Genetics: Early Online, published on October 15, 2014 as 10.1534/genetics.114.171322

Title: Unraveling Additive from Non-additive Effects using Genomic Relationship Matrices

Authors: Muñoz, P.[1,3*]; Resende Jr., M.F.R.[2,3*]; Gezan, S. A.[3]; Resende, M.D.V.[4,5]; de los Campos, G.[6]; Kirst, M.[3,7]; Huber, D[3]; Peter, G.F.[3,7]

[1]Agronomy Department, University of Florida, P.O. Box 110965, Gainesville, FL 32611, USA

[2]Genetics and Genomics Graduate Program, University of Florida, P.O. Box 103610, Gainesville, FL 32611, USA

[3]School of Forest Resources and Conservation, University of Florida, P.O. Box 110410, Gainesville, FL 32611, USA

[4]EMBRAPA Forestry, Estrada da Ribeira, km 111 Caixa Postal 319, Colombo, PR 83411-000 Brazil;

[5]Department of Forest Engineering, Universidade Federal de Viçosa, Viçosa, MG 36571-000 Brazil

[6] Department of Biostatistics, University of Alabama at Birmingham, Birmingham, AL 35294, USA

[7]University of Florida Genetics Institute, University of Florida, P.O. Box 103610, Gainesville, FL 32611, USA

[*]Both authors contributed equally to this work

Authors for correspondence:

Gary F. Peter

School of Forest Resources and Conservation

PO Box 110410

University of Florida

Gainesville, FL 32611-0410

Email: gpeter@ufl.edu

T: 352 846-0896

Patricio R. Munoz

Agronomy Department

PO Box 110965

University of Florida

Gainesville, FL 32611-0965

Email: p.munoz@ufl.edu

T: 352 392-6180

Keywords: Genomic selection, G-BLUP, non-additive, realized relationship matrices, dominance relationship matrix

ABSTRACT

The application of quantitative genetics in plant and animal breeding has largely focused on additive models, which may also capture dominance and epistatic effects. Partitioning genetic variance into its additive and non-additive components using pedigree-based models (P-BLUP) is difficult with most commonly available family structures. However, the availability of dense panels of molecular markers makes possible the use of additive and dominance realized genomic relationships for the estimation of variance components and the prediction of genetic values (G-BLUP). We evaluated height data from a multi-family population of the tree species *Pinus taeda* with a systematic series of models accounting for additive, dominance and first order epistatic interactions (additive-by-additive, dominance-by-dominance, and additive-by-dominance), using either pedigree- or marker-based information. We show that, compared with the pedigree, use of realized genomic relationships in marker-base models yields a substantially more precise separation of additive and non-additive components of genetic variance. We conclude that the marker-based relationship matrices in a model including additive and non-additive effects performed better, improving breeding value prediction. Moreover, our results suggest that, for tree height in this population, the additive and non-additive components of genetic variance are similar in magnitude. This novel result improves our current understanding of the genetic control and architecture of a quantitative trait and should be considered when developing breeding strategies.

INTRODUCTION

Quantitative genetics and its applications in plant and animal breeding have largely focused on additive models. Under idealized conditions, such as those described by Cockerham (1954) and Kempthorne (1954), genetic values due to additive and non-additive effects are orthogonal. However, these conditions are often not met in breeding populations, with the consequence that genetic values due to additive and non-additive effects may be confounded. Under these conditions, a large proportion of variance due to interactions of alleles (dominance and epistasis) can manifest as additive variance (Hill *et al.* 2008). For the same reason, with most commonly used family structures, it is difficult to dissect genetic variance into additive, dominance and epistatic effects. With standard pedigree models, variance estimates of these elements are highly correlated, reflecting confounding effects (Lynch and Walsh 1998; Hill 2010). The proportion of additive variance attributable to interactions of alleles largely depends on the distribution of allele frequencies at causal loci (Hill *et al*. 2008; Lu *et al.* 1999; Zuk *et al.* 2012). This affects the estimation of variance components and breeding value (BV) predictions (Palucci *et al.* 2007; Vanderwerf and Deboer 1989), as well as the ability to dissect the genetic architecture of the trait at the causal level. Understanding the genetic architecture of a trait is also useful for defining breeding strategies and for maximizing genetic gains. For instance, individual genetic differences due to non-additive effects can be exploited by designing mating schemes that maximize favorable allelic combinations, particularly if family or clonal propagation are possible in the breeding program.

Separation of additive and non-additive genetic components with standard pedigree-based models requires specific family structures, which are commonly available in plant or animal breeding programs. In practice, estimation of variance due to dominance and additive effects involves mating designs with large numbers of close, typically full-sib relatives. Partitioning epistasis requires, in addition, either inbreds or vegetatively propagated (clonal) populations. In perennial plants inbreds are not used because of their long generation time and because severe inbreeding depression often occurs. Thus, clonal populations are an alternative to explore the full genetic architecture in these species (Foster and Shaw 1988). Several studies aimed at partitioning genetic variance into its various components detected small dominance and negligeable epistatic effects (Araujo *et al.* 2012; Baltunis *et al.* 2007, 2008, 2009; Costa e Silva *et al.* 2004, 2009; Foster and Shaw 1988; Isik *et al.* 2003, 2005; Mullin *et al.* 1992; Wu 1996). These results don't necessarily imply that such effects are not important. Instead, the contribution of non-additive effects may be masked by effects due to the distribution of allele frequencies (e.g. Hill *et al.* 2008). These results may also reflect the limitations imposed by the data/family structure available, or the genetic information used (pedigrees), which only allows for estimation of the expected degree of genetic similarity.

Genome-wide genotypic data can identify, with a high level of certainty, the actual fraction of allele sharing between pairs of individuals. In pedigree-based genetic relationships, each element in the numerator relationship matrix (**A** matrix) is defined as the expected fraction of shared alleles assuming an infinitesimal model. However, due to Mendelian sampling, the values of the realized genomic relationships ($\mathbf{A_G}$ matrix),

constructed from molecular marker information deviate from their expected value (Hill and Weir 2011; VanRaden 2008). One way of incorporating molecular marker information for prediction of genetic values consists of replacing, in a BLUP analysis, the pedigree-based relationship matrices (P-BLUP) with marker-based counterparts (known as Genomic Best Linear Unbiased Predictor or G-BLUP) (VanRaden 2008). Indeed, G-BLUP is one of the most frequently used methods that combine molecular information to predict BVs and has shown remarkably good predictive performance in animal and plant breeding populations (de los Campos *et al.* 2012; Habier *et al.* 2010; Hayes *et al.* 2009; Heslot *et al.* 2012; Veerkamp *et al.* 2011).

Genomic BLUP is a well-known and easily-understood methodology. In the context of genome-wide selection (GWS), it is equivalent to ridge regression BLUP (RR-BLUP) (de los Campos *et al.* 2012; VanRaden 2008). Similar to P-BLUP, G-BLUP can be extended to account for non-additive effects by replacing pedigree-based relationship matrices due to non-additive effects (Mrode 2005), with their marker-based counterpart. This is because dominance and epistatic interaction (e.g. additive-by-additive, dominance-by-dominance and additive-by-domiance) relationship matrices can also be constructed using molecular information, as is currently done with $\mathbf{A_G}$. Use of dominance and epistasis matrices of realized genetic relationships may increase the precision of estimates derived from data in poorly structured populations, and may also increase the power to dissect genetic variance into components due to main and interaction effects.

Evidence indicates that G-BLUP based on $\mathbf{A_G}$ yields more accurate predictions of breeding value and of future phenotypes than its pedigree-based counterpart (**A**)

(Crossa *et al*. 2010; de los Campos *et al*. 2009; Hayes *et al*. 2009; Heslot *et al.* 2012; Munoz *et al.* 2013; Resende *et al.* 2012b; VanRaden 2008). This suggests that use of the realized genomic similarity ($\mathbf{A_G}$) increases (relative to **A**) the ability of the model to uncover the genetic components of the phenotypic data. However, it is not clear whether the power to partition genetic variance into additive and non-additive components can also be improved by the use of the realized genomic relationships. If so, this would lead to a finer dissection of the genetic architecture of complex traits that could have profound impacts on the future design and implementation of breeding strategies. The objective of this study is to assess the extent to which the use of marker-based additive and non-additive relationship matrices improves the precision of partitioning genetic variance into its components. For this assessment, tree height from a clonal population of *Pinus taeda* L. was evaluated with a series of models that account for additive, dominance and 1st order epistatic interactions (additive-by-additive, dominance-by-dominance, and additive-by-dominance) implemented with either pedigree or molecular marker information.

MATERIALS AND METHODS

Data: Field data from a single experimental trial from the CCLONES population (see Baltunis *et al.* 2007 and Resende *et al.* 2012a for details) was used in this study. The response variable total tree height (HT, m) was used. The population was generated by crossing 32 parents in a circular mating design with additional off-diagonal crosses, resulting in 70 full-sib families with an average of 13.5 individuals per family. Each individual was clonaly replicated (ramet) and a clonal field trial was established

using single-tree plots with eight replicates (one ramet per replicate), in a resolvable alpha-incomplete block design (Williams *et al.* 2002). Four of the replicates were grown under high intensity management while the rest were under an operational like regime.

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

Relationship matrices: A marker-based additive relationship matrix ($\mathbf{A_G}$) was constructed following the method described by Yang *et al.* (2010). The pairwise relationship for individuals *j* and *k* was defined by

$$A^*_{G_{jk}} = \begin{cases} \frac{1}{m}\sum_i \frac{(w_{ij}-2p_i)(w_{ik}-2p_i)}{2p_i(1-p_i)}, j \neq k \\ 1 + \frac{1}{m}\sum_i \frac{w_{ij}^2-(1+2p_i)w_{ij}+2p_i^2}{2p_i(1-p_i)}, j = k \end{cases}$$

where *m* is the total number of markers, *w* is an indicator variable representing the number of copies of a given allele and p_i is the observed allele frequency of the i-th SNP. To reduce the sampling variance of the entries of $\mathbf{A}_G^*$, we expanded the formula proposed by Yang *et al.* (2010) and adjusted each value of the $\mathbf{A}_G^*$ matrix by shrinking it towards it's expectation. The the adjusted $\mathbf{A}_G$ was obtained as

$$A_{G_{jk}} = (1 - \frac{1/m}{var(C_{jk})})\left(A^*_{G_{jk}} - A_{jk}\right) + A_{jk}$$

where C_{jk} represents all values of $\mathbf{A_G^*}$ that belong to the same class in A_{jk} (e.g. full sibs individuals where A_{jk} = 0.5). The resulting $\mathbf{A_G}$ was used to correct the original pedigree as previously detailed (Munoz *et al.* 2013), and it was verified that the estimated

genomic coefficients and their standard deviations were within expectations according to Simeone *et al*. (2011).

In addition, a molecular marker-based dominance relationship matrix ($\mathbf{D_G}$) was constructed. To build a dominance relationship matrix, we created an incidence matrix (**S**) for effects due to dominance $\mathbf{S} = \{s_{ij}\}$, where s_{ij} was parameterized to be coded 1 if the genotype was heterozygous and 0 if the marker genotype was homozygous for either class. The matrix **S** was further standardized to have mean zero by using:

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

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

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

Using the above we expanded the theory for the $\mathbf{A_G}$ matrix to construct $\mathbf{D_G}$ as

$$\boldsymbol{D_G} = \frac{\boldsymbol{SS'}}{\sum_{j=1}^{m} 2p_j q_j (1 - 2p_j q_j)}$$

where the denominator is the sum of the variances of s_{ij} under Hardy-Weinberg equilibrium, and the other terms were previously defined. This extension to construct $\mathbf{D_G}$ was also used by Su *et al.* (2013). Another parameterization has been proposed for the dominance genomic relationship matrix (Vitezica *et al.* 2013), which generates a different partition of the genetic variance. We also evaluated this parameterization and included the results in the supporting material.

Pedigree-based relationship matrices for additive (**A**) and dominance (**D**) effects were computed using standard methods (Lynch and Walsh 1998; Mrode 2005). Following existing theory (Cockerham 1954; Gianola and de los Campos 2008; Henderson 1985; Kempthorne 1954), the co-variance matrices due to 1st degree epistatic terms were computed using Hadamard products (i.e. cell-by-cell product

denoted as #) of the following form: (i) additive-by-additive interactions (**A#A** or $\mathbf{A_G}$#$\mathbf{A_G}$); (ii) dominance-by-dominance interactions (**D#D** or $\mathbf{D_G}$#$\mathbf{D_G}$); and (iii) additive-by-dominance interactions (**A#D** or $\mathbf{A_G}$#$\mathbf{D_G}$) for pedigree and marker-based methods, respectively.

Genetic analyses: All analyses were carried out in the software ASReml v3.0 (Gilmour *et al.* 2009), which fits mixed models with complex datasets using sparse matrix methods. ASReml is equipped with the Residual Maximum Likelihood (REML) for variance component estimation using the average information algorithm (Gilmour *et al.* 1995).

Five models were fit using the pedigree-based matrices (models 1 to 5) and five using the marker-based matrices (models 6 to 10). These models range from a simple additive model to a full model including additive, dominance and epistatic effects. The full model (i.e. model 5 or 10) is described below:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z_1}\boldsymbol{i} + \mathbf{Z_2}\boldsymbol{a} + \boldsymbol{Z_3 t_1} + \boldsymbol{Z_4 d} + \boldsymbol{Z_5 t_2} + \boldsymbol{Z_6 i_{c\#c}} + \boldsymbol{e}$$

where $\mathbf{y}$ is the phenotypic HT response, $\boldsymbol{\beta}$ is a vector of fixed effects (i.e. silvicultural treatment and replicate), $\boldsymbol{i}$~N($\mathbf{0}$, $\mathbf{I}\sigma^2_{\mathrm{i}}$) is a vector of the random incomplete block effects within replication, $\boldsymbol{a}$~N($\mathbf{0}$, $\mathbf{C_1}\sigma^2_{\mathrm{a}}$) is a vector of random additive effects of individuals and $\mathbf{C_1}$ is a relationship matrix due to additive effects either from pedigree (**A**) or markers ($\mathbf{A_G}$), $\boldsymbol{t_1}$~N($\mathbf{0}$, $\mathbf{C_1}\otimes\mathbf{I}\sigma^2_{\mathrm{t1}}$) is a vector of random additive by silviculture type interactions, $\boldsymbol{d}$~N($\mathbf{0}$, $\mathbf{C_2}\sigma^2_{\mathbf{d}}$) is a vector of random individual dominance effects and $\mathbf{C_2}$ is a relationship matrix due to dominance effects that was computed either from pedigree (**D**) or markers ($\mathbf{D_G}$), $\boldsymbol{t_2}$~N($\mathbf{0}$, $\mathbf{C_2}\otimes\mathbf{I}\sigma^2_{\mathrm{t2}}$) is a vector of random dominance by silviculture type interactions, $\boldsymbol{i_{c\#c}}$~N($\mathbf{0}$, $\mathbf{C_1}\#\mathbf{C_1}\sigma^2_{\mathrm{iaa}}$) is either a vector of random additive by additive

interaction, a vector of random dominance by dominance interactions $\boldsymbol{i}_{\boldsymbol{c}\#\boldsymbol{c}}$~N(**0**, $\mathbf{C_2}\#\mathbf{C_2}\sigma^2_{idd}$) or a vector of random additive by dominance interactions $\boldsymbol{i}_{\boldsymbol{c}\#\boldsymbol{c}}$~N(**0**, $\mathbf{C_1}\#\mathbf{C_2}\sigma^2_{iad}$) and $\boldsymbol{e}$~N(**0**, $\mathbf{I}\sigma^2_e$) is a vector of random residual effects. Above, matrices **X** and $\mathbf{Z_1}$-$\mathbf{Z_6}$, are incidence matrices for fixed and random effects, respectively, and **I** denotes an identity matrix, $\otimes$ and # represent the Kronecker and Hadamard (cell-by-cell) product, respectively.

Under the above model, the narrow sense-heritability can be estimated as $h^2 = \frac{\sigma_a^2}{\sigma_p^2}$, the dominance to total variance ratio as $d^2 = \frac{\sigma_d^2}{\sigma_p^2}$, the epistatic to total variance ratio as $i^2 = \frac{\sigma_i^2}{\sigma_p^2}$, and the broad-sense heritability as $H^2 = \frac{\sigma_g^2}{\sigma_p^2}$. The σ_a^2 is the estimated additive variance, σ_d^2 is the estimated dominance variance, and σ_p^2, σ_i^2 and σ_g^2 are the total phenotypic, epistatic and total genetic variance, respectively, that changed accordingly to the model being fit (Table 1).

Model comparisons: Models were compared using the Akaike Information Criterion (AIC, Akaike 1974). Precision of variance components estimates, and their dependency, was assessed using the asymptotic variance-covariance matrix of estimates of variance parameters (**V**). The asymptotic sampling correlation matrix of estimates (**F**) was computed as $\boldsymbol{F} = \boldsymbol{L}^{-1/2}\boldsymbol{V}\boldsymbol{L}^{-1/2}$, where **L** is a diagonal matrix containing the diagonal elements of **V**. Inspection of the off-diagonal elements of the **F** matrix allows assessing sampling correlation of variance estimates. In order to have an overall assessment of dependency between the estimates, eigenvalues of **F** were examined. The standard error of the prediction (SEP) was estimated for each model as the square root of the prediction error variance (PEV), which is obtained by extracting

the elements of the diagonal of the generalized inverse of the coefficient matrix from the linear mixed model equations (left hand side), and scaled by the error variance. In short form, the PEVs correspond to the $Var(a - \hat{a})$ (Mrode, 2005, p. 51).

Predictive ability and stability of the models in estimating breeding and genetic values were evaluated. The predictive ability of a model's breeding value was defined as the correlation between the estimated breeding value and the phenotypic average of all the ramets (clones). These values were calculated when all the data was used without cross validation (BV-all). The predictive ability of a model's total genetic value (sum of BV-all, dominance effect and epistatic effect) was defined as the correlation between the predicted total genetic value and the phenotype average of all the clones using all the data without cross validation (GV-all). Prediction models were assessed under cross validation (Kohavi, 1995) to obtain predicted breeding value (BV-cv) and predicted total genetic value (GV-cv) with a random sub-sampling partitioning, fixed for all models. The stability of the predictive models were evaluated as the correlation between the BV-all and BV-cv, and between GV-all and GV-cv and was defined as a measure of the dependency of the predictive breeding value on the phenotype. The mean square error (MSE) was calculated between BV-all and BV-cv within each model using standard methods. Finally, the capacity of the model to predict ranking position of the top 10% of the individuals, simulating a selection scenario, was evaluated as the correlation between the ranking position using the BV-all and the ranking position using the BV-cv.

RESULTS

The genetic parameters and goodness-of-fit statistics, estimated for each model, are summarized in Table 2. Both P_A and M_A models had narrow-sense heritability (h2) higher than 0.30. After including the dominance effect in the pedigree-based model (P_AD), h^2 decreased by approximately 26% and the dominance ratio (d^2) estimate was small (0.06) and non-significant ($2\times SE(d^2) > 0.06$). When the dominance effect was included with the molecular markers-based model (M_AD), the h^2 decreased 47%, to 0.20, and d^2 increased to 0.12. With the M_AD model the dominance variance represents 60% of the additive value and 39% of the total genetic variation. We further extended these models to include the additive-by-additive, dominance-by-dominance and additive-by-dominance first-order epistatic interaction factors in three separate models. In pedigree-based models, P_A#A, P_D#D and P_A#D, the estimations of variance components for additive and dominance varied only slightly from those of the P_AD model (Table S1). Moreover, epistasis estimates were zero in all three models. When the additive-by-additive, dominance-by-dominance and additive-by-dominance interactions were added (models M_A#A, M_D#D and M_A#D), the narrow-sense heritability dropped by more than 30% and the dominance ratio by 80%, compared to the M_AD model. The epistatic ratio (i^2) was estimated at 0.15, 0.12 and 0.14 for the M_A#A, M_D#D and M_A#D models, respectively (Table 2). The alternative parameterization for the dominance genomic relationship matrix proposed by Vitezica *et al.* (2013) showed similar results regarding the partition of additive and non-additive effects (Table S2).

Goodness-of-fit statistics show that inclusion of non-additive effects improved slightly the model fit for pedigree-based models and substantially for marker-based

models (Table 2). The marker-based models M_A#D and M_D#D yielded the best fit of the data, however, fitting differences among the more complex models was small. Thus, the dependency of the random component estimates was evaluated to further differentiate the best model.

We studied the sampling correlation among the variance component estimates, to assess which of the nine models shows less dependency and thus partitioned the genetic variance better (Table S3). Figure 1 shows the cumulative proportion of variance explained by high order eigenvalues of the sampling variance-covariance matrix of estimates derived from models including additive plus dominance, additive-by-additive, dominance-by-dominance and additive-by-dominance epistatic interactions, for pedigree- and marker-based models. As reference, the distribution of eigenvalues for a perfect orthogonal correlation matrix, representing the ideal model (all of the eigenvalues equal to 1) is included. In all cases, the marker-based cumulative distributions are closer to the orthogonal distribution, suggesting less dependency between estimates of variance components. Indeed, the sampling correlation between estimates of variance components due to additive and dominance effects decreases in absolute value from 0.90 with the P_AD to 0.70 with the M_AD model (Table S2). In general, all the marker-based models that include epistasis outperform their pedigree-based counterpart (Figure 1b,c,d). Models M_D#D and M_A#D showed the smallest sampling correlations between additive and dominance/epistasis, with absolute correlation values below 0.45 (Table S3).

The standard error of the predictions (SEP) of BV and dominance value (DV) were compared for the pedigree and markers models including additive-by-additive,

dominance-by-dominance and additive-by-dominance (Figure 2). Values below the 45 degree reference line indicate that marker-based models have smaller SEPs. The SEPs for BV's from the marker-based models were smaller than the pedigree-based models in 99.2% of the cases (Figure 2a,b,c). In the case of the SEPs of DVs, a clear advantage was observed for marker-based models (y-axis) over pedigree-based models (x-axis), with SEP on average 52% lower for the marker-based models (Figure 2d,e,f).

The predictive ability of breeding value and genetic value for the pedigree-based and marker-based models are shown in Table 3. The highest predictive ability for BV was obtained with the pedigree additive model (P_A). A slight decrease in the BV prediction ability was observed when non-additive effects were included in the pedigree-based model (0.86), and a much larger decrease was observed for the marker-based models (0.76). All models evaluated reached similar GV predictive ability.

Predictive stability can be viewed as a measure of how much the prediction of the breeding value and genetic value using all the data (BV-all and GV-all) depend on the individual phenotype (Table 3). Predictions based on models with markers are more stable than those derived from pedigree models (3% increase when comparing M_A to P_A). In the pedigree-based models, inclusion of non-additive effects increased the stability to predict BV by 13%, 14% and 14% for P_A#A, P_D#D and P_A#D, respectively, while inclusion of non-additive effects in the more complex marker-based models, increased the BV prediction stability by over 29% when compared with the M_A and by over 33% when compared to P_A. The Mean Square Error (MSE) decreased approximately by 50% from the additive models (P_A) to the more complex pedigree-

based models. The addition of non-additive effects to the marker-based bodels decreased the MSE even further by more than 68% and up to 94% decrease in the case of model P_A#A (Table 3).

In a breeding program it is important to predict the trend and magnitude of the complete set of individuals in the population; however, it is often more important to predict the best performing individuals (potential selections). Here we ranked all individuals based on BV-all and BV-cv and evaluate the ranking correlation of the top 10%, emulating the selection of the top 10% of genotypes (Table 3). When the pedigree-based matrix was replaced by the marker-based matrix in the additive models (P_A and M_A), the capacity to predict the top 10% remained the same. However, this capacity increased substantially for the more complex marker-based models where the predictive stability of the top 10% genotypes raised 82-170% (Table 3).

DISCUSSION

Here we assessed the use of marker and pedigree-based models to separate additive from non-additive variances for height, in a structured population of loblolly pine. We showed that the two approaches are dramatically distinct in their capactity to properly partition the genetic variance into its various components, with marker-based models being significantly more effective in accounting for non-additive variances. In the pedigree-based models, inclusion of non-additive effects decreased the estimated narrow-sense heritability by 26%. This result is expected because depending on the distribution of allele frequencies, a sizable proportion of variance due to non-additive effects can be manifest as additive variance (Lu et al. 1999; Zuk et al. 2012). In marker

(pedigree) models 71% (57%) of the decrease of in additive varinance was captured by the dominance variance, suggesting that indeed, dominance is making a substantial contribution to the estimated additive variance obtained when dominance is ignored. This phenomena has been postulated theoretically (e.g., Falconer and Mackay, page 126) and observed in multiple studies (Pante et al. 2002; Rodriguez-Almeida et al. 1995; Wei and van der Werf 1993; Winkelman and Peterson 1994). In addition, when pedigree-based models included non-additive effects, the conclusions were not different from what has been commonly observed that non-additive effects represent a small fraction of the total genetic variation (Araujo *et al.* 2012; Baltunis *et al.* 2007; Costa e Silva *et al.* 2004; Isik *et al.* 2003). In contrast, marker-based models with additive and non-additive effects yield a substantially different variance partitioning than their counterparts using the pedigree models. The additive variance decreased as dominance was included in the model and it further decreased when dominance and epistasis were considered. These models indicate that, for this population and trait, non-additive effects are as important as additive effects, and dramatically larger than predicted by the pedigree-based matrices. These changes in the magnitude of variance components have already been observed when the relationship matrix derived from markers is used instead of the pedigree-based relationship matrix, in the context of additive genomic selection models (Lee *et al.* 2010).

The value for AIC varied modestly for the best models, with no clear advantage of one model relative to the others. This is not surprising, if additive effects capture part of the effects due to dominance and epistasis, the additive model should not suffer much if these components are omitted. However, these models varied considerably in

the partitioning of the genetic variance components, thus changing not only the inference but also the potential decisions taken in the breeding strategy.

We also assessed the dependency of the random effects estimates to discriminate the best model, given the small differences for AIC in the different models. The level of confounding between components was very different in the pedigree- and marker-based models. The most unambiguous dissection of the genetic variance occurs when estimates of variance components are uncorrelated, i.e. the sampling correlation among the model effects will be closer to zero and all eigenvalues of this correlation matrix close to one (Hill 2010). In the models that included additive and dominance, and additive, dominance and epistatic effects, the correlation matrices indicated that those derived from molecular markers partitioned the genetic effects more precisely although the partition is still not fully orthogonal. The parameterizations of these paired models were identical, except for the origin of the relationship matrices (pedigree- or marker-based). The limited capacity of pedigree-based models to partition these components is not surprising, as all relationship matrices are derived from the pedigree additive relationship matrix (Mrode 2005) and, therefore, are strongly correlated (Visscher 2009). The models M_A#A, M_D#D and M_A#D had the lowest correlation between additive and non-additive, showing a partition substantially better than that of pedigree-based models (Table S3). These results support the finding that pedigree-based models are inadequate in separating the additive from non-additive effects, as their results are comparable to those of additive models (Hill *et al.* 2008). On the other hand, the use of the matrix derived from markers has already been related to a better capacity to separate random effects in a model (Lee *et al.* 2010). Thus, we conclude that the use of

the marker-based relationship matrices increase substantially the capacity to separate additive and non-additive genetic effects.

Assessment of prediction accuracy further support the conjecture that in pedigree-based models, additive components can capture a large proportion of the variance due to interaction terms (Hill *et al.* 2008). Consequently, there is limited gain in this scenario, by including non-additive effects. On the other hand, for marker-based models, the ability to predict the mean phenotype with the BV decreased when non-additive effects were included in the model, and the maximum predictability (0.89) was only reached when additive and non-additive values were considered together (GV). This indicates that pedigree-based models potentially overestimate the additive effects, which is likely to be due to an inflated additive variance estimate that also represents some of the non-additive components. Inflation from epistasis, for example, falls apart as recombination breaks down favourable combinations of alleles. This is a problem for breeding programs because overestimates of BV inflates genetic gains, but the portion due to non-additive effects is transient and cannot be captured if controlled sexual reproduction is used. Additionally, the genetic architecture of the trait will be predicted to be simpler than it actually is.

In breeding programs the true breeding value is never known. Thus, the prediction models including all the available data (BV-all) is usually used as the best BV estimation. We evaluated the stability of BV estimates by comparing the results obtained with all data, with the results from cross validation for pedigree- and marker-based models. This is a measure of the influence of an individual's phenotype on the predicted breeding value. We observed that models with non-additive relationship

matrices are more stable and produce estimates of breeding values in independent sets that are more similar to the BV-all. The inclusion of non-additive relationship matrices yields models that predict BVs more stably than additive pedigree-based (Animal Model) and marker-based models (traditional G-BLUP). In addition, in this cross-validation scheme the MSE of the model M_A#A decreased more than 15- and 8-fold when compared with both additive models and full pedigree-based models, respectively. These results indicate that, for this trait, a considerable increase in the stability cannot be reached simply by replacing the **A** matrix by the $\mathbf{A}_G$ matrix but also needs to incorporate non-additive effects in the model.

Overall, our study supports the hypothesis that additive effects can capture a large proportion of the genetic variance from dominance and epistasis. This is in part due to the fact that, in breeding populations, additive and non-additive genetic components are not typically independent. However, we also show that with relationship matrices derived from markers, the genetic variances were partitioned more precisely than using only pedigree information. Moreover, our estimates suggest that in this population, for tree height, the additive and non-additive components of the genetic variance are similar in magnitude. While further research is needed in other species, traits and populations, we show that variance estimates can be inadequately estimated if only pedigree information is used. This study improves our current understanding of the genetic control and architecture of a quantitative trait and should be considered when developing effective breeding strategies.

ACKNOWLEDGEMENTS

The authors wish to thank members of the Forest Biology Research Cooperative (FBRC) at the University of Florida for their support in establishing, maintaining and measuring the field trial used in this study. The work was supported by the National Science Foundation Plant Genome Research Program (award no. 0501763), the Foundational Program (award no. 2013-67013-21159), the Department of Energy (award no. 2013-67009-21200) and the Plant Breeding and Education Program (award no. 2010-85117-20569) from the USDA National Institute of Food and Agriculture.

REFERENCES

Akaike, H. 1974. New look at statistical-model identification. Transactions on Automatic Control. AC19:716-723.

Araujo, J.A., N.M.G. Borralho and G. Dehon. 2012. The importance and type of non-additive genetic effects for growth in *Eucalyptus globulus*. Tree Genetics & Genomes. 8:327-337.

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

Baltunis, B.S., T.A. Martin, D.A. Huber and J.M. Davis. 2008. Inheritance of foliar stable carbon isotope discrimination and third-year height in *Pinus taeda* clones on contrasting sites in Florida and Georgia. Tree Genetics & Genomes. 4:797-807.

Baltunis, B.S., H.X. Wu, H.S. Dungey, T.J.T. Mullin and J.T. Brawner. 2009. Comparisons of genetic parameters and clonal value predictions from clonal

trials and seedling base population trials of radiata pine. Tree Genetics & Genomes. 5:269-278.

Cockerham, C.C. 1954. An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present. Genetics. 39:859-882.

Costa e Silva, J., N.M.G. Borralho, J.A. Araujo, R.E. Vaillancourt and B.M. Potts. 2009. Genetic parameters for growth, wood density and pulp yield in *Eucalyptus globulus*. Tree Genetics & Genomes. 5:291-305.

Costa e Silva, J., N.M.G. Borralho and B.M. Potts. 2004. Additive and non-additive genetic parameters from clonally replicated and seedling progenies of *Eucalyptus globulus*. Theoretical and Applied Genetics. 108:1113-1119.

Crossa, J., G.D.L. Campos, P. Perez, D. Gianola, J. Burgueno et al. 2010. Prediction of Genetic Values of Quantitative Traits in Plant Breeding Using Pedigree and Molecular Markers. Genetics. 186:713-724.

De Los Campos, G., H. Naya, D. Gianola, J. Crossa, A. Legarra et al. 2009. Predicting Quantitative Traits with Regression Models for Dense Molecular Markers and Pedigree. Genetics. 182:375-385.

De los Campos, G., J.M. Hickey, R. Pong-Wong, H.D. Daetwyler and M.P.L. Calus. 2012. Whole Genome Regression and Prediction Methods Applied to Plant and Animal Breeding. Genetics 193(2): 327-345.

Eckert, A.J., J. van Heerwaarden, J.L. Wegrzyn, C.D. Nelson, J. Ross-Ibarra, S.C. Gonzalez-Martinez 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.

Falconer D.S., T.F.C. Mackay. 1996. *Introduction to Quantitative Genetics*, Fourth Edition. Addison Wesley Longman, England.

Foster, G.S. and D.V. Shaw. 1988. Using clonal replicates to explore genetic-variation in a perennial plant-species. Theoretical and Applied Genetics. 76:788-794.

Gianola, D. and G.D.L. Campos. 2008. Inferring genetic values for quantitative traits non-parametrically. Genetics Research, 90, pp 525-540. doi:10.1017/S0016672308009890.

Gilmour, A.R., R. Thompson and B.R. Cullis. 1995. Average information REML: An efficient algorithm for variance parameter estimation in linear mixed models. Biometrics. 51:1440-1450.

Gilmour, A. R., B. J. Gogel, B. R. Cullis and R. Thompson.2009. ASReml User Guide Release 3.0.VSN International Ltd, Hemel Hempstead, UK.

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

Hayes, B.J., P.M. Visscher and M.E. Goddard. 2009. Increased accuracy of artificial

selection by using the realized relationship matrix. (vol 91, pg 47, 2009). Genetics Research. 91:143-143.

Henderson, C. R. 1985. Best Linear Unbiased Prediction of Nonadditive Genetic Merits in Noninbred Populations. Journal of Animal Science. 60:111-117.

Heslot, N., H.-P.Yang, M.E. Sorrells and J.-L. Jannink. 2012. Genomic Selection in Plant Breeding: A Comparison of Models. Crop Science. 52:146-160.

Hill, W. 2010.Understanding and using quantitative genetic variation. Philosophical Transactions of the Royal Society B-Biological Sciences. 365:73-85.

Hill, W., M. Goddard and P. Visscher. 2008. Data and theory point to mainly additive genetic variance for complex traits. Plos Genetics. 4(2): e1000008.

Hill, W. G., and B. S. Weir. 2011. Variation in actual relationship as a consequence of Mendelian sampling and linkage. Genetics Research. 93:47-64.

Isik, F., B. Goldfarb, A. LeBude, B.L. Li and S. McKeand. 2005. Predicted genetic gains and testing efficiency from two loblolly pine clonal trials. Canadian Journal of Forest Research-Revue Canadienne De Recherche Forestiere. 35:1754-1766.

Isik, F., B.L. Li and J. Frampton. 2003. Estimates of additive, dominance and epistatic genetic variances from a clonally replicated test of loblolly pine. Forest Science. 49:77-88.

Kempthorne, O. 1954. The correlation between relatives in a random mating population. Proc. R. Soc. Lond. B. Biol. Sci. 143:103–113.

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

Lee, S.H., M.E. Goddard, P.M. Visscher and J.H.J. van der Werf. 2010. Using the realized relationship matrix to disentangle confounding factors for the estimation of genetic variance components of complex traits. Genetics Selection Evolution. 42:22.

Lu, P.X., D.A. Huber and T.L. White. 1999. Potential biases of incomplete linear models in heritability estimation and breeding value prediction. Canadian Journal of Forest Research-Revue Canadienne De Recherche Forestiere. 29:724-736.

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

Mrode, R.A. 2005. Linear models for the prediction of animal breeding values. CABI Publishing Series.

Mullin, T.J., E.K. Morgenstern, Y.S. Park and D.P. Fowler. 1992. Genetic-parameters from a clonally replicated test of black spruce (*Picea mariana*). Canadian Journal of Forest Research-Revue Canadienne De Recherche Forestiere. 22:24-36.

Munoz, P, M.F. Resende, D.A. Huber, T. Quezada, M.D. Resende, D.B. Neale, J.L.Wegrzyn, M. Kirst and G.F. Peter. 2013. Genomic relationship matrix for correcting pedigree errors in breeding populations: impact on genetic parameters and genomic selection accuracy. Crop Science (*In Press*)

Pante, M.J., B. Gjerde, I. McMillan, I. Misztal. 2002. Estimation of additive and

dominance genetic variances for body weight at harvest in rainbow trout Oncorhynchus mykiss Oncorhynchus mykiss. *Aquaculture* 204:383-392.

Palucci, V., L.R. Schaeffer, F. Miglior and V. Osborne. 2007. Non-additive genetic effects for fertility traits in Canadian Holstein cattle. Genetics Selection Evolution. 39:181-193.

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

Resende, M.F.R., Jr., P. Munoz, J.J. Acosta, G.F. Peter, J.M. Davis, D. Grattapaglia, M.D.V. Resende and M. Kirst. 2012a. Accelerating the domestication of trees using genomic selection: accuracy of prediction models across ages and environments. New Phytologist. 193:617-624.

Resende, M.F., P. Muñoz, M.D. Resende, D.J. Garrick, R.L. Fernando, J.M. Davis, E.J. Jokela, T.A. Martin, G.F. Peter and M. Kirst. 2012b. Accuracy of Genomic Selection Methods in a Standard Data Set of Loblolly Pine (*Pinus taeda* L.). Genetics. 190:1503-1510.

Rodriguez-Almeida, F.A., Van Vleck, L.D., Wilham, R.L., Northcutt, S.L., 1995. Estimation of non-additive genetic variances in three synthetic lines of beef cattle using an animal model. J. Anim. Sci. 73, 1002–1011.

Su, G., O.F. Christensen, T. Ostersen, M. Henryon, and M.S. Lund. 2012. Estimating Additive and Non-Additive Genetic Variances and Predicting Genetic Merits Using Genome-Wide Dense Single Nucleotide Polymorphism Markers. PLoS

ONE 7(9): e45293. doi:10.1371/journal.pone.0045293

Vanderwerf, J.H.J. and W. Deboer. 1989. Influence of nonadditive effects on estimation of genetic-parameters in dairy-cattle. Journal of Dairy Science. 72:2606-2614.

VanRaden, P.M. 2008. Efficient Methods to Compute Genomic Predictions. Journal of Dairy Science. 91:4414-4423.

Veerkamp, R.F., H.A. Mulder, R. Thompson and M.P.L. Calus. 2011. Genomic and pedigree-based genetic parameters for scarcely recorded traits when some animals are genotyped. Journal of Dairy Science. 94:4189-4197.

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

Vitezica, Z.G., L. Varona and A. Legarra. 2013. On the Additive and Dominant Variance and Covariance of Individuals within the Genomic Selection Scope. Genetics. Early Online, published on October 11, 2013 as 10.1534/genetics.113.155176

Wei, M., van der Werf, J.H.J., 1993. Animal model estimation of additive and dominance variances in egg production traits of poultry. J. Anim. Sci. 71, 57–65.

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.

Winkelman, A.M., Peterson, R.G., 1994b. Genetic parameters heritabilities, dominance ratios, and genetic correlations for body weight and length of chinook salmon

after 9 and 22 months of saltwater rearing. Aquaculture 125, 30–36.

Wu, R.L. 1996. Detecting epistatic genetic variance with a clonally replicated design: Models for low- vs high-order non allelic interaction. Theoretical and Applied Genetics. 93:102-109.

Yang, J., B. Benyamin, B.P. McEvoy, S. Gordon, A.K. Henders, D.R. Nyholt, P.A. Madden, A.C. Heath, N.G. Martin, G.W. Montgomery, M.E. Goddard and P.M. Visscher. 2010. Common SNPs explain a large proportion of the heritability for human height. Nature Genetics. 42:565-U131.

Zuk, O., E. Hechter, S.R. Sunyaev and E.S. Lander. 2012. The mystery of missing heritability: Genetic interactions create phantom heritability. Proceedings of the National Academy of Sciences of the United States of America. 109:1193-1198.

FIGURE LEGENDS

Figure 1 – Cumulative proportion of variance explained by eigenvalues for models considering A plus D from pedigree (P_AD) versus markers (M_AD) (a), A#A interaction from pedigree (P_A#A) versus markers (M_A#A) (b), D#D interaction from pedigree (P_D#D) versus markers (M_D#D) (c), and considering A#D interaction from pedigree (P_A#D) versus markers (M_A#D) (d). The diagonal represents an orthogonal correlation matrix.

Figure 2 – Standard error of the prediction (SEP) for pedigree-based (X-axis) against their counterpart marker-based models (Y-axis). SEP for BV prediction model including A#A interaction (a), D#D interaction (b) and including A#D interaction (c). SEP for dominance value (DV) prediction model including A#A interaction (d), D#D interaction (e), and A#D interaction (f).

TABLES

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

Model		Relationship matrix used *(Information Used,* **A**,**D**=*Pedigree,* $\mathbf{A_G}$,$\mathbf{D_G}$=*Markers)*		
Number	Code	*Additive*	*Dominance*	*Epistasis*
1	P_A	**A**		
6	M_A	$\mathbf{A_G}$		
2	P_AD	**A**	**D**	
7	M_AD	$\mathbf{A_G}$	$\mathbf{D_G}$	
3	P_A#A	**A**	**D**	**A#A**
8	M_A#A	$\mathbf{A_G}$	$\mathbf{D_G}$	$\mathbf{A_G\#A_G}$
4	P_D#D	**A**	**D**	**D#D**
9	M_D#D	$\mathbf{A_G}$	$\mathbf{D_G}$	$\mathbf{D_G\#D_G}$
5	P_A#D	**A**	**D**	**A#D**
10	M_A#D	$\mathbf{A_G}$	$\mathbf{D_G}$	$\mathbf{A_G\#D_G}$

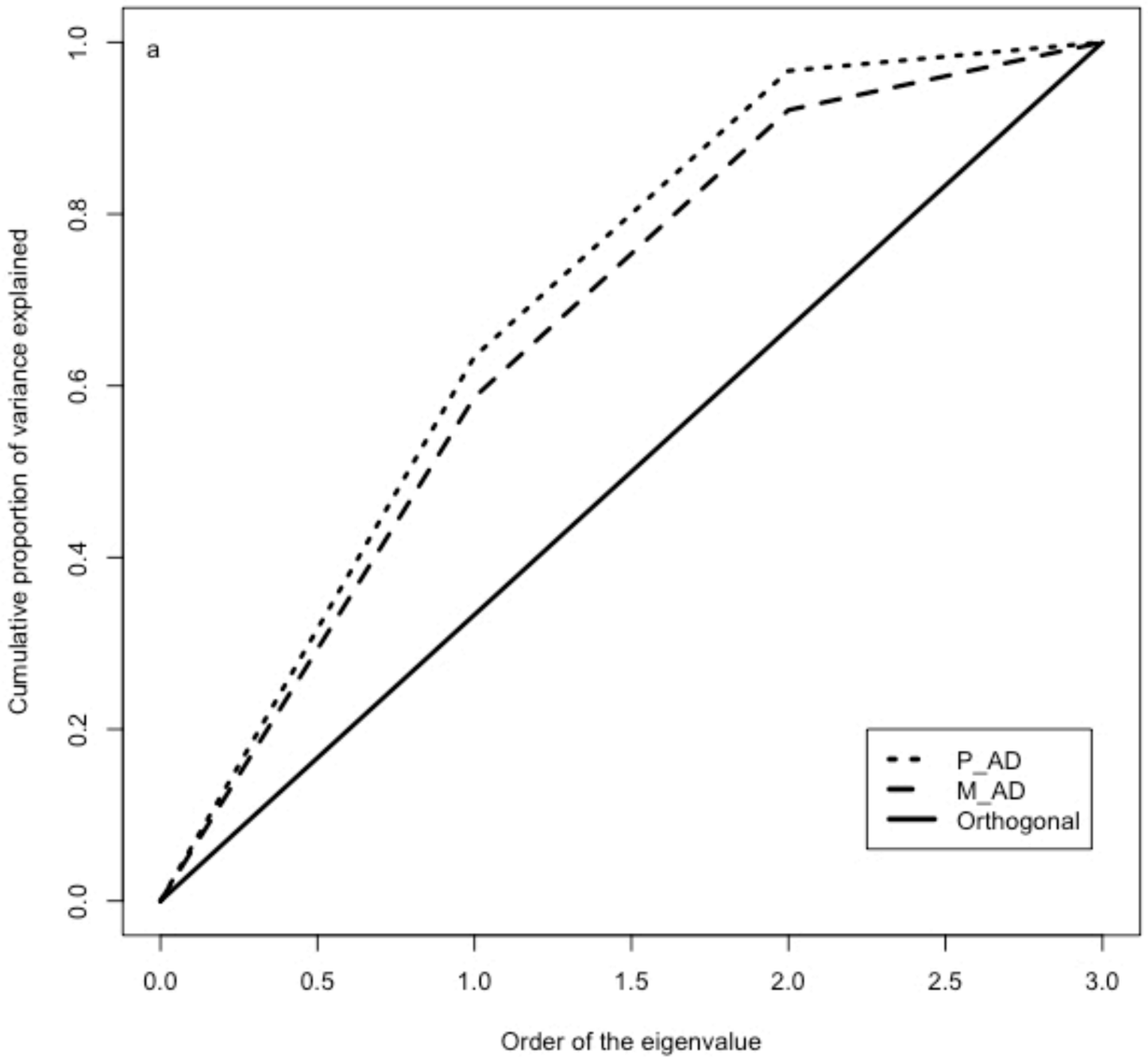
Table 2 – Estimates of genetic parameters (standard errors in parenthesis) and goodness-of-fit measures.

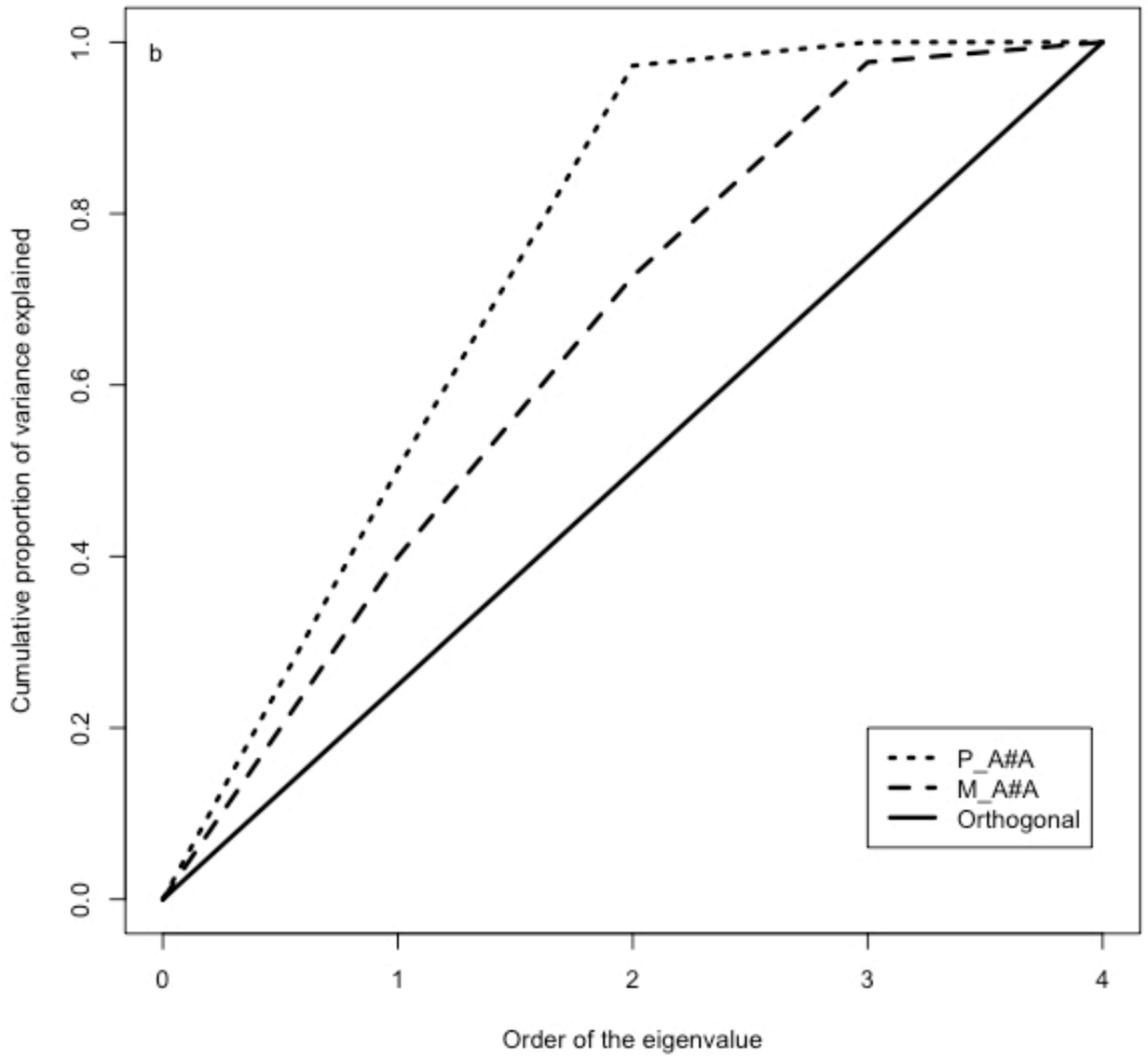
	P_A*	M_A*	P_AD*	M_AD*	P_A#A*	M_A#A*	P_D#D*	M_D#D*	P_A#D*	M_A#D*
h^2	0.32	0.347	0.235	0.199	0.233	0.088	0.228	0.139	0.231	0.1251
SE(h^2)	(0.017)	(0.018)	(0.047)	(0.035)	(0.047)	(0.044)	(0.046)	(0.036)	(0.047)	(0.038)
d^2			0.056	0.117	0.055	0.023	0.058	0.009	0.056	0.006
SE(d^2)	na	na	(0.033)	(0.029)	(0.033)	(0.034)	(0.032)	(0.033)	(0.043)	(0.035)
i^2					0.000	0.154	0.000	0.121	0.000	0.135
SE(i^2)	na	na	na	na	(0.000)	(0.038)	(0.000)	(0.028)	(0.000)	(0.031)
H^2	0.32	0.347	0.29	0.316	0.288	0.264	0.286	0.269	0.288	0.266
SE(H^2)	(0.017)	(0.018)	(0.021)	(0.018)	(0.021)	(0.019)	(0.021)	(0.018)	(0.021)	(0.018)
LogL	-1299.40	-1323.73	-1295.37	-1307.63	-1294.83	-1293.53	-1293.90	-1292.19	-1294.38	-1292.54
AIC	2606.80	2655.46	2602.74	2627.26	2605.66	2603.06	2603.80	2600.38	2604.76	2601.08

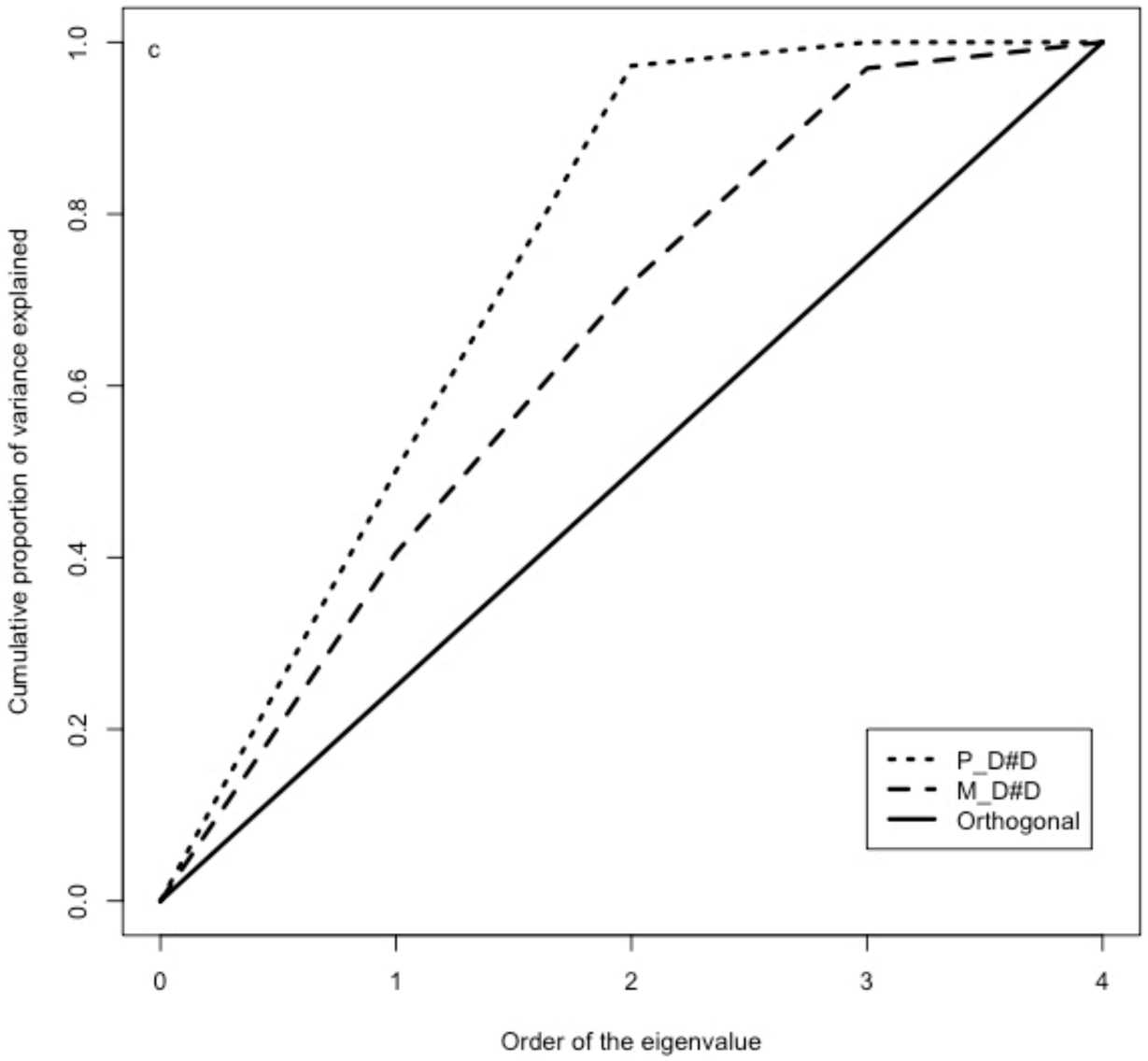
*Each column represents a different model. See Table 1 for model description

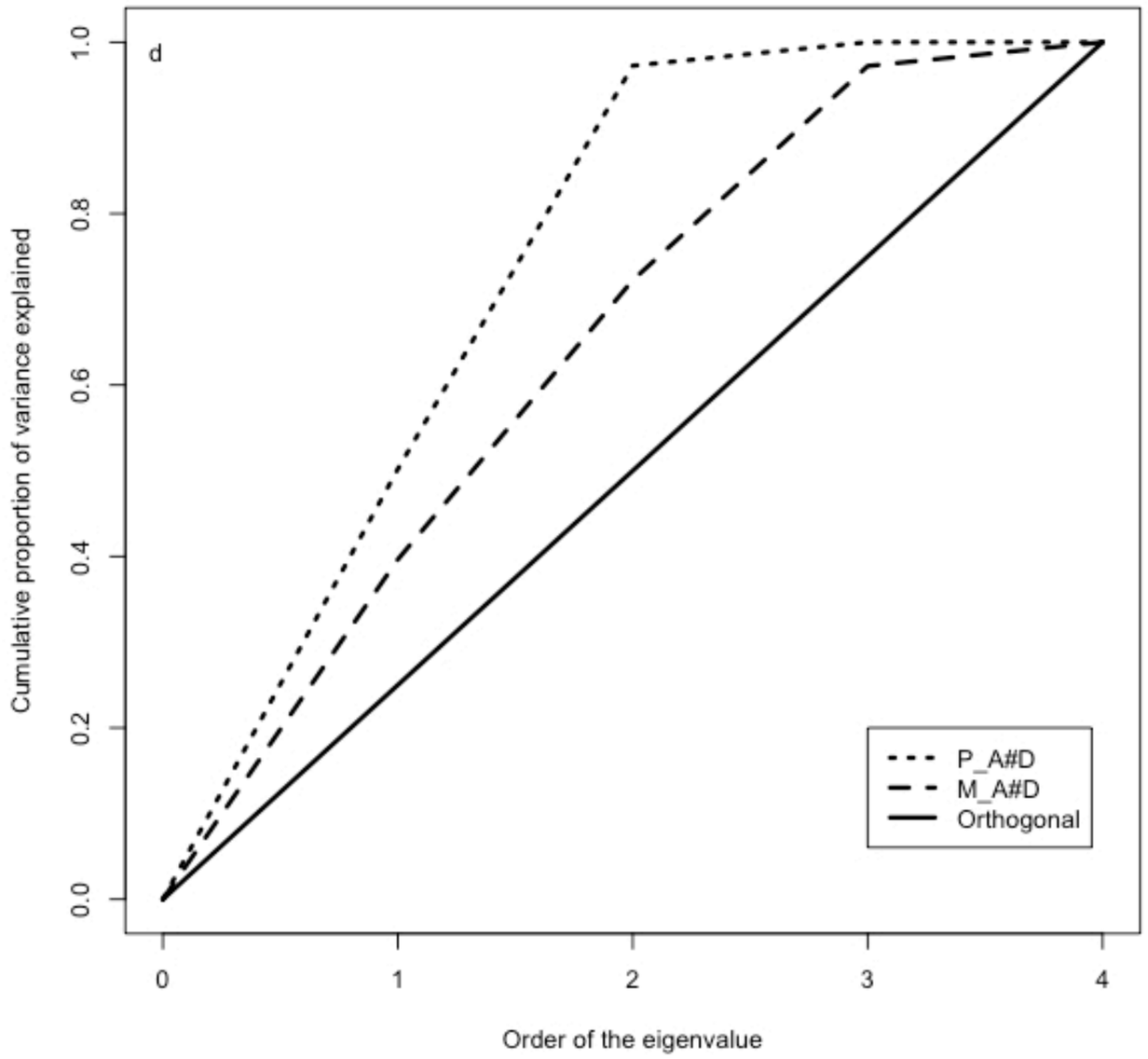
Table 3 – Model predictive ability; correlation between the phenotypic average of all the ramets (phe) and BV-all, and correlation between phe and total genetic value(GV-all=BV-all+DV+epistatic value). Stability in a cross-validation; correlation between BV-all and BV-cv, Mean Square Error (MSE) and correlation of ranking positions for the top 10% individuals (10% Rank cor(BV)).

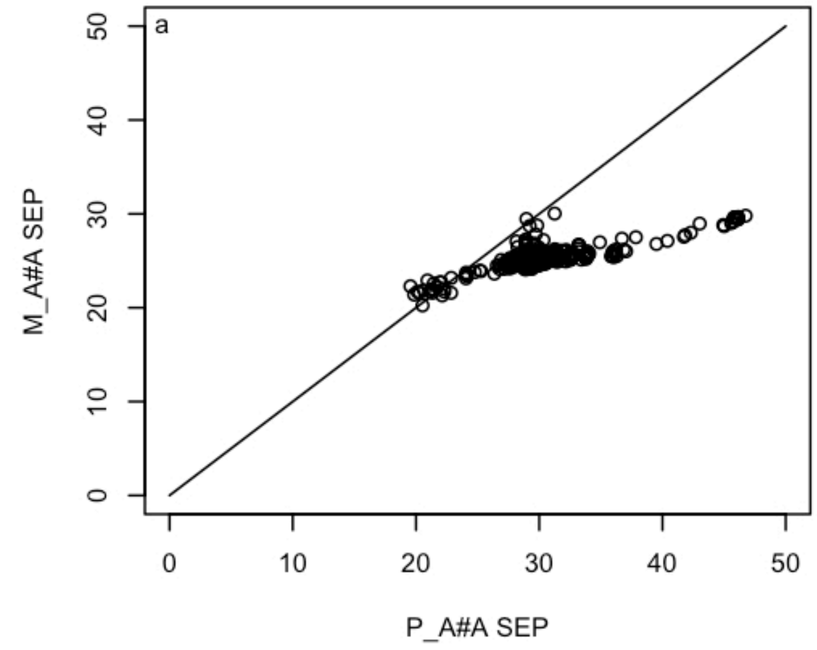
Model	Predictive Ability		Predictive Stability		
	Breeding Value	Genetic Value	Breeding Value	MSE(BV)	10% Rank cor(BV)
P_A	0.89	-	0.64	1335.67	0.17
M_A	0.87	-	0.66	1294.23	0.17
P_AD	0.86	0.89	0.72	681.53	0.12
M_AD	0.82	0.88	0.74	418.83	0.31
P_A#A	0.86	0.89	0.73	669.99	0.15
M_A#A	0.76	0.89	0.85	82.80	0.46
P_D#D	0.86	0.89	0.73	638.58	0.18
M_D#D	0.77	0.89	0.86	161.78	0.43
P_A#D	0.86	0.89	0.73	657.22	0.16
M_A#D	0.76	0.89	0.86	208.15	0.42

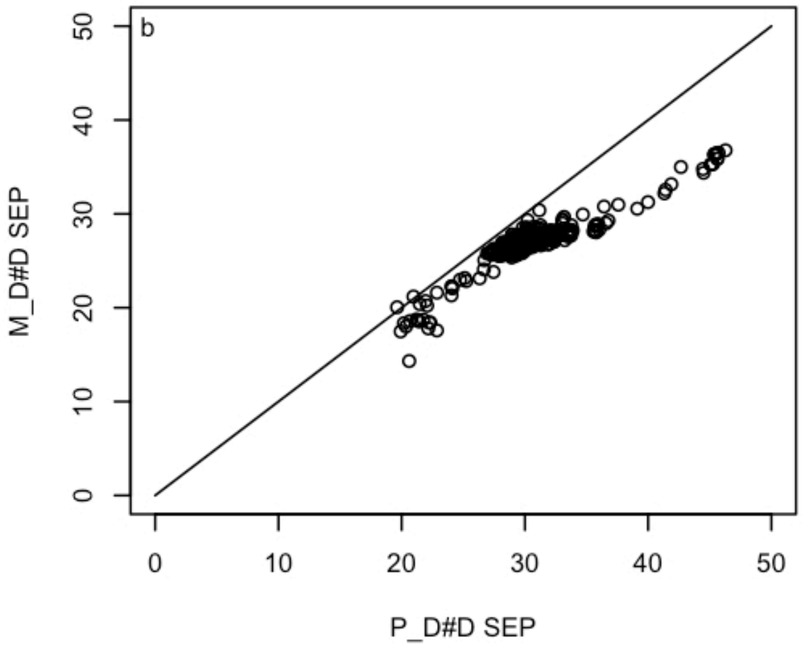

a
Cumulative proportion of variance explained
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Order of the eigenvalue
P_AD
M_AD
Orthogonal

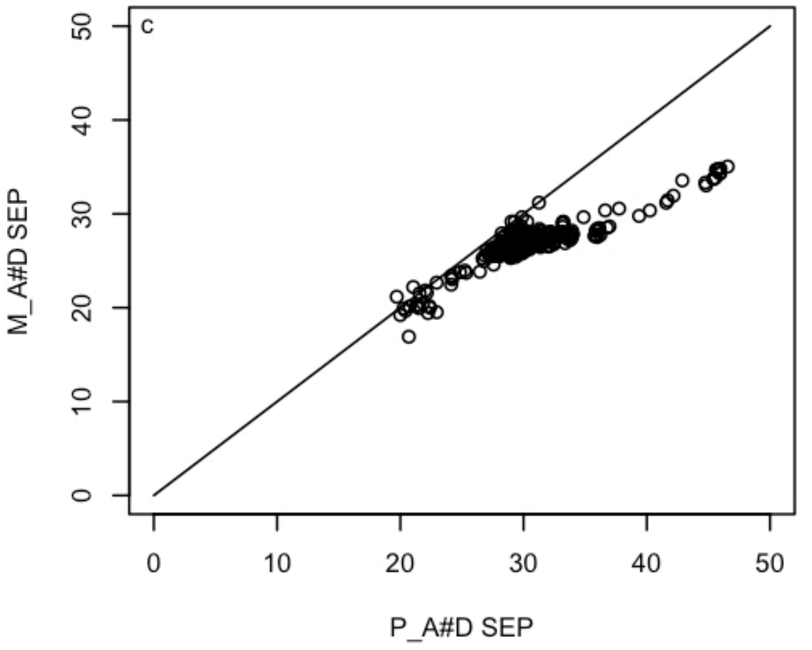

b
Cumulative proportion of variance explained
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
Order of the eigenvalue
P_A#A
M_A#A
Orthogonal

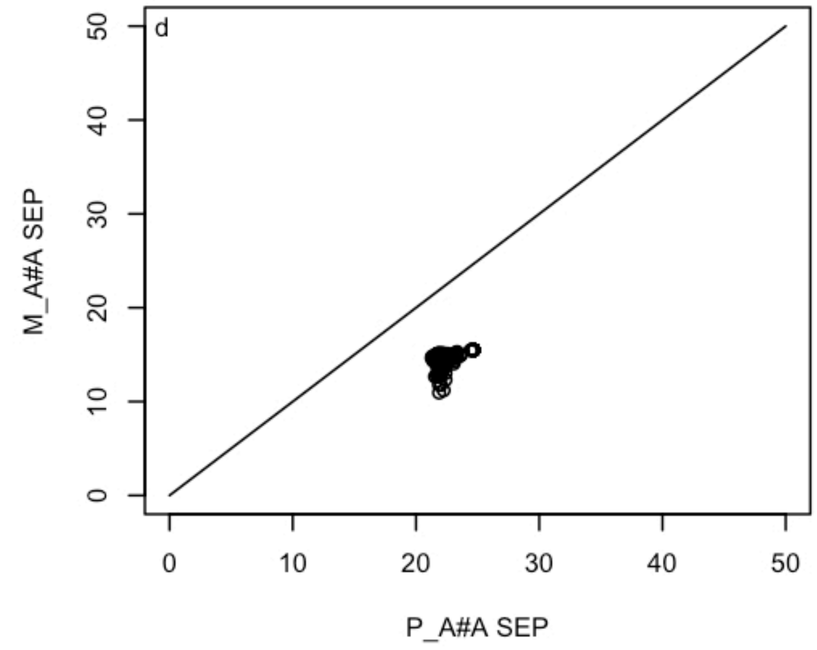

c
Cumulative proportion of variance explained
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
Order of the eigenvalue
P_D#D
M_D#D
Orthogonal

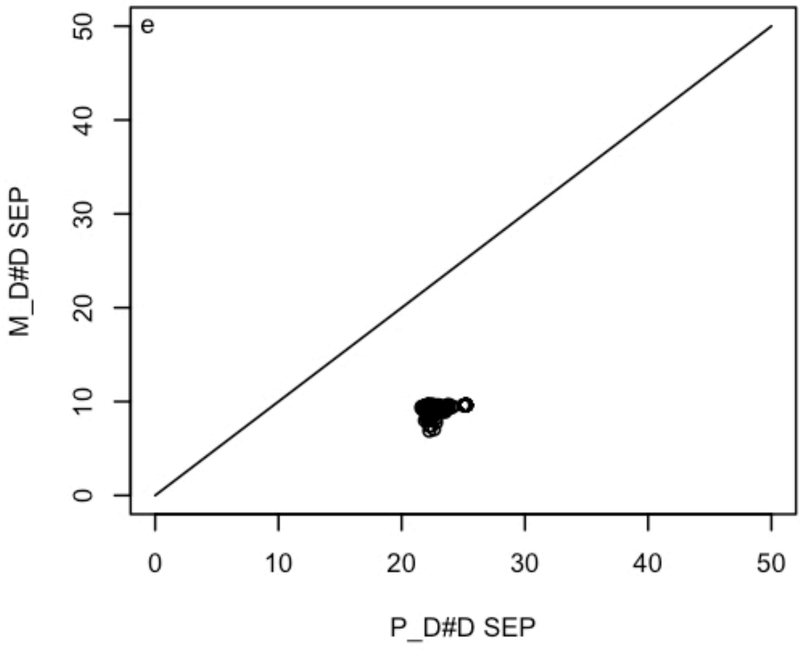

d
Cumulative proportion of variance explained
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
Order of the eigenvalue
P_A#D
M_A#D
Orthogonal

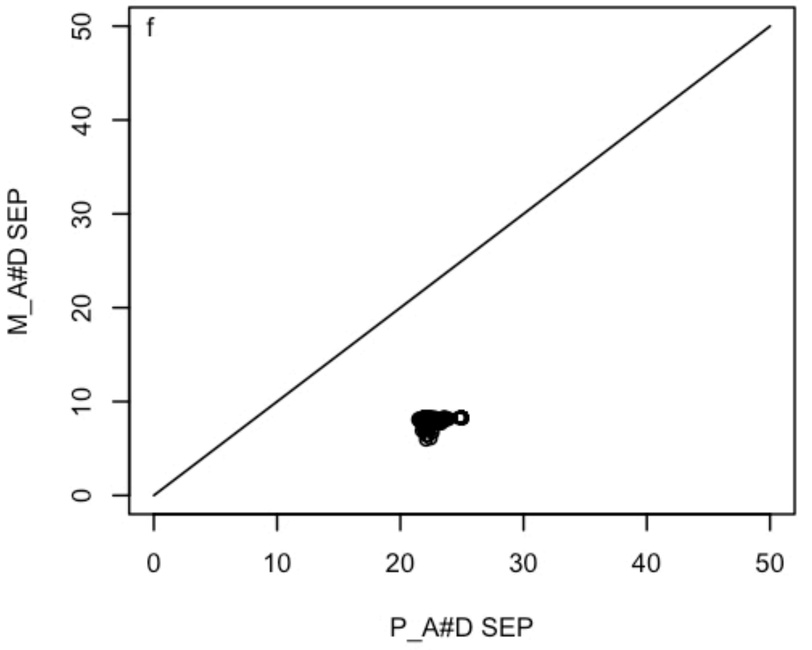

a
M_A#A SEP
P_A#A SEP
0
10
20
30
40
50


b
M_D#D SEP
P_D#D SEP
0
10
20
30
40
50


c
M_A#D SEP
P_A#D SEP
0
10
20
30
40
50


d
M_A#A SEP
P_A#A SEP
0
10
20
30
40
50


e
M_D#D SEP
P_D#D SEP
0
10
20
30
40
50


f
M_A#D SEP
P_A#D SEP
0
10
20
30
40
50