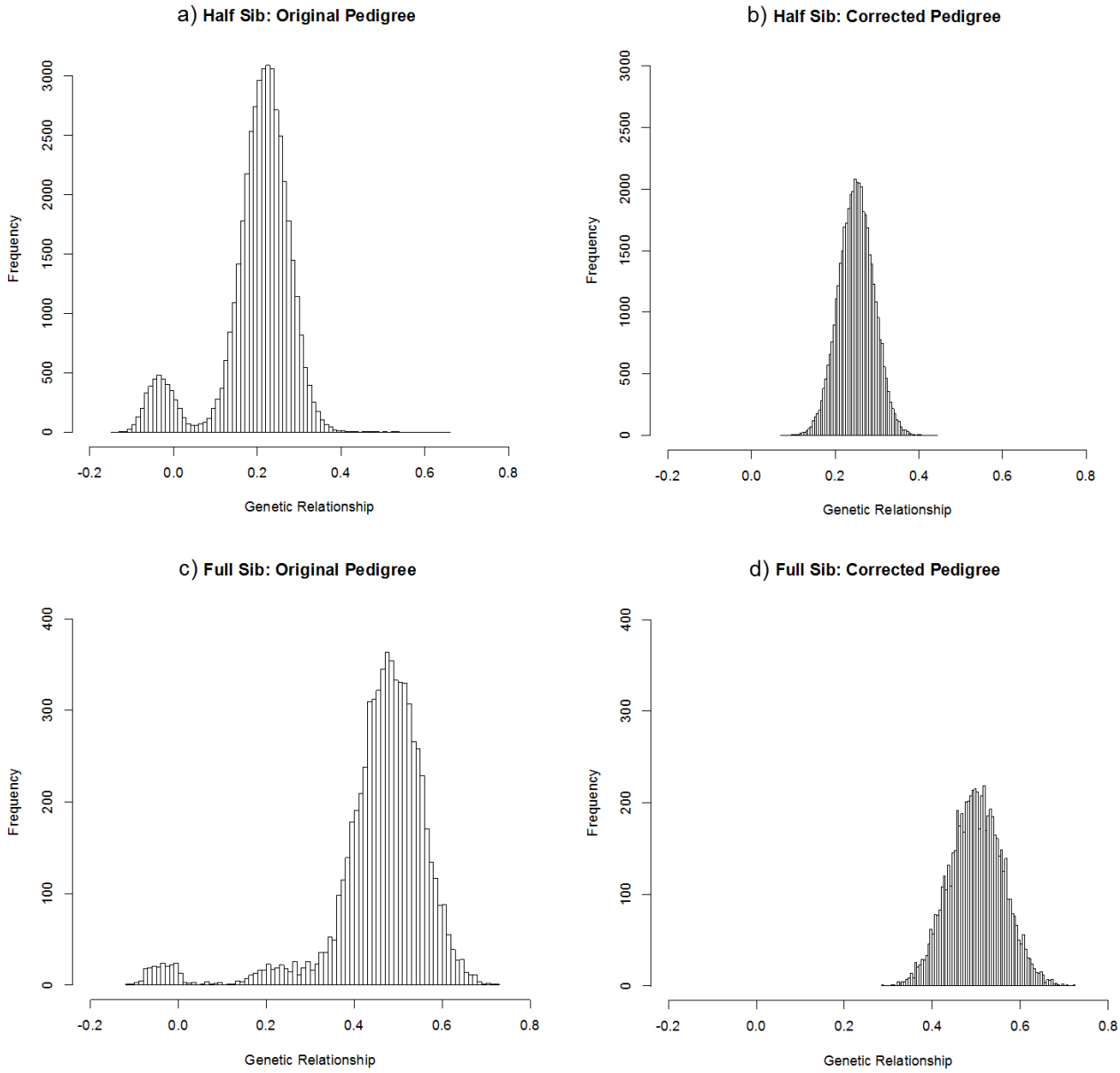

Figure 2-1. Distribution of relationship values for half-sib and full-sib individuals around their expected means 0.25 and 0.5, respectively. Distribution for the original pedigree (a and c) and corrected pedigree (b and d).

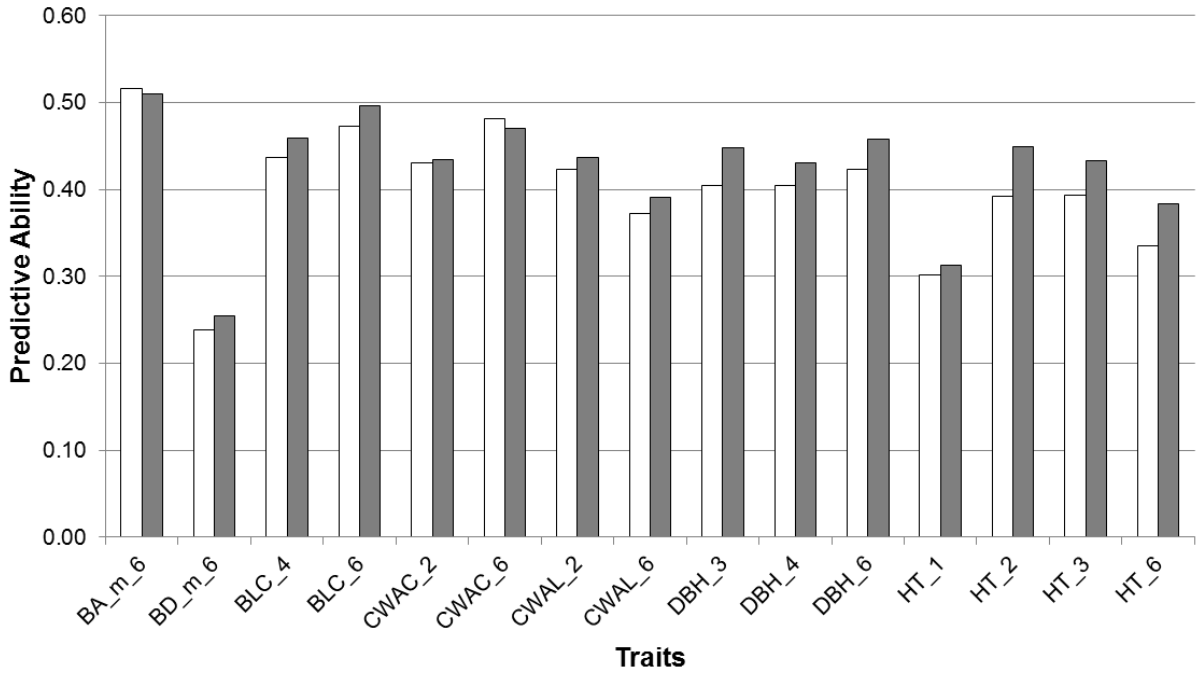

Figure 2-2. Predictive ability for fifteen different traits using the original pedigree derived from historical records (White column) and the corrected version of the pedigree (Grey column).

CHAPTER 3
ACCURACY OF GENOMIC SELECTION METHODS IN A STANDARD DATA SET OF LOBLOLLY PINE (PINUS TAEDA L.)[1]

Background

Plant and animal breeders have effectively used phenotypic selection to increase the mean performance in selected populations. For many traits, phenotypic selection is costly and time-consuming, especially so for traits expressed late in the life cycle of long-lived species. Genome-Wide Selection (GWS) (Meuwissen *et al.* 2001) was proposed as an approach to accelerate the breeding cycle. In GWS, trait-specific models predict phenotypes using dense molecular markers from a base population. These predictions are applied to genotypic information in subsequent generations to estimate Direct Genetic Values (DGV).

Several analytical approaches have been proposed for genome-based prediction of genetic values, and these differ with respect to assumptions about the marker effects (de los Campos *et al.* 2009a; Habier *et al.* 2011; Meuwissen *et al.* 2001). For example, Ridge Regression – Best Linear Unbiased Prediction (RR-BLUP) assumes that all marker effects are normally distributed, and that these marker effects have identical variance (Meuwissen *et al.* 2001). In Bayes A, markers are assumed to have different variances, and are modeled as following a scaled inverse chi-square distribution (Meuwissen *et al.* 2001). The prior in Bayes B (Meuwissen *et al.* 2001) assumes the variance of markers to equal zero with probability π, and the complement with probability (1 – π) follows an inverse chi-square distribution, with *v* degree of freedom and scale parameter *S*. The definition of the probability π depends on the genetic

[1] Chapter published in Genetics, Vol. 190, 1503-1510 April 2012

architecture of the trait, suggesting an improvement to the Bayes B model, known as Bayes Cπ. In Bayes Cπ, the mixture probability π has a prior uniform distribution (Habier *et al.* 2011). A drawback of Bayesian methods is the need for the definition of priors. The requirement of a prior for the parameter π is circumvented in the Bayesian LASSO method, which needs less information (de los Campos *et al.* 2009b; Legarra *et al.* 2011b). Methods for genomic prediction of genetic values may perform differently for different phenotypes (Habier *et al.* 2011; Meuwissen *et al.* 2001; Usai *et al.* 2009) and results may diverge because of differences in genetic architecture among traits (Grattapaglia and Resende 2011; Hayes *et al.* 2009b). Therefore, it is valuable to compare performance among methods with real data and identify those which provide more accurate predictions.

Recently, GWS was applied to agricultural crops (Crossa *et al.* 2010) and trees (Resende *et al.* 2011). Here we report, for an experimental breeding population of the tree species loblolly pine (*Pinus taeda* L.), a comparison of GWS predictive models for 17 traits with different heritabilities and predicted genetic architectures. Genome-wide selection models included RR-BLUP, Bayes A, Bayes Cπ and the Bayesian LASSO. In addition, we evaluated a modified RR-BLUP method that utilizes a subset of selected markers, RR-BLUP B. We show that, for most traits, there is limited difference among these four original methods in their ability to predict GBV. Bayes Cπ performed better for fusiform rust resistance – a disease-resistance trait showed previously to be controlled in part by major genes and the proposed method RR-BLUP B was similar to or better than Bayes Cπ when a subsample of markers was fitted to the model.

Materials and Methods

Training Population and Genotypic Data

The loblolly pine population used in this analysis is derived from 32 parents representing a wide range of accessions from the Atlantic Coastal Plain, Florida and Lower Gulf of the United States. Parents were crossed in a circular mating design with additional off-diagonal crosses, resulting in 70 full-sib families with an average of 13.5 individuals per family (Baltunis *et al.* 2007a). This population is referred to hereafter as CCLONES (Comparing Clonal Lines On Experimental Sites). A subset of the CCLONES population, composed of 951 individuals from 61 families (mean = 15, standard deviation = 2.2) was genotyped using an Illumina Infinium™ assay (Illumina, San Diego, CA; Eckert *et al.* 2010) with 7,216 SNP, each representing a unique pine EST contig. A subset of 4,853 SNPs were polymorphic in this population, and were used in this study. None of the markers were excluded based on minimum allele frequency.

Phenotypic Data

The CCLONES population was phenotyped for growth, developmental and disease resistance traits in three replicated studies. The first was a field study established using single-tree plots in eight replicates (one ramet of each individual is represented in each replicate) that utilized a resolvable alpha-incomplete block design (Williams *et al.* 2002). In that field trial, four replicates were grown under a high intensity and four were grown under a standard silvicultural intensity regime. The traits stem diameter (DBH, cm), total stem height (HT, cm) and total height to the base of the live crown (HTLC, cm) were measured in the eight replicates at years 6, 6 and 4, respectively. At year 6, crown width across the planting beds (CWAC, cm), crown width

along the planting beds (CWAL, cm), basal height of the live crown (BLC, cm), branch angle average (BA, degrees) and average branch diameter (BD, cm) were measured only in the high intensity silvicultural treatment. Phenotypic traits tree stiffness (Stiffness, km^2/s^2), lignin content (Lignin), latewood percentage at year 4 (LateWood), wood specific gravity (Density), and 5- and 6- carbon sugar content (C5C6) were measured only in two repetitions, in the high intensity culture (Baltunis *et al.* 2007a; Emhart *et al.* 2007; Li *et al.* 2007; Sykes *et al.* 2009).

The second study was a greenhouse disease resistance screen. The experimental design was a randomized complete block with single-tree plots arranged in an alpha lattice with an incomplete block (tray container). Fusiform rust (*Cronartium quercuum* Berk. Miyable ex Shirai f. sp. *fusiforme*) susceptibility was assessed as gall volume (Rust_gall_vol) and presence or absence of rust (Rust_bin) (Kayihan *et al.* 2005; Kayihan *et al.* 2010).

Finally, in the third study the rooting ability of cuttings was investigated in an incomplete block design (tray container) with four complete repetitions, in a controlled greenhouse environment. Root number (Rootnum) and presence or absence of roots (Rootnum_bin) were quantified (Baltunis *et al.* 2005; Baltunis *et al.* 2007b).

Breeding Value Prediction

Analyses were carried out using ASReml v.2 (Gilmour *et al.* 2006) with the following mixed linear model:

$$y = Xb + Z_1 i + Z_2 a + Z_3 c + Z_4 f + Z_5 d_1 + Z_6 d_2 + e$$

Where *y* is the phenotypic measure of the trait being analyzed, *b* is a vector of the fixed effects, *i* is a vector of the random incomplete block effects within replication ~N(0,

$\mathbf{I}\sigma^2_{iblk}$), *a* is a vector of random additive effects of clones, ~N(0, $A\sigma^2_a$), *c* is a vector of random non-additive effects of clones ~N(0, $I\sigma^2_c$), *f* is a vector of random family effects ~N(0, $I\sigma^2_f$), d_1 and d_2 are described below, *e* is the vector of random residual effects ~N(0, $DIAG\sigma^2_e$), X and Z_1-Z_6 are incidence matrices and I, A and DIAG are the identity, numerator relationship and block diagonal matrices, respectively. For traits measured in the field study under both high and standard culture intensities, the model also included d_1, a vector of the random additive x culture type interaction ~N(0, $DIAG\sigma^2_{d1}$), and d_2, a vector of the random family x culture type interaction ~N(0, $DIAG\sigma^2_{d2}$). Narrow sense heritability was calculated as the ratio of the additive variance σ^2_a to the total or phenotypic variance (e.g. for the field experiment total variance was $\sigma^2_a+\sigma^2_n+\sigma^2_f+\sigma^2_{d1}+\sigma^2_{d2}+\sigma^2_e$). Prior to use in GWS modeling, the estimated breeding values were deregressed into phenotypes (DP) following the approach described in Garrick *et al.* (2009), to remove parental average effects from each individual.

Statistical Methods

The SNP effects were estimated based on five different statistical methods: RR-BLUP, Bayes A (Meuwissen *et al.* 2001), Bayes Cπ (Habier *et al.* 2011), the Improved Bayesian LASSO (BLASSO) approach proposed by Legarra *et al.* (2011b) and RR-BLUP B (a modified RR-BLUP). In all cases the genotypic information was fitted using the model:

$$DP = 1\beta + Zm + \varepsilon$$

Where DP is the vector of phenotypes deregressed from the additive genetic values (Garrick *et al.* 2009), β is the overall mean fitted as a fixed effect, *m* is the vector of random marker effects and ε is the vector of random error effects, 1 is a vector of ones

and Z is the incidence matrix *m*, constructed from covariates based on the genotypes. No additional information, such as marker location, polygenic effects, or pedigree was used in those models.

Once the marker effects were estimated using one of the methods, the predicted direct genetic value (DGV) of individual *j* for that method was given by:

$\hat{g}_j = \sum_i^n Z_{ij}\hat{m}_i$

Where *i* is the specific allele of the i^{th} marker on individual *j* and *n* is the total number of markers.

Random Regression Best Linear Unbiased Predictor (RR-BLUP)

The RR-BLUP assumed the SNP effects, *m,* were random (Meuwissen *et al.* 2001). The variance parameters were assumed to be unknown and were estimated by restricted maximum likelihood (REML), which is equivalent to Bayesian inference using an uninformative, flat prior. The first and second moments for this model are described below:

$m \sim (0, G = I\sigma_m^2)$ $\quad\quad$ $E(y) = 1\beta$

$\varepsilon \sim (0, R = I\sigma_e^2)$ $\quad\quad$ $Var(y) = V = ZGZ' + R$

where σ_m^2 is the variance common to each marker effect and σ_e^2 is the residual variance.

The mixed model equation for the prediction of *m* is equivalent to:

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + I\,\frac{\sigma_e^2}{\sigma_a^2/\eta} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{m} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

Where σ_a^2 refers to the total additive variance of the trait and η, due to standardization of the Z matrix, refers to the total number of markers (Meuwissen *et al.* 2009). The matrix Z was parameterized and standardized to have a mean of zero and variance of

one as previously described (Resende *et al.* 2010; Resende *et al.* 2012a). The analyses were performed in the software R (available at CRAN, http://cran.r-project.org/).

Bayes A

The Bayes A method proposed by Meuwissen *et al.* (2001) assumes the conditional distribution of each effect (given its variance) to follow a normal distribution. The variances are assumed to follow a scaled inversed chi-square distribution with degrees of freedom *va* and scale parameter S^2a. The unconditional distribution of the marker effects can be shown to follow a t-distribution with mean zero (Sorensen and Gianola 2002). Bayes A differs from RR-BLUP in that each SNP has its own variance. In this study, *va* was assigned the value 4, and S^2a was calculated from the additive variance according to Habier *et al.* (2011) as follows:

$S_a^2 = \frac{\tilde{\sigma}_a^2(v_a-2)}{v_a}$, where:

$$\tilde{\sigma}_a^2 = \frac{\tilde{\sigma}_s^2}{(1-\pi)\sum_{k=1}^{K} 2p_k(1-p_k)}$$

and p_k is the allele frequency of the *k*-th SNP.

Bayes Cπ

Bayes Cπ was proposed by Habier *et al.* (2011). In this method, the SNP effects have a common variance, which follows a scaled inverse chi-square prior with parameters *va*, S^2a. As a result, the effect of a SNP fitted with probability (1-π) follows a mixture of multivariate Student's t-distributions, t(0, *va*, I S^2a), where π is the probability of a marker having zero effect. Parameters *va*, and S^2a were chosen as described for Bayes A. The π parameter is treated as unknown with a uniform (0,1) prior distribution.

Bayes A and Bayes C**π** were performed using the software GenSel (Fernando and Garrick 2008); available at http://bigs.ansci.iastate.edu/bigsgui/) for which an R

package is available in the Supporting Information (File S5). The marker input file was coded as -10, 0 and 10 for marker genotypes 0,1 and 2, respectively. A total of 50,000 iterations were used, with the first 2,000 excluded as the burn-in.

Bayesian LASSO

The Bayesian LASSO method was performed as proposed by Legarra *et al.* (2011b), using the same model equation used previously for the estimation of the markers effects. However, in this case:

$m|\lambda \sim \prod_i^n \frac{\lambda}{2} \exp(-\lambda|a_i|)$; $e|\sigma_e^2 \sim MVN(0, I\sigma_e^2)$

$var(m) = \frac{2}{\lambda^2}$

Using a formulation in terms of an augmented hierarchical model including an extra variance component τ_i^2 associated to each marker locus, we have:

$p(m|\tau) \sim N(0, D); diag(D) = (\tau_1^2 \ldots \tau_n^2)$

$p(\tau|\lambda) = \prod_i (\frac{\lambda^2}{2}) \exp(\frac{-\lambda^2 \tau_i^2}{2})$

therefore, $Var(m_i) = \sigma_{mi}^2 = \tau_i^2$

The prior distribution for σ^2_e was an inverted chi- square distribution with 4 degrees of freedom and expectations equal to the value used in regular genetic evaluation for σ^2_e. Analyses were performed using the software GS3 (Legarra *et al.* 2011a); available in http://snp.toulouse.inra.fr/~alegarra/. The chain length was 100,000 iterations, with the first 2,000 excluded as the burn-in and a thinning interval of 100.

RR-BLUP B

We also evaluated a modified, two-step RR-BLUP method that reduces the number of marker effects estimated. In this case, the DGV for each trait was generated

based on a reduced subset of markers. In order to define the number of markers in the subset, the marker effects from the RR-BLUP were ranked in decreasing order by their absolute values, and grouped in multiples of 10 (10, 20, 30, …, 4800). Each group was used, with their original effects, to estimate DGV. The size, *q*, of the subset that maximized the predictive ability was selected as the optimum number of marker effects to be used in subsequent analyses. Next, markers effects were re-estimated in a second RR-BLUP, using only the selected *q* markers within each training partition (see below). The estimated effects derived from this analysis were used to predict the merit of the individuals in the validation partition that were not present in the training partition. This process was repeated for different allocations of the data into training and validation partitions. In each validation, a different subset of markers was selected, based on the highest absolute effects within that training partition. Therefore, the only restriction applied to the second analysis was related to *q*, the number of markers to be included in each dataset. The same approach was performed with two additional subsets of markers of the same size as a control: the first subset contained randomly selected markers and the second subset contained markers with the smallest absolute effect values.

Validation of the models

Two cross validation schemes were tested in the RR-BLUP method: 10-fold and leave-one-out. For the 10-fold cross validation approach a random sub-sampling partitioning, fixed for all methods, was used (Kohavi 1995). Briefly, the data for each trait were partitioned into two subsets. The first one was composed by the majority of the individuals (90%) and was used to estimate the marker effects. The second one, the validation partition (10%), had their phenotypes predicted based on the marker effects

estimated in the training set. Randomly taken samples of N= (9/10) x N_T individuals were used as training sets, while the remaining individuals were used for validation (N_T= total number of individuals in the population). The process was repeated ten times, each time with a different set of individuals as the validation partition, until all individuals had their phenotypes predicted (Legarra *et al.* 2008; Usai *et al.* 2009; Verbyla *et al.* 2010). In the leave-one-out approach, a model was constructed using N_T-1 individuals in the training population and validated in a single individual that was not used in the training set. This was repeated N_T times, such that each individual in the sample was used once as the validation individual. This method maximized the training population size.

Accuracy of the models

The correlation between the Direct Genetic Values (DGV) and the Deregressed Phenotype (DP) was estimated using the software ASReml v.2 (Gilmour *et al.* 2006) from a bivariate analysis, including the validation groups as fixed effects since each validation group had DGV estimated from a different prediction equation and might have had a different mean. This correlation represented the predictive ability ($r_{y\hat{y}}$) of GS to predict phenotypes, and was theoretically represented (Resende *et al.* 2010) by:

$$r_{y\hat{y}} = r_{g\hat{g}}\, h$$

where $r_{g\hat{g}}$ was the accuracy of GS and h was the square root of the heritability of adjusted phenotypes, which is associated to Mendelian sampling effects and is given by

$$h_m^2 = \frac{n\ 0.5\ \sigma_a^2}{n\ 0.5\ \sigma_a^2 + \sigma_e^2}$$

where n was the number of ramets used in each study. To remove the influence of the heritability upon the predictive ability and thus estimate the accuracy, the following formula was applied

$$r_{g\hat{g}} = \frac{r_{y\hat{y}}}{h}$$

In addition, for each method and trait, the slope coefficient for the regression of DP on DGV was calculated as a measurement of the bias of the DGV. Unbiased models are expected to have a slope coefficient of 1, whereas values greater than 1 indicate a biased underestimation in the DGV prediction and values smaller than indicate a biased overestimation of the DGV.

Results

Cross Validation Method

Testing the effect of cross validation using two methods, 10-fold and leave-one-out (N-fold), showed that their predictive ability was not significantly different (Table A-1). The largest difference was detected for the trait CWAC, where the leave-one-out method outperformed the 10-fold cross validation by 0.02 (standard error = 0.03). Likewise, no significant differences were found for bias of the regressions (slope) in both methods (Table A-2). Thus, the 10-fold approach was selected and used for comparing all methods.

Predictive Ability of the Methods

Four well-established genome-wide selection methods were compared in 17 traits with heritabilities ranging from 0.07 to 0.45. Overall, the ability to predict phenotype ($r_{y\hat{y}}$) ranged from 0.17 for Lignin to 0.51 for BA (Table 3-1). Although the methods differ in *a priori* assumptions about marker effects, their predictive ability was similar – no significant differences were detected for any of the 17 traits. The standard errors for each method and trait are described in (Table A-3).

Bayesian approaches performed better for traits in the disease resistance category. For Rust_bin, the methods Bayes A and Bayes Cπ were 0.05 superior than RR-BLUP and 0.06 superior to BLASSO. For Rust_gall_vol, Bayes Cπ was 0.05 superior to RR-BLUP and BLASSO. The accuracy $(r_{g\hat{g}})$ for each genome-wide prediction method was also estimated, and varied from 0.37–0.77 (Table A-4).

For all methods, the ability to predict phenotypes $(r_{y\hat{y}})$ was linearly correlated with trait heritability. The strongest correlation (0.79) was observed for RR-BLUP (Figure 3-1). The correlation is expected, as traits with lower heritability have phenotypes less reflective of their genetic content, and are expected to be less predictable through genomic selection.

Bias of the Methods

The coefficient of regression (slope) of DP on DGV was calculated as a measurement of the bias of each method. Ideally, a value of beta equal to one indicates no bias in the prediction. For all traits, the slopes of all the models were not significantly different than one, indicating no significant bias in the prediction. In addition, no significant differences among the methods were detected (Table A-5). Although no evidence of significant bias was detected, the value of beta derived from RR-BLUP was slightly higher for all traits (average across traits equal to 1.18).

Markers Subset and RR-BLUP B

Prediction of phenotype was also performed with RR-BLUP, but adding increasingly large marker subsets, until all markers were used jointly in the prediction. The predictive ability was plotted against the size of the subset of markers (Figure 3-2). The pattern of the prediction accuracy was similar for 13 out 17 traits (Figure 3-2 left),

where differences were mainly found in the rate with which the correlation reached the asymptote. In these cases, the size of the subset ranged from 820 to 4,790 markers. However, a distinct pattern was detected for disease resistance related traits, density and CWAL (Figure 3-2 right). In these cases, maximum predictive ability was reached with smaller marker subsets (110-590 markers), and decreased with the addition of more markers. An additional RR-BLUP was performed using as covariates only the marker subset for which maximum predictive ability was obtained. For traits where a large number of markers (> 600) explain the phenotypic variability, RR-BLUP B was similar to RR-BLUP or Bayesian methods (Table A-6). However, for traits where the maximum predictive ability (Density, Rust_bin, Rust_gall_vol) was reached with a smaller number of marker (<600), RR-BLUP B performed significantly better than RR-BLUP. For example, the predictive ability of the trait Rust_gall_vol was 61% higher using RR-BLUP B (0.37) compared to the traditional RR-BLUP (0.23), and also improved relative to BLASSO (0.24), Bayes A (0.28) and Bayes Cπ (0.29).

We also contrasted these results with the predictive ability using a subset of markers of similar size, but selected either randomly or to include those markers with lower effects. As expected, for the three traits the predictive ability was larger for the subset selected by RR-BLUP B over the subsets with lower effects and random effects (Figure 3-3). A significant difference over the lower and random subsets was found for rust resistance related traits (Rust_bin, Rust_gall_vol), while for Density the markers selected by RR-BLUP B were only significantly different than the lower marker subset but not different to the random marker subset.

Discussion

We characterized the performance of RR-BLUP/RR-BLUP B, Bayes A, Bayesian LASSO regression and Bayes Cπ for GWS of growth, developmental, disease resistance and biomass quality traits in common dataset generated from an experimental population of the conifer loblolly pine. In general, the methods evaluated differed only modestly in their predictive ability (defined by the correlation between the DGV and DP).

The suitability of different methods of developing GWS predictive models is expected to be trait dependent, conditional on the genetic architecture of the characteristic. RR-BLUP differs from the other approaches used in this study in that the unconditional variance of marker effects is normally distributed, with the same variance for all markers (Meuwissen *et al.* 2001). This assumption may be suitable when considering an infinitesimal model (Fisher 1918), where the characters are determined by an infinite number of unlinked and non epistatic loci, with small effect. Not surprisingly, BLUP-based methods underperformed relative to Bayesian approaches for oligogenic traits. For instance, Verbyla *et al.* (2009) showed that BLUP-based GWS had lower accuracy, compared to Bayesian methods, in prediction of fat percentage in a population where a single gene explains ~50% of the genetic variation. Similarly, our observation that Bayes A and Bayes Cπ were more accurate in predicting fusiform rust resistance traits, compared to RR-BLUP, may reflect a simpler genetic architecture, with a few loci of large effect. While the causative genes that regulate fusiform rust resistance have not yet been uncovered, several genetic studies support the role of few major genes in the trait variation. For example, the *Fr1* locus confers resistance to

specific fungus aeciospore isolates (Wilcox *et al.* 1996), and at least five families within the CCLONES population segregate for this locus (Kayihan *et al.* 2010).

The under-performance of RR-BLUP for predicting oligogenic traits is a consequence of fitting a large number of makers to model variation at a trait controlled by few major loci, leading to model over-parameterization. In Bayes A and Bayes Cπ, the shrinkage of effects is marker-specific, while in BLUP all markers are penalized equally. To address this limitation, we proposed an alternative, RR-BLUP B, to Bayesian and the traditional RR-BLUP approaches, aimed at reducing the number of parameters. In RR-BLUP B, marker effects are initially estimated and ranked using RR-BLUP. Next, increasing markers subsets that include initially those with larger effect are used to estimate DGV. The number of markers that maximizes the predictive ability is then defined, and used in a second RR-BLUP model. For rust disease resistance and wood density traits, the modified RR-BLUP B approach performed better that traditional RR-BLUP, and as well as the Bayesian methods. Previous studies using simulated data have shown that improvements in predictive ability could be obtained by using a similar approach to the one proposed here (Zhang et al. 2010, Zhang et al. 2011), although with a different strategy of marker selection. While RR-BLUP B may add an additional step to the development of predictive models (i.e. initial marker selection), it is overall simpler and computationally less expensive than Bayesian approaches. Therefore, it may provide a suitable alternative to the use of BLUP-based methods for traits that do not fit an infinitesimal model, and are rather regulated by few major loci.

Table 3-1. Predictive ability of Genomic Selection models using four different methods; h^2 is the narrow-sense heritability of the trait.

Trait category	Trait	h^2	Methods			
			RR-BLUP	BLASSO	Bayes A	Bayes Cπ
Growth	HT	0.31	0.39	0.38	0.38	0.38
	HTLC	0.22	0.45	0.44	0.44	0.44
	BHLC	0.35	0.49	0.49	0.49	0.49
	DBH	0.31	0.46	0.46	0.46	0.46
Development	CWAL	0.27	0.38	0.36	0.36	0.36
	CWAC	0.45	0.48	0.46	0.47	0.47
	BD	0.15	0.27	0.25	0.27	0.27
	BA	0.33	0.51	0.51	0.51	0.51
Disease resistance	Rootnum_bin	0.10	0.28	0.28	0.27	0.28
	Rootnum	0.07	0.24	0.26	0.25	0.24
	Rust_bin	0.21	0.29	0.28	0.34	0.34
	Rust_gall_vol	0.12	0.23	0.24	0.28	0.29
Wood quality	Stiffness	0.37	0.43	0.39	0.42	0.42
	Lignin	0.11	0.17	0.17	0.17	0.17
	LateWood	0.17	0.24	0.24	0.23	0.24
	Density	0.09	0.20	0.22	0.23	0.22
	C5C6	0.14	0.26	0.25	0.25	0.25

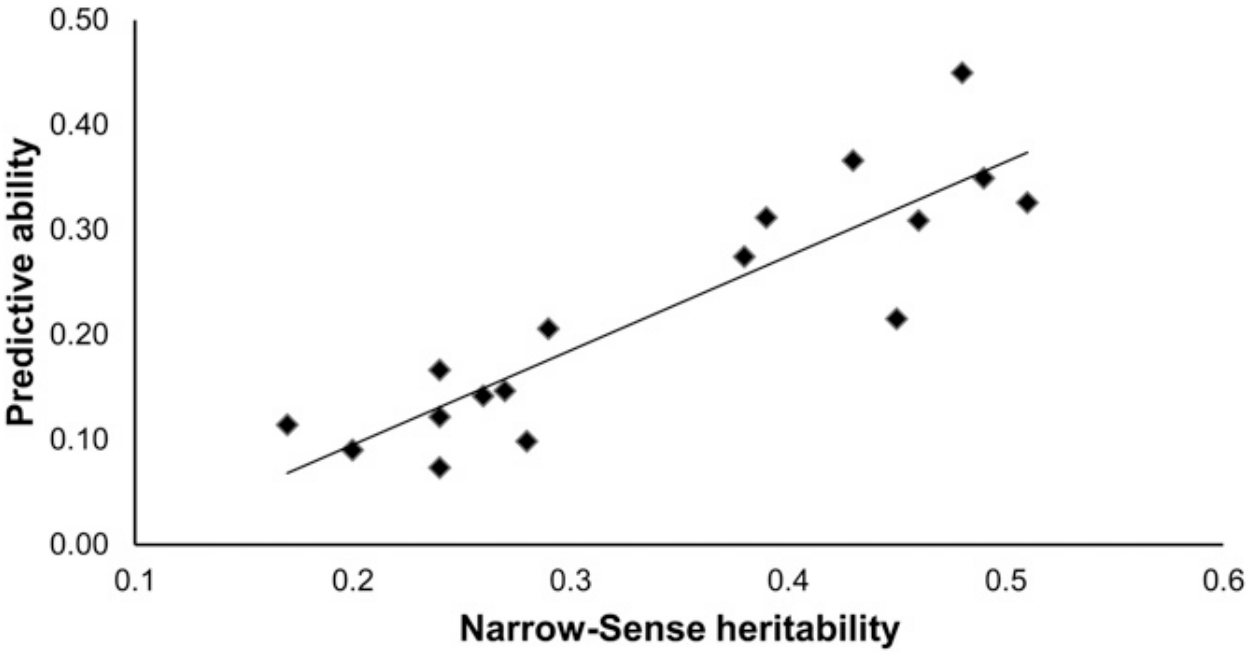

Figure 3-1. Regression of RR-BLUP predictive ability on narrow sense heritability for 17 traits (trend line is shown, R^2=0.79)

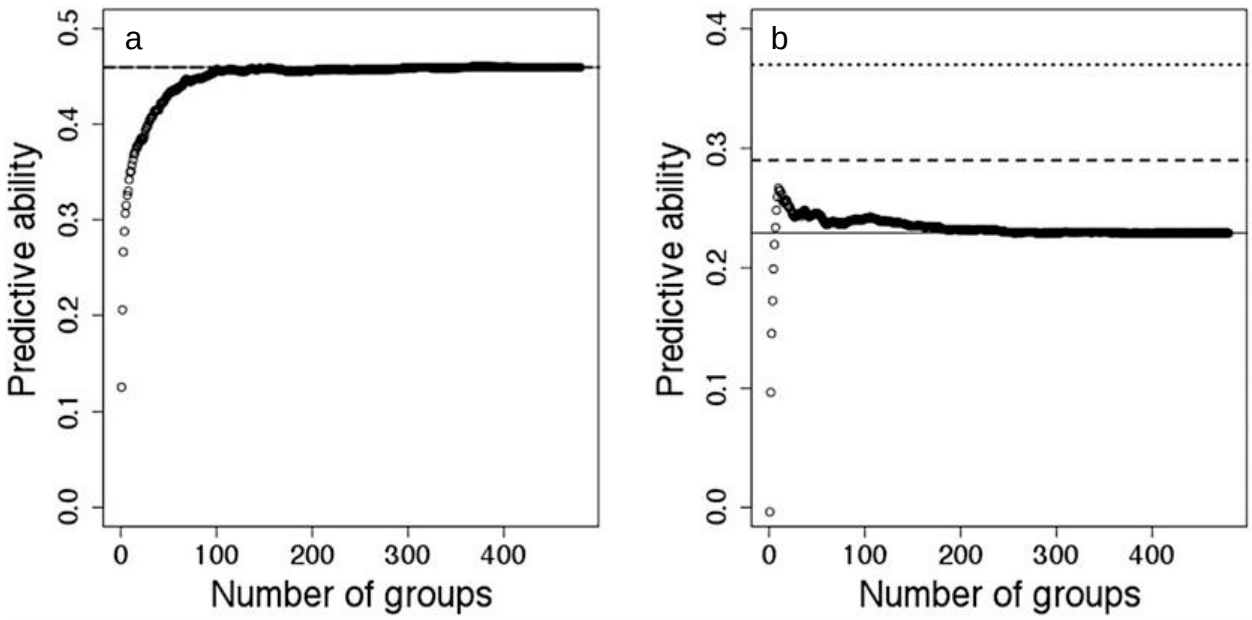

Figure 3-2. Example of the two patterns of predictive ability observed among traits, as an increasing number of markers is added to the model. Each marker group is represented by a set of 10 markers. a) For DBH, the maximum predictive ability was detected when 380 groups of markers (3,800 markers) were included in the model. b) For the trait Rust_gall_vol, predictive ability pattern reached a maximum when only 10 groups (100 markers) were added. Lines indicate the predictive ability of RR-BLUP (filled line), Bayes Cπ (dashed line), and RR-BLUP B (dotted line) as reported on Table 1 and in Supporting Information - Table S6.

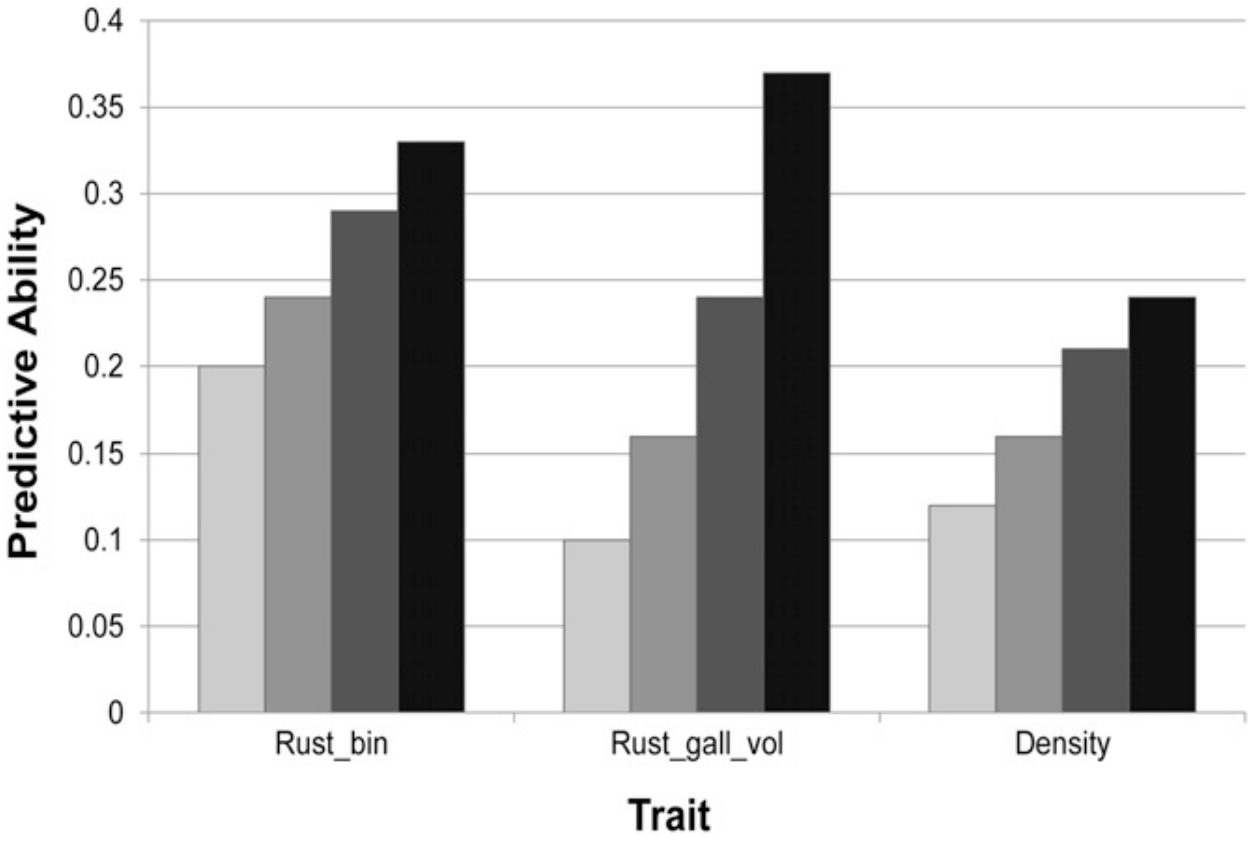

Figure 3-3. Predictive ability for subsets of 310 markers for Rust_bin, 110 markers for Rust_gall_vol and 240 markers for Density. Subsets were generated selecting markers with the lowest absolute effects (light grey), with random values (grey), including all markers (dark grey), and including only those markers with largest absolute effects (black).

CHAPTER 4
THE RE-DISCOVERY OF NON-ADDITIVE EFFECTS WITH GENOMIC RELATIONSHIP MATRICES AND ITS IMPLICATION IN BREEDING[1]

Background

Non-additive effects have been often neglected in animal and plant breeding, mainly because their variance estimates are usually small and non-significant compared to the additive effect (Hill *et al.* 2008). Consequently, many breeding programs rely only on additive values (breeding values) to make progress from selection, completely ignoring dominance and epistasis. Partitioning genetic variance into additive, dominance and epistasis is not always possible because the estimations are highly correlated (Hill 2010) and confounded with each other (Lynch and Walsh 1998). Depending on the gene frequencies involved, non-additive allelic effects could greatly inflate the additive variance estimates (Lu *et al.* 1999; Zuk *et al.* 2012) and therefore impact breeding value (BV) predictions (Palucci *et al.* 2007; Vanderwerf and Deboer 1989). In addition, the proper partitioning of variance components generates a basic understanding of the genetic architecture of traits, which in turn helps to define the breeding strategy that would maximize genetic gains. Thus, for traditional breeding programs, ignoring existing non-additive effects can negatively impact progress in three ways: 1) inflating the heritability and breeding values leading to overestimate of genetic gain, 2) the variability due to non-additive effects cannot be exploited in the next generation as the program selects by breeding values, and 3) if family or clonal propagation is possible in the breeding program, then there will be losses in the potential to exploit this variability and reach the maximum possible genetic gain.

[1] Manuscript prepared to be submitted to Genetics Journal

To partition the genetic variation into additive and dominance in breeding populations, mating designs with at least full-sib relationships are needed. Furthermore, to partition at the epistatic level requires either inbred or vegetatively propagated (clonal) populations. In perennial plants with long rotations (> 5 years), inbreds are not used because of the generation length and the often high amount of inbreeding depression, leaving clonal populations as the only option to explore the full genetic architecture (Foster and Shaw 1988). Several studies with clonal populations have been performed (Foster and Shaw 1988; Mullin *et al.* 1992; Wu1996; Isik *et al.* 2003, 2005; Costa e Silva *et al.* 2004, 2009; Baltunis *et al.*2007b, 2008, 2009; Araujo *et al.*2012) with the aim of partitioning genetic variance using traditional pedigree-based quantitative genetics approaches. All these studies obtained small estimated values for dominance variation and often null or negative values for epistasis variation. Furthermore, work by Hill *et al.* (2008) demonstrated that either all genetic variance is due to additive effects or, if non-additive effects are present they will be at least partially captured as additive variation.

Currently, the use of molecular markers has become a popular way to predict BVs with genomic selection (GS) models. Many analytic methodologies have been proposed with the aim of increasing the accuracy of prediction. However, no significant differences have been detected among methods for quantitative complex traits (Resende *et al.* 2012b; Heslot *et al*. 2012). The use of molecular markers for estimation of the additive relationships among individuals in a population, called realized relationship matrix (A_G, or Observed/Genomic Relationship Matrix) and its use instead of the numerator relationship matrix (A) in BLUP analysis (GBLUP) is one of the

proposed methods to predict BVs (Habier *et al.* 2010; Hayes *et al.* 2009; Veerkamp *et al.* 2011). Given that no large differences exist among methods, GBLUP will likely become the most popular way for predicting genomic BV (Aguilar *et al.* 2010), as BLUP is a well-known easily-understood methodology, and is equivalent to ridge regression BLUP (RR-BLUP) (VanRaden 2008). In addition, as GBLUP has equivalent properties to BLUP, an extended animal model can be fitted to incorporate dominance and epistatic effects, replacing the pedigree-derived non-additive relationship matrices (Mrode 2005) with the marker-derived counterpart. Use of A_G instead of the numerator relationship matrix (A) derived from the pedigree improves estimates because A_G better describes the true relationships among individuals known to be related by providing an estimates of Mendelian sampling and relationships of individuals not previously known to be related. In addition, the A_G matrix combined with the A matrix (Aguilar *et al.* 2010) yields more accurate estimates of variance components (Chen *et al.* 2011; Veerkamp *et al.* 2011). Finally, it has been shown that use of the A_G instead of A matrix better separates genetics from environmental effects (Lee *et al.* 2010). We hypothesize that use of marker-derived relationship matrices will allow better separation of the genetic variance components, revealing the genetic architecture of complex quantitative traits.

The objective of this study was to test the use of molecular-marker derived additive and non-additive relationship matrices for partitioning genetic variance components. For tree height, a trait with complex inheritance, we compared in a clonal population of *Pinus taeda* different BLUP models under additive and full (additive plus non-additive) assumptions with the use of either the pedigree-derived relationship matrix or the marker-derived matrices.

Materials and Methods

Data

Total tree height (HT, m) measured at year 6 from a single field trial of the CCLONES population (see Baltunis *et al.* 2005 for details) was used in the present study. In summary, 32 parents were crossed in a circular mating design with additional off-diagonal crosses, resulting in 70 full-sib families with an average of 13.5 individuals per family (Munoz *et al.* 2012). The clonal field trial was established using single-tree plots with eight replicates (one ramet per replicate) in a resolvable alpha-incomplete block design (Williams *et al.* 2002). Four of the replicates were grown under high intensity silviculture while the rest were under standard silviculture regime.

A subset of the CCLONES population, composed of 951 individuals from 61 families were genotyped using an Illumina Infinium™ assay (Illumina, San Diego, CA (Eckert *et al.* 2010) with 7,216 SNPs, each representing a unique pine EST contig. A subset of 4,853 SNPs were polymorphic in this population.

Relationship Matrices

Out of the polymorphic markers a total of 2,182 SNPs markers had a minor genotype frequency greater than 0.12. This subset was used to estimate A_G following the method proposed by Powell *et al.* (2010) where identity by descent coefficients are determined relative to the parents of the current population as the base population. The relationship values from A_G were adjusted as recommended by Yang *et al.* (2010) to lessen estimation error. The resulting A_G was used to correct the pedigree as detailed in Munoz *et al.* (2012). Also, a molecular marker derived dominance relationship matrix (D_G) was constructed. To build a dominance relationship matrix, a dominance incidence matrix (W) was created where a new codification was established for the genotypic file

containing all the polymorphic SNPs. Thus, W has the dimension n individuals times m markers and was re-parameterized to be coded 1 if the genotype was heterozygous and 0 if the marker genotype was homozygous for either class (MM and mm). The matrix W was further standardized to a mean of 0. This was done by calculating the expectation of $W_{\cdot j}$ for the j-th marker derived as $2p_jq_j$. Thus W_{ij} for the i-th individual and j-th marker received the code:

$W_{ij} = 1 - 2p_jq_j$ if the individual is heterozygous

$W_{ij} = 0$ if the individual has a missing data

$W_{ij} = 0 - 2p_jq_j$ otherwise.

Starting from the matrix W, the dominance relationship matrix was constructed using the following expression:

$$D = \frac{WW'}{\sum_{j=1}^{m} 2p_jq_j(1-2p_jq_j)}$$

where the denominator is the variance of W_{ij}. In addition, relationship matrices were obtained from the pedigree for additive relationships (A) and dominance relationships (D) following traditional methods (Lynch and Walsh 1998; Mrode 2005). The Hadamard product (#) between matrices was used to obtain the epistasis relationship matrices additive-by-additive (A#A), dominance-by-dominance (D#D) and additive-by-dominance (A#D) interaction for pedigree-derived, and markers-derived $A_G\#A_G$, $D_G\#D_G$, and $A_G\#D_G$, respectively.

Genetic Analyses

All analyses were carried out in the software ASReml v3.0 (Gilmour *et al.*2009), a genetic-statistics software for fitting mixed models on complex datasets using the sparse matrix methods and equipped with the Residual Maximum Likelihood (REML) for

variance component estimation using the average information algorithm (Gilmour *et al.* 1995).

Six models were fitted using the pedigree-derived matrices (models 1 to 6) and another equivalent six models using the marker-derived matrices (models 7 to 12), from the simpler (additive) to the more complex (additive plus dominance plus two-way epistasis interaction). The model including all terms (model 6 or 11) is show below

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z_1}\boldsymbol{i} + \mathbf{Z_2}\boldsymbol{a} + \boldsymbol{Z_3 t_1} + \boldsymbol{Z_4 d} + \boldsymbol{Z_5 t_2} + \boldsymbol{Z_6 i_{aa}} + \boldsymbol{Z_7 i_{dd}} + \boldsymbol{Z_8 i_{ad}} + \boldsymbol{e}$$

where $\mathbf{y}$ is the phenotypic HT measure, $\boldsymbol{\beta}$ is a vector of the fixed effects (i.e. silvicultural treatment and replicate), $\boldsymbol{i}$ is a vector of the random incomplete block effects within replication ~N(0, $\mathbf{I}\sigma^2_i$), $\boldsymbol{a}$ is a vector of random additive effects of individuals ~N(0, $\mathbf{B_1}\sigma^2_a$), $\boldsymbol{t_1}$ is a vector of the random additive by silviculture type interaction ~N(0, $\mathbf{B_1}\otimes\mathbf{I}\sigma^2_{t1}$), $\boldsymbol{d}$ is a vector of random dominance effect of individual ~N(0, $\mathbf{B}_2\sigma^2_{\boldsymbol{d}}$), $\boldsymbol{t_2}$ is a vector of the random dominance by silviculture type interaction ~N(0, $\mathbf{B_2}\otimes\mathbf{I}\sigma^2_{t2}$),$\boldsymbol{i_{aa}}$ is a vector of the random additive by additive interaction ~N(0, $\mathbf{B_1}\#\mathbf{B_1}\sigma^2_{iaa}$), $\boldsymbol{i_{dd}}$ is a vector of the random dominance by dominance interaction ~N(0, $\mathbf{B_2}\#\mathbf{B_2}\sigma^2_{idd}$), $\boldsymbol{i_{ad}}$ is a vector of the random additive by dominance interaction ~N(0, $\mathbf{B_1}\#\mathbf{B_2}\sigma^2_{iad}$) and $\boldsymbol{e}$ is the vector of random residual effects ~N(0, $\mathbf{I}\sigma^2_e$).The incidence matrices are $\mathbf{X}$, $\mathbf{Z_1}$-$\mathbf{Z_8}$, while $\mathbf{I}$ is the identity matrix, and $\otimes$ represent the Kronecker product and # the Hadamard product. The matrix $\mathbf{B_1}$ and $\mathbf{B_2}$ corresponded to an additive and dominance relationship matrices either derived from the pedigree replaced for A and D or from the markers replaced for A_G and D_G, respective. Under this model the narrow-sense heritability is $h^2 = \frac{\hat{\sigma}^2_a}{\hat{\sigma}^2_p}$, the dominance to total variance ratio $d^2 = \frac{\hat{\sigma}^2_d}{\hat{\sigma}^2_p}$, the epistatic to total variance ratio $i^2 = \frac{\hat{\sigma}^2_i}{\hat{\sigma}^2_p}$, and

the broad-sense heritability $H^2 = \frac{\hat{\sigma}_g^2}{\hat{\sigma}_p^2}$, where $\hat{\sigma}_a^2$ is the estimated additive variance, $\hat{\sigma}_d^2$ is the estimated dominance variance and $\hat{\sigma}_p^2$, $\hat{\sigma}_i^2$ and $\hat{\sigma}_g^2$ are the total, epistatic and genetic variance, respectively. The definition of the last three variance estimates varied accordingly to the model being fitted (see below for details).

Model 1 (A_ped) and 7 (A_G_MM) were fitted with only the additive effect and its interaction with silviculture. Thus, the estimated additive variance ($\hat{\sigma}_a^2$) equals the genetic variance component ($\hat{\sigma}_g^2$) as in this model non-additive effects are ignored $\hat{\sigma}_g^2 = \hat{\sigma}_a^2$ and the total variance is $\hat{\sigma}_p^2 = \hat{\sigma}_a^2 + \hat{\sigma}_{t1}^2 + \hat{\sigma}_e^2$. Model 2 (D_ped) and 8 (D_G_MM) included additive and dominance effects. Thus, the estimated additive variance plus dominance variance estimation ($\hat{\sigma}_d^2$) equals the genetic variance component as epistasis effect is ignored $\hat{\sigma}_g^2 = \hat{\sigma}_a^2 + \hat{\sigma}_d^2$ and the total variance is $\hat{\sigma}_p^2 = \hat{\sigma}_a^2 + \hat{\sigma}_{t1}^2 + \hat{\sigma}_d^2 + \hat{\sigma}_{t2}^2 + \hat{\sigma}_e^2$. Model 3 (A#A_ped) and 9 (A_G#A_G_MM) expanded model 2 to include epistasis as additive-by-additive interaction. Here the genetic variance component is $\hat{\sigma}_g^2 = \hat{\sigma}_a^2 + \hat{\sigma}_d^2 + \hat{\sigma}_i^2$, the total variance is $\hat{\sigma}_p^2 = \hat{\sigma}_a^2 + \hat{\sigma}_{t1}^2 + \hat{\sigma}_d^2 + \hat{\sigma}_{t2}^2 + \hat{\sigma}_i^2 + \hat{\sigma}_e^2$, and the epistatic variance $\hat{\sigma}_i^2 = \hat{\sigma}_{iaa}^2$. Model 4 (D#D_ped) and model 10 (D_G#D_G_MM) are similar to model 3 but replace the epistasis effect by dominance-by-dominance interaction $\hat{\sigma}_i^2 = \hat{\sigma}_{idd}^2$, while model 5 (A#D_ped) and model 11 (A_G#D_G_MM) are similar to model 3 but replace the epistasis effect by additive-by-dominance interaction $\hat{\sigma}_i^2 = \hat{\sigma}_{iad}^2$. Finally, model 6 (FULL_ped) and model 12 (Full_MM) includes all effects mentioned above, in this model $\hat{\sigma}_g^2 = \hat{\sigma}_a^2 + \hat{\sigma}_d^2 + \hat{\sigma}_i^2$, where $\hat{\sigma}_i^2 = \hat{\sigma}_{iaa}^2 + \hat{\sigma}_{idd}^2 + \hat{\sigma}_{iad}^2$ and the total variance is $\hat{\sigma}_p^2 = \hat{\sigma}_a^2 + \hat{\sigma}_{t1}^2 + \hat{\sigma}_d^2 + \hat{\sigma}_{t2}^2 + \hat{\sigma}_i^2 + \hat{\sigma}_e^2$. A summary of all models is presented in Table 4-1.

Testing and Validation of Models

The Akaike Information Criteria (AIC) was used to compare model fitting (Akaike 1974). The capacity of the different parameterizations to partition the variances estimated by the different models was evaluated using the sampling correlation matrix (**R**) among variance components estimates (Lee *et al.* 2010). The variance-covariance matrix of estimated variance components (**V**) was used to estimate **R** as $\boldsymbol{R} = \boldsymbol{L}^{-1/2}\boldsymbol{V}\boldsymbol{L}^{-1/2}$, where **L** is the diagonal of the **V** matrix.

The prediction ability of the different models was tested with a 10-fold cross validation with a random sub-sampling partitioning fixed for all models (Kohavi 1995). Briefly, the genotypes were partitioned in 10 groups, thus each model were run 10-times with different 9 groups for constructing the model and the 10th group to validate the model. At the end, all genotypes had their BV predicted from the model (PBV). On the other hand the BV derived from the models using the complete dataset in each model were assumed to be the real BV (RBV). The predictive ability of each model was tested as the correlation between RBV and PBV, and between predicted (PR) and real rankings (RR) for the top 10% genotypes, emulating an operational selection scenario.

Results

Table 4-2 summarizes the variance components, genetic parameters and indicators of data fitting, estimated with each of the twelve alternative models. Compared to the A_ped model, the A_G_MM model narrow- and broad-sense heritability (as non-additive effects were ignored) increased slightly, with both larger than 0.40. Including dominance effects in the pedigree-based model (D_ped) decreased h^2 by approximately 26% and the dominance ratio (d^2) was small (0.07) and non-significantly different from zero ($2*SE(d^2)>0.08$). When the dominance effect was included with the

molecular marker-based model (D_G_MM), h^2 decreased 47% to 0.24, d^2 estimation increased to 0.16, a highly significantly difference from zero and a 70% of the additive value. With this model, the dominance variance represents 40% of the total genetic variation. We further extended these models to include the additive-by-additive, dominance-by-dominance and additive-by-dominance two-way epistatic interaction factors in three separate models. In pedigree-based models with epistasis (A#A_ped, D#D_ped and A#D_ped) the estimations of variance components for additive and dominance varied slightly from those of the D_ped model. Moreover, in these three cases the estimation of epistasis was zero. The model including the additive-by-additive epistatic interaction (A_G#A_G_MM) could not be fit and a reduced version dropping the culture by epistatic interaction factor was fitted instead. In the A_G#A_G_MM model, the additive and dominance ratios (h^2 and d^2) dropped considerably, while the epistatic interaction increased to 0.23. When the dominance-by-dominance and additive-by-dominance interactions were added in models D_G#D_G_MM and A_G#D_G_MM, respectively, the additive component dropped more than 30% and the dominance dropped 88%, with respect to the D_G_MM model, while the epistatic ratio (i^2) was estimated at 0.15 and 0.17 for the D_G#D_G_MM and A_G#D_G_MM, respectively. Finally, a full model including additive, dominance and all two-way epistatic interactions was fitted for pedigree (Full_ped) and marker-based (Full_MM). In the case of the pedigree, the full model did not converge and a reduced version including additive, dominance, additive-by-additive and dominance-by-dominance was fitted. In this model the additive component decreased slightly from 0.30 (in the D_ped model) to 0.27, the dominance was estimated to be zero and the epistasis (as the sum of both two-way interactions)

was estimated close to 0.10. The estimation of variance components with the full marker-based model (Full_MM) was similar to the $A_G\#D_G$_MM model. Under this model the epistatic interaction was calculated as the sum of all three two-way epistatic interactions, with dominance-by-dominance having the greatest contribution while additive-by-dominance was almost zero.

Overall, including the non-additive effects improves the fitting of the data. Although this improvement is small for pedigree models, there is a substantial improvement when compared among marker-based models (Table 4-2). Given the differences observed in variance component estimations on the models including non-additive effects, we studied the sampling correlation among the variance component estimation to evaluate which of the 10 models was able to better partition the genetic variance (Appendix B). The correlation of variance components between additive/dominance with epistasis cannot be estimated under the pedigree models, as the estimation of the epistasis variance was zero (Appendix B). The distribution of the eigenvalues of the portion of the correlation matrices with relevance for genetics and breeding (additive, dominance, epistasis and error) was calculated (Figure 4-1) excluding additive-only models. As reference, the distribution of eigenvalues for a perfectly orthogonal correlation matrix (identity matrix) with all of them equal to 1 representing the ideal scenario (Figure 4-1a). The distribution of the eigenvalues is narrower for the matrix of correlation from model D_G_MM, outperforming the D_ped model (Figure 4-1b), as an example, the correlation between additive and dominance variance components decreased from 0.90 with the D_ped to 0.70 with the D_G_MM model (Appendix B). In general, all the marker-based models including epistasis outperformed their pedigree-based counterpart. Models

D_G#D_G_MM and A_G#D_G_MM showed the best performance (Figure 4-1c), with correlation values between additive and dominance/epistasis below 0.4 (Appendix B). In both cases, pedigree- and marker-based, the models including more than one epistasis matrix performed worse than simpler models. These eigenvalue distributions are only comparable for matrices of the same dimension. So the distributions cannot be used to compare between sections b with c in Figure 4-1.

Given the results above, we studied the standard error of the prediction (SEP) for models including dominance-by-dominance and additive-by-dominance for pedigree- and marker-based model. Figure 4-2 shows the SEP pattern found for the BV and DV for each of the 860 individual clones. SEP for BVs from the marker-based models was smaller than the pedigree based models in 99.8% of the cases (Figure 4-2a and c). For DV a clear advantage was observed for the marker- over the pedigree-based models with almost all SEP more than 40% worse in the pedigree based models (Figure 4-2b,d).

If a given model can estimate variance components free of noise (low correlation among effects) the prediction of BV using these estimations should be more accurate. Following this assumption, we tested with cross-validation the additive models (A_ped and A_G_MM) and models including dominance-by-dominance and additive-by-dominance for both pedigree- (D#D_ped and A#D_ped) and marker-based (D_G#D_G_MM and A_G#D_G_MM) models (Table 4-3). With a 10-fold cross validation, replacing the A matrix (model A_ped) by the A_G matrix (model A_G_MM) in the additive models increased BV prediction ability by 4%. However, in the pedigree-based models, inclusion of non-additive effects increases the BV prediction ability by 14% in the D#D_ped model and

13% in the A#D_ped model. Inclusion of non-additive effects in the marker-based models (A_G#D_G_MM and D_G#D_G_MM) increase BV prediction by 30% compared with the A_G_MM and by 36% when compared to A_ped. The Mean Square Error (MSE) decreased by 50% from the additive models (A_ped) to the more complex pedigree-based models (D#D_ped and A#D_ped) and it further decreases to approximately 10-12% of the value observed in the additive models (A_ped and A_G _MM) when the marker-based models (A_G#D_G_MM and D_G#D_G_MM) were used. Furthermore, we calculated the rank correlation between real ranking using all data in each model and predicted ranking from the 10-fold cross validation. The correlation of ranking position including all genotypes showed similar values as the correlation between BVs showed above in Table 4-3 (data not shown). However, in a breeding program it is not only necessary to predict the trend and magnitude of the complete set of selection candidates but also to accurately identify the top performers. Here, we emulated a selection of the top 10% and studied the correlation of the true ranking position, given by the model including all data, and the predicted position in the ranking using the model in a cross-validation scenario. The capacity to predict the top 10% doubled when the A matrix (0.17) was replaced by the A_G matrix (0.34) in the additive models (A_ped and A_G_MM) and further increased for the A_G#D_G_MM model to 0.37 (Table 4-3), doubling the correlations observed in the non-additive pedigree-based models. The correlation for the model D_G#D_G_MM was a little lower than the A_G_MM model; however, still 80% better than the best pedigree model.

Discussion

As expected, when the A matrix was replaced by the A_G matrix in the additive BLUP analysis, the heritability increased with the accuracy of the model (Hayes *et al.*

2009). Inclusion of non-additive effects in the pedigree based full model decreased the narrow-sense heritability by 26%. This result indicates that ignoring non-additive effects overestimated the additive variance, and therefore inflates the narrow-sense heritability (Lu *et al.* 1999; Zuk *et al.* 2012). In addition, when pedigree based models include non-additive effects, the conclusions do not vary from the general trend that non-additive effects are a small fraction of the total genetic variation for trees (Isik *et al.* 2003; Costa e Silva *et al.* 2004; Baltunis *et al.* 2007; Araujo *et al.* 2012). In contrast, marker-based models with additive and non-additive effects yield a totally different variance partitioning than their counterparts using the pedigree. The additive variance decreased as dominance was added to the model and it further decreases when dominance plus epistasis were considered in the models. These models indicate that non-additive effects are as important as additive effects and certainly larger than can be predicted with the pedigree models. Although the models that include non-additive effects differed in the partitioning of the variance components, they estimate a common non-additive variance component. Changes in the magnitude of variance components have already been observed when the relationship matrix derived from markers is used instead of the pedigree-derived relationship matrix (Habier *et al.* 2010).

The values for the model selection criteria (AIC) varied slightly for the best models, D_ped, D#D_ped, A#D_ped, A_G#D_G_MM and D_G#D_G_MM and thus no clear advantage of one over any other model can be declared. Consequently, we compared the capacity of the different models to partition the random effects estimates by using the sampling correlation among them. Under perfect orthogonality of genetic effects, the sampling correlation among the model effects will be closer to zero and eigenvalues of this

correlation matrix close to one (Figure 4-1a), indicating a perfect separation of effects (Hill 2010). In the additive plus dominance and additive plus dominance plus epistasis cases, models derived from molecular markers more efficiently partitioned the genetic effects than their counterpart pedigree based-models. The parameterizations of these paired models were identical except for the origin of the relationship matrices (pedigree- or marker-based). The poor capacity of pedigree models to partition is not surprising as all relationship matrices are derived from the additive relationship matrix derived from the pedigree (Mrode 2005) and so they are strongly correlated (Visscher 2009). The D_G#D_G_MM and A_G#D_G_MM models had the weakest correlations between additive and non-additive below 0.4, showing they partitioned effects substantially better than the best value of pedigree based models of 0.89 (Appendix B), similar to the value of 0.82 reported by Visscher (2009). These results support the finding that pedigree-derived models cannot separate the additive from non-additive effects as their results are comparable to those of additive models (Hill *et al.* 2008). On the other hand, the use of the matrix derived from markers has already been related to a better capacity of separate genetic from environmental effects (Habier *et al.* 2010) and thus we conclude that the use of these matrices also increase the capacity to separate additive from non-additive genetic effects.

We further studied the ability of a subset of models to predict the additive effect (BV) using a 10-fold cross validation strategy. We expected that 1) a better estimate of the additive effect should be more stable in a cross-validation scenario, 2) if non-additive effects exist and are significant, then models including non-additive effects should be able to predict better than additive-only models, and 3) the full model that

separates better additive from non-additive and environment effects should provide a better vector of BV to be predicted. We found that by replacing the A matrix with the A_G matrix in the additive models the prediction ability increases by 4% (Table 4-3), as expected for a quantitative trait of these characteristics (Hayes *et al.* 2009). Furthermore, the inclusion of non-additive relationship matrices (D_G and either $D_G\#D_G$ or $A_G\#D_G$) together with A_G yielded a BV predictive ability 36% larger than pedigree models and 30% larger than the additive model with the A_G matrix, which is the traditional GBLUP. In addition, the mean square error of the last models ($D_G\#D_G$_MM and $A_G\#D_G$_MM) decreased significantly more than 8-fold and 4-fold when compared with non-additive pedigree and additive marker models, respectively. Moreover, the smaller SEPs for marker-derived models (Figure 4-2) with the complete dataset also support these findings. These results indicate that, to increase substantially the accuracy to predict the BV for this trait, replacing the A matrix by the A_G matrix is not enough and non-additive effects need to be considered.

Overall, these results support the idea of Hill *et al.* (2008) that if non-additive effects exist they will be partially captured as additive effects. However, this does not mean that non-additive effects do not exist. We demonstrate that the marker-derived models better separated additive from non-additive effects and also yield better predictions of BV for tree height an important trait in tree breeding. We conclude that the use of relationship matrices derived from markers in a model including additive and non-additive effects had the best performance not only to partition the genetic variances but to improve considerably the breeding value prediction ability in trend, magnitude and top individual selection. Moreover, this study revealed that additive and epistatic effects are

of similar magnitude for height growth in *Pinus taeda,* a novel result that should promote new improvement strategies. Finally, using markers rather than pedigree derived matrices in linear mixed models with additive and non-additive effects provides a more accurate estimation of the variance components that change our understanding of the genetic control and genetic architecture of a quantitative trait inferred by the pedigree information.

Table 4-1. Summary of models, effects fitted and relationship matrices used in the study.

Model Number	Model	Effect fitted	Relationship matrix used
1	A_ped	a	A
7	A_G_MM	a	A_G
2	D_ped	a,d	A,D
8	D_G_MM	a,d	A_G,D_G
3	A#A_ped	a,d,aa	A,D,A#A
9	A_G#A_G_MM	a,d,aa	A_G,D_G,A_G#A_G
4	D#D_ped	a,d,dd	A,D,D#D
10	D_G#D_G_MM	a,d,dd	A_G,D_G,D_G#D_G
5	A#D_ped	a,d,ad	A,D,AD
11	A_G#D_G_MM	a,d,ad	A_G,D_G,A_G#D_G
6	Full_ped	a,d,aa,dd,ad	A,D,A#A,D#D,A#D
12	Full_MM	a,d,aa,dd,ad	A_G,D_G,A_G#A_G,D_G#D_G,A_G#D_G

Table 4-2.Variance estimation, genetic parameters (standard errors in parenthesis) and measure of data fitting.

	A_ped	A_G_MM	D_ped	D_G_MM	A#A_ped	A_G#A_G_MM	D#D_ped	D_G#D_G_MM	A#D_ped	A_G#D_G_MM	Full_ped	Full_MM
LogL	-1299.40	-1336.44	-1295.37	-1311.63	-1294.83	-1297.60	-1293.90	-1294.95	-1294.38	-1294.77	-1295.85	-1296.42
Number effects	4	4	6	6	8	8	8	8	8	8	8	9
AIC	2606.80	2680.88	2602.74	2635.26	2605.66	2611.20	2603.80	2605.90	2604.76	2605.54	2607.7	2610.84
Iblk	2512.10	2491.64	2513.88	2491.69	2514.21	2498.05	2513.72	2504.49	2514.05	2503.37	2513.83	2499.48
Additive (Add)	3682.82	4367.48	2599.18	2130.05	2577.58	559.65	2516.89	1327.21	2553.12	1105.90	2258.46	1051.46
Dominance (Dom)	-	-	622.84	1452.33	606.84	372.46	636.18	195.77	623.13	204.43	0.00275	197.18
Epistasis Add x Add	-	-	-	-	0.01	1868.51	-	-	-	-	883.27	569.48
Epistasis Dom x Dom	-	-	-	-	-	-	0.00	1231.60	-	-	0.00126	957.52
epistasis Add x Dom	-	-	-	-	-	-	-	-	0.00	1432.87	-	0.01
Culture x Add	200.95	138.18	115.88	146.63	0.00	146.03	92.04	80.29	91.83	59.06	112.064	145.25
Culture x Dom	-	-	127.51	4.76	0.00	16.32	0.00	0.00	0.00	0.00	135.65	18.85
culture x (Add x Add)	-	-	-	-	282.85	0.00	-	-	-	-	-	-
culture x (Dom x Dom)	-	-	-	-	-	-	213.40	189.52	-	-	-	-
culture x (Add x Dom)	-	-	-	-	-	-	-	-	196.84	214.24	-	-
Residual	5129.61	5263.04	5095.98	5198.93	5068.76	5149.97	5054.42	5073.86	5065.71	5075.45	5092.09	5142.39
Total Variance	9013.38	9768.70	8561.38	8932.70	8536.04	8112.94	8512.93	8098.25	8530.64	8091.95	8481.54	8082.12
h^2	0.409	0.447	0.304	0.239	0.302	0.069	0.296	0.164	0.299	0.137	0.266	0.130
SE(h^2)	(0.018)	(0.021)	(0.059)	(0.039)	(0.058)	(0.050)	(0.058)	(0.041)	(0.058)	(0.043)	(0.081)	(0.059)
d^2	na	na	0.073	0.163	0.071	0.046	0.075	0.024	0.073	0.025	0.000	0.024
SE(d^2)			(0.044)	(0.032)	(0.043)	(0.039)	(0.042)	(0.039)	(0.043)	(0.040)	(0.000)	(0.040)
i^2	na	na	na	na	0.000	0.230	0.000	0.152	0.000	0.177	0.104	0.189
SE(i^2)					(0.000)	(0.048)	(0.000)	(0.034)	(0.000)	(0.039)	(0.063)	(0.051)
H^2	0.409	0.447	0.376	0.401	0.373	0.345	0.370	0.340	0.372	0.339	0.370	0.343
SE(H^2)	(0.018)	(0.021)	(0.023)	(0.020)	(0.023)	(0.021)	(0.023)	(0.021)	(0.024)	(0.021)	(0.026)	(0.021)

Table 4-3. Predictive ability, Mean Square Error (MSE), top 10% ranking correlation (Top10%RankCor) and AIC for selected models.

Model	Cor(RBV,PBV)	MSE(RBV,PBV)	Top10%RankCor	AIC
A_ped	0.640	1335.800	0.17	2606.80
A_G_MM	0.670	1291.800	0.34	2680.88
A#D_ped	0.727	657.258	0.16	2604.76
A_G#D_G_MM	0.872	108.240	0.37	2605.54
D#D_ped	0.732	638.464	0.18	2603.80
D_G#D_G_MM	0.873	151.199	0.32	2605.90

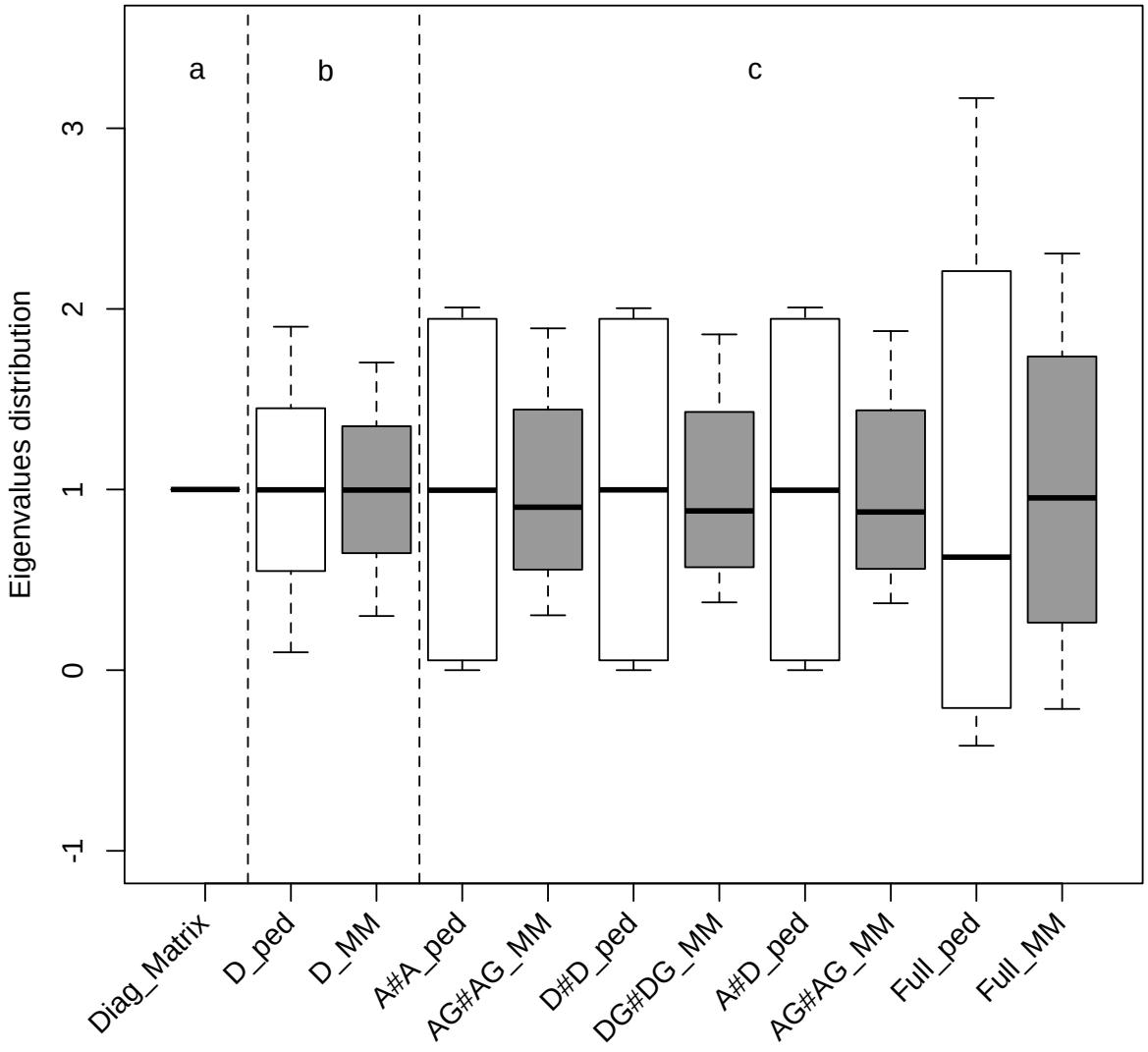

Figure 4-1. Eigenvalue distribution for a perfect orthogonal correlation matrix (a), for models including additive and dominance (b) and for models including additive, dominance and epistasis (c). White box for pedigree- and grey for marker-derived models.

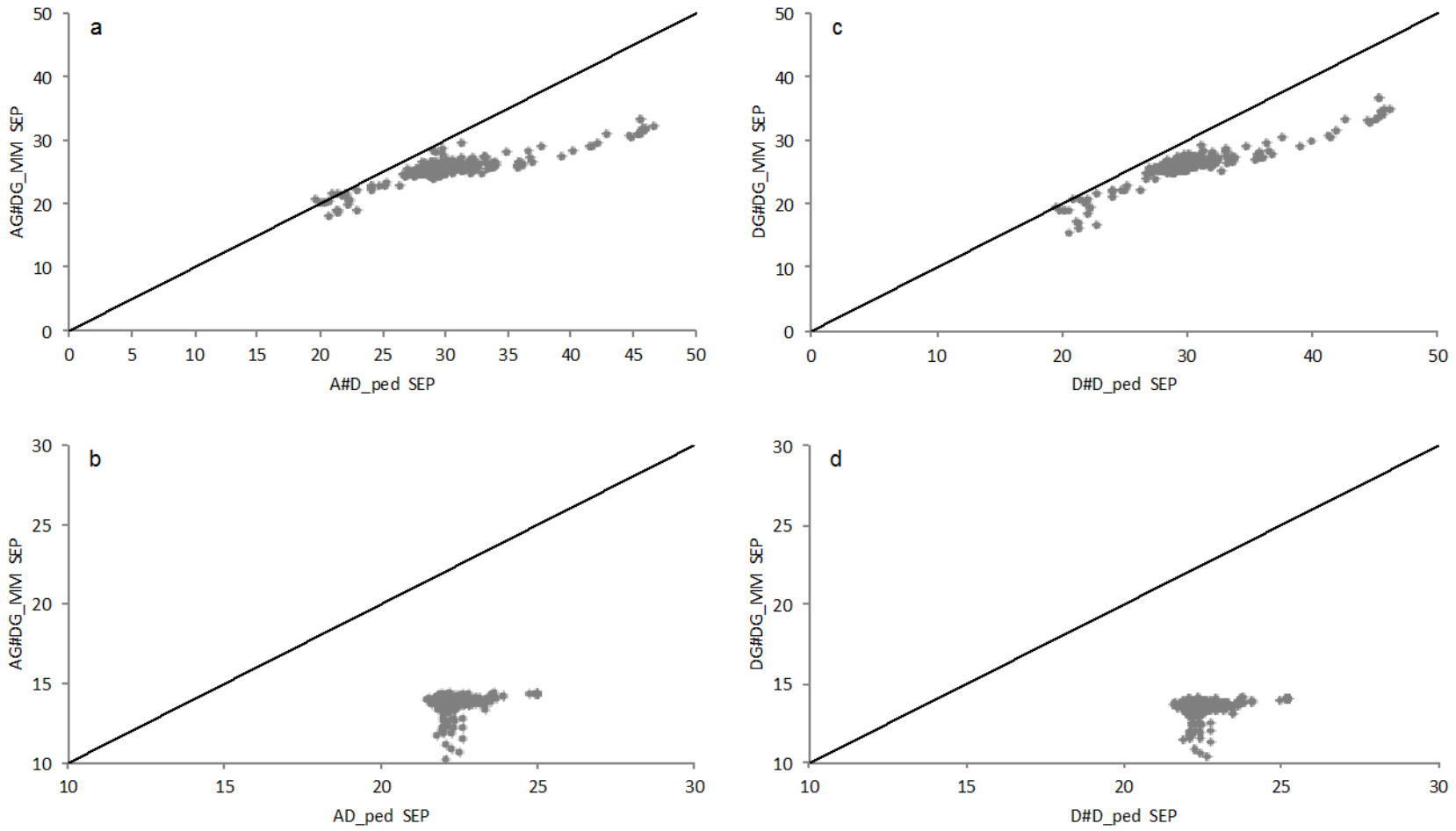

Figure 4-2. Standard error of the prediction (SEP) for pedigree-derived matrices model (X-axis) against their counterpart using marker-derived matrices (Y-axis). a) SEP for BV prediction model including A#D interaction. b) SEP for dominance value (DV) prediction model including A#D interaction. c) SEP for BV prediction model including D#D interaction. d) SEP for DV prediction model including D#D interaction.

CHAPTER 5
CONCLUSIONS

The rapidly decreasing cost of molecular markers (especially SNPs) and the current success of genomic selection in cattle breeding suggest that GS has the potential to become the norm in all breeding programs. In pine breeding, GS has the potential to increase genetic gain by dramatically decreasing the length of the breeding cycle. To apply GS, an important initial investment to obtain a dense panel of SNPs markers in a large number of individuals is required. While recently the use of this dense panel of markers has been studied extensively in breeding for prediction of additive effects, this research extended and maximized the use of molecular markers beyond prediction with direct and positive benefit to the breeding program.

If GS is planned in a breeding program, we strongly recommend constructing an additive-relationship matrix derived from the molecular markers. As shown in this study, this relationship matrix enables detection and correction of pedigree errors (recent and historical) by using the normality property of the different relationship classes. Most types of errors can be corrected by re-assigning individuals, parents or families to known pedigree in the records or by creating new parents for individuals that do not match any parent in the records (i.e. result of pollen or seed contamination). Independently of the stage or the kind of error, our results showed that correcting the pedigree increases the accuracy of the BLUP-BV prediction and dramatically increases the fit of the data. In addition, the corrected pedigree improved the accuracy of GS models in 13 of 15 traits studied, indicating that GS models more efficiently capture associations between markers and QTLs when the correct pedigree is used to estimate BLUP-BV. The use of the normality relationships derived from a large number of SNP

markers is envisioned to be complementary to the use of a small number of microsatellites markers for pedigree validation and error correction in the different levels of the breeding cycle (i.e. field-trial selections, seed orchards, clonal banks, greenhouse and production/deployment).

Following pedigree corrections in the breeding program, it is highly recommended, that the dense panel of markers should be used to perform GS with at least two analytic methodologies that contrast in the context of marker effect distribution. The methods RR-BLUP or GBLUP assume that most genes (markers in the model) will have some effect in the traits being study. This is a suitable assumption when considering an infinitesimal model, where the characters are determined by an infinite number of unlinked and non-epistatic loci, with small effect. In contrast, most Bayesian methods assume that only a portion of the genes will have an effect on the trait. This study shows that trait-specific methodologies will be needed, depending on the genetic architecture of the trait. In general, the methods evaluated here (RR-BLUP/RR-BLUP B, Bayes A, Bayesian LASSO regression and Bayes Cπ) differed only modestly in their predictive ability (defined by the correlation between the direct genetic value and deregressed phenotypes). Not surprisingly, BLUP-based methods underperformed relative to Bayesian approaches for oligogenic traits, as was the case for fusiform rust resistance. Bayes A and Bayes Cπ were more accurate than RR-BLUP in predicting fusiform rust resistance traits, likely reflecting a simpler genetic architecture, with a few loci of large effect involved in gene for gene resistance interactions. In addition to the published models, we developed RR-BLUP B, aimed at reducing the number of parameters, as an alternative to Bayesian and the traditional RR-BLUP approaches. For rust disease

resistance and wood density traits, RR-BLUP B approach performed better that traditional RR-BLUP, as well as the Bayesian methods. While RR-BLUP B may add an additional step to the development of predictive models (i.e. initial marker selection), it is overall simpler and computationally less expensive than Bayesian approaches. Therefore, it may provide a suitable alternative to the use of BLUP-based methods for traits that do not fit an infinitesimal model and are regulated by few major loci. The problem is that the genetic architecture of most traits in breeding programs is not known, thus based on the results of two contrasting GS methods in a cross-validation or directly in a validation population will help to elucidate this.

Finally, the molecular markers should be used to study the existence and magnitude of non-additive effects on the trait of interest. In this study, we showed that compared with pedigree-derived relationship matrices molecular marker-derived relationship matrices (additive and non-additive) in a linear mixed model better partition the genetic variance into additive and non-additive components. The result of this study supports the idea of Hill *et al.* (2008) that if non-additive effects exist additive effects will capture them. Which, does not imply that non-additive effects are non-existent. We showed that the use of marker-derived relationship matrices are able to extract more information from the same phenotypic dataset, due that these matrix better reflect the real additive and non-additive relationship among the individuals. For example height, an important trait in tree breeding, we conclude that the use of relationship matrices derived from markers in a model including additive and non-additive effects had the best performance not only to partition the genetic variances but improving considerably the breeding value prediction ability in trend, magnitude and top individual selection. This

study revealed that additive and non-additive epistatic interactions effects are of similar magnitude for height in *Pinus taeda,* a novel result that that can drive new improvement strategies for the species. In addition, using this model provides a more accurate estimation of the variance components that changed the view of the genetic control and architecture inferred by the pedigree information.

Overall, the results of this study support the idea that molecular markers can be used beyond prediction of breeding values in breeding populations. While the species *Pinus taeda* was used as a model here, our recommendations are not species specific and should be considered in animal and plant breeding programs. In summary, is strongly recommended for breeding programs that will be using genomic selection to use the molecular markers to: 1) confirm the pedigree and check for errors; 2) study the trait genetic architecture; 3) run at least two genomic selection methodologies that differ in the assumptions of number of genes controlling the trait (e.g. GBLUP, Bayes Cpi) to select the appropriate one for each trait, and; 4) predict non-additive effects, given the genetic architecture of the trait. In this way, maximum advantage can be obtained from the molecular markers. Some strategic scenarios should be considered depending on the trait's genetic architecture revealed by the markers. For example, if the trait shows a quantitative pattern with many genes (i.e. markers) controlling the trait, then GS should be considered. Otherwise, if the trait is controlled by major genes (oligogenic) then traditional marker assisted selection (MAS) should be considered together with GWAS, as chances for gene-discovery are higher under this case.

While the study of quantitative genetics is mature, the use of molecular markers in breeding is re-vitalizing the study in this area. A large number of studies have focused

on analyzing different aspects of quantitative genetics but there are still several important areas that need further improvement: genetic-by-environment interaction, genetic architecture of important traits, prediction of non-additive effects, strategies of improvement that better exploit the faster cycles times due to GS and addition of expression data to improve prediction models are some of these. Finally, the use of prediction models that incorporate GS with growth modeling and expression data is envisioned as a target to improve the prediction across areas of deployment and thus predict the genotype-by-environment in addition to genotype performance in pine breeding programs.

APPENDIX A
PREDICTIVE ABILITY, STANDARD ERRORS AND REGRESSION COEFFICIENTS FOR DIFFERENT GENOMIC SELECTION MODELS OF CHAPTER 2

Table A-1. Predictive ability ($r_{y\hat{y}}$) and standard error (SE) of RR-BLUP model under two different cross validation methods: 10-fold cross validation and leave one-out (L1-Out)

Trait Category	Trait	Methods			
		$\boldsymbol{r_{y\hat{y}}}$ - 10-Fold	SE - 10-Fold	$\boldsymbol{r_{y\hat{y}}}$ - L1-Out	SE - L1-Out
Growth	HT	0.39	0.029	0.38	0.029
Growth	HTLC	0.45	0.027	0.46	0.027
Growth	BHLC	0.49	0.026	0.49	0.026
Growth	DBH	0.46	0.027	0.46	0.027
Development	CWAL	0.48	0.029	0.48	0.030
Development	CWAC	0.38	0.026	0.40	0.027
Development	BD	0.27	0.032	0.27	0.032
Development	BA	0.51	0.025	0.52	0.025
Development	Rootnum_bin	0.28	0.030	0.28	0.030
Development	Rootnum	0.24	0.031	0.24	0.031
Disease resistance	Rust_bin	0.29	0.032	0.29	0.033
Disease resistance	Rust_gall_vol	0.23	0.033	0.24	0.033
Wood quality	StiffnessTree	0.43	0.027	0.43	0.028
Wood quality	Lignin	0.17	0.032	0.17	0.032
Wood quality	Latewood%4	0.24	0.031	0.25	0.031
Wood quality	Density	0.20	0.032	0.21	0.032
Wood quality	C5C6	0.26	0.031	0.27	0.031

Table A-2. Regression Beta and standard error (SE) of the RR-BLUP model with two different cross validation methods: 10-fold cross validation and leave one-out (L1-Out)

Trait Category	Trait	Methods			
		Beta - 10-Fold	SE - 10-Fold	Beta - L1-Out	SE - L1-Out
Growth	HT	1.18	0.10	1.12	0.09
Growth	HTLC	1.20	0.08	1.21	0.08
Growth	BHLC	1.11	0.07	1.11	0.07
Growth	DBH	1.19	0.08	1.17	0.08
Development	CWAL	1.04	0.06	1.03	0.06
Development	CWAC	1.10	0.09	1.13	0.09
Development	BD	1.23	0.15	1.19	0.15
Development	BA	1.13	0.07	1.13	0.06
Development	Rootnum_bin	1.36	0.15	1.31	0.15
Development	Rootnum	1.51	0.20	1.48	0.19
Disease resistance	Rust_bin	1.13	0.13	1.12	0.13
Disease resistance	Rust_gall_vol	1.29	0.19	1.23	0.18
Wood quality	StiffnessTree	1.12	0.08	1.10	0.08
Wood quality	Lignin	1.00	0.19	0.96	0.19
Wood quality	Latewood%4	1.01	0.14	1.03	0.13
Wood quality	Density	1.26	0.20	1.27	0.20
Wood quality	C5C6	1.22	0.15	1.23	0.15

Table A-3. Standard error of the prediction models for the different methods tested.

Trait Category	Trait	Methods			
		RR-BLUP	BLASSO	Bayes A	Bayes Cπ
Growth	HT	0.029	0.029	0.029	0.029
Growth	HTLC	0.027	0.027	0.028	0.028
Growth	BHLC	0.026	0.026	0.026	0.026
Growth	DBH	0.027	0.027	0.027	0.027
Development	CWAL	0.029	0.030	0.030	0.030
Development	CWAC	0.026	0.027	0.027	0.027
Development	BD	0.032	0.032	0.032	0.032
Development	BA	0.025	0.025	0.025	0.025
Development	Rootnum_bin	0.030	0.030	0.031	0.031
Development	Rootnum	0.031	0.031	0.031	0.031
Disease resistance	Rust_bin	0.032	0.033	0.031	0.031
Disease resistance	Rust_gall_vol	0.033	0.033	0.033	0.033
Wood quality	StiffnessTree	0.027	0.028	0.028	0.028
Wood quality	Lignin	0.032	0.032	0.032	0.032
Wood quality	Latewood%4	0.031	0.031	0.032	0.031
Wood quality	Density	0.032	0.032	0.032	0.032
Wood quality	C5C6	0.031	0.031	0.031	0.031

Table A-4. Accuracies of genomic selection models in 17 different traits of loblolly pine. h^2_m represents the Mendelian segregation and was the correction factor used to convert predictive abilities into accuracies.

Trait Category	Trait	h^2_m	Methods			
			RR-BLUP	BLASSO	Bayes A	Bayes Cπ
Growth	HT	0.66	0.48	0.47	0.47	0.47
Growth	HTLC	0.53	0.62	0.60	0.60	0.60
Growth	BHLC	0.52	0.68	0.68	0.68	0.68
Growth	DBH	0.66	0.57	0.57	0.57	0.57
Development	CWAL	0.43	0.58	0.55	0.55	0.55
Development	CWAC	0.63	0.60	0.58	0.59	0.59
Development	BD	0.26	0.53	0.49	0.53	0.53
Development	BA	0.50	0.72	0.72	0.72	0.72
Development	Rootnum_bin	0.50	0.40	0.40	0.38	0.40
Development	Rootnum	0.43	0.37	0.40	0.38	0.37
Disease resistance	Rust_bin	0.21	0.63	0.61	0.74	0.74
Disease resistance	Rust_gall_vol	0.18	0.57	0.57	0.66	0.68
Wood quality	StiffnessTree	0.37	0.71	0.64	0.69	0.69
Wood quality	Lignin	0.11	0.51	0.51	0.51	0.51
Wood quality	Latewood%4	0.17	0.58	0.58	0.56	0.58
Wood quality	Density	0.09	0.67	0.73	0.77	0.73
Wood quality	C5C6	0.14	0.69	0.67	0.67	0.67

Table A-5. Regression coefficients estimates of deregressed phenotypes regressed on Direct Genetic Values. Estimates for the model RR-BLUP are presented in the Supporting Material Table A-2.

Trait	BLASSO	SE (BLASSO)	Bayes A	SE (Bayes A)	BayesCπ	SE (Bayes Cπ)
HT	0.97	0.08	0.90	0.08	1.02	0.09
HTLC	0.98	0.07	0.93	0.06	1.00	0.07
BHLC	0.98	0.06	0.98	0.06	1.01	0.06
DBH	0.99	0.07	0.97	0.06	1.02	0.07
CWAL	0.90	0.08	0.87	0.08	0.95	0.08
CWAC	0.91	0.06	0.91	0.06	0.99	0.06
BD	0.97	0.13	0.97	0.12	1.08	0.13
BA	0.96	0.06	0.99	0.06	0.99	0.06
Rootnum_bin	1.01	0.11	0.85	0.10	1.04	0.12
Rootnum	1.01	0.13	1.04	0.14	1.05	0.14
Rust_bin	0.88	0.11	1.04	0.10	1.00	0.10
Rust_gall_vol	1.00	0.14	1.04	0.13	1.11	0.13
StiffnessTree	0.83	0.06	1.04	0.08	0.99	0.07
Lignin	1.01	0.20	0.87	0.17	1.11	0.21
Latewood%4	0.93	0.13	0.84	0.12	0.96	0.13
Density	0.95	0.14	0.99	0.14	1.04	0.15
C5C6	0.95	0.12	0.97	0.13	1.01	0.13

Table A-6. Predictive abilities of RR-BLUP_B when a reduced subset of markers was used compared to the predictive abilities of RR-BLUP

Trait	RR-BLUP	RR-BLUP_B	Subset length
HT	0.38	0.38	4630
HTLC	0.45	0.45	3910
BHLC	0.49	0.50	1380
DBH	0.46	0.46	3800
CWAL	0.38	0.39	820
CWAC	0.48	0.47	590
BD	0.27	0.28	1110
BA	0.51	0.52	1040
Rootnum_bin	0.28	0.28	2550
Rootnum	0.23	0.24	3350
Rust_bin	0.29	0.33	310
Rust_gall_vol	0.23	0.37	100
StiffnessTree	0.43	0.44	1300
Lignin	0.17	0.17	4240
Latewood%4	0.24	0.24	3820
Density	0.20	0.24	240
C5C6	0.25	0.25	4790

APPENDIX B
SAMPLING CORRELATION MATRICES FOR CHAPTER 4 MODELS

Table B-1. Sampling correlation matrix for all models tested. Above diagonal pedigree-based and below diagonal marker-based models. (A) additive models 1 and 7, (B) additive plus dominance models 2 and 8, (C) additive plus dominance plus additive-by-additive models 3 and 9, (D) additive plus dominance plus dominance-by-dominance models 4 and 10, (E) additive plus dominance plus additive-by-dominance models 5 and 11, and (F) full models 6 and 12. Effect estimated effects are I.block = incomplete block, Add=Additive, Dom=dominance and Cult=silviculture type.

A	I.Block	Add	CultxAdd	Residual
I.Block	1.00	0.00	0.01	-0.07
Add	0.00	1.00	-0.14	-0.08
CultxAdd	0.01	-0.08	1.00	-0.24
Residual	-0.07	-0.13	-0.19	1.00

B	I.Block	Add	Dom	CultxAdd	CultxDom	Residual
I.Block	1.00	0.00	0.00	0.00	0.01	-0.07
Add	0.00	1.00	-0.90	-0.08	0.05	-0.04
Dom	0.00	-0.70	1.00	0.09	-0.15	0.02
CultxAdd	0.00	-0.12	0.12	1.00	-0.62	0.02
CultxDom	0.01	0.09	-0.17	-0.69	1.00	-0.26
Residual	-0.07	-0.05	-0.02	0.00	-0.19	1.00

C	I.Block	Add	Dom	AddxAdd	CultxAdd	CultxDom	Cultx(AddxAdd)	Residual
I.Block	1.00	0.00	0.00	-0.07	-0.07	-0.07	0.01	-0.07
Add	0.00	1.00	-0.89	-0.02	-0.02	-0.02	-0.04	-0.02
Dom	-0.01	-0.19	1.00	0.01	0.01	0.01	-0.07	0.01
AddxAdd	0.00	-0.55	-0.56	1.00	1.00	1.00	-0.30	1.00
CultxAdd	0.00	-0.11	0.11	0.00	1.00	1.00	-0.30	1.00
CultxDom	0.01	0.07	-0.16	0.01	-0.69	1.00	-0.30	1.00
Cultx(AddxAdd)	-	-	-	-	-	-	1.00	-0.30
Residual	-0.07	0.03	0.03	-0.07	-0.01	-0.19	-	1.00

D	I.Block	Add	Dom	DomxDom	CultxAdd	CultxDom	Cultx(DomxDom)	Residual
I.Block	1.00	0.00	0.00	-0.07	0.01	-0.07	0.00	-0.07
Add	0.00	1.00	-0.89	-0.01	-0.02	-0.01	-0.05	-0.01
Dom	0.00	-0.35	1.00	0.01	0.04	0.01	-0.07	0.01
DomxDom	0.00	-0.29	-0.62	1.00	0.02	1.00	-0.31	1.00
CultxAdd	0.00	-0.09	0.00	0.10	1.00	0.02	-0.48	0.02
CultxDom	-0.07	0.00	0.01	-0.02	0.04	1.00	-0.31	1.00
Cultx(DomxDom)	0.01	0.04	0.00	-0.18	-0.52	-0.31	1.00	-0.31
Residual	-0.07	0.00	0.01	-0.02	0.04	1.00	-0.31	1.00

E	I.Block	Add	Dom	AddxDom	CultxAdd	CultxDom	Cultx(AddxDom)	Residual
I.Block	1.00	0.00	0.00	-0.07	0.01	-0.07	0.00	-0.07
Add	0.00	1.00	-0.89	-0.02	-0.03	-0.02	-0.02	-0.02
Dom	0.00	-0.29	1.00	0.01	0.06	0.01	-0.10	0.01
AddxDom	0.00	-0.38	-0.62	1.00	0.03	1.00	-0.30	1.00
CultxAdd	0.00	-0.10	0.00	0.11	1.00	0.03	-0.54	0.03
CultxDom	-0.07	0.00	0.01	-0.02	0.08	1.00	-0.30	1.00
Cultx(AddxDom)	0.01	0.05	0.00	-0.17	-0.62	-0.31	1.00	-0.30
Residual	-0.07	0.00	0.01	-0.02	0.08	1.00	-0.31	1.00

F	I.Block	Add	Dom	AddxAdd	DomxDom	AddxDom	CultxAdd	CultxDom	Residual
I.Block	1.00	0.00	-0.07	0.00	-0.07	-	0.00	0.01	-0.07
Add	0.00	1.00	-0.03	-0.94	-0.03	-	-0.09	0.10	-0.03
Dom	-0.01	-0.29	1.00	0.02	1.00	-	0.02	-0.26	1.00
AddxAdd	0.00	-0.68	0.02	1.00	0.02	-	0.11	-0.17	0.02
DomxDom	0.00	0.52	-0.30	-0.90	1.00	-	0.02	-0.26	1.00
AddxDom	-0.07	0.00	0.04	0.01	-0.04	1.00	-	-	-
CultxAdd	0.00	-0.10	0.11	0.00	0.00	-0.01	1.00	-0.62	0.02
CultxDom	0.01	0.07	-0.16	0.00	0.01	-0.19	-0.69	1.00	-0.26
Residual	-0.07	0.00	0.04	0.01	-0.04	1.00	-0.01	-0.19	1.00

LIST OF REFERENCES

Adams, W., D. Neale and C. Loopstra, 1988 Verifying controlled crosses in conifer tree-improvement programs. Silvae Genetica 37: 147-152.

Baltunis, B., D. Huber, T. White, B. Goldfarb and H. Stelzer, 2007a Genetic analysis of early field growth of loblolly pine clones and seedlings from the same full-sib families. Canadian Journal of Forest Research-Revue Canadienne De Recherche Forestiere 37: 195-205.

Baltunis, B., D. Huber, T. White, B. Goldfarb and H. Stelzer, 2007b Genetic gain from selection for rooting ability and early growth in vegetatively propagated clones of loblolly pine. Tree Genetics & Genomes 3: 227-238.

Baltunis, B. S., D. A. Huber, T. L. White, B. Goldfarb and H. E. Stelzer, 2005 Genetic effects of rooting loblolly pine stem cuttings from a partial diallel mating design. Canadian Journal of Forest Research-Revue Canadienne De Recherche Forestiere 35: 1098-1108.

Banos, G., G. R. Wiggans and R. L. Powell, 2001 Impact of paternity errors in cow identification on genetic evaluations and international comparisons. J Dairy Sci 84: 2523-2529.

Bennewitz, J., N. Reinsch and E. Kalm, 2002 Gencheck: A program for consistency checking and derivation of genotypes at co-dominant and dominant loci. J. Anim. Breed. Genet. 119: 350-360.

Bernardo, R., 2008 Molecular markers and selection for complex traits in plants: Learning from the last 20 years. Crop Science 48: 1649-1664.

Bramlett, D., 1997 Genetic gain from mass controlled pollination and top working. Journal of Forestry 95: 15-19.

Buckler, E., J. Holland, P. Bradbury, C. Acharya, P. Brown *et al.*, 2009 The genetic architecture of maize flowering time. Science 325: 714-718.

Collard, B., and D. Mackill, 2008 Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. Philosophical Transactions of the Royal Society B-Biological Sciences 363: 557-572.

Crossa, J., G. E. L. Campos, P. Pérez, D. Gianola, J. Burgueño *et al.*, 2010 Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186: 713-724.

De Los Campos, G., D. Gianola and G. J. M. Rosa, 2009a Reproducing kernel Hilbert spaces regression: A general framework for genetic evaluation. Journal of Animal Science 87: 1883-1887.

De Los Campos, G., H. Naya, D. Gianola, J. Crossa, A. Legarra *et al.*, 2009b Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics 182: 375-385.

Dekkers, J. C., 2004 Commercial application of marker- and gene-assisted selection in livestock: strategies and lessons. J Anim Sci 82 E-Suppl: E313-328.

Doerksen, T., and C. Herbinger, 2010 Impact of reconstructed pedigrees on progeny-test breeding values in red spruce. Tree Genetics & Genomes 6: 591-600.

Eckert, A. J., J. Van Heerwaarden, J. L. Wegrzyn, C. D. Nelson, J. Ross-Ibarra *et al.*, 2010 Patterns of population structure and environmental associations to aridity across the range of Loblolly pine (Pinus taeda L., Pinaceae). Genetics 185: 969-982.

Emhart, V. I., T. A. Martin, T. L. White and D. A. Huber, 2007 Clonal variation in crown structure, absorbed photosynthetically active radiation and growth of loblolly pine and slash pine. Tree Physiol 27: 421-430.

Ericsson, T., 1999 The effect of pedigree error by misidentification of individual trees on genetic evaluation of a full-sib experiment. Silvae Genetica 48: 239-242.

FAO, 2002 World agriculture: toward 2015/2030. Summary report, Food and Agriculture Organization of the United Nations, Rome, Italy.

Fernando, R., and D. Garrick, 2008 GenSel - User manual for a portfolio of genomic selection related analyses., pp. Animal Breeding and Genetics, Iowa State University, Ames.

Fisher, R., 1918 The correlation between relatives on the supposition of Mendelian inheritance. Philosophical Transactions of the Royal Society of Edinburgh 52: 399-433.

Garrick, D. J., J. F. Taylor and R. L. Fernando, 2009 Deregressing estimated breeding values and weighting information for genomic regression analyses. Genet Sel Evol 41: 55.

Geldermann, H., U. Pieper and W. Weber, 1986 Effect of misidentification on the estimation of breeding value and heritability in cattle. J. Dairy Sci. 63: 1759-1768.

Gianola, D., R. L. Fernando and A. Stella, 2006 Genomic-assisted prediction of genetic value with semiparametric procedures. Genetics 173: 1761-1776.

Gilmour, A., B. Gogel, B. Cullis and R. Thompsom, 2009 ASReml User guide, release 3.0. VSN International Ltd, Hemel Hempstead, HP1 1ES, UK.

Gilmour, A., B. Gogel, B. Cullis and R. Thompsom, 2006 ASReml User guide, release 2.0, VSN International Ltd, Hemel Hempstead, UK.

Goddard, M., 2009c Genomic selection: prediction of accuracy and maximisation of long term response. Genetica 136: 245-257.

Goddard, M. E., and B. J. Hayes, 2009b Mapping genes for complex traits in domestic animals and their use in breeding programs. Nat Rev Genet 10: 381-391.

Goddard, M. E., N. R. Wray, K. Verbyla and P. M. Visscher, 2009a Estimating effects and making predictions from genome-wide marker data. Statistical Science 24: 517-529.

Grattapaglia, D., and M. Resende, 2011 Genomic selection in forest tree breeding. Tree Genetics & Genomes 7: 241-255.

Habier, D., R. L. Fernando and J. C. M. Dekkers, 2009 Genomic selection using low-density marker panels. Genetics 182: 343-353.

Habier, D., R. L. Fernando, K. Kizilkaya and D. J. Garrick, 2011 Extension of the bayesian alphabet for genomic selection. BMC Bioinformatics 12:186.

Habier, D., J. Tetens, F. R. Seefried, P. Lichtner and G. Thaller, 2010 The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genetics, Selection, Evolution 42:5.

Hamblin, M. T., E. S. Buckler and J. L. Jannink, 2011 Population genetics of genomics-based crop improvement methods. Trends Genet 27: 98-106.

Hayes, B., and M. Goddard, 2010 Genome-wide association and genomic selection in animal breeding. Genome 53: 876-883.

Hayes, B. J., 2011 Efficient parentage assignment and pedigree reconstruction with dense single nucleotide polymorphism data. J Dairy Sci 94: 2114-2117.

Hayes, B. J., P. J. Bowman, A. C. Chamberlain, K. Verbyla and M. E. Goddard, 2009a Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genetics Selection Evolution 41:51.

Hayes, B. J., P. J. Bowman, A. J. Chamberlain and M. E. Goddard, 2009b Invited review: Genomic selection in dairy cattle: progress and challenges. J Dairy Sci 92: 433-443.

Heffner, E. L., A. J. Lorenz, J.-L. Jannink and M. E. Sorrells, 2010 Plant breeding with genomic selection: Gain per unit time and cost. Crop Science 50: 1681-1690.

Heslot, N., H.-P. Yang, M. E. Sorrells and J.-L. Jannink, 2012 Genomic selection in plant breeding: A comparison of models. Crop Science 52: 146-160.

IPCC, 2007 Climate change 2007: the physical science basis. Contribution of Working Group I to the fourth assessment report of the Intergovernmental Panel on Climate Change, pp., edited by S. SOLOMON, D. QIN, M. MANNING, Z. CHEN, M. MARQUIS et al. Cambridge University Press, Cambridge, UK & New York, NY, USA.

Israel, C., and J. I. Weller, 2000 Effect of misidentification on genetic gain and estimation of breeding value in dairy cattle populations. J Dairy Sci 83: 181-187.

Iwata, H., J.-L. Jannink, 2011 Accuracy of genomic selection prediction in barley breeding programs: A simulation study based on the real single nucleotide polymorphism data of barley breeding lines. Crop Science 51: 1915-1927.

Jannink, J., A. Lorenz and H. Iwata, 2010 Genomic selection in plant breeding: from theory to practice. Briefings in Functional Genomics 9: 166-177.

Jannink, J. L., A. J. Lorenz and H. Iwata, 2010a Genomic selection in plant breeding: from theory to practice. Brief Funct Genomics 9: 166-177.

Kayihan, G., C. Nelson, D. Huber, H. Amerson, T. White *et al.*, 2010 Clonal evaluation for fusiform rust disease resistance: effects of pathogen virulence and disease escape. Canadian Journal of Forest Research-Revue Canadienne De Recherche Forestiere 40: 1042-1050.

Kayihan, G. C., D. A. Huber, A. M. Morse, T. L. White and J. M. Davis, 2005 Genetic dissection of fusiform rust and pitch canker disease traits in loblolly pine. Theor Appl Genet 110: 948-958.

Kohavi, R., 1995 The power of decision tables. Machine Learning: Ecml-95 912: 174-189.

Lambeth, C., 1980 Juvenile-mature correlations in pinaceae and implications for early selection. Forest Science 26: 571-580.

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

Legarra, A., A. Ricarti and O. Filangi, 2011a Genomic selection, Gibbs sampling, Gauss-Seidel (and BayesCp), pp., http://snp.toulouse.inra.fr/~alegarra/.

Legarra, A., C. Robert-Granie, P. Croiseau, F. Guillaume and S. Fritz, 2011b Improved Lasso for genomic selection. Genetics Research 93: 77-87.

Legarra, A., C. Robert-Granie, E. Manfredi and J. Elsen, 2008 Performance of genomic selection in mice. Genetics 180: 611-618.

Li, X., D. Huber, G. Powell, T. White and G. Peter, 2007 Breeding for improved growth and juvenile corewood stiffness in slash pine. Canadian Journal of Forest Research-Revue Canadienne De Recherche Forestiere 37: 1886-1893.

Lynch, M., and B. Walsh, 1998 Genetics and analysis of quantitative traits. Sinauer Associates, Inc. Sunderland, MA.

Meuwissen, T., T. Solberg, R. Shepherd and J. Woolliams, 2009 A fast algorithm for BayesB type of prediction of genome-wide estimates of genetic value. Genetics Selection Evolution 41:2.

Meuwissen, T. H., B. J. Hayes and M. E. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. Genetics 157: 1819-1829.

Mrode, R. A., 2005 Linear models for the prediction of animal breeding values 2nd ed. CABI Publishing Company: Cambridge.

Munoz, P., M. Resende Jr, D. Huber, T. Quezasa, M. Resende *et al.*, Dense panel of markers for correcting pedigree errors in breeding populations: impact on heritability, breeding value and genomic selection accuracy. Genetic Selection Evolution. (*Submitted*)

Neale, D., and A. Kremer, 2011 Forest tree genomics: growing resources and applications. Nature Reviews Genetics 12: 111-122.

Powell, J. E., P. M. Visscher and M. E. Goddard, 2010 Reconciling the analysis of IBD and IBS in complex trait studies. Nature Reviews Genetics 11: 800-805.

Resende, M., M. Resende Jr, A. Aguiar, J. Abad, A. Missiaggia *et al.*, 2010 Comptacao da selecao Genomica Ampla (GWS), pp. 81, edited by E. Florestas. Embrapa, Colombo, Brazil.

Resende Jr, M. F. R., P. Munoz, J. J. Acosta, G. F. Peter, J. M. Davis *et al., 2012*a Accelerating the domestication of trees using genomic selection: accuracy of prediction models across ages and environments (vol 193, 617, 2012). New Phytologist 193: 1099-1099.

Resende, M. F. R., Jr., P. Munoz, M. D. V. Resende, D. J. Garrick, R. L. Fernando *Et Al.*, 2012b Accuracy of genomic selection methods in a standard data set of loblolly pine (Pinus taeda L.). Genetics 190: 1503-1510.

Sanders, K., J. Bennewitz and E. Kalm, 2006 Wrong and missing sire information affects genetic gain in the angeln Dairy Cattle Population J. Dairy Sci. 89: 7.

Schaeffer, L., 2006 Strategy for applying genome-wide selection in dairy cattle. Journal of Animal Breeding and Genetics 123: 218-223.

Simeone, R., I. Misztal, I. Aguilar and A. Legarra, 2011 Evaluation of the utility of diagonal elements of the genomic relationship matrix as a diagnostic tool to detect mislabelled genotyped animals in a broiler chicken population. J Anim Breed Genet 128: 386-393.

Soller, M., 1978 Use of loci associated with quantitative effects in dairy-cattle improvement. Animal Production 27: 133-139.

Sorensen, D., and D. Gianola, 2002 Likelihood, Bayesian and MCMC methods in quantitative genetics. Springer-Verlag, New York.

Stuber, C., M. Goodman and R. Moll, 1982 Improvement of yield and ear number resulting from selection at allozyme loci in a maize population. Crop Science 22: 737-740.

Sykes, R., M. Yung, E. Novaes, M. Kirst, G. Peter *et al.*, 2010 High-Throughput Screening of Plant Cell-Wall Composition Using Pyrolysis Molecular Beam Mass Spectroscopy in *Biofuels: Methods in Molecular Biology*, edited by J. Mielenz. Humana Press, New York.

Usai, M., M. Goddard and B. Hayes, 2009 LASSO with cross-validation for genomic selection. Genetics Research 91: 427-436.

Verbyla, K., M. Calus, H. Mulder, Y. De Haas and R. Veerkamp, 2010 Predicting energy balance for dairy cows using high-density single nucleotide polymorphism information. Journal of Dairy Science 93: 2757-2764.

Verbyla, K. L., B. J. Hayes, P. J. Bowman and M. E. Goddard, 2009 Accuracy of genomic selection using stochastic search variable selection in Australian Holstein Friesian dairy cattle. Genet Res (Camb) 91: 307-311.

Visscher, P., 2008 Sizing up human height variation. Nature Genetics 40: 489-490.

Visscher, P. M., 2009 Whole genome approaches to quantitative genetics. Genetica 136: 351-358.

Visscher, P. M., J. A. Woolliams, D. Smith and J. L. Williams, 2002 Estimation of pedigree errors in the UK dairy population using microsatellite markers and the impact on selection. J Dairy Sci 85: 2368-2375.

White, T., and M. Carson, 2004 Breeding program of conifers, *In* Walter, C and Carson, M (edts) 2004. Plantation Forest Biotechnology for the 21st Century. Research Signpost, Kerala, India.

White, T. L., W. T. Adams and D. B. Neale, 2007a Forest genetics. CABI Publishing: Cambridge, MA, USA.

Wiggans, G. R., P. M. Vanraden, L. R. Bacheller, M. E. Tooker, J. L. Hutchison *et al.*, 2010 Selection and management of DNA markers for use in genomic evaluation. J Dairy Sci 93: 2287-2292.

Wiggans, G. R., P. M. Vanraden and T. A. Cooper, 2011 The genomic evaluation system in the United States: past, present, future. J Dairy Sci 94: 3202-3211.

Wilcox, P., H. Amerson, E. Kuhlman, B. Liu, D. Omalley *et al.*, 1996 Detection of a major gene for resistance to fusiform rust disease in loblolly pine by genomic mapping. Proceedings of the National Academy of Sciences of the United States of America 93: 3859-3864.

Williams, E. R., A. C. Matheson and C. E. Harwood, 2002 Experimental design and analysis for tree improvement 2nd ed. Commonwealth Scientific and Industrial Research Organization, Melbourne, Australia.

Wong, C., and R. Bernardo, 2008 Genome-wide selection in oil palm: increasing selection gain per unit time and cost with small populations. Theoretical and Applied Genetics 116: 815-824.

Yang, J., B. Benyamin, B. P. Mcevoy, S. Gordon, A. K. Henders *et al.*, 2010 Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42: 565-569.

Zhang, Z., X. Ding, J. Liu, Q. Zhang and D. J. De Koning, 2011 Accuracy of genomic prediction using low-density marker panels. J Dairy Sci 94: 3642-3650.

Zhang, Z., J. Liu, X. Ding, P. Bijma, D. J. De Koning *Et al.*, 2010 Best linear unbiased prediction of genomic breeding values using a trait-specific marker-derived relationship matrix. PLOS One 5(9): e12648.

BIOGRAPHICAL SKETCH

Patricio R. Munoz was born in Santiago, Chile. His parents are Margarita Del Valle and Luis Munoz. He is the 7th of 8 siblings. He lived his childhood in Pirque, a town located at the foot of the Andes mountains. He then moved to the south of Chile where he obtained the title of Forestry Engineer from the Universidad Catolica de Temuco. He worked for Forestal Mininco (CMPC Forestal) for about two years as a breeder assistant and data analyst in the tree improvement program. He then moved to the University of Florida (Gainesville) where he obtained a Master of Science on quantitative genetics at the School of Forestry Resources and Conservation (SFRC). Curious on how to better integrate traditional breeding with molecular information, he decided to pursue a PhD in the Plant Molecular and Cellular Biology program (PMCB) at the University of Florida.