Specialty Wood Products Research Partnership

SCION FORESTS ▪ PRODUCTS ▪ INNOVATION

# Implementation of genomic selection in provenance/progeny test of Douglas-fir

Authors: Jaroslav Klápště, Gancho Slavov, Toby Stovold, Heidi Dungey

**Date:** June 2019

Publication No: SWP-T-084

FOREST GROWERS RESEARCH

# TABLE OF CONTENTS

EXECUTIVE SUMMARY ..... 1
INTRODUCTION ..... 2
METHODS ..... 4
RESULTS ..... 8
CONCLUSION ..... 14
ACKNOWLEDGEMENTS ..... 14
REFERENCES ..... 14



FOREST GROWERS RESEARCH

# EXECUTIVE SUMMARY

Two Douglas-fir progeny trials planted in Kaingaroa and Gowan Hill in 1996 were assessed in 2007 and re-assessed in February 2017. Since this material contains genetically broad material coming from populations across Oregon and California, these two progeny trials were also selected as training populations in a genomic selection project for Douglas-fir.

Since the bussines model connected to Douglas-fir marker array allows for genotyping of each individual for 75 USD, the strategy for genotyping was based on a partial genotyping effort equally distributed across all open-pollinated families (~2,200 individuals in total) and thus a single-step genetic evaluation approach combining phenotypic, pedigree and genomic data was implemented in this study.

Implementation of genomic resources developed by Oregon State University in the evaluation of New Zealand Douglas-fir breeding populations in this study resulted in substantially improved prediction accuracy and response to selection compared with pedigree-based analysis. An additional increase in the response to selection was found when only ancestry informative markers were used in the analysis of traits with strong population differentiation (e.g. DBH). Thus, the currently available SNP array appears to be a useful genotyping platform for the New Zealand Douglas-fir breeding program.

Implementation of the metafounders approach (i.e. inference of relatedness between pedigree founders of the provenance/progeny trial) resulted in increased prediction accuracy not only for genotyped but also for non-genotyped individuals. However, more complex modelling of population demography resulted in a reduction in model fit and lower prediction accuracy compared with a simple model with a single metafounder population. Therefore, reliable modelling of population structure in forest trees is challenging, even with the availability of abundant genetic marker data. The important finding from this study is that consideration of the distance of populations from native populations is important when building an implementation strategy for genomics.

# INTRODUCTION

Forest tree populations grow under a variety of environmental conditions, often along steep gradients of temperature and precipitation. Divergent selection has therefore resulted in predictable patterns of population genetic structure for putatively adaptive traits, particularly vegetative and flowering phenology, resulting in reproductive asynchrony (Levin, 1995). The efficient management of genetic resources under changing climates requires the capture of maximum genetic diversity across a species' native distribution. Initial screening is usually performed through the establishment of provenance trials in common gardens, with samples collected across different populations with or without familial structure (Matyas, 1996; Mátyás, 1994). In addition to providing descriptive information about genetic variation, the establishment of multiple common gardens across a variety of environments allows the construction of universal response functions, which can be used to match populations to environments (Wang *et al*, 2006; Wang *et al*, 2010), as well as to guide assisted migration to mitigate the effects of climate change (Aitken and Whitlock, 2013).

The integration of breeding programs into the management of genetic resources for future climates is critical for the maintenance of resilient genetic progress in adaptive as well as non-adaptive but economically important attributes (Borralho and Dutkowski, 1998). Most forest tree breeding programmes are in an early phase of domestication due to the late onset of sexual maturity and delayed expression of economically important traits. Under such circumstances, evolutionary forces such as migration, random drift and natural selection can affect genetic parameters estimates obtained in the early generations of breeding cycles. Since mixed linear models (MLMs) are the preferred statistical tool in genetic evaluations, the fitting of provenance effects as a fixed or random term in MLMs was proposed as a viable solution for the capture of local genetic differences caused by these evolutionary forces (Ugarte *et al*, 1992). Alternatively, the delineation of contemporary genetic groups enables the estimation of relatedness between founders of documented pedigrees to correct for differences between genetically diverse groups of individuals (Westell *et al*, 1988).

The development of genomic resources has impoved our understanding of the population genetic structure captured in breeding/conservation programmes. The tracking of coancestry (relatedness) between individuals in breeding populations improves the precision of genetic parameter estimates in three ways. First, realised relationship coefficients do not follow Mendelian expectations exactly due to random genome sampling errors. Second, individuals from different families are not necessarily unrelated (i.e. due to shared or related ancestros). Finally, pedigree records are not always accurate.

Progeny tests often include up to tens of thousands of individuals, and genotyping the whole population can be costly. In such cases, a combination of pedigree and genetic marker information is recommended thorugh an approach called "single-step evaluation" (Legarra *et al*, 2009; Misztal *et al*, 2009). The approach is based on blending a marker-based into a pedigree-based relationship matrix, and the resulting combined relationship matrix is then used in 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). However, this approach assumes that the base population of the pedigree is well defined, without any hidden relatedness or population structure. This assumption is problematic in the early generations of forest tree breeding cycles, and explicit modelling of the relatedness between pedigree founders can improve the prediction accuracy of single-step genetic evaluation. Consistent with this, modelling population structure increased the precision of breeding values in a previous analysis of a Douglas-fir provenance/progeny test (Klapste *et al*, 2019).

Douglas-fir is the second most important conifer plantation species in New Zealand (i.e. after radiata pine), with a current planted area of ca. 104,000 ha. The economic feasibility of Douglas-fir plantations, however, depends crucially on genetic improvement, with productivity and stem defects being the two most important traits (Magalska and Howe, 2014), followed by resistance to Swiss Needle Cast (Dungey *et al*, 2012). Provenance tests performed on a broad

geographical scale in North America found that local populations grew best, with decreasing performance when the distance between the test site and population origin exceeded 435 km in latitude or 370 m in elevation (Terrance and Jayawickrama, 2014). However, this is not always the case, and in some instances, geographically distant provenances perform as well as the local one (Krakowski and Stoehr, 2009).

Our study evaluated the efficiency of the exome capture based SNP array in New Zealand Douglas-fir population. The SNP array was developed based on transcriptome assembly covering around 25,000 unique gene models (Howe *et al*, 2013). In addition, we tested the effect of accounting for relatedness between pedigree founders on single-step evaluation.

# METHODS

## Material

The materials used to plant at two New Zealand environments (Kaingaroa, latitude 38° 17' S and Gowan Hill, latitude 45° 52' S) in 1996 were collected from populations in two US regions (California and Oregon), ranging in latitude from 36° to 48° N along the western coast of the USA. Each experiment included 30 replications of 7 sets, with each set containing 34 open-pollinated families and two controls. Provenances contributed equally to each set. A detailed description of the materials was provided in a previous study (Dungey *et al*, 2012). Trees were measured for diameter at breast height (DBH, measured in mm), acoustic wave velocity (time of flight between two probes placed on the tree around breast height, VEL, measured in km/s) as asurrogate trait for wood density using the Hitman ST300 (Fibre-gen, Christchurch, New Zealand), straightness (STR, scored on a scale of 1 to 9, (Carson, 1986) and needle retention (NR) as screening for resistance to Swiss needle cast at the age of 21 years. Needle retention was measured only at Kaingaroa due to the favourable conditions for disease outbreak present at this site and scored on 6-grade scale, reflecting the damage of needles of different ages. STR and NR (class variables) were Z-transformed into a normal score (Gianola and Norton, 1981).

Genomic data were generated through the 58K Axium array developed using exome capture sequencing (Howe *et al*, 2013). Markers were filtered using the default filter developed by an external provider (Neogen, Nebraska US), but the threshold for call rate was set at 0.8 instead of the default value of 0.97. In addition, all markers were filtered for minor allele frequency (MAF) > 0.01. Missing data were imputed as genotype means. In total, 25,170 SNPs were used in all downstream analyses for this scenario. Since the exome presumably represents the most evolutionary constrained part of the genomme, an *ad hoc* strategy was implemented for marker selection. Markers were first filtered for MAF > 0.1, and then each locus was checked for its informativeness to infer ancestry (Rosenberg *et al*, 2003) as follows:

$$I = \sum_{j=1}^{N} \left( -p_j \log(p_j) + \sum_{i=1}^{K} \frac{p_{ij}}{K} \log(p_{ij}) \right)$$

where N was the number of alleles (all SNPs were bi-allelic, so N = 2 in our case), $p_j$ was the frequency of the $j^{th}$ allele across all populations, K was the number of populations and $p_{ij}$ was the frequency of the $j^{th}$ allele in the $i^{th}$ population. The top 15,000 markers were selected based on their *I* index for genomic selection and inference of relatedness between founders.

## Statistical analysis

### Recovery of hidden relatedness in the genotyped sample

Because the population was established from open-pollinated seed collected in the wild stands, sib-ship reconstruction was performed to recover hidden familial structure in the sample of genotyped individuals using the COLONY program (Jones and Wang, 2010) using 500 randomly selected markers. Because conifers are investing a considerable amount of energy into reproductive outcome, polyandry and polygamy were assumed in our analysis. Furthermore, non-negligible rates of selfing have been reported in several previous studies of Douglas-fir seed orhards (Slavov *et al*, 2005; Song *et al*, 2018). Therefore, the non-zero rates of inbreeding were assumed in these analyses. The prior for maternal family size was set as 10, reflecting the genotyping effort, and the prior paternal size was set at 3, assuming that fathers contribute to multiple offspring. The relationship matrix between genotyped individuals was initially constructed based on half-sib structure inferred from sib-ship reconstruction with a confidence of 0.9. This matrix was used for K-means clustering with the number of clusters (K) equal to the number of open-pollinated families. The analysis was performed by using “kmeans” function implemented in “stats” package in R programming environment (Team, 2018). Pedigree errors were identified and corrected based on on the results from this analysis.

In addition, phantom parents (fathers) were added for individuals identified as full-sibs. These phantom fathers were assumed to be from the same provenance as mothers. The updated pedigree information, along with contemporary genetic groups included in the pedigree, was used

in the pedigree-based analysis to obtain genetic parameters and estimated breeding values (EBV). A MLM implemented in the statistical package "ASReml-R" (Butler *et al*, 2009) was used to analyse phenotypic data as follows:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zg} + \boldsymbol{Zr} + \boldsymbol{Zrs} + \boldsymbol{e}$$

where y was the vector of measurements, $\boldsymbol{\beta}$ was the vector of fixed effects containing the overall mean, g was the vector of additive genetic effects following $var(\boldsymbol{g}) \sim N(0, \boldsymbol{A}\sigma_g^2)$, where $\boldsymbol{A}$ was the average of the numerator relationship matrix (Wright, 1922) and $\sigma_g^2$ was the additive genetic variance. The design effects were represented by a vector of random replication effects following $var(\boldsymbol{r}) \sim N(0, \boldsymbol{I}\sigma_r^2)$, where $\boldsymbol{I}$ was the identity matrix and $\sigma_r^2$ was the replication variance, and by the vector of sets nested within replications following $var(rs) \sim N(0, \boldsymbol{I}\sigma_{rs}^2)$, where $\sigma_{rs}^2$ was the set nested within replication variance. Similarly, $\boldsymbol{e}$ was the vector of random residuals following $var(\boldsymbol{e}) \sim N(0, \boldsymbol{I}\sigma_e^2)$, where $\sigma_e^2$ was the residual variance. The narrow-sense heritability was estimated as follows:

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}$$

and the theoretical accuracy of breeding values was estimates as:

$$r = \sqrt{1 - \frac{PEV}{\sigma_g^2}}$$

where PEV was the prediction error variance (Mrode, 2014) estimated as the square of standard errors for breeding value estimates.

**Relatedness between metafounders**

Three scenarios of relatedness between metafounders were investigated through the implementation of a generalized least squares (GLS) method (Garcia-Baccino *et al*, 2017).

The first scenario considered a single population represented by the coefficient of relatedness γ in metafounders that was estimated as follows:

$$\gamma = 2\sigma_\mu^2$$

where $\sigma_\mu^2$ was the variance in genotype mean estimates across all loci. The genotype mean was estimated separately for each locus as follows:

$$\hat{\mu} = (1' \boldsymbol{A}_{22}^{-1} 1)^{-1} 1' \boldsymbol{A}_{22}^{-1} m_i$$

where $\boldsymbol{A}_{22}$ was the pedigree-based relationship matrix for genotyped individuals (derived from sib-ship reconstruction) and $\boldsymbol{m}_i$ was the vector of genotypes for teh i$^{th}$ locus in terms of allele dosage (0, 1 and 2).

The second scenario assumed multiple populations with no crosses between populations. This scenario used a matrix **Γ**, representing ancestral relationships between pedigree founders, which was estimated as follows:

$$\boldsymbol{\Gamma} = 2 \begin{bmatrix} \sigma^2_{\mu^{[1]}\mu^{[1]}} & \cdots & \sigma_{\mu^{[1]}\mu^{[n]}} \\ \vdots & \ddots & \vdots \\ \sigma_{\mu^{[n]}\mu^{[1]}} & \cdots & \sigma^2_{\mu^{[n]}\mu^{[n]}} \end{bmatrix}$$

where $\sigma^2_{\mu^{[1]}\mu^{[1]}}$ and $\sigma^2_{\mu^{[n]}\mu^{[n]}}$ were the variances of genotype mean estimates in the 1st and nth population, $\sigma_{\mu^{[1]}\mu^{[n]}}$ and $\sigma_{\mu^{[n]}\mu^{[1]}}$ were the covariances between genotype mean estimates from the 1st and nth populations.
Genotype means were estimated as:

$$\hat{\mu}^{[b]} = (1'\boldsymbol{A}_{22}^{[b]-1}1)^{-1}1'\boldsymbol{A}_{22}^{[b]-1}\boldsymbol{m}_i^{[b]}$$

where $\boldsymbol{A}_{22}^{[b]}$ was the pedigree-based relationship matrix for genotyped individuals coming from the bth population and $\boldsymbol{m}_i^{[b]}$ was the vector of genotypes for the ith locus for individuals coming from the bth population.

The third scenario assumed multiple populations with crosses between populations. In this case, genotype means were estimated simultaneously for all populations as follows:

$$\widehat{\boldsymbol{\mu}} = (\boldsymbol{Q}'\boldsymbol{A}_{22}^{-1}\boldsymbol{Q})^{-1}\boldsymbol{Q}'\boldsymbol{A}_{22}^{-1}\boldsymbol{m_i}$$

where ***Q*** represented the matrix of ancestry fractions coming from each population for each genotyped individual, with columns representing ancestral populations and rows representing each individual. The sum of each row was equal to 1. The ***Q*** matrix was constructed by assigning a coefficient of 0.8 to the maternal population (i.e. ancestral population of mother) of each individual and 0.1 to the two closest provenances to the North and South of the maternal population. In the case of marginal provenances, a coefficient 0.2 was assigned to the closest provenance. This strategy was adopted based on previous results, indicating that assuming a mixture of ancestral populations resulted in better model fit (Klapste *et al*, 2019). However, when sib-ship reconstruction suggested that the father of a particular individual was coming from the same provenance as the mother, the ***Q*** matrix was assigned a coefficient of 1 for the parental provenance and 0 for all other populations.

**Implementation of relatedness between metafounders in the pedigree-based relationship matrix**

The documented pedigree-based relationship matrix was modified to infer relatedness between metafounders. For the single-population scenario, the pedigree-based relationship matrix was modified as follows:

$$\boldsymbol{A}^{\gamma} = \boldsymbol{A}\left(1 - \frac{\gamma}{2}\right) + \gamma\boldsymbol{J}$$

where ***A*** was the documented pedigree-based relationship matrix, γ was the relatedness between pedigree founders, and ***J*** was a matrix of 1's. For pedigree founders from multiple populations, the documented-pedigree based relationship matrix was modified as follows:

$$\boldsymbol{A}^{\Gamma} \approx \boldsymbol{A}\big(\boldsymbol{I} - 0.5diag(\boldsymbol{Q\Gamma Q}')\big) + \boldsymbol{Q\Gamma Q}'$$

This step was performed using software developed by Legarra et al. (Legarra *et al*, 2015), available at: https://github.com/alegarra/metafounders.

**Single-step genetic evaluation**

Phenotypic values were corrected for design effects to reduce the computational burden associated with the implementation of Bayesian approaches using a MLM implemented in the "ASReml-R" package (Butler *et al*, 2009) as follows:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{e}$$

where ***y*** was a vector of phenotypes, ***β*** was the vector of fixed effects including overall mean, replication and set nested within replication and ***e*** was the vector of residual effects with $var(\boldsymbol{e}) \sim N(0, \boldsymbol{I}\sigma_e^2)$, where $\sigma_e^2$ was the residual variance, and ***X*** was the index matrix associating effects from vector ***β*** to phenotypes in vector ***y***.

The single-step genetic evaluation was performed using a MLM based on a Gibbs sampling algorithm implemented in the "BGLR" R package (Pérez and de Los Campos, 2014) as follows:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zg} + \boldsymbol{e}$$

where ***y*** was a vector of phenotypes, ***β*** was the vector of fixed effects including overall mean, ***g*** was a vector of genomic estimated breeding values following $var(\boldsymbol{g}) \sim MVN(0, \boldsymbol{H}\sigma_g^2)$, where $\sigma_g^2$ was the additive genetic variance associated with relatedness inferred from the combination of pedigree and genomic information following the default setting $\sim \chi^{-2}(df = 5, S = var(y) * 0.5)$, and ***H*** was the combined relationship matrix. ***H*** was constructed using the pedigree-based relationship matrix modified for relatedness between pedigree founders either considering only one population $\boldsymbol{A}^{\gamma}$ or multiple populations defined by provenances $\boldsymbol{A}^{\Gamma}$ and the marker-based relationship matrix on the original scale (no rescaling step has been implemented) as follows:

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

where $\boldsymbol{A}_{11}^{\Gamma}$ was the pedigree-based relationship matrix for non-genotyped individuals, $\boldsymbol{A}_{22}^{\Gamma}$ was the pedigree-based matrix for genotyped individuals, $\boldsymbol{A}_{12}^{\Gamma}$ and $\boldsymbol{A}_{21}^{\Gamma}$ were the pedigree-based matrices between genotyped and non-genotyped individuals, and ***G*** was 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*** was the genotype matrix coding reference allele homozygote as 0, heterozygote as 1 and alternative allele homozygote as 2 and ***P*** was double the frequency for the alternative allele.

Independent evaluation of the prediction model was performed through jackknifing with fixed variance components following the approach proposed by (Gianola and Schön, 2016). Predictive ability was estimated as the correlation between predicted genomic breeding values (GEBV) and corrected phenotypes cor(GEBV, y). Prediction accuracy was estimated in two ways: CV1 was the correlation between GEBV and pedigree-based breeding values cor(GEBV, EBV); and CV2 was the predictive ability divided by the square root of heritability:

$$r_p = \frac{cor(GEBV, y)}{\sqrt{h^2}}$$

# RESULTS

From a total of 58,350 markers, 24,010 (41%) had high quality and included all genotype classes (reference homozygote, heterozygote, alternative homozygote), while only 2.6% had high quality but were monomorphic (Table 1). Around 20% of markers showed only two genotype classes, without minor allele homozygotes. In total, 35,584 markers passed initial filtering (PolyHighResolution + MonoHighResolution + NoMinorHom), and 25,171 markers remained after filtering for minor allele frequency (>0.01). The “OTV” confersion type is group of markers which are usually highly reproducible but interfere with the SNPs of interest. The “other” represent markers which do not fit any of other categories and usually suffer from low call rates.

Finally, markers were assessed for their informativeness in inferring ancestry, and the top 15,000 markers based on *I* index were selected for testing in one of our alternative scenarios (AIM).

| Conversion type | Count | Percentage |
|---|---|---|
| **PolyHighResolution** | 24010 | 41.15 |
| **Other** | 16554 | 28.37 |
| **MonoHighResolution** | 1536 | 2.63 |
| **NoMinorHom** | 11182 | 19.16 |
| **CallRateBelowThershold** | 16 | 0.03 |
| **OTV** | 5052 | 8.66 |

**Table 1**: Performance of 58K Axium array on the New Zealand population

Based on principal component analysis of marker-based relationship matrices, the scenario using only ancestry informative markers had a slightly higher proportion of total variance explained by first and second principal (Figure 1).

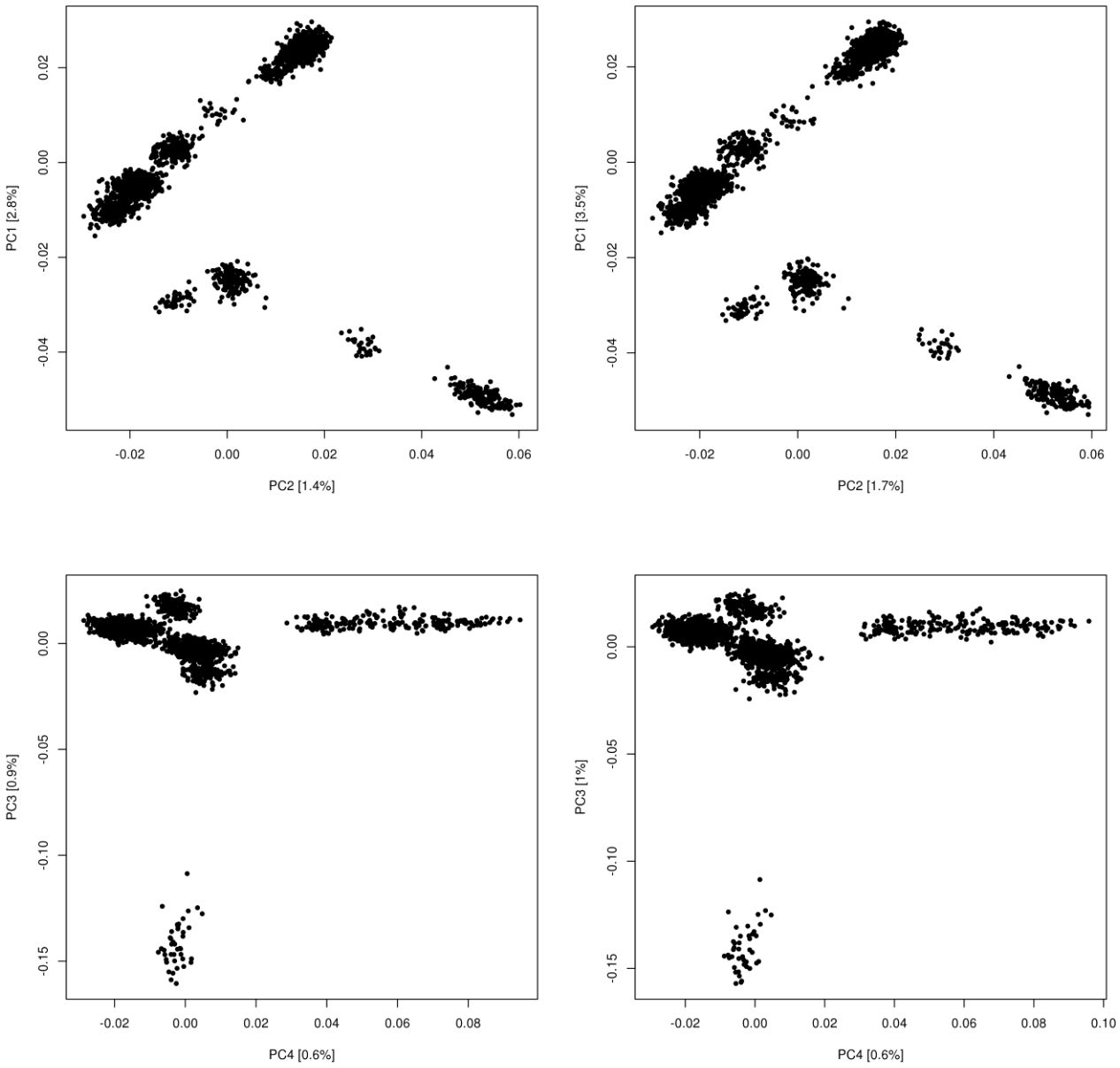


**Figure 1**: First and second principal components (upper row) as well as third and fourth principal component (bottom row) from spectral decomposition of marker-based relationship matrices using all markers (left side) and only ancestry informative markers (right side).

| Model | Gowan Hill | | | Kaingaroa | | | |
|---|---|---|---|---|---|---|---|
| | DBH | STR | VEL | DBH | STR | NR | VEL |
| Add.gen.var | 842 (134) | 0.13 (0.02) | 0.04 (0.01) | 949 (194) | 0.11 (0.03) | 0.05 (0.02) | 0.03 (0.01) |
| Rep var | 81.4 (26.5) | 0.01 (0.00) | 0.04 (0.01) | 80.9 (31.3) | 0.11 (0.03) | 0.25 (0.07) | 0.02 (0.01) |
| Rep(set) var | 11.6 (14.2) | 0.00 (0.00) | 0.00 (0.00) | 20.7 (27.7) | 0.01 (0.01) | 0.02 (0.01) | 0.01 (0.01) |
| Resid var | 2546 (90.9) | 0.21 (0.01) | 0.08 (0.00) | 2751 (136) | 0.41 (0.02) | 0.30 (0.01) | 0.11 (0.01) |
| $h^2$ | 0.25 (0.04) | 0.40 (0.04) | 0.33 (0.04) | 0.26 (0.05) | 0.21 (0.05) | 0.15 (0.04) | 0.22 (0.04) |
| r | 0.67 | 0.72 | 0.66 | 0.63 | 0.43 | 0.42 | 0.54 |
| logL | -24744 | 580 | 2770 | -13475 | -481 | -25.2 | 1368 |

**Table 2:** Variance components, heritability (standard errors in parenthesis), accuracy of breeding value estimates and log-likelihoods from models using an updated pedigree based on results from sibship reconstruction.

Genetic evaluation using the updated pedigree information (i.e. based on sibship reconstruction) resulted in statistically significant additive genetic variances and heritability estimates for all traits. Heritability ranged from 0.15 for needle retention to 0.40 for straightness at Gowan Hill. Similarly, the accuracy of breeding values ranged from 0.42 for needle retention to 0.72 for straightness at Gowan Hill (Table 2).

| Model | Pop | Gowan Hill | | | Kaingaroa | | | |
|---|---|---|---|---|---|---|---|---|
| | | DBH | STR | VEL | DBH | STR | NR | VEL |
| | NG | NA | NA | NA | NA | NA | NA | NA |
| ABLUP | G | NA | NA | NA | NA | NA | NA | NA |
| | T | 0.30 (0.73,0.60) | 0.19 (0.39,0.30) | 0.23 (0.58,0.40) | 0.34 (0.78,0.67) | 0.13 (0.54,0.28) | 0.16 (0.77,0.41) | 0.15 (0.52,0.32) |
| | DIC | 59868.8 | 9195.7 | 4062.2 | 32596.8 | 6366.5 | 5885.1 | 2913 |
| | NG | 0.37 (0.84,0.74) | 0.35 (0.68,0.55) | 0.32 (0.76,0.56) | 0.41 (0.85,0.80) | 0.21 (0.69,0.46) | 0.20 (0.73,0.52) | 0.28 (0.74,0.60) |
| HBLUP1 | G | 0.42 (0.69,0.84) | 0.42 (0.62,0.66) | 0.38 (0.65,0.66) | 0.39 (0.66,0.77) | 0.27 (0.46,0.59) | 0.18 (0.41,0.47) | 0.27 (0.53,0.58) |
| | T | 0.38 (0.80,0.76) | 0.37 (0.67,0.59) | 0.34 (0.72,0.59) | 0.41 (0.77,0.80) | 0.23 (0.59,0.50) | 0.19 (0.59,0.49) | 0.28 (0.65,0.60) |
| | DIC | 56269.9 | 8583.4 | 3802.6 | 32503.3 | 6315.6 | 5860.7 | 2812.7 |
| | NG | 0.36 (0.85,0.72) | 0.36 (0.68,0.57) | 0.32 (0.76,0.56) | 0.41 (0.87,0.80) | 0.20 (0.73,0.44) | 0.19 (0.83,0.49) | 0.28 (0.78,0.60) |
| HBLUP2 | G | 0.41 (0.79,0.82) | 0.40 (0.63,0.63) | 0.38 (0.72,0.66) | 0.40 (0.78,0.78) | 0.29 (0.58,0.63) | 0.20 (0.68,0.52) | 0.26 (0.61,0.55) |
| | T | 0.37 (0.84,0.74) | 0.37 (0.67,0.59) | 0.33 (0.75,0.58) | 0.41 (0.84,0.80) | 0.24 (0.67,0.52) | 0.19 (0.77,0.49) | 0.27 (0.71,0.58) |
| | DIC | 56968.8 | 8607.6 | 3865.4 | 32500.7 | 6323.7 | 5869.6 | 2809.7 |
| | NG | 0.36 (0.86,0.72) | 0.35(0.69,0.55) | 0.31 (0.76,0.54) | 0.41 (0.88,0.80) | 0.19 (0.74,0.42) | 0.20 (0.85,0.52) | 0.28 (0.78,0.72) |
| HBLUP3 | G | 0.41 (0.80,0.82) | 0.40 (0.63,0.63) | 0.37 (0.72,0.64) | 0.40 (0.80,0.79) | 0.29 (0.60,0.63) | 0.21 (0.71,0.54) | 0.27 (0.62,0.58) |
| | T | 0.37 (0.84,0.74) | 0.36 (0.67,0.57) | 0.33 (0.75,0.58) | 0.41 (0.85,0.80) | 0.23 (0.68,0.50) | 0.21 (0.79,0.54) | 0.28 (0.72,0.60) |
| | DIC | 57196 | 8643.9 | 3873.9 | 32502.4 | 6326.4 | 5869.4 | 2814.7 |

**Table 3**: Predictive ability and accuracy (in parenthesis: CV1 - left, CV2 - right) for non-genotyped individuals (NG), genotyped individuals (G) and across all individuals (T) for all phenotypic traits. Model fit was assessed through the Deviation Information Criterion (DIC) for the scenario using all genetic markers.

Across all investigated traits, the Deviation Information Criterion (DIC) indicated a large improvement in model fit when single-step evaluation was implemented compared with pedigree-based analysis. Further improvement in model fit was achieved when only ancestry informative markers were used for the inference of relatedness between pedigree founders, as well as for the construction of the marker-based relationship matrix between all genotyped individuals.

Surprisingly, the best model fit was observed in a scenario using only single metafounder population (HBLUP1) compared to scenarios using multiple populations with or without gene flow (HBLUP2 and HBLUP3) (Table 3 and 4). Therefore, a simple estimation of relatedness and population structure between pedigree founders performed better than scenarios based on specific assumptions about the complexity and dynamics of gene flow among provenances. A deeper

understanding of population structure and gene flow may be required to take advantage of modelling more complex metafounder scenarios.

Predictive ability followed a similar pattern, improving greatly for both genotyped and non-genotyped individuals in single-step evaluation compared with pedigree-based analyses. As expected, additional improvements in predictive ability were observed for genotyped individuals, as genetic markers capture relatedness much better than the rather shallow pedigree information available at the current stage of the breeding programme. Surprisingly, there was a decrease in predictive ability in genotyped individuals at Kaingaroa when all markers were used compared with the scenario using the only ancestry informative markers (AIM) (Table 3 and 4). Thus, a screen for marker informativeness seems to be a sensible way to increase the efficiency of single-step genetic evaluation.

Estimates of prediction accuracy were highly influenced by the method used (CV1 versus CV2). Following the CV1 method, the correlation between predicted genomic breeding values (GEBV) and pedigree-based estimated breeding values (EBV) was lower for genotyped compared with non-genotyped individuals. However, the simple rescaling of predictive ability by the square root of heritability reversed this trend. This could be an artefact driven by the quality of EBV estimates. The shallow pedigree records provide very limited information about related individuals, which ultimately leads to shrinking breeding values toward the family mean. Thus, while genetic markers can to some extent track Mendelian segregation, pedigree-based predictions are able to only predict family means, which results in higher correlations between GEBV and EBV for non-genotyped individuals in a single-step evaluation. Therefore, the Mendelian segregation captured by genetic markers is reflected only in predictive ability in this case, which appears to not be affected by the limitations of breeding values estimated in a population with shallow pedigree information. Prediction accuracy estimates based on CV1 appeared to be more inflated for the traits with lowest heritabilities (NR) and less so for traits with higher heritabilities (STR).

The expected response to selection was generally higher using single-step evaluation compared with the pedigree-based analysis (Figure 2). A further increase in the response to selection was archived when only markers that were informative for ancestry were used in the single-step evaluation (Figure 3), especially for traits with stronger phenotypic differences between provenances (e.g. DBH).

| | | Gowan Hill | | | Kaingaroa | | | |
|---|---|---|---|---|---|---|---|---|
| **Model** | **Pop** | **DBH** | **STR** | **VEL** | **DBH** | **STR** | **NR** | **VEL** |
| | **NG** | NA | NA | NA | NA | NA | NA | NA |
| **ABLUP** | **G** | NA | NA | NA | NA | NA | NA | NA |
| | **T** | 0.30 (0.73,0.60) | 0.19 (0.39,0.30) | 0.23 (0.58,0.40) | 0.34 (0.78,0.67) | 0.13 (0.54,0.28) | 0.16 (0.77,0.41) | 0.15 (0.52,0.32) |
| | **DIC** | 59868.8 | 9195.7 | 4062.2 | 32596.8 | 6366.5 | 5885.1 | 2913.6 |
| | **NG** | 0.37 (0.84,0.74) | 0.37 (0.69,0.59) | 0.32 (0.76,0.56) | 0.40 (0.85,0.79) | 0.19 (0.68,0.42) | 0.20 (0.73,0.52) | 0.28 (0.75,0.60) |
| **HBLUP1** | **G** | 0.43 (0.72,0.86) | 0.42 (0.63,0.66) | 0.40 (0.68,0.70) | 0.43 (0.70,0.84) | 0.29 (0.51,0.63) | 0.20 (0.45,0.52) | 0.29 (0.57,0.62) |
| | **T** | 0.38 (0.81,0.76) | 0.38 (0.66,0.60) | 0.34 (0.74,0.59) | 0.42 (0.79,0.82) | 0.23 (0.61,0.50) | 0.20 (0.61,0.52) | 0.29 (0.68,0.62) |
| | **DIC** | 55811.3 | 8513.0 | 3786.9 | 32474.6 | 6297.6 | 5858.0 | 2797.6 |
| | **NG** | 0.36 (0.86,0.72) | 0.36 (0.69,0.57) | 0.31 (0.76,0.54) | 0.41 (0.88,0.80) | 0.21 (0.74,0.46) | 0.20 (0.84,0.52) | 0.28 (0.78,0.60) |
| **HBLUP2** | **G** | 0.41 (0.80,0.82) | 0.42 (0.65,0.66) | 0.38 (0.73,0.66) | 0.42 (0.80,0.82) | 0.32 (0.62,0.70) | 0.22 (0.71,0.57) | 0.28 (0.64,0.60) |
| | **T** | 0.37 (0.84,0.74) | 0.37 (0.68,0.59) | 0.33 (0.75,0.58) | 0.42 (0.85,0.82) | 0.25 (0.69,0.55) | 0.21 (0.79,0.54) | 0.28 (0.72,0.60) |
| | **DIC** | 56733.1 | 8572.5 | 3849.4 | 32472.0 | 6305.7 | 5866.9 | 2795.6 |
| | **NG** | 0.35 (0.85,0.70) | 0.36 (0.69,0.57) | 0.31 (0.76,0.54) | 0.41 (0.88,0.80) | 0.20 (0.74,0.44) | 0.20 (0.84,0.52) | 0.29 (0.78,0.62) |
| **HBLUP3** | **G** | 0.41 (0.80,0.82) | 0.41 (0.65,0.65) | 0.38 (0.73,0.66) | 0.42 (0.81,0.82) | 0.32 (0.63,0.70) | 0.22 (0.72,0.57) | 0.28 (0.64,0.60) |
| | **T** | 0.36 (0.84,0.72) | 0.37 (0.68,0.59) | 0.33 (0.75,0.58) | 0.42 (0.85,0.82) | 0.25 (0.70,0.55) | 0.21 (0.80,0.54) | 0.28 (0.73,0.60) |
| | **DIC** | 56733.1 | 8573 | 3857.9 | 32473.7 | 6308.4 | 5866.7 | 2800.6 |

**Table 4**: Predictive ability and accuracy (in parenthesis: CV1 - left, CV2 - right) for non-genotyped individuals (NG), genotyped individuals (G) and across all individuals (T) for all phenotypic traits. Model fit was assessed through the Deviation Information Criterion (DIC) for the scenario using only ancestry informative markers (AIM). See comments in previous table

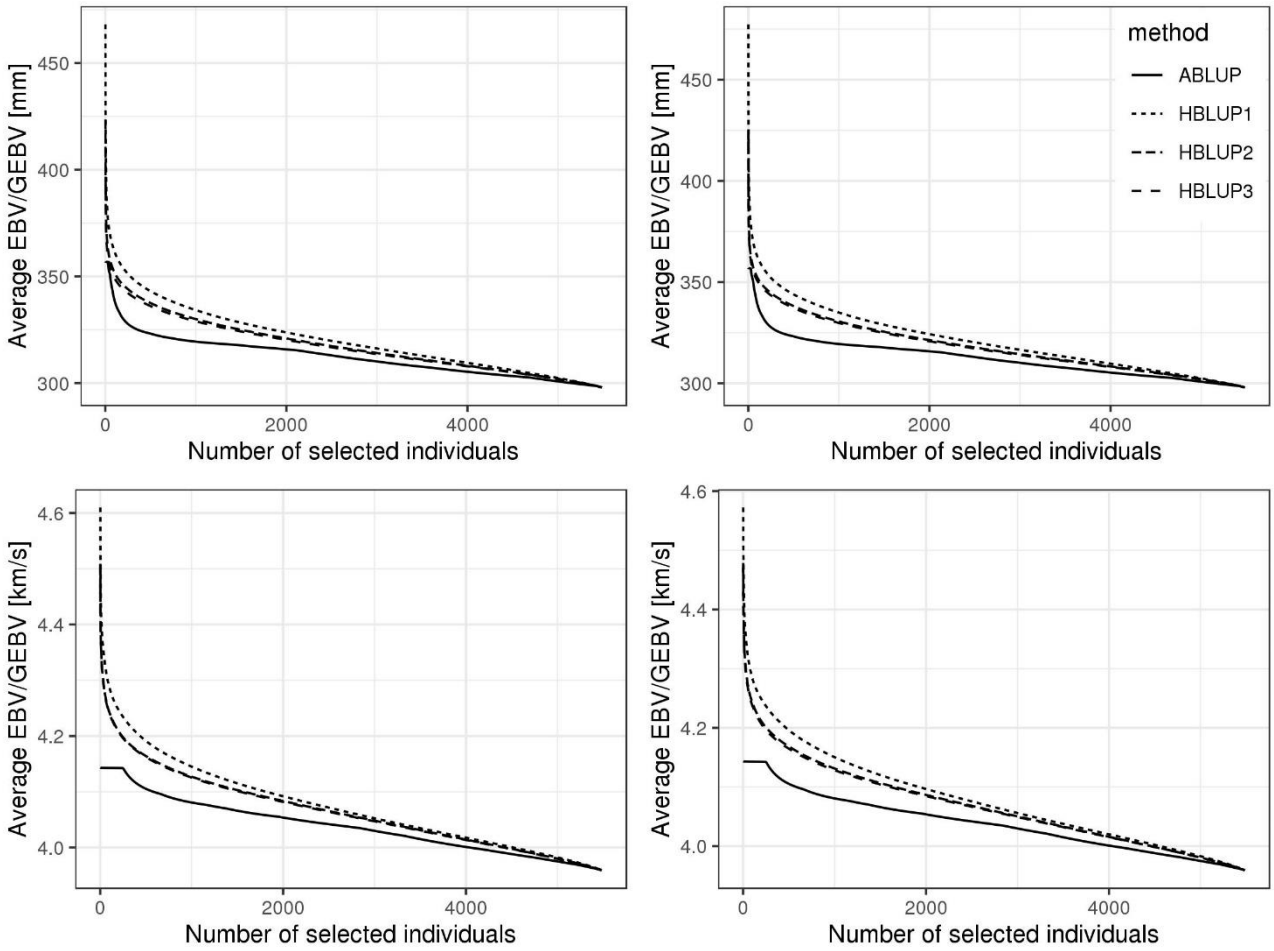


**Figure 2:** Genetic gain (i.e. average breeding values) achieved in each tested scenario for DBH using all markers (upper left) and ancestry informative markers only (upper right), as well as for acoustic velocity using all markers (bottom left) and ancestry informative markers only (bottom right) at the Gowan Hill site.

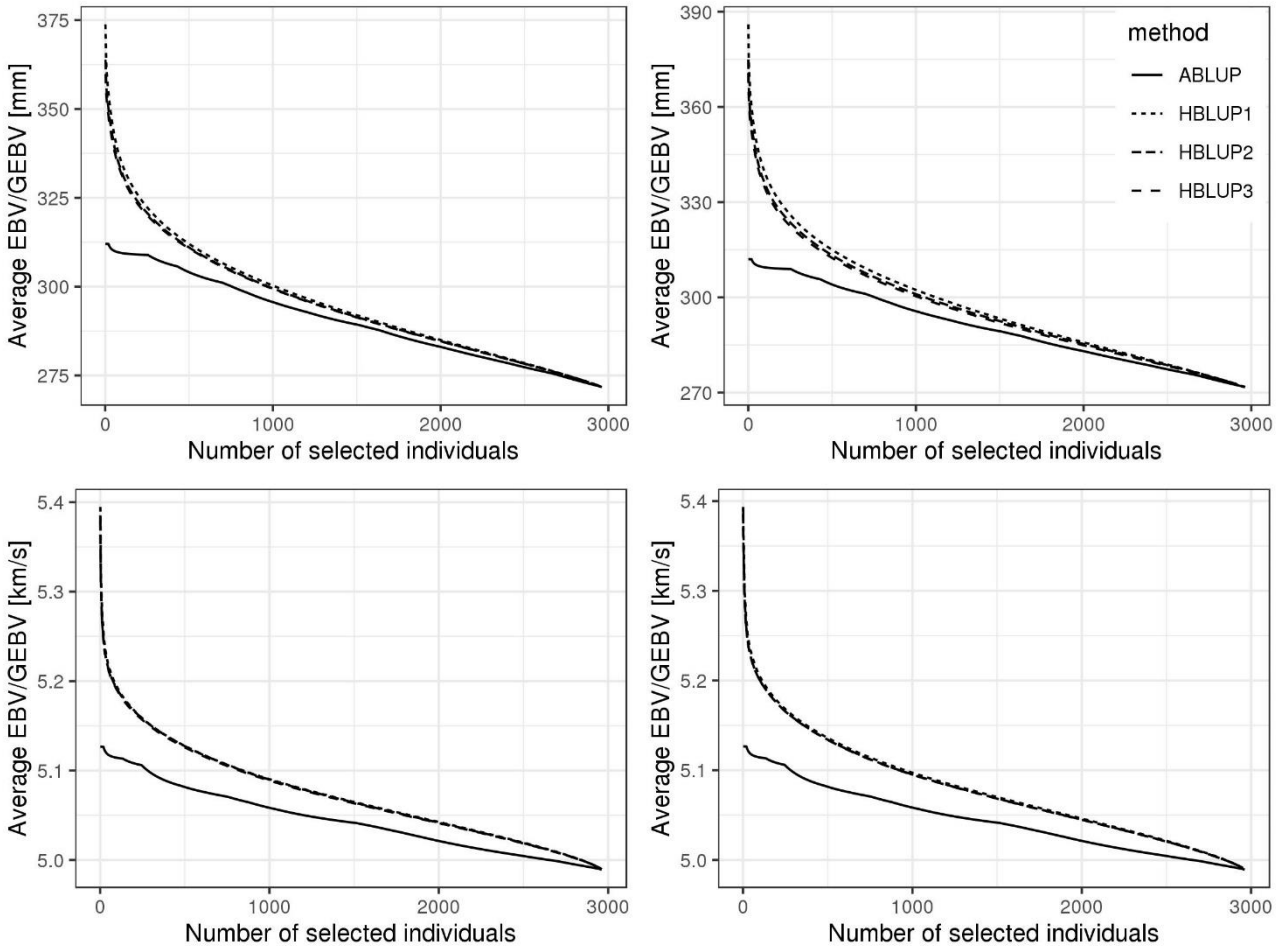


**Figure 3**: Genetic gain ( i.e. average breeding values) achieved in each tested scenario for DBH using all markers (upper left) and ancestry informative markers only (upper right), as well as for acoustic velocity using all markers (bottom left) and ancestry informative markers only (bottom right) at the Kaingaroa site.

# CONCLUSION

Implementation of genomic resources developed by Oregon State University in the evaluation of New Zealand Douglas-fir breeding populations resulted in substantially improved prediction accuracy and response to selection compared to pedigree-based analysis. An additional increase in the response to selection was found when only ancestry informative markers were used in the analysis of traits with strong population differentiation (e.g. DBH; Figures 2 and 3). Thus, the currently available SNP array appears to be a useful genotyping platform for the New Zealand Douglas-fir breeding program.

Implementation of the metafounders approach (i.e. inference of relatedness between pedigree founders) resulted in increased prediction accuracy not only for genotyped but also for non-genotyped individuals. This was not the case when relatedness between pedigree founders was ignored. However, more complex modelling of population demography resulted in worse model fit and lower prediction accuracy compared to a parsimonious model with a single metafounder population (Tables 3 and 4). Therefore, reliable modelling of population structure in forest trees is challenging, even with the availability of abundant genetic marker data.

# ACKNOWLEDGEMENTS

We would like to thank Mari Suontama for the initiation of the Douglas-fir genomic selection project, Dagmar Goeke for DNA extraction and handling, Natalie Graham for coordination of DNA sequencing and genotyping with the service provider and Mark Miller, Kane Fleet and Dylan Hicks for field data collection.

# REFERENCES

Aitken SN, Whitlock MC (2013). Assisted gene flow to facilitate local adaptation to climate change. *Annual review of ecology, evolution, and systematics* **44:** 367-388.

Borralho NM, Dutkowski GW (1998). Comparison of rolling front and discrete generation breeding strategies for trees. *Canadian Journal of Forest Research* **28**(7)**:** 987-993.

Butler D, Cullis BR, Gilmour A, Gogel B (2009). ASReml-R reference manual. *The State of Queensland, Department of Primary Industries and Fisheries, Brisbane*.

Carson MJ (1986). *Control-pollinated seed orchards of best general combiners: A new strategy for radiata pine improvement*. New Zealand Forest Service.

Dungey H, Low C, Lee J, Miller M, Fleet K, Yanchuk A (2012). Developing breeding and deployment options for Douglas-fir in New Zealand: breeding for future forest conditions. *Silvae Genetica* **61**(1-6)**:** 104-115.

Garcia-Baccino CA, Legarra A, Christensen OF, Misztal I, Pocrnic I, Vitezica ZG *et al* (2017). Metafounders are related to F st fixation indices and reduce bias in single-step genomic evaluations. *Genetics Selection Evolution* **49**(1)**:** 34.

Gianola D, Norton H (1981). Scaling threshold characters. *Genetics* **99**(2)**:** 357-364.

Gianola D, Schön C-C (2016). Cross-validation without doing cross-validation in genome-enabled prediction. *G3: Genes, Genomes, Genetics* **6**(10)**:** 3107-3128.

Howe GT, Yu J, Knaus B, Cronn R, Kolpak S, Dolan P *et al* (2013). A SNP resource for Douglas-fir: de novo transcriptome assembly and SNP detection and validation. *BMC genomics* **14**(1)**:** 137.

Jones OR, Wang J (2010). COLONY: a program for parentage and sibship inference from multilocus genotype data. *Molecular ecology resources* **10**(3)**:** 551-555.

Klapste J, Suontama M, Dungey H, Telfer E, Stovold G (2019). Modelling of population structure through contemporary groups in genetic evaluation. *BMC Genetics* **(under review)**.

Krakowski J, Stoehr MU (2009). Coastal Douglas-fir provenance variation: patterns and predictions for British Columbia seed transfer. *Annals of Forest Science* **66**(8)**:** 811.

Legarra A, Aguilar I, Misztal I (2009). A relationship matrix including full pedigree and genomic information. *Journal of dairy science* **92**(9)**:** 4656-4663.

Legarra A, Christensen OF, Vitezica ZG, Aguilar I, Misztal I (2015). Ancestral relationships using metafounders: finite ancestral populations and across population relationships. *Genetics* **200**(2)**:** 455-468.

Levin DA (1995). Metapopulations: an arena for local speciation. *Journal of Evolutionary Biology* **8**(5)**:** 635-644.

Magalska L, Howe GT (2014). Genetic and environmental control of Douglas-fir stem defects. *Forest ecology and management* **318:** 228-238.

Matyas C (1996). Climatic adaptation of trees: rediscovering provenance tests. *Euphytica* **92**(1-2)**:** 45-54.

Mátyás C (1994). Modeling climate change effects with provenance test data. *Tree physiology* **14**(7-8-9)**:** 797-804.

Misztal I, Legarra A, Aguilar I (2009). Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. *Journal of Dairy Science* **92**(9)**:** 4648-4655.

Mrode RA (2014). *Linear models for the prediction of animal breeding values*. Cabi.

Pérez P, de Los Campos G (2014). Genome-wide regression and prediction with the BGLR statistical package. *Genetics* **198**(2)**:** 483-495.

Rosenberg NA, Li LM, Ward R, Pritchard JK (2003). Informativeness of genetic markers for inference of ancestry. *The American Journal of Human Genetics* **73**(6)**:** 1402-1422.

Slavov GT, Howe GT, Adams WT (2005). Pollen contamination and mating patterns in a Douglas-fir seed orchard as measured by simple sequence repeat markers. *Canadian Journal of Forest Research* **35**(7)**:** 1592-1603.

Song J, Ratcliffe B, Kess T, Lai BS, Korecký J, El-Kassaby YA (2018). Temporal quantification of mating system parameters in a coastal Douglas-fir seed orchard under manipulated pollination environment. *Scientific reports* **8**(1)**:** 11593.

Team RC (2018). R Foundation for Statistical Computing; Vienna, Austria: 2014. *R: A language and environment for statistical computing***:** 2013.

Terrance ZY, Jayawickrama K (2014). Geographic Variation and Local Growth Superiority for Coastal Douglas-fir–Rotation-age Growth Performance in a Douglas-fir Provenance Test. *Silvae Genetica* **63**(1-6)**:** 116-124.

Ugarte E, Alenda R, Carabano M (1992). Fixed or random contemporary groups in genetic evaluations. *Journal of Dairy Science* **75**(1)**:** 269-278.

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

Wang T, Hamann A, Yanchuk A, O'neill G, Aitken S (2006). Use of response functions in selecting lodgepole pine populations for future climates. *Global Change Biology* **12**(12)**:** 2404-2416.

Wang T, O'Neill GA, Aitken SN (2010). Integrating environmental and genetic effects to predict responses of tree populations to climate. *Ecological applications* **20**(1)**:** 153-163.

Westell R, Quaas R, Van Vleck LD (1988). Genetic groups in an animal model. *Journal of Dairy Science* **71**(5)**:** 1310-1318.

Wright S (1922). Coefficients of inbreeding and relationship. *The American Naturalist* **56**(645)**:** 330-338.