



**HAL**  
open science

## Performance of genomic prediction within and across generations in maritime pine

Jérôme Bartholomé, Joost van Heerwaarden, Fikret Isik, Christophe C. Boury, Marjorie Vidal, Christophe Plomion, Laurent Bouffier

► **To cite this version:**

Jérôme Bartholomé, Joost van Heerwaarden, Fikret Isik, Christophe C. Boury, Marjorie Vidal, et al.. Performance of genomic prediction within and across generations in maritime pine. *BMC Genomics*, 2016, 17 (1), 14 p. 10.1186/s12864-016-2879-8 . hal-02636107

**HAL Id: hal-02636107**

**<https://hal.inrae.fr/hal-02636107>**

Submitted on 27 May 2020

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.


L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

RESEARCH ARTICLE

Open Access



# Performance of genomic prediction within and across generations in maritime pine

Jérôme Bartholomé<sup>1</sup> , Joost Van Heerwaarden<sup>2</sup>, Fikret Isik<sup>3</sup>, Christophe Boury<sup>1</sup>, Marjorie Vidal<sup>1,4</sup>, Christophe Plomion<sup>1</sup> and Laurent Bouffier<sup>1\*</sup>

## Abstract

**Background:** Genomic selection (GS) is a promising approach for decreasing breeding cycle length in forest trees. Assessment of progeny performance and of the prediction accuracy of GS models over generations is therefore a key issue.

**Results:** A reference population of maritime pine (*Pinus pinaster*) with an estimated effective inbreeding population size (status number) of 25 was first selected with simulated data. This reference population ( $n = 818$ ) covered three generations (G0, G1 and G2) and was genotyped with 4436 single-nucleotide polymorphism (SNP) markers. We evaluated the effects on prediction accuracy of both the relatedness between the calibration and validation sets and validation on the basis of progeny performance. Pedigree-based (best linear unbiased prediction, ABLUP) and marker-based (genomic BLUP and Bayesian LASSO) models were used to predict breeding values for three different traits: circumference, height and stem straightness. On average, the ABLUP model outperformed genomic prediction models, with a maximum difference in prediction accuracies of 0.12, depending on the trait and the validation method. A mean difference in prediction accuracy of 0.17 was found between validation methods differing in terms of relatedness. Including the progenitors in the calibration set reduced this difference in prediction accuracy to 0.03. When only genotypes from the G0 and G1 generations were used in the calibration set and genotypes from G2 were used in the validation set (progeny validation), prediction accuracies ranged from 0.70 to 0.85.

**Conclusions:** This study suggests that the training of prediction models on parental populations can predict the genetic merit of the progeny with high accuracy: an encouraging result for the implementation of GS in the maritime pine breeding program.

**Keywords:** Genomic selection, Growth, Multiple generations, *Pinus pinaster*, Progeny validation, Relatedness, Stem straightness

## Background

The use of genome-wide DNA markers to predict genomic estimated breeding values (GEBV), first proposed by Meuwissen et al. [1], has radically changed perspectives in molecular breeding. Breeders now have access to large numbers of single-nucleotide polymorphisms (SNPs). They have therefore focused their efforts on genomic selection (GS), which is based on a large set of markers expected to be in linkage disequilibrium (LD) with every QTL controlling the phenotype of interest. In

comparison to classical marker-assisted selection, which uses a small set of well-characterized markers tracing a small number of quantitative trait loci (QTLs), each with a medium-to-large effect, GS offers the possibility of a higher genetic gain per unit of time [2–4]. Thus, with the availability of cost-effective genotyping platforms [5], the use of this approach has become widespread in the breeding of animals [4, 6] and plants [7, 8], including forest trees [9, 10]. GS requires the development of a predictive model with a calibration population for which both genotype and phenotype have been characterized. This model is then used to predict GEBV, from marker genotypes alone, in the targeted breeding populations. As in traditional selection based on estimated breeding

\* Correspondence: bouffier@pierreton.inra.fr

<sup>1</sup>BIOGECO, INRA, Univ. Bordeaux, 33610 Cestas, France

Full list of author information is available at the end of the article



values (EBV), prediction accuracy is a key issue in evaluations of the efficiency of GS strategies. The prediction accuracy of GS models is evaluated by assessing the correlation between the GEBV obtained with GS models and the EBV obtained by classical genetic evaluation based on progeny testing. Simulation studies, either general [1, 11–15], or species-based (maize [3], oil palm [16], barley [17], Japanese cedar [18]), have attempted to identify key factors affecting the accuracy of GEBV. In a review on dairy cattle, Hayes et al. [4] highlighted four major factors: i) the heritability of the target trait, ii) the genetic architecture of the trait (number and effect of underlying QTLs), iii) the level of LD between markers and QTLs in the reference and target populations, and iv) the size of the reference population, and the degree to which the reference and target populations are related. The statistical methods used to predict GEBV may also affect the accuracy of this prediction [19], but to a lesser extent.

In forest tree breeding, the duration of a single cycle of selection-recombination is driven by the time at which flowering first occurs (e.g. 7–8 years in maritime pine) and the age at which early indirect selection for mature properties can be carried out (e.g. 10–12 years for total height and stem straightness in maritime pine). A full cycle therefore generally lasts more than two decades. In addition, the low-to-medium heritabilities of most complex traits, such as growth, stem form, and branching

characteristics, limit the response to selection, and, thus, the expected genetic gain. GS may overcome these limitations, by decreasing breeding cycle duration and improving selection efficiency/intensity for traits with a low heritability, thereby increasing the efficiency of breeding strategies. Preliminary studies on major plantation forest trees (eucalyptus, spruces and pines) have given encouraging results [9, 10], with accuracies of up to 0.8 (Table 1), despite the low level of LD in these outcrossing species, which have large population sizes [20, 21], and low marker coverage (i.e. a few thousand loci). These studies showed that GS with DNA markers provided accuracies similar to those obtained for classical genetic evaluation with progeny testing (Table 1). Rather than capturing historical LD associations between markers and QTLs, this approach derives its prediction accuracy from better estimations of realized genomic relationships [22, 23]. The relatively small effective population sizes of the reference populations and validation within the same population clearly contributed to higher accuracies. Indeed, lower accuracies (around 0.5) were obtained for larger reference populations [24, 25] or when GS models were applied to target populations different from the reference population [26]. It is important to assess the prediction accuracy of GS models across generations, because recombination may modify marker-allele phases in subsequent generations, and because selection may change allele frequencies [10].

**Table 1** List of genomic selection studies based on real data sets conducted on forest tree species. Studies are listed in chronological order of publication. This study is the last one listed

Species	Population			Genotyping			Traits analyzed	Models	Prediction accuracy	Reference
	Size	Family type	Family size	G	Method	Number of markers				
<i>Eucalyptus</i> hybrids	738	43 FS	15 to 23	1	DArT array	3129	Growth, wood properties	RR-BLUP	0.54–0.6	[26]
	920	51 FS	10 to 15	1	DArT array	3564			0.38–0.55	
Loblolly pine	790–840	61 FS	-	1	SNP array	4852	Growth	RR-BLUP	0.63–0.75	[51]
Loblolly pine	951	61 FS	15 ± 2.2	1	SNP array	4825	Growth, tree architecture, wood properties, disease resistance	RR-BLUP, Bayes A, Bayes Cπ, B-LASSO, RR-BLUP B	0.17–0.51	[52]
Loblolly pine	149	13 FS	1 to 34	1	SNP array	3406	Growth, wood properties	RR-BLUP	0.30–0.83	[75]
Loblolly pine	165	9 FS	3 to 37	1	SNP array	3461	Growth	ABLUP, GBLUP	0.37–0.74	[71]
White spruce	1694	214 HS	-	1	SNP array	6385	Growth, wood properties	ABLUP, B-RR, B-LASSO	0–0.44	[24]
White spruce	1748	59 FS	25 to 33	1	SNP array	6932	Growth, wood properties	ABLUP, B-RR, Combined	0.33–0.45	[60]
Loblolly pine	956	61 FS	15 ± 2.2	1	SNP array	4825	Growth, tree architecture	ABLUP, RR-BLUP	0.17–0.51	[57]
Maritime pine	661	191 HS	1 to 13	2	SNP array	2500	Growth, stem straightness	GBLUP, B-RR, B-LASSO	0.09–0.73	[25]
Interior spruce	1126	25 HS	<32	1	GBS	8868–62,198	Growth, wood properties	RR-BLUP, GRR	0.34–0.77	[61]
Interior spruce	769	25 HS	-	1	GBS	34,570–50,803	Growth	RR-BLUP, GRR, Bayes Cπ	0.04–0.55	[76]
Maritime pine	817	35 HS	13 to 34	3	SNP array	4332	Growth, stem straightness	ABLUP, GBLUP, B-LASSO	0.24–0.94	This study

FS full-sib family, HS half-sib family, G number of generations included in the study, GBS genotyping-by-sequencing method

These effects may decrease GS accuracy over generations [11, 27]. The validation of GS models across generations, with assessment of the predictive ability of markers, is essential before the implementation of GS strategies in tree breeding. The marker-trait associations established in “parental” populations (the parents or preceding generations) should be validated in progeny populations (i.e., progeny validation) [28, 29]. To our knowledge, no study on forest tree species has yet used empirical data to address this issue. Indeed, in all the studies listed in Table 1, individuals of the same generation were split into calibration and validation sets for the evaluation of GS models.

Maritime pine (*Pinus pinaster*) is a major forest tree species in south-western Europe. A breeding program based on a recurrent selection strategy was initiated in France in the 1960s [30]. A base population of 635 founders (the G0 trees) was selected from the “Landes” ecotype (an ecotype found in South-West France) for growth (height and circumference) and stem straightness. This population was subjected to two cycles of breeding, testing and selection (i.e. the G1 and G2 generations). The potential of GS for use in maritime pine breeding is currently being evaluated alongside the implementation of a forward selection strategy with pedigree reconstruction [31]. A preliminary investigation based on a population of 661 individuals from the first two generations, with low marker coverage (2500 SNPs, i.e. ~1.39 markers/cM), showed the prediction accuracy of GS models to be about 0.50 for growth and stem straightness [25]. In this study, we first selected a reference population on the basis of the following criteria: i) high performance for the main traits of the breeding program, ii) limited effective population size, and iii) combining the three generations of the maritime pine breeding population. Simulations were carried out to optimize the set of individuals to be genotyped for genomic prediction. Finally, using the reference population with real phenotypic and genotypic data, we aimed: i) to compare the predictive power of SNP markers with that of the pedigree-based method, ii) to investigate the effect on prediction accuracy of pedigree depth and relatedness between the calibration and validation sets, and iii) to investigate the impact of the use of third-generation individuals as a validation set (progeny validation) on the prediction accuracy of GS models.

## Methods

### Design of the reference population

The reference population was designed in two steps, as summarized in Fig. 1. A pre-selection step based on pedigree and phenotype information was first applied to G2 individuals and their progenitors. Simulations were then used to select a subset of about 800 individuals (*a priori* on the basis of genotyping constraints) to maximize the expected genomic prediction accuracy.

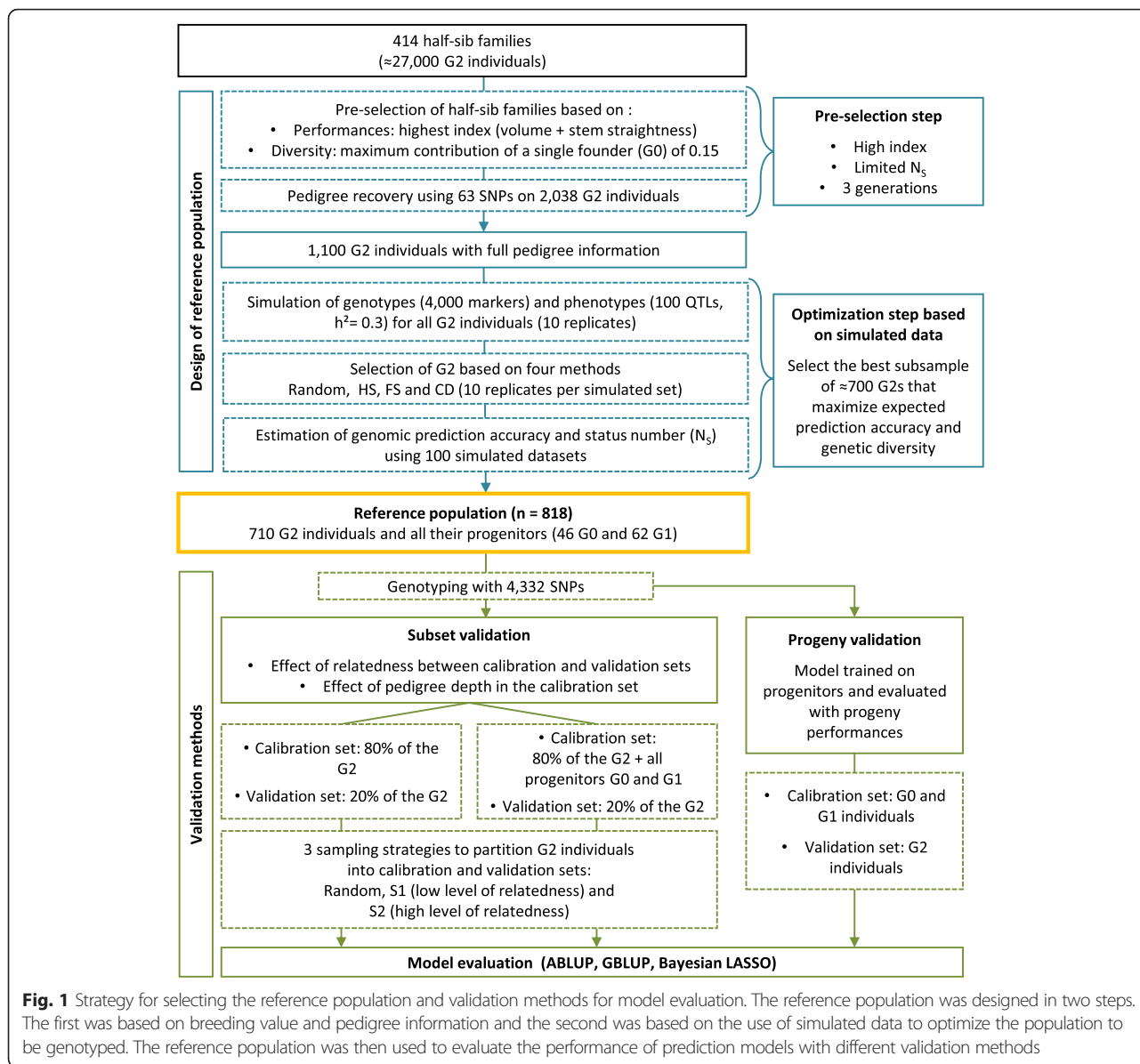
### Pre-selection of G2 individuals

G2 individuals were pre-selected in series of polycross trials involving 414 half-sib families (identified mothers were crossed with a pollen mixture) and 27,265 G2 trees. Breeding values based exclusively on maternal pedigree (as the paternal pedigree was unknown) were estimated for height, stem circumference at breast height and stem straightness, in a mixed model framework. Two criteria were used to select a subset of G2 trees: i) an index combining the best linear unbiased predictions (BLUP) for volume and stem straightness (equal weighting) to select the best half-sib family, ii) a maximum of 40 half-sib families with a maximum contribution of a single founder (G0) of 0.15, to prevent the over-representation of a few founders and to give a limited status number ( $N_S$ , an estimate of effective population size). This procedure resulted in the selection of 2038 G2 trees. Pedigree recovery with 63 SNP markers was carried out on these trees to identify the paternal parent and to check the maternal genotype (see Vidal et al. [31] for a description of the methodology). Maternal identity was confirmed and paternal parents (pollen donors) were identified for 1308 G2 individuals. At least one of the grandparents (G0 individuals) was unknown for 208 of the 1308 G2 individuals. We decided to select only G2 trees for which full pedigree information was available. Thus, 1100 G2 trees and their progenitors (78 G1 and 50 G0) were available for the design of the reference population on the basis of simulation data.

### Simulation to optimize the final selection of the reference population

We used 4000 markers evenly distributed over a 1665 cM composite genetic map of maritime pine [25], including 2965 mapped positions. A gene-dropping algorithm developed in R [32] was used to generate the genotypes of the G1 and G2 offspring. Starting with a set of identified founder haplotypes in generation G0, this algorithm modeled the process of segregation and gamete association over the three generations resulting in known founder alleles at each marker position for each individual in the G2 population. The probability of recombination between adjacent markers was set according to the genetic distance between them. Marker states were assigned randomly to each founder allele, assuming an allele frequency of 0.5 for all markers. The trait of interest was modeled by assigning a non-zero QTL effect, assuming a normal distribution, to 100 random marker positions and setting the environmental error term to give a narrow-sense heritability of 0.3, corresponding to the observed heritability of the target traits [31].

Four methods were applied to the 1100 G2 plants, to establish a reference population of about 800 trees (G0, G1 and G2). In the first method, G2 trees were selected at random (the random method). The second method



was based on sampling within the largest maternal half-sib families, with equal numbers of individuals selected from each half-sib family (the HS method). In the third method, G2 trees were sampled from the largest full-sib families, with a maximum of two individuals selected per family (the FS method). For the fourth method, we maximized the mean generalized coefficient of determination (CD method) [33, 34]. The CD method provides a measurement of the expected reliability of predictions based on the pedigree. Briefly, a specified number (eight in our case) of individuals with the highest CD values are removed one-by-one, with the individuals causing the largest decrease in mean CD being retained. This process is repeated until the desired number of individuals remain. We evaluated these four methods by

simulating 100 replicates corresponding to 10 different datasets (simulated genotypes and phenotypes), each with 10 different samplings of the G2 generation. Status number ( $N_S$ , [35]) was estimated as  $N_S = \frac{1}{2}F$ , where  $F$  is the mean inbreeding value calculated from the realized kinship matrix; see the methods below.

**Phenotypic and genotypic data for the reference population**  
**Traits analyzed**

The estimated breeding values (EBV) for three different traits — circumference and height at 12 years of age and stem straightness at 8 years of age — were obtained from a meta-analysis based on the TREEPLAN framework [36]. The correlations between circumference

and height (Spearman's correlation coefficient  $\rho = 0.61$ ,  $p < 0.01$ ) and between circumference and stem straightness ( $\rho = 0.45$ ,  $p < 0.01$ ) were moderate. A weaker correlation was observed between height and stem straightness ( $\rho = 0.36$ ,  $p < 0.01$ , Additional file 1: Figure S1). EBV reliability was generally high ( $0.97 \pm 0.02$ ) for G0 and G1 individuals, and mean EBV reliability for the G2 population was 0.75. Parental effects on the EBV of individuals can be large and may introduce bias into genomic estimated breeding values. The BLUP method shrinks the breeding values towards the mean and reduces the variation. We addressed the issues of bias and reduced heterogeneity by deregressing the EBV of individuals, as suggested by Garrick et al. [37]. We used the heritabilities estimated from TREEPLAN evaluation for deregression: 0.17, 0.32 and 0.26 for circumference, height and stem straightness, respectively. The resulting deregressed breeding values were used as pseudo-phenotypes for the genomic prediction analysis.

#### Genotyping and linkage disequilibrium analysis

The DNA extraction method and the Illumina Infinium array used to genotype the reference population have been described elsewhere [38]. SNP clustering was performed with GenomeStudio (Genotyping module V1.9, Illumina, San Diego, USA), with the manual checking of each SNP. One G2 individual, with a call rate below 0.98 and a 10 % GenCall score below 0.24, was removed. We analyzed 8411 SNP loci: genotyping failed for 2429 (low fluorescence intensity, GenTrain score below 0.35), 1539 were monomorphic and 4443 were polymorphic (52.8 %). The pattern of SNP inheritance was checked with MERLIN [39]. SNPs presenting an aberrant inheritance pattern or for which more than 2 % of values were missing were removed from subsequent analyses. For the remaining 4436 polymorphic SNPs, the mean GenTrain score was 77.7 %, the mean percentage of missing data was 0.05 % and the repeatability, based on eight duplicated genotypes, was greater than 99.9 %. For genomic prediction models, 4332 SNPs were retained on the basis of their minor allele frequency (MAF > 0.01). Genetic location on the *P. pinaster* composite map [40] was determined for 3962 SNPs (91.5 %, Additional file 1: Figure S2), corresponding to a total of 2548 contigs of the *P. pinaster* unigene [41]. The number of markers per linkage group ranged from 279 to 376, with a mean of 330, corresponding to 2.4 SNPs per cM.

The intra-chromosomal LD between markers was calculated as  $r^2$  with R software and expressed as a function of the genetic distance between markers. The effect of selection (differentiation between generations), resulting in changes in allele frequencies between generations,

was assessed by calculating a fixation index ( $F_{ST}$ ) [42] with the R package *pegas* [43].

#### Methods for genomic prediction

Data for genomic prediction models were handled in the R 3.2.2 environment [32] with the R packages *synbreed* [44] and *BGLR* [45]. The results were visualized with the *ggplot2* package [46].

#### Genetic relationship matrices

Kinship coefficients between individuals of the three-generation pedigree were estimated from pedigree and genomic data. Two expected additive genetic relationship matrices (matrix **A**) based on pedigree were derived. The first, **A<sub>P</sub>**, used only data for the maternal parents and corresponds to polymix breeding, in which only the maternal parents are known. For the second (**A<sub>F</sub>**), the full pedigree was used. G0 plants were considered to be unrelated and no population structure was identified [20]. In parallel to pedigree-based matrices, a realized genomic relationship matrix (matrix **G**) was also calculated, as described by Van Raden [47].

$$\mathbf{G} = \frac{(\mathbf{M}-\mathbf{P})(\mathbf{M}-\mathbf{P})'}{2 \sum p_i(1-p_i)} \quad (1)$$

where **M** and **P** are two matrices of dimension  $n$  (number of individuals)  $\times p$  (number of markers). **M** is the matrix of gene content, with values of -1, 0, and 1, for one homozygote, the heterozygote, and the other homozygote, respectively. **P** is the matrix of allele frequencies in the following form  $2(p_i - 0.5)$ , where  $p_i$  is the observed allele frequency at the marker  $i$  for all individuals. Use of the matrix of minor allele frequency scales **G** such that it lies on the scale of the expected additive genetic relationships matrix derived from the pedigree.

#### Statistical models for genomic prediction

We used genomic BLUP and Bayesian LASSO [48] to predict genomic estimated breeding values. The classical genetic evaluation (BLUP) was used to predict genomic estimated breeding values (GEBV) by a mixed model approach, in which the pedigree-based relationship matrix **A** was replaced with the realized genetic relationship matrix **G**. The methods used have been described in detail elsewhere [25], but we summarize them in brief here.

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Z}\mathbf{u} + \mathbf{e} \quad (2)$$

where **y** is a vector of the pseudo-phenotypes (EBV) (dimension  $n \times 1$ ),  $\mu$  is the overall mean with a vector of **1**, **Z** is a design matrix of the random effects with  $n \times n$  dimensions, **u** is the vector of random tree effect ( $n \times 1$ )  $\mathbf{u} \sim N(0, \mathbf{G}\sigma_u^2)$ , and **e** is the vector of residuals (dimension  $n \times 1$ ) with expectations  $\mathbf{e} \sim N(0, \mathbf{I}_n\sigma_e^2)$ . The diagonal

elements of the residual variance covariance matrix  $\mathbf{R}$  are prediction accuracies. For the prediction of GEBV, the  $\mathbf{G}$  matrix derived from DNA markers is used to solve mixed model equations:

$$\begin{bmatrix} \mathbf{1}'\mathbf{1} & \mathbf{1}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{1} & \mathbf{Z}'\mathbf{Z} + \mathbf{G}^{-1}\alpha \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\mu}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{1}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix} \quad (3)$$

where  $\mathbf{G}^{-1}$  is the inverse of the realized genomic relationship matrix,  $\alpha$  is the residual variance ( $\sigma_e^2$ ) divided by the variance associated with the random tree effect  $\sigma_u^2$ . This ratio is equal to the sum across loci  $2\sum p_i(1 - p_i)$  times the ratio  $\sigma_e^2/\sigma_a^2$  where  $\sigma_a^2$  represents the total genetic variance and  $p_i$  is the minor allele frequency at the  $i^{\text{th}}$  locus. The  $\mathbf{G}^{-1}$  matrix was replaced with the  $\mathbf{A}^{-1}$  matrix for predictions of the breeding values of individuals from expected genetic relationships. GBLUP assumes that markers have the same effects and that each marker has a small effect on the phenotype.

We tested the marker specific shrinkage model, Bayesian LASSO and compared it to GBLUP in terms of GEBV reliability. The linear model has the form:  $\mathbf{y} = \boldsymbol{\mu} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ , where  $\mathbf{X}$  ( $n \times p$ ) is the incidence matrix of markers,  $\boldsymbol{\beta}$  ( $p \times 1$ ) is the vector of marker effects, and  $\boldsymbol{\varepsilon}$  ( $n \times 1$ ) is the random residual effect with expectations  $\boldsymbol{\varepsilon} \sim N(0, \mathbf{I}_n\sigma_\varepsilon^2)$ . The solutions of marker effects are obtained as

$$\hat{\boldsymbol{\beta}}_L = \arg \min_{\boldsymbol{\beta}} \left\{ |\mathbf{y} - \mathbf{X}\boldsymbol{\beta}|^2 + \lambda \sum_{i=1}^p |\beta_i| \right\} \quad (4)$$

The expression outside the curly brackets minimizes the error variance. The shrinkage of markers towards the intercept is marker-specific and regulated by the  $\lambda$  parameter [49]. The coefficients of uninformative markers are shrunk to exactly zero, reducing the complexity of the model and this can be used as the basis of a model selection method. A scaled inverse  $\chi^{-2}$  prior with  $df_\varepsilon$  degrees of freedom and scale parameter  $S_\varepsilon$  was assigned as a flat prior to residual effect as  $\sigma_\varepsilon^2 \sim \chi^{-2}(\sigma_\varepsilon^2, S_\varepsilon)$ . We used the same priors and rate parameters as Isik et al. [25] for the Bayesian LASSO regression coefficients. The vector  $\boldsymbol{\beta}_L$  is assumed to have a multivariate normal distribution with marker-specific prior variances with expectations  $\boldsymbol{\beta}_L \sim N(0, \mathbf{T}(\sigma_\varepsilon^2))$ , where  $\mathbf{T} = \text{diag}(t_1^2, \dots, t_q^2)$ . We assigned  $t_j^2$  parameters independently and used identically distributed exponential priors,  $t_j^2 \sim \text{Exp}(p(\lambda^2))$  for  $j = 1, \dots, q$ , where parameter  $\lambda^2$  is given a gamma prior distribution with hyper-parameters  $r$  (shape) and  $\delta$  (rate), giving  $\lambda^2 \sim \text{gamma}(r, \delta)$  [48, 50].

**Definition of the calibration and validation sets and model evaluation**

Based on the reference population; two different validation methods were used to evaluate the effect of the structure of the calibration set on genomic prediction accuracies: subset validation and progeny validation (Fig. 1).

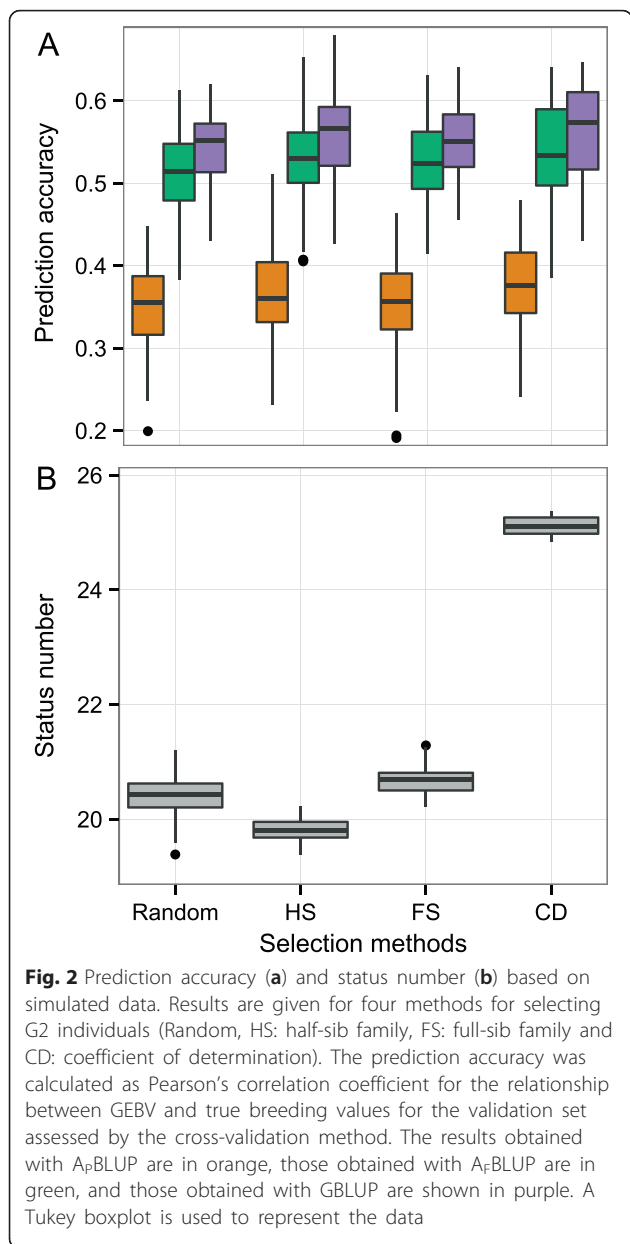
The subset validation method, in which the G2 population was split into calibration and validation sets, evaluated the effect of the relatedness of the calibration and validation sets on prediction accuracy. Three different sampling strategies were used to sample 20 % of the G2 population to form the validation set: i) random selection of G2 trees (random), ii) selection of G2 trees from the same half-sib families, to obtain a low level of relatedness between the calibration and validation sets (S1), iii) sampling of G2 trees from different full-sib families, to obtain a high level of relatedness between the calibration and validation sets (S2). For each sampling strategy, two types of calibration sets were used to evaluate the effect of pedigree depth. The first was the remaining 80 % of the G2 population and the second was the remaining 80 % of the G2 population plus all progenitors (G0 and G1). Model fit statistics were obtained for 100 replications for each scenario.

In addition to subset validation (different sampling approaches applied to G2 trees), we performed progeny validation to evaluate the prediction accuracy of GS models over generations. The individuals of the G0 and G1 generations were used as the calibration set and the individuals of the G2 generation were used as the validation set. This second validation method was used to assess the accuracy of genomic prediction models across generations, with the model trained on ancestral generations (Gn, Gn-1, etc.) and validated on progeny generation (Gn + 1). The prediction accuracy of GS models was estimated as the coefficient of correlation between the genomic estimate breeding values (GEBV) of the validation set and the EBV obtained by TREEPLAN evaluation. The prediction bias was calculated as the slope of the regression line between EBV and GEBV. A slope of  $b > 1$  indicates deflation and a slope of  $b < 1$  indicates inflated predictions.

**Results**

**Design of the reference population**

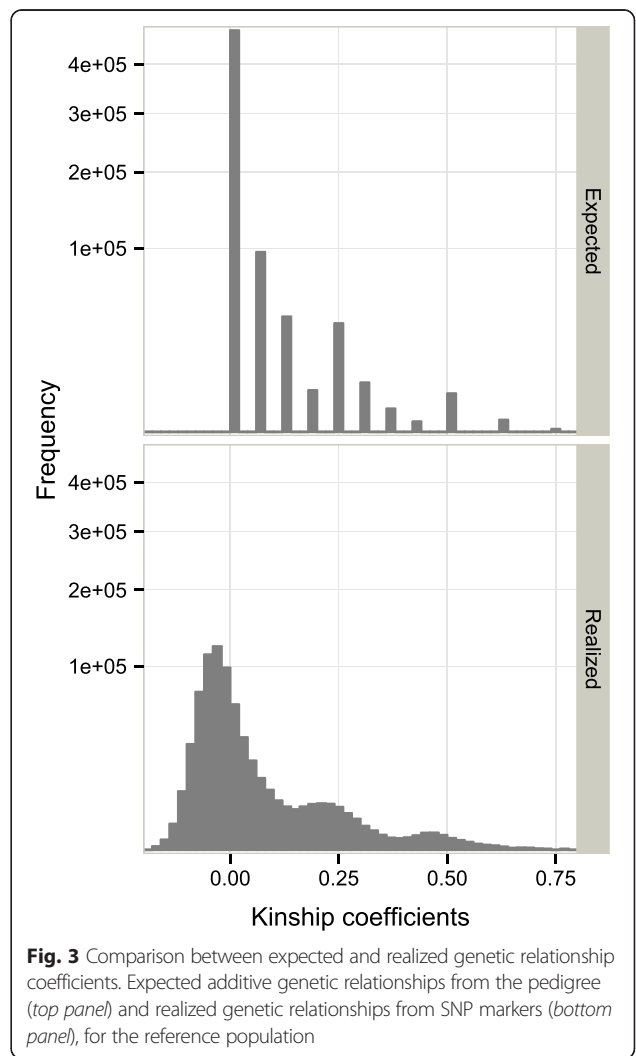
For all pre-selection methods, (Random, HS, FS and CD), use of the full-pedigree information (matrix  $\mathbf{A}_F$ ) substantially increased the prediction accuracy of GS models ( $p < 0.05$ ) over that for the partial pedigree based only on maternal information (matrix  $\mathbf{A}_P$ , Fig. 2a). Small but significant increases in prediction accuracy (0.03 on average,  $p < 0.05$ ) were achieved by using GBLUP rather than  $\mathbf{A}_P$ BLUP. For example, for the HS selection method, the mean accuracies of genomic predictions



were 0.53 for  $A_P$ BLUP and 0.56 for GBLUP (Additional file 1: Table S1). For all relationship matrices, the CD method performed significantly better than the other three methods (Fig. 2a). However, the differences were small: the mean prediction accuracies for GBLUP were 0.54, 0.56, 0.55 and 0.56 for the Random, HS, FS and CD selection methods, respectively. Status number depended on the selection method used (Fig. 2b). The highest  $N_S$  value was obtained for the CD selection method (25.1 on average). This value was significantly higher than those for the HS (19.8), FS (20.7) or Random (20.4) methods (Additional file 1: Table S1). We therefore used the CD method to select the reference population, as it gave the highest prediction accuracy and  $N_S$ .

**Characterization of the reference population**

The reference population selected by the CD method comprised 818 individuals from the three generations (Additional file 1: Figure S3): 710 G2 trees and all their progenitors (62 G1 and 46 G0). The G2 individuals came from 35 maternal half-sib families, corresponding to 355 full-sib families. The number of individuals per half-sib family ranged from 13 to 34, with a mean value of 22.2. As expected, given the low level of relatedness in the population (founder G0 trees are not related), a large majority of the kinship coefficients estimated from the pedigree were zero. The coefficients obtained were grouped into 11 classes and ranged from 0 to 0.75 (Fig. 3). By using markers, we were able to estimate the proportion of the genome shared by different individuals. The relationships predicted from markers were more consistent with the actual relationships than the expected genetic relationships derived from the pedigree (Fig. 3). Unlike the expected genetic relationships derived from the pedigree, the realized genetic relationships in the G matrix





were continuously distributed, with values between -0.18 and 0.77 (Fig. 3). Some of the realized genetic relationships were negative, suggesting that some individuals shared fewer markers than expected on the basis of allele frequencies. Similarly, some pairs of coefficients were positive and close to zero due to the sharing of a larger number of alleles than expected from allele frequencies.

The extent of LD in the reference population was estimated by calculating  $r^2$  from 3962 markers mapped onto the *P. pinaster* composite map. A rapid decrease in intra-chromosomal LD was observed for an inter-marker distance of about 5 cM on all linkage groups (Additional file 1: Figure S4). The overall LD was close to zero (average  $r^2 = 0.016$ ) and only a few marker pairs (0.5 %) had  $r^2$  values greater than 0.4. Most of the markers concerned (96.5 %) were physically linked (on the same contig) or genetically linked (less than 5 cM apart on the composite map). The remaining markers displaying high levels of LD (2.5 %) probably reflected a bias in composite linkage map construction rather than true long-distance LD, as suggested by their positions on component maps. Changes in allele frequencies were observed between G0 and G1, with an  $F_{ST}$  value greater than 0.05 for 19 SNPs, mostly located on chromosomes 5, 6, 9 and 12 (Additional file 1: Figure S5). By contrast, no difference was observed between G1 and G2. Overall, almost no differentiation was found between generations, with a global  $F_{ST} < 0.01$  between G0 and G1 and between G1 and G2.

**Prediction accuracy of genomic selection models for the reference population**

**Effect of calibration set structure on accuracy**

The mean prediction accuracies of models using 80 % of the G2 for the calibration set ranged from 0.52 to

0.87, depending on the trait and the scenario considered (Table 2). When G0 and G1 trees were added to the calibration set, mean prediction accuracies ranged from 0.66 to 0.91. Whatever the calibration set or trait considered, mean prediction accuracies for models using only pedigree information (ABLUP) were higher than those for models using marker information (GBLUP or B-LASSO, Additional file 1: Figure S6). This difference was larger (up to 0.1 larger on average) for the scenario including the progenitors of the G2 trees (generations G0 and G1) in the calibration set, suggesting that it is important to use a deep pedigree to increase prediction accuracy. As expected, when the level of relatedness between the calibration and validation sets was low (S1), mean prediction accuracy was lower than that for random sampling or S2 sampling (Table 2, Additional file 1: Figure S6). Overall, the prediction accuracy of S2 was about 0.17 lower than that for S1, for all traits and all models, if only G2 trees were used for the calibration set. Inclusion of the progenitors of the G2 trees in the calibration set resulted in a much smaller difference in prediction accuracy between S1 and S2 (maximum difference of 0.03). For random or S2 sampling, the gain in prediction accuracy achieved by adding the progenitors of G2 trees to the calibration set was smaller (0.03 and 0.02 on average, for random and S2 sampling, respectively) than that for S1 sampling (0.12 on average). However, not all traits followed this general trend. For example, the increase in prediction accuracy for stem straightness was close to zero when progenitors of the G2 trees were added to the calibration population, for the GBLUP and B-LASSO models (Table 2, Additional file 1: Figure S6).

**Table 2** Comparison of prediction accuracies across three sampling and two calibration strategies. Three sampling strategies for the selection of 20 % of the G2 population as the validation set were applied: random, S1: between half-sib families and S2: within full-sib families. Two calibration strategies were used for each sampling strategy. For predictions for the 20 % of the G2 population selected, we used the remaining 80 % of the G2 plus their progenitors (G0 and G1) as the calibration set. The mean prediction accuracy (and range) for models based on pedigree information (ABLUP) and marker information (GBLUP and B-LASSO), and the results for the three traits studied (tree diameter, height and stem straightness) are presented

		Calibration set: 80 % of the G2			Calibration set: 80 % of the G2 + G0/G1		
		ABLUP	GBLUP	B-LASSO	ABLUP	GBLUP	B-LASSO
Circumference	Random	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 (0.8–0.89)	0.75 (0.68–0.84)	0.75 (0.68–0.82)
Height	Random	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 (0.6–0.76)	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 (0.6–0.77)	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 straightness	Random	0.86 (0.8–0.90)	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)

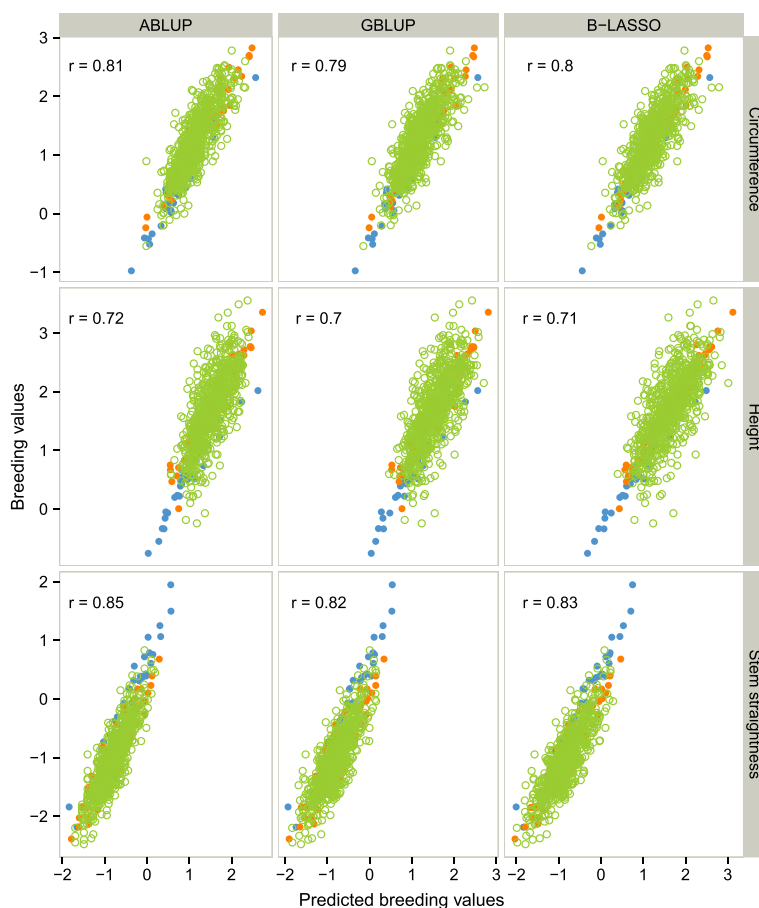
**Predictive value of markers across generations with progeny validation**

The prediction accuracies of models using only the G0 and G1 genotypes for the calibration set and only G2 for the validation set ranged from 0.70 to 0.85, depending on the trait and the method considered (Fig. 4, Additional file 1: Table S2). For all traits, ABLUP had a similar or slightly higher (up to 0.03) prediction accuracy than genomic predictions (GBLUP and B-LASSO). For all models and all three traits (except for circumference with B-LASSO model), a bias greater than one was observed, indicating that GEBV was overestimated relative to EBV. The B-LASSO model had the lowest bias: 0.99, 1.07 and 1.06 for circumference, height and stem straightness, respectively. Conversely, ABLUP had the highest bias, at 1.15, 1.22 and 1.36 for circumference, height and stem straightness, respectively (Fig. 4, Additional file 1: Table S2).

**Discussion**

**Factors affecting the prediction accuracy of GS models**

Our reference population was specifically designed to maximize prediction accuracy given the available genetic material. By contrast to previous GS studies on forest trees, we used simulation to select individuals on the basis of an explicit criterion maximizing the expected prediction accuracy for the population. As a result, we obtained medium-to-high prediction accuracies for all three traits studied (0.52 to 0.91), consistent with published results for forest tree species (Table 1). Indeed, despite differences in species, population structure and GS models between studies, similar accuracies were reported for height in eucalyptus hybrids [26] and loblolly pine [51], with values ranging from 0.66 to 0.79 for eucalyptus and from 0.64 to 0.74 for loblolly pine, depending on effective population size ( $N_e$ ) or environment. No clear trend was observed for the relationship between



**Fig. 4** Relationship between predicted breeding values (x-axis) and empirical breeding values (y-axis) for the progeny validation method. The three traits (circumference, height and stem straightness) and three different models (ABLUP, GBLUP and B-LASSO) are represented. The prediction accuracy ( $r$ ) of genomic prediction models evaluated on the validation set (G2 genotypes are shown as open green circles) is indicated. Closed circles represent the calibration set with G0 genotypes ( $n = 46$ ) in blue and G1 genotypes ( $n = 62$ ) in orange

accuracy and trait heritability. Using a large number of traits, resulting in a wider range of heritability, Resende et al. [52] reported a strong correlation ( $R^2 = 0.79$ ) between predictive ability and narrow-sense heritability. Compared to a previous study on maritime pine with a broader genetic basis [25], our results showed higher prediction accuracies on the same traits. The smaller effective population size in this study, measured as status number ( $N_S = 25$ ), than in a previous study ( $N_S \approx 100$ ) and the inclusion of multiple generations might account for the higher prediction accuracies in this study. Indeed, effective population size, which is directly related to level of LD and relatedness in populations, is known to be an important factor determining GS accuracy [10]. The importance of effective population size for prediction accuracy was highlighted by Resende, MDV et al. [26], in a study using empirical datasets for *Eucalyptus* with contrasting effective population sizes:  $N_e = 11$  and  $N_e = 51$ .

The level of relatedness between the calibration and validation sets also affected GEBV estimates. We found a 0.17 difference in prediction accuracy between low (S1) and high (S2) levels of relatedness. Our findings are consistent with previous results for white spruce [24] and mice [53]. Similar results have been reported for eucalyptus, in which the GS model was applied to populations other than that used to build the prediction model [26]. In cases of a higher marker density, yielding a more stable linkage phase between markers and QTLs across populations, population-specific models have also been described [54]. Similarly, Hayes et al. [55] reported an accuracy close to zero for the use of models developed for the Holstein breed to predict GEBVs for the Jersey breed, and *vice versa*. Thus, in the presence of short-distance LD, as in the maritime pine population in this study, the relatedness of the calibration and validation sets may be the main driver of prediction accuracy [22].

#### Comparison between pedigree- and marker-based models

Given all the possible mechanisms separating genomic variants, such as SNPs, from phenotype expression and the efforts required to identify them, one of the main issues in GS studies is demonstrating the predictive value of markers relative to conventional BLUP. In this study, regardless of the scenario used, the model using pedigree information (ABLUP) had a higher prediction accuracy than marker-based models. The marker density (2.4 SNPs per cM) used to predict GEBV may account for this difference. Indeed, simulation studies have suggested that there may be a positive asymptotic relationship between marker density and prediction accuracy [14, 22, 56]. Using a deterministic approach, Grattapaglia [10] showed that the minimal density at which marker-based models achieve accuracies similar to those of ABLUP was 2–3 SNPs per cM for an effective population size below 60. In

addition, our reference population selection strategy may also have reduced the additional gain of information provided by molecular markers relative to the pedigree. Indeed, one of the steps in the selection process was pedigree recovery, which improved the estimation of BVs [31]. Indeed, Munoz et al. [57] reported that using the G matrix to correct the pedigree and re-estimate EBVs increased prediction accuracy. In the presence of pedigree errors, which are frequently reported in tree breeding programs [31, 58, 59], the differences in prediction accuracy between ABLUP and GBLUP observed in previous GS studies may be biased. However, our results are consistent with previous findings for forest trees based on simulated [14, 18] or empirical [24, 26, 60, 61] data, with conventional BLUP having an accuracy similar to or slightly higher than that of GS models, particularly for traits with a low heritability. The genetic gain per unit of time of the GS approach over conventional BLUP would therefore be dependent solely on the decrease in breeding cycle length. This decrease in breeding cycle length raises questions about the loss of genetic variation and the maintenance of long-term genetic gain relative to conventional BLUP [62–64].

#### GS accuracy over generations

This study is novel because, unlike previous empirical studies on forest trees, we assessed the predictive value of markers across generations, rather than splitting a single population in two for model development and validation [27]. GS in forest trees is likely to be used to select progeny within families without the need for progeny testing, to reduce breeding cycle length. In this case, GS evaluation must be carried out with the progeny population. During the breeding process, recombination between haplotypes should decrease the marker-QTL linkage phase. As a result, prediction accuracy would be expected to decrease over generations [11, 17]. In this study, we assessed the predictive value of the markers, using the parents (G1) and grandparents (G0) as the training set, with validation of the model on the descendants (G2). Interestingly, prediction accuracy remained high (0.70 to 0.85, depending on the trait considered) in the validation set. These accuracies were very similar to those estimated by subset validation with a high level of relatedness between the calibration and validation sets (S2), although the calibration set was larger in this second case (567 vs. 108). These results are consistent with those of Sallam et al. [28] for a five-generation population of barley and with findings for oat breeding lines and cultivars from distant generations [65]. Indeed, both studies reported consistent prediction accuracies over generations for most traits. In sugar beet, Hofheinz et al. [29] reported that prediction accuracy was similar across generations for sugar content but that it decreased by 0.4 for molasses loss. These results suggest that the predictive value

of markers across generations is sensitive to the genetic architecture of the trait. Marker density was low in this study and in the three studies described above. However, a larger number of markers should become available in the near future, because further decreases in the cost of genotyping are anticipated. Additional markers will, therefore, probably be included in GS models over generations to maintain the accuracy of GEBV at an operational level [64].

When progenitors (G0 and G1) of the G2 population were included in calibration models, differences in prediction accuracy between low (S1) and high (S2) levels of relatedness were less than 0.03. Moreover, a slight increase in prediction accuracy was observed for all scenarios, highlighting the importance of genotyping the ancestral populations, which are generally conserved in tree breeding programs, to increase prediction accuracy. Simulation studies have also highlighted the importance of including multiple generations in the calibration set, to update the prediction equation [18, 66]. Indeed, a simulation study carried out on *Cryptomeria japonica* trees generated over a period of 60 years showed that GS outperformed phenotypic selection only if the GS model was updated [18]. Sallam et al. [28] reported contrasting results for empirical data from barley, for which the inclusion of previous generations increased prediction accuracy for some traits, but decreased it for others.

### Prospects for the use of GS in the maritime pine breeding program

The maritime pine breeding program follows a recurrent selection scheme, with breeding value estimated from polycross and bi-parental progeny trials. The genetic gain achieved in the released varieties over the first two generations was estimated at 30 % for both growth and stem straightness. The improved varieties generated by this program in the future will need to be adapted to predict changes in climate, pest and disease outbreaks and the demand for diversified wood-based products. The major challenge faced in this breeding program will therefore be the integration of new traits to deliver suitable varieties. With the rapid decrease in genotyping costs and the promising results obtained for forest trees (Table 1), GS could prove an essential tool for addressing these challenges and overcoming the limitations of marker-assisted selection [27, 67]. One of the main advantages of GS is that it can be included in the framework of current genetic evaluation. Indeed, the currently used pedigree-based BLUP method could be replaced with the "single-step" GS strategy [68] with only minimal changes. This strategy is based on the integration of both genotyped and ungenotyped individuals into the genetic evaluation through a hybrid pedigree-genomic

relationship matrix [69, 70]. As an increasing number of individuals are being genotyped for higher densities of markers, the information obtained could be used, to decrease the error rate in pedigrees. By eliminating pedigree errors and adding more information (concerning the father), this method should increase the accuracy of genetic evaluation [31, 57]. In addition, GS on the progeny population should make it possible to capture the Mendelian segregation effect in families. In forest trees, crossing can generate large numbers of offspring. In the absence of GS models, all the offspring are considered to have the same mid-parent BV at the seed or seedling stage (before progeny testing) [71]. The challenge is thus to select the superior plants without progeny testing. GS models can meet this challenge, by selecting a subset of progeny on the basis of their GEBV. This should greatly shorten the breeding cycle and decrease the costs of progeny testing, which is expensive and time-consuming for forest trees. Furthermore, a more complete knowledge of the genotype of all candidates for selection should improve the management of genetic diversity and inbreeding depression. However, shortening of the breeding cycle in maritime pine should combine GS with artificial flower induction by top-grafting, as in loblolly pine [51], or by growth regulators, as suggested for *Eucalyptus* [72] and white spruce [24]. These techniques have already been successfully implemented in these species [73, 74], but not yet in maritime pine.

### Conclusion

We selected a reference population covering three generations, with a limited status number ( $N_S = 25$ ) and a marker density of 2.5 SNPs per cM, for assessment of the prediction accuracy of GS models within and across generations. We studied three major traits used in maritime pine breeding: circumference, height and stem straightness. These three traits have low heritabilities, from 0.17 to 0.32. Prediction accuracies of up to 0.85 were obtained with progeny validation, confirming the potential of GS to predict progeny performance for low-heritability traits. However, the pedigree-based model had prediction accuracies similar to or greater than that of marker-based models. The optimization of current breeding strategies based on polymix breeding will therefore be required to enhance the potential of the GS approach in the maritime pine breeding program.

### Additional file

**Additional file 1: Figure S1.** Scatter plots (lower diagonal), histograms (diagonal) and correlations, with their significance ( $H_0: r = 0$ , upper diagonal), between breeding values for the traits: circumference, height and stem straightness. Individuals from the G0 generation are in blue, G1 in orange and G2 in green. **Figure S2.** *P. pinaster* composite map [40]. Markers in red

correspond to 3965 of the 4335 SNPs used for genomic prediction analysis. **Figure S3.** Pedigree of the 818 trees comprising the reference population ( $N_S = 25$ ) with the following frequency for each generation:  $G_0 = 46$ ,  $G_1 = 62$  and  $G_2 = 710$ . Links in purple represent mother–progeny relationships and those in orange represent father–progeny relationships. Pedigree Viewer software was used to represent the relatedness between individuals from the three generations. **Figure S4.** Pairwise linkage disequilibrium (LD) based on 3962 single-nucleotide polymorphisms mapped onto the twelve linkage groups (LG) of *P. pinaster*. Only loci with minor allele frequencies greater than 0.01 were included in the analysis. **Figure S5.** Distribution of the fixation index ( $F_{ST}$ ) over the 12 chromosome of maritime pine. The top panel represents the  $F_{ST}$  between  $G_0$  and  $G_1$  and the bottom panel represents the  $F_{ST}$  between  $G_1$  and  $G_2$ . **Figure S6.** Comparison of prediction accuracies across three sampling and two calibration strategies. Three sampling strategies for the selection of 20 % of the  $G_2$  population for use as the validation set were used: random, S1: between half-sib families and S2: within full-sib families. Two calibration scenarios were used for each sampling strategy. For predictions for the 20 % of the  $G_2$  population selected, we used the remaining 80 % of the  $G_2$  (in green) plus their progenitors ( $G_0$  and  $G_1$ , in blue) as the calibration set. The results for models based on pedigree information (ABLUP) and marker information (GBLUP and B-LASSO), and the results for the three traits studied (tree diameter, height and stem straightness) are presented. The data are represented in a Tukey boxplot. **Table S1.** Prediction accuracy and status number ( $N_S$ ) for four methods of selecting  $G_2$  individuals. Prediction accuracy was estimated with three relationship matrices ( $A_P$ ,  $A_F$  and  $G$ ). Mean values (standard deviation in parentheses) are based on 100 replicates per model. The four selection methods were: Random, HS: half-sib family, FS: full-sib family and CD: coefficient of determination. **Table S2.** Prediction accuracy and bias for the use of the progeny population for validation (calibration set =  $G_0$  and  $G_1$ , validation set =  $G_2$ ). The results for the ABLUP, GBLUP and Bayesian LASSO (B-LASSO) models and for the traits tree circumference, height and stem straightness are presented. (PDF 1725 kb)

#### Abbreviations

A, expected additive genetic relationship matrix; B-LASSO, Bayesian least absolute shrinkage and selection operator; BLUP, best linear unbiased prediction; CD, coefficient of determination; EBV, estimated breeding value; FS, full-sibs;  $F_{ST}$ , Fixation index; G, realized genomic relationship matrix;  $G_0$ ,  $G_1$  and  $G_2$ , generations of the breeding program; GEBV, genomic estimated breeding value; GS, genomic selection; HS, half-sibs; LD, linkage disequilibrium;  $N_e$ , effective population size;  $N_S$ , status number; QTL, quantitative trait loci

#### Acknowledgments

The authors would like to thank the experimental unit of INRA Pierroton (UE 0570) and the GIS "Groupe Pin Maritime du Futur" for planting and the field trials and making the necessary measurements. This work was supported by the European Union Seventh Framework Programme (Procogen, no. 289841) and the INRA SelGen metaprogram. JB was awarded a postdoctoral fellowship from the "Conseil Général des Landes".

#### Availability of data and materials

Information concerning the SNP array used to genotyped our reference population are available in [38]. The datasets supporting the conclusions of this article are available upon request.

#### Authors' contributions

CB extracted the DNA. LB, MV and JVH selected the reference population. JB performed the analysis and wrote the first draft of the manuscript. LB, CP, FI and JVH were involved in the writing and editing of the manuscript. LB designed and coordinated the study. All authors have read and approved the manuscript.

#### Competing interests

The authors declare that they have no competing interests.

#### Author details

<sup>1</sup>BIOGECO, INRA, Univ. Bordeaux, 33610 Cestas, France. <sup>2</sup>Biometris, Wageningen University and Research Centre, NL-6700 AC Wageningen, The

Netherlands. <sup>3</sup>Department of Forestry and Environmental Resources, North Carolina State University, Raleigh, NC, USA. <sup>4</sup>FCBA, Biotechnology and Advanced Silviculture Department, Genetics & Biotechnology Team, 33610 Cestas, France.

Received: 18 March 2016 Accepted: 5 July 2016

Published online: 11 August 2016

#### References

1. Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*. 2001;157(4):1819–29.
2. Heffner EL, Lorenz AJ, Jannink J-L, Sorrells ME. Plant breeding with genomic selection: gain per unit time and cost. *Crop Sci*. 2010;50(5):1681–90.
3. Bernardo R, Yu J. Prospects for genomewide selection for quantitative traits in Maize. *Crop Sci*. 2007;47(3):1082–90.
4. Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME. Invited review: genomic selection in dairy cattle: progress and challenges. *J Dairy Sci*. 2009;92(2):433–43.
5. Thomson MJ. High-throughput SNP genotyping to accelerate crop improvement. *Plant Breed Biotechnol*. 2014;2(3):195–212.
6. Goddard ME, Hayes BJ. Genomic selection. *J Anim Breed Genet*. 2007;124(6):323–30.
7. Jannink J-L, Lorenz AJ, Iwata H. Genomic selection in plant breeding: from theory to practice. *Brief Funct Genomics*. 2010;9(2):166–77.
8. Heslot N, Jannink J-L, Sorrells ME. Perspectives for genomic selection applications and research in plants. *Crop Sci*. 2015;55(1):1–12.
9. Isik F, Kumar S, Martínez-García PJ, Iwata H, Yamamoto T. Chapter Three - Acceleration of Forest and Fruit Tree Domestication by Genomic Selection. In: Plomion C, Adam-Blondon A-F, editors. *Advances in Botanical Research*, vol. 74. Academic; 2015. p. 93–124. doi:10.1016/bs.abr.2015.05.002.
10. Grattapaglia D. Breeding Forest Trees by Genomic Selection: Current Progress and the Way Forward. In: Tuberosa R, Graner A, Frison E, editors. *Genomics of Plant Genetic Resources*. Springer: Netherlands; 2014. p. 651–82.
11. Habier D, Fernando RL, Dekkers JCM. The impact of genetic relationship information on genome-assisted breeding values. *Genetics*. 2007;177(4):2389–97.
12. Calus MPL, Meuwissen THE, de Roos APW, Veerkamp RF. Accuracy of genomic selection using different methods to define haplotypes. *Genetics*. 2008;178(1):553–61.
13. Piyasatian N, Fernando RL, Dekkers JCM. Genomic selection for marker-assisted improvement in line crosses. *Theor Appl Genet*. 2007;115(5):665–74.
14. Grattapaglia D, Resende MV. Genomic selection in forest tree breeding. *Tree Genet Genomes*. 2011;7(2):241–55.
15. Pszczola M, Strabel T, Mulder HA, Calus MPL. Reliability of direct genomic values for animals with different relationships within and to the reference population. *J Dairy Sci*. 2012;95(1):389–400.
16. Wong CK, Bernardo R. Genomewide selection in oil palm: increasing selection gain per unit time and cost with small populations. *Theor Appl Genet*. 2008;116(6):815–24.
17. Zhong S, Dekkers JCM, Fernando RL, Jannink J-L. Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a Barley case study. *Genetics*. 2009;182(1):355–64.
18. Iwata H, Hayashi T, Tsumura Y. Prospects for genomic selection in conifer breeding: a simulation study of *Cryptomeria japonica*. *Tree Genet Genomes*. 2011;7(4):747–58.
19. Heslot N, Yang H-P, Sorrells ME, Jannink J-L. Genomic selection in plant breeding: a comparison of models. *Crop Sci*. 2012;52(1):146–60.
20. Plomion C, Chancerel E, Endelman J, Lamy J-B, Mandrou E, Lesur I, Ehrenmann F, Isik F, Bink M, van Heerwaarden J, et al. Genome-wide distribution of genetic diversity and linkage disequilibrium in a mass-selected population of maritime pine. *BMC Genomics*. 2014;15(1):171.
21. Slavov GT, DiFazio SP, Martin J, Schackwitz W, Muchero W, Rodgers-Melnick E, Lipphardt MF, Pennacchio CP, Hellsten U, Pennacchio LA, et al. Genome resequencing reveals multiscale geographic structure and extensive linkage disequilibrium in the forest tree *Populus trichocarpa*. *New Phytol*. 2012;196(3):713–25.
22. Habier D, Fernando RL, Garrick DJ. Genomic BLUP decoded: a look into the black box of genomic prediction. *Genetics*. 2013;194(3):597–607.
23. Hickey JM. Sequencing millions of animals for genomic selection 2.0. *J Anim Breed Genet*. 2013;130(5):331–2.

24. Beaulieu J, Doerken T, Clement S, MacKay J, Bousquet J. Accuracy of genomic selection models in a large population of open-pollinated families in white spruce. *Heredity*. 2014.
25. Isik F, Bartholomé J, Farjat A, Chancerel E, Raffin A, Sanchez L, Plomion C, Bouffier L. Genomic selection in maritime pine. *Plant Sci*. 2016;242:108–19.
26. Resende MDV, Resende MFR, Sansaloni CP, Petroli CD, Missiaggia AA, Aguiar AM, Abad JM, Takahashi EK, Rosado AM, Faria DA, et al. Genomic selection for growth and wood quality in *Eucalyptus*: capturing the missing heritability and accelerating breeding for complex traits in forest trees. *New Phytol*. 2012;194(1):116–28.
27. Isik F. Genomic selection in forest tree breeding: the concept and an outlook to the future. *New For*. 2014;45(3):379–401.
28. Sallam AH, Endelman JB, Jannink J-L, Smith KP. Assessing genomic selection prediction accuracy in a dynamic Barley breeding population. *The Plant Genome*. 2015;8(1).
29. Hofheinz N, Borchardt D, Weissleder K, Frisch M. Genome-based prediction of test cross performance in two subsequent breeding cycles. *Theor Appl Genet*. 2012;125(8):1639–45.
30. Illy G. Recherches sur l'amélioration génétique du Pin maritime. *Ann Sci For*. 1966;1966:765–948.
31. Vidal M, Plomion C, Harvengt L, Raffin A, Boury C, Bouffier L. Paternity recovery in two maritime pine polycross mating designs and consequences for breeding. *Tree Genet Genomes*. 2015;11(5):1–13.
32. R Core Team. R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing; 2015. <https://www.r-project.org/>.
33. Laloë D. Precision and information in linear models of genetic evaluation. *Genet Sel Evol*. 1993;25(6):1–20.
34. Laloë D, Phocas F, Mégnier F. Considerations on measures of precision and connectedness in mixed linear models of genetic evaluation. *Genet Sel Evol*. 1996;28(4):1–20.
35. Lindgren D, Gea L, Jefferson P. Loss of genetic diversity monitored by status number. *Silvae genetica*. 1996;45(1):52–8.
36. McRae T, Dutkowsky G, Pilbeam D, Powell M, Tier B. Genetic evaluation using the TREEPLAN system. Charleston: IUFRO; 2004.
37. Garrick D, Taylor J, Fernando R. Deregressing estimated breeding values and weighting information for genomic regression analyses. *Genet Sel Evol*. 2009;41(1):55.
38. Plomion C, Bartholomé J, Lesur I, Boury C, Rodríguez-Quilón I, Lagravelle H, Ehrenmann F, Bouffier L, Gion JM, Grivet D, et al. High-density SNP assay development for genetic analysis in maritime pine (*Pinus pinaster*). *Mol Ecol Resour*. 2016;16(2):574–87.
39. Abecasis GR, Cherny SS, Cookson WO, Cardon LR. Merlin - rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet*. 2002; 30(1):97–101.
40. de Miguel M, Bartholomé J, Ehrenmann F, Murat F, Moriguchi Y, Uchiyama K, Ueno S, Tsumura Y, Lagravelle H, de Maria N, et al. Evidence of intense chromosomal shuffling during conifer evolution. *Genome Biol Evol*. 2015; 7(10):2799–809.
41. Canales J, Bautista R, Label P, Gómez-Maldonado J, Lesur I, Fernández-Pozo N, Rueda-López M, Guerrero-Fernández D, Castro-Rodríguez V, Benzekri H, et al. *De novo* assembly of maritime pine transcriptome: implications for forest breeding and biotechnology. *Plant Biotechnol J*. 2014;12(3):286–99.
42. Weir BS, Cockerham CC. Estimating F-statistics for the analysis of population structure. *Evolution*. 1984;38(6):1358–70.
43. Paradis E. pegas: an R package for population genetics with an integrated-modular approach. *Bioinformatics*. 2010;26(3):419–20.
44. Wimmer V, Albrecht T, Auinger H-J, Schön C-C. synbreed: a framework for the analysis of genomic prediction data using R. *Bioinformatics*. 2012;28(15):2086–7.
45. Pérez P, de los Campos G. Genome-wide regression and prediction with the BGLR statistical package. *Genetics*. 2014;198(2):483–95.
46. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer Publishing Company, Incorporated; 2009. doi: 10.1007/978-0-387-98141-3.
47. Van Raden PM. Efficient methods to compute genomic predictions. *J Dairy Sci*. 2008;91(11):4414–23.
48. Pérez P, de los Campos G, Crossa J, Gianola D. Genomic-enabled prediction based on molecular markers and pedigree using the Bayesian linear regression package in R. *The Plant Genome*. 2010;3(2):106–16.
49. Park T, Casella G. The bayesian lasso. *J Am Stat Assoc*. 2008;103(482): 681–6.
50. Gianola D. Priors in whole-genome regression: the Bayesian alphabet returns. *Genetics*. 2013;194(3):573–96.
51. Resende MFR, Muñoz P, Acosta JJ, Peter GF, Davis JM, Grattapaglia D, Resende MDV, Kirst M. Accelerating the domestication of trees using genomic selection: accuracy of prediction models across ages and environments. *New Phytol*. 2012;193(3):617–24.
52. Resende MFR, Muñoz P, Resende MDV, Garrick DJ, Fernando RL, Davis JM, Jokela EJ, Martin TA, Peter GF, Kirst M. Accuracy of genomic selection methods in a standard data set of loblolly pine (*Pinus taeda* L.). *Genetics*. 2012;190(4):1503–10.
53. Legarra A, Robert-Granié C, Manfredi E, Elsen J-M. Performance of genomic selection in mice. *Genetics*. 2008;180(1):611–8.
54. Windhausen VS, Atlin GN, Hickey JM, Crossa J, Jannink J-L, Sorrells ME, Raman B, Cairns JE, Tarekegne A, Semagn K, et al. Effectiveness of genomic prediction of maize hybrid performance in different breeding populations and environments. *G3 Genes Genom Genet*. 2012;2(11):1427–36.
55. Hayes BJ, Bowman P, Chamberlain A, Verbyla K, Goddard M. Accuracy of genomic breeding values in multi-breed dairy cattle populations. *Genet Sel Evol*. 2009;41(1):51.
56. Solberg TR, Sonesson AK, Woolliams JA, Meuwissen TH. Genomic selection using different marker types and densities. *J Anim Sci*. 2008;86(10):2447–54.
57. Munoz PR, Resende MFR, Huber DA, Quesada T, Resende MDV, Neale DB, Wegrzyn JL, Kirst M, Peter GF. Genomic relationship matrix for correcting pedigree errors in breeding populations: impact on genetic parameters and genomic selection accuracy. *Crop Sci*. 2014;54(3):1115–23.
58. Hansen OK, Nielsen UB. Microsatellites used to establish full pedigree in a half-sib trial and correlation between number of male strobili and paternal success. *Ann For Sci*. 2010;67(7):703.
59. Kumar S, Richardson TE. Inferring relatedness and heritability using molecular markers in radiata pine. *Mol Breed*. 2005;15(1):55–64.
60. Beaulieu J, Doerken T, MacKay J, Rainville A, Bousquet J. Genomic selection accuracies within and between environments and small breeding groups in white spruce. *BMC Genomics*. 2014;15(1):1048.
61. Gamal El-Dien O, Ratcliffe B, Klapste J, Chen C, Porth I, El-Kassaby Y. Prediction accuracies for growth and wood attributes of interior spruce in space using genotyping-by-sequencing. *BMC Genomics*. 2015;16(1):370.
62. Bastiaansen J, Coster A, Calus M, van Arendonk J, Bovenhuis H. Long-term response to genomic selection: effects of estimation method and reference population structure for different genetic architectures. *Genet Sel Evol*. 2012;44(1):3.
63. Jannink J-L. Dynamics of long-term genomic selection. *Genet Sel Evol*. 2010; 42(1):35.
64. Goddard M. Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica*. 2008;136(2):245–57.
65. Asoro FG, Newell MA, Beavis WD, Scott MP, Jannink J-L. Accuracy and training population design for genomic selection on quantitative traits in Elite North American Oats. *Plant Gen*. 2011;4(2):132–44.
66. Sonesson A, Meuwissen T. Testing strategies for genomic selection in aquaculture breeding programs. *Genet Sel Evol*. 2009;41(1):37.
67. Muranty H, Jorge V, Bastien C, Lepoittevin C, Bouffier L, Sanchez L. Potential for marker-assisted selection for forest tree breeding: lessons from 20 years of MAS in crops. *Tree Genet Genomes*. 2014;1–20.
68. Legarra A, Christensen OF, Aguilar I, Misztal I. Single step, a general approach for genomic selection. *Livest Sci*. 2014;166:54–65.
69. Aguilar I, Misztal I, Johnson DL, Legarra A, Tsuruta S, Lawlor TJ. Hot topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score1. *J Dairy Sci*. 2010;93(2):743–52.
70. Legarra A, Aguilar I, Misztal I. A relationship matrix including full pedigree and genomic information. *J Dairy Sci*. 2009;92(9):4656–63.
71. Zapata-Valenzuela J, Whetten RW, Neale D, McKeand S, Isik F. Genomic estimated breeding values using genomic relationship matrices in a cloned population of loblolly pine. *G3 Genes Genom Genet*. 2013;3(5):909–16.
72. Grattapaglia D. Marker-assisted selection in *Eucalyptus*. Rome: Food and Agriculture Organization of the United Nations (FAO); 2007.
73. Beaulieu J, Deslauriers M, Daoust G. Flower induction treatments have no effects on seed traits and transmission of alleles in *Picea glauca*. *Tree Physiol*. 1998;18(12):817–21.
74. Griffin AR, Whiteman P, Rudge T, Burgess IP, Moncur M. Effect of paclobutrazol on flower-bud production and vegetative growth in two species of *Eucalyptus*. *Can J For Res*. 1993;23(4):640–7.

75. Zapata-Valenzuela J, Isik F, Maltecca C, Wegrzyn J, Neale D, McKeand S, Whetten R. SNP markers trace familial linkages in a cloned population of *Pinus taeda*—prospects for genomic selection. *Tree Genet Genomes*. 2012;8(6):1307–18.
76. Ratcliffe B, El-Dien OG, Klapste J, Porth I, Chen C, Jaquish B, El-Kassaby YA. A comparison of genomic selection models across time in interior spruce (*Picea engelmannii* × *glauca*) using unordered SNP imputation methods. *Heredity*. 2015.

Submit your next manuscript to BioMed Central  
and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)

