American Genetic Association

Journal of Heredity, 2018, 802–810
doi:10.1093/jhered/esy051
Original Article
Advance Access publication October 4, 2018

OXFORD

Original Article

Effect of Hidden Relatedness on Single-Step Genetic Evaluation in an Advanced Open-Pollinated Breeding Program

Jaroslav Klápště, Mari Suontama, Heidi S. Dungey, Emily J. Telfer, Natalie J. Graham, Charlie B. Low, and Grahame T. Stovold

From the Scion (New Zealand Forest Research Institute Ltd.), 49 Sala Street, Whakarewarewa, Rotorua 3010, New Zealand. Mari Suontama is now at Skogforsk, Umeå, Box 3, Sävar SE-918 21, Sweden.

Address correspondence to J. Klápště at the address above, or e-mail: jaroslav.klapste@scionresearch.com.

Received June 11, 2018; First decision August 20, 2018; Accepted September 27, 2018.

Corresponding Editor: John Stommel

Abstract

Open-pollinated (OP) mating is frequently used in forest tree breeding due to the relative temporal and financial efficiency of the approach. The trade-off is the lower precision of the estimated genetic parameters. Pedigree/sib-ship reconstruction has been proven as a tool to correct and complete pedigree information and to improve the precision of genetic parameter estimates. Our study analyzed an advanced generation *Eucalyptus* population from an OP breeding program using single-step genetic evaluation. The relationship matrix inferred from sib-ship reconstruction was used to rescale the marker-based relationship matrix (***G*** matrix). This was compared with a second scenario that used rescaling based on the documented pedigree. The proposed single-step model performed better with respect to both model fit and the theoretical accuracy of breeding values. We found that the prediction accuracy was superior when using the pedigree information only when compared with using a combination of the pedigree and genomic information. This pattern appeared to be mainly a result of accumulated unrecognized relatedness over several breeding cycles, resulting in breeding values being shrunk toward the population mean. Using biased, pedigree-based breeding values as the base with which to correlate predicted GEBVs, resulted in the underestimation of prediction accuracies. Using breeding values estimated on the basis of sib-ship reconstruction resulted in increased prediction accuracies of the genotyped individuals. Therefore, selection of the correct base for estimation of prediction accuracy is critical. The beneficial impact of sib-ship reconstruction using ***G*** matrix rescaling was profound, especially in traits with inbreeding depression, such as stem diameter.

Keywords: *Eucalyptus nitens*, hidden relatedness, inbreeding depression, sib-ship reconstruction, single-step evaluation

Precise estimation of genetic parameters is essential to perform an accurate selection of genetically superior individuals and best practice management of genetic diversity in operational breeding programs. To achieve these goals, pedigrees that are both error-free and complete across generations should be established. Documenting and maintaining complete pedigrees in forest tree breeding is time-consuming and labor-intensive. In many cases achieving crossing, designs are technically challenging due to biological constraints, differential temporal sexual maturation or the differential size of reproduction organs physically preventing a successful cross (Potts and Dungey 2004). Costs of tracking parents mean that progeny tests based on open-pollinated (OP) mating are preferred

(Burdon and Shelbourne 1971). OP strategies can not fully track pedigree and so do suffer from the presence of hidden relatedness, the proportion of which is affected by conditions under which reproduction was performed (i.e., wild stands vs. breeding arboretum vs. polymix breeding). Hidden relatedness can affect the accuracy of genetic parameter estimation and rankings of estimated breeding values (Squillace 1974; Askew and El-Kassaby 1994; Namkoong et al. 1988; Vidal et al. 2015; Tambarussi et al. 2018). The development of highly polymorphic genetic markers, such as simple sequence repeats, has enabled pedigree reconstruction to be performed, eliminating the deleterious effect of hidden relatedness on accuracy of genetic parameters and breeding values in genetic evaluations (Lambeth et al. 2001; Grattapaglia et al. 2004; Doerksen and Herbinger 2010; Hansen and McKinney 2010; El-Kassaby et al. 2011).

More recently, the development of next-generation sequencing technologies has facilitated the development of genomic resources, even for organisms with missing reference genomes such as forest trees (Elshire et al. 2011; Chen et al. 2013; Neves et al. 2013; Plomion et al. 2014; Silva-Junior et al. 2015). These technologies generate abundant genome-wide genetic markers, such as single nucleotide polymorphisms (SNPs), which allow the construction of a marker-based relationship matrix (Nejati-Javaremi et al. 1997; VanRaden 2008). Such matrices provide a tool to track Mendelian segregation (Visscher et al. 2006; Zapata-Valenzuela et al. 2013), historical relatedness before base population defined by pedigree (Powell et al. 2010) and linkage disequilibrium (LD) between markers and quantitative trait loci (QTLs) (Habier et al. 2013). In particular, tracking LD improves the ability to estimate genetic covariance and helps achieve the more accurate estimation of genetic variance (Lippert et al. 2013). The marker-based relationship matrix can then be used as a tool to predict phenotypes for individuals with genotypes through genomic selection (GS) models (Resende et al. 2012; Beaulieu et al. 2014; Muñoz et al. 2014; Gamal El-Dien et al. 2015; Ratcliffe et al. 2015; Bartholomé et al. 2016; Isik et al. 2016).

Forest tree species are a challenge as they are often characterized by high genetic diversity, large effective population size, and rapid LD decay, which requires genotyping of large training populations to fully utilize all the benefits of the genomics approach. The complete genotyping of a forest tree progeny test is currently cost-prohibitive due to their large dimensions (thousands of trees), and a reasonable alternative should be used (Beaulieu et al. 2014). El-Kassaby and Lstibůrek (2009) proposed a partial pedigree reconstruction as an efficient alternative to full pedigree reconstruction to improve the comparative precision of genetic parameters (El-Kassaby et al. 2011). Single-step evaluation (Legarra et al. 2009; Misztal et al. 2009) can be seen as a genomic-based equivalent of the above mentioned partial pedigree reconstruction to reasonably implement genomics into forest tree testing schemes. This strategy has already been successfully applied in animal breeding and also in some forest tree genetic evaluations (Christensen and Lund 2010; Meuwissen et al. 2011; Christensen et al. 2012; Cappa et al. 2017, 2018; Ratcliffe et al. 2017). The rescaling of the marker-based relationship matrix to that inferred from the documented pedigree is the greatest challenge in single-step genetic evaluation to avoid any inaccuracy of genetic parameter estimates. Usually, the marker-based matrix ***G*** is adjusted regarding differences of average diagonal and average off-diagonal elements to its pedigree-based counterpart. Nevertheless, the rescaling effects are highly variable and depend on the method used for ***G*** matrix construction (Forni et al. 2011). Several rescaling approaches have already been developed (Forni et al. 2011; Vitezica et al. 2011; Gao et al. 2012). However, there is lack of knowledge on the effect of incomplete pedigree information on accuracy of predicted breeding values in single-step evaluation. The rescaling of the ***G*** matrix based on incomplete pedigree-based relationship appears to be causing an issue. Individuals with shallow, single-generation pedigrees are causing the ***G*** matrix elements to be larger, on average, compared with the pedigree-based matrix ***A***. In contrast, individuals with deep pedigrees have, on average, ***G*** matrix elements that are smaller (Misztal et al. 2013). The strategy to avoid this issue is through implementing patterns of population history. Misztal et al. (2013) developed a strategy based on implementation of unknown parental groups in a multibreed population. We found this strategy, however, unsuitable in our case due to the lack of isolation in mating events and rather we focused on reconstruction of hidden relatedness. A previous study performed on the material used in the current study was focused on sib-ship reconstruction and found a reasonable proportion of relatedness (including selfing), unrecognized by documented pedigree. The implementation of the relationship matrix based on sib-ship reconstruction improved the precision of genetic parameters and response to selection especially in traits suffering from inbreeding depression (Klápště et al. 2017). This study, therefore, investigates the efficiency of single-step genetic evaluation in an advanced generation of a *Eucalyptus nitens* breeding population, with an only partially tracked pedigree. It compares the effect of using relatedness inferred from sib-ship reconstruction versus the documented pedigree in the process of marker-based relationship matrix rescaling. In addition, the pedigree-based matrix was modified to take into account the probability of selfing in an attempt to further improve the accuracy of this strategy.

Methods

Material

The studied population is a third generation breeding population, derived from 2 seed orchards (Klápště et al. 2017). The experiment includes 3593 individuals structured into 116 half-sib families, of which 691 were randomly selected, representing 72 tested families analyzed through sib-ship reconstruction in previous study (Klápště et al. 2017). The individuals were measured for diameter at breast height (DBH) and scored for straightness (STR) using a 9° scale from 1—crooked to 9—straight and malformation (MAL) coded as a binary trait where 1 is perfectly formed and 0 otherwise.

Genetic markers were generated through EUChip60K SNP chip (Silva-Junior et al. 2015) and filtered for GenTrain score > 0.5, GenCall > 0.15, minor allele frequency (MAF) > 0.05 and SNP call rate > 0.6 which generated 13 844 markers.

Statistical Analysis

Pedigree-Based Analysis

Genetic parameters such as additive genetic variance and heritability were estimated using a linear mixed model, implemented in the ASReml-R package (Butler et al. 2009) as follows:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Za} + \boldsymbol{Zr} + \boldsymbol{Zr}(s) + \boldsymbol{e}$$

where $\boldsymbol{y}$ is the vector of observations, $\boldsymbol{\beta}$ is the vector of fixed effects such as intercept and seed orchard, $\boldsymbol{a}$ is the vector of random effects for breeding values following $\mathrm{var}(\boldsymbol{a}) \sim N(0, \boldsymbol{A}\sigma_a^2)$, where ***A*** is the average numerator relationship matrix (Wright 1922) which is substituted by the combined relationship matrix ***H*** using both pedigree and marker information in the single-step evaluation (see below) and σ_a^2 is the additive genetic variance, $\boldsymbol{r}$ is the vector of random replication effects following $\mathrm{var}(\boldsymbol{r}) \sim N(0, \boldsymbol{I}\sigma_r^2)$, where $\boldsymbol{I}$ is the identity matrix and σ_r^2 is the replication variance, $\boldsymbol{r(s)}$ is the vector of random set nested within replication effects following $\mathrm{var}(\boldsymbol{r}(s)) \sim N(0, \boldsymbol{I}\sigma_{r(s)}^2)$ (set represents incomplete block within replication having fixed number of families from each seed orchard), $\boldsymbol{e}$ is the vector of residuals following $\mathrm{var}(\boldsymbol{e}) \sim N(0, \boldsymbol{I}\sigma_e^2)$, where σ_e^2 is the residual variance, $\boldsymbol{X}$ and $\boldsymbol{Z}$ are incidence matrices assigning fixed and random effects to observations in vector $\boldsymbol{y}$.

Single-Step Genetic Evaluation

Since the marker-based relationship matrix is reflecting both temporal and historical relatedness (Powell et al. 2010), the reference (base) population is different compared with the pedigree-based counterpart. Such discrepancies can result in biased estimations of genetic parameters and reduced accuracy of breeding values (Vitezica et al. 2011). Therefore, the adjustment of the marker-based relationship matrix is the most crucial step in the single-step evaluation. The marker-based relationship matrix $\boldsymbol{G}$ was constructed following (VanRaden 2008):

$$\boldsymbol{G} = \frac{\boldsymbol{ZZ'}}{2\sum_j p_j(1-p_j)}$$

where $\boldsymbol{Z} = \boldsymbol{M} - \boldsymbol{P}$, $\boldsymbol{M}$ is the matrix of genotypes coded 0, 1, and 2 as reference allele homozygote, heterozygote, and alternative allele homozygote, respectively, and $\boldsymbol{P}$ is the vector of doubled frequencies for alternative alleles, p_j is the frequency of the alternative allele at jth loci. The rescaling of the marker-based relationship matrix to adjust for a base population defined by the documented pedigree was performed following (Gao et al. 2012):

$$\begin{cases} \mathrm{Avg.diag}(\boldsymbol{G})\beta + \alpha = \mathrm{Avg.diag}(\boldsymbol{A}_{22}) \\ \mathrm{Avg.offdiag}(\boldsymbol{G})\beta + \alpha = \mathrm{Avg.offdiag}(\boldsymbol{A}_{22}) \end{cases}$$

Since the investigated field experiment is derived from a 3rd generation breeding population in a program with incomplete tracking of relatedness, 2 $\boldsymbol{A}_{22}$ matrices were implemented to rescale the $\boldsymbol{G}$ matrix: 1) based on tracked pedigree (HBLUP1), and 2) based on sib-ship reconstruction performed in a previous study (Klápště et al. 2017) (HBLUP2). We hypothesize that the implementation of a relationship matrix based on sib-ship reconstruction should result in a more precise adjustment of the marker-based relationship matrix to pedigree. The $\boldsymbol{G}$ matrix is usually not positive semi-definite, which is one of the mixed linear model assumptions, and weighting of the genomic and pedigree-based relationship matrices is required as follows:

$$\boldsymbol{G}_w = \boldsymbol{G}(1-w) + \boldsymbol{A}_{22}w$$

Alternatively, the pedigree-based relationship matrix was modified to take into account partial selfing following (Dutkowski et al. 2001; Gilmour and Dutkowski 2004). This pedigree-based matrix was produced by using the "selfing" option in "asreml.Ainverse" function, implemented in the ASReml-R package (Butler et al. 2009).

The $\boldsymbol{H}$ matrix, implementing both marker and pedigree-based information, was constructed as follows:

$$\boldsymbol{H} = \begin{bmatrix} \boldsymbol{A}_{11} + \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}(\boldsymbol{G}_w - \boldsymbol{A}_{22})\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21} & \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{G}_w \\ \boldsymbol{G}_w\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21} & \boldsymbol{G}_w \end{bmatrix}$$

where $\boldsymbol{A}_{11}$ is the relationship matrix for nongenotyped individuals, $\boldsymbol{A}_{12}$ and $\boldsymbol{A}_{21}$ are the relationship matrices between genotyped and nongenotyped individuals and $\boldsymbol{A}_{22}$ is the pedigree-based relationship matrix for genotyped individuals, $\boldsymbol{G}$ is the marker-based relationship matrix which is only available for genotyped individuals.

Narrow-sense heritability for continuous traits was estimated as follows:

$$\hat{h}^2 = \frac{\hat{\sigma}_a^2}{\hat{\sigma}_a^2 + \hat{\sigma}_e^2}$$

and its alternative for binary trait was estimated as follows:

$$\hat{h}^2 = \frac{\hat{\sigma}_a^2}{\hat{\sigma}_a^2 + \phi\dfrac{\pi^2}{3}}$$

where ϕ is the over/under dispersion coefficient. The theoretical accuracy of breeding values was estimated as follows:

$$r = \sqrt{1 - \frac{\mathrm{PEV}}{(1+F_i)\sigma_a^2}}$$

where PEV is prediction error variance (Mrode 2014), and F_i is the inbreeding coefficient of the ith individual. The leave-one-out cross-validation strategy was implemented as an independent evaluation of the tested models. Prediction accuracy for continuous traits was estimated as the correlation between breeding values estimated in the pedigree-based analysis and those predicted in the cross-validation procedure. Additionally, the predicted genomic breeding values for genotyped individuals were correlated with breeding values estimated in the independent analysis using the relationship matrix based on information from sib-ship reconstruction. Correlations were only estimated using the set of genotyped individuals. The area under a ROC curve (AUC) was used to estimate prediction accuracy for binary traits.

Results

The pedigree-based analysis resulted in heritability from 0.05 (MAL) to 0.28 (STR) for form traits and 0.22 (DBH) for the growth trait analyzed. The estimates for all traits were found to be statistically significant with regard to their standard errors (α = 0.05). The accuracy of the breeding values was moderate and reached 0.54 for the growth trait DBH and from 0.32 to 0.58 for form traits (MAL and STR) (Table 1). The LD in our population decayed to an r^2 of 0.2 within 3 kb, which is a common pattern in forest trees (Figure 1). The comparison of marker-based and sib-ship reconstruction-based relationship coefficients showed a clear deflation of marker-based estimates across the whole spectrum of relationship coefficients (Figure 2). The marker-based relationship matrix $\boldsymbol{G}$ was rescaled following Gao et al. (2012), using pedigree-based and sib-ship reconstruction-based relationship matrices. The parameters α and β reached values of 0.005090189 and 1.322984116 in the pedigree-based scenario and 0.01343057 and 1.33787272 in the sib-ship reconstruction scenario.

Table 1. Variance components, heritability, their standard errors in parentheses, breeding values accuracy, their prediction accuracy (PA) in parentheses [2 prediction accuracies are reported for genotyped individuals regarding base to which are correlated (a) documented pedigree-based breeding value estimates; (b) sib-ship–based breeding value estimates—bold], and model fit for pedigree-based model (ABLUP), single-step evaluation where ***G*** matrix is rescaled to documented pedigree (HBLUP1) and single-step evaluation where ***G*** matrix is rescaled to information from sib-ship reconstruction (HBLUP2) under no selfing probability

Model	Parameter	DBH	STR	MAL
ABLUP	Additive genetic var.	132.4 (28.88)	0.488 (0.096)	0.218 (0.078)
	Replicate var.	0.000 (0.000)	0.074 (0.025)	0.011 (0.013)
	Rep(set) var.	0.000 (0.000)	0.016 (0.011)	0.000 (0.000)
	Residual var.	480.5 (26.81)	1.280 (0.052)	1.000 (0.000)
	Heritability	0.216 (0.045)	0.278 (0.052)	0.050 (0.017)
	Acc (PA)—total	0.54 (0.69)	0.58 (0.68)	0.32 (0.56)
	Acc (PA)—mother	0.54 (NA)	0.56 (NA)	0.35 (NA)
	Acc (PA)—offspring	0.54 (0.69)	0.58 (0.68)	0.31 (0.56)
	AIC	25153.28	5290.2	8271.75
HBLUP1	Additive genetic var.	147.7 (31.47)	0.484 (0.086)	0.216 (0.077)
	Replicate var.	0.000 (0.000)	0.073 (0.025)	0.011 (0.013)
	Rep(set) var.	0.000 (0.000)	0.017 (0.013)	0.000 (0.000)
	Residual var.	469.8 (27.75)	1.277 (0.076)	1.000 (0.000)
	Heritability	0.239 (0.048)	0.275 (0.047)	0.050 (0.017)
	Acc (PA)—total	0.57 (0.66)	0.59 (0.64)	0.34 (0.56)
	Acc (PA)—mother	0.55 (NA)	0.56 (NA)	0.36 (NA)
	Acc (PA)—offspring NonGen	0.55 (0.66)	0.58 (0.68)	0.32 (0.56)
	Acc (PA)—offspring Gen	0.63 (0.58, **0.37**)	0.65 (0.47, **0.58**)	0.40 (0.57)
	AIC	25148.78	5286.23	8272.943
HBLUP2	Additive genetic var.	131.9 (28.39)	0.488 (0.088)	0.231 (0.078)
	Replicate var.	0.000 (0.000)	0.074 (0.025)	0.011 (0.013)
	Rep(set) var.	0.000 (0.000)	0.017 (0.011)	0.00 (0.000)
	Residual var.	480.2 (25.85)	1.272 (0.077)	1.000 (0.000)
	Heritability	0.215 (0.044)	0.277 (0.047)	0.053 (0.017)
	Acc (PA)—total	0.55 (0.67)	0.59 (0.64)	0.34 (0.56)
	Acc (PA)—mother	0.54 (NA)	0.56 (NA)	0.36 (NA)
	Acc (PA)—offspring NonGen	0.53 (0.66)	0.58 (0.68)	0.32 (0.56)
	Acc (PA)—offspring Gen	0.62 (0.58, **0.42**)	0.66 (0.47, **0.58**)	0.39 (0.57)
	AIC	25137.04	5283.96	8275.39

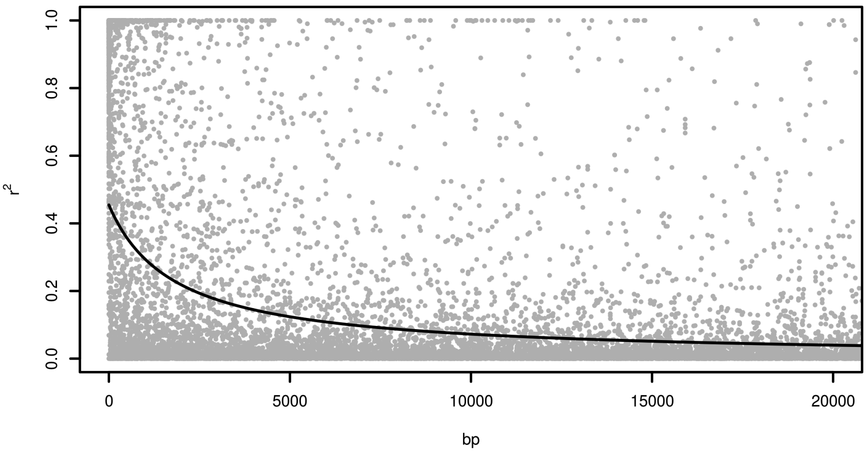

Figure 1. LD decay in population under study.

The single-step evaluation resulted in heritability estimates ranging from 0.05 to 0.28 in the documented pedigree-based scenario and from 0.05 to 0.28 in the sib-ship reconstruction scenario. A slight increase in heritability from 0.22 to 0.24 was observed in the HBLUP1 scenario for DBH but was not accompanied by any concurrent increase in model fit. STR was the only trait to show improvement in the theoretical accuracy of breeding values when using information from sib-ship reconstruction to rescale the ***G*** matrix compared with the pedigree-based scenario. The trend in the theoretical accuracy of the breeding values, however, is a reflection of the trend in heritability, which was not always a reflection of the model fit. The prediction accuracy was investigated through a leave-one-out strategy only in the default scenario (no selfing probability and 0.05 weight on pedigree information). Our study found

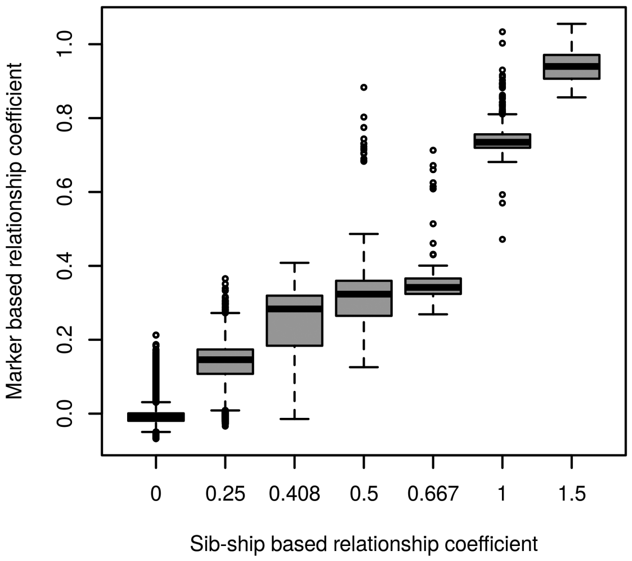

Figure 2. Correspondence of sib-ship and marker-based relatedness/self-relatedness.

the highest prediction accuracy was reached in the pedigree-based analysis (ABLUP), ranging from 0.68 to 0.69. Similar prediction accuracies were found in the HBLUP1 and HBLUP2 scenarios for individuals without genomic information, ranging from 0.66 to 0.68. The lowest prediction accuracy was obtained among individuals with genomic information and ranged from 0.47 to 0.58 in both the HBLUP scenarios when predicted genomic breeding values were correlated with breeding values estimated in ABLUP. However, when the predicted genomic breeding values were correlated with breeding values estimated using relationships from sib-ship reconstruction (performed in only genotyped sample), the prediction accuracy in DBH increased from 0.37 obtained in HBLUP1 to 0.42 obtained in HBLUP2. The prediction accuracy in STR remained constant across both scenarios (Table 1). The reduced accuracy of predicted breeding values was caused by the fact that while sib-ship–based estimated breeding values were estimated using only genotyped individuals, predicted breeding values are biased because nongenotyped individuals are also used in the prediction process and representing 80% of the total population size.

The increase of selfing probability in the pedigree-based relationship matrix resulted in a decrease in heritability across all investigated traits. The different weights applied to the pedigree-based information did not affect the heritability, except for MAL, where a higher weight set on the pedigree-based relationship matrix resulted in a decrease in heritability, with a more obvious pattern in the sib-ship reconstruction scenario (Supplementary File 1). The theoretical accuracy of breeding value estimations was slightly higher in the single-step evaluation compared with the pedigree-based alternative, mainly due to the noted improvement in the accuracy of genotyped individuals. The sib-ship scenario in MAL, however, also improved the accuracy of mothers and nongenotyped offspring. The introduction of selfing probability followed the pattern observed in heritability and decreased with the increase of selfing probability. Similarly, the increased weight of the pedigree-based relationship matrix in the rescaling process resulted in a reduction of breeding value accuracy, with the most noticeable trend for the trait MAL (Supplementary File 1).

Discussion

Controlled pollination in forest tree breeding is expensive, time-consuming, and labor-intensive and its efficiency is affected by both biological and environmental limitations. Therefore, open pollination has been preferred in forest tree breeding programs, such as in the case of the *E. nitens* program in New Zealand (Burdon and Shelbourne 1971). However, this strategy comes at the cost of incomplete knowledge of genealogy, likely to cause the estimation of genetic parameters in quantitative genetic evaluations that are less reliable (Ratcliffe et al. 2017). The development of genetic markers has allowed recovery of missing relatedness and genealogy through pedigree/sib-ship reconstruction (Askew and El-Kassaby 1994; Lambeth et al. 2001; Vidal et al. 2015). Dense marker arrays have also allowed the construction of realized relationship matrices (Nejati-Javaremi et al. 1997; VanRaden 2008), which usually increase the accuracy of genetic parameter estimates and allow for more efficient selection of superior genotypes (Resende et al. 2012; Gamal El-Dien et al. 2015; Ratcliffe et al. 2015, 2017; Suontama et al. 2018).

El-Kassaby and Lstibůrek (2009) and El-Kassaby et al. (2011) found partial pedigree reconstruction as a feasible and cost-effective alternative to full pedigree reconstruction to improve the precision of genetic parameters. Our previous study (Klápště et al. 2017) focused on the effect of sib-ship reconstruction to improve genetic parameters. A significant benefit was demonstrated for those traits suffering from inbreeding depression, achieved by recognizing selfs in the population, consequently leading to an increase in additive genetic variance, heritability, and improvement in estimated genetic gain. Improvement in breeding value accuracy was also observed for traits free of inbreeding depression due to the recovery of hidden relatedness and potential correction of pedigree errors. The analysis found 630 pair-wise relationships originally defined as half-sibs to be unrelated (See figure 2 in Klápště et al. 2017). However, defining all pedigree errors was not possible due to an inability to assign parents to each individual in the sib-ship reconstruction strategy. Similarly, the current study found a benefit when rescaling the marker-based ***G*** matrix according to the relationship matrix based on information from sib-ship reconstruction, rather than the documented pedigree. The benefit seen in the improved model fit (Table 1) and breeding values accuracy (Table 1—bold numbers) was more evident in production trait (DBH), which was more likely to suffer from inbreeding depression (Hardner and Tibbits 1998). Therefore, parentage/sib-ship reconstruction should be performed before ***G*** matrix rescaling in single-step evaluations, when applied in OP breeding programs. However, the low correlations between breeding values estimated on the basis of sib-ship reconstruction with those predicted in single-step evaluation is a result of the high influence of unrecognized relatedness and pedigree errors from nongenotyped individuals (contributing by 80% of the total population size) on breeding values predicted from single-step evaluation. On the other hand, there was no improvement in the accuracy of breeding values estimated in the nongenotyped part of the population after implementation of genomic information. This can be again caused by a high level of uncertainty in relatedness (coming from both the hidden relatedness and pedigree errors) across the population. In this case, we recommend the pedigree/sib-ship reconstruction of the whole population to reach a higher accuracy of predicted breeding values.

Results presented in this study showed that accumulation of unrecognized relatedness and pedigree errors across several generations of breeding cycles resulted in virtually nonexistent between-family variation, with the main source of genetic variation generated by within-family variation (Figure 3). In contrary, the analysis using traits having similar level of heritability but complete pedigree information found large proportion of the genetic variance attributed to between-family variance (Thistlethwaite et al. 2017). Therefore, the missing pedigree information on the paternal side of the current progeny population, as well as for the parents in previous generations, appears to undermine the ability of the REML algorithm to differentiate families, and breeding values are shrunk toward the population mean (Figure 3) (Henderson 1975; Garrick et al. 2009). On the other hand, using genomic markers allowed the recovery of hidden relatedness and pedigree errors, resulting in a more disperse distribution of genomic breeding values compared with their pedigree-based equivalents (Figure 4). There are several strategies developed in animal breeding to overcome uncertain paternity using phenotypic data (Sapp et al. 2007) or construction of a sire probability matrix (Henderson 1988). However, the probability for many possible males (as would be the most likely scenario in forest trees) assigned to each nongenotyped offspring is not sufficient to increase the accuracy of genetic parameter estimates (Konigsberg and Cheverud 1992). The purpose of GS is primarily the approximation of pedigree-based breeding values through the implementation of genetic markers (Meuwissen et al. 2001). When the pedigree-based estimates of breeding values are imprecisely estimated, however, the resulting prediction accuracy (in terms of correlation between pedigree-based estimated and marker-based predicted breeding values) will undermine the efficiency of genomic predictions. Under such conditions, we would highly recommend implementation of genetic markers across the whole population and perform either pedigree or sib-ship reconstruction to obtain relatedness structure approaching the reality. The breeding values estimated on the basis of pedigree/sib-ship reconstruction will reach higher accuracy and provide a better base for the estimation of prediction accuracy.

The construction of a relationship matrix based on information from genetic markers allows tracking of not only temporal relatedness, as defined by the pedigree-based base population, but also Mendelian sampling (Visscher et al. 2006; Zapata-Valenzuela et al. 2013) and historical relatedness (Powell et al. 2010). This is highly beneficial in species in the initial phase of domestication, where pedigrees are shallow and simple, such as forest trees. Additional information from all genotyped individuals increases the precision of breeding values considerably (Table 1). Ratcliffe et al. (2017) investigated the effect of genotyping intensity in a single-step evaluation in white spruce and found continuous improvement in the accuracy of genetic parameters and model fit with increasing genotyping intensity. The study demonstrates the high value of genomic information, implemented in the initial phase of breeding programs, where pedigrees are simple and incomplete. Similarly, we found a large increase in the theoretical accuracy of breeding values for genotyped individuals compared with those without genotypes showing no improvement (Table 1). However, the prediction accuracy of genotyped individuals increased only when sib-ship reconstruction-based breeding values were used as a base. The fact that nongenotyped individuals reached higher prediction accuracy than genotyped individuals can be explained by the highly biased estimates of family means targeted in pedigree-based predictions (Zapata-Valenzuela et al. 2013). On the other hand, within-family variation targeted by genomic-based prediction is largely unreliable due to accumulated unrecognized relatedness and pedigree errors

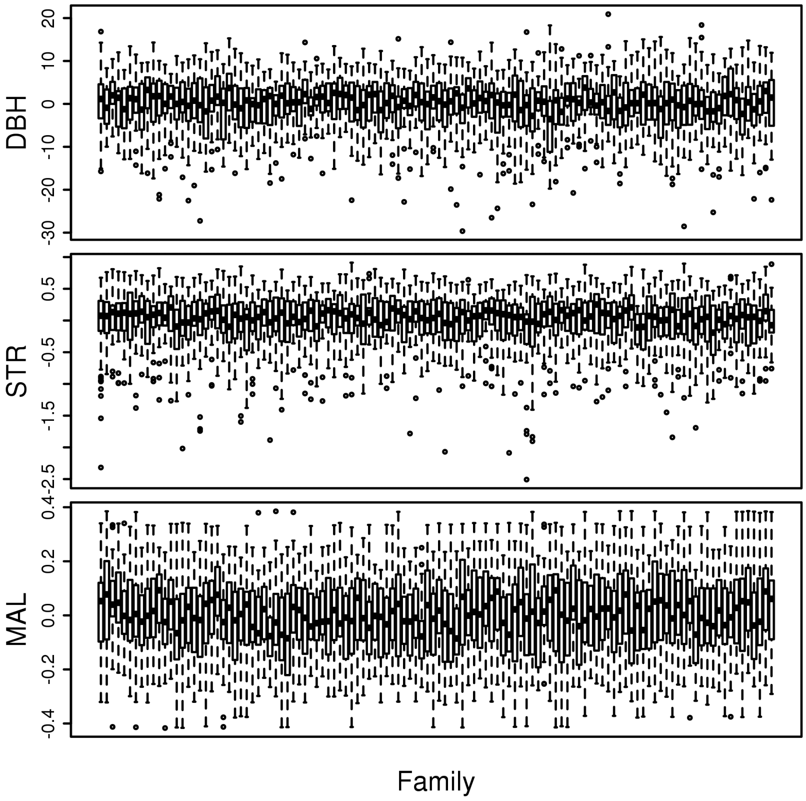

Figure 3. Distribution of pedigree-based breeding values within each family.

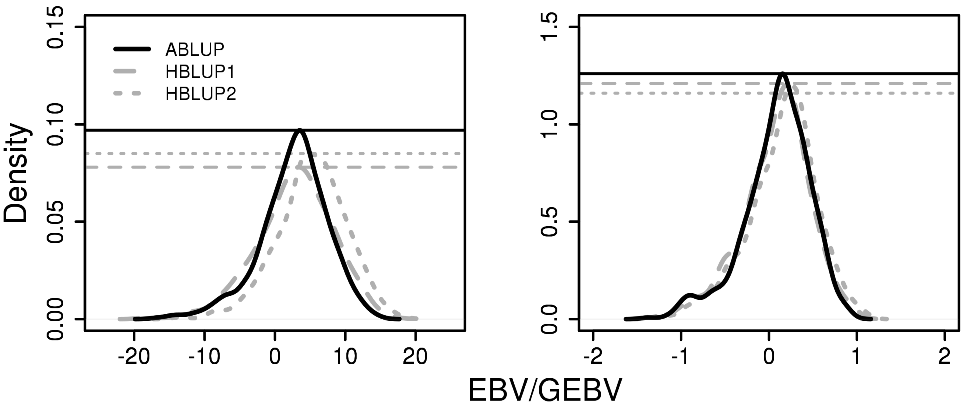

Figure 4. Density of EBV/GEBV values distribution for continuous traits DBH (left plot) and STR (right plot) under the various models tested in population of genotyped individuals. Horizontal lines represent peak of the breeding values distributions for each scenario.

across the breeding cycles. Therefore, genomic approaches remain a very attractive option in forest species even where shortening of the breeding cycle is not possible due to late flowering. Gains can be made instead through a more complete understanding of underlying relationships and more accurate estimation of genetic parameters.

The precision of marker-based estimates of genetic parameters remain sensitive to selection, and a combination of marker and pedigree information is still recommended (Ducrocq and Patry 2010; Vitezica et al. 2011). In addition, the definition of the reference population in marker-based relationship matrices is rather arbitrary (Speed and Balding 2015) and should be rescaled with respect to the pedigree. de los Campos et al. (2015) argued that total heritability can be recovered only when all QTLs are included in the marker array, and is partially lost due to imperfect LD when only SNPs surrounding QTLs are available. Lippert et al. (2013) investigated the effect of using both QTL and non-QTL markers to construct a marker-based relationship matrix and found that only using QTL markers provides the most accurate estimate of additive genetic variance and heritability. Our analysis was performed using a multispecies *Eucalyptus* SNP chip (Silva-Junior et al. 2015), and ~10k markers were informative in this *E. nitens* population. The decay in LD was fast, as is common in forest tree species, and disappeared within ~3 kb (Figure 1). Therefore, capturing markers linked to QTLs is rather unlikely, and relatedness with co-segregation is probably the major source of capturing QTL effects. On the other hand, using an overwhelming amount of the genomic data does not increase the accuracy of prediction model after reaching saturation (Habier et al. 2013) and trait specific SNP prioritization should be applied (Lippert et al. 2013). However, the genetic complexity of the investigated traits can prohibit the reliable selection of causal variants, and therefore, prediction models rely rather on relatedness and co-segregation (Habier et al. 2007, 2013). Our previous analysis (Suontama et al. 2018) found that the sample coming from Tinkers (seed orchard undergoing more intensive selection) had a higher GEBV accuracy compared with the sample coming from the Waiouru seed orchard (seed orchard established as a clonal archive having broader genetic diversity), which was reached thanks to slower LD decay, capturing longer effective chromosomal fragments. Due to the fact that sample from Waiouru seed orchard had twice the sample size and produced lower accuracy of GEBVs, we found that the model didn't reach a saturation point where any additional markers wouldn't increase accuracy of GEBVs. Therefore, there appeared to still be space for improvement of genomic resources in *Eucalyptus* species to create a robust genomic prediction model.

The recovered relatedness through genetic markers in the set of genotyped individuals was underestimated compared with expectations (Figure 2). This can be caused by several cycles of selection and a lack of unrelated individuals to provide a reference for inferring actual relatedness among related individuals (Speed and Balding 2015), and had to be rescaled with respect to the pedigree-based counterpart before blending with the pedigree-based relationship matrix. The rescaling of the marker-based relationship matrix with respect to the pedigree-based equivalent is the most crucial step in single-step genetic evaluation. The difference in the scale of relationship coefficients between marker-based and pedigree-based counterparts causes a decrease in the accuracy of genomic breeding values (Ducrocq and Patry 2010; Forni et al. 2011; Vitezica et al. 2011). We tested 2 scenarios: 1) rescaling of the marker-based relationship matrix with regard to the documented pedigree and 2) rescaling the marker-based relationship matrix with regard to the relationship matrix derived from sib-ship reconstruction. The implementation of information from sib-ship reconstruction in the **G** matrix rescaling process resulted in a considerable improvement in model fit compared with the model that used the documented pedigree. This trend is especially observed in traits suffering from inbreeding depression, such as DBH (Hardner and Tibbits 1998). These improvements were achieved in spite of the fact that the sib-ship reconstruction could only recover higher classes of relatedness, such as full-sibs and half-sibs, but not first and second order cousins as found in the documented pedigree. This means that the greater degree of relatedness recovered by sib-ship reconstruction has a more significant impact on the improvement of genetic parameter estimates through the **G** matrix rescaling process than ignored or undiscovered lower degrees of relatedness. Therefore, pedigree/sib-ship reconstruction is highly recommended prior to **G** matrix rescaling in the single-step genetic evaluation, especially in species with an OP breeding program, where selfs are viable. However, we could not utilize the full potential of relatedness recovered by sib-ship reconstruction due to loss of connectivity with the remainder of the pedigree, as a simple blending of the sib-ship reconstruction-based relationship matrix (sib-ship–based A_{22}) into the pedigree-based relationship matrix would cause the resulting matrix not to be positive definite. Therefore, newly obtained relatedness information should be used only in the rescaling, but not in the weighting step. A more useful strategy would be to perform parentage analysis instead of sib-ship reconstruction when genomic information is also available for parental populations. In this case, consistency between original pedigree and the reconstructed part would remain and positive definite nature of resulting relationship matrix warranted.

In some cases, marker information is not sufficient to capture all additive genetic variance, and residual polygenic effects have to be included in the prediction model (Aguilar et al. 2010; Christensen and Lund 2010). In addition, the implementation of a residual polygenic effect reduces the bias in SNP effects and increases their transferability over generations (Solberg et al. 2009). Similarly, in the single-step genetic evaluation, the weighting of marker and pedigree information is applied. We tested a broad range of weights from 0.05 to 0.5 for the pedigree information, but any increase resulted in a decrease in breeding value accuracies for genotyped individuals, while no effect was observed in nongenotyped individuals (Supplementary File 1).

Our previous analysis identified ~4% selfing in the genotyped sample (Klápště et al. 2017) and, therefore, we modified the pedigree-based matrix for selfing probability as proposed by Dutkowski et al. (2001) before blending with the marker-based ***G*** matrix. The modified selfing probability did not result in any additional improvement in the accuracy of breeding values, with decreases observed once probability exceeded 3% (Supplementary File 1). These results confirm our finding of 4% selfing in previous sib-ship reconstruction analysis (Klápště et al. 2017), it is, therefore, beneficial to implement selfing probability in any single-step genetic evaluation in species where there is strong evidence of viable selfing.

In this study, we have shown how the increase in connectivity between genotyped individuals through genomic similarity has a big impact on the resulting accuracy of breeding values compared with information from a sparse pedigree. In addition, implementation of genomic information in a quantitative genetic evaluation can dissect genetic and environment effects more precisely (Gamal El-Dien et al. 2016). Modification of the relationship matrix for selfing before blending and/or rescaling was found to be important in our population and would be recommended for other OP tree breeding programs.

Supplementary Material

Supplementary data are available at *Journal of Heredity* online.

Funding

This work was supported by the Specialty Wood Partnership Program (SWP) contract number C04X1104 and Scion Strategic Investment Funding.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability

All genomic and phenotypic data is publically available from DRYAD data depository, doi:10.5061/dryad.cb4m96b.

References

Aguilar I, Misztal I, Johnson DL, Legarra A, Tsuruta S, Lawlor TJ. 2010. Hot topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. *J Dairy Sci*. 93:743–752.

Askew GR, El-Kassaby YA. 1994. Estimation of relationship coefficients among progeny derived from wind-pollinated orchard seeds. *Theor Appl Genet*. 88:267–272.

Bartholomé J, Van Heerwaarden J, Isik F, Boury C, Vidal M, Plomion C, Bouffier L. 2016. Performance of genomic prediction within and across generations in maritime pine. *BMC Genomics*. 17:604.

Beaulieu J, Doerksen T, Clément S, MacKay J, Bousquet J. 2014. Accuracy of genomic selection models in a large population of open-pollinated families in white spruce. *Heredity (Edinb)*. 113:343–352.

Burdon R, Shelbourne C. 1971. Breeding populations for recurrent selection: conflicts and possible solutions. *N Z J For Sci*. 1:174–193.

Butler D, Cullis BR, Gilmour A, Gogel B. 2009. *Asreml-r reference manual*. Brisbane: The State Queensland, Department of Primary Industries and Fishery.

Cappa EP, El-Kassaby YA, Munoz F, García MN, Villalba PV, Klápště J, Marcucci Poltri SN. 2017. Improving accuracy of breeding values by incorporating genomic information in spatial-competition mixed models. *Mol Breed*. 37:125.

Cappa EP, El-Kassaby YA, Muñoz F, Garcia MN, Villalba PV, Klápště J, Marcucci Poltri SN. 2018. Genomic-based multiple-trait evaluation in *Eucalyptus grandis* using dominant DArT markers. *Plant Sci*. 271:27–33.

Chen C, Mitchell SE, Elshire RJ, Buckler ES, El-Kassaby YA. 2013. Mining conifers' mega-genome using rapid and efficient multiplexed high-throughput genotyping-by-sequencing (GBS) SNP discovery platform. *Tree Genet Genomes*. 9:1537–1544.

Christensen OF, Lund MS. 2010. Genomic prediction when some animals are not genotyped. *Genet Sel Evol*. 42:2.

Christensen OF, Madsen P, Nielsen B, Ostersen T, Su G. 2012. Single-step methods for genomic evaluation in pigs. *Animal*. 6:1565–1571.

de Los Campos G, Sorensen D, Gianola D. 2015. Genomic heritability: what is it? *PLoS Genet*. 11:e1005048.

Doerksen TK, Herbinger CM. 2010. Impact of reconstructed pedigrees on progeny-test breeding values in red spruce. *Tree Genet Genomes*. 6:591–600.

Ducrocq V, Patry C. 2010. Combining genomic and classical information in national BLUP evaluation to reduce bias due to genomic pre-selection. *Interbull Bull*. 41:33–36.

Dutkowski G, Gilmour A, Borralho N. 2001. Modification of the additive relationship matrix for open pollinated trial. In: S Barros, R Ipinzà, editors. Developing the eucalypt of the future. Proceedings of IUFRO Working Group 2.08.03 Conference, 10-15 September, Valdivia, Chile.

El-Kassaby YA, Cappa EP, Liewlaksaneeyanawin C, Klápště J, Lstibůrek M. 2011. Breeding without breeding: is a complete pedigree necessary for efficient breeding? *PLoS One*. 6:e25737.

El-Kassaby YA, Lstibůrek M. 2009. Breeding without breeding. *Genet Res (Camb)*. 91:111–120.

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One*. 6:e19379.

Forni S, Aguilar I, Misztal I. 2011. Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. *Genet Sel Evol*. 43:1.

Gamal El-Dien O, Ratcliffe B, Klápště J, Chen C, Porth I, El-Kassaby YA. 2015. Prediction accuracies for growth and wood attributes of interior spruce in space using genotyping-by-sequencing. *BMC Genomics*. 16:370.

Gamal El-Dien O, Ratcliffe B, Klápště J, Porth I, Chen C, El-Kassaby YA. 2016. Implementation of the realized genomic relationship matrix to open-pollinated white spruce family testing for disentangling additive from nonadditive genetic effects. *G3 (Bethesda)*. 6:743–753.

Gao H, Christensen OF, Madsen P, Nielsen US, Zhang Y, Lund MS, Su G. 2012. Comparison on genomic predictions using three GBLUP methods and two single-step blending methods in the Nordic Holstein population. *Genet Sel Evol*. 44:8.

Garrick DJ, Taylor JF, Fernando RL. 2009. Deregressing estimated breeding values and weighting information for genomic regression analyses. *Genet Sel Evol*. 41:55.

Gilmour A, Dutkowski G. 2004. Pedigree options in asreml. Available from: https://www.animalgenome.org/bioinfo/resources/manuals/ASReml3/pedigree.pdf

Grattapaglia D, Ribeiro VJ, Rezende GD. 2004. Retrospective selection of elite parent trees using paternity testing with microsatellite markers: an alternative short term breeding tactic for Eucalyptus. *Theor Appl Genet*. 109:192–199.

Habier D, Fernando RL, Dekkers JC. 2007. The impact of genetic relationship information on genome-assisted breeding values. *Genetics*. 177:2389–2397.

Habier D, Fernando RL, Garrick DJ. 2013. Genomic BLUP decoded: a look into the black box of genomic prediction. *Genetics*. 194:597–607.

Hansen OK, McKinney LV. 2010. Establishment of a quasi-field trial in *Abies nordmanniana*—test of a new approach to forest tree breeding. *Tree Genet Genomes*. 6:345–355.

Hardner C, Tibbits W. 1998. Inbreeding depression for growth, wood and fecundity traits in *Eucalyptus nitens*. *For Genet*. 5:11–20.

Henderson CR. 1975. Best linear unbiased estimation and prediction under a selection model. *Biometrics*. 31:423–447.

Henderson CR. 1988. Use of an average numerator relationship matrix for multiple-sire joining. *J Animal Sci*. 66:1614–1621.

Isik F, Bartholomé J, Farjat A, Chancerel E, Raffin A, Sanchez L, Plomion C, Bouffier L. 2016. Genomic selection in maritime pine. *Plant Sci*. 242:108–119.

Klápště J, Suontama M, Telfer E, Graham N, Low C, Stovold T, McKinley R, Dungey H. 2017. Exploration of genetic architecture through sib-ship reconstruction in advanced breeding population of *Eucalyptus nitens*. *PLoS One*. 12:e0185137.

Konigsberg LW, Cheverud JM. 1992. Uncertain paternity in primate quantitative genetic studies. *Am J Primatol*. 27:133–143.

Lambeth C, Lee BC, O'Malley D, Wheeler N. 2001. Polymix breeding with parental analysis of progeny: an alternative to full-sib breeding and testing. *Theor Appl Genet*. 103:930–943.

Legarra A, Aguilar I, Misztal I. 2009. A relationship matrix including full pedigree and genomic information. *J Dairy Sci*. 92:4656–4663.

Lippert C, Quon G, Kang EY, Kadie CM, Listgarten J, Heckerman D. 2013. The benefits of selecting phenotype-specific variants for applications of mixed models in genomics. *Sci Rep*. 3:1815.

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

Meuwissen TH, Luan T, Woolliams JA. 2011. The unified approach to the use of genomic and pedigree information in genomic evaluations revisited. *J Anim Breed Genet*. 128:429–439.

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

Misztal I, Vitezica ZG, Legarra A, Aguilar I, Swan AA. 2013. Unknown-parent groups in single-step genomic evaluation. *J Anim Breed Genet*. 130:252–258.

Mrode RA. 2014. *Linear models for the prediction of animal breeding values*. CABI Publishing, Wallingford, UK.

Muñoz PR, Resende MF Jr, Gezan SA, Resende MD, de Los Campos G, Kirst M, Huber D, Peter GF. 2014. Unraveling additive from nonadditive effects using genomic relationship matrices. *Genetics*. 198: 1759–1768.

Namkoong G, Kang HC, Brouard JS. 1988. *Tree breeding: principles and strategies*. New York: Springer-Verlag. p. 177.

Nejati-Javaremi A, Smith C, Gibson JP. 1997. Effect of total allelic relationship on accuracy of evaluation and response to selection. *J Anim Sci*. 75:1738–1745.

Neves LG, Davis JM, Barbazuk WB, Kirst M. 2013. Whole-exome targeted sequencing of the uncharacterized pine genome. *Plant J*. 75:146–156.

Plomion C, Chancerel E, Endelman J, Lamy JB, Mandrou E, Lesur I, Ehrenmann F, Isik F, Bink MC, van Heerwaarden J, *et al*. 2014. Genome-wide distribution of genetic diversity and linkage disequilibrium in a mass-selected population of maritime pine. *BMC Genomics*. 15:171.

Potts BM, Dungey HS. 2004. Interspecific hybridization of eucalyptus: key issues for breeders and geneticists. *New For*. 27:115–138.

Powell JE, Visscher PM, Goddard ME. 2010. Reconciling the analysis of IBD and IBS in complex trait studies. *Nat Rev Genet*. 11:800–805.

Ratcliffe B, El-Dien OG, Cappa EP, Porth I, Klápště J, Chen C, El-Kassaby YA. 2017. Single-step BLUP with varying genotyping effort in open-pollinated *Picea glauca*. *G3 (Bethesda)*. 7:935–942.

Ratcliffe B, El-Dien OG, Klápště J, Porth I, Chen C, Jaquish B, El-Kassaby YA. 2015. A comparison of genomic selection models across time in interior spruce (*Picea engelmannii* × *glauca*) using unordered SNP imputation methods. *Heredity (Edinb)*. 115:547–555.

Resende MF Jr, Muñoz P, Acosta JJ, Peter GF, Davis JM, Grattapaglia D, Resende MD, Kirst M. 2012. Accelerating the domestication of trees using genomic selection: accuracy of prediction models across ages and environments. *New Phytol*. 193:617–624.

Sapp RL, Zhang W, Bertrand JK, Rekaya R. 2007. Genetic evaluation in the presence of uncertain additive relationships. I. Use of phenotypic information to ascertain paternity. *J Anim Sci*. 85:2391–2400.

Silva-Junior OB, Faria DA, Grattapaglia D. 2015. A flexible multi-species genome-wide 60K SNP chip developed from pooled resequencing of 240 *Eucalyptus* tree genomes across 12 species. *New Phytol*. 206:1527–1540.

Solberg TR, Sonesson AK, Woolliams JA, Odegard J, Meuwissen TH. 2009. Persistence of accuracy of genome-wide breeding values over generations when including a polygenic effect. *Genet Sel Evol*. 41:53.

Speed D, Balding DJ. 2015. Relatedness in the post-genomic era: is it still useful? *Nat Rev Genet*. 16:33–44.

Squillace A. 1974. Average genetic correlations among offspring from open-pollinated forest trees. *Silvae Genet*. 23:149–156.

Suontama M, Klápště J, Telfer EJ, Graham NJ, Stovold GT, Low CB, McKinley RB, Dungey HS. 2018. Efficiency of genomic prediction across two *Eucalyptus nitens* seed orchards with different selection histories. *Heredity (Edinb)*. doi:10.1038/s41437-018-0119-5

Tambarussi EV, Pereira FB, Müller da Silva PH, Lee D, Bush D. 2018. Are tree breeders properly predicting genetic gain? A case study involving *Corymbia* species. *Euphytica*. 214:150.

Thistlethwaite FR, Ratcliffe B, Klápště J, Porth I, Chen C, Stoehr MU, El-Kassaby YA. 2017. Genomic prediction accuracies in space and time for height and wood density of Douglas-fir using exome capture as the genotyping platform. *BMC Genomics*. 18:930.

VanRaden PM. 2008. Efficient methods to compute genomic predictions. *J Dairy Sci*. 91:4414–4423.

Vidal M, Plomion C, Harvengt L, Raffin BC, Bouffier L. 2015. Paternity recovery in two maritime pine polycross mating designs and consequences for breeding. *Tree Genet Genomes*. 11:105.

Visscher PM, Medland SE, Ferreira MA, Morley KI, Zhu G, Cornes BK, Montgomery GW, Martin NG. 2006. Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS Genet*. 2:e41.

Vitezica ZG, Aguilar I, Misztal I, Legarra A. 2011. Bias in genomic predictions for populations under selection. *Genet Res (Camb)*. 93:357–366.

Wright S. 1922. Coefficient of inbreeding and relationship. *Am Nat*. 56:330–338.

Zapata-Valenzuela J, Whetten RW, Neale D, McKeand S, Isik F. 2013. Genomic estimated breeding values using genomic relationship matrices in a cloned population of loblolly pine. *G3 (Bethesda)*. 3:909–916.