Specialty Wood Products Research Partnership

SCION FORESTS ▪ PRODUCTS ▪ INNOVATION

Multi-environment single-step genomic evaluation of *Eucalyptus nitens* progeny test

Authors: Jaroslav Klápště, Steffi Fritsche, Peter Bird, Kane Fleet, Toby Stovold

Date: 30. September 2021

Publication No: SWP-T136

FOREST GROWERS RESEARCH

TABLE OF CONTENTS

EXECUTIVE SUMMARY 1
INTRODUCTION 2
AIM 2
METHODS 3
RESULTS 5
CONCLUSION 9
ACKNOWLEDGEMENTS 10
REFERENCES 10
APPENDIX 12
Breeding values estimated for each individual at each site 12

supported by forestgrowers commodity levy

FOREST GROWERS RESEARCH

EXECUTIVE SUMMARY

The key outcome from this research is the ability to predict genomic breeding values for non-phenotyped individuals and parents using single-step genomic evaluation, especially for traits associated with costly phenotyping such as wood density.

The evaluation was performed on two open pollinated field experiments (“Keen’s block” and “Fortification”) which included material derived from different sources.
Our single-step genomic evaluation found statistically significant heritabilities for all investigated traits which confirms potential for genetic improvement. Our analysis also detected considerable genotype by environment interaction in growth attributes but a lower GxE interaction for wood density. Most growth traits were strongly correlated while growth and wood density were only weakly correlated. The trait stem straightness had a strong relationship with productivity in Keens block while non was found in Fortfication.

The implementated single-step genomic evaluation allows for predicting genomic breeding values, also for non-genotyped individuals and parents. This, in turn, can be utilized in genetic thining/culling in current seed orchards, especially with focus on wood quality which shows high prediction accuracy and is usually costly to phenotype.

INTRODUCTION

Forest tree breeding is a long-term endeavour to find the best match between genetics and environmental conditions. Long breeding cycles prevent testing each genotype across all tested environments. Thus, the predicted genotype's performance in untested environments relies mainly on information from relatives. However, pedigree information in forest tree breeding can be relatively shallow and simple which decreases the accuracy of such predictions. Genomic prediction using genomic markers distributed across the genomes can increase the precision of prediction of unknown phenotypes (Meuwissen et al., 2001).

Affordable genotyping platforms such as SNP arrays (Silva-Junior et al., 2015) or Genotyping-by-sequencing (GBS) (Elshire et al., 2011) led to the establishment of forest genomic resources andcan be also implemented in operational breeding programmes to predict unknown phenotypes (Suontama et al., 2019; Klapste et al., 2018; Klapste et al., 2021). Single point mutation (SNP) markers are highly abundant across the whole genome and can help trace uncovered recent and historical relatedness between individuals (Powell et al. 2010). This canconvert a relatively sparse pedigree-based relationship matrix into a dense marker-based relationship matrix representing the proportion of genomic information shared across any pair of genotyped individuals. Thus, the performance of genotypes in untested environments can be predicted more precisely due to the higher number of related individuals. Additionally, if the genomic markers are in close vicinity of causal variants responsible for phenotypic expression of the tested traits, the precision of performance predicted for untested genotypes can further increased (Habier et al., 2013).
In summary, the accuracy of genomic prediction is dictated by three factors:

1) traits' heritability – higher heritability indicates that environment has a lesser influence on the phenotype (small residual variance related to the measurements),
2) the size of the training population. Larger populations allow for a stronger statistical evidence about the connection between markers and causal variants, and
3) effective population size: smaller effective population size means higher chances that the markers will be connected to causal variants due to lack of recombinations (Goddard 2009, Reseande and Grattapaglia 2011).

Since forest tree breeding is a long-term effort, not every genotype in the breeding programme is available for genotyping. Therefore, a mixture of genotyped and non-genotyped individuals should be used to form a training population. Single-step genomic evaluation allows combining all phenotypic, pedigree and genomic data into a single analysis (Misztal et al., 2009; Legarra et al., 2009). The approach is based on blending a marker-based into a pedigree-based relationship matrix. The resulting combined relationship matrix is then used in mixed linear models (MLMs) to predict genomic breeding values. The blending of the two matrices consists of two critical steps: (1) rescaling the marker-based relationship matrix to the same scale as the pedigree-based relationship matrix and (2) weighting of the marker-based relationship matrix to reflect the fact that not all additive genetic variance is explained by markers, as well as to assure that the matrix is positive-definite (requirement for MLMs).

AIM

The aim of this study was to implement a multi-environment single-step genomic evaluation to explore genotype by environment interaction (GxE) and genomic prediction of untested genotypes for two progeny sites (Keen's block and Fortification) planted in the South Island of New Zealand.

METHODS

The two third-generation of *Eucalyptus nitens* progeny tests (Keen's block and Fortification) were screened for diameter at breast height (DBH [mm]), stem height (HT [m]), wood density (WD [kg/m3]), stem straightness (STR – 8 degree scale from 1 – poor straightness to 9 – straight stem) and malformation (MAL – 8 degree scale from 1 – heavy malformed stem with multiple stems to 9 – no stem malformation). Multiple seed sources were tested at each site: Keen's block included seed material from Tinkers seed orchard, Waiouru seed orchard and the Australian tree seed centre (ATSC); the Fortification site included material from Tinkers seed orchard, Alexandra seed orchard, Drumfern seed orchard, Waikuku seed orchard, Forestry Tasmania seed orchard and ATSC. Genomic data were collected from previous SWP projects for Keen's block site while new data were generated for Fortification site. In total, 5,753 individuals were phenotyped from which 1,011 individuals were genotyped.
To eliminate any discrepancies between data coming from different genotyping platforms the median of genotypes were compared between samples from Keen's block and Fortification sites. In total, 27,013 SNPs were in common across the genotyping platforms, then from those 6,585 SNPs, which had also a match in the median of the genotypes, were selected for constructing a marker-based relationship matrix.

We implemented single-step genomic evaluation combining all phenotypic, genomic and pedigree information to attribute to the partly genotyping of the population. The multivariate single-step genomic analysis of the JWAS package (Cheng et al., 2015) was performed using MCMC algorithm as follows:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{Zg} + \boldsymbol{Zb} + \boldsymbol{e}$$

where $\boldsymbol{Y}$ is a matrix of phenotypes, $\boldsymbol{\beta}$ is the vector of fixed effects including the overall mean, seed source and replication effects, $\boldsymbol{g}$ is a matrix of genomic estimated breeding values following $var(\boldsymbol{g}) \sim N(0, \boldsymbol{G_1})$, following the variance-covariance structure $G_1 = \begin{bmatrix} \sigma^2_{a_{t1}} & \cdots & \sigma_{a_{t1}a_{tn}} \\ \vdots & \ddots & \vdots \\ \sigma_{a_{tn}a_{t1}} & \cdots & \sigma^2_{a_{tn}} \end{bmatrix} \otimes \boldsymbol{H}$, where $\sigma^2_{a_{t1}}$ and $\sigma^2_{a_{tn}}$ are the additive genetic variances for the 1st and nth trait, $\sigma_{a_{t1}a_{tn}}$ and $\sigma_{a_{tn}a_{t1}}$ are the additive genetic covariances between traits, where σ^2_g is additive genetic variance associated with relatedness inferred from combination of pedigree and genomic information, $\otimes$ is the Kronecker product and $\boldsymbol{H}$ is the combined relationship matrix that was constructed using the pedigree-based relationship matrix $\boldsymbol{A}$ and marker-based relationship matrix as follows:

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

where $\mathbf{A_{11}}$ is the pedigree-based relationship matrix for non-genotyped individuals, $\mathbf{A_{22}}$ is the pedigree-based matrix for genotyped individuals, $\mathbf{A_{12}}$ and $\mathbf{A_{21}}$ are pedigree-based matrices between genotyped and non-genotyped individuals, **G** is the marker-based relationship matrix. The marker-based relationship matrix was estimated following (VanRaden 2008):

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

where **Z** = **M** - **P**, **M** is the genotype matrix coding reference allele homozygote as 0, heterozygote as 1 and alternative allele homozygote as 2 and **P** is double the frequency for the alternative allele. Since the marker-based relationship matrix reflects both the temporary relatedness defined by pedigree and the historical relatedness developed prior to the base populationof the pedigree (Powell et al.. 2010, Speed and Balding 2015), which is not on the same scale as the pedigree-based

relationship matrix. Therefore, the rescaling of the marker-based relationship matrix was performed. We adopted a rescaling approach developed in Gao et al. (2012) as follows:

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

A requirement of the mixed linear models for covariance structures is a positive semi-definite marker-based relationship matrix, which is often not given. Thus, it was necessary to undertake a weighting of information (0.05 for pedigree information) originating from genomic markers and from the pedigree which was performed as follows:

$$\boldsymbol{G_w}=\boldsymbol{G}(1\text{-w})+\boldsymbol{A_{22}}\text{w}$$

where **w** is the proposed weight on the pedigree-based relationship matrix. Vector **b** represents a random set-within-rep effect following $var(\boldsymbol{b})\sim N(0,\boldsymbol{G_2})$, following the variance-covariance structure $G_2=\begin{bmatrix} \sigma^2_{b_{t1}} & \cdots & \sigma_{b_{t1}b_{tn}} \\ \vdots & \ddots & \vdots \\ \sigma_{b_{tn}b_{t1}} & \cdots & \sigma^2_{b_{tn}} \end{bmatrix}\otimes\boldsymbol{I}$, where $\sigma^2_{b_{t1}}$ and $\sigma^2_{b_{tn}}$ are the set-within-rep variances for the 1st and nth trait, $\sigma_{b_{t1}b_{tn}}$ and $\sigma_{b_{tn}b_{t1}}$ are the set-within-rep covariances between traits and ***I*** is identity matrix. Similar to the vector ***g***, ***e*** is the matrix of residual effects following $va(\boldsymbol{e})\sim N(0,\boldsymbol{R})$, where ***R*** is a variance-covariance structure for residual effects $\boldsymbol{R}=\begin{bmatrix} \sigma^2_{e_1} & \cdots & \sigma_{e_1e_n} \\ \vdots & \ddots & \vdots \\ \sigma_{e_ne_1} & \cdots & \sigma^2_{e_n} \end{bmatrix}\otimes\boldsymbol{I}$, where $\sigma^2_{e_1}$ and $\sigma^2_{e_n}$ are residual variances for 1st and nth trait and $\sigma_{e_1e_n}$ and $\sigma_{e_ne_1}$ are residual covariances between 1st and nth trait. The convergence of the MCMC algorithm was tested through comparison of 5 Markov chains with a total number of 60,000 runs and 10,000 burnin period using Gelman-Rubin method (Genlman and Rubin 1992). The final model was set for 40,000 burnin period and 140,000 runs, where every 100th run was sampled.In an alternative analysis only genotyped individuals and marker-based relationship matrix (GBLUP) was implemented using the same model but relationship matrix **H** was replace by relationship matrix **G**.

The narrow-sense heritability of continuous traits and traits transformed into normal scores was estimated as:

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

Genetic correlations were estimated by using Pearson's moment product correlation:

$$r_G=\frac{\sigma_{g_1g_2}}{\sqrt{\sigma^2_{g_1}\sigma^2_{g_2}}}$$

where $\sigma_{g_1g_2}$ is the posterior mean of additive genetic covariance between traits 1 and 2, and $\sigma^2_{g_1}$ and $\sigma^2_{g_2}$ are the posterior means of additive genetic variances for traits 1 and 2, respectively.

Independent evaluation of the prediction model was performed through 10-fold cross-validation, where a circularly one fold was proposed as the validation population, and all phenotypes associated with this fold were masked as missing values. Predicted values were then correlated with phenotypes to estimate predictive ability. Prediction accuracy was estimated as predictive ability divided by square root of traits' heritability.

RESULTS

The phenotypic data were checked for normality. While quantitative traits (HT, DBH and WD) showed distribution close to normal, the class variables STR was transformed by power of 2 to approximate the normal distribution required for mixed linear models used in this study. Stem malformation showed very low variability and most trees were scored as 9. Therefore, the MAL trait was not included in the analysis and we recommend using truncation selection for this trait before considering other traits. Additionally, we explored the effect of the seed source on phenotypic performance. We found that the seed material from Tinkers seed orchard showed an exceptional performance in WD across both sites but not for HT and DBH. While this seed material performed superior at Keen's block site, it underperformed at Fortification site (Figure 1). The stem straightness and malformation showed the poorest performance compared to all other Australian seed sources (Figure 1).

The heritabilities of the traits had similar patterns across both models (single-ste genomic evaluation vs. GBLUP) with the highest estimates obtained for WD (0.191 vs. 0.279 in Keen's block and 0.310 vs 0.247 in Fortification) and the lowest estimates obtained for DBH and HT (0.149 vs 0.117 in Keen's block and 0.246 vs 0.105 in Fortification). Genotype by environment interaction between Keens' block site and Fortification site was the strongest for HT and DBH (-0.411 vs. 0.173) and the lowest for WD (0.188 versus 0.406). Therefore, wood density is the most stable trait across investigated sites and traits (Table 1 and 2). Genetic correlations were strong between HT and DBH (0.93 – 0.99 across all scenarios) and weak between productivity traits (DBH and HT) and WD, reaching from slightly negative (-0.07 to -0.08) to slightly positive (0.113 to 0.253) in the single-step evaluation. However, they were not statistically significant in any scenario (Table 1 and 2). Stem straightness was positively correlated with HT and DBH but reached statistically significant estimates only in Keen's block (Table 1).

The predictive ability was the highest for wood density (0.172 at Keen's block and 0.184 at Fortification) which further increased when GBLUP was used instead of the single-step evaluation (0.203 in Keen's block and 0.280 in Fortification). The lowest predictive ability was found for DBH and HT reaching 0.125 in Keen's block and 0.106 – 0.152 in Fortification. When only genotyped individuals were considered in the model (GBLUP scenario) the DBH and HT predictive ability decreased to 0.082 in Keen's block and got negative (-0.05) at Fortification, probably due to low number of individuals and low or negative genetic correlations between sites. However, predictive ability in WD increased to 0.203 in Keen's block and 0.280 in Fortification (Table 3).

Since the prediction accuracy of genomic breeding values is inferred from predictive ability, it follows the same patterns. The highest prediction accuracy of genomic breeding values was reached for WD (0.408 vs. 0.341) which further increased when GBLUP was implemented (0.400 vs. 0.592). The lowest prediction accuracy was found in DBH and HT (from 0.218 to 0.338) which decreased when GBLUP model was implemented (from -0.184 to 0.256). Stem straightness showed mixed results while GBLUP method improved the prediction accuracy for the Fortification site from 0.314 to 0.496 and slightly decreased in Keen's block site from 0.402 to 0.388 (Table 3).

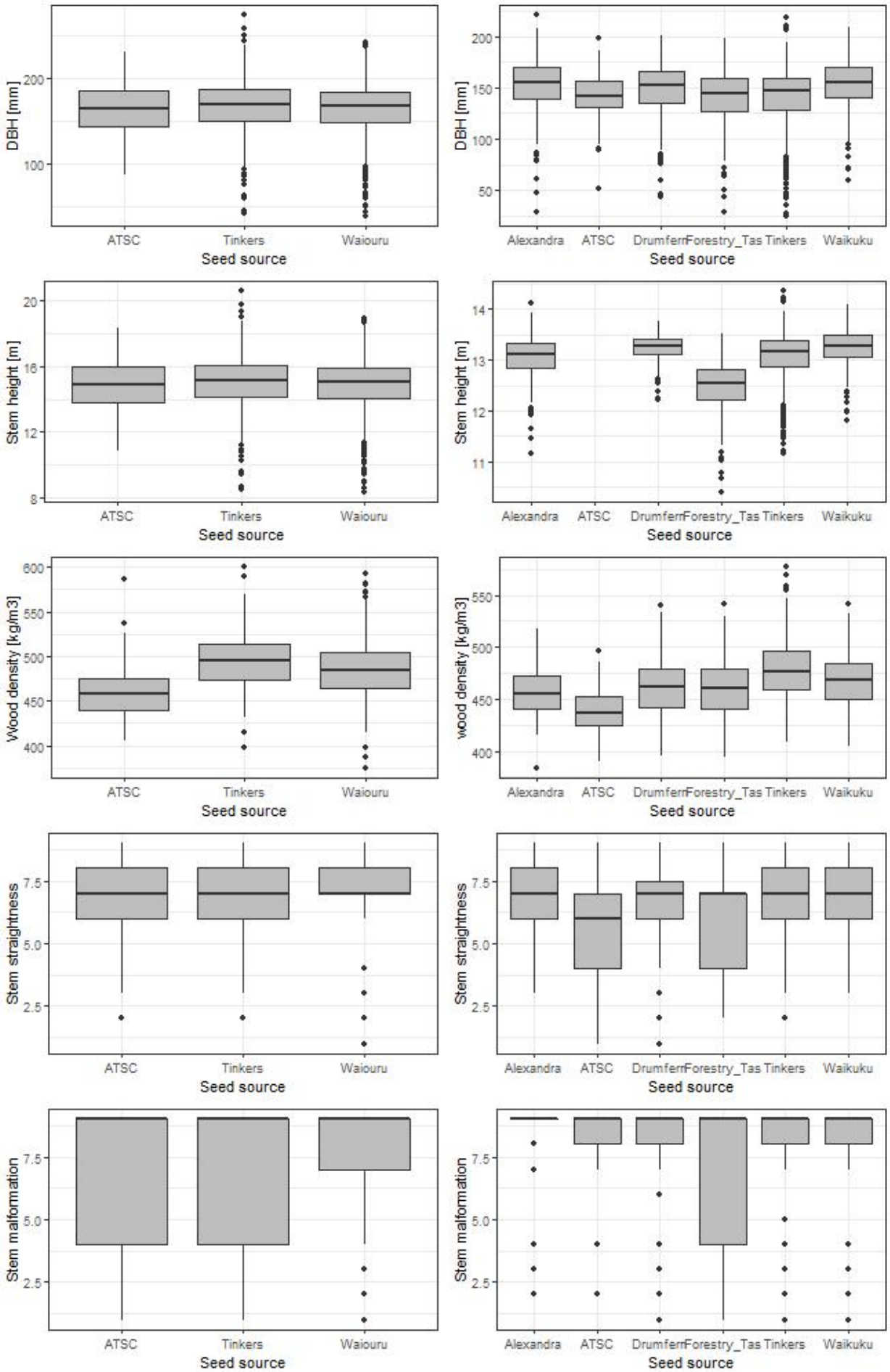

Figure 1: Seed source performance at Keen's block (left plots) and Fortification (right plots) sites.

Table 1: Estimates of genetic parameters such as heritability (diagonal elements), genetic correlations between any pair of traits (off-diagonal elements) estimated through single-step genetic evaluation

Site	Trait	Keen's block				Fortification			
		DBH	**HT**	**WD**	**STR**	**DBH**	**HT**	**WD**	**STR**
Keen's block	**DBH**	**0.149 (0.033)**							
	HT	0.997 (0.001)	**0.149 (0.033)**						
	WD	-0.089 (0.156)	-0.089 (0.156)	**0.191 (0.049)**					
	STR	0.301 (0.109)	0.301 (0.109)	0.020 (0.144)	**0.237 (0.033)**				
Fortification	**DBH**	**-0.411 (0.175)**	-0.411 (0.175)	0.179 (0.205)	-0.483 (0.121)	**0.246 (0.064)**			
	HT	-0.406 (0.173)	**-0.406 (0.173)**	0.177 (0.202)	-0.487 (0.120)	0.954 (0.013)	**0.269 (0.067)**		
	WD	-0.067 (0.223)	-0.067 (0.223)	**0.188 (0.199)**	-0.094 (0.190)	-0.084 (0.160)	-0.073 (0.157)	**0.310 (0.064)**	
	STR	-0.110 (0.205)	-0.145 (0.443)	0.183 (0.188)	**0.210 (0.173)**	0.044 (0.163)	0.034 (0.163)	0.076 (0.156)	**0.179 (0.042)**

.

Table 2: Estimates of genetic parameters such as heritability (diagonal elements), genetic correlations between any pair of traits (off-diagonal elements) estimated through GBLUP.

Site	Trait	Keen's block				Fortification			
		DBH	**HT**	**WD**	**STR**	**DBH**	**HT**	**WD**	**STR**
Keen's block	**DBH**	**0.117 (0.037)**							
	HT	0.981 (0.007)	**0.117 (0.037)**						
	WD	0.253 (0.213)	0.253 (0.213)	**0.279 (0.089)**					
	STR	-0.158 (0.191)	-0.158 (0.192)	-0.155 (0.208)	**0.204 (0.044)**				
Fortification	**DBH**	**0.173 (0.243)**	0.172 (0.243)	0.061 (0.262)	-0.330 (0.199)	**0.105 (0.040)**			
	HT	0.175 (0.243)	**0.174 (0.244)**	0.063 (0.261)	-0.331 (0.197)	0.933 (0.028)	**0.106 (0.040)**		
	WD	0.057 (0.238)	0.057 (0.239)	**0.406 (0.227)**	-0.221 (0.206)	0.114 (0.243)	0.113 (0.241)	**0.247 (0.069)**	
	STR	-0.168 (0.228)	0.097 (0.411)	-0.099 (0.249)	**0.398 (0.178)**	-0.335 (0.197)	-0.336 (0.198)	-0.094 (0.219)	**0.194 (0.058)**

Table 3: Predictive ability and prediction accuracy of genomic breeding values using single-step genomic model (the results from GBLUP analysis are in brackets).

Site	Trait	Predictive ability	Prediction accuracy
Keen's block	DBH	0.125 (0.083)	0.338 (0.259)
	HT	0.125 (0.082)	0.338 (0.256)
	WD	0.172 (0.203)	0.408 (0.400)
	STR	0.193 (0.169)	0.402 (0.388)
Fortification	DBH	0.106 (-0.058)	0.218 (-0.184)
	HT	0.152 (-0.056)	0.299 (-0.177)
	WD	0.184 (0.280)	0.341 (0.592)
	STR	0.130 (0.210)	0.314 (0.496)

CONCLUSION

This study implemented genomic data generated across two genotyping platforms. This greatly reduced the numbers of usuable markers in order to avoid inconsistent patterns across platforms. Thus, we suggest genotyping several individuals on both platforms to achieve a higher confidence of markers with consistent patterns across platforms.

The comparison of seed sources identified Tinkers seed orchard seeds as superior for wood quality across both sites but with large change in ranking regarding the productivity at each site. Australian seed sources performed relatively poor compared to seed sources based on selection in New Zealand environments.

The current study estimated genetic parameters such as heritability and genetic correlations between traits and tested sites. It identified moderate heritability across all tested traits therefore the potential for genetic improvement is present. Additionally, genetic correlations estimated between traits showed that DBH and HT were strongly correlated but not with WD.
Stem straightness had a strong positive correlation with growth traits in the Keen's block site but no relationship in the Fortification site. Also a strong GxE interaction across growing sites, especially for growth traits, was found.

The implementation of genomics and single-step genomic evaluation approach allowed for prediction of genomic breeding values for non-phenotyped individuals and parents at each site. Therefore, the genetic thinning/culling might be implemented especially for wood density showing higher prediction accuracy in current seed orchards using predicted genomic breeding values.

ACKNOWLEDGEMENTS

We would like to thank to John Filmer from Southwood export for great help with tissue sample collection for DNA extraction.

REFERENCES

Cheng H, Fernando RL, Garrick DJ (2018). JWAS: Julia implementation of whole-genome analysis software. Proceedings of the World Congress on Genetics Applied to Livestock Production,11.859. Auckland, New Zealand.

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE 6(5):e19379.

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

Gelman A, Rubin DB (1992). Inference from iterative simulation using multiple sequences. Statistical Science 7:457-472.

Goddard M (2009). Genomics selection: prediction of accuracy and maximisation of long term response. Genetica 136:245-257.

Grattapaglia D, Resende MDV (2011). Genomic selection in forest tree breeding. Tree Genetics and Genomes 7:241-255.

Habier D, Fernando RL, Garrick DJ (2013). Genomic BLUP decoded: a look into the black box of genomic prediction. Genetics 194(3):597-607.

Klapste J, Suontama M, Dungey H, Telfer J, Graham N, Low C, Stovold T (2018). Effect of hidden relatedness on single-step genetic evaluation in an advanced open-pollinated breeding program. Journal of Heredity 109(7):802-810.

Klapste J, Ashby R, Telfer E, Graham N, Dungey H, Drauning R, Clarke S, Dodds K (2021). The use of “genotyping-by-sequencing” to recover shared genealogy in genetically diverse eucalyptus population. Forests 12(7):904.

Legarra A, Aguilar I, Misztal I (2009). A relationship matrix including full pedigree and genomic information. Journal of Dairy Science 92:4656–4663.

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

Misztal I, Legarra A, Aguilar I (2009). Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. Journal of Dairy Science 92:4648–4655.

Powell JE, Visscher PM, Goddard ME (2010). Reconciling the analysis of IBD and IBS in complex trait studies. Nature Review Genetics 11(11):800-805.

Silva-Junior OB, Faria DA, Grattapaglia D (2015). A flexible multi-species genome-wide 60k SNP chip developed from pooled resequencing of 240 Eucalyptus tree genomes across 12 species. New Phytologist 206(4):1527-1540.

Speed D, Balding DJ (2015). Relatedness in the poste-genomic era: it is still useful? Nature Review Genetics 16:33-44.

Suontama M, Klapste J, Telfer E, Graham N, Sovold T, Low C, McKinley R, Dungey H (2019). Efficiency of genomic prediction across two Eucalyptus nitens seed ochards with different selection histories. Heredity 122(3):370-379.

VanRaden PM (2008). Efficient methods to compute genomic predictions. Journal of Dairy Science 91(11):4414-4423.

APPENDIX

Breeding values estimated for each individual at each site (please contact FGR for spreadsheet).