ABSTRACT

OGUT, FUNDA. Predictions of Genetic Merit in Tree Breeding Using Factor Analytic Linear Mixed Models and Blended Genomic Relationship Matrices. (Under the direction of Dr. Fikret Isik and Dr. Ross Whetten).

Increase in computer power and efficiency in DNA sequencing technologies is providing new opportunities to plant and animal breeders to fit more complex statistical models for predictions of genetic merit of individuals. Such models can be powerful to account for heterogeneity in the data and as a result can increase the accuracy of predictions and genetic gains from breeding programs.

In this study, I first evaluated the efficiency of factor analytic (FA) linear mixed models for a large, multi-environmental trial of loblolly pine (*Pinus taeda* L.). Height was assessed on 37,269 trees at age six years in a diallel experiment. Among models fitted, FA models produced the smallest AIC model fit statistics. FA models captured both the variance and covariance at the genetic level better than models with simpler covariance structures, and they provided more accurate predictions of genotypes. The mean narrow-sense heritability estimates for height was about 0.20 when more complex variance structures were used, compared to 0.13 when simpler variance structures were employed. FA models were parsimonious compared to US structures. The FA models provided a natural framework for modeling genotype by environment interaction. Genotype by environment interaction was non-significant as suggested by high genetic correlations both for additive (0.83) and dominance (0.91) effects.

Molecular marker data, especially single-nucleotide polymorphic (SNP) markers have been commonly used to predict genetic merit in animal breeding. However, marker data probably have missing genotypes and they need to be imputed. The effects of percent (level) and pattern (random or structured) of missing data, and mating designs on the accuracy imputation of genotypes were investigated. I used linkage based BLUP to impute missing genotypes for an empirical (unbalanced) data set for loblolly pine. For simulated (balanced) data sets, both BLUP and Hidden Markov Model (HMM) approaches were used. The actual data had 178 clones that were genotyped at 3,461 biallelic SNP markers. The simulated data consist of double-pair and half-diallel mating design with 2880 and 2940 individuals, respectively. For empirical data, accuracy of imputation was higher for the structured pattern of missing data at any level of missing percentages. Regardless of the pattern of the missing data, imputation accuracy was less than 0.70 when the data had greater than 40% missing values. For the simulated data, the imputation accuracy was not affected by mating design for the BLUP approach when the pattern of the missing data was structured. For HMM approach, when the pattern of the missing data was evenly-spaced, the mating design had no effect on the imputation accuracy.

Combining information from pedigree and DNA markers might improve prediction accuracies in tree breeding programs. In the third chapter, a cloned population of loblolly pine and simulated data sets were used to examine the efficiency of blended additive genetic relationships and realized genomic relationships were examined. Cloned 166 individuals were genotyped at 3,461 SNP markers out of a total 354 clones. For simulated data 300 or 600 trees genotyped at 1000 markers out of 1200. For the empirical data, the accuracy of

predictions based on the ABLUP was 0.79. Predictions based on HBLUP had accuracy of 0.72 to 0.76, depending on the genomic relationships used. For the simulated data set, the accuracy values for HBLUP models were higher compared to ABLUP model. Also, as the genotyped population size increased the accuracies increased. HBLUP uses all the available phenotype, pedigree, and genotype data in a single-step and is easy to implement for genomic based selection. The major advantage of using genomic relationships matrices is that markers can capture the Mendelian sampling effect within full-sib families for selection.

Predictions of Genetic Merit in Tree Breeding Using Factor Analytic Linear Mixed Models and Blended Genomic Relationship Matrices

by
Funda Ogut

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Forestry and Environmental Resources

Raleigh, North Carolina

2012

APPROVED BY:

_____	_____
Dr. Fikret Isik Committee Co-Chair	Dr. Ross Whetten Committee Co-Chair
_____	_____
Dr. Steven McKeand	Dr. Christian Maltecca

BIOGRAPHY

Funda Ogut was born in December of 1979 to Asaf and Birsen Ogut in a home of two sisters and two brothers in Bayburt, Turkey. Funda received her B.S. degree in 2002 and her M.S. degree 2006 in Forest Engineering from Karadeniz Technical University. In 2007, she was awarded a full PhD scholarship by the Ministry of National Education of Republic of Turkey to study in the United States. In 2008, she started her PhD in the Department of Forestry and Environmental Resources at North Carolina State University, Raleigh, NC.

ACKNOWLEDGMENTS

I would like to thank my advisors Dr. Fikret Isik and Dr. Ross Whetten for all their invaluable guidance and expertise. Thank you for always being kind and patient whenever I had questions and I had many of them! I am very thankful to my committee member Dr. Steven McKeand for many edits and revisions he made to improve my writing and for his invaluable support and guidance. My committee member Dr. Christian Maltecca was always a source of great ideas and was very helpful and patient with my data analyses.

I also would like to thank North Carolina State University Cooperative Tree Improvement for supporting my study. The Cooperative Program staff and graduate students were always supportive. To Saul Garcia, thank you for all the help with the computer problems and being a good friend. I am particularly thankful to Plum Creek Timber Company, especially Paul Belonger, for providing a valuable data set from their diallel experiments.

To Hicran Koc, thank you for being a good roomie and friend on this journey that we started together. Last but not least, I would like to thank my mom and dad for their unconditional love and support and my siblings for their encouragement and support.

TABLE OF CONTENTS

LIST OF TABLES vii
LIST OF FIGURES x
CHAPTER 1 1
Genetic Analysis of Diallel Progeny Test Data in *Pinus taeda* Using Factor Analytic Linear Mixed Model 1
Abstract 2
Introduction 3
Materials and Methods 11
Genetic material and mating design 11
Field design 11
Statistical Analysis 12
Modeling **B** and **R** variance structures 15
Modeling **G** variance-covariance structure 17
Model fit statistics (model comparison) 22
Estimation of genetic parameters 23
Results 25
Model comparison 25
Genetic parameter estimates 26
Genotype by Environment interactions 27
Discussion 31
Model comparison 31
Percent variance explained by factors in FA models 33
Modeling GxE using FA models 33
Genetic parameter estimates 35
Conclusion 37
References 38
Appendices 63
CHAPTER 2 82
Effects of Patterns of Missing Data and Mating Design on the Genotype Imputation Accuracy 82
Abstract 83
Introduction 84
Materials and Methods 88

Data 88
Imputation with BLUP approach using pedigree 88
Imputation with BEAGLE (parent-offspring trios) 91
Comparison of Models 92
Results 93
Results from BLUP approach 93
Results from HMM approach 94
Computational demands: 95
Discussion 96
Conclusion 100
References 101
Appendices 106
CHAPTER 3 108
Genetic Prediction Using Pedigree and Genomic Relationships in a Cloned Population of Loblolly Pine 108
Abstract 109
Introduction 110
Materials and Methods 114
Numerator relationship matrix 114
Genomic relationship matrix 114
Joint pedigree and genomic relationship matrix 115
Data 119
Statistical Analysis 121
Leave one out cross validation (LOOCV) 122
Comparison of Models 122
Results 124
Results from empirical data 124
Descriptive statistics of pedigree based ($\mathbf{A}_{22}$) and genomic based ($\mathbf{G}$) relationship matrices 124
Variance components 124
Comparison of Models 125
Results from Simulated Data 127

Descriptive statistics of pedigree based (A22) and genomic based (G) relationship matrices 127
Variance components 127
Comparison of Models 128
Discussion 130
Variances and the effect of scaling G matrices 132
Comparison of Models 133
Capturing the Mendelian sampling term 134
Conclusion 135
References 136
Appendices 153

LIST OF TABLES

CHAPTER 1

Table 1. Site IDs, locations and establishment year of the progeny tests, number of crosses (full-sib families), and number of checklots per site. 47

Table 2. Models and variance-covariance structures fit to GCA and SCA genetic effects and block effects (B) and error (R) for height. 48

Table 3. Model fit statistics REML log likelihood (LogL) and Akaike Information Criterion (AIC) when block within site and error variances were modeled as IID for height. When errors are assumed to have simple homogenous structure (all trees have the same identical residuals), models 4 and 5 with the FA1 and XFA1 genetic structures, respectively, were superior with the lowest AIC. 49

Table 4. Model fit statistics LogL and AIC when block and error variances were modeled heterogeneous (Diagonal). Among the models, Model 8 with XFA2 for GCA and XFA1 for SCA genetic effects was superior. 50

Table 5. Estimated model parameters fit to GCA and SCA effects, respectively for Model 8 (XFA2 for GCA and XFA1 for SCA). $\boldsymbol{\sigma a2}$ is GCA variance, $\boldsymbol{\Lambda a1}$ and $\boldsymbol{\Lambda a2}$ are first and second site loadings for the GCA effect, $\boldsymbol{\Psi a}$ is GCA variance not explained by site loadings (lack of fit), $\boldsymbol{\sigma d2}$ is SCA variance, $\boldsymbol{\Lambda d}$ is site loadings for the SCA effect, $\boldsymbol{\Psi d}$ is SCA variance not explained by site loadings and % is the percentage of variance explained by the loadings of first and second factor. 51

Table 6. Individual tree narrow sense heritabilities and standard errors within the parenthesis for the range of models fit to GCA and SCA genetic effects for all 10 progeny test sites for height. The XFA2/XFA1 and US variance-covariance structures gave similar heritability results. 52

Table 7. Individual tree broad sense heritabilities and standard errors within the parenthesis for the range of models fit to genetic effects GCA and SCA. 53

Table 8. Genetic correlations of GCA genetic effects between pair of progeny test sites from the XFA2 structure (above diagonal) based on model 8 and the US model (model 10) (below diagonal) for height. XFA2 and US variance-covariance structures fit to GCA genetic effects

gave similar genetic correlations between pair of sites. For the visual display of the same table see Appendix 2. The values in the diagonal are the average additive genetic correlations for environments based on FA model. 54

Table 9. Genetic correlations of SCA genetic effects between pair of progeny test sites from the XFA1 structure (above diagonal) based on Model 8 and US model (Model 10) (below diagonal). XFA1 and US variance-covariance structures fit to SCA genetic effects gave similar genetic correlations between pair of sites. For the visual display of the same table see Appendix 2. 55

CHAPTER 2

Table 1. Effect of having parental genotypes on the imputation accuracy (r) using simulated data generated as half-diallel mating design when the sampling is random and structured. 107

CHAPTER 3

Table 1. Descriptive statistics of coefficients of relationships estimated using pedigree and SNP markers from actual data. The coefficients for the A matrix are based on 354 trees, whereas the coefficients of G matrices are estimated for 166 trees and 3461SNP markers. 141

Table 2. Variance component obtained from ABLUP and various HBLUP prediction methods for the complete actual data (354 trees) 142

Table 3. Mean breeding values (EBVs), regression coefficients (intercept and slope) of TBVs on EBVs, mean squared error (MSE), bias (the difference between TBVs and EBVs), Pearson moment correlations (r_p) between TBVs and EBVs and Spearman rank correlations (r_s) for ABLUP and HBLUP prediction methods for the leave-one-out cross validation data 143

Table 4. The effect of scaled GOF matrix, based on observed allele frequencies, on breeding values. Different α values were used for the same HBLUP model to make the GOF matrix analogous to the A matrix of the genotyped individuals 144

Table 5. Descriptive statistics of coefficients of relationship estimated using pedigree and genomic information from simulated data for the construction of H matrix for 300 and 600 genotypes 145

Table 6. Variance component estimates obtained from ABLUP and HBLUP prediction methods for the complete simulated data using different number of genotyped individuals (300 genotyped/900 non-genotyped and 600 genotyped/600 non-genotyped) for the construction of H matrix 146

Table 7. Mean breeding values (EBVs), regression coefficients (intercept and slope) of TBVs on EBVs, mean squared error (MSE), bias (the difference between TBVs and EBVs), Pearson moment correlations (r_p) between TBVs and EBVs and Spearman rank correlations (r_s) for ABLUP and HBLUP prediction methods using simulated data used for the construction of H matrix 147

LIST OF FIGURES

CHAPTER 1

Figure 1. The diallel mating design of 19 parent trees used as female and male in the study. Parent trees 1 to 14 are from the first generation breeding population. Parent trees 18 to 22 are from the second generation breeding populations. 57

Figure 2. Pedigree structure of 2nd generation parent trees used in the diallel mating design. Black squares refer to the 1st generation parent trees, grey squares refer to the 2nd generation parent trees and white squares refer to the pollen mixed trees. Parent trees 2 and 6 are from the first generation and are related to the some of the second generation parent trees used as parent trees in diallel. 58

Figure 3. Plot of the site loadings for the first factor against second factor from XFA2 model fitted for GCA genetic effects for height. Site loadings are labeled from s1 to s10. The plot of the first loadings against second loadings shows three clusters, namely sites 7 and 10 consisted of first cluster, sites 1, 3, 6, 8 and 9 consisted of the second cluster and sites 2, 4 and 5 were in the third cluster. So, sites within these clusters were closely similar for this data set. 59

Figure 4. Biplot of XFA2 model for GCA genetic effects. Parents (genotype scores) are labeled p1 to p22, and site loadings are labeled as s1 to s10. Parents 3 and 8 tended to have negative GCA interaction with all the sites (GxE) included in the analysis. Parent 1, 17and 20 had positive response with all the sites. Sites 1, 3, 5, 6, 8 and 9, located farther away from the center, are the ones that discriminate the genotypes most and with higher correlations. Genotypes 4, 10, 14, 18 and 19 are clustered at the center of the biplot, indicating they had average response across all sites in terms of GxE. Parents 3 and 8 were poor parents in all sites. 60

Figure 5. Line plots of the BLUP values of the parents for clustered sites from XFA2 model (Model 8) for height. The patterns of the first three plots (a, b, c) indicated that there were small rank changes of parents among environments with similar loadings (clustered). When parental predictions were plotted across all sites (d), we observe clear rank changes (GxE interaction). The last figure (d) also shows that as the site variances get larger, we observe a clear separation among parents. 61

Figure 6. Plot of the BLUP values of full-sib families for the sites with highest and lowest site loadings (site 3 and 7 respectively) (a) and the line plots of full-sib families across all sites (b) are presented. In general, the ranking of the full-sib families were similar between sites 7 and 3. Also, we see more separation among families at the site 3 which had the highest variance (highest site loading). We also see full-sib families with low predictions show consistency across all sites (b). 62

CHAPTER 2

Figure 1. Pedigree structure of 178 cloned trees. Letters refer to nine family full-sib ID's. Squares with OP were unknown male contribution (wind-pollinated). Numbers refer to 13 half-sib ID's. Number of cloned trees per family ranged from 1 to 39. 109

Figure 2.Patterns of missing genotypes in given data due to assay failure: sporadic missing genotypes (a) or by design: structured missing genotypes (b)..... 110

Figure 3. Schematic presentation of marker genotypes for 10 loci (M1 to M10) for each tree 100

Figure 4. Accuracy of imputation (r) between imputed allele content values and true values for empirical data when sampling was sporadic or structured. Allele content values were imputed by the BLUP method using pedigree. Accuracy of imputation between imputed allele content values and true values are higher for the structured pattern of missing data than for the random pattern. 101

Figure 5. Accuracy of imputation (r) between imputed allele content values and true values for simulated data from half-diallel and double-pair mating designs when sampling was structured for different levels (missing genotypes ranged from 10% to 80%). Allele content values were imputed by the BLUP method using pedigree. Accuracy of imputations between imputed allele content values and true values are virtually equivalent for the two patterns of mating design and for different levels of missing genotypes. 102

Figure 6. Allele content values were imputed by haplotype analysis of parent-offspring trios using BEAGLE software for simulated data from half-diallel and double-pair mating designs when sampling was random. Accuracy of imputation was assessed as percent correctly imputed values relative to the true values. Imputation accuracy is >0.98 for datasets with up to 60% random missing genotypes for both mating designs. 103

Figure 7. Effect of random and evenly-spaced patterns of missing genotypes on the accuracy of imputation for half-diallel and double-pair mating designs at 80% missing genotype data when imputation was carried out with BEAGLE software. For the evenly-spaced pattern of missing data, the imputation accuracy was ~0.96 and was not affected by mating design at 80% missing data. 104

Figure 8. Effect of random, structured (non-genotyped) and evenly spaced patterns of missing genotypes on the accuracy of imputation, evaluated as percent correctly imputed values relative to the true values, for half-diallel mating design at 10% and 80% missing genotype data when imputation was carried out with BEAGLE software. For evenly spaced pattern of missing data, the imputation accuracy was the highest (0.96) at 80% missing.... 105

CHAPTER 3

Figure 1. Pedigree structure of 354 cloned trees (166 genotyped, 188 non-genotyped) used for the data analysis. Parents 2, 4, 5, 7 and 10 were used as female and male. Letters refer nine family full-sib ID's. 150

Figure 2. Scatter plots between true breeding values (TBVs) and estimated breeding values (EBVs) obtained from ABLUP and HBLUP methods using actual data. TBVs were predicted from complete data and EBVs were obtained using leave-one-out data set (354 trees: 166 genotyped and 188 non-genotyped). Genotyped individuals are in dark blue and non-genotyped individuals are in gray. 151

Figure 3. Scatter plots between true breeding values (TBVs) and estimated breeding values (EBVs) obtained from ABLUP and HBLUP prediction methods using simulated data (rep1). Genotyped individuals are in dark blue and non-genotyped individuals are in gray..... 152

CHAPTER 1

Genetic Analysis of Diallel Progeny Test Data in *Pinus taeda* Using Factor Analytic Linear Mixed Model

Abstract

The objective of this study was to test the efficiency of factor analytic (FA) covariance structures for genetic analysis of height in *Pinus taeda* L. diallel progeny trials to account for heterogeneity in variances and covariances between different environments. Progeny were planted across ten sites ranging from Atlantic Coast sites with mild winters and evenly distributed rainfall patterns through the year to more inland, drier and colder sites in Arkansas. The model efficiency was compared with various covariance structures commonly used in mixed models to estimate genetic merits of the parents, variance components, and narrow and broad sense heritabilities and to determine the extent of genotype by environment interaction (GxE). Among models fitted, FA models produced the smallest AIC model fit statistics. An unstructured variance-covariance matrix (US) produced a similar log likelihood value as the FA model, but required a large number of parameters to be estimated. FA models were parsimonious compared to US structures. FA models captured both variance and the covariance at the genetic level better than simpler models, and provided more accurate predictions of genotypes. The mean narrow-sense heritability estimates for height was about 0.20 when more complex variance structures were used, compared to 0.13 when simpler variance structures such as identity and block-diagonal variance structures were employed. The mean additive and non-additive genetic correlations were 0.83 and 0.9, respectively suggesting that GxE should not be a concern for this specific population in the environments where the genotypes were tested.

Introduction

Loblolly pine (*Pinus taeda* L.) is the most widely planted commercial forest tree species in the United States (McKeand et al. 2003, McKeand et al. 2006). Thanks to the efforts of tree improvement programs, every year more than 800 million genetically-improved loblolly pine seedlings are planted (McKeand et al. 2006). As a wood basket, the southeastern region of the United States produces more timber than any other country in the world (McKeand et al. 2003).

The North Carolina State University Cooperative Tree Improvement Program has finished 56years of loblolly pine genetic improvement and is about to start the fourth cycle of breeding and selection (McKeand et al. 2010). During the third cycle of the program diallel mating designs (called 'diallels' hereafter) were widely used because they provide information about general combining ability (GCA) and specific combining ability (SCA) of crosses. Diallel crosses are used to understand the genetic mechanism of a trait, such as the relative magnitudes of additive and non-additive genetic effects (Henderson 1977). Depending on whether selfs and reciprocals are included, diallel designs sometimes named full, half, partial and disconnected diallels (Lynch & Walsh 1998). Using controlled crosses such as diallels allow breeders to control coancestry in a breeding population while making selections for advanced breeding populations (Johnson and King, 1998). Diallel mating designs have also some disadvantages. The number of crosses needed in a diallel design can increase substantially the work required in making controlled crosses. In a diallel cross with N parents, there are N^2 crosses if self-crosses and reciprocal crosses are included (Henderson

1977, Curnow 1980). Many of the genetic assumptions of the diallel models are not met in applied breeding programs. For instance, most analytical methods assume there is no epistasis, no reciprocal effects, and no linkage disequilibrium (LD) occur and gene frequencies are intermediate in the population (Johnson and King 1998). Another drawback of using diallels is random genetic drift due to genetic sampling of parents from a large population (Isik et al. 2005).

Mixed models have become very popular for the analysis of data from series of variety trials. With mixed models, unbalanced data can be handled very easily. Within-trial variation can be modeled appropriately and some effects can be fit as random or fixed (Smith et al. 2005). Mixed linear model approaches allow modeling heterogeneous and correlated variance-covariance structures (Crossa et al. 2006). Linear mixed models give accurate predictions of genotypic effects by exploiting covariance structures that consider correlations between sites, years, plots, and genetic relationships between relatives (Cullis et al. 1998). Using pedigree information through an additive relationship matrix improves the genetic model because correlated information from relatives increase the accuracy of individual genotype effect estimates (Kelly et al. 2009). Mixed models can easily be applied to the incomplete data and they can model genotype by trial interaction and within trial error variation (Smith et al. 2005). Fitting a mixed model, not only the genetic covariance matrix can be estimated but also best linear unbiased predictions (BLUP) of the breeding values can be determined (Crossa et al. 2006) to rank varieties and make selection.

Since estimates of variance components will directly affect future breeding plans, fitting the best model for the data analysis is crucial to correctly predict breeding values, and thus make selections. There are various variance-covariance structures used in linear mixed models to model the heterogeneity in the data and correlations among factor levels. The most restrictive variance-covariance structures assume that genotype by environment interaction variances among sites are equal and all pair-wise correlations between sites and between genotypes are zero (Cullis et al. 1998). This unstructured matrix is the most general variance model, containing $t(t+1)/2$ parameters where t is the number of sites, but in terms of computation efficiency, especially for large number of parameters, it is difficult to fit (Smith et al. 2005).

The restricted maximum likelihood (REML) method has been widely used in mixed models to estimate variance covariance parameters with the assumption that random terms have Gaussian distribution (Gilmour et al. 1995). REML is the basic estimation procedure for the heterogeneous variance components estimation (Welham et al. 2010). Most of the mixed model analysis for multi environment trials (MET) use REML for estimation of covariances and best linear unbiased estimation (BLUE) and best linear unbiased prediction (BLUP) for the estimation of fixed and random effects, respectively (Smith et al. 2005). BLUP is a method for estimating random effects of a mixed model and gives better results in terms of predictive accuracy than other procedures (Piepho et al. 2007). The optimality properties of BLUP are based on the assumption that variance parameters in the model are known (Thomson et al. 2003). For many breeding programs accurate prediction of breeding values (BV's) is the main component for genetic improvement (Mrode 2005). More accurate and

unbiased BLUP results depend on the genetic covariance among individuals (Mrode 2005). Since the BLUP of genetic effects uses the genetic correlation among relatives, the higher the genetic correlation of a genotype of interest with related genotypes, the more information can be gained from the records of relatives (Piepho et al. 2007). BLUPs exploit the information from relatives via coefficient of coancestry (Piepho et al. 2007, Crossa et al. 2006), so BLUPs of breeding values of random genotypes are helpful for selecting superior genotypes (Crossa et al. 2006).

Multi-environmental trials (MET) provide opportunities to evaluate genotypes in different environmental conditions, to compare their responses in these environments, to evaluate their stability and adaptability, to study genotype by environment interaction, and to select the best genotypes in specific and across environments (Cullis et al. 1998, Smith et al. 2001, 2005). MET data generally show heterogeneous variance and covariance structures (Smith et al. 2001, Crossa et al. 2006). It is common for MET data to show unbalanced data structures (Piepho et al. 2007, So and Edwards 2009), because often not all cultivars grown in all trials (So and Edwards 2009). Analysis of MET data has some statistical complexities, because the data are often incomplete due to selection and unbalanced due to missing observations (Kelly et al. 2007, Spilke et al. 2005), and the possibility of genotype by environment interactions must be evaluated (Signor et al. 2001); these complexities often lead to violations of homogeneity of variance assumptions for random effects (Edwards and Jannink 2006).

For the analysis of MET data, many statistical methods have been proposed which don't depend on balanced data (Smith et al. 2005). The objective of statistical analysis of MET

data is to find the best (most appropriate) model that gives reliable predictions of variety performance across environments (Xiang and Li 2001). The choice of the best model depends on the data structure, statistical properties of the data and the statistical approach for the selection of the best model (Smith et al. 2005, Kelly et al. 2007). In traditional MET data analysis, it has been assumed that in each site, individual field plot errors are spatially independent and genotypes are unrelated, but these assumptions are not realistic, because the closer sites and related genotypes can be expected to be similar, and closer field plot observations tend to be correlated (Cullis et al. 1998). In forest tree progeny trials, in addition to measurement errors, there will be tree to tree variation because of the micro-site variation and non-additive genetic effects (Kelly et al. 2009). Since forest genetic trials are generally very large and likely heterogeneous, it is necessary to account for this micro environmental heterogeneity (Costa e Silva et al. 2001, Dutkowski al. 2002, Piepho et al. 2007, Kelly et al. 2009).

In plant breeding programs, due to limited number of genotypes and complex genetic covariance structures, genetic variance estimates are often inaccurate (Piepho et al. 2007). Fitting the appropriate variance-covariance structure to the data provides more accurate and precise variance component estimates and thus more precise prediction of breeding values for selection. Various variance-covariance structures have been proposed for the analysis of MET data. An unstructured variance-covariance matrix (US) is the common way to form a genetic variance. However, this structure has difficulties in terms of computational efficiency and reliability based on the number of parameters (Kelly et al. 2007, Meyer 2009). Unstructured covariances are explored to account for correlations among genotypes between

pair of sites. However, they require larger number of parameters to be estimated and such models can fail to converge for complex MET data. Another model for the genetic variance matrix is the uniform (or compound symmetry) model that evolved from an ANOVA approach. The model involves two parameters; the genetic variance and the variance due to GxE interaction effect. This model is still being used by many researchers (Zas 2008, Meyer 2009).

Smith et al. (2001) proposed a mixed model analysis with a factor analytic (FA) variance structure (multiplicative model) for the genotype by environment interaction effects and separate spatial covariance structures for the errors for each trial. This model is an extension of the spatial mixed model approach suggested by Cullis et al. (1998). This model provides genetic variance for each environment and genetic covariances between environments. FA models were introduced as an alternative to US models. The FA variance-covariance structure might be considered similar (as an approximation) to the completely unstructured variance-covariance matrix (Piepho et al. 1998) with different parameterization. Multiplicative models are the most complex models (Smith et al. 2005). FA models are commonly employed in MET data to handle heterogeneity (Cullis et al. 1998). The FA variance structure obtains information from correlated environments using the correlation of genotype by environment effects (Piepho et al. 2008). The FA model with sufficient multiplicative terms has been found to provide a good and parsimonious approximation to the unstructured form in terms of computation (Smith et al. 2005). If *k* factors are fit, then

$t(k+1) - k(k-1)/2$ parameters are estimated (Kelly et al. 2007). This is a large reduction in number of parameters to be estimated compared to US model.

FA model based on $\mathcal{k}$ factors, denoted as FA$\mathcal{k}$, given by

$$\mathbf{G_e} = \Lambda\Lambda' + \mathbf{\Psi}$$

where, Λ is a *t* by *k* matrix of environment loadings, *t* being the number of trials and *k* being the factors and $\mathbf{\Psi}$ is a *t* by *t* diagonal matrix with elements commonly referred to as specific variances (Smith et al. 2001).

There are clear advantages of FA models. FA models provide a natural framework in the presence of genotype by environment interactions (Meyer 2009). In the presence of substantial GXE interaction which comprises heterogeneous variances and crossovers in the genotype ranking, the FA model is robust with high predictive accuracy because it captures both variance heterogeneity and a more complex covariance structure at the genetic level. Also, fitting a FA model is generally easier than the US model. For example in ASReml software, the FA algorithm can handle the situations where the estimated variance matrix is reduced rank, but US model cannot (Kelly et al. 2007). The FA model provides more accurate estimates of overall variety effects and it allows one to understand and interpret the GxE interaction (Smith et al. 2001). Since transformations (re-parameterization of the model) are implemented at the effects level, different design matrices, missing observations and multiple random effects can be handled in FA models. An FA model can significantly reduce computational requirements of mixed model analysis in terms of estimation of variance effects and genetic evaluation schemes (Meyer 2009). Smith et al. (2001) emphasized that

incomplete data pose no difficulties. However, FA models do have some drawbacks. Factor analytic structures can be difficult to fit (Oakey et al. 2007). They may introduce some computational difficulties, and convergence problems can originate from poor starting values (Smith et al. 2001).

The objectives of this study were to;

1. test the efficiency of factor analytic (FA) covariance structures for analysis of multi-environmental progeny trials in forest tree breeding and compare model efficiency with various variance-covariance structures commonly used in mixed models
2. evaluate predictive accuracy of the models applied using model goodness of fit statistics
3. estimate the heritability of height trait using various models and
4. determine the extent of genotype by environment interaction.

Materials and Methods

Genetic material and mating design

A diallel mating design was used by the Plum Creek Timber Company, Inc. to produce crosses among 19 parent trees from first and second generation breeding populations of loblolly pine (**Figure 1**). In addition to 19 known parents, two seed orchard mixes (bulked seeds called SOM6 and SOM8) were included in the field tests. Some of the second generation parent trees used in mating design were descendents of first generation parents and were genetically related (**Figure 2**). Only 142 crosses were obtained from this mating design in which 139 crosses have known both parents (full-sibs) and three crosses have known female parents and unknown male parents (half-sibs). In addition, 11 parents were selfed to obtain 11 selfed crosses (**Table 1**).

For this study, height growth of 37,269 trees was analyzed as the response variable. Measurement of height growth of trees from 10 sites was carried out in 2009 and 2010 when trees were six years old. Tree height (HT) was measured to the nearest centimeter.

Field design

The study was established on 10 sites located in Georgia, Arkansas, north Louisiana, and south Mississippi in 2003 (three sites) and 2004 (seven sites) (Table 1). A randomized complete block design with single tree plot configuration was used for all sites. One progeny of each cross was randomly assigned in a block. There were 30 blocks per site. Thus, each cross had 30 progeny at one site. Number of crosses per site ranged from 115 to 141. Checklots were also included in each of the progeny test sites. Checklots are unimproved and

improved families used to estimate genetic gains for several levels, such as genetic gains over unimproved seed. The number of checklots ranged from two to seven among sites (Table 1). Two-tree row plots in each block were used for the checklots in the three progeny test sites established in 2003, while three-tree row plots in each block were used for the checklots in the seven sites established in 2004. As a standard practice, two rows of border trees were planted around experimental trees.

Statistical Analysis

The following linear mixed model was used to partition observed variance for growth and stem straightness into causal (additive, dominance, environmental) components.

$$[1] \qquad y_{ijklm} = \mu + T_i + B(T)_{j(i)} + T*G_{ik} + T*G_{il} + T*S_{ikl} + e_{ijklm}$$

where

y_{ijklm} is the m^{th} observation for the kl^{th} cross in the j^{th} block within the i^{th} site;

μ is the overall mean;

T_i is the i^{th} fixed site (location) effect, i=1 to t;

$B(T)_{j(i)}$ is the random effect of the j^{th} block within i^{th} site, j=1 to b ~ Normally and Independently Distributed NID $(0, \sigma^2_B)$;

$T*G_{ik}$, $T*G_{il}$ are the random GCA effect of the k^{th} female or the l^{th} male by i^{th} site interaction effect ~NID $(0, \sigma^2_{TG})$;

$T*S_{ikl}$ is the random SCA effect of the kl^{th} cross by i^{th} site interaction effect ~NID $(0, \sigma^2_{TS})$;

e_{ijklm} is the random error term ~NID $(0, \sigma^2_e)$.

Consider *p* genotypes grown at *t* sites. In terms of the analysis of MET data the general linear mixed model used for the analysis of progeny test data can be written as a matrix of the general form:

$$[2] \qquad \mathbf{y} = \mathbf{X\tau} + \mathbf{Z_b b} + \mathbf{Z_u u} + \mathbf{e}$$

where **y** is the $n \times 1$ vector of observations combined across all sites, $\boldsymbol{\tau}$ is the $t \times 1$ vector of fixed effects, **b** is the $bt \times 1$ vector of random/fixed replicate (block) effects within sites, **u** is the vector of $pt \times 1$ random genotype effects within sites and **e** represents $n \times 1$ vector of residuals combined across sites. $\mathbf{X}^{(\mathbf{n\times t})}$, $\mathbf{Z_b}^{(\mathbf{n\times bt})}$ and $\mathbf{Z_u}^{(\mathbf{n\times pt})}$ are the design matrices that relate observations to the fixed and random effects, with the expectations

$$\begin{bmatrix}\mathbf{b}\\ \mathbf{u}\\ \mathbf{e}\end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix}0\\0\\0\end{bmatrix}, \begin{bmatrix}\mathbf{B} & 0 & 0\\ 0 & \mathbf{G} & 0\\ 0 & 0 & \mathbf{R}\end{bmatrix}\right)$$

The variance-covariance matrix of **y** is

$$[3] \qquad \mathrm{var}(\mathbf{y}) = \mathbf{Z_b B Z'_b} + \mathbf{Z_u G Z'_u} + \mathbf{R}$$

Following the approach of Oakey et al. (2006) and Kelly et al. (2009), the vector of genotype effects (total genetic effect) within sites (**u**) can be partitioned into two components: additive and dominance genetic effects as

$$\mathbf{u} = \mathbf{u_a} + \mathbf{u_d}$$

where $\mathbf{u_a}$ is the *ptx1* vector of random additive genetic effects within sites ($T*G_{ik}$, $T*G_{il}$) and $\mathbf{u_d}$ is the *kltx1* vector of random dominance genetic effects within sites ($T*S_{ikl}$). The random genotype effects are assumed to follow a Gaussian distribution with zero mean and variance matrix

$$\text{var}(\mathbf{u}) = \mathbf{G} = \mathbf{G_a} \oplus \mathbf{G_d} = \begin{pmatrix} \mathbf{G_a} & 0 \\ 0 & \mathbf{G_d} \end{pmatrix}$$

The additive genotype effects variance matrix, as a two-way structure of genotype by environment effects, has the separable form

$$\mathbf{G_a} = \mathbf{G_{e_a}} \otimes \mathbf{G_{v_a}}$$

where $\mathbf{G_{e_a}}$ and $\mathbf{G_{v_a}}$ are *txt* and *pxp* symmetric positive definite matrices, respectively. $\mathbf{G_{e_a}}$ is the additive genetic variance-covariance matrix between sites, and $\mathbf{G_{v_a}}$ is the variance-covariance matrix between parents or additive effects. $\mathbf{G_{v_a}} = \mathbf{A}$ where $\mathbf{A}$ is a numerator relationship matrix derived from pedigree. Thus, $\mathbf{G_a}$ can be written as $\mathbf{G_a} = \mathbf{G_{e_a}} \otimes \mathbf{A}$. This model implies that genotype x environment effects are correlated between sites (Smith et al. 2005).

The dominance genetic effects are modeled as a two-way structure of cross by environment effects, with an associated variance of

$$\mathbf{G_d} = \mathbf{G_{e_d}} \otimes \mathbf{G_{v_d}}$$

where $\mathbf{G_{e_d}}$ and $\mathbf{G_{v_d}}$ are *txt* and *klxkl* symmetric positive definite matrices, respectively. $\mathbf{G_{e_d}}$ is the dominance genetic variance-covariance matrix between sites, and $\mathbf{G_{v_d}}$ is the variance-covariance matrix between full-sib families. It was assumed that dominance genetic effects are independent, $\mathbf{G_{v_d}} = \mathbf{I}_{\text{kl}}$ and can be written as $\mathbf{G_d} = \mathbf{G_{e_d}} \otimes \mathbf{I}_{\text{kl}}$.

When all variance-covariance components are put together the variance-covariance matrix of $\mathbf{y}$ can be rewritten as

$$[4] \qquad \text{var}(\mathbf{y}) = \mathbf{H} = \mathbf{Z_b B Z'_b} + \mathbf{Z_u}[\mathbf{G_{e_a}} \otimes \mathbf{A} + \mathbf{G_{e_d}} \otimes \mathbf{I}_{\text{kl}}]\mathbf{Z'_u} + \mathbf{R}$$

We had data on a set of trees that were genetically linked via a pedigree. Therefore, genetic effects were correlated and, with the assumption of normal modes of inheritance, the correlation expected from additive genetic effects can be derived from the pedigree provided that all the genetic links are in the pedigree. By specifying relationships between ancestors, parents and offspring, all breeding values are predicted on the same scale, and all breeding values are obtained as a single set of rankings (White et al. 2007). The additive genetic relationship matrix, sometimes called the numerator relationship matrix (**A**), indicating the additive genetic relationship among individuals (Mrode 2005), can be calculated from the pedigree. The **A** matrix has elements $A_{ij} = 2 \ominus_{ij}$, where $\ominus_{ij}$ is coefficient of coancestry: the probability that two genes are identical by descent when drawing a gene randomly from two individuals (Lynch and Walsh 1998).

In this study, different **B**, **R** and **G** variance-covariance structures were explored to find the best model fitted to the data based on the AIC criteria.

*Modeling **B** and **R** variance structures*

The variance-covariance matrix of replicates within sites (**B)** is usually assumed to have a block diagonal variance structure (Kelly et al. 2007, 2009) as given below.

$$\mathbf{B} = \sum_{\mathbf{b}} \otimes \mathbf{I_b} = \begin{bmatrix} \sigma^2_{b_1} & 0 & \cdots & 0 \\ 0 & \sigma^2_{b_2} & \cdots & 0 \\ . & . & \cdots & . \\ . & . & \cdots & . \\ 0 & 0 & \cdots & \sigma^2_{b_{10}} \end{bmatrix} \otimes \mathbf{I_b} = \sigma^2_{b1}\mathbf{I_{n1}} \oplus \sigma^2_{b2}\mathbf{I_{n2}} \oplus \ldots \oplus \sigma^2_{b10}\mathbf{I_{n10}}$$

where $\otimes$ is the direct product operator, $\mathbf{I_b}$ is an identity matrix of order *b* and $\sum_{\mathbf{b}}$ is replicate matrix (Crossa et al. 2006).

In this study, the variance matrix for the replicate effects within sites, **B**, was fitted as diagonal (IID) with $\mathbf{B} = \boldsymbol{\sigma}_{\mathbf{b}}^{\mathbf{2}}\mathbf{I_n}$ ($\sigma_{b1}^2 = \sigma_{b2}^2 = \cdots = \sigma_{b10}^2$) and block diagonal with $\mathbf{B} = \sigma_{b1}^2\mathbf{I_{n1}} \oplus \sigma_{b2}^2\mathbf{I_{n2}} \oplus \ldots \oplus \sigma_{b10}^2\mathbf{I_{n10}}$ ($\sigma_{b1}^2 \neq \sigma_{b2}^2 \neq \cdots \neq \sigma_{b10}^2$). For the block diagonal variance structure, replicate effects within sites are independently and identically distributed (IID).

The variance matrix for the error effects, **R**, is assumed to be block diagonal with **R**=diag$(\mathbf{R}_t)$, where $\mathbf{R}_t$ is the error variance matrix for the t^{th} site (Kelly et al. 2007, 2009). Smith et al. (2001) used the spatial variance structure (autoregressive process of order one) of Gilmour et al. (1997) to model the error effects. Thompson et al. (2003) assumed that errors from different trials are independent and the errors from j^{th} trial constitute a two dimensional (row-column) spatially dependent process. So et al. (2009) compared models in which error variance matrices were fitted as block diagonal ($\sigma_{e1}^2 \neq \sigma_{e2}^2 \neq \sigma_{e3}^2$) and diagonal ($\sigma_{e1}^2 = \sigma_{e2}^2 = \sigma_{e3}^2$). The variance-covariance matrix of error, **R,** was assumed to have the simple variance-covariance structure ($\mathbf{R} = \sum_{\mathbf{e}} \otimes \mathbf{I_{bg}}$) (Crossa et al. 2006).

In this study, the variance matrix for the error effects was assumed to be diagonal (IID) with $\mathbf{R} = \sigma_e^2\mathbf{I_n}$ ($\sigma_{e1}^2 = \sigma_{e2}^2 = \cdots = \sigma_{e10}^2$) and block diagonal with $\mathbf{R} = \sigma_{e1}^2\mathbf{I_{n1}} \oplus \sigma_{e2}^2\mathbf{I_{n2}} \ldots \sigma_{e10}^2\mathbf{I_{n10}}$ ($\sigma_{e1}^2 \neq \sigma_{e2}^2 \neq \cdots \neq \sigma_{e10}^2$).

*Modeling **G** variance-covariance structure*

The main focus of this study was modeling the **G** variance structure ($\mathbf{G_{e_a}}$ and $\mathbf{G_{e_d}}$). Genetic variances for each site show the magnitude of variation between genotypes, and genetic correlations between pairs of sites show the agreement in genotype rankings (Smith et al. 2001). We considered diagonal (IID), block diagonal (B-DIAG), compound symmetry (CS), factor analytic (FA) and unstructured (US) variance-covariance structures to model $\mathbf{G_{e_a}}$ and $\mathbf{G_{e_d}}$ matrices.

In a diagonal variance structure, $\mathbf{G_{e_a}}$ / $\mathbf{G_{e_d}}$ is IID; each site has the same GCA/SCA genetic variances in the diagonal and GCA/SCA genetic covariances between pairs of sites in the off-diagonal are 0. The matrix of $\mathbf{G_{e_a}}$ / $\mathbf{G_{e_d}}$ for the diagonal variance-covariance structure can be presented as:

$$\mathbf{G_{e_a}} = \begin{bmatrix} \sigma^2_{e_a} & 0 & 0 & \cdots & 0 \\ 0 & \sigma^2_{e_a} & 0 & \cdots & 0 \\ 0 & 0 & \sigma^2_{e_a} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & \sigma^2_{e_a} \end{bmatrix} \qquad \mathbf{G_{e_d}} = \begin{bmatrix} \sigma^2_{e_d} & 0 & 0 & \cdots & 0 \\ 0 & \sigma^2_{e_d} & 0 & \cdots & 0 \\ 0 & 0 & \sigma^2_{e_d} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & \sigma^2_{e_d} \end{bmatrix}$$

In a block diagonal variance matrix (B-DIAG), each site has its own GCA/SCA genetic variances in the diagonal and genetic covariances between pairs of sites in the off diagonal are 0. The matrix of $\mathbf{G_{e_a}}$ / $\mathbf{G_{e_d}}$ for the B-DIAG variance-covariance structure can be presented as:

$$\mathbf{G_{e_a}} = \begin{bmatrix} \sigma^2_{e_{a1}} & 0 & 0 & \cdots & 0 \\ 0 & \sigma^2_{e_{a2}} & 0 & \cdots & 0 \\ 0 & 0 & \sigma^2_{e_{a3}} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & \sigma^2_{e_{a10}} \end{bmatrix} \quad \mathbf{G_{e_d}} = \begin{bmatrix} \sigma^2_{e_{d1}} & 0 & 0 & \cdots & 0 \\ 0 & \sigma^2_{e_{d2}} & 0 & \cdots & 0 \\ 0 & 0 & \sigma^2_{e_{d3}} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & \sigma^2_{e_{d10}} \end{bmatrix}$$

In compound symmetry (CS) structure, all sites have the same GCA/SCA genetic variance along the diagonal and all pairs of sites have the same GCA/SCA genetic correlation in the off diagonal. The matrix of $\mathbf{G_{e_a}}$ / $\mathbf{G_{e_d}}$ with CS structure can be shown,

$$\mathbf{G_{e_a}} = \begin{bmatrix} \sigma^2_{e_a} & r & r & \cdots & r \\ r & \sigma^2_{e_a} & r & \cdots & r \\ r & r & \sigma^2_{e_a} & \cdots & r \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ r & r & r & \cdots & \sigma^2_{e_a} \end{bmatrix} \qquad \mathbf{G_{e_d}} = \begin{bmatrix} \sigma^2_{e_d} & r & r & \cdots & r \\ r & \sigma^2_{e_d} & r & \cdots & r \\ r & r & \sigma^2_{e_d} & \cdots & r \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ r & r & r & \cdots & \sigma^2_{e_d} \end{bmatrix}.$$

The factor analytic (FA) model is hypothesized in terms of the (unobserved) GCA/SCA genetic effects in different sites assuming dependence on a set of random hypothetical factors $f_r^{(pxp)}$, $r = 1 \ldots . k < p$:

[5] $\quad u_a = \sum_{r=1}^{k} \lambda_{ir} f_{lr} + \delta_{li} \quad / \quad u_d = \sum_{r=1}^{k} \lambda_{ir} f_{(kl)r} + \delta_{(kl)i}$

where u_a is the GCA effect for genotypes *l* at site *i*, u_d is the SCA effect (cross) at site *i*, f_{lr} is the score for the genotype *l* in factor r, $f_{(kl)r}$ is the score for the cross *kl* in factor r, λ_{ir} is the associated loading for site *i* and δ_{li} and $\delta_{(kl)i}$ are the lack of fit of the model. When we write these models in a matrix notation:

$$\mathbf{u_a} = (\lambda_{a1} \otimes \mathbf{A})\mathbf{f_1} + \cdots + (\lambda_{ak} \otimes \mathbf{A})\mathbf{f_k} + \boldsymbol{\delta} = (\Lambda_{\mathbf{a}} \otimes \mathbf{A})\mathbf{f} + \boldsymbol{\delta}$$

$$\mathbf{u_d} = (\lambda_{d1} \otimes \mathbf{I_{kl}})\mathbf{f_1} + \cdots + (\lambda_{dk} \otimes \mathbf{I_{kl}})\mathbf{f_k} + \boldsymbol{\delta} = (\Lambda_{\mathbf{d}} \otimes \mathbf{I_{kl}})\mathbf{f} + \boldsymbol{\delta}$$

where $\lambda_r{}^{(t\times 1)} = \{\lambda_{ir}\}$, $\mathbf{f}_r{}^{p\times 1} = \{f_{lr}\}/\{f_{(kl)r}\}$, $\boldsymbol{\delta}^{(pt\times 1)} = \delta_{li}/\delta_{(kl)i}$, $\Lambda^{(t\times k)} = (\lambda_1 \dots \lambda_k)$ and $\mathbf{f}^{(pk\times 1)} = (\mathbf{f}_1', \mathbf{f}_2' \dots \mathbf{f}_k')'$. The joint distribution of $\mathbf{f}$ and $\boldsymbol{\delta}$ for the GCA/SCA genetic effect is,

$$\begin{pmatrix}\mathbf{f}\\ \boldsymbol{\delta}\end{pmatrix} \sim N\left[\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\mathbf{I_k}\otimes\mathbf{A} & 0\\ 0 & \boldsymbol{\Psi_a}\otimes\mathbf{A}\end{pmatrix}\right] \quad / \quad \begin{pmatrix}\mathbf{f}\\ \boldsymbol{\delta}\end{pmatrix} \sim N\left[\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\mathbf{I_k}\otimes\mathbf{I_{kl}} & 0\\ 0 & \boldsymbol{\Psi_d}\otimes\mathbf{I_{kl}}\end{pmatrix}\right]$$

where, $\boldsymbol{\Psi_a} = \text{diag}(\Psi_{a1} \dots \Psi_{at})$, $\boldsymbol{\Psi_d} = \text{diag}(\Psi_{d1} \dots \Psi_{dt})$, $\boldsymbol{\Psi_a}/\boldsymbol{\Psi_d}$ is diagonal matrix with elements $(\Psi_{a1} \dots \Psi_{at})/ (\Psi_{d1} \dots \Psi_{dt})$ and Ψ_{ai}/Ψ_{di} is known specific variance for the i^{th} site. The variance matrix for the GCA/SCA genetic effects at each site is,

$$\text{var}(\mathbf{u_a}) = (\Lambda_\mathbf{a}\otimes\mathbf{A})\text{var}(\mathbf{f})(\Lambda_a'\otimes\mathbf{A}) + \text{var}(\boldsymbol{\delta}) = (\Lambda_\mathbf{a}\Lambda_\mathbf{a}' + \boldsymbol{\Psi_a})\otimes\mathbf{A} = \mathbf{G_{e_a}}\otimes\mathbf{A}$$

$$\text{var}(\mathbf{u_d}) = (\Lambda_\mathbf{d}\otimes\mathbf{I_{kl}})\text{var}(\mathbf{f})(\Lambda_d'\otimes\mathbf{I_{kl}}) + \text{var}(\boldsymbol{\delta}) = (\Lambda_\mathbf{d}\Lambda_\mathbf{d}' + \boldsymbol{\Psi_d})\otimes\mathbf{I_{kl}} = \mathbf{G_{e_d}}\otimes\mathbf{I_{kl}}$$

[6] $$\text{var}(\mathbf{u}) = \text{var}(\mathbf{u_a} + \mathbf{u_d}) = G = (\Lambda_\mathbf{a}\Lambda_\mathbf{a}' + \boldsymbol{\Psi_a}) \otimes \mathbf{A} + (\Lambda_\mathbf{d}\Lambda_\mathbf{d}' + \boldsymbol{\Psi_d}) \otimes \mathbf{I_k}$$

where $\Lambda_\mathbf{a}/\Lambda_\mathbf{d}$ is a $t \times k$ matrix of environment loadings (on the covariance scale), $\mathbf{f}$ is the associated $pk \times 1$ vector of genotype scores, $\boldsymbol{\Psi_a}/\boldsymbol{\Psi_d}$ is a $t \times t$ diagonal matrix with elements referred to as specific variances and k is the number of components (multiplicative terms). The matrix of $\mathbf{G_{e_a}}$ / $\mathbf{G_{e_d}}$ with FA structure for k=1 can be shown as

$$\Lambda_\mathbf{a} = \begin{bmatrix} c_{a1}\\ c_{a2}\\ c_{a3}\\ \dots\\ c_{a10}\end{bmatrix} \quad \boldsymbol{\Psi_a} = \begin{bmatrix} V_1 & 0 & 0 & \cdots & 0\\ 0 & V_2 & 0 & \cdots & 0\\ 0 & 0 & V_3 & \cdots & 0\\ \dots & \dots & \dots & \cdots & \dots\\ 0 & 0 & 0 & \cdots & V_{10}\end{bmatrix} \quad \mathbf{G_{e_a}} = \begin{bmatrix} c_{a1}^2 + V_1 & c_{a1}c_{a2} & c_{a1}c_{a3} & \cdots & c_{a1}c_{a10}\\ c_{a2}c_{a1} & c_{a2}^2 + V_2 & c_{a2}c_{a3} & \cdots & c_{a2}c_{a10}\\ c_{a3}c_{a1} & c_{a3}c_{a2} & c_{a3}^2 + V_3 & \cdots & c_{a3}c_{a10}\\ \dots & \dots & \dots & \cdots & \dots\\ c_{a10}c_{a1} & c_{a10}c_{a2} & c_{a10}c_{a3} & \cdots & c_{a10}^2 + V_{10}\end{bmatrix}$$

$$\Lambda_{\mathbf{d}} = \begin{bmatrix} c_{d1} \\ c_{d2} \\ c_{d3} \\ \cdots \\ c_{d10} \end{bmatrix} \boldsymbol{\Psi}_{\mathbf{d}} = \begin{bmatrix} V_1 & 0 & 0 & \cdots & 0 \\ 0 & V_2 & 0 & \cdots & 0 \\ 0 & 0 & V_3 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & V_{10} \end{bmatrix} \mathbf{G}_{\mathbf{e_d}} = \begin{bmatrix} c_{d1}^2 + V_1 & c_{d1}c_{d2} & c_{d1}c_{d3} & \cdots & c_{d1}c_{d10} \\ c_{d2}c_{d1} & c_{d2}^2 + V_2 & c_{d2}c_{d3} & \cdots & c_{d2}c_{d10} \\ c_{d3}c_{d1} & c_{d3}c_{d2} & c_{d3}^2 + V_3 & \cdots & c_{d3}c_{d10} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ c_{d10}c_{d1} & c_{d10}c_{d2} & c_{d10}c_{d3} & \cdots & c_{d10}^2 + V_{10} \end{bmatrix}$$

The variance structure for GxE interaction effects, $(\Lambda\Lambda' + \boldsymbol{\Psi})\otimes\mathbf{A}$, is known as a Factor Analytic structure of order *k* (FA*k*). When *k*=1, (*k*, number for factors), the number of variance parameters estimated is $tk + t$, where *t* is the number of sites and *k* is the number factors. However if *k* >1, *k*(*k*-1)/2 independent constraints are required on the elements of Λ for identifiability as suggested by Smith et al. (2001). As a result, rotation is required on the site loadings and genotype scores for a definite interpretation of them as Λ*=ΛΛ'Λ (Smith et al. 2001, Thompson et al. 2003). Then the number of parameters estimated is $tk + t - k(k - 1)/2$, where *t* is the number of sites and *k* is the number factors, when *k*>1. Using a FA model for GxE effects, the genetic covariances among *t* sites are explained in terms of smaller number of unknown factors (*k*<*t*) (Smith et al. 2002). Burgueno et al. (2007) reported that an FA model can be interpreted as a random regression model of genotype and genotype by environment interaction on *k* environmental covariates (environmental loadings), $\lambda_1 \ldots \lambda_k$, in which each genotype has its own slope (genotypic scores) but common intercept. The slopes measure the sensitivity of genotypes to the loadings of each site (Smith et al. 2002).

The FAk and XFAk are different parameterization of Factor Analytic (FA) model. In the correlation form of the FA models, parameters are specified in the order loadings for each factor followed by the variances (Gilmour et al. 2009). We also tested an Extended Factor Analytic (XFAk) form of the FA model. The parameterization of the XFAk model is

specified in the order of specific variances and site loadings (Gilmour et al. 2009). One of the problems with the FA model is the Heywood case in which one or more specific variances tends to 0, which means that the GxE variance matrix is not full rank (Thompson et al. 2003). In the presence of the Heywood case, genotype effects for a site (trial) are completely determined by the multiplicative part of the model (Smith et al. 2001). In the FA model, specific variances are constrained to be 0 when they tend to estimate on the boundary of parameter case.

As in FA*k* models, in XFA*k* models, when *k* is greater than 1, constraints on the elements of Λ are required. XFA*k* models allow some elements of the specific variances ($\mathbf{\Psi}$) to be fixed to 0. Elements of $\mathbf{\Psi}$ might be 0 making $\mathbf{G_{e_a}/G_{e_d}}$ singular which requires a XFA*k* model term. The XFA*k* model inserts *i* columns of 0's into the design matrix corresponding to *i* factors. XFA*k* models are faster than FA*k* models when *k* is much smaller than *t* (Gilmour et al. 2009).

Finally, we used unstructured (US) variance structure to model **G** structure. The unstructured (US) variance-covariance matrix is the most general heterogeneous variance-covariance matrix, containing $t(t+1)/2$ parameters, where *t* is the maximum number of repeated measurements on a subject (Wolfinger 1996, Smith et al. 2001). In the US variance matrix, each site has different GCA/SCA genetic variances in the diagonal and genetic covariances between pairs of sites in the off diagonal. The US variance-covariance matrix of $\mathbf{G_{e_a}}$ / $\mathbf{G_{e_d}}$ can be shown as:

$$\mathbf{G_{e_a}} = \begin{bmatrix} \sigma^2_{a1} & \sigma_{a12} & \sigma_{a13} & \cdots & \sigma_{a110} \\ \sigma_{a21} & \sigma^2_{a2} & \sigma_{a23} & \cdots & \sigma_{a210} \\ \sigma_{a31} & \sigma_{a32} & \sigma^2_{a3} & \cdots & \sigma_{a310} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \sigma_{a101} & \sigma_{a102} & \sigma_{a103} & \cdots & \sigma^2_{a10} \end{bmatrix} \qquad \mathbf{G_{e_d}} = \begin{bmatrix} \sigma^2_{d1} & \sigma_{d12} & \sigma_{d13} & \cdots & \sigma_{d110} \\ \sigma_{d21} & \sigma^2_{d2} & \sigma_{d23} & \cdots & \sigma_{d210} \\ \sigma_{d31} & \sigma_{d32} & \sigma^2_{d3} & \cdots & \sigma_{d310} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \sigma_{d101} & \sigma_{d102} & \sigma_{a103} & \cdots & \sigma^2_{d10} \end{bmatrix}$$

All the models fit in this study as the combination of **G** and **R** structures are summarized in Table 1.2.

Model fit statistics (model comparison)

After fitting all these variance-covariance structures to the linear mixed model, variance components were estimated by REML method using ASReml3 software (Gilmour et al. 2009). The Akaike Information Criterion (AIC) (AIC=-2LogL+2t_i) was used to evaluate the goodness-of-fit of the variance models (Gilmour et al. 2009).

ASReml and GENSTAT have been reported to be the most appropriate software packages for the analysis of MET data (Smith et al. 2005). The most commonly used software for fitting such models is ASReml (Gilmour et al. 1995, 2002) which uses the average information (AI) algorithm for restricted maximum likelihood (REML) estimation of the variance parameters. For the MET data, the mixed model equations are complex and as a result, the speed of the computation of the analysis of datasets with large number of sites or fitting FA variance models with several factors is quite slow. Thompson et al. (2003) proposed a sparse implementation of the AI algorithm for reduced rank or FA variance parameter estimates. Initial values for the FA model is crucial because the initial values effect the convergence (Smith et al. 2001, Burgueno et al. 2007), so the block-DIAG variance structure solution provides a good starting point for fitting FA*k* (k=2,3..) (Burgueno et al.

2007). In our study for FA(1) model, 0.9 was used as starting value for the loadings (Λ), and block-DIAG variance structure results were used as initial values for variances. For FA(2) model, FA(1) model results (loadings and variances) were used as initial values of the first factor loadings and variances and for the second factor loadings, 0.5 was used for each site. Since convergence of the FA(1) or FA(2) model was not obtained with the first initial values, the initial values for loadings and variances were updated using the solutions output file until the model converged. Genetic correlations were obtained directly from the variance covariance matrix.

Estimation of genetic parameters

With the assumption of no epistasis, additive and dominance genetic variances for our fitted models are assumed to be $\sigma_A^2 = 4\sigma_G^2$ and $\sigma_D^2 = 4\sigma_S^2$, respectively (Lynch and Walsh 1998). For the Block-DIAG, CS, FA, XFA and US models, narrow sense and broad sense heritabilities were calculated using estimated variance components according to the following formulas:

Narrow-sense individual heritability:

$$[7] \qquad h^2 = 4\sigma_G^2/(2\sigma_G^2 + \sigma_S^2 + \sigma_e^2)$$

Broad-sense individual heritability:

$$[8] \qquad H^2 = 4(\sigma_G^2 + \sigma_S^2)/(2\sigma_G^2 + \sigma_S^2 + \sigma_e^2)$$

where σ_G^2 is genetic variance of GCA, σ_S^2 is genetic variance of SCA and σ_e^2 is error variance for each site. Mean narrow sense and broad sense heritabilities were calculated for each model, taking the average of each of the estimated parameters (Isik et al. 2005).

We evaluated GxE interaction based on the genetic correlations between pairs of sites obtained from the best model (Model 8). To visualize GxE interaction and rank changes among parents/ families, BLUP of the parents/families were plotted against site loadings for a subset of sites. A biplot of the fitted factor analytic (Model 8) was produced from the scores of the genotypes (parents) and loadings of the sites.

Results

Model comparison

In this study we analyzed MET progeny trials of loblolly pine data for height growth. All the variance-covariance structures fitted to GCA and SCA genetic effects and model fit statistics are presented in Table 3 and Table 4 when **R** side of the mixed model was IID and Diagonal, respectively. When an IID structure was used for the **R** side of the mixed model and when block effect (**B**) is assumed IID, the US genetic variance structure produced the highest REML log-likelihood (LogL) value (**Table 3**). However, when the number of parameters was taken into account, models (4 and 5 in the table) with FA genetic variance structures were superior, as shown by the smaller AIC model fit statistics.

Fitting heterogeneous variance structures for the residuals and for the block effects (diagonal) for the same models given in Table 3 resulted in lower AIC values (**Table 4**). For example, model 4 given in Table 3 and Table 4 had the same genetic variance-covariance structure (FA1) but they differ for the **R** side (IID versus Diagonal) and for the block effect (IID versus Diagonal). Model 4 with the Diagonal **R** (Table 4) had a better fit (lower AIC) than Model 4 with IID **R** in Table 3. A model with the smallest AIC value is assumed to be superior in terms of fit and parsimony (Oakey et al. 2007).

Models given in Table 4 can also be compared among themselves. All the models had the same **R** structure (diagonal) and the same variance structure for the block effect (diagonal). Among the models given in Table 4, Model 1 assumed the same GCA and SCA genetic variances across all sites and no correlations between pair of sites. Not surprisingly, this

model gave the poorest fit statistics among the fitted models. Model 8 with the XFA2 for GCA and XFA1 for SCA genetic effects produced the lowest AIC value.

Model 8 with XFA2 and XFA1 structures for GCA and SCA genetic effects, respectively, was not significantly different from Model 10 with the US genetic variance structure (REMLRT=48.2 on 61 df, p=0.88). The US structure for genetics effects (Model 10) produced the highest LogL. However, in terms of parsimony (49 parameters vs. 110 parameters), Model 8 (XFA2 and XFA1 structures for GCA and SCA genetic effects, respectively) was superior than Model 10 (US structure for GCA and SCA genetic effects) and was a better fit as shown by lower AIC values.

Genetic parameter estimates

The variance explained by GCA and SCA effects, loadings, lack of fit (specific variances), the percent variance explained by the site loadings and the ratio of SCA variance and GCA variance are summarized in **Table 5**. The estimates are for the best fit model (Model 8 in **Table 4**) with the XFA2 for GCA and XFA1 for SCA effects.

There was heterogeneity among the sites for GCA and SCA genetic variances. The first factor site loadings (average 84%) for GCA effect accounted for more variation than the second factor site loadings (average %8). On average, both factor site loadings ($\%\Lambda_{a1} + \%\Lambda_{a2}$) explained about 92% of the GCA variance, varying between 55% and 100%. For the SCA effect, the loadings of the one factor model accounted for 75% to 100% of variation (average 91%) in SCA effect. At sites 4, 5, 8 and 10, specific variances were estimated as 0, the meaning that GCA effects are completely determined by the multiplicative part of the

model at these sites. SCA genetic variance was always larger than GCA genetic variance across the sites as shown by the ratio of SCA variance and GCA variance (**Table 5** last column). The ratio ranged from 1.0 to 5.7, depending on the site.

Genotype by Environment interactions

The loadings (after rotation) from the XFA2 model for the GCA effect are displayed graphically in **Figure 3**. The two-dimensional plot of first loadings against second loadings formed three clusters of sites. The first cluster included sites 7 and 10 with high positive loading2, the second cluster included sites 1, 3, 6, 8 and 9 with loading clustered around 0, and the last cluster had sites 2, 4 and 5 with negative loading2. GCA genetic effect correlations between pair of sites for each cluster varied between 0.91, 0.73-0.98, and 0.92-0.98, respectively.

The biplot of the site loadings and the genotype scores (after rotation) for the GCA genetic effects from XFA2 model (Model 8) illustrates the relationships between sites and parents (GxE pattern) in a two dimensional space (**Figure 4**). Parents 3 and 8 tended to have negative GCA interaction (GxE) with all the sites included in the analysis. They are located on the negative quadrant of the biplot and had negative response (poor performance) in all sites. Parents 1, 17 and 20 had positive interactions (response) with all the sites. Parents 4, 10, 14, 18 and 19 are clustered at the center of the biplot, indicating they had average response across all sites in terms of GxE. Sites 1, 3, 5, 6, 8 and 9, located farther away from the center, are the ones that discriminate the genotypes most and with higher correlations. Sites 2, 10

and 7 are characterized by shorter lines (smaller site loadings) and they do not discriminate the parents as much.

In order to see the comparative parental ranking (GCA effect) among clustered sites (similar sites based on the loadings) and all sites, **GxE interaction**, the estimated BLUP values of the 22 parents are illustrated in **Figure 5** from the XFA2 model (Model 8). Genotypes did not show significant rank changes between environments with similar loadings (low or high) or sites in the same clusters (Figure 5a, 5b, 5c). When parental predictions (GCA) were plotted for environments across sites in different clusters (Figure 5d) we see more rank changes. Similarly, in **Figure 6a**, the change in full-sib family rankings (BLUPs of SCA effect) was plotted for site 7 (a relatively low site loading for SCA effect) and site 3 (a relatively high site loading for SCA effect) and found that they showed similar rankings for XFA1 model (Model 8). Similarly, when BLUP of SCA effects were plotted against all the site loadings, we observed fewer rank changes compared to GCA effects of parents (**Figure 6b**).

The individual narrow sense heritabilities for height based on different variance structures for the **G** side of the mixed model varied between sites (**Table 6**). The narrow sense heritabilities were low for simple variance structures, such as IID and DIAG and the estimates ranged between 0.03 and 0.26 among 10 sites with an average of 0.13. We see a large increase in narrow-sense heritabilities for more complex structures. For example, for XFA and US, the range of narrow-sense heritability was 0.08 to 0.33 with a mean of 0.19.

The broad sense heritabilities were high and varied between 0.25 and 0.90 across the ten sites (**Table 7**). Although not as strikingly different as narrow-sense heritability estimates, we

again observed considerably higher broad-sense heritability estimates for complex variance structures, such as XFA and US. The mean broad-sense heritability estimates across 10 sites were 0.48 for IID and DIAG versus 0.55 for more complex variance structures.

The GCA and SCA genetic correlations between pairs of sites for the best fit model (Model 8 in **Table 4**) are given in **Table 8** and **Table 9**. Genetic correlations from XFA models are given in in the upper triangles, and genetic correlations from US models are given in the lower triangles. The GCA genetic effect correlations between pair of sites were positive and ranged between 0.43 and 0.98 (**Table 8** upper triangle). The correlations of GCA genetic effects between site 2 and other sites were low and they varied between 0.43 and 0.73. The correlations of SCA genetic effects between pair of sites were over 0.80 (**Table 9** upper triangle). The low correlations of site 2 with other sites were obvious. Similarly, site 7 had noticeable low correlations compared to correlations between other sites. Visual illustrations of these tables are given in Appendix 2. Correlations based on XFA and US structures falling on the 45 degree reference line suggest that XFA and US structures produce similar correlations between pair of sites. These correlations show the relative performance of parents between pair of sites. Non-additive genetic correlations suggest correspondence between predicted SCA effects of full-sib families between pairs of sites. A high correlation suggests little interaction of parents or full-sib families with sites; whereas a low correlation suggests considerable rank changes among parents (or full-sib families) between two sites. XFA1 produced higher non-additive genetic correlations compared to US structure for the same pairs of sites as seen dots above the 45 degree reference line. Except a few low

correlations of site 2 with several other sites a high majority of additive genetic correlations between pairs of sites were above 0.70. Scatter plots of parental breeding values between pairs of sites are presented in Appendix 2. In general we observed a good agreement between pairs of sites, except for site 2 and site 7. Predicted parental breeding values for these two sites (with low loadings) clearly do not show strong correlations with predictions at the remaining sites.

Discussion

Model comparison

In tree breeding programs MET data is common and shows heterogeneity in variance estimates. In this study, we tested the efficiency of factor analytic (FA) covariance structures for the analysis of MET progeny trials of loblolly pine. In doing so we compared the model efficiency with various other variance-covariance structures commonly used in analysis of data from tree breeding progeny tests, and addressed GxE interactions. The reduced LogL values for FA and US variance structures compared with the fitted simplistic models suggests that the heterogeneity can be more efficiently modeled using these structures in analysis of data from multi-environment progeny trials of forest trees.

We found that FA models were superior over other variance-covariance structures for MET, and our results are in parallel with previously published papers. Specifically, we found that the XFA2 structure for GCA effect and XFA1 for the SCA effect gave the best fit among tested models. Kelly et al. (2007) analyzed eight MET data sets from Australian crop breeding programs by fitting a range of genetic variance structures. They found that the FA model provided the best fit for six of the eight early generation data sets. In another study of Kelly et al. (2009), a model including a FA3 structure for additive effects, and a FA1 structure for non-additive effects gave the best fit to barley data consisting of 1255 genotypes across 14 sites. Since FA parameterization directly models the pattern in the similarity of genetic response across environments, it reduces the number of parameters to be estimated (Hardner et al 2010). There are few published studies with factor analytic models in forest

progeny tests. Costa e Silva et al. (2006) used a FA structure to investigate 578 open pollinated *Eucalyptus globulus* families from 15 subraces across 15 sites. Similarly, Costa e Silva and Gradual (2008) evaluated 28 *Pinus kesiya* provenances across seven sites by fitting FA models. Hardner et al. (2010) found that FA1 structure for the family by trial effect and FA2 structure for the genotype within family by trial effects provided the most parsimonious model for a highly unbalanced data set collected at 3 years across 21 sites for a total of 841 genotypes from 10 *Eucalyptus spp.* hybrid families. The accuracy of estimation of genetic effects can be improved when a multivariate analysis is used to include information from other sites, especially in cases of large imbalance across sites (Hardner et al. 2010).

The US variance structure was one of the fitted variance structures to the GCA and SCA effects in this study. With the XFA2/XFA1 model, 49 parameters are estimated compared to 110 parameters for the US model. Although the FA structure is a simpler approximation of a US structure, findings in this study are in agreement with other studies. Smith et al. (2001) demonstrated that a FA3 model with common specific variance was not significantly different from a US structure, but more parsimonious, fitted to data from 172 barley lines across seven sites. Similarly, Kelly et al. (2007) showed that a FA model was superior to a US model via simulations. However, Piepho (1998) did cross validation using five data sets and demonstrated that a US model provided better fit than a FA model with common specific variance. The convergence of FA models is sensitive to starting values of loadings and specific variances or variances.

Percent variance explained by factors in FA models

In this study, on average 92% of the variation in GCA effects and 90% of the variation in SCA effects across sites were explained by XFA2 and XFA1 models, respectively. Similarly, Costa e Silva et al. (2006) found that in a *Eucalyptus globulus* progeny test with 15 subraces across 15 sites, one factor model explained 74% of the variation. Similarly, one factor analytic model explained 73% of the variance in a *Pinus kesiya* provenance trial across seven sites (Costa e Silva and Gradual 2008). In *Eucalyptus* hybrid families, a FA1 model explained 83% of the variation in family effects and FA2 model explained 86% of the variation in genotype within family effects across 21 trials (Hardner et al. 2010). Smith et al. (2001), using data from 172 barley lines collected across seven sites, determined that a FA3 model with common specific variance was the best model, where the first factor explained the most of the variation (86.2%) in genetic effects and the second and third factors explained only 9.3% and 4.5% variation in genetic effects, respectively.

Modeling GxE using FA models

GxE interaction occurs when the genetic value of an individual varies across environments. Modeling GxE is an important procedure in determining the magnitude and pattern of GxE, to design breeding programs and draw conclusions for deployment of selected genotypes. Based on the FA model results, plotting of environmental vectors may be used to summarize and understand GxE visually as how genotype response changes across sites (Kelly et al. 2007). For a set of environments, genotype (cultivar) stability can be graphically displayed by plotting cultivar scores against site loadings (Smith et al. 2002). Since the FA model

allows variance heterogeneity and covariance structure at the genetic level, it gives more accurate predictions for individual sites and for net merit across sites, resulting in more precise heritability estimates, and thus more accurate estimates of genetic gain in breeding programs. Kelly et al. (2007) reported that the FA model gave better results in terms of mean squared error of prediction over an independent site analysis or a simple variance model.

In this study, the pattern of GxE interaction for additive effect (GCA) was mostly defined by two environments (sites 2 and 7), and these two sites had clearly low variation explained by the first site loading for the GCA effect and formed a distinct group from the other sites (Figure 4). The most significant changes in ranking were observed between sites 2 and 7 with other sites. The difference between these two environments and others could be explained by micro environmental factors, such as high water table in the soil or drainage conditions, though we did not have details about soil properties. On the other hand, the magnitude of GxE was much smaller as suggested by higher non-additive genetic correlation (SCA effect) between pairs of sites (Table 9).

However, when all the correlations among pairs of sites were taken into account, GxE does not appear to be important enough to justify different breeding zones for this population of loblolly pine. Even the most interactive environments, at sites 2 and 7, had on average 0.64 and 0.71 additive genetic correlations with other environments. Shelbourne (1972) proposed 0.67 as a genetic correlation threshold for GxE interaction and a genetic correlation greater than 0.67 would not be a concern for breeding and deployment programs (McKeand et al. 2006). On average, all 10 environments had an additive correlation of 0.83 and an average of

0.91 non-additive genetic correlation, suggesting that GxE should not be a concern for this specific population in the environments where genotypes were tested. The test sites cover a wide geographic area in the southern USA, from the Atlantic coast with mild winters and evenly distributed rainfall patterns through the year to more inland, drier and colder sites in Arkansas (Appendix 4).

Genetic parameter estimates

For a tree breeding program, modeling the GxE and heterogeneity in MET is crucial, because it affects the variance components, heritability and thus genetic gain estimates. In this study, significant heterogeneity was observed in GCA and SCA genetic variances and in error variances. We observed a stark increase in heritability estimates for individual sites when US and FA models were fit to account for variance heterogeneity and correlations between pairs of environments. Heterogeneity in genetic variances and in heritability could be as a result of differential expression of genotypes across environments (Hardner et al. 2010, Clarke 1993).

In this study, a large proportion of the total genetic variation was due to SCA effects. Tree improvement strategies for using additive and non-additive variation are well known (McKeand et al. 1986, Lambeth et al. 1994). The question is whether to breed for improved non-additive variation as well as additive variation (McKeand et al. 1986). The level of additive and non-additive genetic variance in traits is important for tree breeding programs and has a great impact on the determination of the breeding strategies (Foster and Shaw 1987). Estimation of non-additive genetic variances for loblolly pine might help to develop more efficient breeding strategies to maximize genetic gain (McKeand et al. 1986). Wu and

Matheson (2004) reported significant SCA variance and a high ratio of SCA/GCA variance for diameter using 20 sets of 6x6 half-diallel mating design in *Pinus radiata* planted at 10 sites across Australia. Isik et al. (2003) found that additive genetic effects were the major source of genetic variation for height in early ages and there was a little increase in non-additive variance at age six using nine full-sib loblolly pine families generated from a factorial mating design planted across two sites. In the same study, the ratio of dominance to additive variance for height was reported to be 0.76. McKeand et al. (1986) reported that for height at age five, the ratio of non-additive to additive genetic variance in loblolly pine ranged from 0 to 2.5. In another study in loblolly pine, the ratio of non-additive to additive genetic variances varied between 0 and 3.4 for growth traits at ages one to five (Paul et al. 1997). Substantial SCA variance observed in this study suggest that, with the deployment of specific crosses, genetic gain for growth can be increased substantially.

Conclusion

Among all models tested, FA models were best able to account for heterogeneity in GCA and SCA variances across multiple environments. Factor analytic models provide more accurate predictions of genetic values of individual genotypes, because the models capture both variance and the covariance at the genetic level better then simpler models. FA models also provide a natural framework for modeling GxE interaction, and they reduce computational requirements of mixed model analysis. They should be routinely used for MET trials in forest tree breeding programs.

References

Burgueño, J., Crossa, J., Cornelius, P.L., Trethowan, R., McLaren, G., Krishnamachari, A., 2007. Modeling Additive × Environment and Additive × Additive × Environment Using Genetic Covariances of Relatives of Wheat Genotypes. Crop Science 47, 311-320.

Clarke, G.M., 1993. The genetic basis of developmental stability . I . Relationships between stability , heterozygosity and genomic coadaptation. Genetica 89, 15-23.

Costa e Silva, J., Dutkowski, G.W., Gilmour, A.R., 2001. Analysis of early tree height in forest genetic trials is enhanced by including a spatially correlated residual. Canadian Journal of Forest Research 31, 1887-1893.

Costa e Silva, J., Graudal, L., 2008. Evaluation of an international series of Pinus kesiya provenance trials for growth and wood quality traits. Forest Ecology and Management 255, 3477-3488.

Costa e Silva, J., Potts, B.M., Dutkowski, G.W., 2006. Genotype by environment interaction for growth of Eucalyptus globulus in Australia. Tree Genetics & Genomes 2, 61-75.

Crossa, J., Burgueño, J., Cornelius, P.L., McLaren, G., Trethowan, R., Krishnamachari, A., 2006. Modeling Genotype × Environment Interaction Using Additive Genetic Covariances of Relatives for Predicting Breeding Values of Wheat Genotypes. Crop Science 46, 1722-1733.

Cullis, B., Gogel, B., Verbyla, A., Thompson, R., Verbyla, A., 1998. Spatial Analysis of Multi-Environment Early Generation Variety Trials. Biometrics 54, 1-18.

Dutkowski, G.W., Costa, J., Gilmour, A.R., Lopez, G.A., 2002. Spatial analysis methods for forest genetic trials. Canadian Journal of Forest Research 32, 2201-2214.

Edwards, J.W., Jannink, J.-L., 2006. Bayesian Modeling of Heterogeneous Error and Genotype × Environment Interaction Variances. Crop Science 46, 820-833.

Gilmour, A.R., B.J. Gogel, B.R. Cullis, and R. Thompson. 2009. ASReml user guide release 3.0. VSN International Ltd., Hemel Hempstead, HP1 1ES, United Kingdom. 372 p.

Gilmour, A.R., Thompson, R., Cullis, B.R., 1995. Average Information REML : An Efficient Algorithm for Variance Parameter Estimation in Linear Mixed Models. Biometrics 51, 1440-1450.

Goebel, N.B, and J.R. Warner. 1966. Total and bark volume tables for small diameter loblolly, shortleaf, and Virginia pine in the upper South Carolina piedmont. Forest Research Series No. 9 Clemson University, Clemson, South Carolina, USA.

Hardner, C.M., Dieters, M., Dale, G., DeLacy, I., Basford, K.E., 2010. Patterns of genotype-by-environment interaction in diameter at breast height at age 3 for eucalypt hybrid clones grown for reafforestation of lands affected by salinity. Tree Genetics & Genomes 6, 833-851.

Henderson, C.R. 1984. Applications of linear models in animal breeding. University of Guelph, Guelph, Canada. 462 p.

Isik, F., Boos, D.D., Li, B., 2005. The distribution of genetic parameter estimates and confidence intervals from small disconnected diallels. Theoretical and Applied Genetics 110, 1236-1243.

Johnson, G.R., King, J.N., 1998. Analysis of Half Diallel Mating Designs I – A Practical Analysis Procedure for ANOVA Approximation. Silvae Genetica 47, 74-79.

Kelly, A.M., Cullis, B.R., Gilmour, A.R., Eccleston, J. A, Thompson, R., 2009. Estimation in a multiplicative mixed model involving a genetic relationship matrix. Genetics Selection Evolution 41, 1-9.

Kelly, A.M., Smith, A.B., Eccleston, J.A., Cullis, B.R., 2007. The Accuracy of Varietal Selection Using Factor Analytic Models for Multi-Environment Plant Breeding Trials. Crop Science 47, 1063-1070.

Lynch, M., and Walsh, B. 1998. Genetics and analysis of quantitative traits. Sinauer Associates, Inc., Sunderland, MA, USA, 980 p.

Lambeth, C., Endo, M., Wright, J., 1994. Genetic-analysis of 16 clonal trials of Eucalyptus grandis and comparisons with seedling checks. For. Sci. 40, 397–411.

McKeand, S., Mullin, T., Byram, T., White, T., 2003. Deployment of Genetically Improved Loblolly and Slash Pines in the South. Journal of Forestry 32-37.

McKeand, S.E., Li, B., Grissom, J.E., Isik, F., Jayawikrama, K.J.S., 2006. Genetic Parameter Estimates for Growth Traits from Diallel Tests of Loblolly Pine throughout the Southeastern United States. North 57, 101-110.

McKeand, S.E., Bridgwater, F.E. 1986. When to establish advanced generation seed orchards. Silvae genetica. 35(5-6), p.245.

Meyer, K., 2009. Factor-analytic models for genotype x environment type problems and structured covariance matrices. Genetics Selection Evolution 41, 1-11.

Mrode, R. 2005. Linear models for the prediction of animal breeding values, 2nd edition, CAB international , Wallingford, Oxfordshire, OX10 8DE, UK.

Oakey, H., Verbyla, A., Pitchford, W., Cullis, B., Kuchel, H., 2006. Joint modeling of additive and non-additive genetic line effects in single field trials. Theoretical and applied genetics 113, 809-819.

Oakey, H., Verbyla, A.P., Cullis, B.R., Wei, X., Pitchford, W.S., 2007. Joint modeling of additive and non-additive (genetic line) effects in multi-environment trials. Theoretical and applied genetics. 114, 1319-1332.

Piepho, H.-P., 1998. Empirical best linear unbiased prediction in cultivar trials using factor-analytic variance-covariance structures. Theoretical and Applied Genetics 97, 195-201.

Piepho, H.P., Möhring, J., Melchinger, A.E., Büchse, A., 2008. BLUP for phenotypic selection in plant breeding and variety testing. Euphytica 161, 209-228.

SAS Institute Inc. 2010. SAS 9.1 User‟s guide statistics. SAS Institute Inc., Cary, NC, USA

Shelbourne, C. J. A. (1972): Genotype-environment interaction: its study and its implications in forest tree improvement. In: Proc. IUFRO Genetics-SABRAO, Jt. Symp., Tokyo, Japan. B-1 (I): 1–28.

Signor, C.E.-L., Dousse, S., Lorgeou, J., Bonhomme, R., Carolo, P., Charcosset, A., 2001. Interpretation of Genotype X Environment Interactions for Early Maize Hybrids over 12 Years. Crop Science 41, 663-669.

Smith, A., B. Cullis, and R. Thompson. 2002. Exploring variety–environment data using random effects AMMI models with adjustments for spatial field trend: Part 1: Theory. In: Kang, M.S. (ed.). 2002. Quantitative genetics, genomics and plant breeding. Oxford University Press, Cary, USA. 416 p.

Smith, A., Cullis, B., Luckett, D., Hollamby, G., Thompson, R., 2002. 22 Exploring Variety – Environment Data Using Random Effects AMMI Models with Adjustments for Spatial Field Trend : Part 2 : Applications. Quantitative genetics, genomics and plant breeding 337-351.

Smith, A., Cullis, B., Thompson, R., 2001. Analyzing Variety by Environment Data Using Multiplicative Mixed Models and Adjustments. Biometrics 57, 1138-1147.

Smith, A.B., Cullis, B.R., Thompson, R., 2005. The analysis of crop cultivar breeding and evaluation trials: an overview of current mixed model approaches. Journal of Agricultural Science 143, 449-462.

So, Y.-S., Edwards, J., 2009. A Comparison of Mixed-Model Analyses of the Iowa Crop Performance Test for Corn. Crop Science 49, 1593-1601.

Society, I.B., Statistics, E., 1997. Accounting for Natural and Extraneous Variation in the Analysis of Field Experiments. Journal of Agricultural, Biological, and Environmental Statistics 2, 269-293.

Spilke, J., Piepho, H.P., Hu, X., 2005. Analysis of Unbalanced Data by Mixed Linear Models Using the mixed Procedure of the SAS System. Journal of Agronomy and Crop Science 191, 47-54.

Thompson, R., Cullis, B., Smith, A., Gilmour, A., 2003. A Sparse Implementation of the Average Information Algorithm for Factor Analytic and Reduced Rank Variance Models. Aust. N.Z.J. Stat. 45, 445-459.

Wolfinger, R.D., 1996. Heterogeneous Variance-Covariance Structures for Repeated Measures. Journal of Agricultural, Biological, and Environmental Statistics 1, 205-230.

Wu, H.X., Matheson, A.C., 2004. General and specific combining ability from partial diallels of radiata pine: implications for utility of SCA in breeding and deployment populations. Theoretical and applied genetics 108, 1503-1512.

Wu, H.X., Matheson, A.C., 2005. Genotype by Environment Interactions in an Australia-Wide Radiata Pine Diallel Mating Experiment : Implications for Regionalized Breeding. Forest Science 51, 29-40.

White, T.L., W.T. Adams, and D.B. Neale. 2007. Forest genetics. CABI Publishing CAB International. Cambridge, MA 02139 UK, USA. 702 p.

Xiang, B., Li, B., 2001. A new mixed analytical method for genetic analysis of diallel data. Canadian Journal of Forest Research 31, 2252-2259.

Yan, W., Kang, M.S., Ma, B., Woods, S., Cornelius, P.L., 2007. GGE Biplot vs. AMMI Analysis of Genotype-by-Environment Data. Crop Science 47, 641-653.

Yanchuk, A.D., 1996. General and Specific Combining Ability from Disconnected Partial Diallels of Coastal Douglas-fir. Silvae Genetica 45, 37-45.

Yang, R.-C., Crossa, J., Cornelius, P.L., Burgueño, J., 2009. Biplot Analysis of Genotype × Environment Interaction: Proceed with Caution. Crop Science 49, 1564-1576.

Zas, R., 2008. The impact of spatial heterogeneity on selection: a case study on Pinus pinaster breeding seedling orchards. Canadian Journal of Forest Research 38, 114-124.

Zobel, B.J., and Talbert, J.T. 1984. Applied Forest Tree Improvement. John Wiley & Sons, New York, 505 p.

TABLES

Table 1. Site IDs, locations and establishment year of the progeny tests, number of crosses (full-sib families), and number of checklots per site.

Site ID	State	Year established	# of Cross	# of Checklots
1	Georgia	2003	140	5
2	N Louisiana	2003	115	7
3	Arkansas	2003	122	7
4	Georgia	2004	140	4
5	Georgia	2004	140	4
6	Georgia	2004	140	4
7	S Mississippi	2004	141	3
8	S Mississippi	2004	141	2
9	Arkansas	2004	125	2
10	N Louisiana	2004	125	3

*Checklots are either unimproved bulked seedlings or wind pollinated (half-sib) families included in the progeny tests to estimate genetic gains.

Table 2. Models and variance-covariance structures fit to GCA and SCA genetic effects and block effects (**B**) and error (**R**) for height.

Model	GCA effect	SCA effect	B	R
1	IID	IID	IID/DIAG	IID/DIAG
2	DIAG	DIAG	IID/DIAG	IID/DIAG
3	CS	CS	IID/DIAG	IID/DIAG
4	FA1	FA1	IID/DIAG	IID/DIAG
5	FA2	FA2	IID/DIAG	IID/DIAG
6	XFA1	XFA1	IID/DIAG	IID/DIAG
7	XFA1	XFA2	IID/DIAG	IID/DIAG
8	XFA2	XFA1	IID/DIAG	IID/DIAG
9	XFA2	XFA2	IID/DIAG	IID/DIAG
10	US	US	IID/DIAG	IID/DIAG

*IID, identity variance; DIAG, diagonal variance; CS, compound symmetry; US, unstructured variance; FA, factor analytic variance; XFA, extended factor analytic variance.

Table 3. Model fit statistics REML log likelihood (LogL) and Akaike Information Criterion (AIC) when block within site and error variances were modeled as IID for height. When errors are assumed to have simple homogenous structure (all trees have the same identical residuals), models 4 and 5 with the FA1 and XFA1 genetic structures, respectively, were superior with the lowest AIC.

Model	GCA effect	SCA effect	# parameters	LogL	AIC
1	IID	IID	2	-3421.7	6847.4
2	DIAG	DIAG	20	-3394.9	6829.7
3	CS	CS	4	-2912.9	5833.8
4	**FA1**	**FA1**	**40**	**-2862.8**	**5805.6**
5	**XFA1**	**XFA1**	**40**	**-2862.8**	**5805.6**
6	US	US	110	-2815.9	5851.8

Table 4. Model fit statistics LogL and AIC when block and error variances were modeled heterogeneous (Diagonal). Among the models, Model 8 with XFA2 for GCA and XFA1 for SCA genetic effects was superior.

Model	GCA effect	SCA effect	# parameters	LogL	AIC
1	IID	IID	2	-3183.5	6370.9
2	DIAG	DIAG	20	-3156.7	6353.4
3	CS	CS	4	-2664.2	5336.4
4	FA1	FA1	40	-2614.2	5308.3
5	FA2	FA2	58	-2605.2	5326.4
6	XFA1	XFA1	40	-2614.2	5308.3
7	XFA1	XFA2	49	-2607.8	5313.6
8	**XFA2**	**XFA1**	**49**	**-2599. 5**	**5296.9**
9	XFA2	XFA2	58	-2592.9	5301.9
10	US	US	110	-2575.4	5370.8

Table 5. Estimated model parameters fit to GCA and SCA effects, respectively for Model 8 (XFA2 for GCA and XFA1 for SCA). $\boldsymbol{\sigma_a^2}$ is GCA variance, $\boldsymbol{\Lambda_{a1}}$ and $\boldsymbol{\Lambda_{a2}}$ are first and second site loadings for the GCA effect, $\boldsymbol{\Psi_a}$ is GCA variance not explained by site loadings (lack of fit), $\boldsymbol{\sigma_d^2}$ is SCA variance, $\boldsymbol{\Lambda_d}$ is site loadings for the SCA effect, $\boldsymbol{\Psi_d}$ is SCA variance not explained by site loadings and % is the percentage of variance explained by the loadings of first and second factor.

	GCA effect						**SCA effect**				
Site	σ^2_a	Λ_{a1}	Λ_{a2}	Ψ_a	% Λ_{a1}	% Λ_{a2}	σ^2_d	Λ_d	Ψ_d	% Λ_d	$\sigma^2_d/\ \sigma^2_a$
1	13	-3.6	0.2	0.3	97	0	16	4.0	0.5	97	**1.2**
2	7	-1.8	-0.5	3.0	50	5	17	3.8	2.8	84	**2.6**
3	12	-3.4	-0.1	0.6	95	0	32	5.5	1.9	94	**2.6**
4	10	-2.9	-1.3	0.0	83	17	12	3.5	0.0	100	**1.2**
5	14	-3.7	-0.7	0.0	97	3	15	3.9	0.3	98	**1.0**
6	15	-3.8	0.2	0.3	98	0	26	4.9	2.4	91	**1.7**
7	2	-1.1	0.8	0.3	58	27	12	3.0	3.0	75	**5.7**
8	12	-3.4	0.7	0.0	96	4	22	4.2	4.0	82	**1.8**
9	12	-3.2	0.5	1.0	89	2	21	4.5	1.2	94	**1.8**
10	4	-1.8	1.0	0.0	78	21	19	4.3	1.1	94	**4.6**
Mean	10				84	8	19			91	

*The estimated loadings for GCA effects from XFA2 model given in the table obtained after the rotation

Table 6. Individual tree narrow sense heritabilities and standard errors within the parenthesis for the range of models fit to GCA and SCA genetic effects for all 10 progeny test sites for height. The XFA2/XFA1 and US variance-covariance structures gave similar heritability results.

SiteID	IID	DIAG	CS	XFA2/XFA1	US
1	0.15 (0.02)	0.15 (0.06)	0.23 (0.06)	0.27 (0.04)	0.23 (0.08)
2	0.12 (0.02)	0.12 (0.05)	0.18 (0.05)	0.11 (0.03)	0.14 (0.05)
3	0.12 (0.02)	0.24 (0.08)	0.19 (0.06)	0.20 (0.04)	0.23 (0.08)
4	0.15 (0.02)	0.14 (0.05)	0.23 (0.06)	0.22 (0.04)	0.22 (0.08)
5	0.15 (0.02)	0.23 (0.08)	0.24 (0.07)	0.31 (0.05)	0.28 (0.09)
6	0.17 (0.02)	0.26 (0.09)	0.26 (0.07)	0.33 (0.05)	0.35 (0.11)
7	0.11 (0.02)	0.03 (0.02)	0.17 (0.05)	0.04 (0.02)	0.06 (0.03)
8	0.12 (0.02)	0.03 (0.02)	0.18 (0.05)	0.19 (0.04)	0.18 (0.08)
9	0.13 (0.02)	0.07 (0.04)	0.20 (0.06)	0.21 (0.04)	0.20 (0.08)
10	0.13 (0.02)	0.05 (0.03)	0.20 (0.06)	0.08 (0.02)	0.11 (0.05)
Mean	0.13	0.13	0.20	0.19	0.20

Table 7. Individual tree broad sense heritabilities and standard errors within the parenthesis for the range of models fit to genetic effects GCA and SCA.

SiteID	IID	DIAG	CS	XFA2/XFA1	US
1	0.54 (0.03)	0.48 (0.07)	0.62 (0.06)	0.61 (0.05)	0.58 (0.08)
2	0.43 (0.02)	0.39 (0.07)	0.49 (0.05)	0.41 (0.04)	0.44 (0.07)
3	0.45 (0.02)	0.74 (0.09)	0.52 (0.06)	0.73 (0.05)	0.73 (0.09)
4	0.53 (0.03)	0.38 (0.06)	0.61 (0.06)	0.48 (0.04)	0.48 (0.08)
5	0.55 (0.03)	0.55 (0.08)	0.63 (0.07)	0.62 (0.05)	0.61 (0.09)
6	0.62 (0.03)	0.82 (0.09)	0.70 (0.07)	0.90 (0.05)	0.89 (0.09)
7	0.41 (0.02)	0.22 (0.05)	0.47 (0.05)	0.25 (0.04)	0.26 (0.06)
8	0.42 (0.02)	0.41 (0.06)	0.49 (0.05)	0.54 (0.04)	0.53 (0.08)
9	0.46 (0.02)	0.44 (0.07)	0.53 (0.06)	0.58 (0.05)	0.57 (0.08)
10	0.46 (0.02)	0.37 (0.06)	0.53 (0.06)	0.45 (0.04)	0.44 (0.07)
Mean	0.48	0.48	0.55	0.55	0.55

Table 8. Genetic correlations of GCA genetic effects between pair of progeny test sites from the XFA2 structure (above diagonal) based on model 8 and the US model (model 10) (below diagonal) for height. XFA2 and US variance-covariance structures fit to GCA genetic effects gave similar genetic correlations between pair of sites. For the visual display of the same table see Appendix 2. The values in the diagonal are the average additive genetic correlations for environments based on FA model.

SiteID	1	2	3	4	5	6	7	8	9	10
1	0.89	0.69	0.96	0.88	0.96	0.98	0.77	0.98	0.94	0.89
2	0.61	**0.64**	0.69	0.73	0.73	0.68	0.43	0.65	0.64	0.53
3	0.92	0.77	0.88	0.90	0.97	0.96	0.73	0.95	0.92	0.85
4	0.88	0.75	0.91	0.79	0.97	0.88	0.48	0.81	0.80	0.62
5	0.96	0.70	0.93	0.96	0.87	0.96	0.66	0.93	0.90	0.79
6	0.97	0.71	0.94	0.88	0.94	0.91	0.78	0.98	0.94	0.90
7	0.73	0.54	0.72	0.55	0.63	0.83	**0.71**	0.85	0.79	0.91
8	0.97	0.62	0.92	0.81	0.92	0.97	0.83	0.90	0.95	0.96
9	0.93	0.66	0.94	0.80	0.90	0.90	0.73	0.95	0.86	0.90
10	0.84	0.65	0.90	0.68	0.77	0.91	0.92	0.93	0.90	0.82

Table 9. Genetic correlations of SCA genetic effects between pair of progeny test sites from the XFA1 structure (above diagonal) based on Model 8 and US model (Model 10) (below diagonal). XFA1 and US variance-covariance structures fit to SCA genetic effects gave similar genetic correlations between pair of sites. For the visual display of the same table see Appendix 2.

SiteID	1	2	3	4	5	6	7	8	9	10
1		0.90	0.96	0.99	0.98	0.94	0.86	0.89	0.96	0.96
2	0.90		0.89	0.91	0.91	0.87	0.79	0.83	0.89	0.89
3	0.95	0.91		0.97	0.96	0.92	0.84	0.88	0.94	0.94
4	0.94	0.83	0.92		0.99	0.95	0.87	0.90	0.97	0.97
5	0.93	0.83	0.89	0.98		0.94	0.86	0.89	0.96	0.96
6	0.92	0.86	0.93	0.92	0.92		0.83	0.86	0.93	0.92
7	0.81	0.78	0.73	0.82	0.83	0.80		0.78	0.84	0.84
8	0.88	0.74	0.81	0.85	0.90	0.90	0.78		0.88	0.88
9	0.94	0.81	0.92	0.94	0.93	0.91	0.86	0.88		0.94
10	0.98	0.90	0.94	0.91	0.89	0.92	0.81	0.86	0.92	

FIGURES

F/M	1	2	3	4	5	6	7	8	9	18	19	20	21	22	10	11	12	13	14
1		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X		X	
2		S	X	X	X	X	X	X	X	X	X	X	X	X	X	X		X	X
3			S	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	
4				S	X	X	X	X	X	X	X	X	X	X	X	X		X	
5					S	X	X	X	X	X	X	X	X	X	X	X		X	
6							X	X	X	X	X	X	X	X	X	X		X	
7							S	X	X	X	X	X	X	X	X	X		X	
8								S	X	X	X	X	X	X	X	X		X	
9									S	X	X	X	X	X	X	X		X	
18										S	X	X	X	X	X	X		X	
19											S	X	X	X	X	X		X	
20												S	X	X	X	X	X	X	
21													S	X	X	X		X	
22															X	X		X	
10																X		X	
11																		X	

*S is self; X is cross made. In addition to crosses made among parents two seed orchard mixes (bulked seeds called SOM6 and SOM8) were included in the field tests.

Figure 1. The diallel mating design of 19 parent trees used as female and male in the study. Parent trees 1 to 14 are from the first generation breeding population. Parent trees 18 to 22 are from the second generation breeding populations.

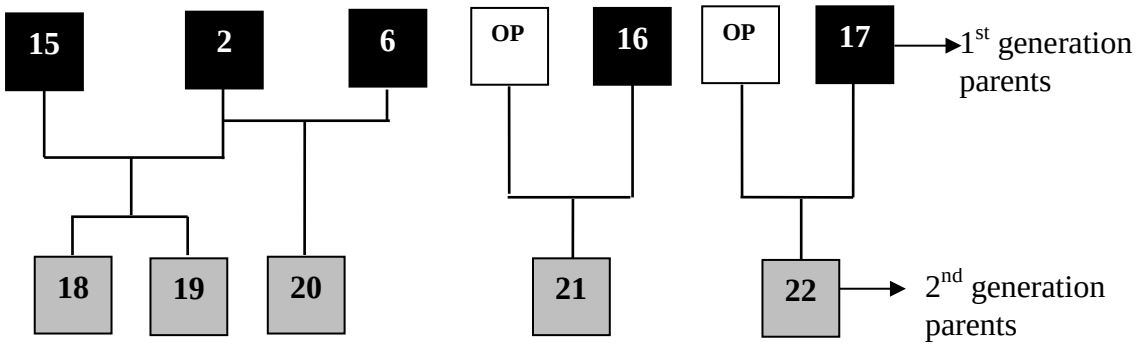


Figure 2. Pedigree structure of 2nd generation parent trees used in the diallel mating design. Black squares refer to the 1st generation parent trees, grey squares refer to the 2nd generation parent trees and white squares refer to the pollen mixed trees. Parent trees 2 and 6 are from the first generation and are related to the some of the second generation parent trees used as parent trees in diallel.

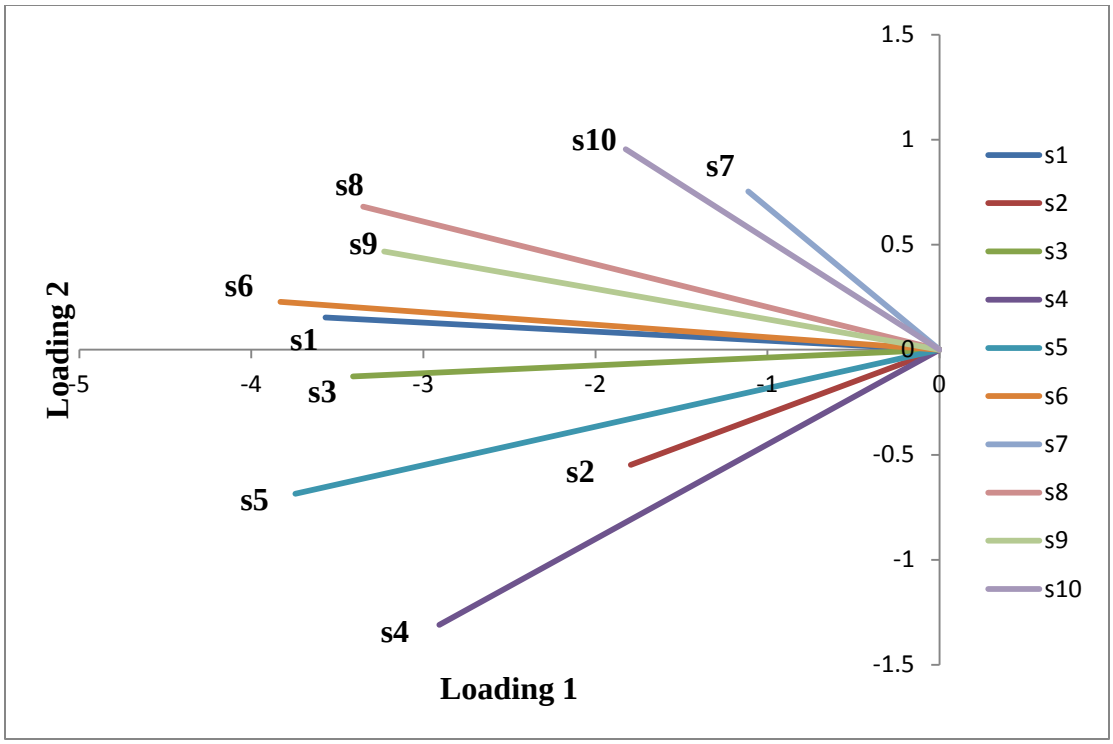


Figure 3. Plot of the site loadings for the first factor against second factor from XFA2 model fitted for GCA genetic effects for height. Site loadings are labeled from s1 to s10. The plot of the first loadings against second loadings shows three clusters, namely sites 7 and 10 consisted of first cluster, sites 1, 3, 6, 8 and 9 consisted of the second cluster and sites 2, 4 and 5 were in the third cluster. So, sites within these clusters were closely similar for this data set.

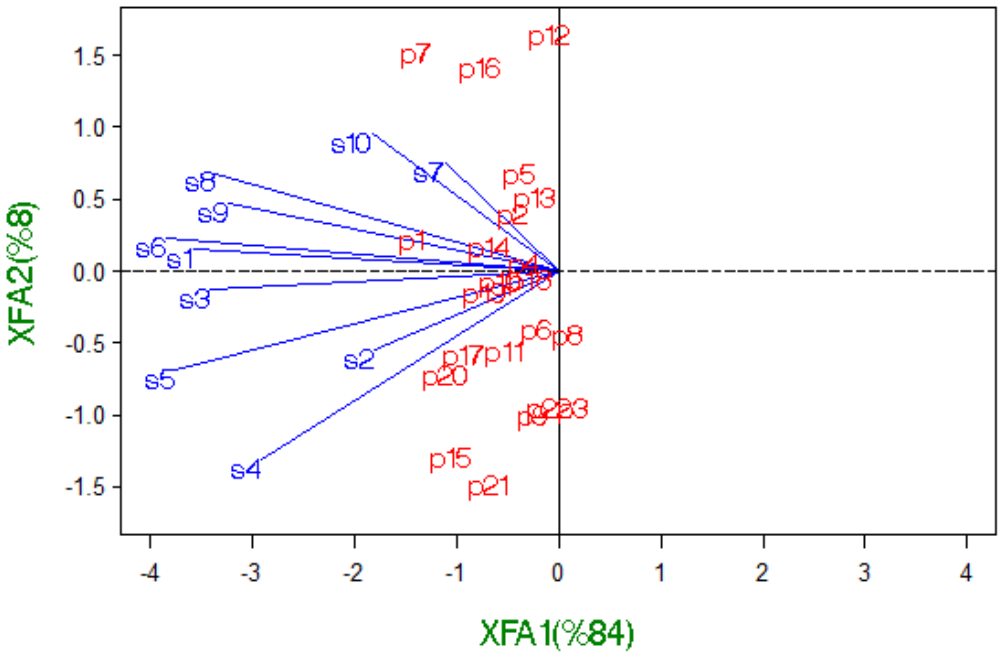


Figure 4. Biplot of XFA2 model for GCA genetic effects. Parents (genotype scores) are labeled p1 to p22, and site loadings are labeled as s1 to s10. Parents 3 and 8 tended to have negative GCA interaction with all the sites (GxE) included in the analysis. Parent 1, 17and 20 had positive response with all the sites. Sites 1, 3, 5, 6, 8 and 9, located farther away from the center, are the ones that discriminate the genotypes most and with higher correlations. Genotypes 4, 10, 14, 18 and 19 are clustered at the center of the biplot, indicating they had average response across all sites in terms of GxE. Parents 3 and 8 were poor parents in all sites.

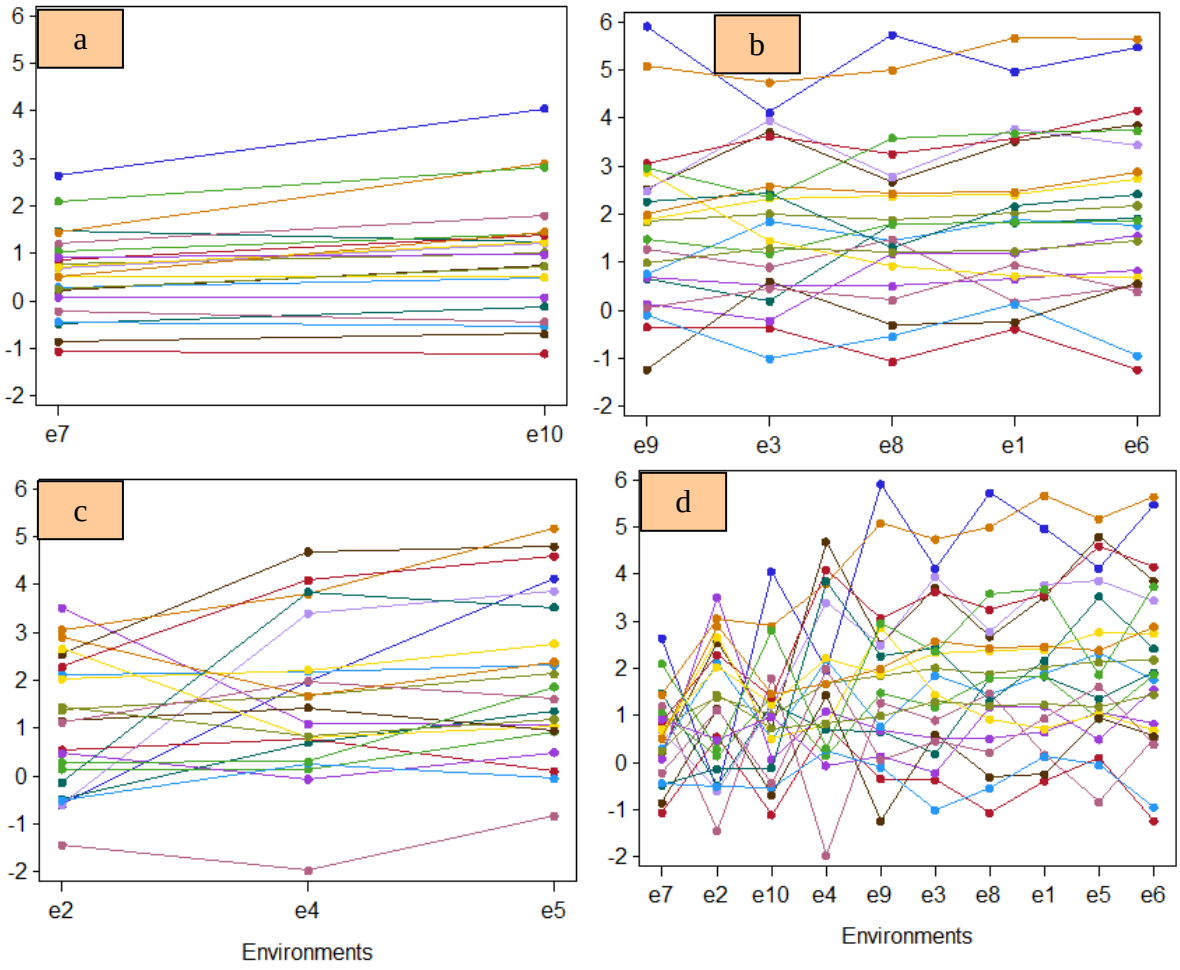


Figure 5. Line plots of the BLUP values of the parents for clustered sites from XFA2 model (Model 8) for height. The patterns of the first three plots (a, b, c) indicated that there were small rank changes of parents among environments with similar loadings (clustered). When parental predictions were plotted across all sites (d), we observe clear rank changes (GxE interaction). The last figure (d) also shows that as the site variances get larger, we observe a clear separation among parents.

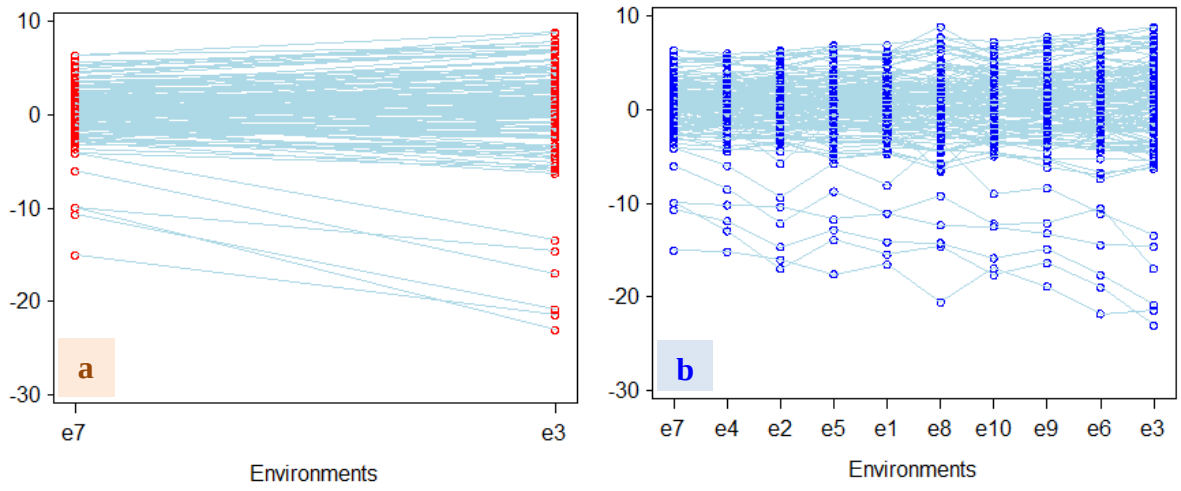


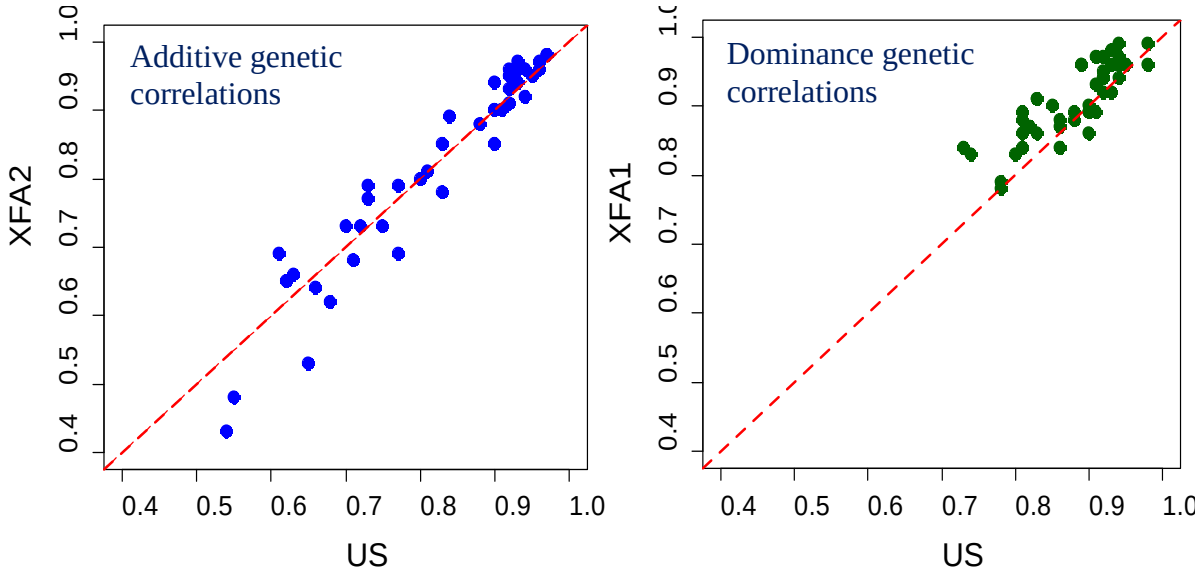
Figure 6. Plot of the BLUP values of full-sib families for the sites with highest and lowest site loadings (site 3 and 7 respectively) (a) and the line plots of full-sib families across all sites (b) are presented. In general, the ranking of the full-sib families were similar between sites 7 and 3. Also, we see more separation among families at the site 3 which had the highest variance (highest site loading). We also see full-sib families with low predictions show consistency across all sites (b).

APPENDICES

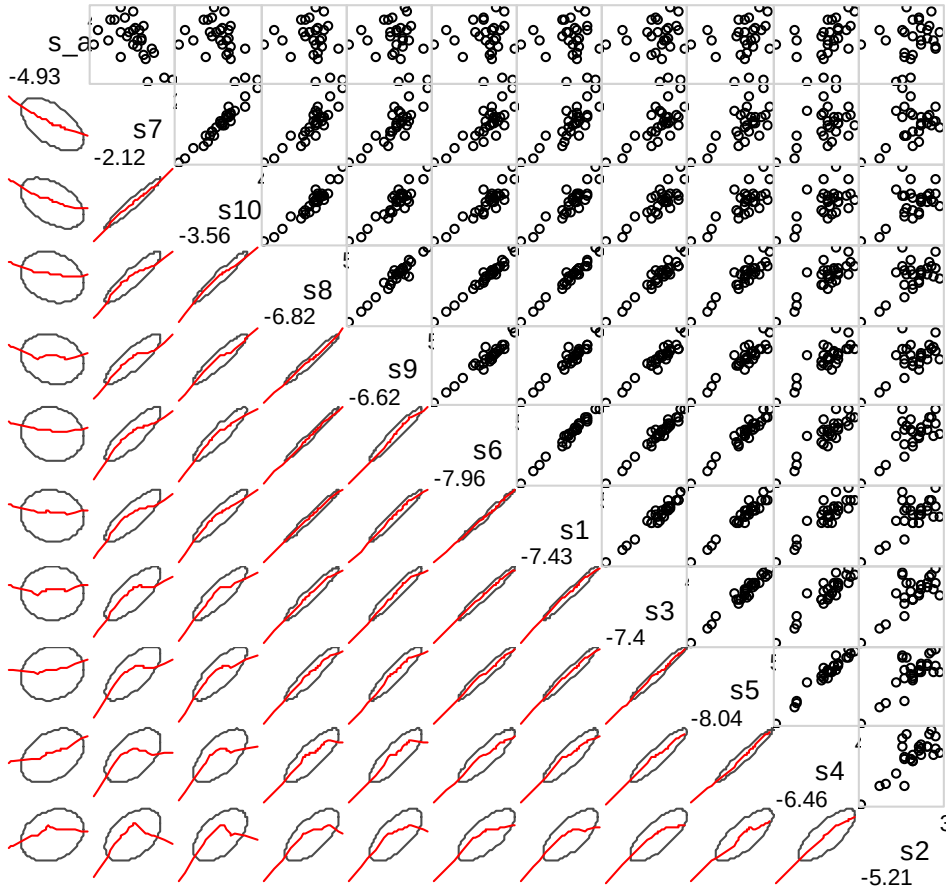
Appendix 1. Detailed information about the test sites. (D.W.T, depth to water table; F.F.P, frost free period)

Site ID	Water capacity	D.W.T (cm)	Drainage	Mean (m)	Soil Series	Precip. (mm)	Temp. (°C)	F.F.P
1	Low	0-30	Poor	7.9	Pelham loamy sand	1118-1321	18-21	230-290
2	High/ Moderate	46-122/ 30-76	Poor / Well	6.7	Frizzell-Providence-Guyton/ Savannah-Sacul Associations	1067-1575	16-21/ 16-19	200-350/ 190-230
3	High	0-30	Poor	7.7	Amy silt loam	1194-1676	9-23	205-275
4	Low	0-30	Poor	6.4	Plummer	1118-1321	18-21	230-290
5	Low/ Moderate	46-76/ 15-46	Poor	8.6	Olustee fine sand/Mascotte sand	1118-1321	18-21	230-290
6	Moderate	91-152	Well	8.3	Dothan loamy sand	1118-1321	18-21	230-290
7	High	15-30	Poor	8.0	Trebloc silt loam	1524-1905	18-21	270-335
8	Moderate	76-107	Well	9.1	Malbis fine sandy loam	1524-1905	18-21	270-335
9	Low	30-61	Well	6.7	Grenada silt loam	1194-1676	9-23	205-275
10	Very high/High	0-46/ 46-122	Poor	6.4	Guyton association/Frizzell silt loam	1067-1575	18-21	220-350/ 200-300

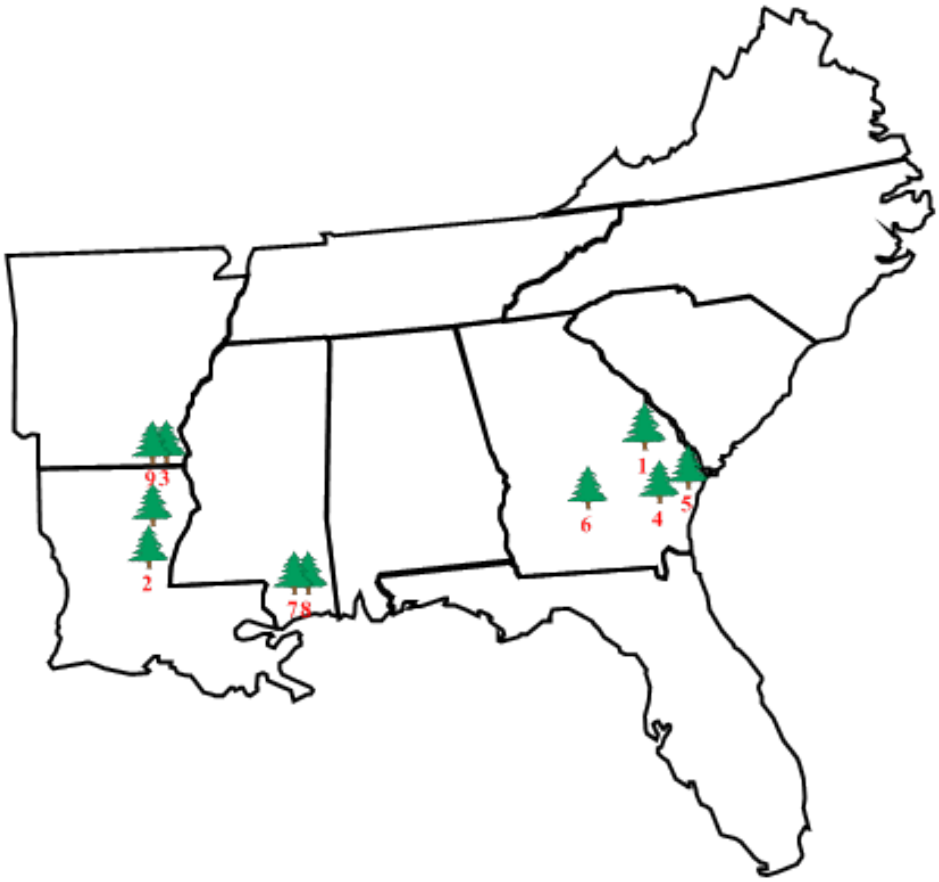
Appendix 2. Scatter plot of additive genetic correlations and dominance genetic correlations between pair of progeny test sites. XFA2 and US variance-covariance structures fitted to additive genetic effects gave similar genetic correlations between pair of sites. In contrast, XFA1 structure fit to dominance genetic effects tend to produce higher correlations between pair of sites compared to US structure.



Appendix 3. Pair-wise plots of parental breeding values for each site (labeled from s1-s10). In the diagonal of the figure the lowest and the highest breeding values for that specific site are reported. Notice that pair-wise plots of parental breeding between did not show any clear pattern between site 2 and other sites (right column or the bottom row matrix).



Appendix 4. Locations of 10 test sites.



Appendix 5. ASReml job file used for fitting DIAG, Block-DIAG, CS, FA*k*, XFA*k* and US variance structures to the additive and dominance genetic effects

```
title: diallel data
 tree      !A 37095
 female    !P
 male      !P
 cross     !A
 crosstype !A
 loc       !A
 test      *
 rep       *
 HT06 DBH06 VOL06 HTs VOLs
parent_ped.csv !SKIP 1   !ALPHA !SORT
PC.csv !SKIP 1 !DOPART $A !continue !MAXIT 50!MVINCLUDE


!PART 1
! DIAG(IID) G structure
$B ~ mu test !r test.rep female - male and(male) test.female -
test.male and(test.male) cross test.cross
10 1 1
3777  0 IDEN !s2=150.10
3336  0 IDEN !s2=197.44
3533  0 IDEN !s2=185.72
4144  0 IDEN !s2=155.68
4109  0 IDEN !s2=156.59
4166  0 IDEN !s2=149.74
2637  0 IDEN !s2=208.38
3910  0 IDEN !s2=200.38
3491  0 IDEN !s2=181.57
3370  0 IDEN !s2=182.16

test.rep 2
test 0 DIAG 28.1 16.0 11.1 54.8 35.4 12.6 11.0 9.26 27.5 33.5
rep 0 IDEN
```

Appendix 5. Continued

```
!PART 2
! Block-DIAG G structure
$B ~ mu test  !r  test.rep test.female -test.male
and(test.male) test.cross
10 1 3
3777  0 IDEN !s2=150.10
3336  0 IDEN !s2=197.44
3533  0 IDEN !s2=185.72
4144  0 IDEN !s2=155.68
4109  0 IDEN !s2=156.59
4166  0 IDEN !s2=149.74
2637  0 IDEN !s2=208.38
3910  0 IDEN !s2=200.38
3491  0 IDEN !s2=181.57
3370  0 IDEN !s2=182.16

test.rep 2
test 0 DIAG 28.1 16.0 11.1 54.8 35.4 12.6 11.0 9.26 27.5 33.5
rep 0 IDEN

test.female 2
test 0 DIAG 7 7 7 7 7 7 7 7 7 7
female 0 AINV

test.cross 2
test 0 DIAG 10 10 10 10 10 10 10 10 10 10
cross 0 IDEN


!PART 3
! COR_UV
$B ~ mu test  !r  test.rep test.female -test.male
and(test.male) test.cross
10 1 3
3777  0 IDEN !s2=150.10
3336  0 IDEN !s2=197.44
3533  0 IDEN !s2=185.72
4144  0 IDEN !s2=155.68
4109  0 IDEN !s2=156.59
4166  0 IDEN !s2=149.74
2637  0 IDEN !s2=208.38
3910  0 IDEN !s2=200.38
```

Appendix 5. Continued

```
3491  0 IDEN !s2=181.57
3370  0 IDEN !s2=182.16

test.rep 2
test 0 DIAG 28.1 16.0 11.1 54.8 35.4 12.6 11.0 9.26 27.5 33.5
rep 0 IDEN

test.female 2
test 0 CORUV 0.9 3
female 0 AINV

test.cross 2
test 0 CORUV 0.9 3
cross 0 IDEN


!PART 4
! FA_structure
!Step 0.001    !slow
$B ~ mu test  !r  test.rep test.female -test.male
and(test.male) test.cross
10 1 3
3777  0 IDEN !s2=150.10
3336  0 IDEN !s2=197.44
3533  0 IDEN !s2=185.72
4144  0 IDEN !s2=155.68
4109  0 IDEN !s2=156.59
4166  0 IDEN !s2=149.74
2637  0 IDEN !s2=208.38
3910  0 IDEN !s2=200.38
3491  0 IDEN !s2=181.57
3370  0 IDEN !s2=182.16

test.rep 2
test 0 DIAG 28.1 16.0 11.1 54.8 35.4 12.6 11.0 9.26 27.5 33.5
rep 0 IDEN

test.female 2
test 0 FA1 !GP
10*0.9
6.7 6.8 14.7 6.0 10.3 11.0 1.6 1.8 3.7 2.2
female 0 AINV
```

Appendix 5. Continued

```
test.cross 2
test 0 FA1 !GP
10*0.9
14.7 14.9 30.7 10.7 14.3 24.3 10.4 21.5 19.3 16.4
cross 0 IDEN


!PART 5
! FA_structure
!Step 0.001    !slow
$B ~ mu test  !r  test.rep test.female -test.male
and(test.male) test.cross
10 1 3
3777  0 IDEN !s2=150.10
3336  0 IDEN !s2=197.44
3533  0 IDEN !s2=185.72
4144  0 IDEN !s2=155.68
4109  0 IDEN !s2=156.59
4166  0 IDEN !s2=149.74
2637  0 IDEN !s2=208.38
3910  0 IDEN !s2=200.38
3491  0 IDEN !s2=181.57
3370  0 IDEN !s2=182.16

test.rep 2
test 0 DIAG 28.1 16.0 11.1 54.8 35.4 12.6 11.0 9.26 27.5 33.5
rep 0 IDEN

test.female 2
test 0 FA2 !GP
0.97 0.72 0.97 0.97 0.99 0.95 0.21 0.99 0.90 0.99
0.00 -0.38 -0.13 -0.18 1.05E-08 0.23 0.80 0.11 8.27E-02 1.05E-
08
10.11 8.17 11.33 11.62 13.47 10.70 0.22 5.43 6.76 0.61
female 0 AINV

test.cross 2
test 0 FA2 !GP
0.84 0.76 0.82 0.99 0.99 0.87 0.85 0.84 0.89 0.81
0.00 0.54 0.53 9.26E-02 0.20 0.38 0.22 0.32 0.39 0.55
17.54 18.08 34.57 12.41 16.10 28.00 12.90 25.52 22.62 21.59
```

Appendix 5. Continued

```
cross 0 IDEN


!PART 6
! XFA structure
!Step 0.001    !slow
$B ~ mu test  !r  test.rep  xfa(test,1).female -
xfa(test,1).male and(xfa(test,1).male) xfa(test,1).cross
10 1 3
3777  0 IDEN !s2=150.10
3336  0 IDEN !s2=197.44
3533  0 IDEN !s2=185.72
4144  0 IDEN !s2=155.68
4109  0 IDEN !s2=156.59
4166  0 IDEN !s2=149.74
2637  0 IDEN !s2=208.38
3910  0 IDEN !s2=200.38
3491  0 IDEN !s2=181.57
3370  0 IDEN !s2=182.16

test.rep 2
test 0 DIAG 28.1 16.0 11.1 54.8 35.4 12.6 11.0 9.26 27.5 33.5
rep 0 IDEN

xfa(test,1).female 2
xfa(test,1) 0 XFA1 !GP
0.49 3.83 0.62 0.94 0.00 0.60 0.23 0 1.16 0.00
3.19 2.08 3.42 3.36 3.92 3.54 0.23 2.41 2.53 0.91
female 0 AINV

xfa(test,1).cross 2
xfa(test,1) 0 XFA1 !GP
0.39 2.87 2.00 0.12 0.38 2.31 3.06 4.59 1.10 1.16
4.31 3.86 5.67 3.44 3.96 5.19 3.24 4.69 4.72 4.57
cross 0 IDEN


!PART 7
! XFA
!Step 0.001    !slow
$B ~ mu test  !r  test.rep  xfa(test,1).female -
xfa(test,1).male and(xfa(test,1).male) xfa(test,2).cross
```

Appendix 5. Continued

```
10 1 3
3777  0 IDEN !s2=150.1
3336  0 IDEN !s2=197.44
3533  0 IDEN !s2=185.72
4144  0 IDEN !s2=155.68
4109  0 IDEN !s2=156.59
4166  0 IDEN !s2=149.74
2637  0 IDEN !s2=208.38
3910  0 IDEN !s2=200.38
3491  0 IDEN !s2=181.57
3370  0 IDEN !s2=182.16

test.rep 2
test 0 DIAG 28.1 16.0 11.1 54.8 35.4 12.6 11.0 9.26 27.5 33.5
rep 0 IDEN

xfa(test,1).female 2
xfa(test,1) 0 XFA1 !GP
0.49 3.83 0.62 0.94 0.00 0.60 0.23 0.00 1.16 0.00
3.19 2.08 3.42 3.36 3.92 3.54 0.23 2.41 2.53 0.91
female 0 AINV

xfa(test,2).cross 2
xfa(test,2) 0 XFA2 !GP
0.42 1.88 0.00 0.00 0.00 2.43 2.66 3.25 1.14 0.83
4.32 4.86 6.81 3.45 3.59 5.13 2.73 3.91 4.59 4.86
0.00 1.15 1.54 2.31 3.05 2.76 2.47 3.17 2.60 1.67
cross 0 IDEN


!PART 8
! XFA
!Step 0.001    !slow
$B ~ mu test  !r  test.rep  xfa(test,2).female -
xfa(test,2).male and(xfa(test,2).male) xfa(test,1).cross
10 1 3
3777  0 IDEN !s2=150.10
3336  0 IDEN !s2=197.44
3533  0 IDEN !s2=185.72
4144  0 IDEN !s2=155.68
4109  0 IDEN !s2=156.59
4166  0 IDEN !s2=149.74
```

Appendix 5. Continued

```
2637  0 IDEN !s2=208.38
3910  0 IDEN !s2=200.38
3491  0 IDEN !s2=181.57
3370  0 IDEN !s2=182.16

test.rep 2
test 0 DIAG 28.1 16.0 11.1 54.8 35.4 12.6 11.0 9.26 27.5 33.5
rep 0 IDEN

xfa(test,2).female 2
xfa(test,2) 0 XFA2 !GP
0.40 2.86 0.37 0.00 0.00 0.34 0.20 0.00 1.00 0.00
3.08 1.23 2.83 1.77 2.81 3.38 1.39 3.29 3.10 2.18
0.00 -1.48 -2.04 -2.68 -2.59 -1.87 7.54E-01 -1.24 -1.42 -0.30
female 0 AINV

xfa(test,1).cross 2
xfa(test,1) 0 XFA1 !GP
0.39 2.87 2.00 0.12 0.38 2.31 3.06 4.59 1.10 1.16
4.31 3.86 5.67 3.44 3.96 5.19 3.24 4.69 4.72 4.57
cross 0 IDEN


!PART 9
! XFA
!Step 0.001    !slow
$B ~ mu test  !r  test.rep  xfa(test,2).female -
xfa(test,2).male and(xfa(test,2).male) xfa(test,2).cross
10 1 3
3777  0 IDEN !s2=150.1
3336  0 IDEN !s2=197.44
3533  0 IDEN !s2=185.72
4144  0 IDEN !s2=155.68
4109  0 IDEN !s2=156.59
4166  0 IDEN !s2=149.74
2637  0 IDEN !s2=208.38
3910  0 IDEN !s2=200.38
3491  0 IDEN !s2=181.57
3370  0 IDEN !s2=182.16

test.rep 2
test 0 DIAG 28.1 16.0 11.1 54.8 35.4 12.6 11.0 9.26 27.5 33.5
```

Appendix 5. Continued

```
rep 0 IDEN

xfa(test,2).female 2
xfa(test,2) 0 XFA2 !GP
0.40 2.86 0.37 0.00 0.00 0.34 0.20 0.00 1.00 0.00
3.08 1.23 2.83 1.77 2.81 3.38 1.39 3.29 3.10 2.18
0.00 -1.48 -2.04 -2.68 -2.59 -1.87 7.54E-01 -1.24 -1.42 -0.30
female 0 AINV

xfa(test,2).cross 2
xfa(test,2) 0 XFA2 !GP
0.42 1.88 0.00 0.00 0.00 2.43 2.66 3.25 1.14 0.83
4.32 4.86 6.81 3.45 3.59 5.13 2.73 3.91 4.59 4.86
0.00 1.15 1.54 2.31 3.05 2.76 2.47 3.17 2.60 1.67
cross 0 IDEN


!PART 10
! US
!Step 0.001    !slow
$B ~ mu test  !r  test.rep test.female -test.male
and(test.male) test.cross
10 1 3
3777  0 IDEN !s2=150.1
3336  0 IDEN !s2=197.44
3533  0 IDEN !s2=185.72
4144  0 IDEN !s2=155.68
4109  0 IDEN !s2=156.59
4166  0 IDEN !s2=149.74
2637  0 IDEN !s2=208.38
3910  0 IDEN !s2=200.38
3491  0 IDEN !s2=181.57
3370  0 IDEN !s2=182.16

test.rep 2
test 0 DIAG 28.1 16.0 11.1 54.8 35.4 12.6 11.0 9.26 27.5 33.5
rep 0 IDEN
```

Appendix 5. Continued

```
test.female 2
test 0 US !+55 !GP
10.86
05.67 7.90
11.43 8.16 14.22
09.20 6.75 10.96 10.26
11.37 7.10 12.70 11.13 13.03
12.59 7.86 14.09 11.22 13.42 15.74
03.87 2.40 04.37 02.70 03.57 05.35 2.78
10.43 5.64 11.28 08.44 10.84 12.53 4.38 10.65
10.14 6.07 11.75 08.40 10.68 11.73 3.86 10.20 10.95
06.62 4.34 08.08 05.17 06.62 08.61 3.64 07.21 07.09 5.77
female 0 AINV

test.cross 2
test 0 US !+55 !GP
16.05
14.99 17.16
21.00 20.77 29.88
13.23 12.24 17.43 11.72
14.82 13.42 18.99 13.38 15.48
18.38 17.75 25.29 16.03 17.99 24.83
11.13 11.17 13.88 9.758 11.30 13.71 12.03
16.26 14.17 20.57 14.05 16.15 20.61 12.54 21.32
17.10 15.23 22.76 14.79 16.71 20.57 13.30 18.38 20.46
16.15 15.55 21.65 13.17 14.59 18.98 11.59 16.60 17.28 17.31
cross 0 IDEN
```

Appendix 6. The following SAS code was used for plotting BLUP values of parents and families against sites to understand GxE interaction, rank changes, obtained from Model 8 (XFA2 model and XFA1 model of the GCA and SCA genetic effects, respectively)

```
data FA.predict;
length test $ 8.  genotype  $ 8. BLUP 8. SE 8. ;
infile "&folder\parent_BLUPc.csv" dsd missover firstobs=2
lrecl=55000;
input test genotype BLUP SE  ;
run;

data pred ; set FA.predict;
if test=1  then loading=2.18;
if test=2  then loading=0.58;
if test=3  then loading=1.86;
if test=4  then loading=0.60;
if test=5  then loading=1.59;
if test=6  then loading=2.39;
if test=7  then loading=1.26;
if test=8  then loading=2.49;
if test=9  then loading=2.24;
if test=10 then loading=1.83;
run;

/* SAS general options */
options helpbrowser=sas;

/* General graphics options */
GOPTIONS reset=global htext=1.6 vsize=10 hsize=10;

AXIS1 LABEL = (ANGLE=90 ) MINOR=NONE ;
AXIS2 LABEL = ('Environmental Loadings') MINOR=NONE
      order=(2.18 0.58 1.86 0.60 1.59 2.39 1.26 2.49 2.24
1.83);
symbol1 v=dot i=join r=160;
run;

proc gplot data=pred;
where loading in(0.58, 1.59, 2.49);
plot BLUP*loading=genotype /nolegend  ;
run;
```

Appendix 6. Continued

```
proc gplot data=pred;
where loading in(0.58, 1.26, 2.49);
plot BLUP*loading=genotype /nolegend  ;
run;

proc gplot data=pred;
where loading in(2.18, 0.58, 1.86, 0.60, 1.59, 2.39, 1.26,
2.49, 2.24, 1.83);
plot BLUP*loading=genotype /nolegend  ;
run;

data FA.predict;
length test $ 8.  genotype $ 8. BLUP 8. SE 8. ;
infile "&folder\cross_BLUP.csv" dsd missover firstobs=2
lrecl=55000;
input test genotype BLUP SE  ;
run;

data pred ; set FA.predict;
if test=1  then loading=3.97;
if test=2  then loading=3.79;
if test=3  then loading=5.46;

if test=4  then loading=3.46;
if test=5  then loading=3.86;
if test=6  then loading=4.88;
if test=7  then loading=3.03;
if test=8  then loading=4.19;
if test=9  then loading=4.45;
if test=10 then loading=4.26;
run;

/* SAS general options */
options helpbrowser=sas;
```

Appendix 6. Continued

```
/* General graphics options */

GOPTIONS reset=global htext=1.6 vsize=10 hsize=10;

AXIS1 LABEL = (ANGLE=90)   MINOR=NONE ;
AXIS2 LABEL = ('Environmental Loadings')   MINOR=NONE
      order=(3.97 3.79 5.46 3.46 3.86 4.88 3.03 4.19 4.45
4.26);
symbol1 v=dot  i=join r=160;
run;

proc gplot data=pred;
where loading in(3.03, 4.19, 5.46 );
plot BLUP*loading=genotype /nolegend  ;
run;

proc gplot data=pred;
where loading in(3.03, 5.46);
plot BLUP*loading=genotype /nolegend  ;
run;

proc gplot data=pred;
where loading in(3.97, 3.79, 5.46, 3.46, 3.86, 4.88, 3.03,
4.19, 4.45, 4.26 );
plot BLUP*loading=genotype /nolegend  ;
run;
```

Appendix 7. The following R code was used for Biplot of the site loadings and genotype scores from fitted XFA2 model (Model 8) to the GCA genetic effects.

```
# Read specific variances and loadings
XFAm <- matrix(scan('/additive.txt'),10,3)
ss <- svd(XFAm[,-1])
# rotation of site loadings with singular value decomposition
Lam <- XFAm[,-1] %*% ss$v
colnames(Lam) <- c("Load1","Load2")
# GCA variances for each site
Gvar <- Lam %*% t(Lam) + diag(XFAm[,1])
# loadings on the correlation scale
cLam <- diag(1/sqrt(diag(Gvar))) %*% Lam
scores <- XFAsln[,11:12] %*% ss$v
dimnames(scores)<-list(paste('p',1:22,sep=''),
paste('XFA',1:2,sep=''))
biplot(scores[,1:2],Lam[,1:2])
```

Appendix 7. Parent IDs used in the figures and their corresponding TIP (Tree Improvement Program)IDs

Parent IDs	TIP IDs
P1	TI1164622
P2	TI138245
P6	TI150187
P7	TI137174
P8	TI2387750
P9	TI2619739
P11	TI143184
P12	TI1123708
P13	TI216059
P14	TI2523434
P16	TI2751739
P17	TI1212177
P18	TI1306258
P19	TI1940090
P20	TI11065116
P21	TI1121708
P22	TI1532234

CHAPTER 2

Effects of Patterns of Missing Data and Mating Design on the Genotype Imputation Accuracy

Abstract

We used empirical and simulated data sets to investigate the effects of level of missing data, pattern of missing data and mating design on the accuracy of imputation when using BLUP (Gengler et al. 2007) and haplotype Hidden Markov Model (HMM) (BEAGLE software, Browning and Browning et al. 2009) approaches for imputation of missing genotypes. Due to lack of knowledge of which SNPs were in linkage disequilibrium (LD) with any particular SNPs that had missing value, only the BLUP approach (given pedigree and molecular data) was employed for empirical data. Accuracy of imputation was higher for the structured pattern of missing data at any level of missing data. Regardless of the pattern of the missing data, imputation accuracy was less than 0.70 when the data had greater than 40% missing values. Imputation was carried out with BLUP and HMM approaches for the simulated data set. Imputation accuracy was greater with HMM approach than BLUP approach regardless of the mating design and level and pattern of missing data. Mating design did not have any effect on the imputation accuracy for the BLUP approach when the pattern of the missing data was structured. Also, including parental genotypes in the analysis had no effect on the imputation accuracies. For the HMM approach, mating design effect became clear on the imputation accuracy when missing data had an 80% for random missing pattern. Among three patterns of missing data, imputation accuracy was the highest (0.96) for the evenly-spaced pattern at 80% missing data. This study demonstrated that the BLUP method is fairly accurate and fast for imputation of missing genotype with known pedigree information and balanced data set for current tree breeding programs due to lack of the accurate maps of marker order for loblolly pine (*Pinus taeda* L.).

Introduction

High-throughput SNP genotyping has introduced new analytical approaches to forest tree breeding programs for genetic evaluations. Costs of high-throughput genotyping have decreased to the point where it is possible to use molecular genetic markers in applied breeding programs. However, genotyping is still not reasonable for all populations due to high cost and logistic limitations (Legarra et al. 2009). For a genotype data with missing genotypes, imputation is implemented to predict unknown genotypes (Druet and Georges, 2010, Meuwissen and Goddard, 2010, Daetwyler et al., 2011). Genotype imputation is a statistical method for predicting unobserved genotypes (Browning 2008, Li et al. 2009; Marchini and Howie 2010, Marchini 2011, Howie et al. 2011) to increase the power of genome-wide association studies (Browning and Browning 2009, Howie et al. 2011). Genotype imputation is described as the procedure of predicting genotypes that are not directly assayed (Marchini and Howie 2010, Marchini 2011) in a subset of individuals.

Several factors such as population structure, linkage disequilibrium, sampling design and laboratory methods have effects on the patterns of missing data and the ease of imputation (Sun and Kardia 2008).

In a genotype data set, data can be missing due to genotyping assay failures at some loci in some samples (**Figure 1a**), in which there are no clear patterns (Lin et al. 2008, Habier et al. 2009). Many of the imputation programs allow imputing missing genotypes due to assay failure which is usually very low (Marchini 2011). Genotype call accuracy can be greater than 0.999 for SNPs in genome-wide association studies (Browning and Yu 2009).

Missing genotype data can be due to by design (not directly assayed) as a result of lowering the cost. The pattern can be structured such a way that some individuals do not have any genotype for all the loci (**Figure 1b**). The goal is to genotype a subset of population for high density markers and genotype a larger subset of population for low density markers (Lin et al. 2008, Habier et al. 2009).

There are two major imputation methods to infer missing genotypes. The first is using linkage (family based) and the second is using LD (population based) information. Imputation methods depending on known relationships between individuals exploit linkage, capturing the identical by descent (IBD) information in the pedigree structure. Imputation methods depending on unknown relationships in population use linkage disequilibrium among flanking markers (Johnston et al. 2011). There are also other methods which use both linkage and LD information.

In the presence of a reference panel, unobserved genotypes in the study sample are predicted based on the haplotype patterns in the reference panel (Howie et al. 2011). The accuracy of imputation is using a reference panel is affected by similarity of haplotypes in the study sample and reference panel (Huang et al. 2009). The allele frequencies and levels of linkage disequilibrium (LD) in the study sample (Huang et al. 2009). Marchini and Howie (2010) investigated the effect of reference panel, study sample and genotyping chips on the imputation accuracy. They concluded that minor allele frequency, reference panel size, genotyping chip design have effect on the imputation accuracy. Also, genetic differences between study sample and reference panel affects the accuracy of imputation (Marchini 2011).

Imputation can be carried out using only the information contained within the dataset of interest, or in the presence of a reference panel of haplotypes using it. Such a reference panel and genetic map is available for few if any forest tree species (Neale and Kremer 2011), although that situation is likely to change within the next few years.

The extent of linkage disequilibrium is very important in genome-wide association studies. Due to LD among nearby markers, non-genotyped SNP markers might be correlated with one or more nearby genotyped markers (Servin and Stephens 2007). Imputed genotypes are given probabilistic instead of discrete and using probabilistic allele counts without transforming to a discrete value is suggested to prevent loss of information particularly when the allele is less common (Li et al. 2009).

Many methods and software packages used for genetic evaluations require complete genotype data. In a genotype data set, some individuals might have some missing genetic markers, even if the missing rate per marker is low (Yu and Schaid et al. 2007). This compromises association analyses, with varying numbers of subjects contributing to analyses when performing single-marker or multi-marker analyses (Yu and Schaid et al. 2007). Dropping the individual or locus records in which these missing values occur is an option, but it is not desired because it is not cost-effective to drop individuals that are missing data at any one of the loci being analyzed. Also, it causes loss of valuable information and decreases power. In the extreme case in which all marker loci in the genome are analyzed at once, the approach of excluding individuals with missing genotype data can mean discarding hundreds or thousands of correct data values in order to remove a few missing values. Therefore

imputation of missing genotypes is very important to avoid losing information and decreasing the power of a study (Sun and Kardia 2008).

The objective of this study was to test the effects of

1. levels and patterns (sporadic and structured) of missing genotypes
2. mating design (double-pair and half-diallel)
3. statistical approach (mixed model BLUP and BEAGLE parent-offspring trio)

on the accuracy of imputation using empirical data from a loblolly pine breeding population as well as simulated data.

Materials and Methods

Data

In this study, actual and simulated were used to investigate the accuracy of imputation. The actual data used for the analysis had 178 cloned individuals genotyped at 3461 biallelic SNP markers. Complete pedigree was available for the actual data. A total of 13 families, that were mostly single-pair mating design were obtained using 15 parents as female and male (**Figure 1**). Number of cloned individuals per crosses ranged from 1 to 39.

The simulated data set was designed to mimic a tree breeding population for two different mating designs, double-pair mating design and half-diallel mating design. Simulated data were generated using an R script. For genome, 12 chromosome 1500 cM long was simulated. In total, 3500 biallelic SNP markers were distributed at random across the chromosome with equal allele frequency. 10 QTL loci out of 150 QTL were randomly chosen to be SNP QTL. For the half diallel mating design, with 21 parents, 210 full-sib families were generated to produce 14 progeny per family. For the double-pair mating design, 60 parents were crossed to obtain 120 families of 24 progeny. For the simulation scenario all parents and offspring genotype map positions were known.

Imputation with BLUP approach using pedigree

In a pedigreed population, missing genotype of individuals can be predicted by imputation (Ellinghaus et al. 2009). Gengler et al. (2007) described a method to impute missing genotypes of an individual using numerator relationship matrix (**A**). Pedigree information and all available genotype values for a given locus is an alternative approach to impute

missing genotypes when LD information among SNPs is not a choice (Gengler et al. 2007). The method described by Gengler et al. (2007) predicts conditional expectations of gene contents for non-genotyped individuals based on available genotypic information of relatives. Linear model methods for estimating genomic relationships and genetic values from SNP genotype data utilize numerically coded SNP genotypes (**Figure 2b**) rather than the nucleotide alleles provided by most genotyping services (**Figure 2a**). The number of copies of a particular allele (say B2) in a genotype (N_2=0, 1, 2 for diploids) is referred to as the gene content (Lynch and Walsh 1998). Gene content can be used to study additive gene action of a candidate gene. Under the assumption of no additive interaction of an allele with all other alleles, the relationship between the gene content and the genotypic value will be nonlinear (Lynch and Walsh, 1998).

As an imputation method, we followed Gengler et al. (2007) mixed model BLUP approach. In the first step, marker genotypes were converted to gene contents based on minor allele frequencies. Gene content was treated as the dependent variable in a mixed model approach. This method relies on of genetic covariance among relatives, namely the covariance between gene contents is linear and proportional to the additive relationship between individuals (Gengler et al. 2007). Genetic covariance occurs because two related individuals have alleles that are identical by descent (IBD).

In this method, a single biallelic locus with alleles A and a was considered. If we assume the q (0, 1 or 2) is the number of copies of allele (gene content) for a genotyped individual the expected value of q_p for a progeny of genotyped parents is $E(q_p) = 0.5(q_s+q_d)$, where q_p, q_s,

and q_d are the contents of allele A for progeny, sire and dam, respectively. The gene content (q) of genotyped individuals can be replaced by its deviation (d) from the population mean (μ), d = q-μ. Then the expected gene content of a progeny of genotyped parents is $E(q_p) = \mu + (d_s + d_d)/2$, where $q_s=d_s + \mu$ and $q_d=d_d + \mu$ (Gengler et al. 2007).

If a mother has not been genotyped, her expected value is replaced by population mean, $q_d = \mu$. And if a genotyped relative is not a parent or offspring, the expected values of q can be obtained as: if maternal grand-sire (mgs) has been genotyped, the expected value of q for mother is $E(q_d|d_{mgs}) = \mu + 0.5d_{mgs}$ because the mother receives half of its genes from its parents, and the expectation for progeny is $E(q_p|d_s,d_{mgs}) = \mu + 0.5d_s + 0.25d_{mgs}$.

The following linear mixed model was used to predict allele gene contents for empirical and simulated genotype data

$$\mathbf{[1]} \qquad \mathbf{y} = \mathbf{Xb} + \mathbf{Mu} + \mathbf{e}$$

where **y** is the vector of gene content with nx1, **b** is intercept with 1xn dimension, and **u** is individual tree effect with txt. The **X** is the design matrix (vector of 1s) for the mean, **M** is the design matrix connecting trees to the gene content vector **y**, and **e** is the error variance. The solutions of mixed model equations produce predicted allele gene contents for trees with missing genotypes. The solutions would be continuous in nature, centered on 1 since the gene content number could be 0, 1 or 2. In matrix format the same model can be presented as

$$\mathbf{[2]} \qquad \begin{bmatrix} \mathbf{1'1} & \mathbf{1'M} \\ \mathbf{M'1} & \mathbf{M'M} + \mathbf{A^{-1}}\lambda \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\mu}} \\ \hat{\mathbf{d}}_{\mathbf{y}} \\ \hat{\mathbf{d}}_{\mathbf{x}} \end{bmatrix} = \begin{bmatrix} \mathbf{1'q_y} \\ \mathbf{M'q_y} \end{bmatrix}$$

where $\mathbf{d_y}$ is a vector of gene content deviations for trees with genotype records, $\mathbf{d_x}$ is a vector of gene content deviations for trees with no genotype records , **M** is incidence matrix linking $\mathbf{q_y}$ to $\mathbf{d_y}$ and $\mathbf{d_x}$ and **A** is the additive relationship matrix. **A** matrix can be separated into sub-matrices as $\mathbf{A} = \begin{bmatrix} \mathbf{A_{yy}} & \mathbf{A_{yx}} \\ \mathbf{A_{xy}} & \mathbf{A_{xx}} \end{bmatrix}$, where $\mathbf{A_{yy}}$ is the sub-matrix of covariances among trees with known genotypes, $\mathbf{A_{xx}}$ is the sub-matrix of covariances among trees with unknown genotypes and $\mathbf{A_{xy}}$ and $\mathbf{A_{yx}}$ are the covariances between trees with known and unknown genotypes. And λ is the shrinkage factor, a ratio of error and additive genetic variance ($\lambda = \sigma_e^2/\sigma_d^2$). We were not interested in estimation of variances but a small error variance is needed to solve the mixed model equations. The authors suggest an error variance of 0.001. Setting λ equal to 0.01 or less value allows accounting for possible genotyping and pedigree errors (Gengler et al. 2007). The analysis was carried out in ASReml software (Gilmour et al. 2009).

Imputation with BEAGLE (parent-offspring trios)

For simulated data, we used a haplotype hidden Markov model conditional on observed genotype data to infer haplotype phase and impute non-genotyped markers in unrelated individuals and parent-offspring trio data (Browning and Browning 2009). The hidden Markov model (HMM) employed by BEAGLE doesn't model recombination and mutation (Browning and Browning 2011). The model clusters similar haplotypes at each locus and defines a HMM to find the most likely haplotype pairs conditional on observed genotype data. Missing genotypes are imputed based on these haplotypes. The number of clusters varies depend on the sample size and linkage disequilibrium (LD) (Browning and Browning 2011). Imputation was performed with BEAGLE 3.0 package (Browning and Browning

2007) with default settings. BEAGLE is computationally efficient and relies on genotypes nearby markers when imputing a missing genotype (Li et al. 2009). The trio option in BEAGLE was used, since our simulated data structure was appropriate (known parents and progeny). The HMM algorithm requires the order of markers on the chromosomes or map position of all markers. We did not have genetic map of loblolly pine and thus we did not analyze empirical data with BEAGLE.

Comparison of Models

Allele content values imputed by the BLUP method are continuous rather than discrete, so accuracy of imputation was assessed by correlation (**r**) of the true allele contents with the imputed values. BEAGLE was not tested on the empirical data because it requires the map position of all SNPs, and not all SNPs in the empirical dataset have been mapped. Genotype values imputed by BEAGLE are discrete, so accuracy of imputation was assessed as percent (**%**) correctly imputed values relative to the true values. Results for the simulated data were given as the average of three replicates.

Results

Results from BLUP approach

Empirical data

Accuracy of the imputation (r) between imputed allele content values and true values was higher for the structured missing pattern (ranged from 0.96 to 0.23) of missing data than for the random pattern (ranged from 0.95 to 0.12) for any level of missing genotypes (**Figure 4**). Imputation accuracies decreased as the percent level of missing genotypes increased regardless of the pattern of the missing data. The dramatic change in imputation accuracy between structured and random pattern of missing genotypes in the actual data was observed when missing data had more than 40% missing values.

Simulated data

For simulated data set, we observed that mating design (half-diallel and double-pair) had almost no effect on the accuracy of imputation (**Figure 5**). Imputation accuracies (r) between imputed allele content values and true values were virtually the same for the mating designs, regardless of the level of the missing data for the structured pattern. Imputation accuracy was 0.95 at 20% missing data, but dropped to 0.75 at 80% missing data.

For the BLUP approach, we also evaluated the effect of having parental genotypes in the data on the imputation accuracy for two patterns (random and structured) of missing data when mating design was half-diallel (**Table 1**). Having parental genotypes had nearly no effect on the imputation accuracy regardless of the pattern of missing data. Also random and structured

pattern of missing data gave the same imputation accuracies when mating was the half-diallel.

Results from HMM approach

Imputation accuracy (% correctly imputed) was greater than 0.98 for datasets with up to 60% random missing values for both mating designs (**Figure 6**). However when genotype data had 80% missing data, the mating design showed a noticeable effect on the imputation accuracy. Imputation was 0.82 and 0.91 for double-pair and half-diallel mating designs, respectively, at 80% missing data.

Imputation accuracy was not affected by mating design (half-diallel and double-pair) when the pattern of the missing data was evenly-spaced at 80% missing data (**Figure 7**). However, when pattern of the missing data was random, the accuracy of imputation was affected the mating design. For random pattern missing data, imputation accuracy was 0.91 and 0.82 for half-diallel and double-pair mating designs, respectively, at 80% missing genotype data. The effect of random, structured and evenly spaced patterns of missing genotypes on the accuracy of imputation was evaluated for half-diallel mating design at 10% and 80% missing genotype data (**Figure 8**). Among the patterns of missing data, structured pattern of missing data had the lowest imputation accuracy. Imputation accuracy was not noticeably different at 10% missing. However, we observed a dramatic decrease on the imputation accuracy for the structured pattern at 80% missing marker dataset. Among the missing data patterns, the evenly-spaced pattern gave the highest imputation accuracy when 80% of the markers were

missing in the progeny genotype data. Imputation accuracies were 0.73, 0.91 and 0.96 for structured, random and evenly spaced patterns, respectively at 80% missing data.

Computational demands:

The BLUP approach was reasonably fast in ASReml software using the CYCLE qualifier. One limiting factor for the CYCLE qualifier was that not all marker data were run simultaneously to impute missing genotypes. The best way to run the program was splitting marker data in several sub files. With this method the time required for imputation did not change with the level of the missing data as suggested by Gengler et al. (2007). BEAGLE software imputed missing genotypes on chromosomes individually and each chromosome was imputed separately. In this study we used the BEAGLE trio option. We observed that imputation time increases as the level of missing data and effective population size (21 parents vs. 60 parents) increased.

Discussion

In application of markers for predictions of genetic merit missing genotypes is a limiting factor in terms of power and methodology for the estimation of genomic estimated breeding values. Software used for predictions require complete genotype data. Missing genotype could be due to having a subset of the population not being genotyped or it could be due to random missing genotype information of some individuals at some loci. Most software available for imputation requires genetic map of the order of marker loci. Since a loblolly pine genetic map has not been completed, we explored methods that can use information from the data set itself, such as using pedigree to estimate missing genotypes.

We used the BLUP method described by Gengler et al. (2007) to investigate the effect of pattern and level of missing data on the accuracy of imputation given the pedigree and genotype data. The proposed method predicts gene contents for bi-allelic markers, such as SNPs using a mixed model given pedigree and genotype data (BLUP method). For empirical data, we found that imputation of missing genotypes using the numerator relationship matrix and mixed model was relatively more powerful when the pattern of missing data was structured (all genotypes missing) than random pattern with the same levels of missing data. As the proportion of missing genotypes increased in the data, the power of imputation decreased. The empirical data used for imputation was highly unbalanced where the number of clones per family ranged from 1 to 39. The BLUP method was not suitable for imputation of random missing genotypes when family sizes were unbalanced. With simulated data set, it was observed that the imputation accuracy was less affected by the distribution of missing

genotypes in a balanced mating design with families of equal size. We also found that including parental genotype data in the analysis did not have any effect on the imputation accuracy for the data generated for half-diallel mating design when the patterns of missing data were random and structured. Similarly, Gengler et al. (2007) using simulated data set for real bovine pedigrees found that imputation accuracy increased as the proportion of missing data decreased. Imputation accuracies were 0.67 and 0.54 at 50% and 90% missing data when the pattern of missing data was structured. In another study, using cross validation data set on 961 sires, allele gene content imputation accuracy was 0.93 (Gengler et al. 2008). The accuracy of the gene content predictions depends on the pedigree structure (Gengler et al. 2008). To investigate the accuracy of genomic selection when using predicted genotypes to increase the reference population, Pszczola et al (2011) employed Gengler et al. (2007) method to predict 1000 additional non-genotyped individuals' gene contents. However, they found that adding individuals with imputed genotypes did not have effect on the accuracy of genomic estimated predictions.

In dairy cattle breeding, due to high linkage disequilibrium as a result of small effective population size, (3K) low-density SNP panel can be imputed to obtain (50K) high-density SNP panel with very high accuracy (low error rates) (Hayes et al. 2011). Calus et al. (2011) compared the effect of pattern of missing genotypes on the accuracy of imputation for low and high density marker data using information from surrounding markers as a correlated trait with a multivariate mixed model (BLUP method) using simulated data. This method was an extension of the method described by Gengler et al. (2007) including information from

flanking markers. In the same study, imputation accuracy of the proposed method was compared with BEAGLE and fastPhase for random and evenly-spaced pattern of missing data, where BEAGLE assumed animals were unrelated. They found that BEAGLE was superior for high density SNP data, however for low density marker data and when LD was low among flanking markers the proposed method using pedigree was superior.

Even though the BEAGLE approach requires a parent-offspring trio reference panel for imputation of structured pattern of missing data, we imputed structured missing genotypes assuming random missing pattern. Among three patterns of missing data, imputation accuracy was the lowest for structured missing data at 10% and 80% levels of missing data. BEAGLE was accurate enough to allow implementation of evenly spaced genotyping strategies for some mating designs at very low density genotyping. BEAGLE is a data dependent program which uses a local-haplotype cluster model and as a result BEAGLE use the actual LD in the data (Browning and Browning 2007). In the presence of pedigree, imputation accuracy can be increased by tracking the chromosome segments within families (Habier et al. 2009, Hayes et al. 2011).

The cost of genotyping is still not reasonable in forest tree breeding programs. Therefore, genotype imputation is powerful statistical tool to reduce the cost by genotyping a large subset of the population for a low-density SNP marker panel and relatively small subset of the population for high-density panel. The BLUP method was fast and reasonably accurate for imputation of genotypes missing due to random or structured pattern, depending on the level of missing data and data structure (unbalanced or balanced data). BLUP is currently the

choice of imputation method for current tree breeding programs. However, based on the simulation data results, when the accurate maps of marker order are available for the species of interest (loblolly pine), evenly-spaced genotyping in progeny and high-density genotyping in the parents, the trio option in BEAGLE is a good option in tree breeding programs.

Conclusion

For empirical and simulated data sets, imputation accuracy decreased as the level of missing data increased regardless of the method of imputation, level (percent missing) and pattern (random, structured) of missing data and mating design.

Accuracy of imputation from the BLUP method for empirical data set was higher for the structured pattern of missing data at any level of missing data.

For simulated data set, imputation accuracy was greater with the HMM (BEAGLE parent-offspring trio) than with the BLUP approach regardless of the mating design and level and pattern of missing data. Mating design (half-diallel and double-pair) did not have any effect on the imputation accuracy for the BLUP approach when pattern of the missing data was structured. For the HMM approach, the mating design effect became clear on the imputation accuracy when missing data was at 80% for random missing pattern. Among three pattern of missing data, imputation accuracy was the highest (0.96) for evenly-spaced pattern at 80% missing data. For the comparison of mating design effect on the imputation accuracy when the pattern of the missing data was evenly-spaced, we found that mating design had no effect on the imputation accuracy.

References

Browning, B.L., Browning, S.R., 2009. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. American Journal of Human Genetics 84, 210–23.

Browning, B.L., Yu, Z., 2009. Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. American Journal of Human Genetics 85, 847–61.

Browning, S.R., Browning, B.L., 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. American Journal of Human Genetics 81, 1084–97.

Browning, S.R., Browning, B.L., 2011. Haplotype phasing: existing methods and new developments. Nature reviews: Genetics 12, 703–14.

Calus, M.P.L., Veerkamp, R.F., Mulder, H.A, 2011. Imputation of missing single nucleotide polymorphism genotypes using a multivariate mixed model framework. Journal of Animal Science 89, 2042–9.

Daetwyler, H.D., Wiggans, G.R., Hayes, B.J., Woolliams, J. A, Goddard, M.E., 2011. Imputation of missing genotypes from sparse to high density using long-range phasing. Genetics 189, 317–27.

Druet, T., Georges, M., 2010. Hidden Markov model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fine mapping. Genetics 184, 789–798.

Ellinghaus, D., Schreiber, S., Franke, A., Nothnagel, M., 2009. Current software for genotype imputation. Human Genomics 3, 371–80.

Gengler, N., Abras, S., Verkenne, C., Vanderick, S., Szydlowski, M., Renaville, R., 2008. Accuracy of prediction of gene content in large animal populations and its use for candidate gene detection and genetic evaluation. Journal of dairy science 91, 1652–9.

Gengler, N., Mayeres, P., Szydlowski, M., 2007. A simple method to approximate gene content in large pedigree populations: application to the myostatin gene in dual-purpose Belgian Blue cattle. Animal : An International Journal of Animal Bioscience 1, 21–8.

Gilmour, A.R., B.J. Gogel, B.R. Cullis, R. Thompson. 2009. ASReml user guide release 3.0. VSN International Ltd., Hemel Hempstead, HP1 1ES, United Kingdom. 372 p.

Habier, D., Fernando, R.L., Dekkers, J.C.M., 2009. Genomic selection using low-density marker panels. Genetics 182, 343–53.

Hayes, B.J., Bowman, P.J., Daetwyler, H.D., Kijas, J.W., van der Werf, J.H.J., 2012. Accuracy of genotype imputation in sheep breeds. Animal Genetics 43, 72–80.

Howie, B., Marchini, J., Stephens, M., 2011. Genotype imputation with thousands of genomes. G3: Genes, Genomes, Genetics 1, 457–70.

Huang, L., Li, Y., Singleton, A.B., Hardy, J. a, Abecasis, G., Rosenberg, N. a, Scheet, P., 2009. Genotype-imputation accuracy across worldwide human populations. American Journal of Human Genetics 84, 235–50.

Johnston, J., Kistemaker, G., Sullivan, P.G., 2011. Comparison of Different Imputation Methods. Interbull Bulletin 44, 25-33.

Legarra, a, Aguilar, I., Misztal, I., 2009. A relationship matrix including full pedigree and genomic information. Journal of Dairy Science 92, 4656–63.

Li, Y., Willer, C., Sanna, S., Abecasis, G., 2009. Genotype imputation. Annual Review of Genomics and Human Genetics 10, 387–406.

Lin, D.Y., Hu, Y., Huang, B.E., 2008. Simple and Efficient Analysis of Disease Association with Missing Genotype Data. American Journal of Human Genetics 82, 444–452.

Lynch, M., Walsh, B. 1998. Genetics and analysis of quantitative traits. Sinauer Associates, Inc., Sunderland, MA, USA, 980 p.

Marchini, J., 2011. Genotype imputation. In E. Zeggini and A. Morris (Eds), Analysis of complex disease association studies: a practical guide (pp. 157-175). Elsevier Inc.

Marchini, J., Howie, B., 2010. Genotype imputation for genome-wide association studies. Nature reviews. Genetics 11, 499–511.

Meuwissen, T., Goddard, M., 2010. The use of family relationships and linkage disequilibrium to impute phase and missing genotypes in up to whole-genome sequence density genotypic data. Genetics 185, 1441–9.

Neale, D.B., Ingvarsson, P.K., 2008. Population, quantitative and comparative genomics of adaptation in forest trees. Current opinion in plant biology 11, 149–55.

Neale, D.B., Kremer, A., 2011. Forest tree genomics: growing resources and applications. Nature Reviews. Genetics 12, 111–22.

Pszczola, M., Mulder, H.A, Calus, M.P.L., 2011. Effect of enlarging the reference population by (un)genotyped animals on the accuracy of genomic selection in dairy cattle. Journal of Dairy Science 94:431–441.

Servin, B., Stephens, M., 2007. Imputation-based analysis of association studies: candidate regions and quantitative traits. PLoS Genetics 3, 7.

Sun, Y.V., Kardia, S.L., 2008. Imputing missing genotypic data of single-nucleotide polymorphisms using neural networks. European Journal of Human Genetics 16, 487–95.

Yu, Z., Schaid, D.J., 2007. Methods to impute missing genotypes for population data. Human Genetics 122, 495–504.

TABLES

Table 1. Effect of having parental genotypes on the imputation accuracy (r) using simulated data generated as half-diallel mating design when the sampling is random and structured.

	Structured Sampling					**Random Sampling**				
% missing	%10	%20	%40	%60	%80	%10	%20	%40	%60	%80
Parent+progeny	0.97	0.94	0.89	0.82	0.75	0.97	0.95	0.89	0.82	0.75
progeny	0.97	0.94	0.88	0.82	0.74	0.97	0.95	0.89	0.82	0.75

FIGURES

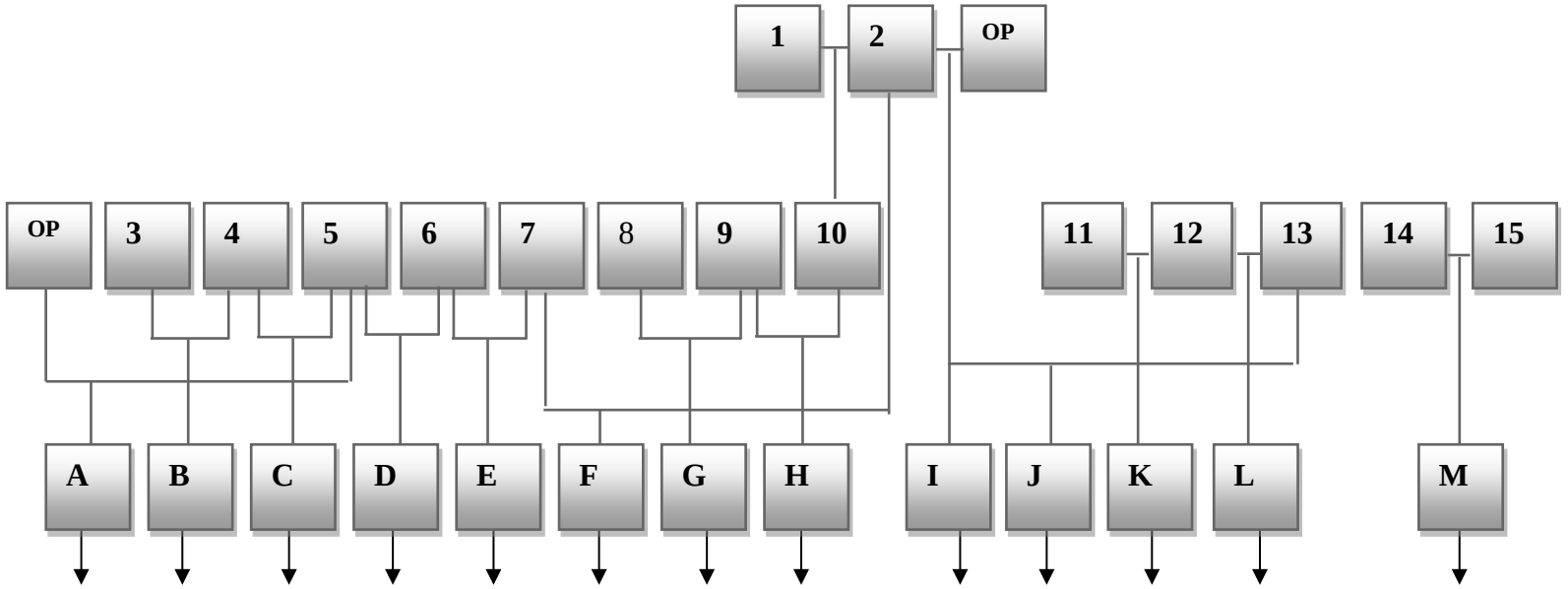


Figure 1. Pedigree structure of 178 cloned trees. Letters refer to nine family full-sib ID's. Squares with OP were unknown male contribution (wind-pollinated). Numbers refer to 13 half-sib ID's. Number of cloned trees per family ranged from 1 to 39.

a) Sporadic missing genotypes

clone	parent1	parent2	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10
26	1	12	NA	GG	NA	AG	AG	AC	AA	AA	AA	NA
27	1	12	AG	NA	GG	AA	AG	AC	AG	NA	AA	NA
28	1	12	AG	GG	GG	AA	NA	AC	AG	NA	AA	AA
29	1	12	GG	NA	AG	AG	GG	CC	NA	AA	AA	AA
30	1	12	GG	GG	NA	AG	AG	NA	AG	AA	AA	NA

b) Structured missing genotypes (non-genotyped)

clone	parent1	parent2	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10
26	1	12	GG	GG	GG	AG	AG	AC	AA	AA	AA	AA
27	1	12	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
28	1	12	AG	GG	GG	AA	AG	AC	AG	AA	AA	AA
29	1	12	GG	GG	AG	AG	GG	CC	AG	AA	AA	AA
30	1	12	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

*NA: missing genotypes.

Figure 2. Patterns of missing genotypes in given data due to assay failure: sporadic missing genotypes (a) or by design: structured missing genotypes (b)

a) Genotype data

clone	parent1	parent2	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10
26	1	12	GG	GG	GG	AG	AG	AC	AA	AA	AA	AA
27	1	12	AG	GG	GG	AA	AG	AC	AG	AA	AA	AA
28	1	12	AG	GG	GG	AA	AG	AC	AG	AA	AA	AA
29	1	12	GG	GG	AG	AG	GG	CC	AG	AA	AA	AA
30	1	12	GG	GG	GG	AG	AG	AC	AG	AA	AA	AG

b) Gene content of 10 loci using minor allele frequency for loci

Clone	parent1	parent2	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10
26	1	12	0	0	0	1	1	1	0	0	0	0
27	1	12	1	0	0	2	1	1	1	0	0	0
28	1	12	1	0	0	2	1	1	1	0	0	0
29	1	12	0	0	1	1	0	0	1	0	0	0
30	1	12	0	0	0	1	1	1	1	0	0	1

Figure 3. Schematic presentation of marker genotypes for 10 loci (M1 to M10) for each tree

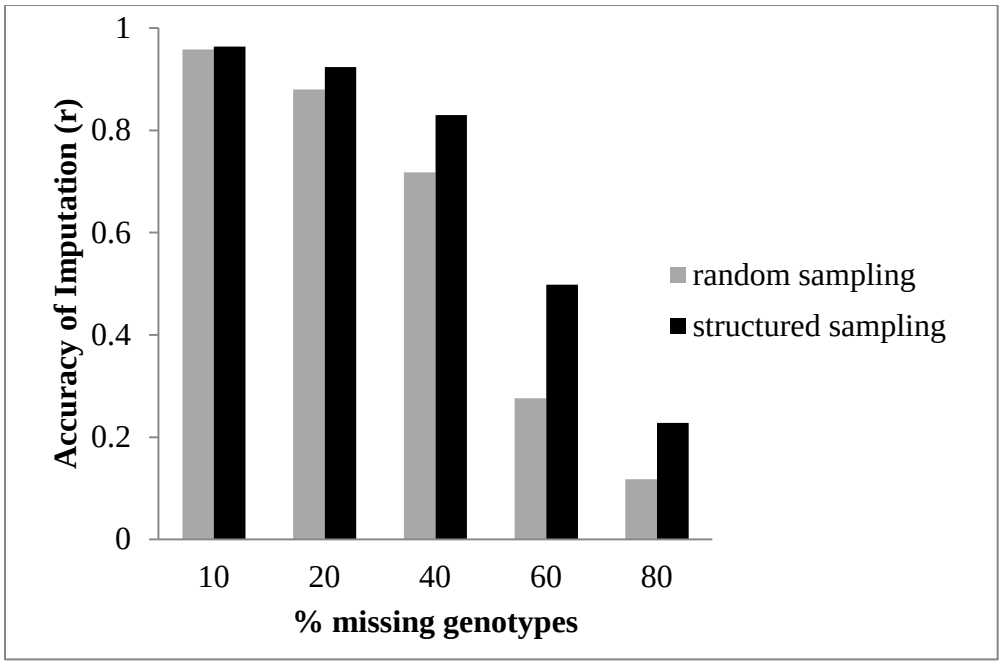


Figure 4. Accuracy of imputation (r) between imputed allele content values and true values for empirical data when sampling was sporadic or structured. Allele content values were imputed by the BLUP method using pedigree. Accuracy of imputation between imputed allele content values and true values are higher for the structured pattern of missing data than for the random pattern.

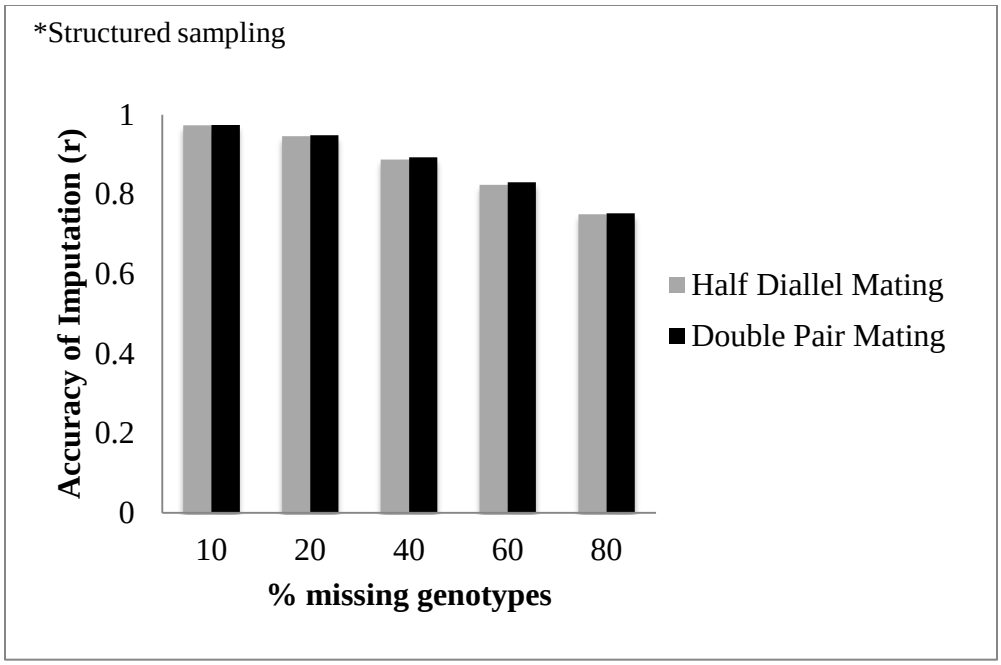


Figure 5. Accuracy of imputation (r) between imputed allele content values and true values for simulated data from half-diallel and double-pair mating designs when sampling was structured for different levels (missing genotypes ranged from 10% to 80%). Allele content values were imputed by the BLUP method using pedigree. Accuracy of imputations between imputed allele content values and true values are virtually equivalent for the two patterns of mating design and for different levels of missing genotypes.

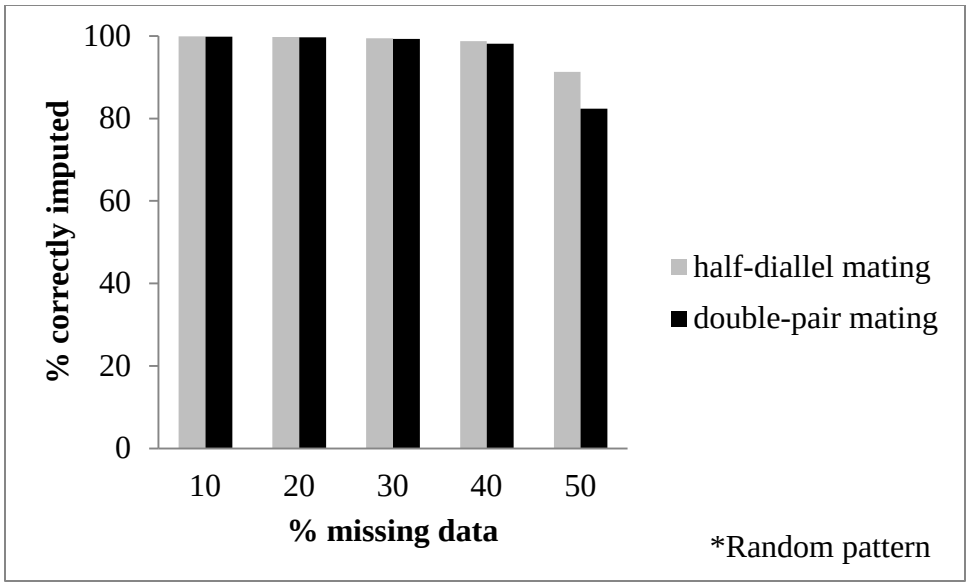


Figure 6. Allele content values were imputed by haplotype analysis of parent-offspring trios using BEAGLE software for simulated data from half-diallel and double-pair mating designs when sampling was random. Accuracy of imputation was assessed as percent correctly imputed values relative to the true values. Imputation accuracy is >0.98 for datasets with up to 60% random missing genotypes for both mating designs.

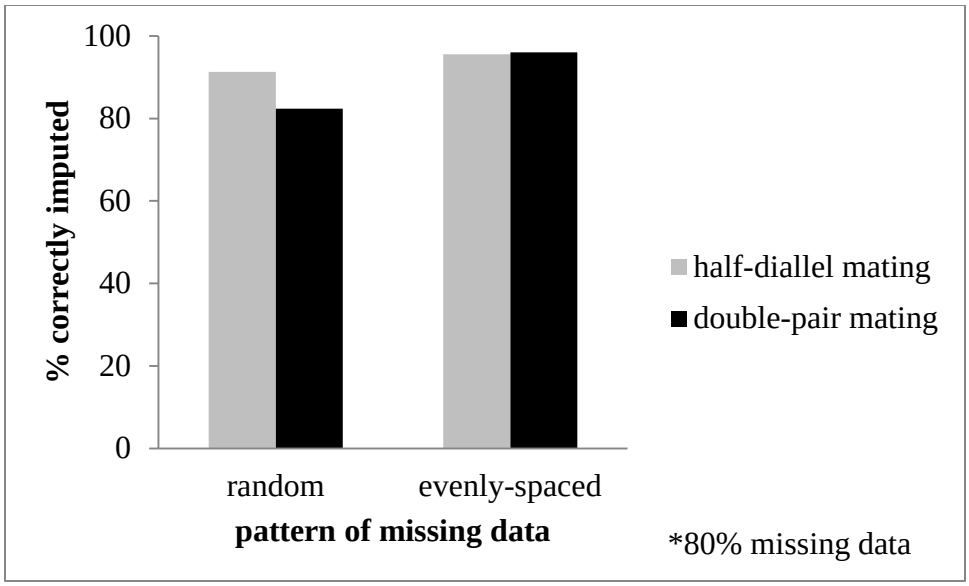


Figure 7. Effect of random and evenly-spaced patterns of missing genotypes on the accuracy of imputation for half-diallel and double-pair mating designs at 80% missing genotype data when imputation was carried out with BEAGLE software. For the evenly-spaced pattern of missing data, the imputation accuracy was ~0.96 and was not affected by mating design at 80% missing data.

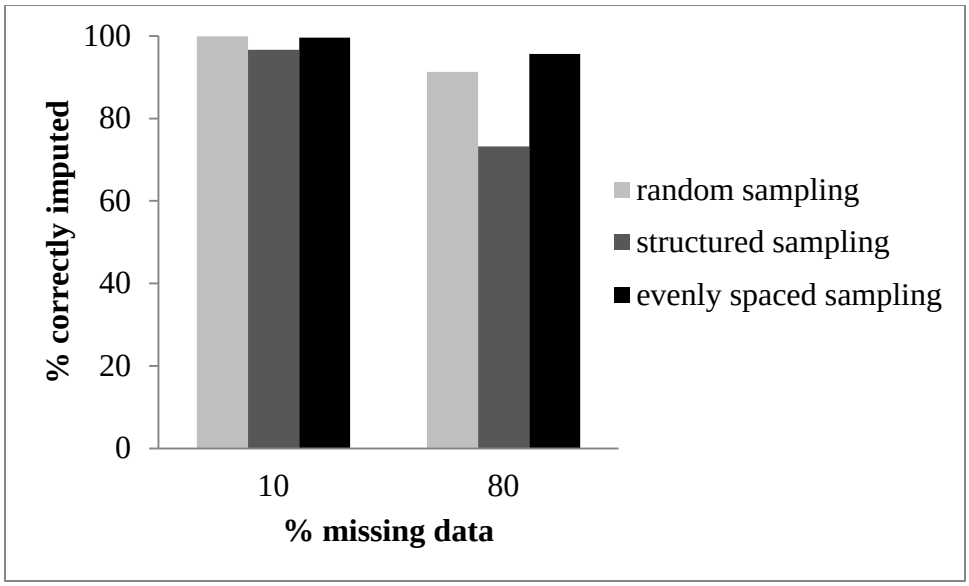


Figure 8. Effect of random, structured (non-genotyped) and evenly spaced patterns of missing genotypes on the accuracy of imputation, evaluated as percent correctly imputed values relative to the true values, for half-diallel mating design at 10% and 80% missing genotype data when imputation was carried out with BEAGLE software. For evenly spaced pattern of missing data, the imputation accuracy was the highest (0.96) at 80% missing

APPENDICES

Appendix 1. ASReml job file used for the imputation of missing genotypes with Gengler et al. (2007) approach

```
1       !ARGS 1
2       Title: Imputing missing genotypes using A matrix.
3       #clone,par1,par2,M1,M2,M3,M4,M5,M6,M7,M8,M9,M10
4       #26,1,12,0,0,0,1,1,1,0,0,0,0
5       #27,1,12,1,0,0,NA,1,1,1,0,0,0
6       clone  !P
7       par1   !I
8       par2   !I
9       M1 M2 M3 M4 M5 M6 M7 M8 M9 M10
10      pedigree.txt !SKIP 1 !sort
11      genecontent.csv !SKIP 1 !DOPATH $A

12      !PART 1
13      !CYCLE  M1 M2 M3 M4 M5 M6 M7 M8 M9 M10
14      $I ~ mu !r clone
15      1 1 1
16      0 !S2 ==.001
17      clone 1
18      clone 0 AINV 0.1 !GF
```

Explanation of the ASREML code

```
13   !CYCLE qualifier and $I process markers one by one and
create one     prediction file (.sln)
15   1 1 1 indicates there is one R structure with 1 dimension
and  there is 1 G structure
16   !S2==.001 fixes the error variance to 0.001
18   AINV explicitly requests to use the A inverse derived
from      pedigree and 0.1 !GF  fixes the clonal variance to
0.1
```

CHAPTER 3

Genetic Prediction Using Pedigree and Genomic Relationships in a Cloned Population of Loblolly Pine

Abstract

A cloned population of loblolly pine (*Pinus taeda* L.) and simulated data set was used to compare accuracies of genomic estimated breeding values based on the blended genomic and pedigree relationship matrix (**H**) with the estimates based on the additive genetic relationship matrix (**A**) derived from pedigree via the BLUP method. The genomic relationship matrices, **G** were constructed with different allele frequencies and combined with numerator relationship matrix (**A**) for the construction of **H** matrix. The effect of these **G** matrices on genomic estimated breeding values (GEBVs) and variance components was also evaluated. For the empirical data, the accuracy of predictions based on ABLUP matrix had accuracy of 0.79 and ranged from 0.72 to 0.76 among HBLUP methods. For the simulated data set, the accuracy of the predictions was 0.65 for ABLUP model and 0.68 and 0.73 for HBLUP when 300 and 600 individuals were used for the construction of **G** matrices, respectively. The accuracy of predictions was the same for all **G** matrices. Also, as the genotyped population size increased the accuracies increased.

Introduction

Genomic data has been used in animal breeding programs for genetic evaluations in addition to phenotypic and pedigree data. Two general methods have been used to estimate genomic breeding values: 1) estimating individual SNP marker effects simultaneously and 2) using a genomic relationship matrix (VanRaden et al. 2009, Misztal et al. 2009, Legarra et al. 2009). These models show differences from each other due to prior distribution of the marker effects (Meuwissen et al. 2001).

Even though these methods are very promising, extensive genotyping is not reasonable for all populations due to high cost and logistic limitations (Legarra et al. 2009). As a result, genomic evaluations are often implemented as a multi-step (two- or three-step) procedure where, first, data are obtained with a classic genetic evaluation from pedigree and phenotypic data, then pseudo-phenotype data are used in the second step where the whole genome assisted selection model is applied (VanRaden et al. 2008).

Genomic evaluation implemented as a multi-step procedure is similar to the classic evaluation, and has simple steps for predicting genomic values for genotyped individuals (VanRaden 2008, Hayes et al. 2009, Misztal et al. 2009). There are, however, multiple potential problems with this multi-step approach. Incorrect parameters such as known variances and weights and strong assumptions regarding prior distribution of marker effects may be used in multiple-step procedures (Aguilar et al. 2010), and to date the procedure has only been implemented with single trait analysis (no multi-trait analysis so far) (Legarra et al. 2009, Gray et al. 2012). In some cases the pseudo-phenotype data may be less accurate due

to small numbers of progeny for which phenotypic data are available (Vitezica et al. 2011, Gray et al. 2012). Also, the effects of incorrect parameters and assumptions such on the genomic evaluation results are difficult to confirm, especially in the presence of selection (Aguilar et al. 2010).

The multi-step procedure can be simplified by modifying the numerator relationship matrix **A** with genomic information (Misztal et al. 2009, Legarra et al. 2009). The mixed model equations require inverse of the numerator relationship matrix ($\mathbf{A}^{-1}$), which is very easy to construct for large populations (Henderson 1976). Even though $\mathbf{A}^{-1}$ is a sparse matrix, inversion of the modified matrix is likely to be difficult for large populations, because **A** is a dense matrix and is too large to store for large pedigrees (Misztal et al. 2009). One way to modify the **A** matrix is to substitute the relationships of genotyped individuals with the genomic relationship matrix **G**, but this can result in incoherencies because **G** includes information on relationships among ancestors and descendants (Legarra et al. 2009). Ignoring the genomic information and using pedigree based covariances between non-genotyped and genotyped individuals also leads to inconsistencies (Legarra et al. 2009). To solve this problem, Legarra et al. (2009) proposed to condition the genetic value of non-genotyped animals on the genetic value of genotyped animals via selection index. This is a joint distribution of genotyped and non-genotyped genetic values as a hybrid pedigree-genomic relationship matrix **H**. In the **H** matrix, which is semi (positive) definite by construction, genomic information is transmitted to the covariances among all non-genotyped individuals (Legarra et al. 2009). Misztal et al. (2009) also proposed a single-step computing procedure

based on a non-symmetric system of mixed model equations that was suitable for millions of animals. The single-step procedure uses all the available phenotype, pedigree and genotype data in one step (Aguilar et al. 2010, Chen et al. 2011), eliminates several assumptions and parameters used in the multi-step procedure, and enables to obtain more accurate genomic based estimates (Aguilar et al. 2010).The incorporation of genomic relationship matrix into the numerator relationship matrix allows estimation of breeding values using all phenotypic, pedigree and genomic information simultaneously (Forni et al. 2011).

Legarra et al. (2009) presented several **H** matrices, which are based on transmissions from genotyped individuals to their offspring or selection indexes from genotyped to non-genotyped individuals. Values in the **H** matrix have a multivariate distribution due to the fact that it is constructed based on selection index principles, and uses the assumption that **A** is multivariate normal (Legarra et al. 2009). Aguilar and Misztal (2008) presented the method which computes the sub-matrix of genotyped animals $\mathbf{A_{22}}$ explicitly.

Genotyping cost is still a bottle-neck in implementation of genomic prediction methods in forest trees. Combining information from different sources, such as pedigree data available for all the population and DNA markers available for a subset of individuals in the same population, can be an attractive strategy to improve predictions accuracies and increase genetic gain. We used a cloned population of loblolly pine (*Pinus taeda* L.) to examine the effect of a hybrid genetic relationship matrix **H** on prediction accuracies of genetic merit of trees and compared the results with a traditional genetic evaluation system.

The objective of this study was to investigate the effect of using a blended genomic and pedigree relationship matrix (**H**) on prediction accuracies of height growth in a cloned loblolly pine population. We also compared the results from empirical data with simulated data sets via BLUP method. Predicted breeding values and accuracies based on genomic information were compared with traditional method (based on pedigree). We employed different genomic relationship matrices, **G**, for the construction of the **H** matrix and evaluated the impact of those **G** matrices on estimated breeding values (EBVs) and variance components.

Materials and Methods

Numerator relationship matrix

The additive genetic relationship matrix, sometimes called the numerator relationship matrix (**A**), indicating the additive genetic relationship among individuals, can be calculated from the pedigree (Mrode 2005). **A** matrix has elements $A_{ij} = 2\ominus_{ij}$, where $\ominus_{ij}$ is the coefficient of coancestry: the probability of two genes to be identical by decent when drawing a gene randomly from two individuals (Lynch and Walsh 1998). By specifying relationships between ancestors, parents and offspring via pedigree, all breeding values are predicted on the same scale and all breeding values are obtained as a single set of rankings (White et al. 2007).

Genomic relationship matrix

Traditional genetic evaluations in animal and tree breeding programs are based on phenotypic and pedigree data. Genetic markers across the genome can be used to measure genetic resemblance and might be more accurate than pedigree (VanRaden 2008), by estimating the proportion of chromosome segments shared by individuals including identification of genes identical by state that might be shared by common ancestors not recorded in the pedigree (Forni et al. 2011). While the genomic relationship matrix **G** provides the actual proportion of the genome (observed relationships), the numerator relationship matrix, **A** gives the predicted proportion of the genome (average relationships) (Hayes et al. 2009). As a result, a **G** matrix can account for Mendelian sampling effects that

distinguish full-siblings within a single family, and can also detect unknown or distant relationships between individuals (Legarra et al. 2009).

Genomic relationship matrices were constructed by using observed allele frequencies (GOF), equal allele frequencies (G05), minor allele frequency (GMF) and normalized (GN) allele frequencies (VanRaden 2008, Forni et al. 2011). Let $\mathbf{u}_2 = \mathbf{Za}$, where $\mathbf{u}_2$ is a vector of genetic merit of individuals based on markers, **Z** is an incidence matrix and **a** is the SNP marker effect for an individual *i* and SNP *j* with allele frequency p_j. If matrix **Z** is centered by allele frequencies, then $\text{var}(\mathbf{u}_2) = \mathbf{ZZ}'\sigma_a^2 = \frac{\mathbf{ZZ}'}{k}\sigma_u^2 = \mathbf{G}\sigma_u^2$, where *k* is the scaling parameter ($k = 2\sum p_j(1-p_j)$), which is twice the sum of heterozygosities of the markers and σ_u^2 marker variance (VanRaden 2008).

Joint pedigree and genomic relationship matrix

Misztal et al. (2009) proposed to modify the numerator relationship matrix **A** with a **G** matrix. This modification resulted in an **H** matrix, which includes both pedigree and genomic relationships as

$$\textbf{[1]} \qquad \mathbf{H} = \begin{vmatrix} \mathbf{A_{11}} & \mathbf{A_{12}} \\ \mathbf{A_{21}} & \mathbf{G} \end{vmatrix} = \mathbf{A} + \begin{vmatrix} 0 & 0 \\ 0 & \mathbf{G} - \mathbf{A_{22}} \end{vmatrix}$$

where subscripts 1 and 2 stand for non-genotyped and genotyped individuals, respectively, and **G** is a genomic relationship matrix. Since off-diagonals in this matrix are not functions of the **G** matrix, using **G** to derive var($\mathbf{u}_2$) but not to derive cov($\mathbf{u}_1$, $\mathbf{u}_2$,) is not rational. Legarra et al. (2009) proposed deriving the joint density of non-genotyped and genotyped

individuals. The selection index used as a conditional distribution accounts for **G** to derive covariances between $\mathbf{u}_1$ and $\mathbf{u}_2$ and allows parsimonious inclusion of all information as full pedigree and genomic relationships. A blended relationship matrix allows a joint evaluation and all the information is explained automatically in just one step (Legarra et al. 2009).

Let **u** be a vector of genetic effects. Under a polygenic infinitesimal model of inheritance, $\text{var}(u) = \mathbf{A}\sigma_u^2$, where **A** is the numerator relationship matrix based on pedigree. Since there is no distinction between ancestor and progeny of genotyped individuals, there are two types of individuals in **u**: $\mathbf{u}_1$ non-genotyped individuals and $\mathbf{u}_2$ genotyped individuals. Then **A** can be partitioned as:

$\mathbf{A} = \begin{vmatrix} \mathbf{A_{11}} & \mathbf{A_{12}} \\ \mathbf{A_{21}} & \mathbf{A_{22}} \end{vmatrix}$ with the inverse $\mathbf{A^{-1}} = \begin{vmatrix} \mathbf{A^{11}} & \mathbf{A^{12}} \\ \mathbf{A^{21}} & \mathbf{A^{22}} \end{vmatrix}$

So the joint density of $\mathbf{u}_1$ and $\mathbf{u}_2$ would be $p(\mathbf{u}_1|\mathbf{u}_2) = p(\mathbf{u}_1|\mathbf{u}_2)p(\mathbf{u}_2)$, where $p(\mathbf{u}_2)$ is obtained from genomic data as a genomic relationship matrix (**G**). Based on pedigree via selection index theory and multivariate normal principles, the conditional distribution of $p(\mathbf{u}_1|\mathbf{u}_2)$ is:

$p(\mathbf{u}_1|\mathbf{u}_2) = N(\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{u}_2, \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})$

$\mathbf{u}_1 = E(\mathbf{u}_1|\mathbf{u}_2) + \varepsilon = \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{u}_2 + \varepsilon$

$\text{var}(\varepsilon) = \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} = (\mathbf{A^{11}})^{-1}$,

Substituting $\mathbf{u}_2$ with **Za,** where **Z** is an incidence matrix relating individuals to markers and **a** is the marker effects, then

$\mathbf{u}_1 = \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{Z}\mathbf{a} + \varepsilon$

$$\text{var}(\mathbf{u}_1) = \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} + \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} = \mathbf{A}_{11} + \mathbf{A}_{12}\mathbf{A}_{22}^{-1}(\mathbf{G} - \mathbf{A}_{22})\mathbf{A}_{22}^{-1}\mathbf{A}_{21}$$

$$\text{var}(\mathbf{u}_2) = \frac{\mathbf{Z}\mathbf{Z}'}{\text{k}} = \mathbf{G}$$

$$\text{cov}(\text{u}_1, \text{u}_2) = \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G}$$

The final covariance matrix of the joint distribution of $\mathbf{u}_1$ and $\mathbf{u}_2$ as an **H** matrix can be shown as:

$$[2] \qquad \mathbf{H} = \begin{vmatrix} \mathbf{A}_{11} + \mathbf{A}_{12}\mathbf{A}_{22}^{-1}(\mathbf{G} - \mathbf{A}_{22})\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G} \\ \mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{G} \end{vmatrix}$$

The **H** matrix is a semi-positive or positive definite matrix by construction (Legarra et al. 2009). The **G** matrix is semi-positive definite, but due to reasons such as the use of centered allelic coding (Stranden and Christensen 2011), if the number of markers is limited, if the number of individuals is more than the number of genotypes, or two individuals have identical genotypes across all markers it can be singular (VanRaden 2008). In the case when **G** is singular, an improved and non-singular **G** matrix can be obtained as $\text{p(u)} \sim \text{N}(0, \text{G}_\text{w})$, $\mathbf{G}_\text{w} = \text{w}\mathbf{G} + (1 - \text{w})\mathbf{A}$, where **A** is the numerator relationship matrix and *w* is the weight (VanRaden 2008). Using $k = 2\sum p_j(1 - p_j)$ produces **G** similar to the numerator relationship matrix **A**, approximately 1's on the diagonal as explained by VanRaden (2008). In other words, using observed allelic frequencies for the construction of **G** matrix, the population is assumed to be in Hardy-Weinberg equilibrium, and will be in linkage equilibrium conditions where **G** has an average value of 1's on the diagonal and 0's on the off-diagonal (Legarra 2011). Since **G** matrices in our study were singular we used weighted

G matrix as $\mathbf{G}_w$ ($\mathbf{G_w} = 0.95\mathbf{G} + 0.05\mathbf{A}_{22}$) to obtain inverse of the $\mathbf{G}^{-1}$ for the construction of **H** matrix as proposed by VanRaden (2008).

The numerator relationship matrix **A** assumes that the genetic values of the base population are 0 (Vitezica et al. 2011). The genomic relationship matrix **G** also assumes that the mean genetic value of the genotyped population is 0 if current allele frequencies are used for the calculation of **G** matrix (VanRaden 2008). In the presence of selection, the mean genetic value for genotyped individuals ($\mathbf{u}_2$) relative to the base population (**u**) might be different from 0 (Vitezica et al. 2011). If observed allele frequencies are used for the construction of **G** matrix, to account for the mean difference between genomic relationship matrix, **G** and numerator relationship matrix $\mathbf{A}_{22}$ for the genotyped individuals, different α values, which scales differences between genomic and pedigree-based information, were used for the construction of **H** matrices as, $p(u_2 I\, \alpha)(N{\sim}(0, G + 11'\alpha)\sigma_u^2)$, where 1 is a vector of one's (Vitezica et al. 2011). In this study we used α values varied between 0 to 2, including the difference between $\mathbf{A_{22}}$ and **G** off-diagonal means.

The appropriate choice of **G** matrix, used for the construction of **H** matrix, is significant because genomic and pedigree based relationship matrices should be compatible in terms of scale and structure (Aguilar et al. 2010). The accuracy of the single step approach depended on **G** matrix (Aguilar et al. 2010) and the adjustment placed on the difference between **G** and $\mathbf{A_{22}}$ matrices (Vitezica et al. 2011, Chen et al. 2011). They demonstrated that with the best choice, the accuracy of single-step approach was superior to multi-step approach. So, **H**

matrix might provide correct inferences if the matrix is reasonable and numerical errors are negligible (Legarra et al. 2009).

Data

In this study two data sets, actual and simulated, were used to evaluate the effectiveness of the single step genomic selection (HBLUP) approach.

Actual phenotypic data of 354 loblolly pine clones were provided by Plum Creek Timber Company, Inc. A total of nine crosses, which were mostly single-pair mating design were obtained using 13 parents as female and male (**Figure 1**). The number of clones per cross ranged from 16 to 83 (**Appendix 1**). The study was established on 16 sites located in South Carolina, Georgia, and Mississippi in 2000, 2001 and 2003. An alpha-lattice incomplete block design with single tree plot configuration was used for all sites. Tree height was measured to the nearest centimeter. Measurement of the height on a total of 7652 trees of 354 cloned full-sib individuals was carried out when trees were five years old. Complete pedigree was available for all data. All genotyped clones had phenotypic records for the height trait. Among 354 clones, 166 clones were genotyped at 3461 biallelic SNP markers. These markers were used to estimate genomic relationship coefficients.

The simulated data set was designed to mimic a tree breeding population. The simulator QMSim (Sargolzaei & Schenkel, 2009) was used to generate historical and recent populations. A genome of composed by a single100 cM chromosome was simulated. 1000 biallelic SNP markers were distributed at random across the chromosome with equal allele frequency. 10 loci out of 1000 were randomly chosen to be QTL and QTL allele effects were

sampled from a Gamma distribution with a shape parameter of 0.4. The mutation rate of the SNP markers and QTL were assumed to be $2.5x10^{-5}$ per locus per generation. A historical population with an effective population size of 3000 trees was generated over 100 generations and then 10 generations of selection for a trait were simulated where the trait heritability was 0.2 and phenotypic variance was 1. In each generation best 10 females and 10 males were selected and mated based on their estimated breeding values to produce 260 progeny per female by a poly-cross (pollen-mix) mating design. Ten replicate datasets were produced using these parameters. For each simulation true genetic values (TBVs), pedigree information (2600 individuals per generation), phenotypes and genotype data were available. For this simulation scenario, no fixed effects were simulated. In tree breeding programs, inbreeding is a big concern and selections are made to minimize the inbreeding. The QMSim simulator does not provide an option to control inbreeding, and inbreeding accumulated in the advanced generations, we analyzed only the first generation data sets for two replicates. Also among 2600 individuals, 1200 individuals were randomly selected and used for the analysis. As a result the number of progeny per full-sib family ranged from 5 to 18. The simulated data sets were used in two scenarios to determine the effect of the number of genotyped individuals on the accuracy. First, 600 individuals with phenotype only and 600 genotyped and phenotyped individuals were analyzed, and second, 900 individuals with phenotype only and 300 genotyped and phenotyped individuals were used to create **H** matrix.

Statistical Analysis

The following linear mixed model was used to estimate variance components and predict breeding values of trees for height growth for actual data

$$\mathbf{[3]} \qquad \mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e}$$

where **y** is the $n \times 1$ vector of observations (dependent variables) combined across all sites, **b** is the $t \times 1$ vector of site fixed effects, **u** is the vector of $a \times 1$ random combined genetic effect as the sum of polygenic effect and genomic genetic effect and **e** represents $n \times 1$ vector of residuals. $\mathbf{X}^{(\mathbf{n}\times\mathbf{t})}$, and $\mathbf{Z}^{(\mathbf{n}\times\mathbf{a})}$ are the design matrices which relate observations to the fixed and random effects, respectively. For the simulated data, no fixed effects were simulated, so **b** is a scalar (intercept).

Breeding values were predicted for the actual data of height trait and simulated data via BLUP method using ASREML (Gilmour et al., 2009). Mixed model equations were solved by using the numerator relationship matrix (**A**) and joint pedigree and genomic relationship matrix (**H**):

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z} \\ \mathbf{Z'X} & \mathbf{Z'Z} + \lambda\mathbf{A^{-1}} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{Z'y} \end{bmatrix}$$ where λ is the shrinkage factor as $\lambda = \alpha_e^2/\alpha_u^2$.

The $\mathbf{A}^{-1}$ was substituted with $\mathbf{H}^{-1}$ for genomic predicted breeding values. For traditional BLUP analysis (ABLUP) the variance-covariance matrix was assumed

$$\text{var}\begin{bmatrix} \mathbf{u} \\ \mathbf{e} \end{bmatrix} = \begin{bmatrix} \mathbf{A}\sigma_u^2 & 0 \\ 0 & \mathbf{I}\sigma_e^2 \end{bmatrix}$$

where **A** is the numerator relationship matrix, **I** is the identity matrix and σ_u^2 and σ_e^2 were additive and residual variances, respectively. In single step genomic selection (HBLUP), **A** matrix was replaced with **H** matrix

$$\text{var}\begin{bmatrix}\mathbf{u}\\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}\mathbf{H}\sigma_u^2 & 0\\ 0 & \mathbf{I}\sigma_e^2\end{bmatrix}$$

which combines the pedigree and genomic information (Aguilar et al. 2010, Christen and Lund, 2010).

Leave one out cross validation (LOOCV)

The predictive ability of the methods (ABLUP versus HBLUP) on the actual data was evaluated using a leave-one-out cross validation (LOOCV) approach. In the LOOCV approach, the model used *n*-1 trees (where *n* is the number of trees in the data) as a training data set. The coefficients from the prediction model were used to predict the genetic merit of the single omitted tree as validation. The procedure was repeated *n* times. The LOOCV method is a clever way to increase the power of prediction models when the genotyped population is small, as in our actual population.

Comparison of Models

For actual data set, the inflation of the prediction method in breeding values predicted from ABLUP and HBLUP was evaluated by regressing the true breeding values (TBVs) on the estimated breeding values (EBVs) using the complete and validation data sets, where a regression coefficient (slope) of 1 denotes no inflation. The values different from 1 are indication of under or over estimation of the breeding values. Bias was measured as the mean

differences between TBV and EBV. Accuracy of the methods (predictive ability) was determined as the correlation (r_p) between TBV and EBV. The Spearman rank correlation (r_s) was also used to see the rank changes. For the simulated data sets, bias and accuracy were obtained using the known TBVs and EBVs obtained from ABLUP and HBLUP methods using complete data sets.

Results

Results from empirical data

Descriptive statistics of pedigree based ($\mathbf{A}_{22}$) and genomic based ($\mathbf{G}$) relationship matrices

Statistics of pedigree-based and genomic relationship coefficients for the genotyped individuals ($\mathbf{A}_{22}$ or $\mathbf{G}$) for the actual data are given in **Table 1**. In the G05 and GMF genomic relationship matrices, the averages of both diagonal and off-diagonal elements were greater than the coefficients in $\mathbf{A}_{22}$. For GOF, the average diagonal and off-diagonal coefficients were smaller than the pedigree based coefficients, $\mathbf{A}_{22}$. For all genomic matrices, diagonal coefficients had greater variance than the pedigree-based diagonal coefficients, while the off-diagonal coefficients of the genomic matrices had smaller variance than the corresponding pedigree-based coefficients, except for the GN matrix, which had the same variance for off-diagonals elements as the pedigree-based matrix. The difference between average diagonal and off-diagonal elements of GOF matrix was similar to the difference between average diagonal and off-diagonal elements of A_{22} matrix.

Variance components

Variance components estimates from ABLUP and HBLUP prediction methods for the complete actual (354 clones) data are given in **Table 2**. Compared to estimates obtained with ABLUP method, most of the additive variance estimates using the HBLUP method were inflated regardless of the **G** matrix used to construct the **H** matrix, whereas the residual variance estimates were similar. For the ABLUP method, the variance explained by the clone effect was 2.81. Among HBLUP methods in which different **H** matrices were used, the

variance explained by the clone effect ranged from 10.42 ($HBLUP_{GN}$) to 14.88 ($HBLUP_{GO5}$). The error variance of the ABLUP method was 10.01 and ranged from 9.98 to 10.01 among HBLUP methods. ABLUP produced a variance component for clone identity effect, measuring environmental variation among ramets of the clones, whereas HBLUP models had zero estimates for the term. The effects of scaling parameter (α) and the different scaled GOFα matrices for the H matrix on breeding values are presented in Table 4. Different averages of off-diagonal elements of **G** would affect bias in single-step genomic selection (Chen et al. 2011). The scaling tends to increase the average BVs, but in terms of bias, we did not see any bias as shown by the differences between BVs of genotyped and all the population (Table 3).

Comparison of Models

Mean breeding values (Ave_EBVs), regression coefficients of TBVs on EBVs, the accuracies (r_p) of EBVs and Spearman rank correlations (r_s) between TBVs and EBVs for prediction methods for the validation population are given in **Table 3**. We assumed that BVs obtained by using all the phenotypic information and full pedigree are true BVs. The mean EBVs based on hybrid relationship matrices **H** for all the population ranged from -0.23 to 7.41.

Regression of TBVs on EBVs was used as a measure of the inflation of the prediction method, where a regression coefficient of 1 indicates no inflation. All the regression coefficients were lower than 1 for all the HBLUP validation methods. The slopes obtained from ABLUP method were high, approaching 1 for all the population and for the genotyped

population. The validation method based on G05 matrix (using equal allele frequency) had considerably larger bias than any other validation HBLUP models. However, despite a large bias, we did not observe a noticeable difference between models for accuracies of BVs or a difference for the rank correlations.

Accuracy (r_p) of the prediction methods was computed as the correlation between TBVs and EBVs. The accuracy of the predictions for the genotyped sub-population estimated with ABLUP method was 0.90, noticeably higher than the accuracies of predictions based on HBLUP methods regardless of the **G** matrix used for the construction of **H** matrices. However, this was not the case for all the population or the subset of the population that were not genotyped. We also examined accuracy of the prediction methods by scatter plots for the validation population (**Figure 2**). It was obvious that in the absence of phenotype, the ABLUP method did not discriminate among siblings within each full-sib cross. Full-siblings coming from a specific cross had very similar if not the same predictions when the numerator relationship matrix (ABLUP) was used. In contrast, two HBLUP prediction models using both markers and pedigree predicted different BVs for the full-siblings within each cross. The *ide(clone)* term in the model was statistically significant at 0.045 level based on a log likelihood ratio test. While having *ide(clone)* term in the model had a great impact on the estimates of additive variance obtained from ABLUP, HBLUP methods were not affected by this term (**Appendix 2**). The additive variance obtained from ABLUP was 10.54 and ranged from 10.42 to 14.88 among HBLUP methods. Fitting a model without *ide(clone)* term had also affected the accuracy of the BVs and rank correlations obtained from ABLUP and

HBLUP methods (**Appendix 3**). The slopes obtained from ABLUP method were high for all the population and 1.01 for the genotyped population. HBLUP methods, regardless of the **G** matrix used for the construction of **H** matrices, gave higher accuracies of BVs and rank correlations than ABLUP method.

Results from Simulated Data

Descriptive statistics of pedigree based (A22) and genomic based (G) relationship matrices

Statistics of pedigree-based and genomic relationship coefficients for the genotyped individuals ($\mathbf{A}_{22}$ or **G**) calculated from the simulated data are given in **Table 5** (**Appendix 4** for the Rep 2 results). When either 300 or 600 genotyped individuals were used for the construction of G05 genomic relationship matrix, the average of both diagonal and off-diagonal elements were greater than the coefficients in $\mathbf{A}_{22}$. For GOF and GN matrices, the average diagonal and off-diagonal coefficients were very similar to the pedigree based coefficients, $\mathbf{A}_{22}$. For all genomic matrices constructed for 300 or 600 genotypes, the diagonal and off-diagonal coefficients had greater variances than the pedigree-based coefficients, $\mathbf{A}_{22}$.

Variance components

Variance components estimates from ABLUP and HBLUP prediction methods for the simulated data are given in **Table 6** (**Appendix 5** for the Rep 2 results). The additive variance obtained from ABLUP method was the lowest (0.14) compared to additive variance estimates obtained with HBLUP methods for both simulated populations. ABLUP method also gave the slightly higher error variance (0.90) compared to the HBLUP methods.

Increasing the validation population size from 300 to 600 individuals seems to increase the additive variances using H matrices but error variances were similar.

Comparison of Models

As with the empirical data, we also compared predictions from ABLUP and HBLUP methods using various statistics and by regressing the TBVs on EBVs (**Table 7**). In general, the regression slopes of various models were similar to slopes of ABLUP models and were close to 1 when a smaller validation population was used (300 genotyped individuals) for the construction of **G** matrices (**Table 7a**). When a larger population was used (600 genotyped individuals) for the calculations of **G** matrices the slopes were slightly lower than 1 for HBLUP methods (**Table 7b**).

Bias was measured as the mean difference between TBVs and breeding values obtained using ABLUP and HBLUP methods (EBVs) (**Appendix 6** for the Rep 2 results). We did not see any difference in bias between all the models regardless of population size or the **G** matrix used, except for the **G** matrix based on GN method. The accuracy values for HBLUP models were higher compared to ABLUP models, particularly for the genotyped individuals. This was true for the simulated larger population as well. The Spearman rank correlations (r_s) between TBVs and EBVs from HBLUP method were again higher than the estimates for the ABLUP models, regardless of the population size.

Scatter plots of predictions are quite informative to understand the nature of correlation coefficients, and are presented in **Figure 3** (**Appendix 7** for the **Rep 2** results). Unlike the result from the empirical data, we did not observe discontinuous distributions of EBVs from

the ABLUP model for the two simulated validation populations. For illustration, scatter plots of TBVs and EBV based on GOF genetic relationship matrix are presented for both simulated populations. When the size of the prediction population increased from 300 to 600 individuals, we observe a tighter distribution (better agreement) between TBV and EBV predictions.

Discussion

In tree breeding populations, traditional genetic evaluations are based on phenotypic data and pedigree. Recently, with the availability of DNA markers, tree breeding may employ new methods to integrate marker data in the genomic evaluations. Blending genetic relationships from pedigree and DNA markers could be an appealing procedure for forest tree breeding because genotyping cost is still prohibitively expensive to genotype large number of trees. In creating blended genetic relationship **H** matrices, different methods were employed in the literature for estimation of genomic relationships matrices (Forni et al. 2011). The inverse of these blended matrices are then used in a BLUP approach to predict breeding values of individuals (Van Raden 2008). One of the criterions used to judge the success of genetic covariances based on DNA markers is to compare with genetic covariances based on pedigree for the same individuals.

Using actual and simulated data sets, we found that when genomic relationships were estimated using observed allele frequencies (GOF), and normalized allele frequencies (GN), the average diagonal and off-diagonal coefficients were very similar to the pedigree based coefficients. Forni et al. (2011) also reported similar average diagonal elements between **G** and **A** matrices when observed allele frequencies (GOF) were used for a pig breeding population. Chen et al. (2011) compared the diagonal and off-diagonal elements of numerator relationship matrix with the elements of **G** matrices based on equal allele frequencies (GO5) and observed allele frequencies (GOF) for broiler chickens. They found that diagonal

elements were close to 1 and off-diagonal elements were 0, regardless of the threshold set for the minor allele frequency.

Using actual and simulated data set in this study, we found greater variance among diagonal coefficients of **G** than the numerator relationship matrix ($\mathbf{A}_{22}$), which is a sub matrix of individuals with marker genotypes. Higher variance among diagonal elements was the case for all the genomic matrices. Similarly, Forni et al. (2011) found greater variance between the elements of **G** than $\mathbf{A}_{22}$. The genomic relationship is an estimate of the realized proportion of the genome shared by two individuals, but the pedigree-based relationship is the expectation of this proportion (Goddard et al. 2011). Since genomic relationships reflect the actual fraction of genome shared, greater variance was expected between the elements of **G** than $\mathbf{A}_{22}$ which are predictions (Forni et al. 2011).

The difference between the average diagonal elements of $\mathbf{A_{22}}$ and **G** matrices is as a result of additional genomic information in **G** compared with $\mathbf{A_{22}}$, and also the noise due to the number of SNP markers used to construct **G** and incorrect scaling of **G** (Chen et al. 2011). Using correct scaling and large number of SNP markers might help to increase genomic information and decrease the noise (Chen et al. 2011). Unfortunately we did not have a large enough number of markers to be able to sub-sample them to see the effect of genomic information on the predictions accuracies as Chen et al (2011) did. However, in this study, we looked at the effect of validation population size on the accuracies for the simulated data set (Table 7). For the simulated data, we found that when we increased the number of

genotyped individuals from 300 to 600, using the same number of markers, the accuracy of predictions from HBLUP models increased.

Variances and the effect of scaling G matrices

For the actual data, the additive variance estimates using the HBLUP method were relatively higher compared with ABLUP method and error variances were similar. For the simulated data set, ABLUP method gave smaller additive variance and higher error variance estimates compared to the HBLUP method. The simulation results suggest that as the validation population increases, additive variances tend to be higher and error variances tend to be smaller. For genomic relationship matrices, with diagonal elements smaller than 1, additive genetic variance estimates were larger in a small data set (Forni et al. 2011). The deviation from 1 is associated with upward biases in estimates of the additive variance. Upward biased additive genetic variances might cause greater (biased) accuracy values of predictions (less shrinkage in HBLUP). When we increased the scale factor (α) in HBLUP based on observed allele frequencies (GOF), the mean EBVs for genotyped individuals increased. However, the scale factor had no effect on the average difference of EBVs between genotyped and all trees (Table 4). Genomic and pedigree-based matrices should be compatible in scale to be integrated (Aguilar et al. 2010, Christensen and Lund 2010, Chen et al. 2011, Vitezica et al. 2011). For the actual data, α had effect on the average breeding values and regression coefficients but not on the accuracy of predictions and Spearman rank correlations. We did not observe any effect of scale factor on accuracy of predictions or on the regression coefficients from the simulated data set. This result might be due to our population structure,

or also due to lack of selection effect. When genomic relationships were scaled by a constant, the single-step method was unbiased and the most accurate for a simulation study (Vitezica et al. 2011), and for broiler chicken (Chen et al. 2011).

Comparison of Models

In any breeding program, the most important aim is increasing genetic gain. Accuracy is currently used to assess genomic prediction methods in cross-validation studies (VanRaden et al. 2009). In the presence of bias, genetic gain might be overestimated or underestimated. Since MSE also includes bias, it should be the criterion of choice (Vitezica et al. 2011). Although all the phenotypic, pedigree and genomic information are used in HBLUP, only genotyped (and mostly selected) individuals are included in **G**. Since genetic values of non-genotyped animals are conditioned on genetic values of genotyped animals (Legarra et al., 2009), the bias is decreased but may not be fully corrected (Vitezica et al. 2011). In a single-step method with correction α, bias is eliminated by referencing genomic and pedigree-based relationship matrices to the base population (Vitezica et al. 2011). As a result single-step genomic selection method with scaling gave the most accurate results (Vitezica et al. 2011). In our study using the actual data set, ABLUP gave smaller MSE than HBLUP methods regardless of the **G** matrix used. This could be due to a smaller number of individuals (166) genotyped, and also smaller number of markers (3461) used to estimate genomic relationships. In contrast, we observed less bias (smaller MSE) for HBLUP methods using the simulated datasets. The MSE value decreased as the number of genotyped individuals increased (300 vs. 600).

Regression of TBVs on EBVs was used as a measure of the differences in predicted breeding values from ABLUP and HBLUP methods, where a regression coefficient of 1 indicates no inflation. In general, the slopes were lower than the expected value of 1 for ABLUP and HBLUP methods for actual data. For the simulated data HBLUP methods, regardless of the **G** matrices used, were superior to the ABLUP method. The number of genotypes used for the construction of **G** matrices affected the slopes, MSE, accuracies and Spearman rank correlations.

Capturing the Mendelian sampling term

In traditional genetic evaluations based on pedigree, Mendelian sampling is ignored among full-sibs within a family because average additive genetic covariances cannot discriminate among full-siblings in the absence of phenotype data. However, genomic markers allow us to capture segregation among full-siblingss even if we do not have phenotype data. Genomic analysis resulted in higher accuracies due to accounting for Mendelian sampling (VanRaden et al. 2009, Forni et al. 2011, Wolc et al. 2011, Gray et al. 2012). In another study by Daetwyler et al. (2007) the accuracy of genomic predicted breeding values were higher as a result of better estimation of Mendelian sampling. Capturing the Mendelian sampling effect can be an efficient method to control the average inbreeding level in breeding populations (Daetwyler et al. 2007).

Conclusion

The leave-one-out cross validation method of empirical analysis showed that HBLUP approach based on observed allele frequency matrix (**G**) gave similar prediction accuracies and rank correlations of genomic estimated breeding values compare to the ABLUP method. The blended genetic relationship matrices in a BLUP approach enabled us to capture segregation of Mendelian sampling effect within full-sib crosses when phenotypes were not available in validation populations. For simulated data, we observed higher prediction accuracies, rank correlations and similar bias and less inflation regardless of the choice of genomic relationship matrix (**G**) used for the construction of **H** matrix compare to ABLUP method. For both empirical and simulated data the scaling factor had no effect on the prediction accuracies and bias of EBVs. HBLUP uses all the available phenotype, pedigree and genotype data in a single-step and is easy to implement for genomic predictions. Blending genetic relationships obtained from pedigree and DNA markers is an attractive strategy for forest tree breeding evaluations to capture the Mendelian sampling effect, which is important to manage inbreeding and maximize genetic gain.

References

Aguilar, I., Misztal, I., Johnson, D.L., Legarra, A, Tsuruta, S., Lawlor, T.J., 2010. Hot topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. Journal of Dairy Science 93, 743–52.

Chen, C.Y., Misztal, I., Aguilar, I., Legarra, A, Muir, W.M., 2011. Effect of different genomic relationship matrices on accuracy and scale. Journal of Animal Science 89, 2673–2679.

Chen, C.Y., Misztal, I., Aguilar, I., Tsuruta, S., Meuwissen, T.H.E., Aggrey, S.E., Wing, T., Muir, W.M., 2011. Genome-wide marker-assisted selection combining all pedigree phenotypic information with genotypic data in one step: An example using broiler chickens. Journal of Animal Science 89, 23–28.

Christensen, O.F., Lund, M.S., 2010. Genomic prediction when some animals are not genotyped. Genetics Selection Evolution 42, 2.

Daetwyler, H.D., Villanueva, B., Bijma, P., Woolliams, J.A, 2007. Inbreeding in genome-wide selection. Journal of Animal Breeding and Genetics 124, 369–76.

Forni, S., Aguilar, I., Misztal, I., 2011. Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. Genetics Selection Evolution 43, 1.

Gilmour, A.R., B.J. Gogel, B.R. Cullis, R. Thompson. 2009. ASReml user guide release 3.0. VSN International Ltd., Hemel Hempstead, HP1 1ES, United Kingdom. 372 p.

Goddard, M.E., Hayes, B.J., Meuwissen, T.H.E., 2011. Using the genomic relationship matrix to predict the accuracy of genomic selection. Journal of Animal Breeding and Genetics 128, 409–21.

Gray, K. A., Cassady, J.P., Huang, Y., Maltecca, C., 2012. Effectiveness of genomic prediction on milk flow traits in dairy cattle. Genetics Selection Evolution 44, 24.

Hayes, B.J., Visscher, P.M., Goddard, M.E., 2009. Increased accuracy of artificial selection by using the realized relationship matrix. Genetics Research 91, 47–60

Henderson, C.R. 1984. Applications of linear models in animal breeding. University of Guelph, Guelph, Canada. 462 p.

Legarra, A, Aguilar, I., Misztal, I., 2009. A relationship matrix including full pedigree and genomic information. Journal of Dairy Science 92, 4656–63.

Lynch, M., Walsh, B. 1998. Genetics and analysis of quantitative traits. Sinauer Associates, Inc., Sunderland, MA, USA, 980 p.

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

Misztal, I., Legarra, A, Aguilar, I., 2009. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. Journal of Dairy Science 92, 4648–55.

Mrode, R., 2005. Linear models for the prediction of animal breeding values, 2nd edition, CAB international , Wallingford, Oxfordshire, OX10 8DE, UK.

R Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

SAS Institute Inc. 2010. SAS 9.1 User‟s guide statistics. SAS Institute Inc., Cary, NC, USA

Sargolzaei, M., Schenkel, F., 2009. QMSim: a large-scale genome simulator for livestock. Bioinformatics 25, 680–681.

Strandén, I., Christensen, O.F., 2011. Allele coding in genomic evaluation. Genetics, selection, evolution 43, 25.

VanRaden, P.M., 2008. Efficient methods to compute genomic predictions. Journal of Dairy Science 91, 4414–4423.

VanRaden, P. M., C. P. Van Tassell, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor, and F. S. Schenkel. 2009. Invited review: Reliability of genomic predictions for North American Holstein bulls. Journal of Dairy Science. 92,16–24.

Vitezica, Z.G., Aguilar, I., Misztal, I., Legarra, a, 2011. Bias in genomic predictions for populations under selection. Genetics Research 93, 357–366.

White, T.L., W.T. Adams, D.B. Neale. 2007. Forest Genetics. CABI Publishing CAB International. Cambridge, MA 02139 UK, USA. 702 p.

Wolc, A., Stricker, C., Arango, J., Settar, P., Fulton, J.E., O'Sullivan, N.P., Preisinger, R., Habier, D., Fernando, R., Garrick, D.J., Lamont, S.J., Dekkers, J.C.M., 2011. Breeding value prediction for production traits in layer chickens using pedigree or genomic relationships in a reduced animal model. Genetics Selection Evolution 43, 5.

TABLES

Table 1. Descriptive statistics of coefficients of relationships estimated using pedigree and SNP markers from actual data. The coefficients for the A matrix are based on 354 trees, whereas the coefficients of G matrices are estimated for 166 trees and 3461SNP markers

	Diagonal elements				**Off diagonal elements**			
Matrix	Mean	Min	Max	Variance	Mean	Min	Max	Variance
A_2	1.000	1.000	1.000	0.000	0.112	0.000	0.500	0.032
GOF	0.962	0.830	1.141	0.007	0.047	-0.188	0.949	0.028
G05	3.378	3.263	3.612	0.002	2.950	2.274	3.419	0.008
GMF	2.699	2.429	3.029	0.018	1.935	1.688	2.734	0.022
GN	1.025	0.884	1.216	0.008	0.047	-0.205	1.014	0.032

Table 2. Variance component obtained from ABLUP and various HBLUP prediction methods for the complete actual data (354 trees)

	Variance components		
Prediction Method	$*\sigma^2_c$	$\sigma^2_{c\_ide}$	σ^2_e
ABLUP	2.81	3.98	10.01
HBLUP$_{GOF}$	10.82	0	10.01
HBLUP$_{GOF\alpha}$	10.81	0	10.01
HBLUP$_{GO5}$	14.88	0	9.98
HBLUP$_{GMF}$	11.57	0	10.00
HBLUP$_{GN}$	10.42	0	10.01

$*\sigma^2_c$ is the variance explained by the tree effect, $\sigma^2_{c\_ide}$ is the variance explained by ide(clone) term in the model to remove genetic relationship matrix from copies of the same trees (to make the relationship matrix as identity) and σ^2_e is the residual variance

Table 3. Mean breeding values (EBVs), regression coefficients (intercept and slope) of TBVs on EBVs, mean squared error (MSE), bias (the difference between TBVs and EBVs), Pearson moment correlations (r_p) between TBVs and EBVs and Spearman rank correlations (r_s) for ABLUP and HBLUP prediction methods for the leave-one-out cross validation data

Method	Data	Ave_EBV	Intercept	Slope	MSE	Bias	r_p	r_s
ABLUP (ide)	all	0.17	0.00	0.99	0.38	0.00	0.79	0.73
	geno	0.46	0.10	1.03		0.11	0.90	0.85
	non-geno	-0.08	-0.09	0.81		-0.10	0.58	0.33
HBLUP$_{GOF}$	all	-0.23	0.33	0.68	0.49	0.40	0.73	0.68
	geno	0.02	0.44	0.64		0.43	0.79	0.78
	non-geno	-0.45	0.24	0.71		0.38	0.57	0.29
HBLUP$_{GOF\alpha}$	all	0.22	0.02	0.68	0.48	-0.04	0.73	0.68
	geno	0.46	0.14	0.64		-0.01	0.79	0.78
	non-geno	0.00	-0.08	0.72		-0.08	0.58	0.32
HBLUP$_{GO5}$	all	7.41	-4.77	0.67	0.43	-7.24	0.76	0.77
	geno	7.89	-4.52	0.63		-7.44	0.77	0.78
	non-geno	6.99	-5.87	0.83		-7.07	0.72	0.67
HBLUP$_{GMF}$	all	4.06	-2.63	0.69	0.44	-3.89	0.76	0.75
	geno	4.46	-2.47	0.65		-4.00	0.79	0.76
	non-geno	3.71	-2.99	0.79		-3.79	0.65	0.56
HBLUP$_{GN}$	all	-0.24	0.34	0.67	0.49	0.41	0.72	0.68
	geno	-0.01	0.46	0.63		0.46	0.79	0.78
	non-geno	-0.45	0.24	0.71		0.37	0.57	0.30

*geno: genotyped

*non-geno: non-genotyped

Table 4. The effect of scaled GOF matrix, based on observed allele frequencies, on breeding values. Different α values were used for the same HBLUP model to make the GOF matrix analogous to the A matrix of the genotyped individuals

	α	**$G0F_{\alpha}$-A_{22}**	**EBV_{gen}**	**EBV_{all}**	**EBV_{gen}-EBV_{all}**
$ABLUP_{ide}$	-	-	0.46	0.17	0.28
$HBLUP_{GOF}$	0.00	-0.650	0.07	-0.50	0.57
	0.06	-0.005	0.37	-0.20	0.57
	0.11	0.045	0.58	0.01	0.57
	0.12	0.055	0.63	0.06	0.57
	0.13	0.065	0.67	0.10	0.57
	0.14	0.075	0.72	0.15	0.57
	0.15	0.085	0.74	0.17	0.57
	0.20	0.105	0.97	0.40	0.57

α, scale parameter

$G0F_{\alpha}$ - A_{22}, The difference between average of off-diagonal elements of scaled **GOF** and **A_{22}**

Table 5. Descriptive statistics of coefficients of relationship estimated using pedigree and genomic information from simulated data for the construction of H matrix for 300 and 600 genotypes

a) Rep 1: 300 genotypes

	Diagonal elements				Off diagonal elements			
Matrix	Mean	Min	Max	Variance	Mean	Min	Max	Variance
A_2	1.000	1.000	1.000	0.000	0.050	0.000	0.500	0.011
GOF	0.996	0.911	1.188	0.002	0.048	-0.142	1.071	0.020
G05	1.503	1.332	1.652	0.003	0.593	0.344	1.583	0.020
GMF	0.004	0.000	0.011	0.000	-0.000	-0.005	0.002	0.000
GN	1.025	0.937	1.224	0.002	0.048	-0.148	1.103	0.022

b) Rep 1: 600 genotypes

	Diagonal elements				Off diagonal elements			
Matrix	Mean	Min	Max	Variance	Mean	Min	Max	Variance
A_2	1.000	1.000	1.000	0.000	0.050	0.000	0.500	0.012
GOF	0.999	0.904	1.183	0.002	0.049	-0.139	1.144	0.020
G05	1.507	1.333	1.762	0.003	0.594	0.341	1.704	0.021
GMF	0.005	0.000	0.011	0.000	-0.000	-0.004	0.003	0.000
GN	1.025	0.927	1.215	0.002	0.050	-0.145	1.175	0.022

Table 6. Variance component estimates obtained from ABLUP and HBLUP prediction methods for the complete simulated data using different number of genotyped individuals (300 genotyped/900 non-genotyped and 600 genotyped/600 non-genotyped) for the construction of H matrix

	300 genotypes		600 genotypes	
Prediction Method	* σ^2_a	σ^2_e	σ^2_a	σ^2_e
ABLUP	0.14	0.90	0.14	0.90
HBLUP$_{GOF}$	0.17	0.89	0.19	0.88
HBLUP$_{GOF\alpha}$	0.17	0.89	0.19	0.88
HBLUP$_{GO5}$	0.19	0.89	0.20	0.87
HBLUP$_{GMF}$	0.18	0.90	0.18	1.00
HBLUP$_{GN}$	0.16	0.89	0.19	0.88

* σ^2_a Where genetic variance due to trees

Table 7. Mean breeding values (EBVs), regression coefficients (intercept and slope) of TBVs on EBVs, mean squared error (MSE), bias (the difference between TBVs and EBVs), Pearson moment correlations (r_p) between TBVs and EBVs and Spearman rank correlations (r_s) for ABLUP and HBLUP prediction methods using simulated data used for the construction of H matrix

a) Rep1 (1200 individuals in total, 300 genotyped and 900 non-genotyped)

Method	**data**	**intercept**	**Slope**	**Bias**	**MSE**	**r_p**	**r_s**
ABLUP	all	0.62	1.05	0.01	0.092	0.65	0.63
	geno	0.62	0.98	0.01		0.66	0.64
	non-geno	0.62	1.08	0.01		0.65	0.62
HBLUP$_{GOF}$	all	0.61	1.02	0.01	0.086	0.68	0.65
	geno	0.61	0.97	0.01		0.74	0.73
	non-geno	0.61	1.05	0.01		0.65	0.63
HBLUP$_{GOF\alpha}$	all	0.61	1.02	0.01	0.086	0.68	0.65
	geno	0.61	0.97	0.01		0.74	0.73
	non-geno	0.61	1.05	0.01		0.65	0.63
HBLUP$_{GO5}$	all	0.54	1.02	0.01	0.086	0.68	0.65
	geno	0.54	0.96	0.01		0.74	0.73
	non-geno	0.53	1.04	0.01		0.65	0.63
HBLUP$_{GN}$	all	0.61	1.03	0.00	0.086	0.68	0.65
	geno	0.61	0.97	-0.02		0.74	0.73
	non-geno	0.61	1.06	0.00		0.65	0.63

*geno: genotyped

*non-geno: non-genotyped

Table 7. Continued.

b) Rep1 (1200 individuals in total, 600 genotyped and 600 non-genotyped)

Method	data	Intercept	Slope	Bias	MSE	r_p	r_s
ABLUP	all	0.62	1.05	0.01	0.092	0.65	0.63
	geno	0.62	1.06	0.00		0.66	0.64
	non-geno	0.63	1.04	0.02		0.65	0.62
$HBLUP_{GOF}$	all	0.62	0.96	0.01	0.073	0.73	0.70
	geno	0.62	0.95	0.01		0.81	0.78
	non-geno	0.62	0.97	0.01		0.64	0.61
$HBLUP_{GOF\alpha}$	all	0.62	0.96	0.01	0.073	0.73	0.70
	geno	0.62	0.95	0.01		0.81	0.78
	non-geno	0.63	0.97	0.01		0.64	0.61
$HBLUP_{GO5}$	all	0.61	0.96	0.01	0.073	0.73	0.70
	geno	0.60	0.95	0.01		0.81	0.78
	non-geno	0.61	0.97	0.01		0.64	0.61
$HBLUP_{GN}$	all	0.62	0.96	0.01	0.073	0.73	0.70
	geno	0.62	0.95	0.01		0.81	0.78
	non-geno	0.62	0.97	0.01		0.64	0.61

*geno: genotyped

*non-geno: non-genotyped

FIGURES

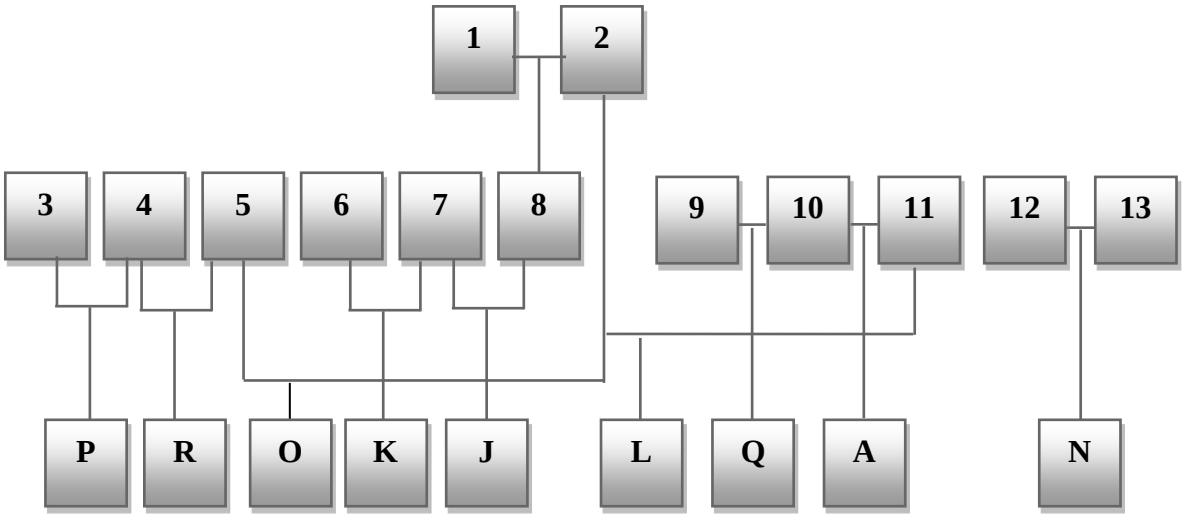


Figure 1. Pedigree structure of 354 cloned trees (166 genotyped, 188 non-genotyped) used for the data analysis. Parents 2, 4, 5, 7 and 10 were used as female and male. Letters refer nine family full-sib ID's.

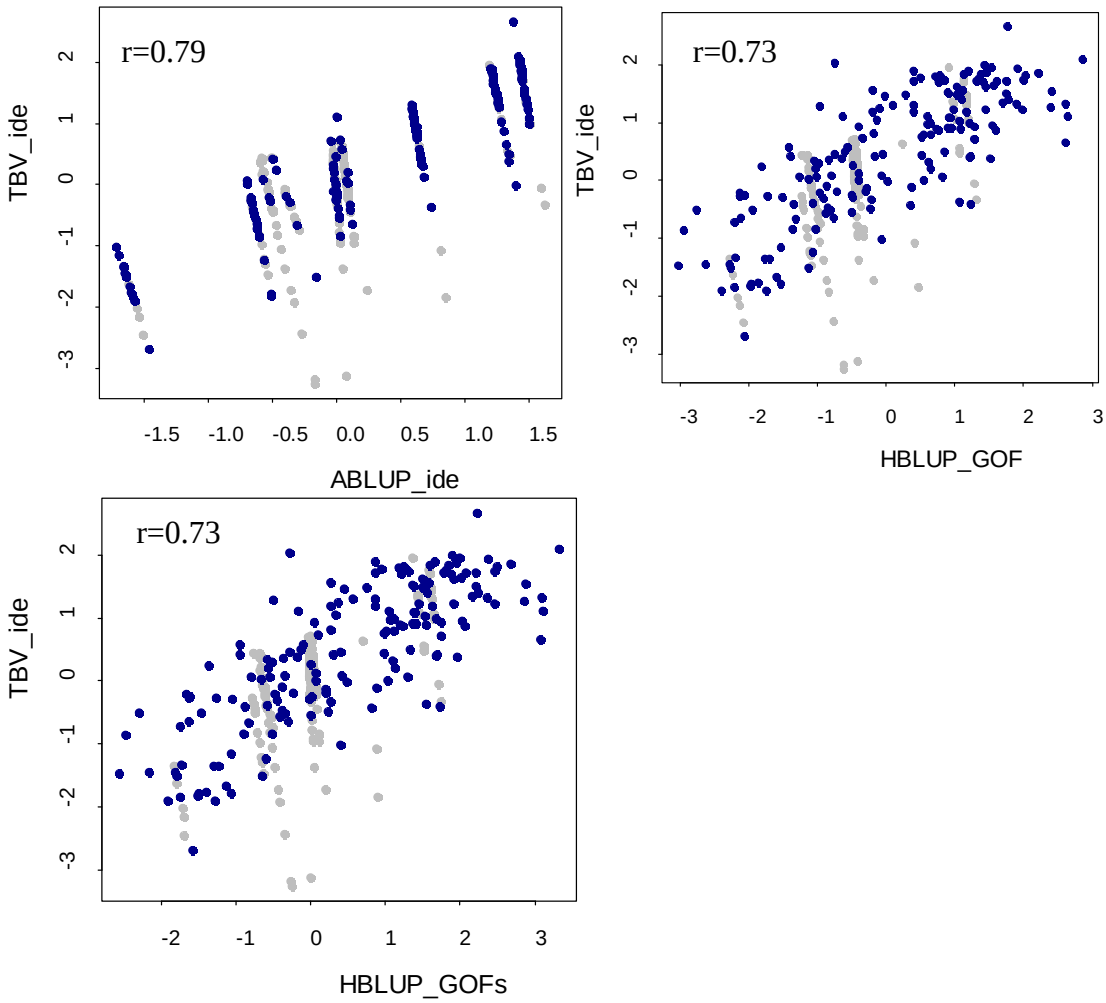
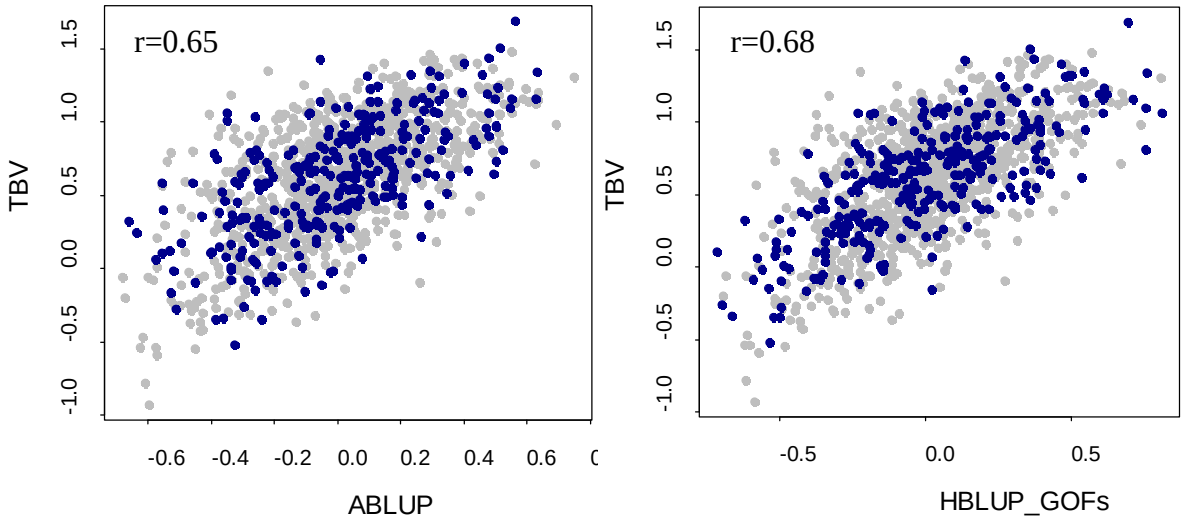


Figure 2. Scatter plots between true breeding values (TBVs) and estimated breeding values (EBVs) obtained from ABLUP and HBLUP methods using actual data. TBVs were predicted from complete data and EBVs were obtained using leave-one-out data set (354 trees: 166 genotyped and 188 non-genotyped). Genotyped individuals are in dark blue and non-genotyped individuals are in gray.

a) When 300 genotyped and 900 non-genotyped individuals were used for the **H** matrix



b) When 600 genotyped and 600 non-genotyped individuals were used for the **H** matrix

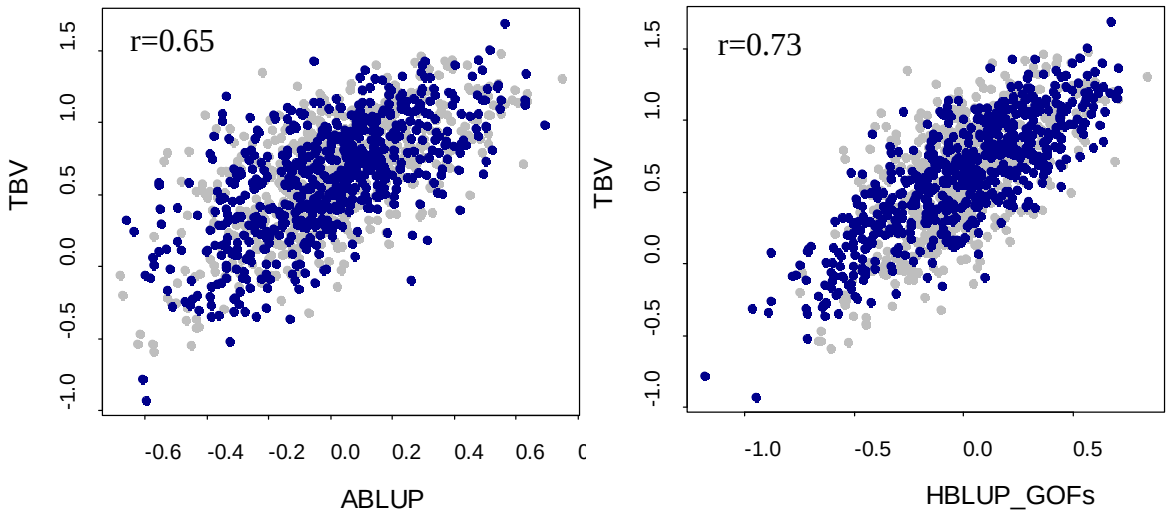


Figure 3. Scatter plots between true breeding values (TBVs) and estimated breeding values (EBVs) obtained from ABLUP and HBLUP prediction methods using simulated data (rep1). Genotyped individuals are in dark blue and non-genotyped individuals are in gray.

APPENDICES

Appendix 1. Detail information about the actual data: family ID's, number of clones per family and number of ramets per family.

	#clones			**# ramet**		
Fam_ID	non-genotyped	genotyped	total	non-genotyped	genotyped	total
P	5	16	21	69	246	315
R	10	22	32	83	344	427
K	28	3	31	491	145	636
A	10	6	16	171	227	398
N	67	16	83	969	496	1465
L	44	9	53	878	403	1281
J	3	37	40	39	1122	1161
O	7	27	34	98	780	878
Q	14	30	44	318	793	1111
total	188	166	354	3116	4556	7672

Appendix 2. Variance component obtained from ABLUP and various HBLUP prediction methods for the complete **actual data** (354 trees)

	Variance components	
Prediction Method	σ^2_c	σ^2_e
ABLUP	10.54	10.01
HBLUP$_{GOF}$	10.82	10.01
HBLUP$_{GOF\alpha}$	10.81	10.01
HBLUP$_{GO5}$	14.88	9.98
HBLUP$_{GMF}$	11.57	10.00
HBLUP$_{GN}$	10.42	10.01

σ^2_c is the variance explained by the tree effect, and σ^2_e is the residual variance

Appendix 3. Regression coefficients (intercept and slope) of TBVs on EBVs, mean squared error (MSE), bias (the difference between TBVs and EBVs), Pearson moment correlations (r_p) between TBVs and EBVs and Spearman rank correlations (r_s) for ABLUP and HBLUP prediction methods for the leave-one-out cross validation data

Method	data	intercept	slope	Bias	MSE	r_p	r_s
ABLUP (no_ide)	all	0.02	0.87	0.00	4.84	0.34	0.29
	geno	0.37	1.01	0.37		0.54	0.46
	non-geno	-0.32	0.28	-0.33		0.08	
HBLUP$_{GOF}$	all	0.37	0.78	0.42	4.75	0.36	0.37
	geno	0.74	0.83	0.74		0.56	0.57
	non-geno	-0.20	0.26	0.14		0.07	-0.11
HBLUP$_{GOF\alpha}$	all	0.01	0.79	-0.02	4.73	0.37	0.37
	geno	0.34	0.83	0.30		0.56	0.57
	non-geno	-0.31	0.32	-0.31		0.09	-0.08
HBLUP$_{GO5}$	all	-6.32	0.88	-7.22	4.43	0.44	0.48
	geno	-5.50	0.79	-7.13		0.53	0.54
	non-geno	-6.65	0.91	-7.30		0.27	0.32
HBLUP$_{GMF}$	all	-3.29	0.86	-3.87	4.56	0.41	0.45
	geno	-2.96	0.83	-3.70		0.55	0.56
	non-geno	-2.65	0.63	-4.02		0.18	0.19
HBLUP$_{GN}$	all	0.38	0.78	0.44	4.76	0.36	0.37
	geno	0.77	0.83	0.77		0.56	0.57
	non-geno	-0.2	0.26	0.14		0.07	-0.11

Appendix 4. Descriptive statistics of relationship coefficients estimated using pedigree and genomic information from **simulated data** for the construction of **H** matrix

Rep 2: 300 genotypes

	Diagonal elements				**Off diagonal elements**			
Matrix	Mean	Min	Max	Variance	Mean	Min	Max	Variance
A_2	1	1	1	0	0.050	0	0.5	0.011
GOF	0.999	0.886	1.149	0.002	0.048	-0.131	1.037	0.020
G05	1.509	1.378	1.705	0.002	0.595	0.365	1.571	0.019
GMF*	0.004	0.00	0.010	0.000	-0.00	-0.005	0.005	0.000
GN	1.025	0.909	1.179	0.002	0.048	-0.136	1.065	0.021

Rep 2: 600 genotypes

	Diagonal				**Off diagonal**			
Matrix	Mean	Min	Max	Variance	Mean	Min	Max	Variance
A_2	1	1	1	0	0.050	0	0.5	0.011
GOF	0.999	0.899	1.151	0.002	0.049	-0.143	1.129	0.020
G05	1.508	1.356	1.705	0.002	0.593	0.365	1.647	0.019
GMF*	0.005	0.00	0.011	0.000	-0.00	-0.000	0.006	0.000
GN	1.025	0.922	1.181	0.002	0.005	-0.149	1.159	0.021

Appendix 5. Variance component estimates obtained from ABLUP and HBLUP prediction methods for the complete simulated data using different number of genotyped individuals (300 genotyped/900 non-genotyped and 600 genotyped/600 non-genotyped) for the construction of **H** matrix (**rep2**)

	300 genotypes		**600 genotypes**	
Prediction Method	σ^2_a	σ^2_e	σ^2_a	σ^2_e
ABLUP	0.11	0.82	0.11	0.82
HBLUP$_{GOF}$	0.13	0.81	0.11	0.82
HBLUP$_{GOF\alpha}$	0.12	0.81	0.11	0.82
HBLUP$_{GO5}$	0.13	0.81	0.11	0.82
HBLUP$_{GN}$	0.13	0.81	0.30	0.77

*Model did not converge.

Appendix 6. Mean breeding values (EBVs), regression coefficients (intercept and slope) of TBVs on EBVs, mean squared error (MSE), bias (the difference between TBVs and EBVs), Pearson moment correlations (r_p) between TBVs and EBVs and Spearman rank correlations (r_s) between TBVs and EBVs for ABLUP and HBLUP prediction methods using simulated data used for the **H** matrix

a) Rep2 (1200 individuals in total, 300 genotyped and 900 non-genotyped individuals)

Method	**data**	**intercept**	**Slope**	**Bias**	**MSE**	**r_p**	**r_s**
ABLUP	all	0.60	1.07	0.03	0.095	0.58	0.55
	geno	0.60	1.04	0.03		0.56	0.51
	non-geno	0.60	1.08	0.03		0.59	0.56
$HBLUP_{GOF}$	all	0.60	0.98	0.03	0.093	0.60	0.56
	geno	0.60	0.88	0.03		0.62	0.56
	non-geno	0.60	1.03	0.03		0.59	0.55
$HBLUP_{GOF\alpha}$	all	0.60	0.98	0.03	0.092	0.60	0.56
	geno	0.60	0.88	0.03		0.62	0.56
	non-geno	0.60	1.03	0.03		0.59	0.55
$HBLUP_{GO5}$	all	0.61	0.98	0.03	0.093	0.60	0.56
	geno	0.60	0.88	0.03		0.62	0.56
	non-geno	0.60	1.03	0.03		0.59	0.55
$HBLUP_{GN}$	all	0.61	0.98	0.03	0.093	0.60	0.56
	geno	0.61	0.88	0.03		0.62	0.56
	non-geno	0.61	1.03	0.03		0.59	0.55

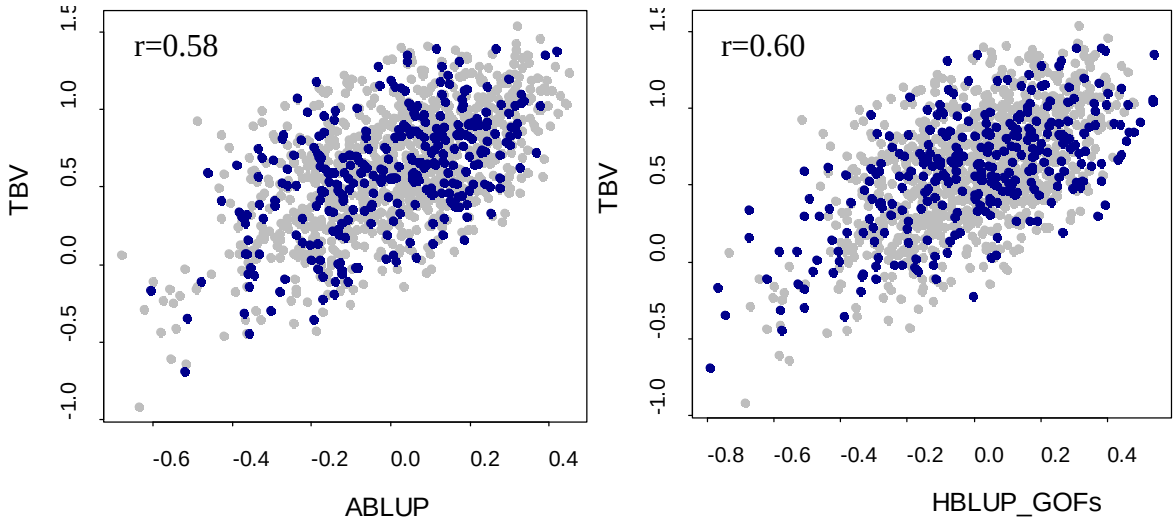
Appendix 6. Continued

b) Rep2= in total 1200 individuals, 600 genotyped and 600 non-genotyped individuals

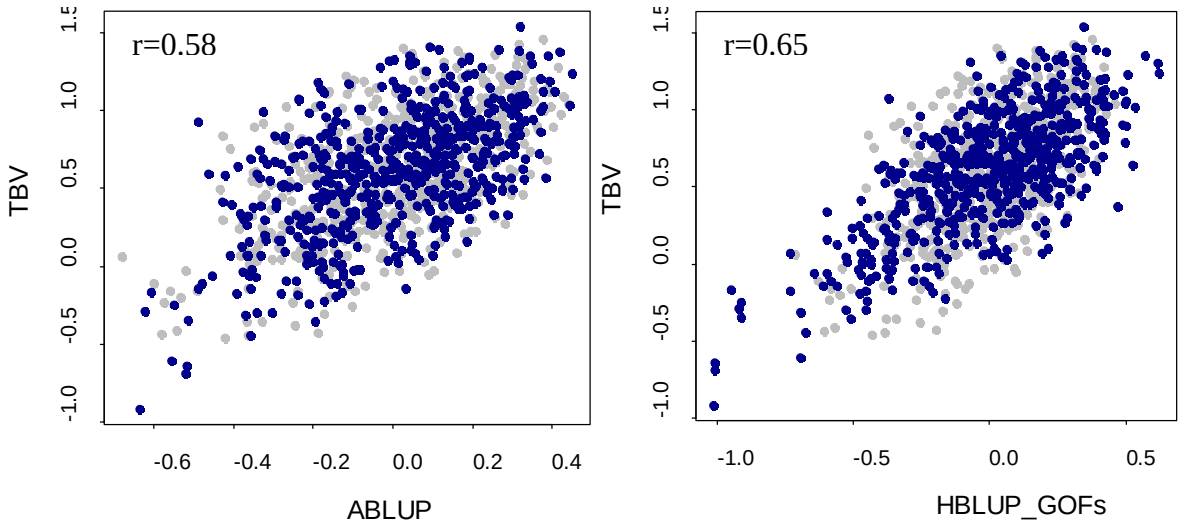
Method	data	intercept	Slope	Bias	MSE	r_p	r_s
ABLUP	all	0.60	1.07	0.03	0.095	0.58	0.55
	geno	0.61	1.06	0.04		0.58	0.53
	non-geno	0.59	1.08	0.02		0.59	0.56
HBLUP$_{GOF}$	all	0.60	1.01	0.03	0.082	0.65	0.61
	geno	0.61	0.98	0.04		0.71	0.65
	non-geno	0.60	1.07	0.02		0.59	0.56
HBLUP$_{GOF\alpha}$	all	0.61	1.01	0.03	0.082	0.65	0.61
	geno	0.61	0.99	0.04		0.71	0.65
	non-geno	0.61	1.07	0.02		0.59	0.56
HBLUP$_{GO5}$	all	0.61	1.02	0.03	0.083	0.65	0.60
	geno	0.61	0.98	0.04		0.71	0.65
	non-geno	0.60	1.07	0.02		0.59	0.56
HBLUP$_{GN}$	all	0.61	0.96	0.03	0.083	0.65	0.60
	geno	0.61	1.02	0.04		0.71	0.65
	non-geno	0.60	0.88	0.02		0.59	0.56

Appendix 7. Scatter plots between true breeding values (TBVs) and estimated breeding values (EBVs) obtained from ABLUP and HBLUP prediction methods using simulated data. Genotyped individuals are in dark blue and non-genotyped individuals are in gray (**rep2**).

a) When 300 genotyped and 900 non-genotyped individuals were used for the **H** matrix



b) When 600 genotyped and 600 non-genotyped individuals were used for the **H** matrix



Appendix 8. R script used for the construction of G matrices

```
# option
#    1= observed allele frequencies, GOF
#    2= normalized matrix having diagonal elements equal to 1,
GN
#    3= allele frequencies fixed at 0.5, GO5
#    4= allele frequencies fixed at mean for each locus, GMF
function(option,data,ped){
     options(warn=-1)
     library(MASS)
     library(GeneticsPed)
     library("corpcor")
     if(option==1){
          M1=data
          M= M1[,2:ncol(M1)]-1
          p1=round((apply(M,2,sum)+nrow(M))/(nrow(M)*2),3)
          p=2*(p1-.5)
          P = matrix(p,byrow=T,nrow=nrow(M),ncol=ncol(M))
          Z = as.matrix(M-P)
          b=1-p1
          c=p1*b
          d=2*(sum(c))
          ZZt = Z %*% t(Z)
          G = (ZZt/d)
}
if(option==2){
     M1=data
     M= M1[,2:ncol(M1)]-1
     p1=round((apply(M,2,sum)+nrow(M))/(nrow(M)*2),3)
     p=2*(p1-.5)
     P = matrix(p,byrow=T,nrow=nrow(M),ncol=ncol(M))
     Z = as.matrix(M-P)
     ZZt = Z %*% t(Z)
     tr=sum(diag(ZZt))/n    # n number of genotyped
individuals
     G = (ZZt/tr)
}
if(option==3){
     M1=data
     M= M1[,2:ncol(M1)]-1
     p1=array(0.5,ncol(M))
     p=p1
     P = matrix(p,byrow=T,nrow=nrow(M),ncol=ncol(M))
```

Appendix 8. Continued.

```
     Z = as.matrix(M-P)
     b=1-p1
     c=p1*b
     d=2*(sum(c))
     ZZt = Z %*% t(Z)
     G = (ZZt/d)
}
if(option==4){
     M1=data
     M= M1[,2:ncol(M1)]
     p1=round((apply(M,2,mean)),3)
     p=2*(p1-.5)
     P = matrix(p,byrow=T,nrow=nrow(M),ncol=ncol(M))
     Z = as.matrix(M-P)
     b=1-p1
     c=p1*b
     d=2*(sum(c))
     ZZt = Z %*% t(Z)
     G = (ZZt/d)
}
```

Appendix 9. R code for the construction of A matrix

```
ped <- read.table('~/pedigree.txt', header=F) #pedigree data
s <- ped[,2]  # sire (male) vector
d <- ped[,3]  # dam (female) vector

`createA` <-
function(s, d){
     if (nargs()==1){
          stop("sire vector and dam vector are required")
     }
     if (length(s) != length(d)){
          stop("size of the sire vector and dam vector are
different!")
     }
     n <- length(s)
     N <- n + 1
     A <- matrix(0, ncol=N, nrow=N)
     s <- (s == 0)*(N) + s
     d <- (d == 0)*N + d
     for(i in 1:n){
          A[i,i] <- 1 + A[s[i], d[i]]/2
          for(j in (i+1):n){
               if (j > n) break
               A[i,j] <- ( A[i, s[j]] + A[i,d[j]] )/2
               A[j,i] <- A[i,j]
          }
     }
          return(A[1:n, 1:n])
}
A <- createA(s,d)
A11 <- A[1:201,1:201]
A22 <- A[202:367,202:367]
A12 <- A[1:201,202:367]
A21 <- A[202:367,1:201]

# A, full pedigree for non-genotyped and genotyped individuals
# A11, subset of the A for non-genotyped individuals
# A22, subset of the A for genotyped individuals
# A12, covariance matrix between non-genotyped & genotyped
individuals
# A21, covariance matrix between genotyped & non-genotyped
individuals
```

Appendix 10. R code used for construction of H matrix

```
G=GenomicRel(option,data,ped) #returns genomic relationship

matrix (G)


Z=solve(A22)                #inverse of the A22
T=(G-A22)*1
A=A12%*%Z%*%T%*%Z%*%A21
B=A12%*%Z%*%T
C=T%*%Z%*%A21
D=T
E=G%*%Z%*%A21
p <- cbind(A,B)
o <- cbind(C,D)
J <- rbind(p,o)
H=A + J                # pedigree & genomic relationship matrix
(H)
Hinv <- solve(H)       # inverse of the H matrix

#appropriate format for the ASReml(sparse format)
     col1=NA
     col2=NA
     col3=NA
     print("Still working, this takes awhile")
     for (i in 1:nrow(Hinv)){
          for (j in 1:ncol(Hinv)){
               if (i >= j){
                    col1=cbind(col1,i)
                    col2=cbind(col2,j)
                    col3=cbind(col3,Hinv[i,j])
               }
          }
     }
invH=cbind(t(col1),t(col2),t(col3))
row.names(invH)=c(0:(nrow(invH)-1))

write.table(invH,"Hinv.giv",row.names=F,col.names=F,sep=" ")
```

Explanation of the H matrix code

GenomicRel(option,data,ped): creates **G** matrix using data and pedigree based on the option

Appendix 11. ASReml job files for ABLUP and HBLUP (actual data)

```
1       !ARGS 1 HT05 !RENAME 2
2       Title: ABLUP for the actual data.
3       #clone,female,male,cross,site,rep,HT05
4       #14,1,2,A,1,2,30.1,5
5       #14,1,2,A,1,6,29.8,5.1
6        clone   !P
7       #female  !P
8       #male    !P
9       #cross   !A
10       site    !I  !SKIP 3
11       rep     !I
12       HT05
13      ped.txt  !SORT                    # pedigree file (A)
14      data.csv !SKIP 1 !DOPART $A      # data file

15      !PART 1
16      $B~ mu site !r  clone ide(clone)
17      1 1 1
18      0 0 IDEN !S2=10

19      clone 1
20      clone 0 AINV 11 !GP
```

Explanation of the ASREML code for the actual data

```
1    !ARGS 1 changing parts to run. PART 1 is run for HT
(height)  trait. !RENAME 2 qualifier rename the output files
at two levels.
2    Title, description of the job.
3-5  Field definitions in the data
6    !P qualifier for the corresponding data field is
connected      pedigree
10   !I qualifier for the numeric but not integer factor.
!SKIP 3   qualifier, skips 3 data fields (female, male and
cross) before  reading this field
13   !SORT      the pedigree file
14   $A points the part to analyze
16   $B is the trait to analyze. ide(clone) fit the pedigree
     factor clone by replacing the genetic relationship matrix
(A)  with identity matrix to eliminate relationships between
copies of      a clone
```

Appendix 11. Continued.

17 is variance header line. 1 R structure for the error with 1 component and 1 G structure for the genetic effects
18 defines the dimension of R: number of data points, place holder for spatial analysis, IDEN (**R** is an identity matrix (IID) across all sites)
19 **G** structure header line involving 1 variance model
20 **AINV** relates clone effect with numerator relationship matrix

```
1       !ARGS 1  HT05 !RENAME 2
2       Title: HBLUP for the actual data.
3       #clone,female,male,cross,site,rep,HT05
4       #14,1,2,A,1,2,30.1,5
5       #14,1,2,A,1,6,29.8,5.1
6        clone   !P
7       #female  !P
8       #male    !P
9       #cross   !A
10       site    !I  !SKIP 3
11       rep     !I
12       HT05
13      ped.txt !SORT              # pedigree file (A)
14      Hinv.giv                  # inverse of the H matrix
15      data.csv !SKIP 1 !DOPART $A    # data file

16      !PART 1
17      $B ~ mu site !r clone ide(clone)
18      1 1 1
19      0 0 IDEN !S2=10

20      clone 1
21      clone 0 GIV 11 !GP
```

Explanation of the ASREML code for the actual data

1-20 Explanation of the HBLUP code is the same as the previous ABLUP code except explicit **H** inverse matrix
14 **GIV** relates the clone effect with $\mathbf{H}^{-1}$ matrix prepared explicitly

Appendix 12. The following codes for the ASReml job file were used for the data analysis of simulated data

```
!ARGS 1  x !RENAME 2
Title: ABLUP for the simulated data.
#indiv,par1,par2,x
#21,2,14,0.709177
#22,1,17,0.059232
 indiv  !P
 par1   !P
 par2   !P
 x
ped.txt !SORT                  # pedigree file (A)
data.csv !SKIP 1 !DOPART $A    # data file

!PART 1
$B~ mu   !r indiv
1 1 1
0 0 IDEN !S2=1

indiv 1
indiv 0 AINV 1 !GP

!ARGS 1  x !RENAME 2
Title: HBLUP for the simulated data.
#indiv,par1,par2,x
#21,2,14,0.709177
#22,1,17,0.059232
 indiv  !P
 par1   !P
 par2   !P
 x
ped.txt !SORT             # pedigree file (A)
Hinv.giv                  # inverse of the H matrix
data.csv !SKIP 1 !DOPART $A  # data file

!PART 1
$B~ mu !r indiv
1 1 1
0 0 IDEN !S2=1

indiv 1
indiv 0 GIV 1 !GP
```