
http://www.diva-portal.org

This is the published version of a paper published in *BMC Plant Biology*.

Citation for the original published paper (version of record):

Tan, B., Grattapaglia, D., Martins, G S., Ferreira, K Z., Sundberg, B. et al. (2017)
Evaluating the accuracy of genomic prediction of growth and wood traits in two Eucalyptus species and their F-1 hybrids.
BMC Plant Biology, 17: 110
https://doi.org/10.1186/s12870-017-1059-6

Access to the published version may require subscription.

N.B. When citing this work, cite the original published paper.

Permanent link to this version:
http://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-138559

Tan et al. BMC Plant Biology (2017) 17:110
DOI 10.1186/s12870-017-1059-6

BMC Plant Biology

RESEARCH ARTICLE Open Access

Evaluating the accuracy of genomic prediction of growth and wood traits in two Eucalyptus species and their F_1 hybrids

Biyue Tan[1,2], Dario Grattapaglia[3,4], Gustavo Salgado Martins[5], Karina Zamprogno Ferreira[5], Björn Sundberg[2] and Pär K. Ingvarsson[1,6*]

Abstract

Background: Genomic prediction is a genomics assisted breeding methodology that can increase genetic gains by accelerating the breeding cycle and potentially improving the accuracy of breeding values. In this study, we use 41,304 informative SNPs genotyped in a *Eucalyptus* breeding population involving 90 *E.grandis* and 78 *E.urophylla* parents and their 949 F_1 hybrids to develop genomic prediction models for eight phenotypic traits - basic density and pulp yield, circumference at breast height and height and tree volume scored at age three and six years. We assessed the impact of different genomic prediction methods, the composition and size of the training and validation set and the number and genomic location of SNPs on the predictive ability (PA).

Results: Heritabilities estimated using the realized genomic relationship matrix (GRM) were considerably higher than estimates based on the expected pedigree, mainly due to inconsistencies in the expected pedigree that were readily corrected by the GRM. Moreover, the GRM more precisely capture Mendelian sampling among related individuals, such that the genetic covariance was based on the true proportion of the genome shared between individuals. PA improved considerably when increasing the size of the training set and by enhancing relatedness to the validation set. Prediction models trained on pure species parents could not predict well in F_1 hybrids, indicating that model training has to be carried out in hybrid populations if one is to predict in hybrid selection candidates. The different genomic prediction methods provided similar results for all traits, therefore either GBLUP or rrBLUP represents better compromises between computational time and prediction efficiency. Only slight improvement was observed in PA when more than 5000 SNPs were used for all traits. Using SNPs in intergenic regions provided slightly better PA than using SNPs sampled exclusively in genic regions.

Conclusions: The size and composition of the training set and number of SNPs used are the two most important factors for model prediction, compared to the statistical methods and the genomic location of SNPs. Furthermore, training the prediction model based on pure parental species only provide limited ability to predict traits in interspecific hybrids. Our results provide additional promising perspectives for the implementation of genomic prediction in *Eucalyptus* breeding programs by the selection of interspecific hybrids.

Keywords: Genomic relationship, Genomic heritability, Two-generation, Genome annotation, High-density SNP-chip, Bayesian LASSO, GBLUP, rrBLUP

* Correspondence: par.ingvarsson@slu.se
[1]Umeå Plant Science Centre, Department of Ecology and Environmental Science, Umeå University, Umeå SE-90187, Sweden
[6]Present address: Department of Plant Biology, Uppsala BioCenter, Swedish University of Agricultural Sciences, Uppsala SE-75007, Sweden
Full list of author information is available at the end of the article

BioMed Central

© The Author(s). 2017 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

Background

Eucalyptus species and their hybrids are the most widely planted hardwoods in tropical, subtropical and temperate regions, due to their fast growth, short rotation times, wide environmental adaptability and suitability for commercial pulp and paper production [1, 2]. Interspecific hybrids of *E.grandis* and *E.urophylla*, in particular, are generally superior to their parents in growth, wood quality and biotic and abiotic stresses resistance, by inheriting both the fast growth and good rooting abilities of *E.grandis* while by maintaining disease tolerance and wide adaptability of *E.urophylla* [3]. A conventional breeding cycle toward clonal selection in hybrid populations involves mating, progeny trials, a small-scale clonal trial and a second expanded clonal trial, that together typically take between 12 and 18 years [1, 4]. To accelerate the genetic gain per unit time, new methods that can help shorten the breeding cycles are greatly needed.

Genomic prediction or genomic selection (GS) is one of the most recent developments in genomics-assisted methods that are aimed at improving breeding efficiency and genetic gains. Genomic prediction provides a genome-wide paradigm for marker-assisted selection (MAS) [5, 6]. In GS all genome-wide markers are fitted simultaneously in a model that relies on the principle of linkage disequilibrium (LD) to capture most of the relevant variation throughout the genome, whereas MAS focuses on discrete quantitative trait loci (QTLs) that have previously been detected, usually in underpowered experiments, and thus leaving most of the phenotypic variation unaccounted for [7]. GS are generally performed in three steps: (1) genotyping and phenotyping a 'reference' or 'training population' combined with the development of genomic prediction models that allow for prediction of phenotypes from genotypes; (2) validation of the predictive models in a 'validation population', i.e. a set of individuals that did not participate in model training; (3) application of the models to predict the genomic estimated breeding values (GEBVs) of unphenotyped individuals which are then selected according to their GEBVs [6]. GS has been successfully implemented in the breeding of livestock [7, 8] and crops [9, 10] and several recent papers have also exemplified its great potential in forest tree breeding [11, 12].

The accuracies of genomic prediction models vary depending on the statistical methods employed. Several methods have been developed for GS, including ridge-regression best linear unbiased prediction (rrBLUP), genomic best linear unbiased prediction (GBLUP), BayesA, BayesB, Bayesian LASSO, BayesR and reproducing kernel Hilbert space (RKHS) regression [7, 13]. These methods mainly differ in the assumptions of the distribution and variances of marker effects. For rrBLUP all loci are a priori assumed to explain an equal amount of variance and thus assumes that marker effects follow a normal distribution where all effects are shrunk to a similar and small size. [6, 14] In Bayesian methods (BayesA, BayesB, Bayesian LASSO and BayesR) the genetic variance explained by the ith locus, V_{gi}, is assumed to themselves follow a prior distribution, $p(V_{gi})$. Therefore, the variance can vary across loci, and combining the information from the prior distribution with that of the data yields an estimate of V_{gi} [6, 15]. For instance, BayesA assumes that the genetic variance follow an inverted chi-square distribution whereas Bayesian LASSO assume the genetic variance follow a double exponential distribution. The GBLUP method computes the additive genetic merits from a genomic relationship matrix and is equivalent to rrBLUP under conditions that are generally met in practice [16]. The RKHS regression model is a linear combination of the basic function provided by the reproducing kernel [17]. Recent studies have indicated that the selection of suitable statistical methods relies on the actual data at hand and the pattern of phenotypic variation in the traits of interest and with reference population used [9, 18].

Beside statistical methods, other factors are known to influence the accuracy of genomic prediction models, such as the size of the training population, number of markers employed, and relatedness between the training and validation population and, by extension, to the future selection candidates. Hayes et al. [19] found that for a given effective population size (N_e), increasing the size of reference population leads to improved accuracy of genomic predictions. Closer relationship between training population and selection candidates has also been reported to lead to a higher accuracy of genomic predictions, while enlarging the genetic diversity of the training population resulted in lower accuracy [20]. A number of simulation and empirical studies have shown that increasing the number of markers may improve the predictive accuracy as N_e also increased [9, 21–23]. However, increasing the number of markers in small N_e populations provides little or no improvement on predictive accuracy [24, 25].

Going one step further from previous studies in forest trees, where individuals of the same breeding generation were allocated to training and validation sets for the evaluation of genomic prediction models, in this study we used both the parental and progeny generations of *E. grandis*, *E. urophylla* and their F_1 hybrids to build prediction models using different subsets of parents and progeny for training and validation sets. A multi-species single-nucleotide polymorphism (SNP) chip containing 60,904 SNPs [26] were used to provide high-density genotyping of the two generations. Based on these data, we developed genomic prediction models for height, circumference at breast height (CBH), volume, wood

basic density and pulp yield, using a number of statistical methods and compared their performance to the traditional pedigree-based prediction. Furthermore, we evaluated the impact of varying the number of SNPs and the composition and size of training and validation sets on the predictive ability (PA) of genomic prediction.

Methods

Breeding population

The breeding population in this study was obtained through controlled crossings of 86 *E. urophylla* and 95 *E. grandis* trees (G0 population) following a incomplete diallel mating design, resulting in 16,660 progeny individuals (G1 population) comprising 476 full-sib families with 35 individuals per family. In 2009, the progenies were deployed in a field trial under a randomized complete block design with single-tree plots and 35 replicates per family in Belmonte (Brazil, 39.19 W, 16.06 S, 210 m above the sea level) at Veracel Celulose S.A. (Eunápolis, BA, Brazil). Our experimental population consists of 168 parents (78 of *E.urophylla* and 90 of *E.grandis*) (G0), as not all parents were still alive at the time of the study, and 958 progeny individuals (G1) sampled across 338 full-sib families by avoiding low performing trees. The number of individuals in each full-sib family ranged from one to 13 with an average of 2.8 individuals per family.

Phenotyping

For the 958 G1 samples, height, volume, and circumference at breast height (CBH) were measured at age three and six years, respectively, and wood traits (basic density and pulp yield) were measured at age five years. For the 168 G0 parents, the same traits had been measured at age seven years for *E. grandis* and at age five years for *E. urophylla*. Briefly, height was measured using a Suunto hypsometer/height meter (PM-5/1520 series) and CBH was measured with a centimetre tape at 130 cm above ground. Wood properties were estimated by employing near-infrared reflectance spectra of sawdust samples collected at breast height using a FOSS NIRSystem 5000-M and applying calibration models developed earlier by Veracel S.A..

A mixed linear model was applied to minimize the impacts of environmental and age differences on each trait.

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{u} + \mathbf{W}\boldsymbol{b} + \boldsymbol{e}$$

where $\boldsymbol{Y}$ is a vector of observations of a single trait; $\boldsymbol{\beta}$ is a vector of fixed effects, including overall mean, experimental sites and age differences; $\boldsymbol{u}$ is a vector of random additive genetic effect of individuals with a normal distribution, $\boldsymbol{u} \sim N(0,\ \mathbf{A}\ \sigma_u^2)$, $\mathbf{A}$ is a matrix of additive genetic relationships among individuals; $\boldsymbol{b}$ is a vector of random incomplete block effect nested in each experimental site; and $\boldsymbol{e}$ is a heterogeneous random residual effect in each experimental site. $\mathbf{X}$, $\mathbf{Z}$ and $\mathbf{W}$ are incidence matrices for $\boldsymbol{\beta}$, $\boldsymbol{u}$ and $\boldsymbol{b}$, respectively. The phenotypes of each trait were then corrected by subtracting variation of sites, ages and blocks effects for all individuals, and were referred to adjusted phenotypes. The adjusted phenotypic traits were used for calculating the heritability of traits and for building genomic prediction models.

Genotyping and quality control

The 168 G0 and 958 G1 populations were genotyped using the Illumina Infinium EuCHIP60K [24] that contains probes for 60,904 SNPs. EUChip60K intensity data (.idat files) were obtained through GENESEEK (Lincoln, NE, USA). SNP genotypes were called using GenomeStudio (Illumina Inc., San Diego, CA, USA) following standard genotyping and quality control procedures with no manual editing of clusters as described earlier [26]. Further quality control of the genotyped samples was performed using PLINK [27]. Nine G1 individuals with sample call rate less than 70% or inbreeding coefficients greater than one were removed for further analyses. 10,240 SNPs were excluded due to low call rates (less than 70%) and 9243 SNPs were filtered out due to monomorphism or by having minor allele frequency (MAF) less than 0.01. Finally 117 SNPs were removed because they showed strong deviations from Hardy-Weinberg equilibrium (p-value $<1 \times 10^{-6}$).

After quality control, missing genotypes of the remaining individuals were filled in by imputation. We first tested the accuracy of imputation methods across a range of missing data (2% - 30%) by artificial removing SNPs from a fraction of our genotypes. Among the available family-based and population-based methods we assessed the following programs for imputation accuracy: BEAGLE [28], fastPHASE [29], MENDEL [30], random forest, SVD Impute, k-nearest neighbors [31], BLUP A matrix, Bayesian PCA, NIPALS, Probabilistic PCA [32]. BEAGLE provided the best accuracy for all missing data percentages, with accuracies exceeding 95% in all cases (Additional file 1). We therefore used BEAGLE to impute missing genotypes at 41,304 SNPs retained after the filtering steps discussed above, across all 168 G0 and 949 G1 individuals. The imputed genotype data was subsequently used in all genomic prediction analyses. LD between SNP pairs were measured using the squared correlation coefficient (r^2) for SNPs located on the same chromosome. Following Remington et al. [33], the decay of LD versus physical distance was then modelled using a nonlinear regression method.

We further estimated population structure and pairwise genomic relationships among the 1117 individuals by performing principal components analysis (PCA) [34] and by calculating genomic relationships among individuals [14] using 10,213 independent SNPs (LD-pruned) ($r^2 < 0.2$) calculated in PLINK [27]. Pedigree-based genetic relationship was estimated by using ABLUP in ASReml (see below for further information).

Statistical methods for genomic prediction

Four statistical methods were assessed for their ability to estimate the parameters in eq. (1) and for predicting GEBVs. These methods include *genomic best linear unbiased predictor* (GBLUP) [5], *ridge regression BLUP* (rrBLUP) [6], *Bayesian LASSO* (BL) [35], and *reproducing kernel Hilbert space* (RKHS) regression [17]. These methods were chosen to represent the variety of available approaches for genomic prediction. GBLUP represents a method which does not rely on marker effect estimation; rrBLUP estimates marker effects using linear and penalized parameters; BL represents a linear, parametric and Bayesian method for marker effect estimation; whereas RKHS represents a non-linear semi-parametric method. The performance of the four genomic prediction methods was compared with that of the commonly used pedigree-based BLUP (ABLUP) [36].

The GEBVs were estimated using the following mixed linear model:

$$\boldsymbol{y} = \mathbf{1}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{a} + \boldsymbol{e} \tag{1}$$

where $\boldsymbol{y}$ is the vector of adjusted phenotypes of single trait, $\boldsymbol{\beta}$ is the vector of overall mean fitted as a fixed effect, $\boldsymbol{a}$ is the vector of random effects, and $\boldsymbol{e}$ is the vector of random residual effects. $\mathbf{1}$ and $\mathbf{Z}$ are incident matrix of $\boldsymbol{\beta}$ and $\boldsymbol{a}$, respectively.

ABLUP

ABLUP is the standard method for predicting breeding values using the expected relatedness among individuals based on pedigree information [36]. For ABLUP, the vector of random additive effects ($\boldsymbol{a}$) in Eq. (1) is assumed to follow a normal distribution $\boldsymbol{a} \sim N\left(0, \boldsymbol{A}\sigma_a^2\right)$, where $\mathbf{A}$ is the additive numerator relationship matrix estimated from pedigree information and the σ_a^2 is the additive genetic variance. The residual vector $\boldsymbol{e}$ is assumed as $\boldsymbol{e} \sim N\left(0, \boldsymbol{I}\sigma_e^2\right)$, where $\boldsymbol{I}$ is the identity matrix. Under these assumptions, Eq. (1) can be re-written as:

$$\begin{bmatrix} \boldsymbol{X}^T\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{Z} \\ \boldsymbol{Z}^T\boldsymbol{X} & \boldsymbol{Z}^T\boldsymbol{Z} + \boldsymbol{A}^{-1}\frac{\sigma_e^2}{\sigma_a^2} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^T\boldsymbol{y} \\ \boldsymbol{Z}^T\boldsymbol{y} \end{bmatrix} \tag{2}$$

where σ_e^2 and σ_a^2 are estimated using a restricted maximum likelihood method. The estimated breeding values ($\hat{\mathbf{a}}$) and fixed effects ($\hat{\boldsymbol{\beta}}$) can be calculated directly from Eq. (2). ABLUP calculations were performed using ASReml 3.0 [37].

GBLUP

The GBLUP method is derived from ABLUP, but differs in that the matrix $\mathbf{A}$ in Eq. (2) is replaced with the genomic relationship matrix ($\mathbf{G}$) that is calculated from genotypic data using $\mathbf{G} = \frac{(\mathbf{M}-\mathbf{P})(\mathbf{M}-\mathbf{P})^{\mathrm{T}}}{2\sum_{j=1}^{p} p_j\left(1-p_j\right)}$, where $\mathbf{M}$ is the matrix of samples with SNPs encoded as 0, 1, 2 (i.e. the number of minor alleles), $\mathbf{P}$ is the matrix of allele frequencies with the j-th column given by $2(p_j - 0.5)$, where p_j is the observed allele frequency of the samples [5]. In GBLUP, the random additive effects ($\boldsymbol{a}$) in the Eq. (1) is assumed to follow $\boldsymbol{a} \sim N\left(0, \mathbf{G}\sigma_g^2\right)$, where σ_g^2 is the genomic-based genetic variance and GEBVs ($\hat{\mathbf{a}}$) are again calculated from equation (2) but with $\boldsymbol{A}^{-1}$ replaced by $\boldsymbol{G}^{-1}$ and σ_a^2 replaced by σ_g^2. The GBLUP calculations were performed using ASReml 3.0 [37] and the G matrix was estimated using the "A.mat" function from the rrBLUP package in R [14].

rrBLUP

As opposed to the previous two methods, rrBLUP alters the notations of parameters $\boldsymbol{a}$ and $\boldsymbol{Z}$ in the Eq. (1), where $\boldsymbol{Z}$ now refers to a design matrix for SNP effects, rather than an incident matrix and $\boldsymbol{a}$ refers to SNP effects that are assumed to follow $\boldsymbol{a} \sim N\left(0, \mathbf{I}\sigma_m^2\right)$, where σ_m^2 denotes the proportion of the genetic variance contributed by each SNP [6]. With these alterations, Eq. (2) becomes:

$$\begin{bmatrix} \boldsymbol{X}^T\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{Z} \\ \boldsymbol{Z}^T\boldsymbol{X} & \boldsymbol{Z}^T\boldsymbol{Z} + \boldsymbol{I}\lambda \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^T\boldsymbol{y} \\ \boldsymbol{Z}^T\boldsymbol{y} \end{bmatrix} \tag{3}$$

where $\lambda = \sigma_e^2/\sigma_u^2$ is the ratio between the residual and marker variances. A prediction for the GEBV for each individual is calculated as $\hat{g}_i = \mathbf{Z}_\mathbf{i}^\mathbf{T}\hat{\mathbf{a}}$ from equation (3), where $\mathbf{Z}_\mathbf{i}^\mathbf{T}$ is the SNP vector for individual i and $\hat{\mathbf{a}}$ is the vector of estimated SNP effects. All calculations were performed using the "mixed.solve" function from the rrBLUP package in R [14].

Bayesian LASSO

The Bayesian LASSO (BL) method is the Bayesian treatment of LASSO regression as proposed by Legarra et al. [34]. In BL the vector of SNP effects, $\boldsymbol{a}$ in equation (1), is assumed to follow a hierarchical prior distribution with $\boldsymbol{a} \sim N\left(0, \mathbf{T}\sigma_m^2\right)$, where $\mathbf{T} = \text{diag}\left(\tau_1^2, \ldots, \tau_p^2\right)$. τ_j^2 is assigned as $\tau_j^2 \sim Exp\left(\lambda^2\right)$, $j = 1,\ldots,p$. λ^2 is assigned as $\lambda^2 \sim Gamma(r, \delta)$. The residual variance σ_e^2 is assigned as $\sigma_e^2 \sim \chi^{-2}(df_e, S_e)$.

We implemented the BL method using the "BLR" function from the BLR package in R [38]. Here a Monte Carlo Markov Chains sampler was applied and prior parameters (df_e, S_e, r, δ, and λ^2) were defined following the guidelines proposed by de los Campos et al. [39]. The chain length was 20,000 iterations, with the first 2000 excluded as burn-in and with a subsequent thinning interval of 100.

RKHS

RKHS assumes that the random additive effects in Eq. (1) are $\boldsymbol{a} \sim N\left(0, \mathbf{K}\sigma_g^2\right)$, where $\mathbf{K}$ is computed by means of a Gaussian kernel that is given by $K_{ij} = \exp(-hd_{ij})$ [17]. h is a semi-parameter that controls how fast the prior covariance function declines as genetic distance increase and d_{ij} is the genetic distance between two samples computed as $d_{ij} = \sum_{k=1}^{p}\left(x_{ik} - x_{jk}\right)^2$, where x_{ik} and $x_{jk}$are kth SNPs ($k = 1,\ldots,p$) for the ith and jth samples, respectively. We implemented the RKHS method through the "BGLR" function from the BGLR package in R [40], which use a Gibbs sampler for the Bayesian framework and assigns the prior distribution of σ_g^2 and σ_e^2 as $\sigma_g^2 \sim \chi^{-2}\left(df_g, S_g\right)$ and $\sigma_e^2 \sim \chi^{-2}(df_e, S_e)$, respectively. Here we chose a multi-kernel model as suggested by Perez [40], where three h values were defined as $h_1 = 2/\left(5*\overline{d}\right)$, $h_2 = 2/\overline{d}$, $h_3 = 2*5/\overline{d}$, $\overline{d}$ was the median of d_{ij}. The Gibbs chain length was 20,000 iterations with the first 2000 iterations discarded as burn-in and a thinning interval set to 100.

Heritability estimation

We estimated the pedigree-based narrow-sense heritability (h_a^2) using the relationship matrix from the ABLUP method, and the narrow-sense genomic heritability (h_g^2) using the genomic relationship matrix from GBLUP (details in [41]). The respective heritabilities were calculated as:

$$h_a^2 = \frac{\sigma_a^2}{\sigma_y^2} \qquad h_g^2 = \frac{\sigma_g^2}{\sigma_y^2}$$

where σ_a^2 is the additive variance estimated from ABLUP, while σ_g^2 is the marker-based genetic variance estimated from GBLUP. σ_y^2 is the phenotypic variance of the population.

Size and genetic composition of the training and validation sets

We simultaneously assessed the impact of the size and genetic participation of G0 and G1 individuals in the training set (TS) and validation set (VS) of the genomic prediction models. Regarding relative TS/VS sizes, we divided all 1117 (G0 and G1) individuals into five different size groups with a TS:VS ratio of 1:1, 2:1, 3:1, 4:1 or 9:1. The corresponding sizes of the TS/VS were respectively 558/559, 743/374, 836/281, 892/225 and 1003/114 individuals. Within these pre-established size compositions, four scenarios were employed where the participation of G0 and G1 individuals were evaluated to assess the impact of varying the degrees of relationship and diversity between TS and VS. In the first scenario (CV_1) assignment of individuals to either TS or VS was random. For the second scenario (CV_2) all G0 parents were assigned to the TS and complemented with a random selection of G1 individuals up to the required number in the set, while the VS was composed exclusively of the remaining G1 individuals. The third (CV_3) and fourth (CV_4) scenarios were built based on minimizing and maximizing relatedness between TS and VS. The relatedness-based assignment of individuals was determined using the procedure described in Spindel et al. [9]. Briefly, 1117 individuals were assigned to 182 clusters based on their genotypes using a k-means clustering algorithm implemented in the "pamk" function from the fpc package in R. This method attempts to minimize the distance between individuals in a cluster and the centre of that cluster. Using the relatedness estimates, CV_3 was then built by assigning individuals to TS and VS based on dissimilarity, such that individuals from the same cluster were not allowed to be both in the same TS or VS. For CV_4 individuals from same cluster were forced to be either in the TS or VS to increase relatedness within TS and VS [9].

Genomic prediction models

We evaluated the effects of the five statistical methods (GBLUP, rrBLUP, BL, RKHS and ABLUP), five TS/VS sizes and four TS/VS composition scenarios (5*5*4 = 100 models in total) on the predictive ability (PA) of genomic prediction. For each of the 100 models, 200 replicate runs were carried out for

each trait and the performance of the models were evaluated in terms of their PA ($r_{y,\ \hat{g}}$), which is defined as the Pearson correlation between the adjusted phenotypes and the GEBVs of the samples in the VS. ANOVA was performed with all effects declared as fixed on 80 out of the 100 models tested (the 20 ABLUP models were excluded) to partition the total variance into different sources (genomic prediction method, TS/VS size and genetic composition). The significant differences we found were further assessed by means of a paired t tests (α = 5%), adjusted by a Bonferroni correction. The 80 models, as described above, were used for assessing the impact of TS/VS composition and TS/VS size, while all 100 models were used to evaluate the statistical methods against ABLUP. All available SNPs were used in all the analyses of these models.

Numbers and genomic location of SNPs subsets

We finally assessed the impact of the number of SNPs and their locations (gene vs. intergenic region) on the PA of genomic prediction models. 12 subsets with different numbers of SNPs were generated by randomly selecting 10, 20, 50, 100, 200, 500, 1000, 2000, 5000, 10,000, 20,000 and 41,304 SNPs from all the available SNPs. For SNP location, SNPs subsets located in different regions of the genome were established by including SNPs located in four different regions: (i) coding sequences (CDS) only (11,786 SNPs); (ii) entire genic regions including CDS, UTRs, introns, and sequences 2 kb up and downstream of the gene (30,405 SNPs); (iii) intergenic regions (10,899 SNPs), and (iv) all 41,304 SNPs. The location and classification of each SNP was obtained by mapping SNPs onto the *E.grandis* genome using SnpEff [42]. Genomic prediction models were built for all four TS/VS compositions using only the two statistical methods (GBLUP and RKHS) that showed the best predictive performance in the previous analyses, using a TS/VS size ratio of 4:1 (892/224).

Results

Phenotypic trait correlations

Growth (height, volume, and CBH) and wood properties (basic density and pulp yield) were measured for all 168 G0 and 949 G1 individuals. The raw phenotypic data were adjusted using a mixed linear model to minimize the impacts of environment and age differences. The pairwise correlations between the adjusted traits were described by calculating Pearson correlation coefficients (Fig. 1). Growth traits were correlated with each other. Interestingly, however, while CBH and volume at age three and six years were highly correlated ($r = 0.92$ and 0.95 respectively), height at age three was only weakly correlated with height at age 6 ($r = 0.36$). For wood properties traits, basic density was negatively correlated with pulp yield, although only weakly so ($r = -0.28$). Growth traits showed no correlations with wood traits ($r = -0.1$ to 0.1).

Breeding population structure and relatedness

Population structure across G0 and G1 individuals was assessed by PCA based on 10,213 LD-pruned, independent SNPs ($r^2 < 0.2$). The first two PCs explained 6.07% and 3.8% of the total genetic variance (Fig. 2a) and clearly separated the G0 individuals of the two species, *E.grandis* and *E.urophylla,* with the *E.grandis* individuals further subdivided into two subgroups likely representing the two main provenances used in breeding programs in Brazil. The G1 individuals were generally projected into the space defined by their parents, but with a few outliers. The expected pedigree-based and realized genomic-based genomic relationships among G0 and G1 individuals were visualized using heatmaps (blue and red in Fig. 2b, respectively). The result of the genomic relationship analysis corroborated the PCA result, in which *E. urophylla* was clustered into a single group, whereas *E. grandis* formed two subgroups. The average values of the realized genomic relationships among what were considered to be full-sibs, half-sibs and unrelated individuals from the pedigree data were generally lower than the expected relationships values (0.309 vs. 0.5, 0.131 vs. 0.25 and 0.0056 vs. 0, respectively) (Table 1). This result suggests that pedigree errors were likely present in this population. These putative pedigree errors in turn negatively affected our ability to estimate the heritability of traits based on pedigree information, which were considerably lower than those estimated using genomic-based realized genetic relationships (Table 2).

Predictive abilities with different statistical methods

Estimates of PAs were obtained using different statistical methods, compositions and sizes of TS/VS for each trait (Additional file 2). An ANOVA showed that all these factors had a significant effect on the PA (*P*-value <0.005) (Additional file 3). Across the four genomic prediction methods used (GBLUP, rrBLUP, BL, and RKHS) the average PA varied from 0.27 to 0.274 (Additional file 4). All the four methods outperformed the pedigree-based ABLUP prediction (mean PA = 0.121) by an average of 80%–200% across the eight traits (Fig. 3). RKHS yielded a slightly better PAs for six out of eight traits and this method was particularly suitable for predicting traits that displayed lower heritabilities, such as CBH and height. The other three methods generally gave similar results across all traits, although with a slightly better performance than RKHS for pulp yield (Fig. 3).

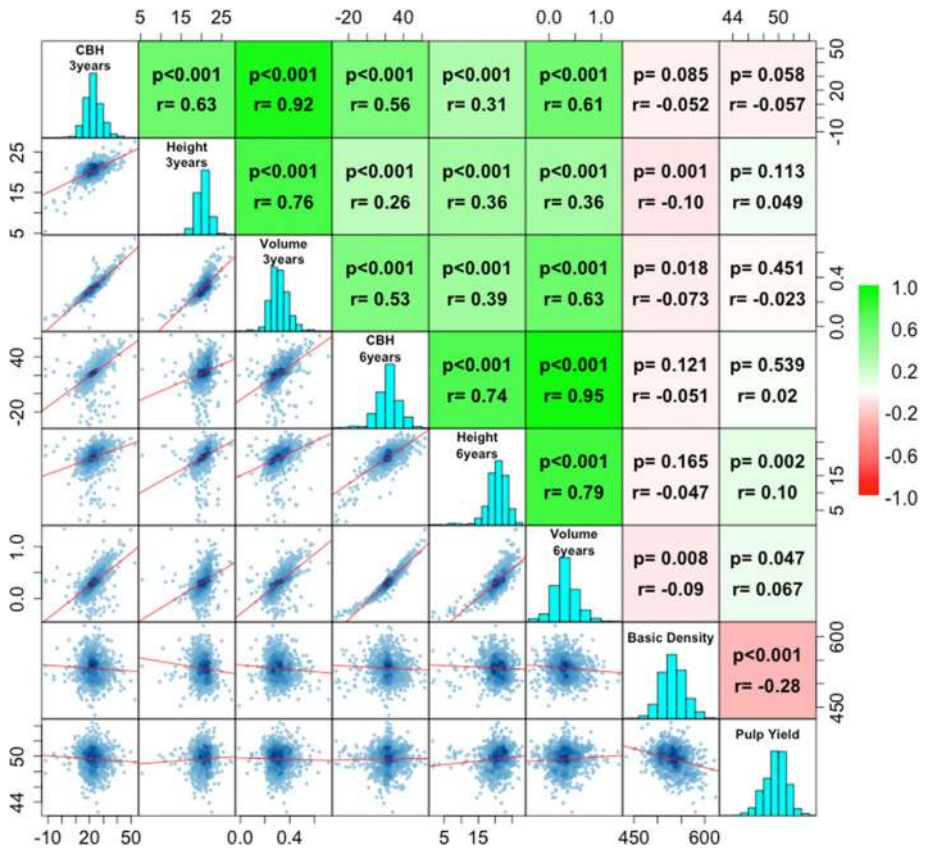

Fig. 1 Correlation and distribution of phenotypes. *Scatter plots* (*lower* off-diagonal) and correlations with probability values (*upper* off-diagonal; H_0: $r = 0$) for adjusted phenotypes between pairs of traits. *Color* key on the *right* indicates the strength of the correlations. Diagonal: histograms of the distribution of adjusted phenotypes values

Impact of TS/VS compositions and relative sizes on predictive ability

The average PAs differed significantly for the different TS/VS compositions tested and varied from 0.253 to 0.286 (Additional file 5). The genomic prediction model built with CV_2 (all G0 parents in the TS) showed the highest PAs for all traits except pulp yield, whereas models based on CV_3 (minimum relatedness between TS and VS) gave the worst predictions. The models based on CV_1 (random assignment) and CV_4 (maximum relatedness between TS and VS) showed no significant differences in PA (Fig. 4, Additional file 5). The average PA was significantly improved from 0.251 to 0.285, as the TS/VS ratio increased from 1:1 (558/559) to 9:1 (1003/113) (Additional file 6), irrespective of the prediction method (Fig. 3) or the genetic composition of TS/VS used (Fig. 4), clearly showing the importance of an adequate size of the training set to build prediction models. Furthermore, there was a steeper increase in PA when TS/VS ratio increased from 1:1 (558/559) to 2:1 (743/374) than from 2:1 (743/374) to 9:1 (1003/114) for all traits (Figs. 3 and 4).

Impact of the number of SNPs and their genomic location on predictive ability

Estimates of PA using different numbers of SNPs (Additional file 7) and subsets of SNPs in different genomic locations (Additional file 8) were obtained with two prediction methods, using a TS/VS ratio of 892/225 and using all the four different TS/VS compositions. An ANOVA showed that both the number of SNPs and their genomic location significantly affect the PA for both prediction methods (GBLUP and RKHS) (*P*-value <0.005), and that the number of SNPs has a larger impact than their genomic location (Additional file 9). The average PAs across all traits decreased from 0.278 to 0.113 when the number of SNPs used in the prediction models dropped from 41,304 to only 10, and the reduction was especially strong when the number of SNPs went below 5000 (Additional file 10). On the other hand, no significant improvement was generally seen in the average of PA when more than 5000 SNPs were used (Additional file 10, Fig. 5). The results obtained for the different traits suggest that traits with lower heritability are more sensitive to the reduction in the number of SNPs (Fig. 5). For instance, PA for basic density ($h^2 = 0.35$)

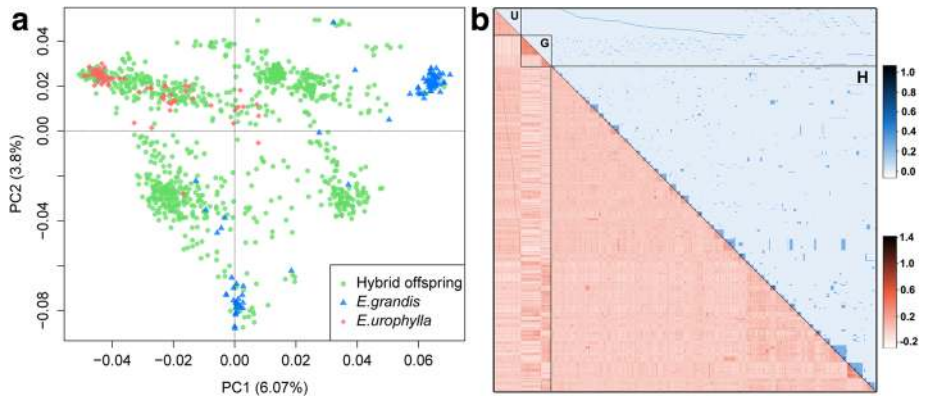

Fig. 2 Genetic structure and relatedness in the breeding population. (**a**) First two principal components of a PCA revealing population structure. *Dots* represent *E.grandis* (*blue*), *E.urophylla* (*red*) and their F_1 (*green*) individuals. (**b**) Heatmaps of the pairwise pedigree-expected relationships (*blue, upper* off-diagonal) and genomic-realized relationship (*red, lower* off-diagonal) of the 1117 individuals assigned to *E.grandis* (G), *E.urophylla* (U) and their hybrid progenies (H)

went from 0.47 to 0.24 (a 50% decrease) when the number of SNPs dropped from 40,000 to 10, whereas CBH of age three ($h^2 = 0.113$) decreased from 0.128 to 0.03 (a 77% decrease). Overall, slight significant differences were seen in PAs by using SNP sets located in different genomic regions (Fig. 6), the average PAs range from 0.270 to 0.284 (Additional file 11). Predictions using SNPs located in intergenic regions were marginally better than using SNPs in genic regions or all SNPs, except for pulp yield that could be better predicted based on models using SNPs from coding and gene regions (Fig. 6). When comparing the PA of models using SNPs in coding versus entire gene regions, the latter had a slightly better performance, most likely due to the larger number of SNPs used (30,504 vs. 11,786) and not due to any specific effect of genomic location. When we assessed the pairwise LD (r^2) among SNPs in the four regions tested, the extent of LD differed among them, with LD showing the most rapid decay in coding regions and the slowest decay in intergenic regions (Additional file 12).

Discussion

This study presents the results of an empirical evaluation of the accuracy of genomic prediction on growth and wood quality traits in *Eucalyptus* using data from a high-density SNP array. Our results are based on data from a two generations breeding population and provide additional encouraging results on the prospects of using genomic prediction to accelerate breeding. We have assessed a range of factors, including the statistical methods used to estimate predictive ability, the size and composition of the training and validation sets as well as the number and genomic locations of SNPs used in the prediction model. Hereafter we will discuss how these factors influenced the prediction accuracy.

Genomic data corrected pedigree inconsistencies

All four genomic prediction methods performed significantly better than the pedigree-based evaluations for all complex traits assessed (Fig. 3). While similar results have been reported for animals [18, 43] and crop species [9, 36] across a number of traits, in forest trees prediction accuracies using genomic data have generally been similar or up to 10–30% lower than accuracies obtained using pedigree-estimated breeding values, including *Eucalyptus* [4], loblolly pine (*Pinus taeda*) [44], white spruce (*Picea glauca*) [45, 46], interior spruce (*Picea engelmannii* × *glauca*) [47, 48] and maritime pine (*Pinus pinaster*) [49]. Genomic predictions with lower accuracies than pedigree-based predictions could arise from insufficient marker density, such that not all casual variants are captured in the genomic estimate [41], or an overestimate of the pedigree-based prediction due to its inability of ascertaining the true genetic relationships in half-sib families [47]. Our result however differ from previous studies in forest trees due to the fact that the

Table 1 Pairwise expected pedigree-based and realized genomic-based relationships in the different family types

	Full-sib families (961)[a]			Half-sib families (12718)			Unrelated individuals (434252)		
	Min	Mean	Max	Min	Mean	Max	Min	Mean	Max
Pedigree-expected relationship	0.5	0.5	0.5	0.25	0.25	0.25	0	0	0
Genomic-realized relationship	−0.274	0.309	0.933	−0.464	0.131	0.908	−0.467	−0.056	0.891

[a]Number in parentheses indicate the number of pairwise estimates

Table 2 Pedigree-based and genomic heritabilities for each trait

	CBH (3)[a]	Height (3)	Volume (3)	CBH (6)	Height (6)	Volume (6)	Basic density	Pulp yield
h_a^{2b}	0.051(0.03)	0.074(0.04)	0.057(0.03)	0.085(0.04)	0.097(0.05)	0.068(0.04)	0.23(0.04)	0.27(0.05)
h_g^2	0.113(0.04)	0.171(0.05)	0.162(0.04)	0.184(0.04)	0.193(0.05)	0.196(0.04)	0.35(0.05)	0.46(0.05)

[a]Number in parentheses correspond the age at measurement;
[b]h_a^2 and h_g^2 correspond to the pedigree and genomic narrow-sense heritability, respectively, with their standard deviation in parenthesis

average pairwise estimates of genetic relationship among individuals were substantially lower using SNP data than expectations based on pedigree information (Table 1), clearly suggesting that the expected pedigrees, and consequently the pairwise relationships, had considerable inconsistencies that were corrected by the SNP data. We speculate that these inconsistencies likely derived from pollen contamination and/or mislabelling in the process of generating the full and half-sib families. Besides correcting potential pedigree errors, the relatively dense SNP data used in our study also was able to accurately capture the Mendelian sampling variation within families so that genetic variances estimates were based on the true proportion of the genome that is identical by descent (IBD) or state (IBS) among half- or full-sib individuals, resulting in improved estimates of trait heritability (Table 2).

Genomic predictions show that traits adequately fit the infinitesimal model

Overall, the different genomic prediction methods provided similar results for the all traits evaluated, with only a slight advantage for RKHS which showed better PAs for the low-heritability growth traits (Fig. 3). However, for pulp yield, RKHS was instead the worst performing method, and it is possible that the definition of a kernel simply was not suitable for this particular trait [17]. Our results corroborate previous reports from both crops and animals [18, 50, 51], as well as forest trees. In loblolly pine, for example, the performance of rrBLUP and three Bayesian methods were only marginally different when compared across 17 traits with distinct heritabilities, with a small improvement seen for BayesA only for fusiform rust resistance where loci of relatively larger effect have been described [44]. Similar results were obtained for growth and wood traits in other forest trees showing no performance difference between rrBLUP and Bayesian methods [46, 48, 49]. This occurs despite simulation studies suggesting that Bayesian methods, like BL, should outperform univariate methods such as rrBLUP and GBLUP [6, 52, 53]. One possible reason for the apparent disagreement between simulations and empirical data sets could be that the true QTL effects for most of traits are relatively small and the distribution is less extreme than in simulated data [54]. Our results therefore support the proposal that either rrBLUP or GBLUP are effective methods in providing the best compromise between computational time and prediction efficiency [55] and that the quantitative traits assessed in our study adequately fit the assumption of the infinitesimal model.

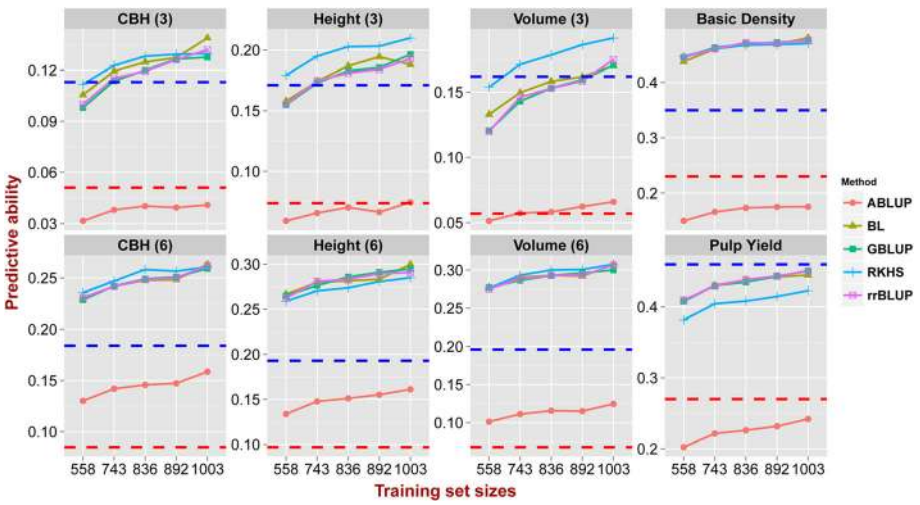

Fig. 3 Predictive abilities with different methods and increasing sizes of training sets. Predictive ability (y axis) estimated using five methods across five training set/validation set sizes in numbers of individuals (x axis) 558/559, 743/374, 836/281, 892/225 and 1003/114. *Red* and *blue dashed lines* indicate the pedigree-based (h_a^2) and genomic-realized (h_g^2) narrow-sense heritability respectively

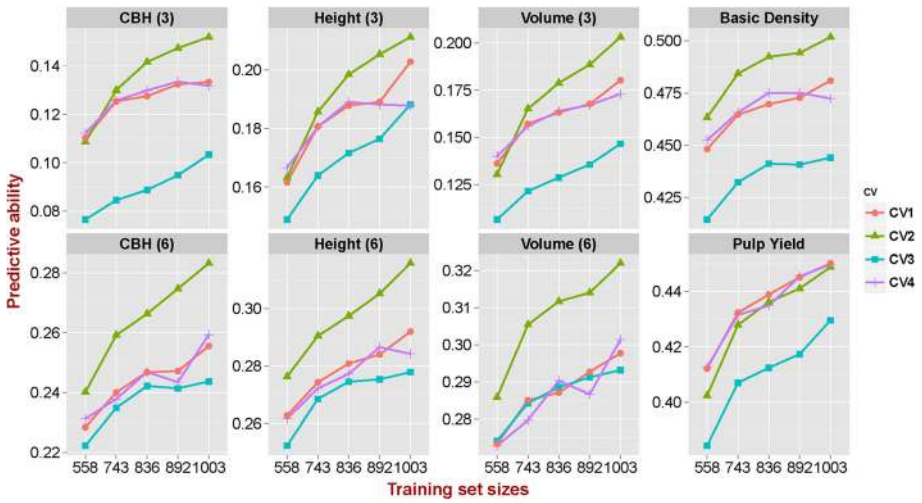

Fig. 4 Predictive abilities with variable levels of relatedness between training and validation sets. CV_1: random assignment of individuals to either training set (TS) or validation set (VS); CV_2: all the G0 pure species parents assigned to the TS; CV_3: minimum relatedness between TS and VS individuals; CV_4: maximum relatedness between TS and VS individuals. Estimates were obtained using GBLUP and RKHS across five TS/VS sizes in numbers of individuals (x axis): 558/559, 743/374, 836/281, 892/225 and 1003/114

Training set size, composition and relatedness strongly affect predictive ability

Our results show that the size and compositions of training and validation sets had the largest impact on the PA, irrespective of the analytical method used (Fig. 4). The average PA rapidly increased with increasing sizes of the TS and did not show any sign of plateauing. Earlier simulations of *Eucalyptus* breeding scenarios had in fact shown that with up to $N = 1000$ individuals in the TS, the accuracy would rapidly increase, and additional gains were seen up to $N = 2000$ individuals for traits with low heritabilities, for larger numbers of QTLs involved in traits and for larger effective population size (N_e). After $N = 2000$ the predictive accuracy would tend to plateau irrespective of the N_e and genotyping density [22]. Simulations [19, 56] and proof-of-concept studies [57] in crop species also show improved PA with larger TS sizes. Larger training populations alleviate the probability of losing rare favourable alleles from the breeding population as generations of selection advance. Additionally, by sampling more individuals for training, a larger diversity is captured and better estimates of the marker effects are obtained which in turn positively

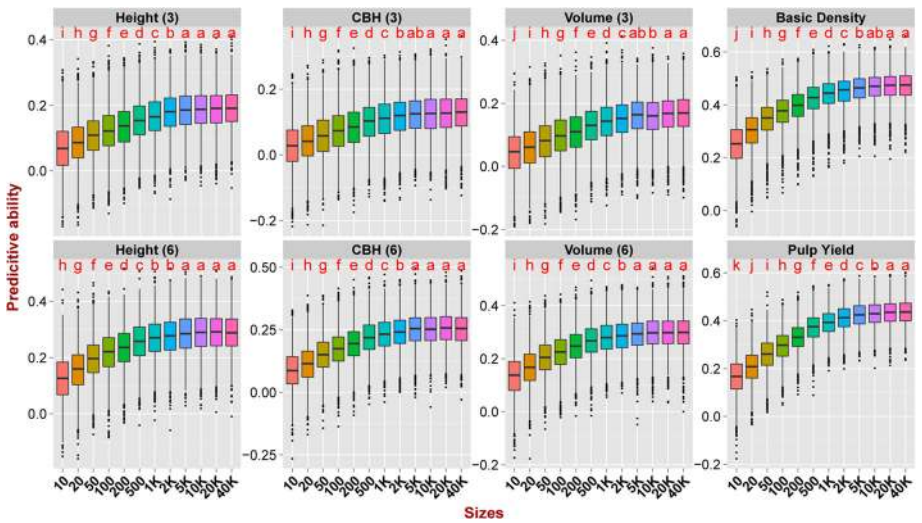

Fig. 5 Predictive abilities with increasing numbers of SNPs. Predictive ability estimated with GBLUP and RKHS with increasingly larger sets of SNP sampled at random from the total of 41,304 SNPs. Outliers are indicated by *black dots*. *Letters* indicate significant difference between the different models after Bonferroni adjustment ($P < 0.05$)

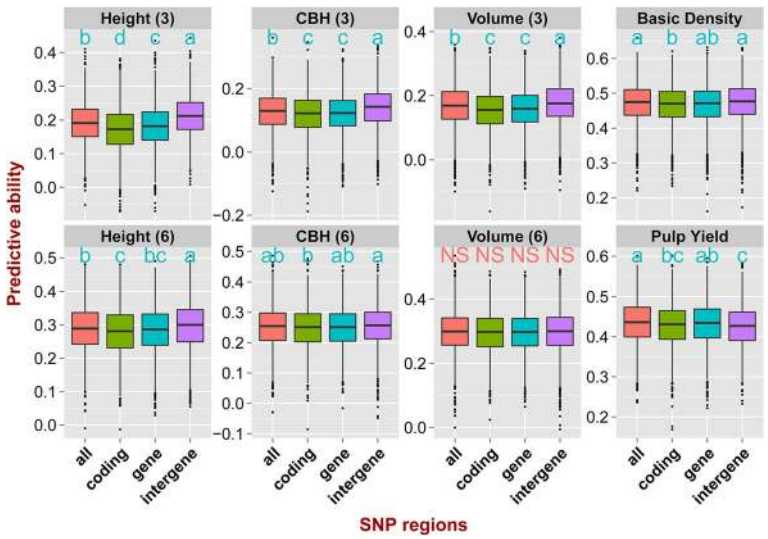

Fig. 6 Predictive abilities using SNPs located in different genomic regions. Predictive ability estimated with GBLUP and RKHS using 11,786 SNPs in coding DNA, 30,405 SNPs in genic regions (CDS, UTR, intron, and within 2 kb upstream and downstream of genes), 10,899 SNPs in intergenic regions and all 41,304 SNPs. *Letters* indicate significant difference between the different models after Bonferroni adjustment ($P < 0.05$)

impact predictions in cross-validations and future genomic selection candidates.

As expected, relatedness between TS and VS had a large impact on PAs for all traits. Prediction models built under scenario CV_3 (minimized relatedness between TS and VS) resulted in significantly worse predictions than in scenario CV_4 when relatedness was maximized. Our results are in line with previous reports in forest trees, such as white spruce [45, 46] and *Eucalyptus* [4], where models developed for one population had limited or no ability of predicting phenotypes in unrelated populations, suggesting that prediction models are largely population specific. With a lower relationship between TS and VS, the extent of LD is shorter and not stable across distantly related individuals in populations and the predictive ability of genomic prediction model is therefore reduced. Recent simulations show that the accuracy of genomic prediction models decline approximately linearly with increasing genetic distance between training and prediction populations [58]. Increased relatedness reduce the number of independently segregating chromosome segments and therefore increase the probability that chromosome segments that are IBD and which are sampled in the training population are also represented in the selection candidates. Our results provide additional experimental evidence that for successful implementation of GS the selection candidates have to show a close genetic relationship to the training population.

PAs were considerably higher when all the G0 parents were kept in the TS (scenario CV_2). This result could be due to two reasons. On one hand, by keeping all G0 parents in TS, we ensure that a large genetic diversity is available for model training, which could explain the positive impact of G0 inclusion on predictions. On the other hand, it is possible that by allocating all G0 individuals to the TS the positive effect we observe is strictly not due to increased predictive power but rather because we avoid the potentially negative impact of having pure species parents in the validation set in combination with G1 progeny that were largely F_1 hybrids. In order to evaluate this, we estimated PA of genomic prediction models by using GBLUP and RKHS, having only the 168 G0 parents for TS and randomly selected 168 G1 individuals in VS. To control for the effect of the strongly reduced TS size, we compared this setup with random assignment of individuals to TS or VS but keeping the size of each at $N = 168$. The results showed considerably lower PAs (even zero or negative) when using only pure species parents to predict G1 hybrid progeny phenotypes (Additional file 13). This observation, together with the fact that PAs for scenario CV_4 (maximum relatedness between TS and VS) were also generally lower than CV_2, suggest that the higher PAs we observe for scenario CV_2 is mostly due to avoiding the negative effect of having pure species parents in the VS.

The issue of genomic prediction in hybrid breeding has been investigated so far only within species and only for domestic animals, more specifically for bovine and pig breeding in which selection is carried out in pure breeds but with the aim to improve crossbred performance [43, 59]. Results from simulations show that training on crossbred data provides good PAs by selecting purebred individuals for crossbred performance,

although PAs drop with increasing distances between breeds [60]. When crossbred data is not available, separate purebred training populations can be used either separately or combined depending on the correlation of LD phase between the pure lines [61], which in turn is in part determined by the time of divergence between the populations. Compared to bovine breeds that belong to the same species and have diverged relatively recently (<300KYA) [62], the estimated divergence time between the two *Eucalyptus* species used in our study is much older, estimated at 2–5 MYA [63]. We therefore don't expect much correlation of LD phase between the two species and it is thus not surprising that training on the combined pure species sets with validation in F_1 hybrids resulted in poor PA. To the best of our knowledge, our results are the first ones to provide an initial look at the issue of genomic prediction from pure species to interspecific hybrids and our results indicate that, consistent with theoretical expectations, models have to be trained using hybrids if one is to predict phenotypes in hybrid selection candidates.

Number of SNPs is more important than SNP genomic location

Across all traits, no major improvement was detected in PA when more than 5000 SNPs were used (Additional file 10, Fig. 5), although a slight increase was observed for height of age three, basic density and pulp yield when using GBLUP based on 20,000 SNPs. Several studies have previously shown that considerably lower numbers of SNPs provided PAs equivalent to those observed using all SNPs available [24, 64]. The necessary number of SNPs needed for genomic prediction model depends on the extent of LD, which is strictly dependent on N_e. Our results, where we achieve equivalent PAs using either all or only 10–20% of the genotyped markers suggests that it represents a closed breeding population with a relatively modest N_e. This has been a common approach in domestic animals with the intent of developing low-density genotyping chips to reduce genotyping costs [8]. The main advantage of using reduced SNP panels is cost-effectiveness, although it is expected that using a higher density of markers will be necessary to mitigate the decay of PAs over generations due to the combined effect of recombination and selection on the patterns of LD [65]. It is also questionable whether it will be more cost effective to have targeted low-density SNP chips for specific populations or a full SNP chip that can be used across breeding populations of several organizations. By having a SNP chip that will accommodate several populations the cost-effectiveness and economy of scale of amassing many more samples to be genotyped with the same chip will likely be much larger than the cost reduction observed by using a smaller number of SNPs on each specific population.

SNP location also contributed to the predictive ability of genomic prediction model although the effects were rather modest. PAs using SNPs in intergenic regions were slightly better than using SNPs in genic regions or using all SNPs, except for pulp yield that could be somewhat better predicted with SNPs in coding and gene regions (Fig. 6). This likely represents a random sampling effect and not any specific enrichment for functional variants for this trait. However, the decline of LD was slower for SNPs in intergenic regions than for SNPs in genic and/or coding regions (Additional file 12) and the slightly longer range of LD might help explain why using SNPs in intergenic regions provided better PAs. With slower LD decay, SNPs in intergenic regions might better capture QTLs across longer genomic segments than SNPs in coding regions where LD decays more rapidly.

Further issues affecting the accuracy of model prediction

Several issues remain to be investigated for successful adoption of genomic prediction in operational eucalypt breeding. First, how does the accuracy of genomic prediction decline over successive generations of selection due to the effects of recombination? Simulation studies illustrated that the prediction accuracy decline rapidly during early generations but this decline slows down in later generations [6, 16]. A GS model should therefore be updated after the phenotypes of next generation individuals become available. Second, how stable are genomic prediction models across multiple environments and how important is it to consider genotype by environment interactions in the models? The interaction between genomic prediction and environmental effects will essentially follow conventional G x E strategies. Prediction models are expected to be accurate across sites within the same breeding zone (an area within which a single population of improved trees can be planted without fear of maladaptation) but not necessarily across different breeding zones [12]. Furthermore, with genomic prediction, individuals are not evaluated on the basis of their own phenotypic performance, but on the basis of genomic information across other individuals, years and environments, which given an opportunity to evaluate the effect of particular genomic segments that are shared between individuals across multiple environments. Burgueno et al. [66] showed that models incorporating pedigree and marker data on wheat lines from multiple environments can substantially enhance prediction accuracy relative to only pedigree-based prediction or relative to genomics prediction models derived from single environments. Finally, we have only considered the additive genetic variance for building genomic prediction models in our eucalypt population, but it is

possible, and perhaps even likely, that non-additive genetic effects play an important role in many breeding populations and specifically in populations consisting of early generation hybrids. A recent simulation study of genomic prediction in *Eucalyptus* breeding reported that genomic prediction including dominance effects performed better for clone selection where as non-additive effects did not improve the estimation of breeding value for parental selection [67]. To the best of our knowledge, no experimental data exist in forest trees regarding the ability of GS to predict the total genotypic value of individual trees, including both additive and non-additive effects.

Conclusions

Our experimental results provide further promising perspectives for the implementation of genomic prediction in *Eucalyptus* breeding programs. Genomic prediction largely outperformed pedigree-based prediction in our experiment, mainly due to the fact that our expected pedigree had major inconsistencies, resulting in gross underestimation of all pedigree-based estimates. This rather unexpected result illustrated an additional advantage of using SNP data and genomic prediction in breeding programs. While the main advantage of genomic prediction in eucalypt breeding will likely be the reduction of the breeding cycle length [4], the use of a genomic relationship matrix allowed us to obtain precise estimates of genetic relationship and heritabilities that we would otherwise not have had access to. Furthermore, our results corroborated the key role of relatedness as a driver of PA, the potential of using lower density SNP panels, and the fact that growth and wood traits adequately fit the infinitesimal model such that either GBLUP or rrBLUP represents a good compromise between computational time and prediction efficiency. In contrast to previous studies in *Eucalyptus*, we had accessed to both the pure species parents (*E. grandis* and *E. urophylla*) and their F_1 progeny. We show that models trained on pure species parents do not allow for accurate prediction in F_1 hybrids, likely due to the strong genetic divergence between the two species and lack of consistent patterns of LD between the two species and their hybrids.

Additional files

Additional file 1: Average accuracy of SNP imputation methods with increasing proportions of missing data. SNPs on chromosomes 6 and 8 were randomly removed from the dataset to generate specific missing data proportions. Accuracy between imputed and true SNP genotypes were subsequently calculated with the different methods. (DOCX 1714 kb)

Additional file 2: Predictive abilities on genomic selection model that comprises of statistical methods, genetic compositions and relative sizes of Training Set/Validation Set for each trait. (XLSX 16 kb)

Additional file 3: ANOVA analysis of sources of variation affecting the predictive ability. (DOCX 48 kb)

Additional file 4: Mean and standard deviation of predictive ability with the five prediction methods for the eight traits. (DOCX 96 kb)

Additional file 5: Mean and standard deviation of predictive ability estimated with the four Training Set/Validation Set compositions. (DOCX 84 kb)

Additional file 6: Mean and standard deviation of predictive ability estimated with the five relative sizes of Training Set/Validation Set expressed in proportions and numbers of individuals. (DOCX 89 kb)

Additional file 7: Mean and standard deviation of predictive ability across increasing numbers of SNPs, statistical methods (RKHS and GBLUP), four Training Set/Validation Set compositions for each of eight traits. (XLSX 61 kb)

Additional file 8: Mean and standard deviation of predictive ability estimated with SNPs in four genomic locations, with two statistical methods (RKHS and GBLUP), four Training Set/Validation Set compositions for each of eight traits. (XLSX 58 kb)

Additional file 9: ANOVA of predictive ability with SNP genomic location and SNP number as sources of variation. (DOCX 62 kb)

Additional file 10: Average predictive ability estimated with different numbers of SNPs fitted into the model. (DOCX 136 kb)

Additional file 11: Average predictive abilities estimated using SNP sets located in different genomic regions. (DOCX 82 kb)

Additional file 12: Decay of linkage disequilibrium (LD) with physical distance estimated with SNPs in different genomic locations. (a) A comparison of the decay of LD with physical distance in four classes of SNPs located with coding, genic, intergenic and all regions, respectively. Dots of pairwise LD versus physical distance and the LD decay for SNPs located in all regions (b), coding region (c), genic region (d) and intergenic region (e), respectively. (DOCX 1375 kb)

Additional file 13: Predictive abilities by training in pure species eucalypt parents and predicting in their F_1 hybrids. Predictive ability estimated under three training/validation sets (TS/VS) scenarios with two methods (GBLUP and RKHS) for each trait. PO168 (red boxes): all 168 *E. grandis* and *E. urophylla* pure species G0 parents used for training and 168 G1 random selected hybrid progeny for validation; random168 (green): randomly selected 168 individuals from all 1117 for TS and 168 randomly also for VS; random558 (blue): randomly divided all 1117 individuals into TS and VS of same size (558/558). Outlier estimates are indicated by black dots. (DOCX 174 kb)

Abbreviations

BL: Bayesian LASSO; CBH: Circumference at breast height; CDS: Coding sequences; GBLUP: Genomic best linear unbiased predictor; GEBV: Genomic estimated breeding values; GRM: Genomic relationship matrix; GS: Genomic selection; IBD: Identity by descent; IBS: Identity by state; LD: Linkage disequilibrium; MAS: Marker-assisted selection; N_e: Effective population size; PA: Predictive ability; PCA: Principal components analysis; QTLs: Quantitative trait loci; RKHS: Reproducing kernel Hilbert space; rrBLUP: Ridge-regression best linear unbiased prediction; SNP: Single-nucleotide polymorphism; TS: Training set; VS: Validation set

Acknowledgements

We would like to thank Michelle Bayerl Fernandes for her contribution on phenotyping the breeding population. The computations were performed on resources provided by the Swedish National Infrastructure for Computing (SNIC) at UPPMAX and HPC2N.

Funding

The study has partly been funded through grants from Vetenskapsrådet and the Kempestiftelserna to PKI. BT gratefully acknowledges financial support from the Umeå Plant Science Centre (UPSC) "The Research School of Forest Genetics, Biotechnology and Breeding".

Availability of data and materials
The data that support the findings of this study are available from Veracel but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are available from the authors upon reasonable request and with permission of Veracel.

Authors' contributions
BT, BS and PKI conceived and designed the experiment; GSM phenotyped data; GSM and KZF collected samples for genotyping; DG was responsible for genotyping; BT analysed the data under DG and PKI's guidance; BT drafted the first version of the manuscript and BT, DG, BS and PKI critically contributed to the final version of the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details
[1]Umeå Plant Science Centre, Department of Ecology and Environmental Science, Umeå University, Umeå SE-90187, Sweden. [2]Biomaterials Division, Stora Enso AB, Nacka SE-13104, Sweden. [3]EMBRAPA Genetic Resources and Biotechnology – EPqB, Brasilia, DF 70770-910, Brazil. [4]Universidade Católica de Brasília- SGAN, 916 modulo B, Brasilia, DF 70790-160, Brazil. [5]Veracel Celulose S.A., Eunápolis, BA 45.820-970, Brazil. [6]Present address: Department of Plant Biology, Uppsala BioCenter, Swedish University of Agricultural Sciences, Uppsala SE-75007, Sweden.

Received: 13 October 2016 Accepted: 15 June 2017
Published online: 29 June 2017

References
1. Rezende GDSP, Resende MDV, Assis TF. *Eucalyptus* breeding for clonal forestry. In: Fenning T, editor. Challenges and opportunities for the world's forests in the 21st century. Dordrecht: Springer Netherlands; 2014. p. 393–424.
2. Myburg AA, Potts BM, Marques CM, Kirst M, Gion JM, Grattapaglia D, Grima-Pettenati J. Eucalyptus. Genome Mapping and Molecular Breeding in Plants. Volume 7. Edited by: Kole CR. New York: Springer, Forest trees; 2007. pp. 115-160.
3. Bison O, Ramalho M, Rezende G, Aguiar A, De Resende M. Comparison between open pollinated progenies and hybrids performance in *Eucalyptus grandis* and *Eucalyptus urophylla*. Silvae Genet. 2006;55(4–5):192–6.
4. Resende MD, Resende MF Jr, Sansaloni CP, Petroli CD, Missiaggia AA, Aguiar AM, 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.
5. Goddard ME, Hayes BJ, Meuwissen THE. Using the genomic relationship matrix to predict the accuracy of genomic selection. J Anim Breed Genet. 2011;128(6):409–21.
6. Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157(4):1819–29.
7. Meuwissen T, Hayes B, Goddard M. Accelerating improvement of livestock with genomic selection. Annu Rev Anim Biosci. 2013;1:221–37.
8. Van Eenennaam AL, Weigel KA, Young AE, Cleveland MA, Dekkers JCM. Applied animal genomics: results from the field. Annu Rev Anim Biosci. 2014;2:105–39.
9. Spindel J, Begum H, Akdemir D, Virk P, Collard B, Redona E, et al. Genomic selection and association mapping in rice (*Oryza sativa*): effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. PLoS Genet. 2015;11(2):e1004982.
10. Windhausen VS, Atlin GN, Hickey JM, Crossa J, Jannink JL, Sorrells ME, 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.
11. Isik F. Genomic selection in forest tree breeding: the concept and an outlook to the future. New Forest. 2014;45(3):379–401.
12. Grattapaglia D. Breeding Forest Trees by Genomic Selection: Current Progress and the Way Forward. In: Genomics of Plant Genetic Resources: Volume 1 Managing, sequencing and mining genetic resources. Edited by Tuberosa R, Graner A, Frison E. Dordrecht: Springer Netherlands; 2014. pp. 651–82.
13. de los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, MPL C. Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics. 2013;193(2):327–45.
14. Endelman JB. Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome. 2011;4(3):250–5.
15. Silva FF E, Viana JM, Faria VR, de Resende MD. Bayesian inference of mixed models in quantitative genetics of crop species. Theor Appl Genet. 2013; 126(7):1749–61.
16. Habier D, Fernando RL, Dekkers JCM. The impact of genetic relationship information on genome-assisted breeding values. Genetics. 2007;177(4):2389–97.
17. De los Campos G, Gianola D, Rosa GJ, Weigel KA, Crossa J. Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. Genet Res. 2010;92(4):295–308.
18. Neves HH, Carvalheiro R, Queiroz SA. A comparison of statistical methods for genomic selection in a mice population. BMC Genet. 2012;13(1):100.
19. Hayes B, Daetwyler H, Bowman P, Moser G, Tier B, Crump R, Khatkar M, Raadsma H, Goddard M. Accuracy of genomic selection: comparing theory and results. In: Proceedings of the 18th Conference: Association for the Advancement of Animal Breeding and Genetics, Barossa Valley, Australia; 2009. pp. 34–37.
20. Wu X, Lund MS, Sun D, Zhang Q, Su G. Impact of relationships between test and training animals and among training animals on reliability of genomic prediction. J Anim Breed Genet. 2015;132(5):366–75.
21. Zhong S, Dekkers JC, Fernando RL, Jannink JL. Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a barley case study. Genetics. 2009;182(1):355–64.
22. Grattapaglia D, Resende MDV. Genomic selection in forest tree breeding. Tree Genet Genomes. 2011;7(2):241–55.
23. Moser G, Khatkar MS, Hayes BJ, Raadsma HW. Accuracy of direct genomic values in Holstein bulls and cows using subsets of SNP markers. Genet Sel Evol. 2010;42
24. Su G, Brondum RF, Ma P, Guldbrandtsen B, Aamand GR, Lund MS. Comparison of genomic predictions using medium-density (similar to 54,000) and high-density (similar to 777,000) single nucleotide polymorphism marker panels in Nordic Holstein and red Dairy cattle populations. J Dairy Sci. 2012;95(8):4657–65.
25. MacLeod IM, Hayes BJ, Goddard ME. The effects of demography and long-term selection on the accuracy of genomic prediction with sequence data. Genetics. 2014;198(4):1671–84.
26. Silva-Junior OB, Faria DA, Grattapaglia D. A flexible multi-species genome-wide 60K SNP chip developed from pooled resequencing of 240 *Eucalyptus* tree genomes across 12 species. New Phytol. 2015;206(4): 1527–40.
27. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559–75.
28. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet. 2007;81(5):1084–97.
29. Stephens M, Scheet P. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. Am J Hum Genet. 2005;76(3):449–62.
30. Candes EJ, Recht B. Exact matrix completion via convex optimization. Found Comput Math. 2009;9(6):717–72.
31. Rutkoski JE, Poland J, Jannink JL, Sorrells ME. Imputation of unordered markers and the impact on genomic selection accuracy. G3-Genes Genom Genet. 2013;3(3):427–39.
32. Stacklies W, Redestig H, Scholz M, Walther D, Selbig J. pcaMethods - a bioconductor package providing PCA methods for incomplete data. Bioinformatics. 2007;23(9):1164–7.

33. Remington DL, Thornsberry JM, Matsuoka Y, Wilson LM, Whitt SR, Doeblay J, et al. Structure of linkage disequilibrium and phenotypic associations in the maize genome. P Natl Acad Sci USA. 2001;98(20): 11479–84.
34. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet. 2006;2(12):2074–93.
35. Legarra A, Robert-Granie C, Croiseau P, Guillaume F, Fritz S. Improved Lasso for genomic selection. Genet Res. 2011;93(1):77–87.
36. Crossa J, Campos Gde L, Perez P, Gianola D, Burgueno J, Araus JL, et al. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics. 2010;186(2):713–24.
37. Gilmour AR, Gogel B, Cullis B, Thompson R, Butler D. ASReml user guide release 3.0. UK https://www.vsni.co.uk/: VSN International Ltd, Hemel Hempstead; 2009.
38. Perez 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. Plant Genome. 2010;3(2):106–16.
39. los Campos G, Pérez P, Vazquez AI, Crossa J. Genome-enabled prediction using the BLR (Bayesian linear regression) R-package. In: Genome-wide association studies and genomic prediction. Edited by Gondro C, van der Werf J, Hayes B. Totowa, NJ: Humana Press; 2013: 299-320.
40. Perez P. De los Campos G. Genome-wide regression and prediction with the BGLR statistical package. Genetics. 2014;198(2):483–95.
41. de Los CG, Sorensen D, Gianola D. Genomic heritability: what is it? PLoS Genet. 2015;11(5):e1005048.
42. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms. SnpEff Fly. 2012;6(2):80–92.
43. Hidalgo AM, Bastiaansen JWM, Lopes MS, Harlizius B, Groenen MAM, de Koning DJ. Accuracy of predicted genomic breeding values in purebred and crossbred pigs. G3-Genes Genom Genet. 2015;5(8):1575–83.
44. Resende MF Jr, Munoz P, Resende MD, Garrick DJ, Fernando RL, Davis JM, et al. Accuracy of genomic selection methods in a standard data set of loblolly pine (*Pinus taeda* L.). Genetics. 2012;190(4):1503–10.
45. Beaulieu J, Doerksen 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;113(4):343–52.
46. Beaulieu J, Doerksen TK, MacKay J, Rainville A, Bousquet J. Genomic selection accuracies within and between environments and small breeding groups in white spruce. BMC Genomics. 2014;15:1048.
47. El-Dien OG, Ratcliffe B, Klapste J, Chen C, Porth I, El-Kassaby YA. Prediction accuracies for growth and wood attributes of interior spruce in space using genotyping-by-sequencing. BMC Genomics. 2015;16:370.
48. Ratcliffe B, El-Dien OG, Klapste J, Porth I, Chen C, Jaquish B, et al. A comparison of genomic selection models across time in interior spruce (*Picea engelmannii* x *glauca*) using unordered SNP imputation methods. Heredity. 2015;115(6):547–55.
49. Isik F, Bartholome J, Farjat A, Chancerel E, Raffin A, Sanchez L, et al. Genomic selection in maritime pine. Plant Sci. 2016;242:108–19.
50. Crossa J, Perez P, Hickey J, Burgueno J, Ornella L, Ceron-Rojas J, et al. Genomic prediction in CIMMYT maize and wheat breeding programs. Heredity. 2014;112(1):48–60.
51. Onogi A, Ideta O, Inoshita Y, Ebana K, Yoshioka T, Yamasaki M, et al. Exploring the areas of applicability of whole-genome prediction methods for Asian rice (*Oryza sativa* L.). Theor Appl Genet. 2015;128(1):41–53.
52. Clark SA, Hickey JM, van der Werf JHJ. Different models of genetic variation and their effect on genomic evaluation. Genet Sel Evol. 2011;43(1):1–9.
53. Honarvar M, Rostami M. Accuracy of genomic prediction using RR-BLUP and Bayesian LASSO. Eur J Exp Biol. 2013;3:42–7.
54. Daetwyler HD, Pong-Wong R, Villanueva B, Woolliams JA. The impact of genetic architecture on genome-wide evaluation methods. Genetics. 2010;185(3):1021–31.
55. Lorenz AJ, Chao S, Asoro FG, Heffner EL, Hayashi T, Iwata H, et al. Genomic selection in plant breeding: knowledge and prospects. Adv Agron. 2011;110
56. Lorenz AJ. Resource allocation for maximizing prediction accuracy and genetic gain of genomic selection in plant breeding: a simulation experiment. G3-Genes Genom Genet. 2013;3(3):481–91.
57. Riedelsheimer C, Endelman JB, Stange M, Sorrells ME, Jannink JL, Melchinger AE. Genomic predictability of interconnected biparental maize populations. Genetics. 2013;194(2):493–503.
58. Scutari M, Mackay I, Balding D. Using genetic distance to infer the accuracy of genomic prediction. PLoS Genet. 2016;12(9):e1006288.
59. Esfandyari H, Bijma P, Henryon M, Christensen OF, Sørensen AC. Genomic prediction of crossbred performance based on purebred landrace and Yorkshire data using a dominance model. Genet Sel Evol. 2016;48(1):1–9.
60. Ibánēz-Escriche N, Fernando RL, Toosi A, Dekkers JC. Genomic selection of purebreds for crossbred performance. Genet Sel Evol. 2009;41(1):1–10.
61. Esfandyari H, Sørensen AC, Bijma P. Maximizing crossbred performance through purebred genomic selection. Genet Sel Evol. 2015;47(1):1–16.
62. Murray C, Huerta-Sanchez E, Casey F, Bradley DG. Cattle demographic history modelled from autosomal sequence variation. Philos T R Soc B. 2010; 365(1552):2531–9.
63. Silva-Junior OB, Grattapaglia D. Genome-wide patterns of recombination, linkage disequilibrium and nucleotide diversity from pooled resequencing and single nucleotide polymorphism genotyping unlock the evolutionary history of *Eucalyptus grandis*. New Phytol. 2015;208(3):830–45.
64. Zhang Z, Ding X, Liu J, Zhang Q, de Koning DJ. Accuracy of genomic prediction using low-density marker panels. J Dairy Sci. 2011;94(7):3642–50.
65. Solberg TR, Sonesson AK, Woolliams JA, Meuwissen THE. Genomic selection using different marker types and densities. J Anim Sci. 2008;86(10):2447–54.
66. Burgueno J, de los Campos G, Weigel K, Crossa J. Genomic prediction of breeding values when modeling genotype x environment interaction using pedigree and dense molecular markers. Crop Sci. 2012;52(2):707–19.
67. Denis M, Bouvet J-M. Efficiency of genomic selection with models including dominance effect in the context of Eucalyptus breeding. Tree Genet Genomes. 2012;9(1):37–51.

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