Published March 28, 2014

RESEARCH

Genomic Relationship Matrix for Correcting Pedigree Errors in Breeding Populations: Impact on Genetic Parameters and Genomic Selection Accuracy

Patricio R. Munoz, Marcio F. R. Resende Jr., Dudley A. Huber, Tania Quesada, Marcos D. V. Resende, David B. Neale, Jill L. Wegrzyn, Matias Kirst, and Gary F. Peter*

ABSTRACT

Quantitative genetic analyses aim to estimate genetic parameters and breeding values to select superior parents, families, and individuals. For these estimates a relationship matrix derived from the pedigree typically is used in a mixed model framework. However, breeding is a complex, multistep process and errors in the pedigree are common. Because errors reduce the accuracy of genetic parameter estimates and affect genetic gain, it is important to correct these errors. Here we show that a realized relationship matrix (RRM) derived from single nucleotide polymorphism markers based on the normality of the relationship coefficients can be used to correct pedigree errors. For a loblolly pine (*Pinus taeda* L.) breeding population, errors in the pedigree were detected and corrected with the RRM. With the corrected pedigree, best linear unbiased predictor (BLUP) models fit the data significantly better for 14 out of 15 traits evaluated, and the predictive ability of the genomic selection models using ridge regression BLUP increased for 13 traits. The corrected pedigree based on the normality of the relationship coefficients improves accuracy of traditional estimations of heritability and breeding values as well as genomic selection predictions. As more breeding programs begin to use genomic selection, we recommend first using the dense panel of markers to correct pedigree errors and then using the improved information to develop genomic selection prediction models.

P.R. Munoz, Agronomy Dep., Univ. of Florida, P.O. Box 110965, Gainesville, FL 32611, USA; MF.R. Resende Jr., D.A. Huber, T. Quesada, M. Kirst, and G.F. Peter, School of Forest Resources and Conservation, 136 Newins-Ziegler Hall, Univ. of Florida, Gainesville, FL 32611; M.D.V. Resende, EMBRAPA Forestry and Dep. of Forest Engineering, Universidade Federal de Viçosa-UFV, Brazil; D.B. Neale and J.L Wegrzyn, Dep. of Plant Sciences, 262C Robbins Hall, One Shields Ave., Univ. of California, Davis, CA 95616. Received 1 Dec. 2012. *Corresponding author (gfpeter@ufl.edu).

Abbreviations: **A**, additive numerator relationship matrix; AIC, Akaike information criteria; BA, average branch angle; BD, average branch diameter; BLC, basal height of the live crown; BLUP, best linear unbiased predictor; BV, breeding value; CWAC, crown width across the planting beds; GS, genomic selection; MAF, minor allele frequency; RRM, realized relationship matrix; SNP, single nucleotide polymorphism; TAV, total additive value.

Genetic tests are designed to provide phenotypic information for estimation of genetic parameters such as variance components, heritability, genetic correlations, and breeding values. In breeding, this information is used for the selection of elite parents, families, and individuals for commercial production and subsequent generations of genetic improvement. For traits with complex inheritance, breeding values (BVs) are typically estimated with best linear unbiased prediction (BLUP) and used to rank the population for selection (Piepho et al., 2008). Best linear unbiased predictions are based on the theory of resemblance between relatives due to genetic factors (Lynch and Walsh, 1998), which are almost always derived from the pedigree (Mrode, 2005). Consequently, when the pedigree information is accurate better estimates of genetic parameters are obtained. Unfortunately, pedigree errors are common in

Published in Crop Sci. 53:1115–1123 (2013).
doi: 10.2135/cropsci2012.12.0673

breeding, averaging 10% in animal and tree breeding populations (Banos et al., 2001; Visscher et al., 2002; Doerksen and Herbinger, 2010). The presence of such errors can lead to incorrect estimates of the additive variance, causing a decrease in the BLUP-BV prediction accuracy (Ericsson, 1999; Banos et al., 2001; Sanders et al., 2006). In traditional BLUP-based selection, it has been reported that decreased BV accuracy reduces genetic gains by 4.3 to 17% (Geldermann et al., 1986; Israel and Weller, 2000).

To correct errors in the pedigree, molecular markers can be used. Most strategies rely on parent–progeny genotyping data (Bennewitz et al., 2002; Wiggans et al., 2010) or more recently in the diagonal of the realized relationship matrix (RRM) (Simeone et al., 2011). When dense panels of molecular markers are available they can be used to empirically estimate the actual relationships between relatives (Powell et al., 2010) and provide precise estimates of the proportion of the genome that is shared among individuals. If a dense panel of markers is used in breeding populations with a complex pedigree, the RRM values among individuals are normally distributed around the expectation for a given class (i.e., expectation [unrelated] = 0.0) (Yang et al., 2010; Simeone et al., 2011). Therefore, the current progeny population RRM diagonal and off-diagonal elements can be used to correct pedigree errors. This corrected pedigree should improve the accuracy of the BLUP-BV predictions and increase genetic gain.

Increasing the accuracy of BLUP-BV not only improves gains from traditional phenotypic selection but should also improve the accuracy of genomic selection models. Genomic selection (GS) models are developed to predict BV using only information from estimated marker effects (Meuwissen et al., 2001). Ideally, GS models should be fit with the best phenotypic values available and corrected for known environmental effects so the resultant value closely resembles the total additive value (TAV). This is because the goal of GS is to partition the TAV in pieces due to marker effects and then sum them under different genotype configurations (e.g., in a validation population or future generations) to estimate the genomic BV. To correct for such environmental effects, a realistic model encompassing, usually, fixed and random effects needs to be fit. This model splits the phenotype value into genetic (random) and environmental effects. At the same time, because breeding populations are used, where individuals are related, this model needs to correct for this known covariation with a correct pedigree relationship matrix (additive numerator relationship matrix [**A**] matrix). This approach properly corrects for any known environmental effects and generates an estimate of BVs (BLUP-BV) that are regressed based on an expected relationship value assuming an infinitesimal model (**A** matrix). Typically in GS prediction models, the BLUP-BVs are deregressed (Garrick et al., 2009) before regressing with the marker data, which is equivalent to single-step methodologies proposed by VanRaden (2008) and Misztal et al. (2009). The models are then tested in a validation population to obtain GS predicted BVs (GS-BV) and estimate the accuracy of genomic prediction (Goddard et al., 2009). The utility of GS in plant and animal breeding depends on the accuracy of the GS models developed to predict BV (Goddard and Hayes, 2009; Habier et al., 2010; Jannink et al., 2010; Grattapaglia and Resende, 2011; Heffner et al., 2010). Recently, a number of analytical approaches (Gianola et al., 2006; de los Campos et al., 2009; Habier et al., 2011) have been developed to study factors that contribute to GS accuracy (Habier et al., 2009, 2010; Iwata and Jannink, 2011) and to increase GS accuracy relative to the original approaches proposed by Meuwissen et al. (2001). Higher accuracy and less bias in the estimated BLUP-BVs are expected to improve the accuracy of all GS models. However, the effect of correcting pedigree errors on BLUP-BVs used to develop GS-BV prediction models has not been assessed.

Here we report for a loblolly pine (*Pinus taeda* L.) breeding population the effect of pedigree correction based on construction of a RRM from a dense panel of single nucleotide polymorphism (SNP) markers. The original and corrected pedigrees were used to generate BLUP-BVs and posteriorly GS models using ridge regression BLUP. The accuracies of the uncorrected and corrected pedigrees on BLUP-BV and GS-BV were compared.

MATERIALS AND METHODS

Data

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

Phenotype measurements were taken for basal height of the live crown (BLC) (cm), crown width across the planting beds (CWAC) (cm), crown width along the planting beds (cm), stem diameter at chest height (cm), and total stem height (cm), as described in Baltunis et al. (2007) and Resende et al. (2012b). The traits average branch angle (BA) (degrees), average branch diameter (BD) (cm), and BLC (year 6) were measured only in the high intensity silvicultural treatment. The age for each measurement is listed in Table 1, together with the trait–age combination used hereafter.

Genomic DNA was extracted from needle tissue using the QIAGEN DNeasy Plant Kit, and quantified with a NanoDrop microvolume spectrophotometer (Thermo Fisher Scientific Inc.). One microgram of DNA from each clone was genotyped using an Illumina Infinium assay (Illumina, Inc.) designed to

Table 1. Age of trait measurement and code trait–age combination.

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

†BA, average branch angle; BD, average branch diameter; BLC, basal height of the live crown; CWAC, crown width across the planting beds; CWAL, crown width along the planting beds; DBH, stem diameter at chest height; HT, total stem height.

detect 7216 SNPs that were identified through the resequencing of 7535 uniquely expressed sequence tag contigs in 18 loblolly pine haploid megagametophytes (Eckert et al., 2010). After filtering for monomorphic markers a total of 4825 SNPs were selected for analysis.

Realized Relationship Matrix and Pedigree Corrections

Molecular markers were preselected in a previous study (Quesada et al., 2010; Quesada, 2010) based on the quality and reliability of the called genotypes using BeadStudio version 3.1.3.0 software (Illumina, Inc.) as well as frequency of polymorphism across genotypes yielding a set of 2182 SNP markers. This subset of SNPs has a minor allele frequency (MAF) of >0.12, similar to the 0.10 successfully used in the barley (*Hordeum vulgare* L.) study of Zhong et al. (2009). The realized relationship for each pair of individuals was calculated as the sum of the products of SNP coefficients between two individuals scaled by SNP heterozygosity as described in Powell et al. (2010). Relatedness estimates were adjusted for sampling error and shrunk toward the expected values to lessen error as recommended by Yang et al. (2010). Using relationships estimated in the RRM, the pedigree was corrected based on the normality of the distribution of the relationship coefficients around their expected values (i.e., 0.5 for full-sib). First, the RRM was paired with the numerator relationship matrix derived from the pedigree (**A**). Second, duplicated individuals (different label but same genotype) were identified, and ones with fewer missing values were kept. Third, the relationship coefficient limits for the full-sib and half-sib classes were defined based on the normal distribution using all relationships in each class. Fourth, individual or groups of individuals not matching the expected pattern were identified. Fifth, conflictive individuals were reassigned by searching across all relationships in the dataset for the parent or family where these individuals match the expectation. In this last step, an individual was reassigned to a new parent or family only if the conflictive individual matched the expectation, given by the defined boundaries, with all individuals from the new parent or family. Once the new parent or family was identified, the individuals were relabeled generating the corrected pedigree. This process was iterative, as every time the pedigree of an individual was corrected the relationship class distributions changed across the database and were recalculated.

Best Linear Unbiased Predictor Analysis: Variance Component Estimation and Breeding Values Prediction

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

Original Pedigree Best Linear Unbiased Predictor

This model assumes no errors in the original pedigree:

$$y = \mathbf{Xb} + \mathbf{Z}_1\mathbf{i} + \mathbf{Z}_2\mathbf{a} + \mathbf{Z}_3\mathbf{f} + \mathbf{Z}_4\mathbf{n} + \mathbf{Z}_5\mathbf{d}_1 + \mathbf{Z}_6\mathbf{d}_2 + \mathbf{Z}_7\mathbf{d}_3 + e, \quad [1]$$

in which y is the measure of the trait being analyzed (see above), **b** is a vector of fixed effects (i.e., culture type and replication within culture type), **i** is a vector of random incomplete block effect within replication ~N(0, $\mathbf{I}\sigma^2_{iblk}$), **a** is a vector of random additive effects of clones ~N(0, $\mathbf{A}\sigma^2_a$) that corresponds to the general combining ability, **f** is a vector of random family effect ~N(0, $\mathbf{I}\sigma^2_f$) that corresponds to the specific combining ability, **n** is a vector of random nonadditive effects of clones ~N(0, $\mathbf{I}\sigma^2_n$) that corresponds to the remainder of the genetic effects, $\mathbf{d}_1$ is a vector of random additive by culture type interaction ~N(0, $\mathbf{DIAG}\sigma^2_{d1}$), $\mathbf{d}_2$ is a vector of random family by culture type interaction ~N(0, $\mathbf{DIAG}\sigma^2_{d2}$), $\mathbf{d}_3$ is a vector of random nonadditive by culture type interaction ~N(0, $\mathbf{DIAG}\sigma^2_{d3}$), e is the random residual effect ~N(0, $\mathbf{DIAG}\sigma^2_e$) as one specific error for each treatment was fitted, **X** and $\mathbf{Z}_1$ through $\mathbf{Z}_7$ are incidence matrices, and **I**, **A**, and **DIAG** are the identity, numerator relationship, and block diagonal matrices, respectively.

Corrected Pedigree Best Linear Unbiased Predictor

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

Genomic Selection and Validation

For GS analysis, the BV estimated in each of the above models was deregressed and the parental average of each family removed (Garrick et al., 2009). The deregressed phenotypes obtained with the original and the corrected pedigrees were used as input for a ridge regression BLUP with the 4825 markers used as covariates as described previously (Resende et al., 2012b). Each analysis was repeated 10 times in a cross-validation scheme (Kohavi, 1995). The predictive ability of each model was estimated as the correlation between the GS predicted BVs (GS-BVs) and the deregressed phenotype that were used as input in the generation of the GS-BVs.

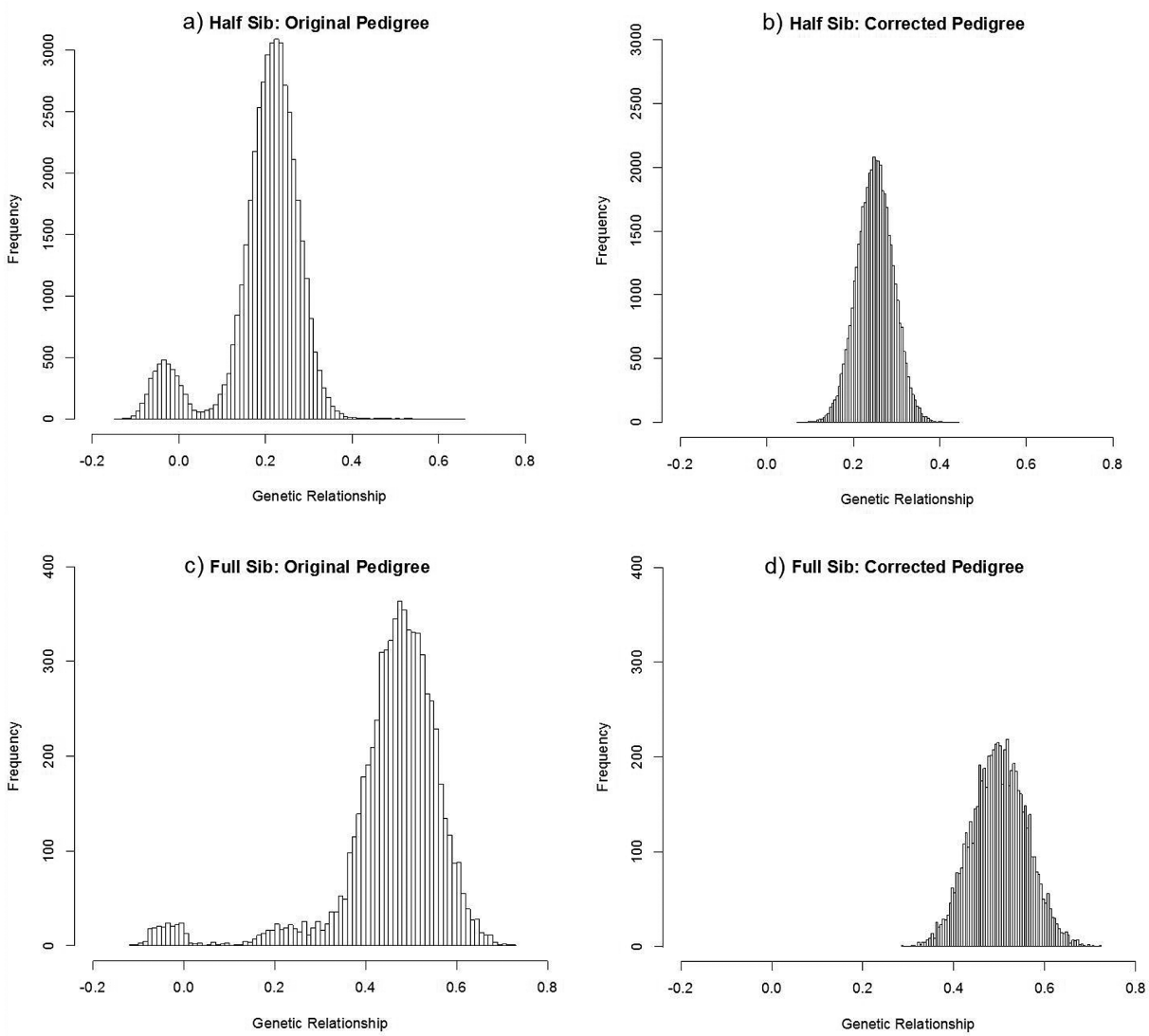

Figure 1. Distribution of relationship values for half-sib (upper panels) and full-sib (lower panel) individuals around their expected means 0.25 and 0.5, respectively. Distribution for the original pedigree (left panels) and corrected (right panels).

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

Relationship class	Expected relationship coefficient	Original pedigree		Corrected pedigree	
		Mean	Standard deviation	Mean	Standard deviation
Unrelated	0.0000	−0.0382	0.044	−0.0005	0.015
Half-sibs	0.2500	0.1974	0.089	0.2500	0.042
Full-sibs	0.5000	0.4563	0.121	0.5001	0.061
Self	1.0000	1.0121	0.055	0.9997	0.040

RESULTS

Pedigree Correction

The relationship coefficients derived from the molecular markers is expected to be a normal distribution centered at 0.5 and 0.25 for full- and half-sib families, respectively. With our original pedigree, a bimodal and asymmetrical distribution was observed for half-sibs, with the largest frequency close to the expected 0.25 value and a second peak close to zero (Fig. 1, top left panel). For the full-sib class a trimodal asymmetrical distribution was observed: the highest peak (mode) around the 0.5 expectation value, with the second and third peaks around the 0.25 and a zero relationship, respectively (Fig. 1, bottom left panel). In the original pedigree before corrections, the most frequent relationship found in the dataset yielded biased average relationship coefficients (Table 2), with unrelated, half-sibs, and full-sibs individuals being underestimated and the diagonal of the matrix being slightly overestimated. The standard deviations for full-sib and half-sib individuals were the largest (Table 2). However, correcting the

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

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

pedigree gave mean values that agreed with the expectations for the given classes, decreasing the standard deviation by 27 to 67% (Table 2; Fig. 1, right panels).

By using the RRM, different types of pedigree errors were detected and corrected, including duplicated genotypes (clones) with different labels, from which only one was kept. Individuals with either one or both incorrect parents (69 in total) were reassigned to the correct parent using the coefficients from the RRM. Eleven new parents, one female and 10 male, were added, as they did not exist in the pedigree records. Parents of four complete families and two grandparents were reassigned. Finally, three individuals were removed because they yielded inconsistent relationships across the pedigree (Table 3).

Estimation of Breeding Values with Original and Corrected Pedigree Relationship Matrices

Genetic parameters were estimated with both the original and corrected pedigrees. Compared with the original pedigree, heritability estimates derived from a traditional BLUP analysis using the corrected pedigree decreased slightly for eight of the traits (maximum decrease of 5%) and increased for seven traits by a maximum of 21% for average branch diameter measured at age 6 (BD_6) (Table 4). With the corrected pedigree, BLUP BV accuracy decreased slightly in only four traits, with a maximum reduction of 0.94% for average branch angle measured at age 6 (BA_6), and increased for 11 traits (maximum increase of 5.8% for total stem height measured at age 6 [HT_6]). Importantly, in all but one trait (basal height of the live crown measured at age 6 [BLC_6]) the models with the corrected pedigree fit the data substantially better, measured by the Akaike information criteria (AIC) (Table 4).

Accuracy of Genomic Selection Predictive Models with Original and Corrected Pedigrees

Breeding values obtained from BLUP analyses with the original and corrected pedigrees were posteriorly deregressed and used as response variables to generate GS models for the 15 trait–age combinations. The predictive ability of the GS models increased for 13 of the 15 traits when the corrected version of the pedigree was used (Fig. 2). For the two traits where accuracies decreased with the corrected pedigree [BA_6 and crown width across planting bed measured at age 6 (CWAC_6)] they were reduced by 1.1 and 2.3%, respectively, whereas the predictive ability of the remaining 13 traits increased from 1 to 15% with an average of 7.2%.

Table 4. Narrow-sense heritability (h^2), accuracy of breeding values [Acc(BV)] and fitting of models (maximum of log(likelihood) [Log L]), and Akaike information criteria (AIC) by traditional best linear unbiased predictor (BLUP) analysis using a full genetic model with original pedigree or using a full genetic model with corrected pedigree (from Eq. [1]) on 15 trait–age combination.

	Full original pedigree BLUP				Full corrected pedigree BLUP			
Trait†	h^2	Acc(BV)	Log L	AIC	h^2	Acc(BV)	Log L	AIC
BA_6	0.33 (0.08)‡	0.82	−9,056.1	18,122.3	0.33 (0.08)	0.81	−9,015.3	18,040.7
BD_6	0.12 (0.04)	0.69	−3,920.6	7,831.2	0.15 (0.05)	0.72	−3,847.5	7,685.0
BLC_4	0.19 (0.06)	0.78	−8,167.9	16,353.8	0.22 (0.02)	0.81	−8,044.6	16,107.1
BLC_6	0.31 (0.08)	0.79	−3,842.4	7,674.8	0.35 (0.03)	0.82	−3,731.1	7,452.2
CWAC_2	0.23 (0.02)	0.82	−5,355.0	10,728.0	0.22 (0.02)	0.82	−5,251.8	10,521.6
CWAC_6	0.43 (0.10)	0.85	−4,898.4	9,806.8	0.45 (0.02)	0.85	−4,834.6	9,679.1
CWAL_2	0.21 (0.02)	0.82	−4,779.5	9,577.0	0.21 (0.02)	0.82	−4,673.5	9,365.0
CWAL_6	0.27 (0.08)	0.79	−3,898.8	7,807.6	0.27 (0.03)	0.79	−3,838.3	7,686.7
DBH_3	0.27 (0.02)	0.83	−4,304.4	8,626.9	0.26 (0.02)	0.83	−4,292.3	8,602.6
DBH_4	0.28 (0.02)	0.83	−6,165.2	12,348.5	0.27 (0.02)	0.83	−6,146.8	12,311.6
DBH_6	0.32 (0.02)	0.85	−7,996.2	16,010.3	0.31 (0.02)	0.85	−7,971.0	15,959.9
HT_1	0.11 (0.03)	0.75	−3,727.4	7,472.8	0.12 (0.03)	0.77	−3,622.3	7,262.6
HT_2	0.27 (0.02)	0.82	−29,071.5	58,160.9	0.27 (0.02)	0.84	−28,950.9	57,919.7
HT_3	0.28 (0.08)	0.83	−2,593.0	5,203.9	0.27 (0.02)	0.84	−2,456.6	4,931.2
HT_6	0.26 (0.07)	0.80	−5,091.4	10,194.8	0.31 (0.02)	0.85	−4,944.6	9,901.1

†BA_6, average branch angle measured at age 6; BD_6, average branch diameter measured at age 6; BLC_4, basal height of the live crown measured at age 4; BLC_6, basal height of the live crown measured at age 6; CWAC_2, crown width across the planting beds measured at age 2; CWAC_6, crown width across the planting beds measured at age 6; CWAL_2, crown width along the planting beds measured at age 2; CWAL_6, crown width along the planting beds measured at age 6; DBH_3, stem diameter at chest height measured at age 3; DBH_4, stem diameter at chest height measured at age 4; DBH_6, stem diameter at chest height measured at age 6; HT_1, total stem height measured at age 1; HT_2, total stem height measured at age 2; HT_3, total stem height measured at age 3; HT_6, total stem height measured at age 6.

‡Standard error for the heritability appears in parentheses.

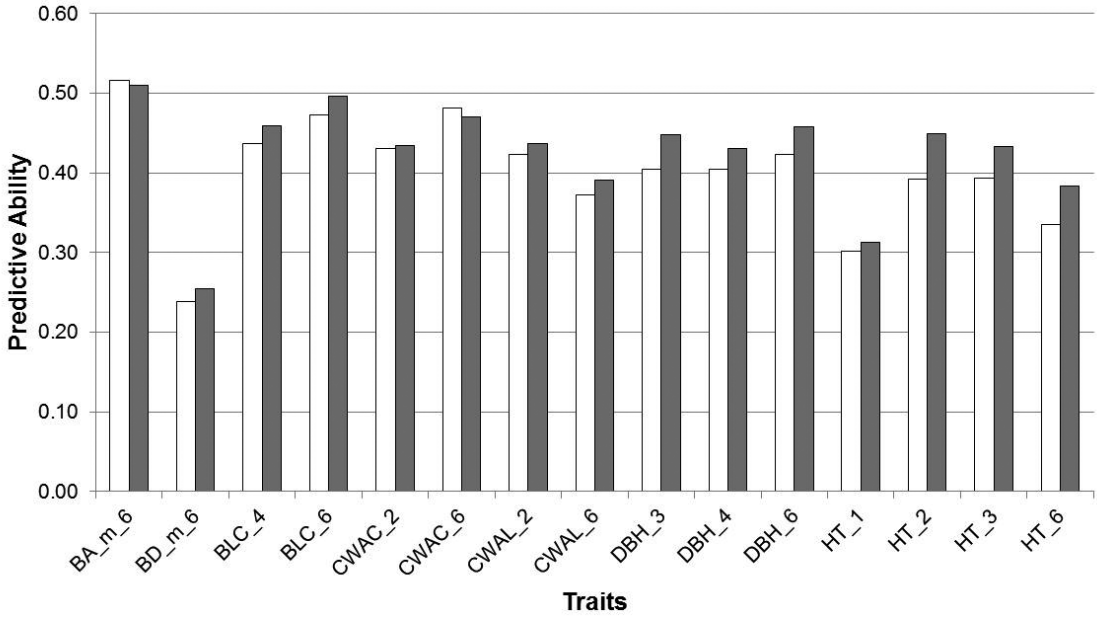

Figure 2. Predictive ability for fifteen different traits using the original pedigree derived from historical records (white column) and the corrected version of the pedigree (grey column). BA_m_6, average branch angle measured at age 6; BD_m_6, average branch diameter measured at age 6; BLC_4, basal height of the live crown measured at age 4; BLC_6, basal height of the live crown measured at age 6; CWAC_2, crown width across the planting beds measured at age 2; CWAC_6, crown width across the planting beds measured at age 6; CWAL_2, crown width along the planting beds measured at age 2; CWAL_6, crown width along the planting beds measured at age 6; DBH_3, stem diameter at chest height measured at age 3; DBH_4, stem diameter at chest height measured at age 4; DBH_6, stem diameter at chest height measured at age 6; HT_1, total stem height measured at age 1; HT_2, total stem height measured at age 2; HT_3, total stem height measured at age 3; HT_6, total stem height measured at age 6.

DISCUSSION

Pedigree Correction

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

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

To correct errors, a RRM (Powell et al., 2010) is constructed for the breeding population. The use of markers with a MAF > 0.12 to construct the relationship matrix should not affect the properties of the matrix; as pointed out by Chen et al. (2011), markers selected with MAF between 0 and 0.20 do not affect either the matrix parameters or the prediction accuracy when current frequency is used as was the case of this study. In the relationship matrix, a normal symmetric and unimodal distribution for each relationship class (i.e., unrelated, half-sib, or full-sib) is expected because of Mendelian sampling (Simeone et al., 2011). This has been observed with 294,831 SNPs markers on 3925 individual humans with a standard deviation between 0.004 and 0.005 (Yang et al., 2010). As more markers are added, more precise estimations of the Mendelian sampling will be obtained and, thus, smaller standard deviations are observed (Hayes et al., 2009). In our case, we detected a bimodal asymmetrical distribution for half-sibs, indicating problems in the recorded pedigree and showing a bias for the mean relationship (see Fig. 1). The additional peak observed in the distribution centered on zero indicated that unrelated individuals were misclassified as half-sibs. After reassignment of individuals and correction of the pedigree, the expected

normal distribution was observed as well as a considerable decrease in the standard deviation. This also was the case for the full-sib relationship and other relationship classes in the population. Although a large decrease in the standard error was obtained, our estimations are still high compared with those obtained by Yang et al. (2010) or Simeone et al. (2011), probably due to the reduced number of SNPs markers (approximately 2300) genotyped on a smaller population (approximately 860 individuals) with many different relationship classes derived from the circular mating design (i.e., unrelated, half-sibs, full-sibs, etc.). Better estimations are expected as more markers and individuals are added in future studies.

The extended length of a pine (*Pinus* spp.) breeding cycle and their reproductive biology contribute to a high likelihood of pedigree errors. Pines are wind pollinated and pollen from foreign genotypes is commonly present during controlled pollination. Similarly, the length of the breeding cycle implies that record keeping is prone to include errors as many people are involved across the long period (White et al., 2007). Most errors can be corrected by reassigning individuals, parents, or families present in the known pedigree although the necessity of adding new parents indicates pollen contamination (Adams et al., 1988). In our case, three individuals were dropped from further analysis as they yielded inconsistent relationships. The inconsistent relationships of these three individuals were due to large amounts of missing SNP data, indicating genotyping problems.

Estimation of Breeding Value with Original and Corrected Relationship Matrices

Independently of the stage when the errors originated, our results show that pedigree errors decrease the accuracy of the BLUP-BV prediction, as previously reported in pines and dairy cattle (*Bos taurus*) (Ericsson, 1999; Banos et al., 2001; Sanders et al., 2006). In addition to improved BLUP-BV accuracy, using the corrected instead of the original pedigree dramatically increased the fit of the data (Table 4) and showed that the heritability was slightly overestimated in eight traits and underestimated in seven with the original pedigree (Table 4). The impact of correcting the pedigree on the BLUP analysis not only depends on the number of errors but also on how much difference existed between the phenotypic value of the individual and the average of the family where the individual was incorrectly assigned. This happens because the traditional BLUP analysis shrinks the individual records towards the parental average of the family defined in the **A** matrix. When the phenotype of the mislabeled individual is similar to the family average in which this individual was misassigned, the estimated BV will be less biased than in a situation where the difference between the phenotypic value and the average of the family is large. However, even in these less biased cases, there are some practical considerations regarding inbreeding and selection. If the best performing individuals are mislabeled, then related individuals may be selected inadvertently or, conversely, selection of superior unrelated individuals may be avoided because they are labeled as the same family. Both cases will impact the potential genetic gain, the first through inbreeding depression and the second in the loss of opportunity to select one of the best individuals. In addition, as pointed by Goddard et al. (2011) and Meuwissen et al. (2011) a pedigree-derived relationship matrix will be still needed even when using the RRM as proposed by Misztal et al. (2009), to provide unbiased predictions.

Accuracy of Genomic Selection Predictive Models with Original and Corrected Pedigrees

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

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

In conclusion, pedigree errors are a common concern among breeders because of their detrimental effect on parameter estimates and reduction in short- and long-term genetic gain. In this work, using a *P. taeda* breeding population as a model, we demonstrate that pedigree errors can weaken the accuracy of traditional estimations (i.e., BLUP) and genomic selection predictions. Because errors in the breeding population are cumulative, as a wrong individual may be used as a parent in the next generation, this can

compromise the long-term breeding strategy. Additionally, we showed that estimation of a genomic relationship matrix can be used to correct such errors based on the normality of the different relationship coefficients. While all individuals are connected to each other in the RRM, using only the most frequent relationship in the matrix of complex pedigrees (full-sib and half-sib in our case) will ensure population-wide pedigree correction that includes all individuals and relationships. Furthermore, this methodology has the advantage that no molecular markers from parents are needed and is a byproduct of information needed (i.e., dense panel of markers) to perform genomic selection. As many breeding programs (annual and perennial) are beginning to test genomic selection, our methodology can readily be applied to these new pedigrees. The utility of the proposed method needs to be investigated under deeper pedigrees (i.e., several generations), where the higher levels of relationships among genotypes create continuous relationship coefficients that may be difficult to separate one relationship class from another. As more breeding programs begin to use genomic selection, we recommend first using the dense panel of markers to correct pedigree errors. The corrected pedigree and markers should only then be applied for developing genomic selection prediction models, thus taking additional advantage of the genotyping investment needed to perform genomic selection.

Acknowledgments

The authors wish to thank members of the Forest Biology Research Cooperative (FBRC) for their support in establishing, maintaining, and measuring field trials used in this study. The work was supported by the National Science Foundation Plant Genome Research Program (award no. 0501763) and the U.S. Department of Agriculture National Institute of Food and Agriculture Plant Breeding and Education Program (award no. 2010-85117-20569). Thanks to Salvador Gezan for his comments and discussion.

References

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

Baltunis, B.S., D.A. Huber, T.L. White, B. Goldfarb, and H.E. Stelzer. 2005. Genetic effects of rooting loblolly pine stem cuttings from a partial diallel mating design. Can. J. For. Res. 35:1098–1108. doi:10.1139/x05-038

Baltunis, B., D. Huber, T. White, B. Goldfarb, and H. Stelzer. 2007. Genetic analysis of early field growth of loblolly pine clones and seedlings from the same full-sib families. Can. J. For. Res. 37:195–205. doi:10.1139/x06-203

Banos, G., G.R. Wiggans, and R.L. Powell. 2001. Impact of paternity errors in cow identification on genetic evaluations and international comparisons. J. Dairy Sci. 84:2523–2529. doi:10.3168/jds.S0022-0302(01)74703-0

Bennewitz, J., N. Reinsch, and E. Kalm. 2002. Gencheck: A program for consistency checking and derivation of genotypes at co-dominant and dominant loci. J. Anim. Breed. Genet. 119:350–360. doi:10.1046/j.1439-0388.2002.00357.x

Chen, C.Y., I. Misztal, I. Aguilar, A. Legarra, and W.M. Muir. 2011. Effect of different genomic relationship matrices on accuracy and scale. J. Anim. Sci. 89:2673–2679. doi:10.2527/jas.2010-3555

de los Campos, G., D. Gianola, and G.J.M. Rosa. 2009. Reproducing kernel Hilbert spaces regression: A general framework for genetic evaluation. J. Anim. Sci. 87:1883–1887. doi:10.2527/jas.2008-1259

Doerksen, T., and C. Herbinger. 2010. Impact of reconstructed pedigrees on progeny-test breeding values in red spruce. Tree Genet. Genomes 6:591–600. doi:10.1007/s11295-010-0274-1

Eckert, A.J., J. Van Heerwaarden, J.L. Wegrzyn, C.D. Nelson, J. Ross-Ibarra, S.C. González-Martínez, and D.B. Neale. 2010. Patterns of population structure and environmental associations to aridity across the range of loblolly pine (*Pinus taeda* L., Pinaceae). Genetics 185:969–982. doi:10.1534/genetics.110.115543

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

Garrick, D.J., J.F. Taylor, and R.L. Fernando. 2009. Deregressing estimated breeding values and weighting information for genomic regression analyses. Genet. Sel. Evol. 41:55. doi:10.1186/1297-9686-41-55

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

Gianola, D., R.L. Fernando, and A. Stella. 2006. Genomic-assisted prediction of genetic value with semiparametric procedures. Genetics 173:1761–1776. doi:10.1534/genetics.105.049510

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

Goddard, M.E., and B.J. Hayes. 2009. Mapping genes for complex traits in domestic animals and their use in breeding programs. Nat. Rev. Genet. 10:381–391. doi:10.1038/nrg2575

Goddard, M.E., B.J. Hayes, and T.H.E. Meuwissen. 2011. Using the genomic relationship matrix to predict the accuracy of genomic selection. J. Anim. Breed. Genet. 128:409–421. doi:10.1111/j.1439-0388.2011.00964.x

Goddard, M.E., N.R. Wray, K. Verbyla, and P.M. Visscher. 2009. Estimating effects and making predictions from genome-wide marker data. Stat. Sci. 24:517–529. doi:10.1214/09-STS306

Grattapaglia, D., and M. Resende. 2011. Genomic selection in forest tree breeding. Tree Genet. Genomes 7:241–255. doi:10.1007/s11295-010-0328-4

Habier, D., R.L. Fernando, and J.C.M. Dekkers. 2009. Genomic selection using low-density marker panels. Genetics 182:343–353. doi:10.1534/genetics.108.100289

Habier, D., R.L. Fernando, K. Kizilkaya, and D.J. Garrick. 2011. Extension of the Bayesian alphabet for genomic selection. BMC Bioinf. 12:186. doi:10.1186/1471-2105-12-186

Habier, D., J. Tetens, F.R. Seefried, P. Lichtner, and G. Thaller. 2010. The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genet. Sel. Evol. 42:5. doi:10.1186/1297-9686-42-5

Hayes, B.J., P.J. Bowman, A.C. Chamberlain, K. Verbyla, and M.E. Goddard. 2009. Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genet. Sel. Evol. 41:51. doi:10.1186/1297-9686-41-51

Heffner, E.L., A.J. Lorenz, J.L. Jannink, and M.E. Sorrells.

2010. Plant breeding with genomic selection: Gain per unit time and cost. Crop Sci. 50:1681–1690. doi:10.2135/cropsci2009.11.0662

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

Israel, C., and J.I. Weller. 2000. Effect of misidentification on genetic gain and estimation of breeding value in dairy cattle populations. J. Dairy Sci. 83:181–187. doi:10.3168/jds.S0022-0302(00)74869-7

Iwata, H., and J.L. Jannink. 2011. Accuracy of genomic selection prediction in barley breeding programs: A simulation study based on the real single nucleotide polymorphism data of barley breeding lines. Crop Sci. 51:1915–1927. doi:10.2135/cropsci2010.12.0732

Jannink, J.L., A.J. Lorenz, and H. Iwata. 2010. Genomic selection in plant breeding: From theory to practice. Brief. Funct. Genomics 9:166–177. doi:10.1093/bfgp/elq001

Kohavi, R. 1995. The power of decision tables. Mach. Learn.: ECML 95(912):174–189. doi:10.1007/3-540-59286-5_57

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

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

Meuwissen, T.H., T. Luan, and J.A. Woolliams. 2011. The unified approach to the use of genomic and pedigree information in genomic evaluations revisited. J. Anim. Breed. Genet. 128:429–439. doi:10.1111/j.1439-0388.2011.00966.x

Misztal, I., A. Legarra, and I. Aguilar. 2009. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. J. Dairy Sci. 92:4648–4655. doi:10.3168/jds.2009-2064

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

Piepho, H.P., J. Mohring, A.E. Melchinger, and A. Buchse. 2008. BLUP for phenotypic selection in plant breeding and variety testing. Euphytica 161:209–228. doi:10.1007/s10681-007-9449-8

Powell, J.E., P.M. Visscher, and M.E. Goddard. 2010. Reconciling the analysis of IBD and IBS in complex trait studies. Nat. Rev. Genet. 11:800–805. doi:10.1038/nrg2865

Quesada, T. 2010. Association genetics of pitch canker resistance in loblolly pine (*Pinus taeda* L.). (publication no. 3447054.) PhD diss., University of Florida, Gainesville FL.

Quesada, T., V. Gopal, W.P. Cumbie, A.J. Eckert, J.L. Wegrzyn, D.B. Neale, et al. 2010. Association mapping of quantitative disease resistance in a natural population of loblolly pine (*Pinus taeda* L.). Genetics 186:677–686. doi:10.1534/genetics.110.117549

Resende, M.F.R., Jr., P. Munoz, J.J. Acosta, G.F. Peter, J.M. Davis, D. Grattapaglia, et al. 2012a. Accelerating the domestication of trees using genomic selection: Accuracy of prediction models across ages and environments. New Phytol. 193:1099–1099. doi:10.1111/j.1469-8137.2011.03895.x

Resende, M.F.R., Jr., P. Munoz, M.D.V. Resende, D.J. Garrick, R.L. Fernando, J.M. Davis, et al. 2012b. Accuracy of genomic selection methods in a standard data set of loblolly pine (*Pinus taeda* L.). Genetics 190:1503–1510. doi:10.1534/genetics.111.137026

Sanders, K., J. Bennewitz, and E. Kalm. 2006. Wrong and missing sire information affects genetic gain in the Angeln dairy cattle population. J. Dairy Sci. 89:7.

Simeone, R., I. Misztal, I. Aguilar, and A. Legarra. 2011. Evaluation of the utility of diagonal elements of the genomic relationship matrix as a diagnostic tool to detect mislabelled genotyped animals in a broiler chicken population. J. Anim. Breed. Genet. 128:386–393. doi:10.1111/j.1439-0388.2011.00926.x

VanRaden, P.M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414–4423. doi:10.3168/jds.2007-0980

Visscher, P.M., J.A. Woolliams, D. Smith, and J.L. Williams. 2002. Estimation of pedigree errors in the UK dairy population using microsatellite markers and the impact on selection. J. Dairy Sci. 85:2368–2375. doi:10.3168/jds.S0022-0302(02)74317-8

White, T.L., W.T. Adams, and D.B. Neale. 2007. Forest genetics. CABI Publishing, Wallingford, UK.

Wiggans, G.R., P.M. VanRaden, L.R. Bacheller, M.E. Tooker, J.L. Hutchison, T.A. Cooper, and T.S. Sonstegard. 2010. Selection and management of DNA markers for use in genomic evaluation. J. Dairy Sci. 93:2287–2292. doi:10.3168/jds.2009-2773

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

Yang, J., B. Benyamin, B.P. Mcevoy, S. Gordon, A.K. Henders, D.R. Nyholt, et al. 2010. Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. 42:565–569. doi:10.1038/ng.608

Zhong, S., J.C.M. Dekkers, R.L. Fernando, and J.-L. Jannink. 2009. Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: A barley case study. Genetics 182:355–364. doi:10.1534/genetics.108.098277