Genomic selection proof-of-concept in two major conifer species

L. Bouffier[1], J. Klápště[2], J. Bartholomé[3], C. Plomion[1], E. Telfer[2], N. Graham[2], H. Dungey[2]

[1] BIOGECO, INRA, Univ. Bordeaux, 33610 Cestas, France
laurent.bouffier@inra.fr (Corresponding Author)
[2] Scion (New Zealand Forest Research Institute Ltd.), 49 Sala Street, Rotorua 3046, New Zealand
[3] AGAP, CIRAD, 34398 Montpellier, France

Keywords: genomic selection, forestry, prediction accuracy

Summary

Genomic selection has been successfully implemented in both animal and agricultural crop breeding programmes. Recently, forest tree breeding has also implemented genomic tools to improve the efficiency of breeding and selection procedures. Due to high genetic complexity of most economically important traits, genomic selection shows promise in operational breeding programmes for forest trees. Our two case studies, performed in economically important conifer species, showed that a genomic-based approach can reach similar prediction accuracies compared to the pedigree-based alternative. However, larger training population sample sizes should be used to increase the efficiency of genomic selection and outperform the traditional pedigree-based scenario. Moreover, broad genetic diversity is needed to successfully estimate genetic correlations and perform multivariate analyses.

Introduction

Conifers are long-lived organisms with rotation ages from 20 to 100 years depending on the species and economic context. Genomic selection offers new opportunities to accelerate genetic gain per unit time with the ability to rapidly integrate new traits in response to adaptation to climate change. Most conifer breeding programs were established in the 1950s or 1960s from base populations with a large effective size and a high level of genetic diversity (the base population consists of superior trees, called “plus-trees”, selected in unimproved plantations or from wild stands). These breeding programmes follow a recurrent selection scheme, focused mainly on general combining ability. Generally, only 2-3 selection cycles have been completed to date due to low investment and logistical complexity (late sexual maturity and large genetic evaluation trials). Although molecular markers have not yet been implemented in operational tree breeding, several projects have explored the potential for genomic selection in conifers. The main obstacle to the implementation of genomic selection in conifers is the limited number of genomic tools (low number of markers per cM due to the large genome, and no reference genome for most conifers). Nevertheless, successive breeding populations are available in clonal archives which should facilitate the calibration of genomic models.

In this paper, we present genomic prediction proof-of-concept for two major conifer species: maritime pine and radiata pine. Maritime pine (*Pinus pinaster*) is the most important plantation species in France (44 millions seedlings in 2016) and radiata pine (*Pinus radiata*) constitutes 90% of the planted forest estate in New Zealand. In this paper we present case

studies for both species, followed by general recommendations based on our learnings, for genomic selection in conifers.

Maritime pine genomic prediction proof-of-concept

The maritime pine breeding program follows a recurrent selection scheme from a base population of 635 "plus-trees" (called G0 trees) selected during the 1960s. The main population has cycled through three generations (G0, G1 and G2). In order to reduce selection cycle length, to increase the selection intensity and implement new selection traits, genomic selection (GS) is under study for future implementation in the maritime pine breeding program (Bartholomé et al., 2016).

Given the low level of linkage disequilibrium and the low number of markers available (<5,000) to cover the large genome of maritime pine (24 Gb), we have set up a proof-of-concept experiment with a limited effective size population (N_S = 25) over the three generations of the breeding program (G0, G1 and G2). The reference population (n=818) was established with 710 G2 trees and all their progenitors (ie. 62 G1 and 46 G0). Thus each G2 of the reference population has its 2 parents and its 4 grand-parents represented in this population. This reference population was successfully genotyped with 4,332 polymorphic SNPs (MAF>0.01). The pseudo-phenotypes (EBV) considered for genomic selection analyses consisted of breeding values (height, diameter and stem straightness) evaluated from a BLUP analysis of the breeding population from more than 500,000 trees.

Two validation methods were considered to evaluate the effect of calibration and validation sets on genomic prediction accuracy: the "subset validation method" and the "progeny validation method" (Figure1). In the first method, the G2 population was split into calibration and validation sets with three sampling strategies (high or low level of relatedness). In the second method, the progenitors (G0 and G1 trees) were used for the calibration set and the G2 trees as the validation set.

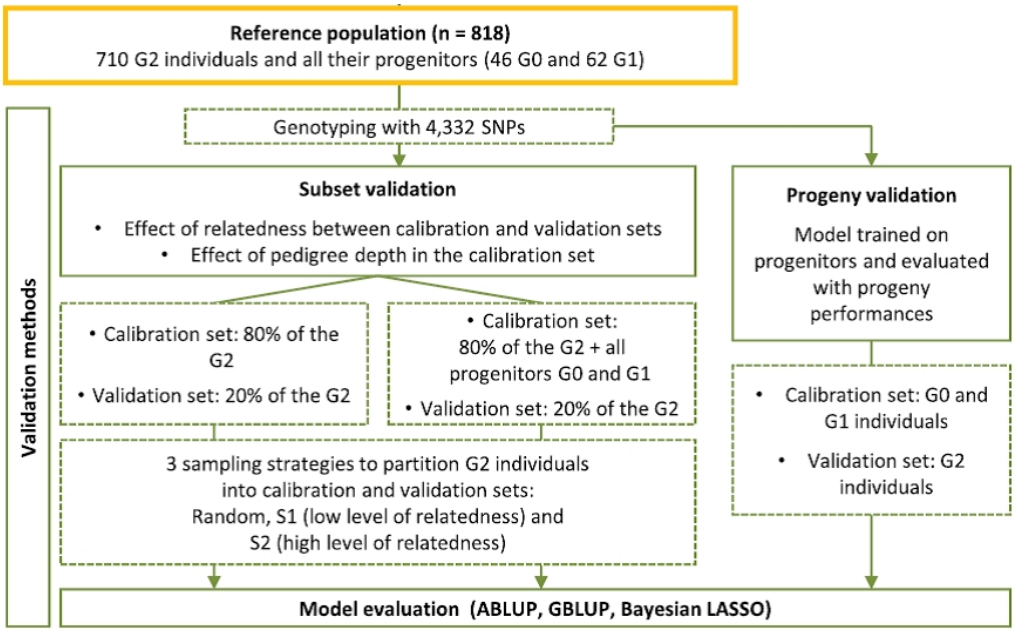

Figure 1: Validation methods considered to evaluate the performance of prediction models.

Three prediction models were considered to estimate breeding values: ABLUP (based

only on pedigree information), GBLUP (all markers are assumed to have the same effects) and Bayesian LASSO (marker effects distributed following a Gaussian likelihood method). Table 1 ("subset validation method") and Figure 2 ("progeny validation method") show results for prediction accuracy which is the coefficient of correlation between the EBV and the GEBV obtained from the three models (ABLUP, GBLUP, and Bayesian LASSO).

Table 1: Prediction accuracy results for the "subset validation method".

		Calibration set 80% of the G2			Calibration set 80% of the G2 + G0/G1		
		ABLUP	**GBLUP**	**B-LASSO**	**ABLUP**	**GBLUP**	**B-LASSO**
Circum - ference	**Rand**	0.78 (0.68-0.85)	0.73 (0.62-0.80)	0.72 (0.62-0.80)	0.83 (0.79-0.89)	0.74 (0.67-0.81)	0.74 (0.67-0.81)
	S1	0.55 (0.34-0.74)	0.52 (0.24-0.67)	0.52 (0.24-0.67)	0.81 (0.65-0.89)	0.69 (0.51-0.81)	0.69 (0.51-0.81)
	S2	0.80 (0.73-0.85)	0.74 (0.67-0.81)	0.74 (0.67-0.80)	0.84 (0.80-0.89)	0.75 (0.68-0.84)	0.75 (0.68-0.82)
Height	**Rand**	0.68 (0.54-0.78)	0.66 (0.56-0.77)	0.66 (0.56-0.77)	0.75 (0.66-0.82)	0.68 (0.60-0.76)	0.68 (0.59-0.75)
	S1	0.58 (0.46-0.77)	0.58 (0.43-0.75)	0.58 (0.38-0.74)	0.74 (0.63-0.87)	0.67 (0.54-0.79)	0.66 (0.53-0.79)
	S2	0.70 (0.60-0.77)	0.69 (0.60-0.76)	0.68 (0.59-0.76)	0.75 (0.66-0.83)	0.70 (0.59-0.79)	0.69 (0.59-0.79)
Stem straight - ness	**Rand**	0.86 (0.80-0.90)	0.81 (0.75-0.86)	0.82 (0.76-0.86)	0.90 (0.86-0.94)	0.82 (0.74-0.88)	0.82 (0.75-0.88)
	S1	0.67 (0.51-0.79)	0.65 (0.48-0.77)	0.66 (0.48-0.77)	0.88 (0.78-0.93)	0.77 (0.62-0.87)	0.77 (0.63-0.87)
	S2	0.87 (0.84-0.91)	0.81 (0.77-0.87)	0.81 (0.77-0.88)	0.91 (0.88-0.94)	0.80 (0.76-0.85)	0.80 (0.76-0.86)

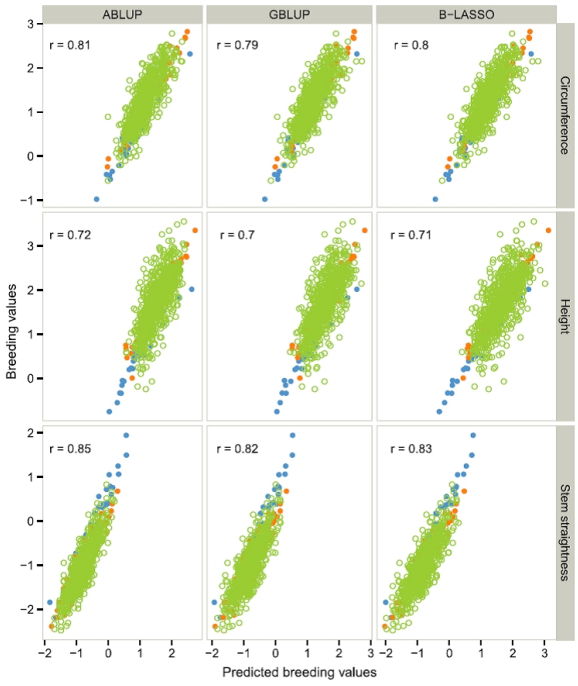

Figure 2: Prediction accuracy results for the "progeny validation method". Closed circles represent the calibration set (46 G0 genotypes in blue and 62 G1 genotypes in orange). The validation set is represented in open green circles (710 G2 genotypes).

The subset validation method highlights the importance of relatedness between calibration and validation sets. The progeny validation method shows that higher prediction accuracies are obtained over multiple generations which is a key issue when applying genomic selection. The pedigree-based model, however, had prediction accuracies similar or greater than that of marker-based models. This means that our genomic models are not better than a basic pedigree recovery analysis, and thus, are not able to predict within-family variability. This can be explained by several factors such as the low number of markers, the low number of trees per family in the calibration set or the poor accuracy of pseudo-BLUP for G2 trees.

Radiata pine genomic prediction proof-of-concept

The current radiata pine breeding program was established by the selection of a large number of "plus-trees" from forest stands planted across New Zealand. A subset of these (588) were selected in 1968, which were subsequently tested through open-pollinated progeny tests and forward selection to create a new generation. Forward selection from these breeding populations, along with the original plus trees, represent the parental population for the genomic selection training population. The training population consists of 988 individuals, representing 85 families. It includes two sub-populations POP2 and POP3 (Li et al., 2016) selected from the same gene pool (N_S=32 for POP2 and N_S=35 for POP3). The POP2 sub-population was created from two selection lines: 1) individuals selected for growth and form through GF score combining breeding values from diameter at breast height (DBH) and acceptability considering straightness, malformation and branching pattern (POP2GF), and 2) individuals selected through tandem selection, weakly for growth and form and strongly for high wood density, with the aim of breaking the population-level negative genetic correlation between growth and wood density (POP2HD). The sub-population POP3 was selected only for growth and form. The training population was phenotyped on 4 ramets per genotype (a ramet is vegetative copy of a genotype) for diameter at breast height (DBH), wood density (WD), straightness (STR9) and branching (BR9). Genomic data were generated through an exome-capture genotype-by-sequencing (GBS) genotyping platform (Neves et al 2012) producing 80,160 single nucleotide polymorphisms (SNPs). An Eigen-decomposition of the marker-based relationship matrix found structure that reflects the selection history of each sub-population (Figure 3). The pedigree-based (ABLUP) and marker-based (GBLUP) models were performed separately for each population (using only genotyped individuals) at a single-trait or multi-trait level using the MTG2 package (Lee & van der Werf, 2016) and breeding value reliabilities were estimated Additionally, the Krzanowski test (Krzanowski 1979) was performed to investigate correlations between correlation matrices estimated in each population and found strong correlations of 0.88 and 0.95 between populations passing the same selection history (POP2GF and POP3) while low correlations from 0.13 to 0.24 were

reached between population following different selection regimes.

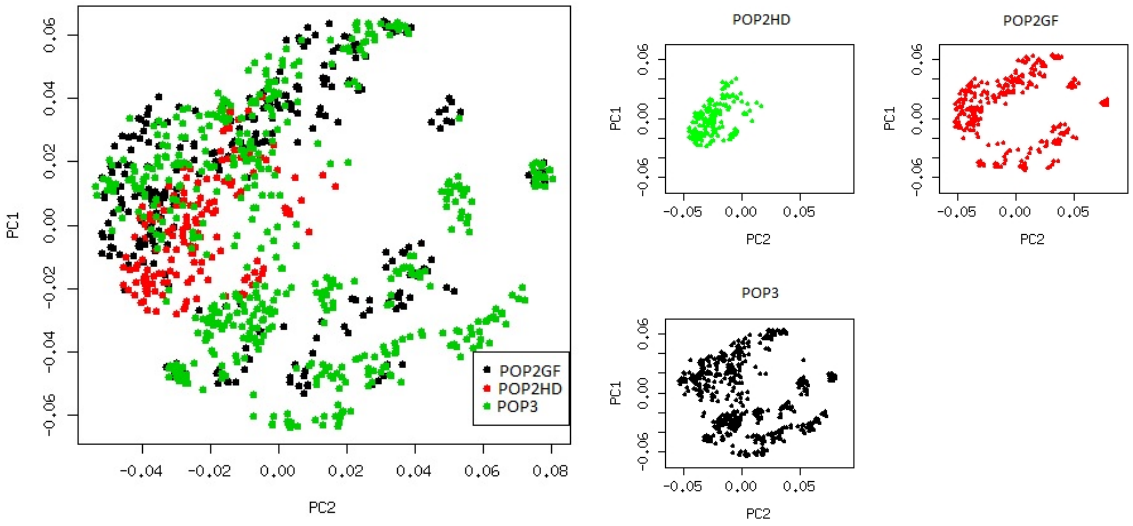

Figure 3: Eigen-decomposition of the marker-based relationship matrix

Table 2: Reliability of breeding values obtained from the single trait model

Site	Population	Pedigree-based model				Marker-based model			
		BR9	DBH	ST9	WD	BR9	DBH	ST9	WD
Tarawera	POP2GF	0.52	0.50	0.40	0.63	0.51	0.48	0.37	0.60
	POP2HD	0.63	0.58	0.40	0.62	0.57	0.54	0.35	0.61
	POP3	0.60	0.54	0.52	NA	0.63	0.55	0.54	NA
Woodhill	POP2GF	0.53	0.51	0.35	0.11	0.52	0.51	0.31	0.08
	POP2HD	0.62	0.62	0.52	0.05	0.58	0.58	0.50	0.03
	POP3	0.47	0.38	0.42	0.71	0.49	0.35	0.43	0.75
Kinleith	POP3	0.49	0.37	0.37	0.38	0.53	0.38	0.34	0.42

Currently, several studies have explored the benefit of utilizing multivariate over univariate analysis (Jia & Jannink, 2012; Marchal et al., 2016) to improve the accuracy of genomic predictions, in traits with low heritability, through genetic correlations. We

investigated the likely benefit of multivariate analysis in the context of radiata pine populations under different selection histories. The multivariate analysis uncovered that the tandem selection strategy broke most of the commonly expected correlations within the population. The expected improvement in the accuracy of genomic breeding values in the multivariate analysis was surprisingly realized only in the POP2HD populations for BR9, ST9 and WD (Table 2 and 3). This could be due to the small sample size used to estimate genetic correlations as the requirements for the reliable estimation of genetic correlations are more demanding compared with heritability estimates (White et al., 2007). As reported in Bijma & Bastiaansen, (2014), at least 100 families should be employed for the estimation of genetic correlations with reasonable precision. Generally, the multivariate analysis was found to not produce any benefit and instead gave rather inferior results among the uncorrelated traits (Jia & Jannink, 2012) but this was not the case in the present study.

Table 3: Reliability of breeding values obtained from the multi-trait model

Site	Population	Pedigree-based model				Marker-based model			
		BR9	DBH	ST9	WD	BR9	DBH	ST9	WD
Tarawera	POP2GF	0.50	0.40	0.39	0.56	0.48	0.38	0.36	0.54
	POP2HD	0.67	0.56	0.49	0.68	0.58	0.50	0.42	0.67
	POP3	0.57	0.51	0.50	NA	0.60	0.51	0.52	NA
Woodhill	POP2GF	0.52	0.48	0.35	NA	0.52	0.43	0.34	NA
	POP2HD	0.62	0.62	0.57	NA	0.62	0.56	0.58	NA
	POP3	0.44	0.31	0.34	0.68	0.43	0.28	0.33	0.71
Kinleith	POP3	0.42	0.32	0.39	0.40	0.45	0.31	0.36	0.47

Recommendations

In these two conifer case studies, genomic-based approaches reach similar prediction accuracies compared with the pedigree based alternatives. This means that, at this stage, Mendelian sampling is not correctly predicted. Several hypotheses can explain these results:

- a low number of markers in comparison to the large size of conifer genome (more than 20 Gb)
- the selection of the SNP set (mainly SNP from EST with low MAF)
- the reliability of the pseudo-phenotype considered as the reference to estimate the genomic selection accuracy
- the design of the training population, generally constituted with a low number of tree per full-sib family

The last hypothesis is probably a key point to outperform the traditional pedigree-based scenarios. The decreasing cost of genotyping should allow to genotype larger training population sample sizes with a focus on increasing the number of genotyped trees per family. In conifer breeding, crossing can generate large numbers of offspring. Capturing Mendelian sampling would allow to select superior trees without testing and thus deliver genetic gain improvements through acceleration of the breeding cycle.

List of References

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

BMC Genomics 17: e604.
Bijma, P. & Bastiaansen, J.W., 2014. Standard errors of genetic correlation: how much data do we need to estimate a purebred-crossbred genetic correlation? Gen Sel Evol 46(1): 1-6.
Jia, Y. & Jannink, J.-L., 2012. Multiple-trait genomic selection methods increases genetic value prediction accuracy. Genetics 192(4): 1513-1522.
Krzanowski W., 1979. Between-groups comparison of principal components. Journal of American Statistical Association 74(367): 703-707.
Lee, S. H., & van der Werf, J., 2016. MTG2: An efficient algorithm for multivariate linear mixed model analysis based on genomic information. Bioinformatics, 32(9): 1420-1422.
Li, Y., Wilcox, P., Telfer, E., Graham, N., Stanbra, L., 2016. Association of single nucleotide polymorphisms with form traits in three New Zealand populations of radiata pine in the presence of genotype by environment interactions. Tree Genetics & Genomes 12:63.
Marchal, A., Legarra, A., Tisne, S., Carasco-Lacombe, C., Manez, A., Suryana, E., Omore, A., Nouy, B., Durand-Gasselin, T., Sanchez, L., Bouvet, J.-M.,Cros, D., (2016). Multivariate genomic model improves analysis of oil palm (*Elaeis guineensis* Jacq.) progeny test. Mol Breeding 36 (1): e2.
White, T. L., Adams, W. T., & Neale, D. B., 2007. Forest genetics. Cabi, pp:682.