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

Biyue Tan, Umeå Plant Science Centre, Department of Ecology and Environmental Science, Umeå University, SE-90187, Umeå, Sweden; Biomaterials Division, Stora Enso AB, SE-13104, Nacka, Sweden. biyue.tan@umu.se

Dario Grattapaglia, EMBRAPA Genetic Resources and Biotechnology – EPqB, 70770-910, Brasilia, DF, Brazil; Universidade Católica de Brasília- SGAN, 916 modulo B, Brasilia, DF, 70790-160, Brazil. dario.grattapaglia@embrapa.br

Gustavo Salgado Martins, Veracel Celulose S.A., 45.820-970, Eunápolis, BA, Brazil. gustavosalgadomartins@gmail.com

Karina Zamprogno Ferreira, Veracel Celulose S.A., 45.820-970, Eunápolis, BA, Brazil. Karina.Zamprogno@veracel.com.br

Björn Sundberg, Biomaterials Division, Stora Enso AB, SE-13104, Nacka, Sweden. Bjorn.Sundberg@storaenso.com

Pär K. Ingvarsson, Umeå Plant Science Centre, Department of Ecology and Environmental Science, Umeå University, SE-90187, Umeå, Sweden. par.ingvarsson@umu.se.
corresponding author.

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 used 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 thee and six years. Based on different genomic prediction methods we assessed the impact of the composition and size of the training/validation sets 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, GRM more precisely capture Mendelian sampling among related individuals, such that the genetic covariance was based on the actual 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 GBLUP or rrBLUP represents better compromises between computational time and prediction efficiency. Only slight improvement was observed in PA when more than 5,000 SNPs were used for all traits. Using SNPs in intergenic regions provided slightly better PA than using SNPs sampled exclusively in genic regions.

Conclusions: Effects of training set size and composition and number of SNPs used are the most important factors for model prediction rather than prediction method and the genomic location of SNPs. Furthermore, training the prediction model on pure parental species 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.

Keywords

Genomic relationship, genomic heritability, two-generation, genome annotation, high-density SNP-chip, Bayesian LASSO, GBLUP, rrBLUP

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, 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* and the disease tolerance and wide adaptability of *E.urophylla* [3]. A conventional breeding cycle toward clonal selection in hybrid populations involves mating, progeny trial, 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 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 had previously been detected, usually in underpowered experiments and thus leaving most of the variation unaccounted for [7]. GS are generally performed in three steps: (1) genotyping and phenotyping a ‘reference’ or ‘training population’ and developing 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 suggest that has great potential also in forest trees [11, 12].

The accuracy of genomic prediction models can vary depending on the statistical method 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 vary in the assumptions of the distribution and variances of marker effects. rrBLUP assumes that marker effects follow a normal distribution where all effects are shrunk to a similar and small size, while Bayesian methods (BayesA, BayesB, Bayesian LASSO and BayesR) assume that genetic variances specific to the marker effects and including a priori data on the probability distributions of marker effects. 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 [14]. The RKHS regression model is a linear combination of the basic function provided by the reproducing kernel [15]. Recent studies have indicated that the selection of suitable statistical methods depends on the actual data at hand and the pattern of phenotypic variation in the traits of interest and with reference population used [9, 16].

Besides 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. [17] found that for a given effective population size

(N_e), increasing the size of the reference population leads to improved accuracy of GS based predictions. Closer relationship between training population and selection candidates has been reported to lead to a higher accuracy of genomic predictions, while enlarge genetic diversity of the training population resulted in lower accuracy [18]. A number of simulation and empirical studies have shown that increasing the number of markers may improve the predictive accuracy as the N_e also increases [9, 19-21]. However, increasing the number of markers in small N_e populations has little or no improvement on predictive accuracy [22, 23].

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. A multi-species single-nucleotide polymorphism (SNP) chip containing 60,904 SNPs [24] 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 training set/validation set composition and size on the predictive ability (PA) of genomic prediction.

Methods

Breeding population

The breeding population in this study was established by 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

in a randomized complete block design with single-tree plots and 35 reps per family in Belmonte (Brazil, 39.19W, 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 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 the 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 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

147 site. **X**, **Z** and **W** are incidence matrices for $\boldsymbol{\beta}$, $\boldsymbol{u}$ and $\boldsymbol{b}$, respectively. The phenotypes of each
148 trait were then corrected by subtracting variation of sites, ages and blocks effects for all
149 individuals, and are referred to as adjusted phenotypes. The adjusted phenotypic traits were
150 used for calculating the heritability of traits and for building genomic prediction models.

151 **Genotyping and quality control**

152 The 168 G0 and 958 G1 populations were genotyped using the Illumina Infinium
153 EuCHIP60K [24] that contains probes for 60,904 SNPs. EUChip60K intensity data (.idat files)
154 were obtained through GENESEEK (Lincoln, NE, USA). SNP genotypes were called using
155 GenomeStudio (Illumina Inc., San Diego, CA, USA) following standard genotyping and
156 quality control procedures with no manual editing of clusters as described earlier [24].
157 Further quality control of the genotyped samples was performed using PLINK [25]. Nine G1
158 individuals were removed due to low sample call rate (<70%) or high inbreeding coefficient
159 (F>1). 10,240 SNPs were excluded due to low call rate (<70%), 9,243 SNPs were filtered out
160 due to monomorphism or minor allele frequency (MAF) < 0.01, and 117 SNPs were removed
161 due strong deviations from Hardy-Weinberg equilibrium (p-value < 1×10^{-6}).

162 After quality control, missing genotypes of the remaining individuals were filled in by
163 imputation. We first tested the accuracy of imputation methods across a range of missing data
164 (2% - 30%) by artificial removing SNPs from a fraction of our genotypes. Among the
165 available family-based and population based methods we assessed the following programs for
166 imputation accuracy: BEAGLE [26], fastPHASE [27], MENDEL [28], random forest, SVD
167 Impute, k-nearest neighbors [29], BLUP A matrix, Bayesian PCA, NIPALS, Probabilistic
168 PCA [30]. BEAGLE provided the best accuracy for all missing data percentages, with
169 accuracies exceeding 95% in all cases (Additional file 1). We therefore used BEAGLE to
170 impute missing genotypes at the retained 41,304 SNPs across the 168 G0 and 949 G1
171 individuals. The imputed genotypic data was subsequently used in all genomic prediction

172 analyses. LD between SNP pairs was measured using the squared correlation coefficient (r^2)
173 for SNPs located on the same chromosome. The decay of LD versus physical distance was
174 then modelled using the nonlinear regression method described in Remington et al. [31].
175 We further studied the population structure and pairwise genomic relationship among the
176 1117 individuals by performing principal components analysis (PCA) [32] and kinship
177 analysis [33] using 10,213 independent SNPs (LD-pruned) ($r^2 < 0.2$) calculated in PLINK
178 [25]. Pedigree-based genetic relationship was estimated from ABLUP (see below for further
179 information).

180 **Statistical methods for genomic prediction**

181 Four statistical methods were assessed to estimate the parameters in equation (1) and for
182 predicting GEBVs, including genomic best linear unbiased predictor (GBLUP) [5], ridge
183 regression BLUP (rrBLUP) [6], Bayesian LASSO (BL) [34], and reproducing kernel Hilbert
184 space (RKHS) regression [15]. The performance of the four genomic prediction methods was
185 compared with that of the commonly used pedigree-based BLUP (ABLUP) [35].
186 The GEBVs were estimated using the following mixed linear model:

187 $$\boldsymbol{y} = \mathbf{1}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{a} + \boldsymbol{e} \qquad (1)$$

188 where $\boldsymbol{y}$ is the vector of adjusted phenotypes of single trait, $\boldsymbol{\beta}$ is the vector of overall mean
189 fitted as a fixed effect, $\boldsymbol{a}$ is the vector of random effects, and $\boldsymbol{e}$ is the vector of random
190 residual effects. **1** and **Z** are incident matrix of $\boldsymbol{\beta}$ and $\boldsymbol{a}$, respectively.
191 **ABLUP.** ABLUP is the standard method for predicting breeding values using the expected
192 relatedness among individuals based on pedigree information [35]. For ABLUP, the vector of
193 random additive effects ($\boldsymbol{a}$) in the equation (1) is assumed to follow a normal distribution,
194 $\boldsymbol{a} \sim N(\mathbf{0}, \boldsymbol{A}\sigma_a^2)$, where **A** is the additive numerator relationship matrix estimated from
195 pedigree information and the σ_a^2 is the additive genetic variance. The residual vector $\boldsymbol{e}$ is

assumed as $\boldsymbol{e} \sim N(\mathbf{0}, \boldsymbol{I}\sigma_e^2)$, where $\boldsymbol{I}$ is the identity matrix. Under these assumptions, equation (1) can be re-written as:

$$\begin{bmatrix} \boldsymbol{X^T X} & \boldsymbol{X^T Z} \\ \boldsymbol{Z^T X} & \boldsymbol{Z^T Z} + \boldsymbol{A}^{-1}\frac{\sigma_e^2}{\sigma_a^2} \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{a}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X^T y} \\ \boldsymbol{Z^T y} \end{bmatrix} \quad (2)$$

where $\frac{\sigma_e^2}{\sigma_a^2}$ is estimated using a restricted maximum likelihood method. The estimated breeding values ($\widehat{\boldsymbol{a}}$) can be calculated directly from equation (2). ABLUP calculations were performed using ASReml 3.0 [36].

GBLUP. The GBLUP method is derived from ABLUP, but differs in that the matrix **A** in equation (2) is replaced with the genomic relationship matrix (**G**) that is calculated from the genotypic data as $\mathbf{G} = \frac{(\mathbf{M}-\mathbf{P})(\mathbf{M}-\mathbf{P})^{\mathrm{T}}}{2\sum_{j=1}^{p} p_j(1-p_j)}$, where **M** is the matrix of samples and their corresponding SNPs denoted as 0, 1, 2, **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 equation (1) is assumed to follow $\boldsymbol{a} \sim N(\mathbf{0}, \mathbf{G}\sigma_g^2)$, where σ_g^2 is the genetic variance and GEBVs 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 [36].

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

$$\begin{bmatrix} \boldsymbol{X^T X} & \boldsymbol{X^T Z} \\ \boldsymbol{Z^T X} & \boldsymbol{Z^T Z} + \boldsymbol{I}\lambda \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{a}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X^T y} \\ \boldsymbol{Z^T y} \end{bmatrix} \quad (3)$$

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 $\hat{g}_i = \mathbf{Z}_i^{\mathbf{T}}\hat{\mathbf{a}}$ from equation (3), where $\mathbf{Z}_i^{\mathbf{T}}$ is the SNP vector for individual *i* and $\hat{\mathbf{a}}$ is the vector of estimated SNP effects. All calculations were performed using the rrBLUP package in the R environment [33].

Bayesian LASSO. The Bayesian LASSO (BL) method is the Bayesian treatment of LASSO regression 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 $\boldsymbol{a} \sim N(\mathbf{0}, \mathbf{T}\sigma_m^2)$, where $\mathbf{T} = \mathrm{diag}(\tau_1^2, \ldots, \tau_p^2)$. τ_j^2 is assigned as $\tau_j^2 \sim Exp(\lambda^2)$, *j*=1,…,*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 package in R [37]. Here a Monte Carlo Markov Chains sampler was applied and prior parameters ($df_e, S_e, r, \delta,$ and λ^2) were defined following the guidelines proposed by de los Campos *et al.* [38]. The chain length was 20,000 iterations, with the first 2,000 excluded as burn-in and with a subsequent thinning interval of 100.

RKHS. RKHS assumes that the random additive effects in equation (1) are $\boldsymbol{a} \sim N(\mathbf{0}, \mathbf{K}\sigma_g^2)$, where **K** is computed by means of a Gaussian kernel that is given by $\mathrm{K}_{ij} = exp(-hd_{ij})$ [15]. *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 $d_{ij} = \sum_{k=1}^{p}(x_{ik} - x_{jk})^2$, where x_{ik} and x_{jk} are *k*th SNPs (*k*=1,…,*p*) for the *i*th and *j*th samples, respectively. We implemented the RKHS method through the BGLR package in R [39], which uses a Gibbs sampler for the Bayesian framework and assigns the prior distribution of σ_g^2 and σ_e^2 as $\sigma_g^2 \sim \chi^{-2}(df_g, S_g)$ and $\sigma_e^2 \sim \chi^{-2}(df_e, S_e)$, respectively. Here we chose a multi-kernel model suggested by Perez [39], where three h were defined as $h_1 = 2/(5 * \bar{d})$,

$h_2 = 2/\bar{d}$, $h_3 = 2 * 5/\bar{d}$, $\bar{d}$ 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 [40]. The respective heritabilities were calculated as:

$$h_a^2 = \frac{\sigma_a^2}{\sigma_a^2+\sigma_{ea}^2} \qquad h_g^2 = \frac{\sigma_g^2}{\sigma_g^2+\sigma_{eg}^2}$$

where σ_a^2 is the additive genetic variance and σ_{ea}^2 is the residual variance estimated with ABLUP, while σ_g^2 is the genetic variance and σ_{eg}^2 is the residual variance estimated with GBLUP.

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 TS/VS sizes, we divided all 1,117 (G0 and G1) individuals into five different size groups with a ratio of TS to VS of 1:1, 2:1, 3:1, 4:1 and 9:1. The corresponding sizes of the TS/VS were respectively 558/559, 743/374, 836/281, 892/225 and 1003/114. Within these pre-established size compositions, four scenarios of 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 G1 individuals up to the required number in the set, while the VS was composed exclusively of G1 individuals. The third and fourth 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, 1,117 individuals were assigned to 182 clusters based on genotypes using the k-means clustering algorithm, a method that 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 [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 on 80 out of 100 models tested (20 ABLUP models excluded) to partition the variance into different sources, with all effects declared as fixed, comparing all the sources of variation (genomic prediction method, TS/VS size and genetic composition). Significant differences 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, 1,000, 2,000, 5,000, 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 2kb up and downstream of the gene (30,405 SNPs); (iii) intergenic regions (10,899 SNPs), and (iv) all 41,304 SNPs. The location of each SNP was obtained by mapping SNPs onto *E.grandis* genome database using SnpEff [41]. Genomic prediction models were built for all four TS/VS compositions using only the two statistical methods (GBLUP and RKHS) that showed optimal predictive performance in the previous analyses, and the TS/VS size ratio of 4:1 (892/224) were used on the PA evaluations.

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 (Figure 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 weakly (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 (Figure 2a) and clearly separated the G0 individuals of the two species,

315 *E.grandis* and *E.urophylla*, with the *E.grandis* individuals further subdivided into two
316 subgroups likely representing the two main provenances used in breeding programs in Brazil.
317 The G1 individuals were generally projected into the space defined by their parents, but with
318 a few outliers. The expected pedigree-based and realized genomic-based genomic
319 relationships among G0 and G1 individuals were visualized in heatmaps (blue and red in
320 Figure 2b, respectively). The result of the genomic relationship analysis corroborated the
321 PCA result, in which *E. urophylla* was clustered into a single group, whereas *E. grandis*
322 formed two subgroups. The average values of the realized genomic relationships among what
323 were considered to be full-sibs, half-sibs and unrelated individuals from the pedigree data
324 were generally lower than the expected relationships values (0.309 vs. 0.5, 0.131 vs. 0.25
325 and .0056 vs. 0, respectively) (Table 1). This result suggests that pedigree errors were likely
326 present in this population. These putative pedigree errors in turn negatively affected the
327 pedigree-based trait heritability, which were considerably lower than those estimated using
328 genomic-based realized genomic relationships (Table 2).

329 **Table 1.** Pairwise expected pedigree-based and realized genomic-based relationships in the
330 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

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

332 **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^2[b]	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)

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

336

337 **Predictive abilities with different statistical methods**

338 Estimates of PAs were obtained using different statistical methods, compositions and sizes
339 of TS/VS for each trait (Additional file 2). An ANOVA showed that all these factors had a
340 significant effect on the PA (P-value < 0.005) (Additional file 3). Across the four genomic
341 prediction methods used (GBLUP, rrBLUP, BL, and RKHS) the average PA varied from
342 0.27 to 0.274 (Additional file 4). All the four methods outperformed the pedigree-based
343 ABLUP prediction (mean PA = 0.121) by an average of 80%-200% across the eight traits
344 (Figure 3). RKHS yielded a slightly better PAs for six out of eight traits and this method was
345 particularly suitable for predicting traits that displayed a lower heritability such as CBH and
346 height. The other three methods generally gave similar results across all traits, although with
347 a slightly better performance than RKHS for pulp yield (Figure 3).

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

349 The average PAs differed significantly for the different TS/VS composition tested varying
350 from 0.253 to 0.286 (Additional file 5). The genomic prediction model built with CV_2 (all G0
351 parents in the TS) showed the highest PAs for all traits except pulp yield, whereas models
352 based on CV_3 (minimum relatedness between TS and VS) gave the worst predictions. The
353 models based on CV_1 (random assignment) and CV_4 (maximum relatedness between TS and
354 VS) showed no significant differences in PA (Figure 4, Additional file 5). The average PA
355 was significantly improved from 0.251 to 0.285, as the TS/VS ratio increased from 1:1
356 (558/559) to 9:1 (1003/113) (Additional file 6), irrespective of the prediction method (Figure
357 3) or the genetic composition of TS/VS used (Figure 4), clearly showing the importance of an
358 adequate size of the training set used to build prediction models. Furthermore, there was a
359 steeper increase in PA when TS/VS ratio increased from 1:1 (558/559) to 2:1 (743/374) than
360 from 2:1 (743/374) to 9:1 (1003/114) for all traits (Figure 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 sets of SNPs in different genomic locations (Additional file 8) were obtained with two prediction methods for all the different TS/VS compositions. An ANOVA showed that both the number of genotyped 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 5,000 (Additional file 10). On the other hand, no significant improvement was generally seen in the average of PA when more than 5,000 SNPs were used (Additional file 10, Figure 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 (Figure 5). For instance, PA for basic density ($h^2 = 0.35$) 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, few and only slight significant differences were seen in PAs by using SNP sets located in different genomic regions (Figure 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 (Figure 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 to any specific effect of genomic location. When we assessed the pairwise LD (r^2) amongst the SNPs in the four regions tested, the

385 extent of LD differed among them, with LD showing the most rapid decay in coding regions
386 and the slowest one in intergenic regions (Additional file 12).

387 ## Discussion

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

396 **Genomic data corrected pedigree inconsistencies**

397 All four genomic prediction methods performed significantly better than the pedigree-
398 based evaluations for all complex traits assessed (Figure 3). While similar results have been
399 reported for animals [16, 42] and crop species [9, 35] across a number of traits, in forest trees
400 prediction accuracies using genomic data have generally been similar or up to 10-30% lower
401 than accuracies obtained using pedigree-estimated breeding values, including *Eucalyptus* [4],
402 loblolly pine (*Pinus taeda*) [43], white spruce (*Picea glauca*) [44, 45], interior spruce (*Picea*
403 *engelmannii* × *glauca*) [46, 47] and maritime pine (*Pinus pinaster*) [48]. Genomic predictions
404 with lower accuracies than pedigree-based predictions could arise from insufficient marker
405 density, such that not all casual variants are captured in the genomic estimate [40], or an
406 overestimate of the pedigree-based prediction due to its inability of ascertaining the true
407 genetic relationships in half-sib families [46]. Our result however differ from previous studies
408 in forest trees due to the fact that the average pairwise estimates of genetic relationship
409 among individuals were substantially lower using SNP data than expectations based on

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

419 **Genomic predictions show that traits adequately fit the infinitesimal model**

420 Overall, the different genomic prediction methods provided similar results for the traits
421 evaluated with only a slight advantage for RKHS showing better PAs for growth traits that
422 had lower heritability (Figure 3) although for pulp yield, RKHS instead was the worst
423 performing method. It is possible that the definition of a kernel simply was not suitable for
424 this particular trait [15]. Our results corroborate previous reports both in crops and animals
425 [16, 49, 50], as well as in forest tree studies. In loblolly pine, for example, the performance of
426 rrBLUP and three Bayesian methods was only marginally different when compared across 17
427 traits with distinct heritabilities, with a small improvement using BayesA only for fusiform
428 rust resistance where loci of relatively larger effect have been described [43]. Similar results
429 were obtained for growth and wood traits in other forest trees studies showing no
430 performance difference between rrBLUP and Bayesian methods [45, 47, 48]. This occurs
431 despite simulation studies suggesting that Bayesian methods, like BL, should outperform
432 univariate methods such as rrBLUP and GBLUP [6, 51, 52]. One possible reason for the
433 apparent disagreement between simulations and empirical data sets is that the true QTL
434 effects for most of traits are relatively small and the distribution is less extreme than

435 simulated data [53]. Our results therefore support the proposal that either rrBLUP or GBLUP
436 are effective methods in providing the best compromise between computation time and
437 prediction efficiency [54] and that the quantitative traits assessed in our study adequately fit
438 the assumption of the infinitesimal model.

439 **Training set size, composition and relatedness strongly affect predictive ability**

440 Our results show that the size and the variable TS/VS compositions in terms of relatedness
441 between training and validation sets had the largest impact on the PA irrespective of the
442 analytical method used (Figure 4). The average PA rapidly increased with increasing sizes of
443 the TS and did not show any sign of plateauing. Earlier simulations of *Eucalyptus* breeding
444 scenarios had in fact shown that with up to N= 1,000 individuals in the TS, the accuracy
445 would rapidly increase, and additional gains would be seen up to N= 2,000 individuals for
446 lower heritability traits, larger numbers of QTLs involved and larger effective population size
447 (N_e). After N= 2,000 the predictive accuracy would tend to plateau irrespective of the N_e and
448 genotyping density [20]. Later simulations mirroring a eucalypt breeding scheme also
449 showed a considerable improvement of genomic predictions with increasing training
450 population sizes by consolidating phenotypic and genotypic data of individuals from previous
451 breeding cycles [55]. Simulations [19, 56] and proof-of-concept studies [57] in crop species
452 also show improved PA with larger TS sizes. Larger training populations alleviate the
453 probability of losing rare favourable alleles from the breeding population as generations of
454 selection advance. Additionally by sampling more individuals for training, a larger diversity
455 is captured and better estimates of the marker effects are obtained which in turn positively
456 impact predictions in cross-validations and future genomic selection candidates.

457 As expected, relatedness between TS and VS had a large impact on PAs for all traits.
458 Prediction models built under scenario CV_3 (minimized relatedness between TS and VS)
459 resulted in significantly worse predictions than in scenario CV_4 when relatedness was

460 maximized. Increasing the genetic relationships between training and selection candidates
461 effectively has the same consequence as reducing the N_e such that the stronger the
462 relationship, the higher in the predictive accuracy. Our results are in line with previous
463 reports in forest trees showing that models developed for one population had limited or no
464 ability of predicting phenotypes in an unrelated one in white spruce [44, 45] and *Eucalyptus*
465 [4], indicating that prediction models will be population specific. With lower relationship
466 between TS and VS, the extent of LD is shorter and not stable across distantly related
467 populations and the predictive ability of genomic prediction model is reduced. Recent
468 simulations show that the accuracy of genomic prediction models decline approximately
469 linearly with increasing genetic distance between training and prediction populations [58].
470 Increased relatedness reduce the number of independently segregating chromosome segments
471 and therefore increase the probability that chromosome segments IBD sampled in the training
472 population are also found in the selection candidates. Our results provide additional
473 experimental evidence that for successful implementation of GS the selection candidates have
474 to show a close genetic relationship to the training population.

475 PAs were considerably higher when all the G0 parents were kept in the TS (scenario CV_2).
476 This result could be due to two reasons. On one hand, by keeping all G0 parents in TS, we
477 had a large diversity available for training, which could explain the positive impact of G0
478 inclusion on predictions. On the other hand, it is possible that by allocating all G0 individuals
479 to the TS the positive effect we observe could strictly not be due to increased predictive
480 power of including G0 individuals but rather a way to avoid the potentially negative impact
481 of having pure species parents in the validation set in combination with G1 progeny that were
482 largely F_1 hybrids. In order to evaluate this, we estimated PA of genomic prediction models
483 by using GBLUP and RKHS, having only the 168 G0 parents for TS and randomly selected
484 168 G1 individuals in VS. To control for the effect of the strongly reduced TS size, we

485 compared this setup with random assignment of individuals to TS or VS but keeping the size
486 of each at N=168. The results showed considerably lower PAs (even zero or negative) when
487 using only pure species parents to predict G1 hybrid progeny phenotypes (Additional file 13).
488 This observation, together with the fact that PAs with scenario CV_4 (maximum relatedness
489 between TS and VS) were also generally lower than CV_2, suggesting that the higher PAs we
490 observe for scenario CV_2 is mostly due to avoiding the negative effect of having pure species
491 parents in the VS.

492 The issue of genomic prediction in hybrid breeding has been investigated so far only within
493 species and only for domestic animals, more specifically for bovine and pig breeding in
494 which selection is carried out in pure breeds but the aim is to improve crossbred performance
495 [42, 59]. Results from simulations show that training on crossbred data provides good PAs by
496 selecting purebred individuals for crossbred performance, although PAs drop with increasing
497 distances between breeds [60]. When crossbred data is not available, separate purebred
498 training populations can be used either separately or combined depending on the correlation
499 of LD phase between the pure lines [61], which in turn is in part determined by the time of
500 divergence between the populations. Compared to bovine breeds that belong to the same
501 species and have diverged relatively recently (<300KYA) [62], the estimated divergence time
502 between the two *Eucalyptus* species used in our study is much older, estimated at 2-5 MYA
503 [63]. Therefore, no correlation of LD phase between these two species is expected and it is
504 not surprising that training on the combined pure species sets to validate on the F_1 hybrids
505 resulted in poor PA. To the best of our knowledge, our results are the first ones to provide an
506 initial look at the issue of genomic prediction from pure species to interspecific hybrids
507 indicating that, consistent with expectations, models have to be trained in hybrids if one is to
508 predict phenotypes in hybrid selection candidates.

509 **Number of SNPs is more important than SNP genomic location**

Across all traits, no major improvement was detected in PA when more than 5,000 SNPs were used (Additional file 10, Figure 5), although a slight increase were observed for height of age three, basic density and pulp yield when using GBLUP based on 20,000 SNPs. Several studies have also shown that considerably lower numbers of SNPs provided PAs equivalent to those observed using all SNPs available [22, 64]. The necessary number of SNPs needed for genomic prediction model depends on the extent of LD, which strictly related to N_e. Our results, where we achieve equivalent PAs using either all of 12-20% of the genotyped markers suggests that it represents a closed breeding population with a relatively limited 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 (Figure 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 when

compared to SNPs in gene and 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.

Conclusions

Our experimental results provide further promising perspectives for the implementation of genomic prediction in *Eucalyptus* breeding programs. Genomic predictions largely outperformed the pedigree-based ones in our experiment, mainly due to the fact that our expected pedigree had major inconsistencies, such that all pedigree-based estimates were grossly underestimated. This 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 heritability 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 GBLUP or rrBLUP represent a good compromise between computation time and prediction efficiency. In contrast to previous studies in *Eucalyptus*, we had access 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.

Several issues remain to be investigated for the operational adoption of genomic prediction in eucalypt breeding. First, how does the accuracy of genomic prediction decline over

successive generations of selection due to subsequent recombination? Second, how stable are genomic prediction models across multiple environments and how important is it to consider genotype by environment interactions in the models? Finally, we have only considered additive genetic variance for building genomic prediction models in our population, but it is possible and perhaps even likely that non-additive genetic effects will play an important role in many breeding populations and specifically in populations consisting of early generation hybrids.

List of 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.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Data availability

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

588 **Competing interests**

589 The authors declare that they have no competing interests.

590 **Funding**

591 The study has partly been funded through grants from Vetenskapsrådet and the
592 Kempestiftelserna to PKI. BT gratefully acknowledges financial support from the UPSC
593 "Industrial graduate school in forest genetics, biotechnology and breeding".

594 **Authors' contributions**

595 BT, BS and PKI conceived and designed the experiment; GSM phenotyped data; GSM and
596 KZF collected samples for genotyping; DG prepared the DNA for genotyping; BT analysed
597 the data and drafted the first version of the manuscript; DG and PKI provided guidance
598 during data analyses; BT, DG, BS and PKI critically contributed to the final version of the
599 manuscript. All authors read and approved the final manuscript.

600 **Acknowledgements**

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

604

605

606 ## References

607 1. Rezende GDSP, Resende MDV, Assis TF. *Eucalyptus* breeding for clonal forestry. In:
608 Challenges and opportunities for the world's forests in the 21st century. Edited by
609 Fenning T. Dordrecht: Springer Netherlands; 2014: 393-424.

2. Myburg AA, Potts BM, Marques CM, Kirst M, Gion J-M, Grattapaglia D, Grima-Pettenatti J. Eucalypts. In: Forest Trees. Springer; 2007: 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):192-196.
4. Resende MD, Resende MF, Jr., 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):116-128.
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):409-421.
6. Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001; 157(4):1819-1829.
7. Meuwissen T, Hayes B, Goddard M. Accelerating improvement of livestock with genomic selection. Annu Rev Anim Biosci. 2013; 1:221-237.
8. Van Eenennaam AL, Weigel KA, Young AE, Cleveland MA, Dekkers JCM. Applied animal genomics: results from the field. Annu Rev Anim Biosci. 2014; 2:105-139.
9. Spindel J, Begum H, Akdemir D, Virk P, Collard B, Redona E, Atlin G, Jannink JL, McCouch SR. 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, 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):1427-1436.
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. Springer; 2014: 651-682.
13. de los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MPL. Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics. 2013; 193(2):327-345.
14. Habier D, Fernando RL, Dekkers JCM. The impact of genetic relationship information on genome-assisted breeding values. Genetics. 2007; 177(4):2389-2397.
15. 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. Genetics Research. 2010; 92(4):295-308.
16. Neves HH, Carvalheiro R, Queiroz SA. A comparison of statistical methods for genomic selection in a mice population. BMC Genet. 2012; 13(1):100.
17. 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: Proc Assoc Advmt Anim Breed Genet*:* 2009. 34-37.
18. 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):366-375.
19. 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):355-364.

20. Grattapaglia D, Resende MDV. Genomic selection in forest tree breeding. Tree Genet Genomes. 2011; 7(2):241-255.
21. 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.
22. 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):4657-4665.
23. 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):1671-1684.
24. 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):1527-1540
25. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ *et al*. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007; 81(3):559-575.
26. 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):1084-1097.
27. Stephens M, Scheet P. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. Am J Hum Genet. 2005; 76(3):449-462.
28. Candes EJ, Recht B. Exact matrix completion via convex optimization. Found Comput Math. 2009; 9(6):717-772.
29. 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):427-439.
30. Stacklies W, Redestig H, Scholz M, Walther D, Selbig J. pcaMethods - a bioconductor package providing PCA methods for incomplete data. Bioinformatics. 2007; 23(9):1164-1167.
31. Remington DL, Thornsberry JM, Matsuoka Y, Wilson LM, Whitt SR, Doeblay J, Kresovich S, Goodman MM, Buckler ES. Structure of linkage disequilibrium and phenotypic associations in the maize genome. P Natl Acad Sci USA. 2001; 98(20):11479-11484.
32. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet. 2006; 2(12):2074-2093.
33. Endelman JB. Ridge Regression and other kernels for genomic selection with R package rrBLUP. Plant Genome. 2011; 4(3):250-255.
34. Legarra A, Robert-Granie C, Croiseau P, Guillaume F, Fritz S. Improved Lasso for genomic selection. Genetics Research. 2011; 93(1):77-87.
35. Crossa J, Campos Gde L, Perez P, Gianola D, Burgueno J, Araus JL, Makumbi D, Singh RP, Dreisigacker S, Yan J *et al*. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics. 2010; 186(2):713-724.
36. Gilmour AR, Gogel B, Cullis B, Thompson R, Butler D. ASReml user guide release 3.0. VSN International Ltd, Hemel Hempstead, UK www.vsni.co.uk. 2009.
37. 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):106-116.
38. 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.

39. Perez P, de los Campos G. Genome-wide regression and prediction with the BGLR statistical package. Genetics. 2014; 198(2):483-495.
40. de Los Campos G, Sorensen D, Gianola D. Genomic heritability: what is it? PLoS Genet. 2015; 11(5):e1005048.
41. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu XY, Ruden DM. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly. 2012; 6(2):80-92.
42. 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):1575-1583.
43. Resende MF, Jr., Munoz P, Resende MD, 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):1503-1510.
44. 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):343-352.
45. 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.
46. 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.
47. 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* x *glauca*) using unordered SNP imputation methods. Heredity. 2015; 115(6):547-555.
48. Isik F, Bartholome J, Farjat A, Chancerel E, Raffin A, Sanchez L, Plomion C, Bouffier L. Genomic selection in maritime pine. Plant Sci. 2016; 242:108-119.
49. Crossa J, Perez P, Hickey J, Burgueno J, Ornella L, Ceron-Rojas J, Zhang X, Dreisigacker S, Babu R, Li Y *et al*. Genomic prediction in CIMMYT maize and wheat breeding programs. Heredity. 2014; 112(1):48-60.
50. Onogi A, Ideta O, Inoshita Y, Ebana K, Yoshioka T, Yamasaki M, Iwata H. Exploring the areas of applicability of whole-genome prediction methods for Asian rice (*Oryza sativa* L.). Theor Appl Genet. 2015; 128(1):41-53.
51. 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.
52. Honarvar M, Rostami M. Accuracy of genomic prediction using RR-BLUP and Bayesian LASSO. Eur J Exp Biol. 2013; 3:42-47.
53. Daetwyler HD, Pong-Wong R, Villanueva B, Woolliams JA. The impact of genetic architecture on genome-wide evaluation methods. Genetics. 2010; 185(3):1021-1031.
54. Lorenz AJ, Chao S, Asoro FG, Heffner EL, Hayashi T, Iwata H, Smith KP, Sorrells ME, Jannink J-L. Genomic selection in plant breeding: knowledge and prospects. Adv Agron. 2011; 110.
55. 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.
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):481-491.

758 57. Riedelsheimer C, Endelman JB, Stange M, Sorrells ME, Jannink JL, Melchinger AE.
759 Genomic predictability of interconnected biparental maize populations. Genetics. 2013;
760 194(2):493-503.
761 58. Scutari M, Mackay I, Balding D. Using Genetic Distance to Infer the Accuracy of
762 Genomic Prediction. PLoS Genet. 2016; 12(9):e1006288.
763 59. Esfandyari H, Bijma P, Henryon M, Christensen OF, Sørensen AC. Genomic prediction
764 of crossbred performance based on purebred Landrace and Yorkshire data using a
765 dominance model. Genet Sel Evol. 2016; 48(1):1-9.
766 60. Ibáñez-Escriche N, Fernando RL, Toosi A, Dekkers JC. Genomic selection of purebreds
767 for crossbred performance. Genet Sel Evol. 2009; 41(1):1-10.
768 61. Esfandyari H, Sørensen AC, Bijma P. Maximizing crossbred performance through
769 purebred genomic selection. Genet Sel Evol. 2015; 47(1):1-16.
770 62. Murray C, Huerta-Sanchez E, Casey F, Bradley DG. Cattle demographic history
771 modelled from autosomal sequence variation. Philos T R Soc B. 2010; 365(1552):2531-
772 2539.
773 63. Silva-Junior OB, Grattapaglia D. Genome-wide patterns of recombination, linkage
774 disequilibrium and nucleotide diversity from pooled resequencing and single nucleotide
775 polymorphism genotyping unlock the evolutionary history of *Eucalyptus grandis*. New
776 Phytol. 2015; 208(3):830-845.
777 64. Zhang Z, Ding X, Liu J, Zhang Q, de Koning DJ. Accuracy of genomic prediction using
778 low-density marker panels. J Dairy Sci. 2011; 94(7):3642-3650.
779 65. Solberg TR, Sonesson AK, Woolliams JA, Meuwissen THE. Genomic selection using
780 different marker types and densities. J Anim Sci. 2008; 86(10):2447-2454.

781

782 Figures

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

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

Figure 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.

Figure 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.

Figure 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 ($P < 0.05$).

Figure 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 2kb 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 ($P < 0.05$).

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 1.8Mb)

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 17 kb)

Additional file 3: ANOVA analysis of sources of variation affecting the predictive ability. (DOCX 50 kb)

Additional file 4: Mean and standard deviation of predictive ability with the five prediction methods for the eight traits. (DOCX 99kb)

Additional file 5: Mean and standard deviation of predictive ability estimated with the four Training Set/Validation Set compositions. (DOCX 87kb)

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 91kb)

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 62kb)

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 59kb)

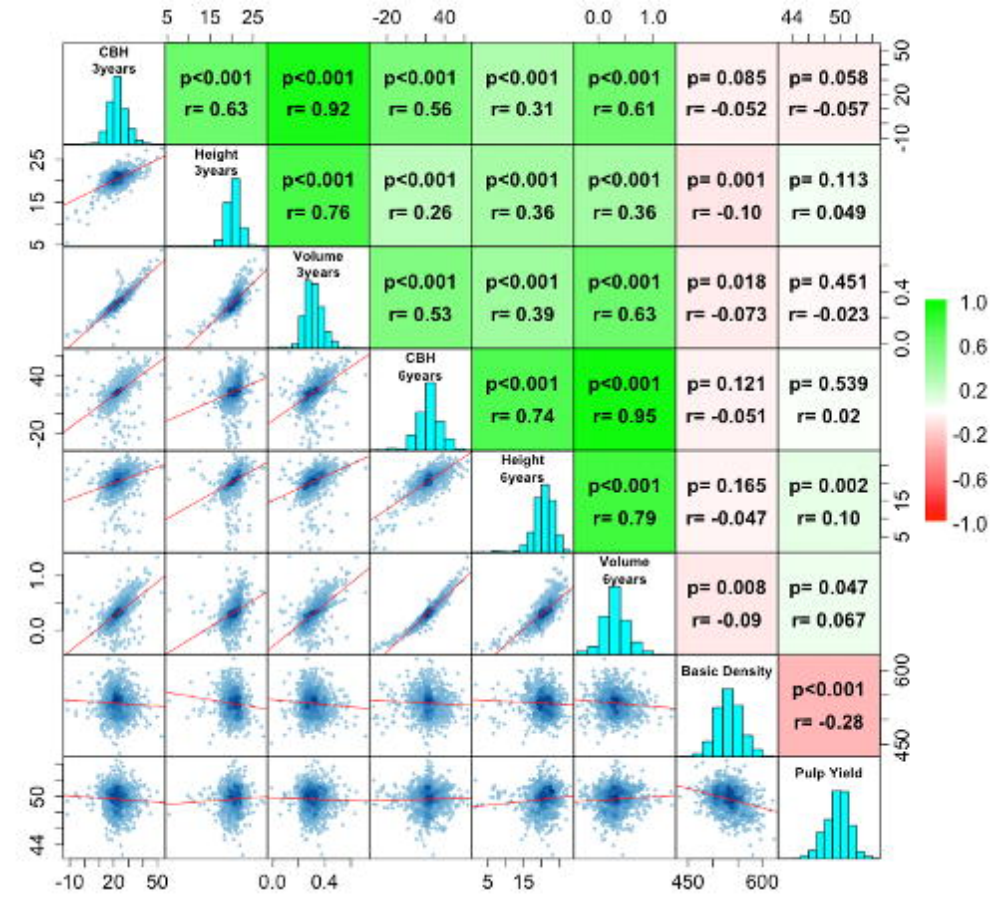
Additional file 9: ANOVA of predictive ability with SNP genomic location and SNP number as sources of variation. (DOCX 63kb)

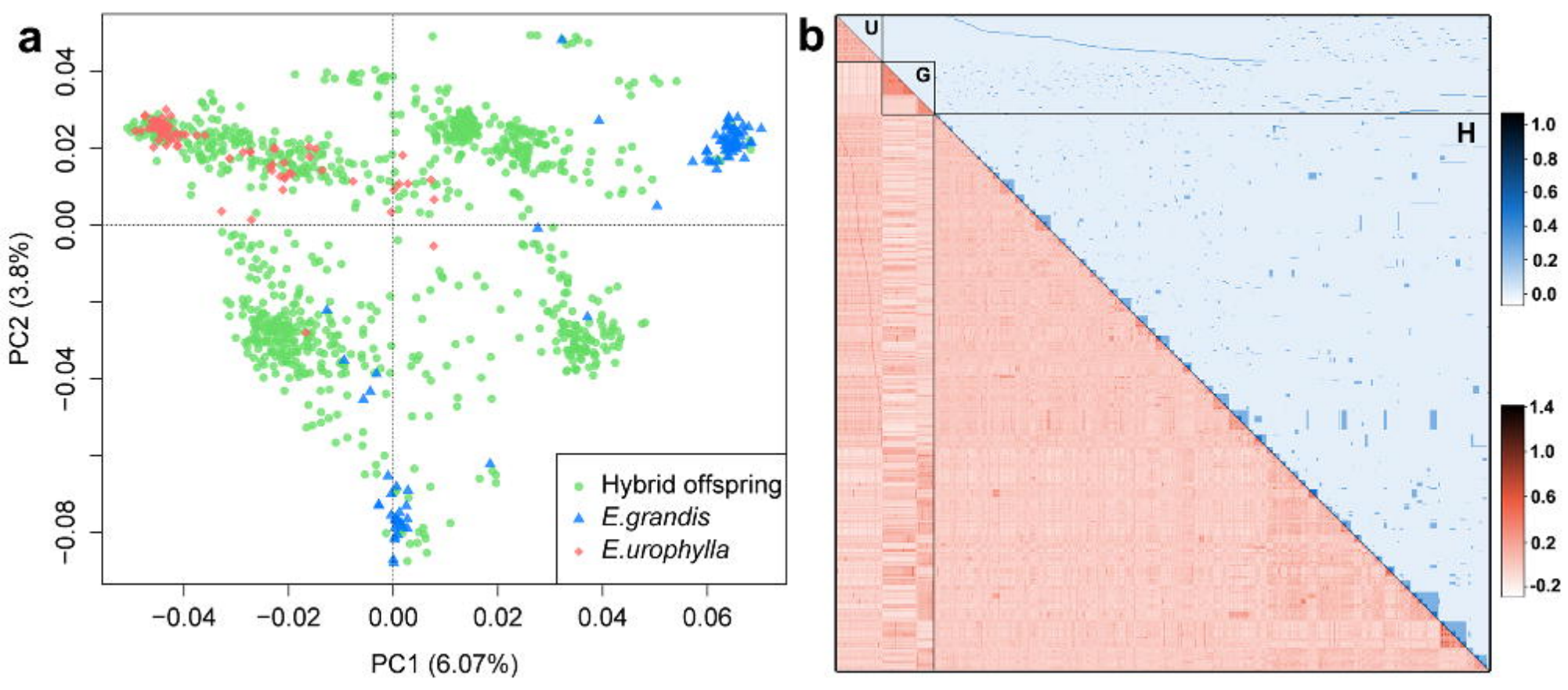
Additional file 10: Average predictive ability estimated with different numbers of SNPs fitted into the model. (DOCX 138kb)

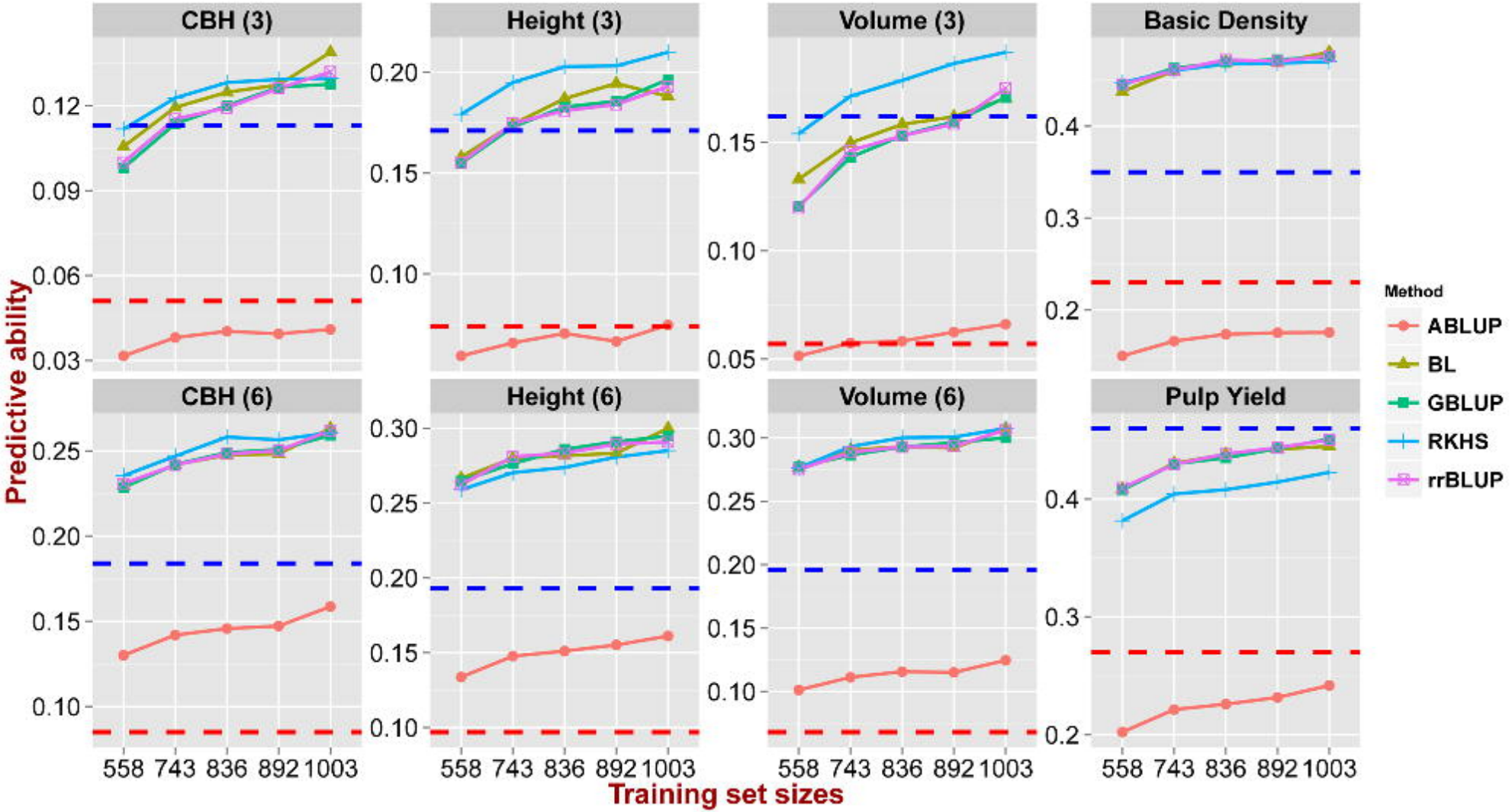
841 **Additional file 11:** Average predictive abilities estimated using SNP sets located in different
842 genomic regions. (DOCX 83kb)

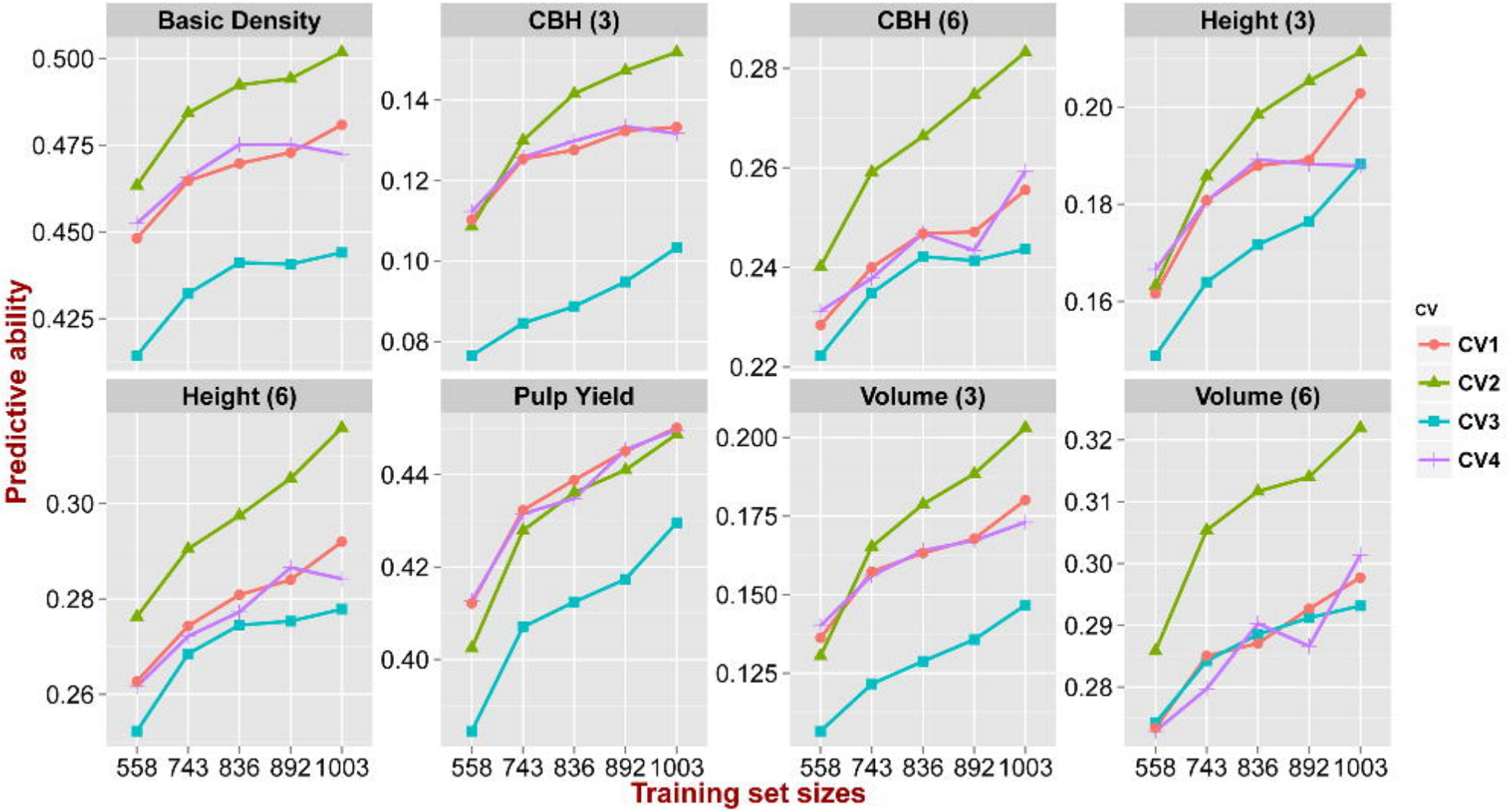
843 **Additional file 12:** Decay of linkage disequilibrium (LD) with physical distance estimated
844 with SNPs in different genomic locations. (a) A comparison of the decay of LD with physical
845 distance in four classes of SNPs located with coding, genic, intergenic and all regions,
846 respectively. Dots of pairwise LD versus physical distance and the LD decay for SNPs
847 located in all regions (b), coding region (c), genic region (d) and intergenic region (e),
848 respectively. (DOCX 1.4Mb)

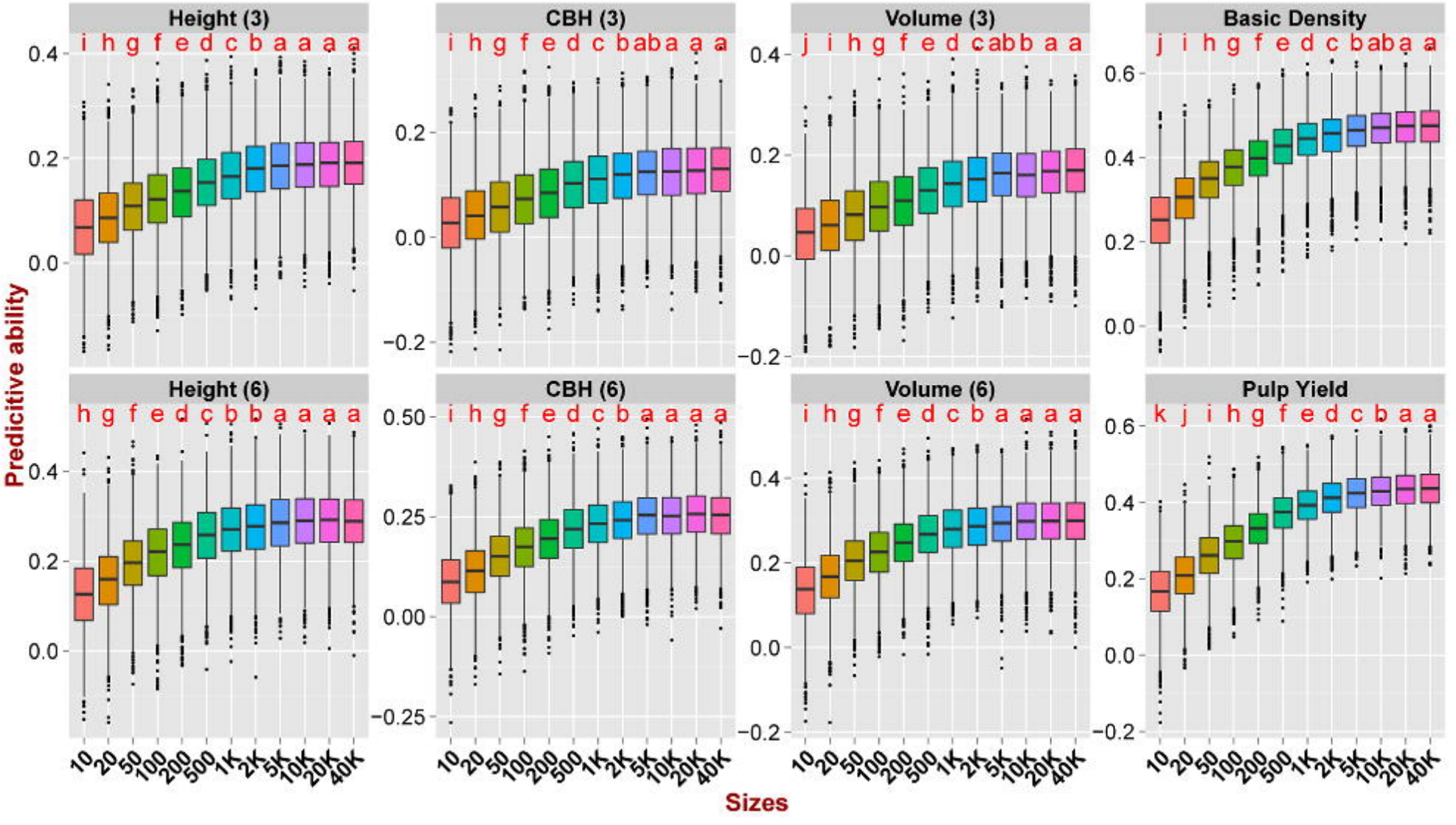
849 **Additional file 13:** Predictive abilities by training in pure species eucalypt parents and
850 predicting in their F_1 hybrids. Predictive ability estimated under three training/validation sets
851 (TS/VS) scenarios with two methods (GBLUP and RKHS) for each trait. PO168 (red boxes):
852 all 168 *E. grandis* and *E. urophylla* pure species G0 parents used for training and 168 G1
853 random selected hybrid progeny for validation; random168 (green): randomly selected 168
854 individuals from all 1,117 for TS and 168 randomly also for VS; random558 (blue):
855 randomly divided all 1,117 individuals into TS and VS of same size (558/558). Outlier
856 estimates are indicated by black dots. (DOCX 179kb)

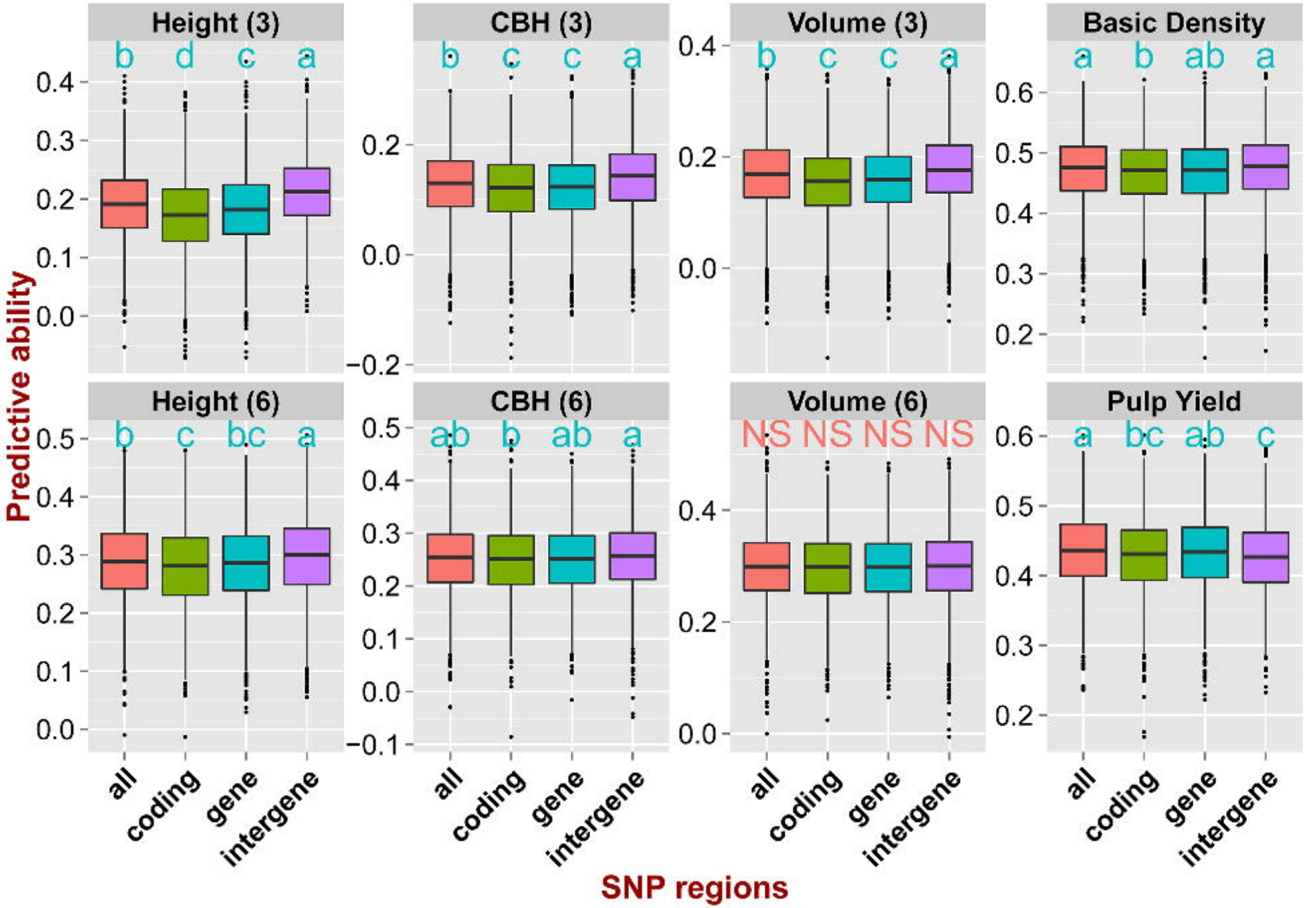

CBH 3years
p<0.001 r= 0.63
p<0.001 r= 0.92
p<0.001 r= 0.56
p<0.001 r= 0.31
p<0.001 r= 0.61
p= 0.085 r= -0.052
p= 0.058 r= -0.057
Height 3years
p<0.001 r= 0.76
p<0.001 r= 0.26
p<0.001 r= 0.36
p<0.001 r= 0.36
p= 0.001 r= -0.10
p= 0.113 r= 0.049
Volume 3years
p<0.001 r= 0.53
p<0.001 r= 0.39
p<0.001 r= 0.63
p= 0.018 r= -0.073
p= 0.451 r= -0.023
CBH 6years
p<0.001 r= 0.74
p<0.001 r= 0.95
p= 0.121 r= -0.051
p= 0.539 r= 0.02
Height 6years
p<0.001 r= 0.79
p= 0.165 r= -0.047
p= 0.002 r= 0.10
Volume 6years
p= 0.008 r= -0.09
p= 0.047 r= 0.067
Basic Density
p<0.001 r= -0.28
Pulp Yield
1.0
0.6
0.2
-0.2
-0.6
-1.0


a
b
PC2 (3.8%)
PC1 (6.07%)
0.04
0.02
0.00
−0.04
−0.08
−0.04
−0.02
0.00
0.02
0.04
0.06
Hybrid offspring
E.grandis
E.urophylla
U
G
H
1.0
0.8
0.6
0.4
0.2
0.0
1.4
1.0
0.6
0.2
-0.2


CBH (3)
Height (3)
Volume (3)
Basic Density
CBH (6)
Height (6)
Volume (6)
Pulp Yield
Predictive ability
Training set sizes
558 743 836 892 1003
Method
ABLUP
BL
GBLUP
RKHS
rrBLUP


Basic Density
CBH (3)
CBH (6)
Height (3)
Height (6)
Pulp Yield
Volume (3)
Volume (6)
Predictive ability
Training set sizes
558 743 836 892 1003
cv
CV1
CV2
CV3
CV4


Height (3)
i h g f e d c b a a a a
CBH (3)
i h g f e d c b a b a a a
Volume (3)
j i h g f e d c a b b a a
Basic Density
j i h g f e d c b a b a a
Height (6)
h g f e d c b b a a a a
CBH (6)
i h g f e d c b a a a a
Volume (6)
i h g f e d c b a a a a a
Pulp Yield
k j i h g f e d c b a a
Predicitive ability
Sizes
10 20 50 100 200 500 1K 2K 5K 10K 20K 40K


Height (3)
CBH (3)
Volume (3)
Basic Density
Height (6)
CBH (6)
Volume (6)
Pulp Yield
b d c a
b c c a
b c c a
a b ab a
b c bc a
ab b ab a
NS NS NS NS
a bc ab c
Predictive ability
SNP regions
all
coding
gene
intergene