A comparison of collinearity mitigation techniques used in predicting (BLUP) breeding values and genetic gains over generations

Karen A. Eatwell[1*], Stephen D. Verryn[2], Carl Z. Roux[1], and Peter J.M. Geerthsen[3]

[1]Department of Genetics, University of Pretoria, Pretoria, 0002, South Africa.

[2]Creation Breeding Innovations cc, 75 Kafue Street, Lynnwood Glen Pretoria, 0081, South Africa.

[3]11 Poole Street, Florida, Johannesburg, 1709, South Africa

Corresponding author, email: keatwell@csir.co.za

Collinearity potentially has a negative impact on the prediction of genetic gains in tree breeding programs. The study investigated the reliability and impact of BLUP using various collinearity mitigation techniques and of two computational numerical precisions on the genetic gains in breeding populations. Multiple-trait, multiple-trial BLUP selection scenarios were run on *Eucalyptus grandis* (F_1, F_2 and F_3) and *Pinus patula* (F_1 and F_2) data, comparing predicted breeding values of parents (forward prediction) with those realised in progeny (backward prediction of parents).

Numeric precision had an impact on inter-generational correlations of BLUPs of some scenarios, indicating that it may not always be optimal to use higher precision when there is collinearity in the data. The relative difference in genetic gains between techniques varied by up to 0.38 standard deviation units in the less stable pine population. This highlights the potentially large impact that instability can have on the efficiency of a breeding programme. BLUP performed close to expected in the relatively stable (less collinear) population (eucalypt F_1), and performed poorly in the other two populations. In the unstable pine data, some of the techniques resulted in improved inter-generational correlations coming in line with expected performance.

This study indicates that BLUP can perform as expected and also confirms the potential problem of instability and consequences thereof. BLUP users should examine the nature of the population of predicted values and should these be outside expectation, various mitigation techniques should be explored.

Keywords: Multi-generational, breeding values, BLUP, realised genetic gains, collinearity

Introduction

In most tree breeding programmes, use is made of data from breeding field trials, to rank the parents or progeny in order of breeding worth, and to select the best trees to breed with or for production purposes. The breeding field trials are usually established over a number of years and locations in order to sample a wide range of environmental conditions (White and Hodge 1989). The data generated from such breeding trials are often unbalanced (White and Hodge 1989), thereby requiring appropriate selection strategies. Best Linear Unbiased Prediction (BLUP) is theoretically well-suited to predicting breeding values from data that come from a wide range of sources, qualities, quantities and ages and is particularly useful for unbalanced or messy data (White and Hodge 1989, Furlani et al. 2005). BLUP requires non-collinear data (Piepho et al. 2008) however, this is often not the practical reality. Forward selection is important in forest tree breeding for advanced generation breeding as the best individuals are used as the base material for the next generation of breeding (Ruotsalainen and Lindgren 1998). The effectiveness of the predictions thus influences the breeding progress.

In simulation research, where 60 randomly generated breeding populations of 1000 trees were created and various predictive techniques used, Best Linear Prediction (BLP) did not obtain the predicted gains in 80% of the cases studied (Verryn 1994). The underperformance of BLP using the White and Hodge (1989) BLP models and methodology, was attributed to instability caused by collinearity (lack of variable independence). This simulation study revealed the need to investigate the effects of instability in experimental data, and possible mitigation methods, prompting the investigation of these effects using BLUP in multi-generational eucalypt and pine field data.

Collinearity has been described as an 'approximate' linear relationship or a shared variance among the predictor variables in the data (Belsey et al. 1980). Collinearity, or ill conditioning, has also been described as a problem that arises when there are correlated 'independent' variables. This phenomenon is not uncommon in tree breeding trials and BLUP (regression) models. It is however, generally accepted that such correlations are permissible as long as there is a justifiable amount of new information from such variables. Inclusion of correlated variables in BLUP models is one of the main attractions of the technique for tree breeders.

When there is a perfect correlation between the 'independent' variables included in the BLUP model, the phenotypic variance-covariance matrix becomes singular and its inverse cannot be computed uniquely, resulting in the inability to calculate the estimates of β (Mitchell-Olds and Shaw 1987). When inverting a matrix that is nearly singular, the solution is computationally unstable, with different matrix inversion methods yielding different results, and the instability is reflected in large sampling variances of the estimates of β (Mitchell-Olds and Shaw 1987). This problem may be encountered in BLUP and could adversely affect the predictions (Verryn 1994).

Regression models containing highly correlated variables may give unstable parameter estimates because small changes in the observed values of the dependent variables could lead to large changes in regression coefficient estimates (McGriffin et al. 1988). Collinearity leads to regression coefficients with round off errors, unstable estimates, incorrect signs and inflated variances (McGriffin et al. 1988). Although collinear predictors may adversely affect the variance of a specific coefficient, they do not operate in isolation and the effects of sample size, overall fit of the regression models, and the interactions between these factors and collinearity occur (Mason and Perreault 1991). Sample sizes (family frequencies) vary between families and between trial sites in most tree breeding populations, which can cause collinearity to 'randomly' occur in such datasets.

The study compared the predicted breeding values of parents (forward prediction) with those realised in their progeny (backward prediction), using eucalypt and pine breeding trials. The impact on the realised genetic gains in the populations was also investigated.

Materials and methods

The approach

A series of forward BLUP breeding values, using a range of economic weightings of traits applied to historic F_1 and F_2 data were made. The realised, backward (BLUP) breeding values of the next generation (F_2 and F_3) data (of progeny from the F_1 and F_2 selections) using the same range of economic weightings, were also made. These backward predictions were regarded as the best available empirical measure of the realised breeding performance (gains) of the open pollinated F_1 and F_2 parents.

Field trials

The *Eucalyptus grandis* breeding population used in this study consisted of six trials of F_1 generation material, seven F_2 generation trials and 13 trials of F_3 material at two sites. The F_2 and F_3 material were progeny from open pollinated selections in the F_1 and F_2 trials respectively (based on historical selection criteria and methodology such as the use of family means and independent culling). The trials were established in plantations in the Limpopo and KwaZulu-Natal provinces in South Africa. The *Pinus patula* breeding population trials included 14 trials of F_1 material and six of F_2 material, the latter being progeny from open pollinated selections in the F_1 trials. The pine trials were established in plantations in the Mpumalanga province in South Africa. Details of these trials are in Table 1. In the eucalypt trials there were large differences in assessment ages between the different generations. Although not ideal, it was the only data available for the study and still proved to be valuable for this study.

[Table 1]

Data analysis

The traits used for this study were diameter at breast height (DBH in centimetres), height (metres) and stem form (subjective eight point scale). Data of all the traits were corrected for fixed effects (trial, site) using the Generalised Least Squares Method (GLM), and standardised in SAS (method of Blom 1958 as in SAS Institute Inc. 2004), giving the data a mean of 0 and a standard deviation of 1, for each trial site. The statistical analyses used SAS/STAT software, Version 9.1 of the SAS System for Windows (Copyright © 2002-2003 SAS Institute Inc.).

Predicting breeding values

The Mixed Model Least-Squares and Maximum Likelihood programme (LSMLMW and MIXMDL PC-2 Version) of Harvey (1990) was used to estimate the genetic variance components needed for the calculation of breeding values and narrow-sense heritabilities for each trial. A coefficient of relationship of 0.25 was used for the *P. patula* data. In the open pollinated half-sib *E. grandis* trials a coefficient of relationship of 0.3 was used due to the expected natural inbreeding (as much as 20%) taking place in the trials (Griffin et al. 1987, Griffin and Cotterill 1988, Verryn 1993, Hodgson 1976a, 1976b).

Two versions of a software package (Matgen), which calculates the Best Linear Unbiased Predictions for unbalanced data in tree breeding (Verryn and Geerthsen 2006), were used for the prediction of breeding values for each generation and series of trials. One of the programmes was created in Delphi (Borland Software Corporation 1983-2002) with 32-bit numerical precision and the other in Clipper (Computer Associates 1993) with 16-bit numerical precision. The analytical and mathematical procedures in both programmes were identical, the key difference being the operational level of numerical precision. The Generalised Least Squares Means estimates for fixed effects were used in this study and thus BLUE (Best Linear Unbiased Estimates) values were input into the programmes and the solution for the predicted breeding values were considered to be the BLUP solutions (White and Hodge 1989, Verryn et al. 1997).

The following equation (White and Hodge 1989) was used in predicting the breeding values:

$$\hat{g} = \boldsymbol{a}\boldsymbol{C}'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{\alpha}) \qquad [1]$$

where

$\hat{g}$ = the predicted breeding value for the individual within a particular family (forward prediction) or of the parent (backward prediction)

$\boldsymbol{a}$ = the vector (1xq) of q economic weights

$\mathbf{C}$ = the mxq matrix of genotypic variances and covariances between observations on a single candidate and its siblings, of each selection trait

$\mathbf{V}$ = the mxm matrix of phenotypic variances and covariances among observations for a single candidate and also of the means of its siblings at each trial site, of each selection trait

$\mathbf{y}$ = the mx1 vector of phenotypic observations, relating to a candidate for selection, which may include observations such as individual measurements and family means at each trial site, of the selection traits

$\boldsymbol{\alpha}$ = E(y) is the mx1 vector of expected values of observed data relating to each candidate.

Different techniques for the mitigation of collinearity were included in the programmes for calculating the BLUP values, for comparison purposes. The mitigation techniques used in the predicted breeding value ($\hat{g}$ values) calculations were Gaussian elimination with partial pivoting (Verryn 1994) in both programmes, Singular Value Decomposition (SVD) (Press et al. 1989) only in the high precision programme and Gaussian elimination with full pivoting (Press et al. 1989) in both programmes. An adapted ridge regression technique (Hoerl and Kennard 1970, Verryn 1994) was also included in the high precision programme. The partial pivoting technique serves as the control where no collinearity mitigation technique is applied.

The variances and measures of normality were estimated for the predicted breeding values in the forward prediction runs using the Univariate procedure in SAS.

Prediction scenarios

Multiple-trial and multiple-trait analyses using DBH, stem form and height in combined data were run with forward and backward models. The number of traits in the multiple-trait scenarios was balanced for all scenarios. Tree stem volume is typically used as a selection trait, however, in this study the components of stem volume were separated into DBH and height, and stem form added in order to construct the multiple-trait test scenarios. The study potentially contained selection bias as it centred on the selections that formed part of the next generation and their families. This may result in lower than expected intergenerational correlations.

Single trait scenarios were also run to compare the occurrence of instability in simple models with the multiple-trait scenarios.

In order to test the BLUP performance over generations under different economic weight scenarios, a set of ten different economic weighting vectors (Table 2) were used to make the forward and backward predictions in the trials.

Inter-generational correlations of the BLUP values (r_{fb})

Pearson correlation coefficients between the backward predicted breeding values ($\hat{g}_b$) and the forward predicted breeding values ($\hat{g}_f$) for each economic weighting scenario and mitigation technique were calculated and compared. These correlations (r_{fb}) serve as an indication of the reliability of the BLUP predictions and whether the relative predicted performance of each generation materialized in the next generation.

A merged dataset of the two generations (F_1F_2 and F_2F_3 for *E. grandis* and F_1F_2 for *P. patula*), was created for the families which were represented in both generations. The number of common families which were represented in both generations is given in Table 2. The correlation between parent and offspring, in the absence of selection, is expected to be equal to (½)h^2 (Falconer 1989). Hence, excluding the bias due to historic selection, 2r_{fb} is expected to be of the order of the heritability of the compound weighted trait (h_c^2) of the F_1 or F_2, where:

$$h_{c_i}^2 = \sum_{it} a_{it} \sum_s (h_{ts}^2 / n) \qquad [2]$$

and

$h_{c_i}^2$ = heritability of the compound weighted trait of the F_1 or F_2 for scenario i

a_{it} = the economic weight applied to trait t for scenario i, t=1 to 3

h_{ts}^2 = heritability of trait t at trial site s (s=1 to n) for scenario i

n = number of trial sites.

The compound heritability is calculated to serve as a benchmark value (Tables 2 and 3) against which the correlations between the F_1 and F_2 breeding values and F_2 and F_3 breeding values in the eucalypt and pine populations can be evaluated. As a compound heritability is calculated for the balanced case, and the data is unbalanced, it may possibly be biased upwards.

Rank correlations

Spearman rank correlations were calculated in SAS between the BLUP forward predictions of the various techniques, as a measure of the predictive ability and stability of the rankings acquired by the techniques.

Realised genetic gains

Realised genetic gains, in standard deviation units, were calculated for each economic weighting scenario and each mitigation technique, using the backward prediction breeding values. The top and bottom five percent (*E. grandis*) or ten percent (*P. patula*) of the forward prediction parents were used and hypothetical realised gain calculated as the mean of the breeding values for the progeny respectively in the F_2 or F_3 trials. Ten percent selection was used in the pine trials as there were fewer common families over generations. The variance of realised genetic gains among (mitigation) techniques within scenarios was calculated.

Results and discussion

The predicted breeding values

The single site heritability estimates for the F_2 and F_3 eucalypt trials for all assessment traits ranged from the lowest heritability of 0.125 to the highest values being 0.547. In the pine F_2 trials the single site heritability estimates ranged from 0.107 to 0.436 over all the assessment traits. The mean heritability over the trials for DBH was 0.321 (F_2 eucalypt), 0.278 (F_3 eucalypt) and 0.336 (F_2 pine). Height had mean heritability estimates of 0.322 (F_2 eucalypt), 0.283 (F_3 eucalypt) and 0.332 (F_2 pine) and stem form 0.229(F_2 eucalypt), 0.251 (F_3 eucalypt) and 0.282 (F_2 pine) over the trials.

In the three population scenarios the variances of the predicted breeding values ($\hat{g}_f$) were lowest (mean values among techniques ranged from 0.053 to 0.120 across economic weighting scenarios) in the relatively stable eucalypt F_1 population scenarios. The variances increased steadily as the populations became less stable. Variances ranged from 0.100 to 40.950 in the F_2 eucalypt scenarios and values in the least stable pine F_1

exceeded 100. The measures of deviation from normality of $\hat{g}_f$ (e.g. kurtosis and skewness) followed a similar pattern of increase as the population became less stable. Kurtosis and skewness values were much closer to the expected zero level of normally distributed populations in the F_1 eucalypt scenarios (values as low as 0.001 in some techniques). In the other two less stable populations these values were much higher (F_2 eucalypt kurtosis from 3.58 to 32.08 and in the pines kurtosis exceeded 150 and skewness was as much as -1.67 in F_2 eucalypt and 23.96 in the pines). The relatively more stable F_1 eucalypt population also had fewer $\hat{g}_f$ outliers than the other two populations' scenarios.

Instability was also measured by observing 'wrong sign' β-coefficients in the forward prediction of the F_1 and F_2 trial datasets (these 'wrong sign' coefficients would, for instance, result in the negative of an observation/family mean being included in the prediction of a breeding value where the value should intuitively be positively weighted), the magnitude of the β-coefficients and the magnitude of the predicted breeding values (large values indicating instability). Varying the economic weights had an effect on the number of cases/families of such detected instability. The F1 eucalypt cases ranged from 0.23% to 86.5% and one case of 100% in low precision partial pivoting. Unstable cases in the more unstable populations and their scenarios were generally higher (pine ranged from 8.8 % to 91.8 % and F_2 eucalypt from 14.2 % to 57.9 %. The cases of instability were generally associated with certain families, and it is thought that this is due to these families having particular frequencies of individuals in the various trials, as all other parameters remained constant in a population-scenario.

In the relatively stable situation (eucalypt F_1F_2 scenarios), the variance of the forward predictions ($\hat{g}_f$) (using standardised values) was moderate, there were no extreme $\hat{g}_f$ outliers, and the kurtosis was approaching zero, indicating normality in the predictions. In addition, the BLUP techniques are able to function as expected and the correlation between the genetic rankings over the generations (r_{fb}) compared favourably with the heritability of the compound weighted trait (h_c^2). Therefore BLUP performed close to expected here.

In contrast, the two other inter-generational suites of comparisons displayed high to very high variances and many outliers of $\hat{g}_f$. These predictions displayed high kurtosis values and deviated significantly from normality. The r_{fb} - h_c^2 relationship was far removed from the theoretically expected 1:2 ratio in these populations.

Comparison of the inter-generational correlations of BLUPs* (r_{fb}) *to the heritabilities

The comparison in the F_1 and F_2 eucalypt population showed that the correlations between the forward prediction and backward prediction breeding values (r_{fb}) obtained were of an acceptable (to high) magnitude, since $2r_{fb}$ are broadly similar (or larger) in magnitude to h_c^2 (Table 2). The effect of potential bias due to historical selection in producing the F_2 eucalypt population was therefore assumed to be negligible in this eucalypt population. In the other population comparisons there was a much wider range of correlations of which many were much smaller than (½)h_c^2 (Tables 2 and 3). This may be due to the higher incidence of instability in the matrix calculations and resulting large index ($\hat{g}_f$) values that contributed to the lower correlations with the predicted performance.

[Table 2, 3]

Fisher's Least Significant Difference (LSD) multiple range tests (α = 0.05) were run (Table 4) to determine whether significant differences existed between the mean r_{fb} correlations (from Table 2) of the different mitigation techniques and numerical precision programmes. In the eucalypt F_1F_2 scenario and pine F_1F_2 scenario a significant difference between the high and the low numerical precision programmes was observed. In the eucalypt F_2F_3 a significant difference was found between partial pivoting (both precision levels) and the rest of the techniques.

[Table 4]

A comparison was made between the mean correlations across the techniques and the compound heritabilities for each economic weighting scenario for each population (Figures 1 to 3). In Figure 1 the F_1F_2 eucalypt population illustrates that the $2r_{fb}$ against h_c^2 is approximately the expected relationship, further indicating stability here. The F_2F_3 eucalypt data show a substantial under performance of the $2r_{fb}$ relative to h_c^2 (Figure 2). The pine population (Figure 3) also deviates from expected, with the relationship points scattered from the linear regression line (the lower right-hand side scatter in Figure 3). The range in h_c^2 was small and it was difficult to obtain a good trend line. There was, however, a large range in the correlations in the pine data where some techniques and scenarios achieved the theoretical correlation whereas many did not. Plotting the r_{fb} of the best techniques (highest r_{fb}) in each scenario with the compound heritabilities in the pines resulted in a better fit, $2r_{fb}$ being within the expected order of magnitude (the upper right-hand side scatter of Figure 3). The best techniques also had better kurtosis and variance values for $\hat{g}_f$. The latter performance and that of the F_1F_2 eucalypt population serve as a confirmation that the methodology and data used here can perform according to expected genetic theory.

[Figure 1, 2, 3]

Rank correlation comparisons

In the F_1 eucalypt trials only small rank changes and in some cases no rank changes were observed within the economic weighting scenarios (Table 5) between the different techniques. More rank changes were observed in the F_2 compared to those of the F_1 eucalypt population (e.g. r = 0.896 in low precision for economic weight scenario 9 F_2, Table 5). In both the F_1 and F_2 eucalypt populations the single trait scenarios showed very few or no rank changes between techniques (Table 5). Much larger rank differences were found in the pine population and more were observed in the higher precision programme than the lower precision programme (Table 6). There was a large range in rank correlation coefficients with values as low as 0.468. The single trait scenarios in the pines showed very few or no rank changes among techniques (r = 0.951 to r = 1.000), indicating that the instability was occurring in the multiple-trait scenarios and that the pine data had the potential to perform in a stable fashion.

[Table 5,6]

The higher rank correlations in the F_1 eucalypt, compared to those of the other populations, again highlighted the stability of this population. The pine population in contrast is less stable (lower rank correlations between techniques) and the discrepancy between the different techniques used in the two programmes was also more pronounced than in the eucalypts.

The correlations between the $\hat{g}_f$ ranks of the various BLUPs with collinearity mitigation techniques were very high (in the order of 0.9 to 1) in the more stable population, and decreased to as low as approximately 0.5 in the most unstable population. The simple single trait scenarios tended to show high correlations between techniques in all populations.

Realised genetic gains

The range in mean r_{fb} of -0.094 to 0.182 and LSDs indicated that the different methods could have a meaningful effect on realised genetic gains. Similarly, significant differences in the realised genetic gains were found between techniques (Table 4). The variance of the genetic gains among mitigation techniques within scenarios is shown in Table 7.

There was a trend of increasing variability in genetic gains among mitigation techniques in the less stable populations.

[Table 7]

Small changes in gain could result in substantial improvement in the economic impact in the long run (Weir 1973, Todd et al. 1995). The range in realised genetic gains among techniques within scenarios differed by up to 0.06 standard deviation units between techniques in the relatively stable F_1F_2 eucalypts, was up to 0.21 standard deviation units between techniques in the F_2F_3 eucalypts and was as much as 0.38 standard deviation units in the pines. The latter observed differences in genetic gains highlight the importance of exploring alternative prediction techniques in the case of instability.

Comparing the realised genetic gains from the techniques, the more stable population had a lower variability of genetic gains between mitigation techniques, than those of the unstable populations. The mitigation techniques displayed greater differences in realised genetic gains in the less stable populations (up to 0.38 standard deviation units' difference).

Conclusion

The results of this study of the predicted and realised breeding value rankings over three populations and 10 scenarios each provide the first empirical evidence of the potential negative impact of collinearity in tree breeding, confirming the simulation studies of Verryn (1994).

The occurrence of instability was sensitive to the economic weightings used to calculate BLUP, and to the particular nature and structure of the data. Certain families displayed instability more readily than others, and this is thought to be as a result of the different frequencies of progeny in the various trial sites in the model (as the narrow-sense heritability and economic weightings were constant for all families of a scenario). This makes the occurrence of instability/collinearity potentially variable within datasets.

Collinearity mitigation techniques had a significant effect in all populations, however the relative performance of technique varied from case to case, and no one technique performed best over all scenarios. The effect of numerical precision showed that it can cause significant differences in r_{fb}. It may not always be optimal to use a higher numerical precision programme for BLUP index calculations. Full pivoting can be recommended over partial pivoting. If the performance of the best prediction technique in each scenario in the most unstable population is considered, the r_{fb} - h_c^2 ratio recovers to the expected range, and there is an improvement in the variance and kurtosis measures.

This study indicates that BLUP can perform as expected, however, it also confirms the potential problem of instability and the consequences thereof. It is suggested that users of BLUP should take careful note of the nature of the population of predicted values (such as kurtosis, variance, outliers and other measures of normality), and should these be outside expectation, various mitigation techniques should be explored.

Acknowledgements

We wish to thank the CSIR and the University of Pretoria for creating the opportunity for this study and investigations. The authors wish to acknowledge the CSIR technical field team from Nelspruit for the measurement of the trials and the data capture of the field measurements.

References

Belsey DA, Kuh E, and Welsh RE. 1980. Detecting and assessing collinearity. In: *Regression Diagnostics. Identifying influential data and sources of collinearity*. New York: John Wiley and Sons Inc. pp. 85 – 191.

Borland Delphi Professional Version 7.0 (Build 4.453) 2002. Borland Software Corporation.

CA-Clipper Version 5.2c 1993. Computer Associates International Inc.

Falconer DS. 1989. *Introduction to quantitative genetics*. London: Longman Scientific and Technical.

Furlani RCM, de Moraes MLT, de Resende MDV, Furlani Junior E, de Souza Gonçalves P, Filho WVV, de Paiva JR. 2005. Estimation of variance components and prediction of breeding values in rubber tree breeding using the REML/BLUP procedure. *Genetics and Molecular Biology* 28: 271-276.

Griffin AR, Cotterill PP. 1988. Genetic variation in growth of outcrossed, selfed and open-pollinated progenies of *Eucalyptus regnans* and some implications for breeding strategy. *Silvae Genetica* 37: 124-131.

Griffin AR; Moran GF, Fripp YJ. 1987. Preferential Outcrossing in *Eucalyptus regnans* F.Muell.. *Australian Journal of Botany* 35: 465-475.

Harvey WR. 1990. Users Guide for LSMLMW and MIXMDL PC-2 Version.

Hodgson LM. 1976a. Some Aspects of Flowering and Reproductive Behaviour in *Eucalyptus grandis* (Hill) Maiden at J.D.M Keet Forest Research Station. 2. The fruit, seed, seedlings, self fertility, selfing and inbreeding effects. *Southern African Forestry Journal* 98: 32-43.

Hodgson LM. 1976b. Some aspects of flowering and reproductive behaviour in *Eucalyptus grandis* (Hill) Maiden at J.D.M. Keet Forest Research Station: 3. Relative yield, breeding systems, barriers to selfing and general conclusions. *Southern African Forestry Journal* 99: 53-60.

Hoerl AE, Kennard RW. 1970. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 12: 55-67.

Mason CH, Perreault WD. Jr. 1991. Collinearity, power, and interpretation of multiple regression analysis. *Journal of Marketing Research* 28: 268-280.

McGriffin ME, Carmer SG, Ruesink WG. 1988. Diagnosis and Treatment of Collinearity Problems and Variable Selection in Least-Squares Models. *Journal of Economic Entomology* 81: 1265-1270.

Mitchell-Olds T, Shaw RG. 1987. Regression analysis of natural selection: statistical inference and biological interpretation. *Evolution* 41: 1149-1161.

Piepho HP, Möhring J, Melchinger AE, Büchse A. 2008. BLUP for phenotypic selection in plant breeding and variety testing. *Euphytica* 161: 209-228.

Press WH, Flannery BP, Teukolsky SA, Vetterling WT. 1989. *Numerical Recipes in Pascal. The art of Scientific Computing*. New York: Cambridge University Press.

Ruotsalainen S, Lindgren D. 1998. Predicting Genetic Gain of Backward and Forward Selection in Forest Tree Breeding. *Silvae Genetica* 47: 42-50.

SAS Institute Inc. 2004. SAS Online Doc® 9.1.2. SAS Institute Inc., Cary, NC, USA.

SAS. Release 9.1 2004. Copyright © 2002-2003 SAS Institute Inc., Cary, NC, USA.

Todd D, Pait J, Hodges J. 1995. The impact and value of tree improvement in the south. In: *Proceedings of the 23rd Southern Forest Tree Improvement Conference, Asheville, NC, 20-22 June 1995*. pp 9-16.

Verryn SD. 1993. Research and development report compiling estimates of genetic parameters in open-pollination and diallel trials. Report FOR-DEA 00601. Pretoria: Environmentek, CSIR.

Verryn SD. 1994. Improving Best Linear Prediction for Tree Breeding. PhD Thesis, University of Pretoria, South Africa.

Verryn SD, Field CL, Garcia J, Griffin R. 1997. A discussion on the relationship between heritabilities and genetic correlations and the standard errors of these parameters with a case study example of GEI in *E. grandis* over two sites in South Africa and one site in Uruguay. In: *Proceedings of the IUFRO Conference on Silviculture and Improvement of Eucalypts*. *24-29 August 1997, Salvador, Brazil*. pp 43-49.

Verryn SD, Geerthsen JMP. 2006. Matgen Version 7.2. Development Version. A BLP package for unbalanced index selection in tree breeding. Natural Resources and the Environment, CSIR, PO Box 395, Pretoria, RSA.

Weir RJ. 1973. Realizing genetic gains through second-generation seed orchards. In: *Proceedings of the 12th Southern Forest Tree Improvement Conference, Baton Rouge, 12-13 June 1973, LA, US.* pp 14-23.

White TL, Hodge GR (eds). 1989. *Predicting Breeding Values with Applications in Forest Tree Improvement.* Kluwer Academic Publishers.

Table 1: Trial information for F_1, F_2 and F_3 *Eucalyptus grandis* and F_1 and F_2 *Pinus patula* trials in South Africa used in the study

Trial name and site	Genetic material (No. Families)	Experimental design	Age Assessed
EA6206; EA6209; EA6210; EA6215; EA6218; EA6221; JDM	99 F_1 *	RCB; 9 replicates; 2x2 tree plots	48 to 72 months
EA62A1; EA62A2 (2 trials); EA62A3; EA62A4; EA62A5; EA62A6; JDM Keet & Kwambonambi	A1-A2: 64 F_2; A3 & A6: 72 F_2; A4-A5: 99 F_2; all thinned to 1 tree per plot	Alpha lattice; 9 replicates; 2x2 tree plots	62 to 91 months
EA62B4-B16 Dukuduku	50 F_3; single tree plots	Alpha lattice; 20 replicates	40 months; 51 months (B15 – 16)
EA62B4-B14 Silwerfontein EA62B15-16 Westfalia	50 F_3; single tree plots	Alpha lattice; 20 replicates	38 & 25 months
PF4002 Belfast & Tweefontein	42 F_1	6x7 Lattice; 3 replicates;4x4 tree plots	156 months
PF4003 Rietfontein	41 F_1	Random complete block; 10 replicates; 4x4 tree plots	108 months
PF4004 Wilgeboom	72 F_1	6x6 lattice; 2 sets; 4 replicates per set; 1x10 tree	102 months
PF4005 Wilgeboom & Tweefontein	49 F_1	7x7 lattice; 4 replicates; 1x10 tree row plots	100 & 166 months
PF4006 Jessievale & Tweefontein	42 F_1	6x7 lattice; 2 sets; 3 replicates per set; 1x10 tree	96 & 65 months
PF4007 Jessievale & Frankfort	42 F_1	6x7 lattice; 6 replicates; 1x6 tree row plots	104 & 97 months
PF4008 Tweefontein & Jessievale	282 F_1 49 F_1	7x7 lattice; 6 sets Tweefontein; 8 replicates per set; 1x6 tree row plots	96 months
PF4009 Jessievale	182 F_1	7x7 lattice; 4 sets; 8 replicates per set; 1x6 tree	96 months
PF4010 Jessievale	64 F_1	8x8 lattice; 9 replicates; 1x6 tree row plots	72 months
PF4011 Tweefontein, Mac-Mac, Frankfort & Wilgeboom	64 - 89 F_2; single tree plots	RCB; 20 replicates	84 months
PF4015 Tweefontein & Wilgeboom	59 F_2; single tree plots	RCB; 20 replicates	96 months

* Trials thinned to 1 tree per plot and rogued to between 59 and 61 families; site EA6218 thinned to 2 trees per plot
RCB = Randomized Complete Block; Trials with EA = *Eucalyptus grandis* and PF = *Pinus patula*

Table 2: A comparison of twice the mean Pearson Correlations (r_{fb}) between the backward prediction ($\hat{g}_b$) and the forward prediction ($\hat{g}_f$) with the heritability of the compound weighted trait in the eucalypt and pine populations

Economic weighting scenario				Twice the mean of mitigation technique r_{fb} correlations			Heritability of compound weighted trait h_c^2		
No.	DBH	Height	Stem	F_1F_2 eucalypt	F_2F_3 eucalypt	F_1F_2 pine	F_1 eucalypt	F_2 eucalypt	F_1 pine
1	0.2	0.4	0.4	0.3507	0.0927	-0.0071	0.310	0.285	0.269
2	0.4	0.4	0.2	0.4163	0.1068	0.1214	0.327	0.303	0.279
3	0.15	0.6	0.25	0.4841	0.1485	0.0669	0.349	0.299	0.285
4	0.1	0.7	0.2	0.5336	0.1581	0.1003	0.367	0.303	0.292
5	0.7	0.2	0.1	0.3740	0.0783	0.2051	0.309	0.312	0.275
6	0.2	0.1	0.7	0.1931	0.0313	0.1581	0.245	0.257	0.242
7	0.3	0.2	0.5	0.2443	0.0506	0.0276	0.275	0.275	0.256
8	0.5	0.3	0.2	0.3777	0.0849	0.1300	0.314	0.303	0.274
9	0.8	0.1	0.1	0.3410	0.0545	0.2107	0.296	0.312	0.270
10	0.1	0.1	0.8	0.2017	0.0296	0.1599	0.237	0.248	0.237

F_1F_2 eucalypt: 451 common families

F_2F_3 eucalypt: 318 common families

F_1F_2 pine: 71 common families (only two trial series data of F_2 population available for study)

Table 3: Mean (over economic weight scenarios) and single trait Pearson Correlation Coefficients ($2r_{fb}$) between the backward prediction ($\hat{g}_b$) and the forward prediction ($\hat{g}_f$) comparing collinearity mitigation techniques together with the mean compound heritability in the different populations

Scenarios	Species	Collinearity Mitigation Method used with BLUP								Mean heritability of compound weighted trait(h_c^2)
		PP	FP	SVD^3	RR	Low PP	Low FP	SVD^1	SVD^2	
Mean over 10	F_1F_2 eucalypt	0.36128	0.36128	0.36128	0.36296	0.30566	0.35154	0.36318	0.35564	0.303
multiple-trait	F_2F_3 eucalypt	-0.00488	0.14346	0.12826	0.11332	-0.00726	0.12822			0.290
scenarios:	F_1F_2 pine	-0.18887	0.21503	0.05388	0.00527	0.32666	0.32346			0.268
Single traits:										
DBH	F_1F_2 eucalypt	0.36682***	0.36654***	0.36654***		0.35026**	0.35026**			0.292
Height		0.54974***	0.54974***	0.54974***		0.55398***	0.55398***			0.423
Stem form		0.38396***	0.38814***	0.38814***		0.38744***	0.38744***			0.207
DBH	F_2F_3 eucalypt	0.13034^{ns}	0.13284^{ns}	0.13034^{ns}		0.13046^{ns}	0.13284^{ns}			0.321
Height		0.39522**	0.39850**	0.39522**		0.39526**	0.39850**			0.322
Stem form		0.28846*	0.28846*	0.28846*		0.28850*	0.28846*			0.229
DBH	F_1F_2 pine	-0.02686^{ns}	-0.03358^{ns}	-0.03358^{ns}		0.02672^{ns}	0.01994^{ns}			0.270
Height		0.21848^{ns}	0.25648^{ns}	0.25648^{ns}		0.32190^{ns}	0.33758^{ns}			0.315
Stem form		0.52930*	0.52930*	0.52930*		0.50366*	0.53894*			0.223

Correlation coefficient significant effect: *** $p<0.0001$ ** $p<0.01$ * $p<0.05$ ns non significant

Significance not calculated for twice the mean correlation coefficients among techniques over economic weighting scenarios

SVD = Singular value decomposition

SVD^1 = SVD with threshold of 1x10^{-2} ; SVD^2 = SVD with threshold of 1x10^{-1}; SVD^3 = SVD with threshold of 1x10^{-6} (standard threshold);

PP = partial pivoting control; FP = full pivoting; RR = ridge regression; Low = lower precision

Table 4: Least significant difference (LSD) multiple range test of the forward-backward correlations for BLUP predictions and the mean realised genetic gains in standard deviation units of the different economic weight scenarios using different mitigation techniques in the eucalypt and pine populations (means with the same letter are not significantly different from each other at α = 0.05)

Species	Mitigation Method	Pearson		Realised Genetic	
		LSD	Mean	LSD	Mean
Eucalypt F_1F_2	SVD(1x10^{-2} threshold)	A	0.18159	A	0.11795
	Ridge	A	0.18148	A	0.12107
	Full pivoting high	A	0.18065	A	0.11768
	Partial pivoting high	A	0.18064	A	0.11710
	SVD(1x10^{-1} threshold)	AB	0.17782	A	0.11316
	Full pivoting low	B	0.17577	A	0.11756
	Partial pivoting low	C	0.15301	B	0.10066
Eucalypt F_2F_3	Full pivoting high	A	0.07173	A	0.08849
	SVD	A	0.06413	A	0.08543
	Full pivoting low	A	0.06411	A	0.08514
	Ridge	A	0.05666	A	0.06474
	Partial pivoting high	B	-0.00244	B	0.01892
	Partial pivoting low	B	-0.00363	B	0.00765
Pine F_1F_2	Partial pivoting low	A	0.16333	A	0.13523
	Full pivoting low	A	0.16173	A	0.15918
	Full pivoting high	B	0.10752	AB	0.09291
	SVD	C	0.02694	B	0.04874
	Ridge	C	0.00264	C	-0.04276
	Partial pivoting high	D	-0.09443	C	-0.03546

N=10 for mean correlations; N=20 for mean realised genetic gains;
low = low precision; high= high precision
SVD = Singular value decomposition; RIDGE = Adapted ridge regression

Table 5: Spearman rank correlation coefficients for the eucalypt population forward predictions

Mitigation Methods	Trials	Single traits			Economic weighting* Multiple-trait									
		DBH	Height	Stem form	1	2	3	4	5	6	7	8	9	10
PP–FP	F1	1.000	1.000	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	F2	1.000	1.000	1.000	0.936	0.962	0.944	0.944	0.946	0.950	0.946	0.938	0.941	0.949
PP–SVD	F1	1.000	1.000	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	F2	1.000	1.000	1.000	0.937	0.962	0.944	0.945	0.946	0.950	0.946	0.939	0.942	0.949
PP–RIDGE	F1				0.997	0.998	0.998	0.998	0.998	0.995	0.996	0.998	0.997	0.994
	F2				0.913	0.938	0.906	0.897	0.933	0.914	0.920	0.922	0.929	0.908
PP– low PP	F1	0.935	0.943	0.937	0.944	0.937	0.940	0.938	0.933	0.935	0.938	0.937	0.933	0.934
	F2	0.908	0.897	0.944	0.922	0.943	0.919	0.915	0.940	0.914	0.918	0.932	0.816	0.912
PP– low FP	F1	0.935	0.943	0.937	0.952	0.944	0.945	0.943	0.942	0.933	0.951	0.945	0.942	0.943
	F2	0.908	0.897	0.944	0.936	0.961	0.944	0.944	0.946	0.949	0.946	0.939	0.942	0.949
FP–SVD	F1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	F2	1.000	1.000	1.000	0.999	0.999	0.999	0.999	0.999	1.000	0.999	0.999	0.999	1.000
FP–RIDGE	F1				0.997	0.998	0.998	0.998	0.998	0.995	0.996	0.998	0.997	0.994
	F2				0.976	0.977	0.966	0.959	0.985	0.974	0.980	0.981	0.987	0.971
FP– low PP	F1	0.935	0.943	0.937	0.944	0.937	0.940	0.938	0.933	0.935	0.938	0.937	0.932	0.934
	F2	0.908	0.897	0.944	0.922	0.961	0.933	0.923	0.945	0.915	0.916	0.935	0.895	0.914
FP– low FP	F1	0.935	0.943	0.937	0.952	0.945	0.945	0.943	0.942	0.933	0.951	0.945	0.942	0.943
	F2	0.908	0.897	0.944	0.998	0.999	0.999	0.999	0.998	0.999	0.998	0.999	0.999	0.999
SVD–RIDGE	F1				0.997	0.998	0.998	0.998	0.998	0.995	0.996	0.998	0.997	0.994
	F2				0.977	0.977	0.966	0.959	0.985	0.975	0.981	0.982	0.988	0.971
SVD– low PP	F1	0.935	0.943	0.937	0.944	0.937	0.940	0.938	0.933	0.936	0.938	0.937	0.932	0.934
	F2	0.908	0.897	0.944	0.923	0.962	0.934	0.923	0.946	0.915	0.917	0.936	0.895	0.914
SVD– low FP	F1	0.935	0.943	0.937	0.952	0.945	0.945	0.943	0.942	0.933	0.951	0.945	0.942	0.943
	F2	0.908	0.897	0.944	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999
RIDGE– low PP	F1				0.938	0.933	0.934	0.933	0.929	0.925	0.929	0.932	0.928	0.923
	F2				0.900	0.939	0.898	0.880	0.932	0.881	0.892	0.919	0.892	0.876
RIDGE– low FP	F1				0.948	0.942	0.942	0.940	0.938	0.928	0.946	0.942	0.938	0.935
	F2				0.976	0.977	0.966	0.958	0.985	0.974	0.980	0.981	0.987	0.971
Low PP– low FP	F1	1.000	1.000	1.000	0.991	0.991	0.993	0.994	0.990	0.974	0.985	0.991	0.990	0.990
	F2	0.908	0.897	0.944	0.923	0.963	0.934	0.924	0.946	0.916	0.917	0.937	0.896	0.914

Correlation coefficient significant effect: all rank correlations significant $p < 0.0001$.

PP = Partial pivoting high precision control; FP = Full pivoting high precision; SVD = Singular value decomposition; RIDGE = Adapted ridge regression; low PP = Partial pivoting low precision control; low FP = Full pivoting low precision

*Economic weighting scenarios as detailed in Table 2

Table 6: Spearman rank correlation coefficients for the pine population forward predictions

	Single traits			Economic weighting* multiple-trait									
Mitigation Methods	DBH	Stem form	Height	1	2	3	4	5	6	7	8	9	10
PP–FP	1.000	1.000	0.997	0.621	0.640	0.614	0.604	0.650	0.658	0.633	0.643	0.660	0.668
PP–SVD	1 .000	1.000	0.997	0.551	0.543	0.537	0.522	0.542	0.572	0.562	0.546	0.552	0.572
PP–RIDGE				0.587	0.575	0.582	0.569	0.585	0.591	0.584	0.582	0.593	0.595
PP–low PP	0.963	0.963	0.951	0.479	0.483	0.490	0.468	0.480	0.540	0.507	0.491	0.497	0.551
PP–low FP	0.967	0.970	0.953	0.572	0.599	0.565	0.554	0.608	0.602	0.583	0.603	0.618	0.607
FP–SVD	1.000	1.000	1.000	0.871	0.831	0.835	0.824	0.814	0.807	0.866	0.828	0.815	0.783
FP–RIDGE				0.863	0.853	0.844	0.837	0.848	0.837	0.862	0.851	0.848	0.820
FP–low PP	0.963	0.964	0.949	0.714	0.699	0.711	0.699	0.706	0.738	0.736	0.708	0.716	0.741
FP–low FP	0.967	0.971	0.955	0.906	0.907	0.901	0.898	0.903	0.891	0.905	0.908	0.902	0.886
SVD–RIDGE				0.885	0.856	0.858	0.854	0.846	0.859	0.876	0.850	0.846	0.855
SVD–low PP	0.963	0.964	0.949	0.649	0.607	0.618	0.600	0.603	0.623	0.658	0.617	0.614	0.604
SVD–low FP	0.968	0.971	0.955	0.814	0.772	0.781	0.773	0.754	0.775	0.808	0.767	0.755	0.770
RIDGE–low PP				0.615	0.614	0.607	0.598	0.618	0.641	0.631	0.621	0.627	0.637
RIDGE–low FP				0.800	0.788	0.785	0.782	0.778	0.767	0.794	0.784	0.778	0.760
Low PP–low FP	0.998	0.994	0.992	0.707	0.704	0.714	0.704	0.709	0.710	0.728	0.712	0.720	0.705

Correlation coefficient significant effect: All rank correlations significant $p<0.0001$

PP = Partial pivoting high precision control; FP = Full pivoting high precision; SVD = Singular value decomposition; RIDGE = adapted ridge regression

low PP = Partial pivoting low precision control; low FP = Full pivoting low precision

*Economic weighting scenarios as detailed in Table 2

Table 7: The variance of realised genetic gains (in standard deviation units) between techniques within scenarios in the eucalypt and pine populations

Species	Selection Population	Performance measured in	Variance of genetic gains*	
			Top %	Bottom %
E. grandis	F_1	F_2	0.0016	0.0014
E. grandis	F_2	F_3	0.0028	0.0032
P. patula	F_1	F_2	0.0125	0.0269

* Eucalypts top and bottom percentage is 5% and pines top and bottom percentage is 10 %

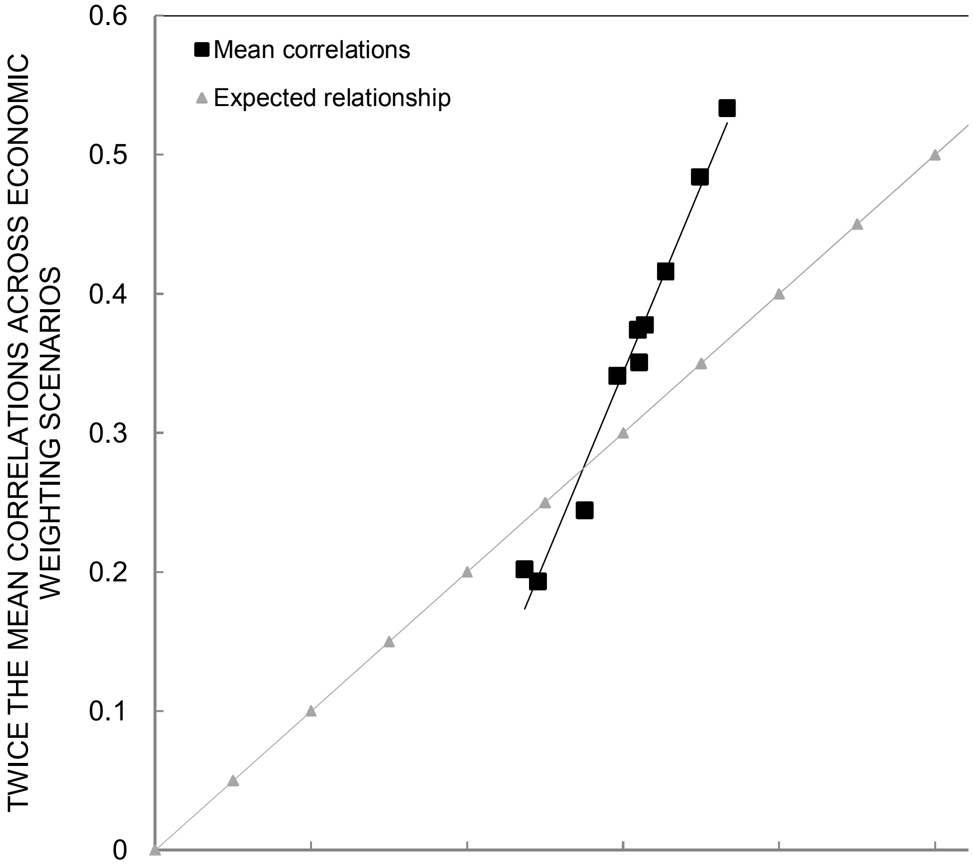

Figure 1: Twice the mean correlations across the economic weighting scenarios relative to the heritability of the compound weighted trait across the same economic weighting scenarios for the F_1F_2 eucalypt population. The diagonal line represents the expected linear relationship between the correlations and the heritability of the compound weighted trait

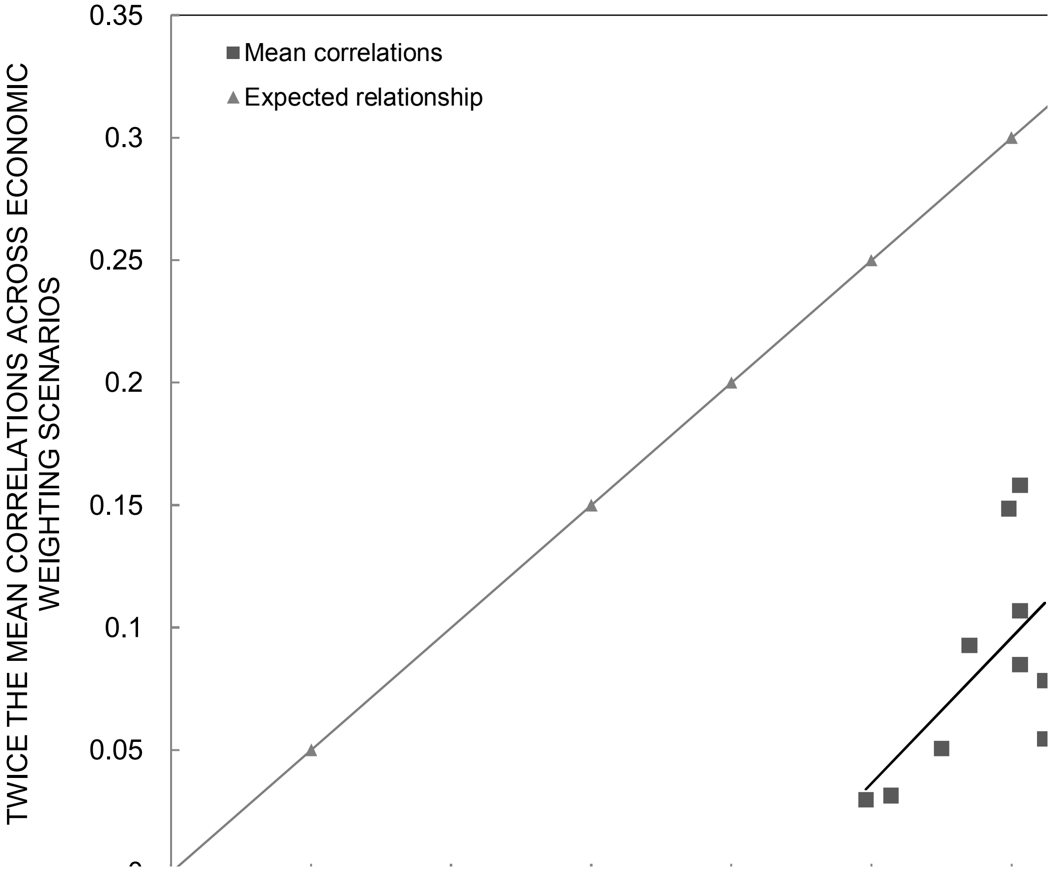

Figure 2: Twice the mean correlations across the economic weighting scenarios relative to the heritability of the compound weighted trait across the same economic weighting scenarios for the F_2F_3 eucalypt populations. The diagonal line represents the expected linear relationship between the correlations and the heritability of the compound weighted trait

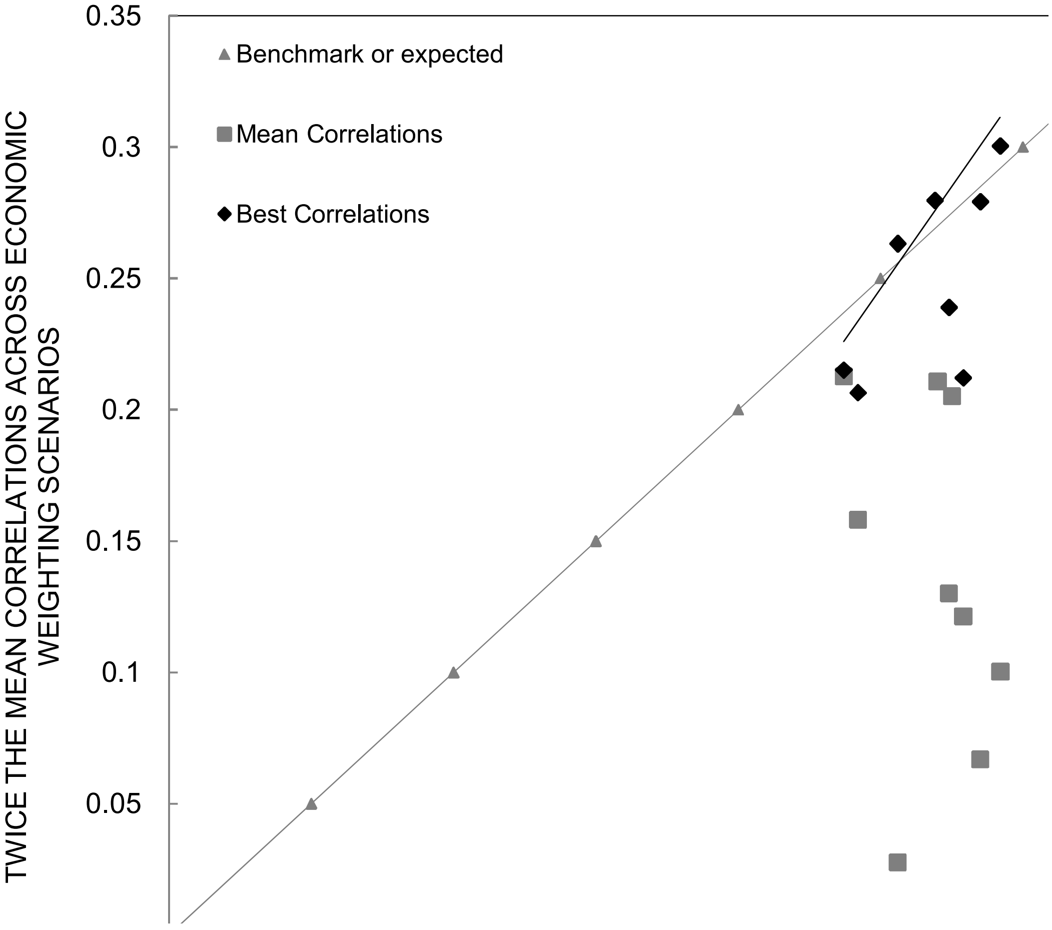

Figure 3: Twice the mean correlations across the economic weighting scenarios and the best correlation within each economic weighting scenario relative to the heritability of the compound weighted trait across the same economic weighting scenarios for the F_1F_2 pine population. The lines represent the linear relationships between the correlations and the heritability of the compound weighted trait