Genomic prediction in a multi-generation Eucalyptus globulus breeding population

Geoffrey Haristoy, Laurent Bouffier, Luis Fontes, Luis Leal, Jorge Paiva, João-Pedro Pina, Jean-Marc Gion

Title: Genomic prediction in a multi-generation *Eucalyptus globulus* breeding population

Authors:

Geoffrey Haristoy **(first author)**

Affiliation: CIRAD, UMR AGAP Institut, F-34398 Montpellier, France

Email: geoffrey.haristoy@cirad.fr

ORCID: 0000-0001-9914-7712

Laurent Bouffier

Affiliation: INRAE, UMR BIOGECO, F-33610 Cestas, France

Email: laurent.bouffier@inrae.fr

ORCID: 0000-0001-7493-5077

Luis Fontes

Affiliation: Altri Florestal, 2510-582 Olho Marinho, Portugal

Email: luis.fontes@altri.pt

Luis Leal

Affiliation: Altri Florestal, 2510-582 Olho Marinho, Portugal

Email: luis.leal@altri.pt

Jorge A. P. Paiva

Affiliations: Associação CECOLAB (Collaborative Laboratory Towards Circular Economy), 3405-155 Oliveira do Hospital, Portugal ; iBET, 2780-157 Oeiras, Portugal

Email: jorge.pinto.paiva@gmail.com

ORCID: 0000-0003-3162-4396

João-Pedro Pina

Affiliation: Altri Florestal, 2510-582 Olho Marinho, Portugal

Email: pina.jp@gmail.com

Jean-Marc Gion **(corresponding author and last author)**

Affiliation: CIRAD, UMR AGAP Institut, F-34398 Montpellier, France

Email: jean-marc.gion@cirad.fr

ORCID: 0000-0003-3958-7796

Keywords: *Eucalyptus globulus*, genomic selection, GBLUP, progeny validation, pedigree error, breeding programme

Statements and Declarations: The authors have no conflict of interest to declare.

Abstract

Genomic selection is a promising approach for reducing the length of the selection cycle in forest tree breeding. Its efficiency must be evaluated across generations for this purpose, but such studies have been performed for multi-generational breeding programmes in only a few forest tree species to date. We analysed a subset of the *Eucalyptus globulus* breeding population from the Portuguese company Altri Florestal. In total, 412 genotypes from three successive breeding generation were genotyped with 14,716 SNP markers. A comparison of pedigree-based and marker-based relationship coefficients allowed to correct several documented pedigree errors. Deregressed breeding values were estimated from phenotypic records for growth traits (height and diameter) and survival for 31 field trials distributed in one breeding zone in Portugal, and used as pseudo-phenotypes for genomic prediction models. Accuracy was assessed by cross-validation according to two main scenarios: i) a scenario based on a five random fold number, not taking generation into account ; ii) scenarios investigating progeny validation using parental generations to predict the progenies. Accuracy was highest after pedigree correction, and ranged from 0.46 to 0.60 for the first scenario, from -0.56 to 0.72 for parent/progeny scenarios, and from 0.34 to 0.78 when progenies were added to the calibration population. This genomic selection study provides promising insight for the Altri Florestal *Eucalyptus* breeding programme.

Introduction

Eucalyptus globulus (Tasmanian blue gum) is an evergreen broadleaf tree species endemic to southern Australia. This forest tree species is one of the most planted hardwood species in temperate regions worldwide. Its economic importance is mainly due to the suitability of its wood for pulp and paper production. This species is also characterized by rapid growth, challenging vegetative propagation by cuttings since it is a rooting recalcitrant species and its ability to adapt to harsh environmental conditions. *E. globulus* covers an area of 0.84 million hectares in Portugal (ICNF 2019), and 0.64 million hectares in Spain (MAPA 2019). Since the 1960s, *E. globulus* breeding programmes have been developed in Portugal, to support intensive silviculture. The Portuguese company Altri Florestal selects varieties of this species, mostly on the basis of wood productivity-related traits, such as growth, and traits related to adaptation (survival). The Altri Florestal breeding programme is still in the early stages of *E. globulus* domestication with only three generations so far, like most of the advanced genetic materials currently available in forestry (Jones et al. 2006; Borralho et al. 2007).

Breeding programmes were initially based on a few key genetic trials, but have gradually expanded, with an ever-increasing number of trials and phenotyped trees. The genetic performance of trees is commonly evaluated on the basis of genetic co-variances in known relatives arising from the pedigree, according to the individual mixed model (Henderson 1950, 1975). This method is based on a numerator relationship matrix (A matrix) derived from the pedigree, which provides information about the proportion of alleles expected to be identical by descent between two individuals (Mrode 2013). Such models have already proved effective for the estimation of genetic components, especially in cases of unbalanced data, a situation frequently encountered in forest tree breeding (Borralho 1995; Jarvis et al. 1995), but their application is hindered by approximations. Indeed, the base population is assumed to consist of unrelated founders, although some cryptic relatedness may exist (Powell et al. 2010). In addition, pedigree information is rarely fully documented, particularly in open crossing strategies (Klápště et al. 2014). Finally, various identity or pedigree errors may occur during the breeding process, from the greenhouse to the field (pollination, seedlings, cuttings, plantation, etc.), compromising the genetic evaluation (Ericsson 1999). Such errors are cumulative over generations, and the earlier they occur in breeding cycles, the more likely they are to have a significant and damaging long-term impact on genetic evaluation. Recent decades have seen considerable advances in molecular biology, leading to the development of new tools for forest tree breeders. Pedigree-based relationships are based on expectations of the sharing of genomic material between individuals, but high-throughput genotyping has made it possible to estimate the proportion of alleles common to individuals (actual or realised relationship) precisely, including the within-family variability arising from Mendelian sampling (Hill and Weir 2011). This information can be summarised in a genomic relationship matrix (G matrix) (VanRaden 2008), and can be used in a genomic selection (GS) strategy through the so-called GBLUP (Genomic Best Linear Unbiased Prediction) methodology (Meuwissen et al. 2001), which involves replacing the A matrix with a G matrix in the individual mixed model. The deviation of A and G matrices is one of the key factors promoting the use of GS to obtain more accurate breeding values (Hayes et al. 2009b). GS exploits the linkage disequilibrium (LD) between high-throughput molecular data and targeted traits. Based on a calibration population of several hundred phenotyped and genotyped individuals, a predictive model is built to predict the genetic values of genotyped individuals. *Eucalyptus* breeding should benefit greatly from GS, as this approach makes early selection possible. Indeed, age-age correlations for both height and diameter are generally low-to-moderate between the ages of one

to three years, delaying progeny evaluation and clonal trials (Salas et al. 2014). *Eucalyptus* breeding programmes therefore require about 12-16 years, from initial recombination to clonal selection and operational deployment (Rezende et al. 2014).

Over the last decade, GS has benefited from many proof-of-concept studies for the genus *Eucalyptus,* with more than 20 publications and promising results, at least as good as those obtained by conventional phenotypic selection in most studies (Lebedev et al. 2020; Ahmar et al. 2021). However, such studies have generally focused on a limited number of progeny trials, whereas most breeding programmes involve at least several dozen, if not hundreds of trials. Furthermore, only a few studies have explored GS across generations, whereas such approach would be required in current breeding programmes based on recurrent selection strategies (Grattapaglia 2017). This study aims to fill this gap by applying GS in the context of the advanced *Eucalyptus globulus* breeding programme of the Portuguese company Altri Florestal. Our main objectives were: i) to highlight pedigree errors by comparing pedigree-based (A matrix) and marker-based relationship coefficients (G matrix) and to investigate the consequences of such errors for breeding value prediction; ii) to assess the accuracy of GS for three major traits in forest tree breeding (height, diameter and survival); and iii) to investigate GS accuracy over generations.

Materials and methods

1. Biological resources

This study was performed on a subset of the *Eucalyptus globulus* breeding population of Altri Florestal, referred to here as POP_{TOT}. In total, 81,520 *E. globulus* genotypes were evaluated in 31 progeny and clonal trials (92,679 trees) planted in the coastal region of Portugal between 1986 and 2009. This region, which is one of the four breeding zones defined by Altri Florestal based on temperatures and precipitations, is considered to be an "unrestricted" environment as it is not subjected to major drought or cold events. According to the documented pedigree, POP_{TOT} encompassed three breeding generations, with 2.8% of the genotypes (2,250) of unknown parentage belonging to the base population called G0. The next generation, corresponding to the first improved generation (G1), contained 41.6% of the genotypes (33,909), each with at least one documented G0 parent. The remaining 55.6% of the genotypes (45,361 genotypes) had at least one documented grandparent and belonged to the second improved breeding generation (G2). The G0 genotypes were documented as originating from selections from populations native to Australia or growing in various stands in Portugal, USA, Chile, Spain and Uruguay. A subsample of 412 POP_{TOT} genotypes corresponding to the genotypes available in clonal archives was selected for high-throughput molecular genotyping: 46 from G0, 292 from G1 and 74 from G2 (Fig. 1). Three full-sib (FS) families accounted for 81% of the genotyped G2 trees: FS1 (15 genotypes), FS2 (29 genotypes) and FS3 (16 genotypes).

2. SNP genotyping

Total genomic DNA was extracted from dried leaves from mature trees with an adapted CTAB protocol (Doyle 1991). DNA concentration was determined with a Quantit fluorometer (Invitrogen, Carlsbad, USA). Single-nucleotide polymorphism (SNP) genotyping was performed with the commercial Axiom Euc72K SNP Array (Affymetrix, Santa Clara, USA), anchored on the 11 linkage groups of the reference *Eucalyptus* genome (Myburg et al. 2014). A first set of thresholds was applied with the default settings of Axiom Suite Analysis v5.0.1 software (Affymetrix, Santa Clara, USA) and an SNP call rate threshold of 97% (detailed in Table S1). SNPs were, thus, classified as polymorphic (PolyHighResolution and NoMinorHom), monomorphic, or failed SNPs (CallRateBelowThreshold, OTV, Other). The highest quality polymorphic SNPs were then selected according to a second set of thresholds: a Fisher's linear discriminant (FLD) above 5, a homozygote ratio offset (HomRo) above 1, a heterozygous strength offset (HetSO) above 0, and a minor allele frequency (MAF) higher than 0.05. Eight genotypes were genotyped twice to assess the repeatability of the genotyping experiment, calculated as the proportion of identical SNP alleles between two samples of a given genotype.

3. Comparison of pedigree-based and genomic relationship matrices

A dedicated R script was used to check the uniqueness of each genotype based on SNPs. Pairs of genotypes with different identities and more than 99% SNP alleles in common were considered synonymous. In this case, the highest call rate profile was retained, and pairs of synonymous genotypes were renamed under the same identity

label in the pedigree and field measurement files. The parents were considered to be unknown if the initially documented parents of two synonymous genotypes were different. After identity correction, the final set of unique genotypes was named POP_{GS}. An initial expected additive numerator relationship matrix (A_I matrix) was calculated for POP_{GS} with the R package kinship2 (Sinnwell et al. 2014) based on the documented pedigree. A genomic relationship matrix (G matrix) was also calculated from the observed allele frequencies (VanRaden 2008), with the R package AGHMatrix (Amadeu et al. 2016):

$$G = \frac{(M-P)(M-P)'}{2\sum_{i=1}^{m} p_i(1-p_i)} \qquad \text{(Eq. 1)}$$

where M is a matrix of dimension n×m (n is the number of individuals and m the number of loci) giving the genotype at each locus *i,* coded as 1 for minor allele homozygous, 2 for heterozygous, and 3 for major allel homozygous. P is the matrix of allele frequencies (n×m) for all individuals, which takes the following form $2 \times (p_i - 0.5)$ where p_i is the frequency of the least frequent allele at the considered locus i. The documented parent-progeny relationship (P/P) were checked by counting the number of non-concordant SNPs for each documented P/P. P/P highlighting fewer than 115 non-concordant SNPs were considered as "true", whereas P/P harbouring more than 115 non-concordant SNPs were considered "false". P/P involving non-genotyped parent were considered to be "undetermined". For each progeny involved in an "undetermined" P/P, G coefficients for all documented full and half siblings were compared to the expected A_I coefficients. If more than 40% of the pairwise differences between G and A_I coefficients exceeded the threshold of 0.2 (Thumma et al. 2022), the pedigree was considered to be inconsistent. For such individuals, the identity of the initial parents was either considered to be unknown or was replaced by a new parent identity if allelic patterns were found to be concordant. The corrected pedigree was used to generate the corrected numerator relationship matrix A_C. The A_I, A_C and G matrices were visualised with the ggplot2 (Wickham 2016) and ComplexHeatmap (Gu et al. 2016) packages in the R statistical environment (Rstudio Team 2021).

4. Pseudo-phenotype estimates

Trials have been measured at various ages for growth traits and survival, but only the most recent measurements were considered for each trial (i.e. ages ranging from 6 to 17 years depending on the trial). The diameter over-bark at breast height (DBH) was calculated as the mean of two tree calliper measurements taken at right angles. Height (HT) was measured with a telescopic pole. Survival (SV) was equal to 1 for living trees, and 0 for dead trees. For each trait, estimated breeding values (EBV) were obtained with ASREML 4.0 (Butler et al. 2017) from a BLUP meta-analysis routinely implemented at Altri Florestal as detailed in Borralho et al. (2018). This mixed-model included fixed effects (trial and replicates within trial) and random effects (incomplete blocks within replicates, additive genetic effect, full-sib family, clone within full-sib family), as well as pedigree relationships across all trees (numerator relationship matrix). The variance components and heritability estimates considered for the meta-analysis were described in Table 1 following Borralho et al. (2018). Considering this mixed-model, an initial estimated breeding value (EBV_I) was calculated with the A_I matrix, and a corrected estimated breeding value (EBV_C) was calculated with the A_C matrix. EBV accuracy was estimated as follows (Isik et al. 2017):

$$r = \sqrt{1 - \frac{S^2}{(1+F)\sigma^2{}_A}} \qquad \text{(Eq.2)}$$

where S is the standard error of the EBV, F is the coefficient of inbreeding and σ^2_A is the additive genetic variance. Additional EBV_C values were calculated with the BLUP meta-analysis based on truncated phenotypic data: i) the $EBV_{C\text{-}T01}$ values considering only phenotypic data from the G0 and G1 genotypes, and ii) the $EBV_{C\text{-}T2}$ values considering only phenotypic data from G2 genotypes. $EBV_{C\text{-}T01}$ and $EBV_{C\text{-}T2}$, were, thus, estimated with independent phenotypic data sets. Considering EBV as phenotypes in genomic prediction may introduce bias and heterogeneity (Garrick et al. 2009). EBV was therefore deregressed and weighted (dEBV) following Garrick et al. (2009) based on estimates of heritability and without removal of the parent average effect, as many individuals had unknown fathers. The resulting $dEBV_I$, $dEBV_C$, $dEBV_{C\text{-}T01}$, $dEBV_{C\text{-}T2}$ were used as pseudo-phenotypes for GS.

Table 1 Variances associated with each random effects and heritability (h^2) estimated for HT, DBH and SV.

σ^2_A, σ^2_r, σ^2_f, σ^2_c and σ^2_e are the variances associated with the following random effects: additive genetic effect, incomplete block within replicate, full-sib family, clone within full-sib family and residuals, respectively. Heritability was estimated as: $h^2 = \frac{\sigma^2_A}{\sigma^2_r+\sigma^2_A+\sigma^2_f+\sigma^2_c+\sigma^2_e}$

Trait	σ_A^2	σ_r^2	σ_f^2	σ_c^2	σ_e^2	h^2
HT	0.25	0.19	0.03	0.05	0.72	0.20
DBH	0.15	0.08	0.04	0.04	0.80	0.14
SV	0.45	-	0.05	0.25	3.29	0.11

5. Genomic prediction models

Genomic estimated breeding values ($GEBV_I$, $GEBV_C$, $GEBV_{C\text{-}T01}$ and $GEBV_{C\text{-}T2}$) were estimated for each trait, from $dEBV_I$, $dEBV_C$, $dEBV_{C\text{-}T01}$, and $dEBV_{C\text{-}T2}$, respectively. The following GBLUP model was implemented with the BreedR R package (Munoz and Rodriguez 2020):

$$y = X\mu + Za + e \qquad \text{(Eq. 3)}$$

in which y is the vector of pseudo-phenotypes ($dEBV_I$, $dEBV_C$, $dEBV_{C\text{-}T01}$, or $dEBV_{C\text{-}T2}$), μ the population mean, a the vector of random additive genetic effects and e the vector of residuals effects. X and Z are the incidence matrices for μ and a effects. The vector a was assumed to follow a normal distribution $a \sim N(0, G\sigma_a^2)$, where G is the realised relationship matrix and σ^2_a the variance of additive effects. The vector e followed a normal distribution with $e \sim N(0, I\sigma^2_e)$, where σ^2_e is the residual variance. The solutions for the random genetic effects of Eq. 3 are the genomic estimated breeding values, GEBV. GBLUP shrunk marker effects uniformly, assuming a centred normal distribution and a common variance for marker effects.

6. Cross-validation scenarios

Nine cross-validation scenarios were tested to assess the accuracy of GS for each trait (Table 2). In the random cross-validation scenario (S0), the three generations were all included in the calibration population (CP) and the validation population (VP), whereas the others cross-validation scenarios were designed for the investigation of GS accuracy over breeding generations by using different configurations of pseudo-phenotypes for both the CP and the VP.

Table 2 Description of the nine GS scenarios (S0, $S1_a$, $S1_b$, $S2_a$, $S2_b$, $S2_c$, $S2_d$, $S3_a$ and $S3_b$) regarding the pseudo-phenotype used, the number of genotypes in the calibration (CP) and validation (VP) populations (the number of genotypes indicated is after pedigree correction), the number of iterations performed, and the corresponding accuracy

* randomly selected in POP_{GS} ; ** 20% of the G1 from POP_{GS} ; *** 20% of the three main G2 FS, and 20% of all the other G2 families ; **** mean scenario accuracy over the 100 iterations.

Scenarios	CP		VP		Iterations	Accuracy
	Pseudo-phenotypes	Genotypes	Pseudo-phenotypes	Genotypes		
S0	$dEBV_C$	321 G0/G1/G2 *	$dEBV_C$	80 G0/G1/G2 *	100	$\rho(GEBV_C, dEBV_C)$ ****
$S1_a$	$dEBV_C$	44 G0	$dEBV_C$	284 G1	-	$\rho(GEBV_C, dEBV_C)$
$S1_b$	$dEBV_C$	44 G0, 57 G1 **	$dEBV_C$	227 G1	100	$\rho(GEBV_C, dEBV_C)$ ****
$S2_a$	$dEBV_C$	44 G0, 284 G1	$dEBV_C$	73 G2	-	$\rho(GEBV_C, dEBV_C)$
$S2_b$	$dEBV_C$	44 G0, 284 G1	$dEBV_{C\text{-}T2}$	73 G2	-	$\rho(GEBV_C, dEBV_{C\text{-}T2})$
$S2_c$	$dEBV_{C\text{-}T01}$	44 G0, 284 G1	$dEBV_C$	73 G2	-	$\rho(GEBV_{C\text{-}T01}, dEBV_C)$
$S2_d$	$dEBV_{C\text{-}T01}$	44 G0, 284 G1	$dEBV_{C\text{-}T2}$	73 G2	-	$\rho(GEBV_{C\text{-}T01}, dEBV_{C\text{-}T2})$
$S3_a$	$dEBV_C$	44 G0, 284 G1, 14 G2 ***	$dEBV_C$	59 G2	100	$\rho(GEBV_C, dEBV_C)$ ****
$S3_b$	$dEBV_C$	44 G0, 284 G1, 14 G2 ***	$dEBV_{C\text{-}T2}$	59 G2	100	$\rho(GEBV_C, dEBV_{C\text{-}T2})$ ****

For the S0 scenario, 80% of POP_{GS} (321 genotypes) were randomly assigned to the CP, with the remaining 20% (80 genotypes) were used as the VP (5 folds, 100 iterations). For $S1_a$, only G0 genotypes were included in the CP (44 genotypes) for prediction of all the G1 genotypes in the VP. $S1_b$ was similar to $S1_a$ but 20% of the G1 genotypes (57 genotypes) were added to the CP (100 iterations). For $S2_a$, all genotyped G0 and G1 were included in the CP (328 genotypes), and $GEBV_C$ were predicted in the VP for the 73 G2 genotypes. In $S2_b$, prediction accuracy was estimated relative to $dEBV_{C\text{-}T2}$. The $S2_c$ scenario used $dEBV_{C\text{-}T01}$ in the CP to predict the G2 $GEBV_{C\text{-}T01}$ of the VP, which was compared to $dEBV_C$ for accuracy estimation. The $S2_d$ used the same pseudo-phenotype in the CP to estimate $GEBV_{C\text{-}T01}$, which was compared to $dEBV_{C\text{-}T2}$. In $S3_a$ and $S3_b$, 14 additional G2 genotypes were added to the G0 and G1 genotypes for the CP (342 genotypes), considering 20% of the three main FS families and 20% of all the other families used to estimate the $GEBV_C$ for the remaining G2 (59) in the VP (100 iterations). Prediction accuracy was estimated relative to either $dEBV_C$ ($S3_a$) or $dEBV_{C\text{-}T2}$ ($S3_b$). When 100 iterations were considered (S0, $S1_b$, $S3_a$ and $S3_b$), the accuracy was calculated as the mean accuracy over the 100 iterations.

Results

1. Genotyping results

The first filtering of 412 genotypes with 68,055 SNPs classified 28,242 SNPs as polymorphic SNPs (41.5%), 23,165 as monomorphic (34.0%), and 16,648 as failed (24.5%). The second filtering step selected the highest quality polymorphic SNPs: 14,716 SNPs uniformly distributed over the 11 chromosomes of the reference genome (Table 3). This final set of SNPs was used for pedigree correction and genomic predictions. The mean sample call rate over all genotype samples was 99.6%. Based on the replicated samples, repeatability was estimated at 99.98%.

Table 3 Marker coverage of the 14,716 SNPs, anchored onto the 11 chromosomes (Chr) of the *E. grandis* v2.0 reference genome (Myburg et al. 2014; Bartholomé et al. 2015). SNP positions were retrieved from Affymetrix documentation to determine if they were located within gene or not. * Chromosome physical length in Mb from Myburg et al. (2014) and their corresponding genetic length in cM based on the *Eucalyptus* composite map from Hudson et al. (2012)

Chr	Number of SNPs	Chr length *		Density		Mean distance between SNPs (kb)	Number of SNPs within gene	Proportion of SNPs within gene (%)
		(Mb)	(cM)	(SNP/Mb)	(SNP/cM)			
1	1,192	40.3	93.8	29.6	12.7	38	805	68
2	1,938	64.2	102.1	30.2	19.0	31	1,253	65
3	1,932	80.1	105.6	24.1	18.3	43	1,249	65
4	786	42.0	80.9	18.7	9.7	51	479	61
5	1,596	74.7	95.9	21.4	16.6	48	1,002	63
6	1,569	54.0	125.3	29.0	12.5	37	1,025	65
7	1,275	52.4	87.7	24.3	14.5	43	872	68
8	1,568	74.3	137.3	21.1	11.4	46	950	61
9	934	39.0	82.9	24.0	11.3	41	569	61
10	955	39.4	97.8	24.2	9.8	39	614	64
11	971	45.5	97.3	21.4	10.0	46	625	64
Genome	14,716	605.9	1,106.5	24.4	13.3	42	9,443	64

The number of SNPs per chromosome (Table 3) ranged from 786 (Chr4) to 1,938 (Chr2) with a mean of 1,338 SNPs per chromosome (mean density of 24.4 SNPs per Mb). The largest distance observed between two neighbouring SNPs on the same chromosome ranged from 970 kb (Chr11) to 3760 kb (Chr5), and the smallest distance between SNPs on the same chromosome ranged from 30 bp (Chr2, Chr3, Chr5, Chr9) to 40 bp (Chr11). Considering the composite linkage map of *Eucalyptus* (Hudson et al. 2012), the number of SNPs per cM ranged from 9.7 (Chr4) to 19.0 (Chr2). Based on the *E. grandis* reference genome, 64% of the SNPs (9,443 SNPs) were located within a gene corresponding to 6,273 different genes out of the 36,376 genes estimated in this species.

2. Pedigree correction and its effect on estimated breeding value (EBV)

Eleven pairs of genotypes were identified as synonymous and were renamed as 11 unique genotypes, accounting for 5.3% of the genotyping set (2 in G0, 8 in G1, and 1 in G2). Thus, POP_{GS} contained 401 unique genotypes spread over three generations: 44 G0, 284 G1, and 73 G2. After the correction of synonymous genotypes, the G matrix (dimension n=401) derived from POP_{GS} was compared to the corresponding subsample

A_I matrix of the same dimensions (Fig. 2a). The A_I matrix coefficients were discrete variables with six relationship classes (0, 0.0625, 0.125, 0.25, 0.5 and 1), whereas the G matrix coefficients followed a continuous distribution ranging from -0.20 to 1.44. Negative coefficients in the G matrix suggested that some individuals had fewer markers in common than expected on the basis of allele frequencies. The diagonal elements of the A_I matrix indicate an absence of inbreeding (relationship of 1), whereas the diagonal elements of the G matrix ranged from 0.82 to 1.44.

Overall, 238 of the 401 genotypes were involved in at least one relationship with A_I=0.25, and 264 were involved in at least one relationship with A_I=0.5 (and 198 were involved in both types of relationships). For both the 0.25 and 0.5 A_I classes, bimodal and asymmetric distributions were observed (Fig. 3a and Fig. 3b) with the largest peak close to the expected value, and a second peak close to zero, suggesting the existence of errors in the documented pedigree. P/P consistencies were first assessed in POP_{GS} by evaluating the compatibility of SNP between parents and progenies. A gap between P/P associated with less than 115 non-concordant SNPs and P/P associated with more than 911 non-concordant SNPs was highlighted (Fig. 4) which justified the threshold considered to make the distinction between "true" and "false" P/P. Both parents were considered to have been correctly documented for 112 genotypes ("true" P/P) which highlighted between 46 and 113 non-concordant SNPs (i.e. between 0.31% and 0.77% of the available SNPs). The presence of non-concordant SNPs for "true" P/P was explained by the genotyping repeatability below 100%. In the present study, the repeatability (99.98%) was probably overestimated as based on only eight replicates. At least one parent was incorrectly documented for 37 genotypes ("false" P/P) which were associated with 912 to 1405 non-concordant SNPs (6.2% to 9.5% of the available SNPs). As some parents were not genotyped, 226 individuals had at least one "undetermined" P/P. Their consistencies were then analysed by comparing A_I and G matrix coefficients for full-sib and half-sib relationships. The distribution of the percentage of relationships differing from more 0.2 was represented in Fig. 5a. Our strategy was to identify and correct pedigree for genotypes which clearly highlighted pedigree inconsistencies. In addition to the "false" P/P previously detected, we identified eight additional genotypes for which more than 40% of G coefficients were at least 0.2 lower than the A_I coefficient, suggesting an incorrect parent. In total, for 45 genotypes (41 G1 and 4 G2), corresponding to 11.2% of POP_{GS}, at least one wrong parent was identified. Reassignment of the correct parents was possible for 14 genotypes (14 G1), and the wrong parents were replaced by unknown parents in the documented pedigree for the remaining 31 genotypes (27 G1 and 4 G2). Finally, an A_C matrix was built from the corrected pedigree and compared to the G matrix. As illustrated in Fig. 2b, the A_C matrix better matched the G matrix than did the A_I matrix. The G matrix coefficient statistics for each A matrix coefficient class were reported in Table 4. As expected, the A_C matrix coefficients fitted better the G matrix coefficients for the two classes considered for pedigree correction (0.25 and 0.5), bringing the mean G matrix coefficient closer to the A matrix coefficient, and reducing the corresponding standard deviations. The applied pedigree correction resolved the bimodal distribution, resulting in a single peak (Fig. 3c and Fig. 3d) and highly decreased the percentage of relationships with A_C and G differing from more than 0.2 (Fig. 5b). There was also a minor effect in other classes, due to the resulting changes in the number of relationships per A class (Fig. S1). After correction, G2 genotypes from the three main FS families accounted for 77% of G2 genotypes in POP_{GS}: FS1 (14 genotypes), FS2 (27 genotypes) and FS3 (15 genotypes).

The G matrix also uncovered relatedness between individuals, that was not expected based on the documented pedigree, i.e. hidden relationships. For example, for the 145,672 pairwise relationships from the A_C=0 class of

unrelated individuals, a total of 1,668 G matrix coefficients (1.1%) were greater than 0.2, mostly in G0 and G1. For the A_C=0.25 class, 52 of the 10,482 relationships (0.5%) had G matrix coefficients greater than 0.5, suggesting the existence of a few undocumented FS or P/P relationships. For hidden relationships not involved in pedigree error, the documented pedigree was not modified. For the three traits considered, Pearson correlation coefficients (ρ) between EBV_I and EBV_C revealed a small but non-negligible effect of pedigree modification on EBV estimates (Fig. S2). At the POP_{GS} level, ρ was 0.99 for HT, 0.97 for DBH and 0.98 for SV. However, when we considered only the 45 genotypes for which parentage errors were highlighted, slightly higher deviations between EBV_I and EBV_C were observed, with ρ equal to 0.92 for HT, 0.81 for DBH and 0.73 for SV (red dots in Fig. S2).

Table 4 G relationship coefficients according to the expected A_I and A_C matrix coefficients, for the 401 genotypes of POP_{GS}

Relationship matrices	Expected coefficient	Number of relationships	Mean	Sd	Min	Max
A_I	0	144,000	-0.02	0.07	-0.20	0.76
	0.0625	58	0.12	0.03	0.00	0.18
	0.125	148	0.14	0.09	-0.05	0.30
	0.25	11,870	0.11	0.11	-0.16	0.72
	0.5	4,324	0.40	0.16	-0.15	0.77
	1	401	1.04	0.13	0.82	1.44
A_C	0	145,672	-0.03	0.07	-0.20	0.76
	0.0625	54	0.13	0.02	0.08	0.18
	0.125	140	0.15	0.09	-0.05	0.30
	0.25	10,482	0.15	0.07	-0.04	0.72
	0.5	4,052	0.44	0.10	0.10	0.77
	1	401	1.04	0.13	0.82	1.44

3. Genetic values and correlations over generations

We investigated the trends in genetic value over generations, and compared genetic values obtained independently from G0/G1 and G2 phenotypic data through a truncation process. Table 5 presents the descriptive statistics for EBV_C for each trait in each generation of POP_{GS}. For growth traits, the lowest EBV_C means were obtained for G1 (0.15 for HT and 0.01 for DBH), with higher values obtained in G2 (0.20 for HT and 0.12 for DBH). For SV, mean EBV_C decreased slightly over generations, with mean values of 0.23 for G0, 0.21 for G1 and 0.16 for G2. The mean accuracy of all EBV_C was high, regardless of the generation considered, ranging from 0.72 to 0.92 in POP_{GS}. This high accuracy was explained by the use of a meta-analysis based on a high number of clonal copies and a high level of connectivity in the pedigree.

Table 5 Descriptive statistics (mean and standard deviation) for EBV_C and its mean accuracy (r) for tree height (HT), diameter at breast height (DBH) and survival (SV) for each generation of POP_{GS}. The number of genotypes in each generation is given (Size).

Generations	Size	HT			DBH			SV		
		Mean	Sd	r	Mean	Sd	r	Mean	Sd	r
All	401	0.17	0.34	0.88	0.03	0.33	0.84	0.2	0.33	0.73
G0	44	0.20	0.35	0.92	0.04	0.31	0.89	0.23	0.38	0.75
G1	284	0.15	0.34	0.87	0.01	0.32	0.83	0.21	0.34	0.72
G2	73	0.20	0.35	0.89	0.12	0.35	0.85	0.16	0.30	0.76

Correlations between the EBV_C for the three traits were assessed using POP_{GS} and within each breeding generation (G0, G1 and G2) (Fig. 6). Strong, highly significant positive genetic correlations (p-value < 0.001) between growth traits were observed for the whole sample (ρ=0.85), whatever the generation considered (0.86 for G0, 0.83 for G1, and 0.93 for G2). HT and SV were significantly but weakly correlated in POP_{GS} (0.18), but strongly correlated in G2 (0.55). Correlation between DBH and SV in POP_{GS} was not significant at the 5% level, even though the G2 correlation was similar to that between HT and SV.

These results highlighted that POP_{GS} was not representative of POP_{TOT} for which an increase of EBV_C over the generation was observed whatever the trait (Table S2). Based on EBV_C correlations, slight differences were also observed between POP_{TOT} and the subsample POP_{GS} (Fig. S3), although the same overall trends were found with the strongest correlations between growth traits (0.87) and weak correlations between growth traits and SV (0.25 for HT and 0.21 for DBH). The differences in EBV_C and trait correlations between POP_{TOT} and POP_{GS} could be explained by a sampling effect, with a smaller number of genotypes from G0 (44), G1 (284), and G2 (73). G2 genotypes in POP_{GS} consisted of 13 FS families, three of which accounted for 77% of G2 genotypes. By contrast, G2 in POP_{TOT} was composed of 45,360 genotypes with 914 FS. The genotyped population was therefore poorly representative of the total diversity of POP_{TOT}, particularly for the first breeding generation (G0) and the last one (G2).

Genetic values of G1 genotypes were estimated based on POP_{GS} through two independent processes based on truncated phenotypic data either $EBV_{C\text{-}T01}$ or $EBV_{C\text{-}T2}$, but keeping relatedness between genotypes. The correlations between G1 $EBV_{C\text{-}T01}$ and G1 $EBV_{C\text{-}T2}$ were weak (0.26 for HT, 0.41 for DBH, 0.14 for SV) (Fig. S4), suggesting that the G0/G1 and G2 generations made slightly different contributions to global EBV_C estimates. This was confirmed by the strong correlations between G1 $EBV_{C\text{-}T01}$ and G1 EBV_C (0.94 for HT, 0.83 for DBH, and 0.84 for SV) whereas the correlations between G1 $EBV_{C\text{-}T2}$ and G1 EBV_C were only moderate (0.47 for HT, 0.59 for DBH and 0.56 for SV). This trend was also observed for genetic value in G0. In contrast, the correlation between G2 $EBV_{C\text{-}T2}$ and G2 EBV_C (0.98 for HT, 0.99 for DBH and 0.96 for SV) was stronger than that between G2 $EBV_{C\text{-}T01}$ and G2 EBV_C (0.46 for HT, 0.43 for DBH and -0.45 for SV). Thus, for G2, global genetic value (G2 EBV_C) was determined principally from the data collected for the G2 generation.

4. Accuracy in cross-validation scenarios

The correlation between EBV and dEBV were very strong (>0.99 for all traits). We therefore used only dEBV as input variables for all GS scenarios. We first checked the effect of pedigree correction on GS accuracy. We plotted the correlation between GEBV and dEBV for the three traits according to the S0 scenario with and without pedigree correction (Fig. 7). For HT, DBH and SV, accuracy was slightly higher for $dEBV_C$ (0.46, 0.60 and 0.48, respectively) than for $dEBV_I$ (0.44, 0.55 and 0.48, respectively), with correction increasing accuracy by 5% for

HT and 9% for DBH (no change for SV). This benefit was also observed for the $S1_a$ and $S2_a$ scenarios (see Table S3), for which $dEBV_C$ gave slightly higher accuracies for HT (16% improvement for $S1_a$, and 6% for $S2_a$), DBH (39% for $S1_a$ and 1% for $S2_a$), and SV (3% for $S1_a$ and 8% in $S2_a$). Pedigree correction (11.2% of POP_{GS}) had a significant effect on GS accuracy: i) mainly for the more precisely measured growth traits, and ii) for the $S1_a$ scenario generating estimates for G1, the generation for which the largest number of corrections were made (41 of the 45 genotypes corrected belonged to G1). The $S2_a$ scenario remained the most accurate for growth traits, with a smaller effect of correction, probably due to the smaller number of pedigree corrections for G2 genotypes. Whatever the scenario, our findings suggest that pedigree correction should be performed to correct pseudo-phenotypes before applying GS.

We then compared the nine scenarios (S0, $S1_a$, $S1_b$, $S2_a$, $S2_b$, $S2_c$, $S2_d$, $S3_a$ and $S3_b$) considering only pseudo-phenotypes after pedigree correction, as shown in Fig. 8 (and Table S4). In the S0 scenario, genotypes were randomly assigned to either the CP (321 genotypes) or VP (80 genotypes), and the corresponding accuracies were higher for DBH (0.60), than for HT (0.46) or SV (0.48). This higher accuracy for DBH was observed in most of the scenarios tested. In the S1 scenarios, the addition of 20% of the G1 genotypes to the G0 genotypes used in the CP ($S1_b$) improved accuracy to 55% for HT, 16% for DBH and 17% for SV, for the prediction of G1 genotypes. Overall, S1 accuracies were lower than the accuracies achieved for scenarios S0, $S2_a$ and $S2_b$, which also used $dEBV_C$ as an input variable. In the S2 scenarios, $dEBV_C$ or $dEBV_{C\text{-}T01}$ for all G0 (44) and G1 (284) genotypes were chosen for the prediction of G2 genetic values in the VP (73). Accuracy was higher for $dEBV_C$ ($S2_a$ and $S2_b$) than for $dEBV_{C\text{-}T01}$ ($S2_c$ and $S2_d$), for which accuracy was non-significant for HT and, surprisingly, negative for SV. S2 scenarios using $dEBV_{C\text{-}T2}$ in the VP ($S2_b$ and $S2_d$) were slightly more accurate than those using $dEBV_C$ (e.g. the $S2_b$ scenario gave increases in accuracy of 4% for HT, 6% for DBH and 23% for SV relative to the $S2_a$ scenario).

The poor accuracy of scenarios $S2_c$ and $S2_d$ in comparison to scenarios $S2_a$ and $S2_b$ can be explained by the pseudo-phenotypes considered in CP. Indeed, even if $dEBV_C$ and $dEBV_{C\text{-}T01}$ were highly correlated for generation G1, this was not the case when considering specifically the parents of the three main G2 families (which represents 77% of the VP). Scenarios $S2_a$ and $S2_b$ calibrated with $dEBV_C$ (including information from G2 phenotypes) were more efficient to predict G2 genotypes than scenarios $S2_c$ and $S2_d$ based on $dEBV_{C\text{-}T01}$ (including only G0 and G1 phenotypes). In the two S3 scenarios, following the addition of G2 genotypes to the CP (14 G2), accuracy was highest for the three traits when $dEBV_{C\text{-}T2}$ was used: 0.65 for HT, 0.78 for DBH, and 0.59 for SV (Fig. 8). This suggests that the addition of G2 genotypes to the CP greatly influences the quality of prediction for the remaining G2 genotypes in the VP due to increased relatedness between CP and VP. In scenarios $S3_a$ and $S3_b$, the accuracies for the three traits were more dispersed than in S0 and similar to those in $S1_b$, scenarios for which iterations were also performed. This suggests that composition of CP was affecting prediction accuracy and thus its optimisation could maximise the accuracy of GS predictions.

Discussion

GS implementation in an advanced forest tree breeding programme requires a better knowledge of the change in genomic prediction accuracy for quantitative traits over breeding generations. Such investigations have been conducted in conifers (Bartholomé et al., 2016; Isik et al., 2016; Thistlethwaite et al., 2019), but not in eucalypts. The advanced *E. globulus* breeding programme of Altri Florestal focuses on growth and survival, providing a great opportunity to evaluate GS accuracy over three breeding generations.

The quantitative genetics of growth traits have been described in detail for *E. globulus,* with low to medium heritabilities, suggesting polygenic determinism for both primary and secondary growth traits (Lopez et al. 2002; Raymond 2002; Potts et al. 2004). The genetic architecture of this species has been studied and a large number of quantitative trait loci (QTLs) have been localised to different linkage groups, varying over time and/or environments (Freeman et al. 2013; Bartholomé et al. 2013, 2020). Genetic control has been shown to be mostly additive, although non-negligible dominance effects have also been identified (Denis and Bouvet 2013; Tan et al. 2018; Thavamanikumar et al. 2020). In *E. globulus*, the reported coefficients for genetic correlations between height and diameter range from 0.55 to 0.93 (Volker et al. 1998; Hamilton et al. 2010; Rojas 2017). We observed strong additive genetic correlations between height and diameter for each generation (G0, G1, G2). Fewer data have been published on tree survival, even though this has become a key breeding objective for companies wishing to expand their planting areas to less optimal climatic conditions (Costa e Silva et al. 2008). Survival can be defined as the ability of a genotype to cope with a set of undefined environmental factors (both biotic and abiotic constraints), which is particularly important in the early stages of the tree's life (Lopez et al. 2002). Reported heritabilities for survival in *E. globulus* are low to moderate, ranging from 0.02 to 0.38 (Chambers et al. 1996; Lopez et al. 2002; Hamilton et al. 2015; Mora and Serra 2014). The genetic determinism of survival may varies over time, with the age of the tree (Dutkowski and Potts 1999) and the breeding generation considered. For long-lived species, such as forest trees, the survival recorded at the start of a breeding programme may depend on genetic drivers different from those in contemporary measurements, due to changes in climate and the emergence of new pests. Conflicting results concerning the correlation between survival and growth have been published for *E. globulus,* from weak negative genetic correlations (Lopez et al. 2002; Mora and Serra 2014) to highly positive genetic correlations (Hamilton et al. 2010), and from non-significant phenotypic correlations (Lopez et al. 2002) to significant phenotypic correlations in a wide-ranging collection of open-pollinated *E. globulus* seeds from parent trees growing in native stands in Australia (Dutkowski and Potts 1999). We found weak genetic correlations between growth traits and survival in the studied breeding zone characterised by no major constraints related to coldness and drought. These correlations suggest the opportunity to select these two traits without trade-offs through this specific breeding zone, highlighting the importance of evaluating both of them in GS-based breeding strategies.

Increasing numbers of studies are investigating the presence of pedigree errors in breeding programmes for forest trees (Isik 2014). Two principal types of error have been highlighted: i) identity errors (synonymous labels or genetically different clonal replicates), and ii) parentage errors, when either one or both documented parents are incorrect. In eucalypts, mislabelled ramets were found for four of 10 commercial clones from several organisations (Keil and Griffin 1994). Reported rates of parentage errors are highly variable, ranging from 2.8% in *Picea rubens* (Doerksen and Herbinger 2008), to 30.2% in *Pseudotsuga menziesii* and 33.3% in *Pinus taeda* (Adams et al. 1988).

In *Pinus sylvestris* L. seed orchards, ramet assignment error rates range from 5.8% to 37.7% (Przybylski et al. 2019). In a breeding population of *Pinus radiata*, 10% of documented relationships were found to be incorrect after pedigree verification (Kumar and Richardson 2005). Some progenies from open-pollinated families of *Picea glauca* were found to have a relationship coefficient of zero, suggesting pedigree errors (Gamal El-Dien et al. 2016). Most of these pedigree verifications compared allelic consistencies between parents and progenies. We propose here an additional method based on comparison between A and G matrices applied to siblings for detecting pedigree errors. Our original approach identified 11.2% parentage errors in addition to the 5.3% identity errors detected by genomic fingerprinting. Our results confirmed the efficacy of SNP analyses for revealing incorrectly inferred relationships between individuals and for identifying previously unknown relationships as shown previously (Munoz et al. 2014; Tan et al. 2017; Lenz et al. 2020; Thumma et al. 2022). It remains tricky to define suitable thresholds for parentage errors in cases of large variances of pairwise relatedness estimators (Blouin 2003), as observed here within FS families. However, individuals for which more than 40% of G matrix coefficients deviated by more than 0.20 from the expected A coefficient were considered to have parentage errors. Our conservative approach allowed the most glaring pedigree errors to be corrected, thereby keeping the risk of false correction low.

In most forest tree breeding programmes, genetic evaluations are performed with a BLUP analysis based on the mixed model methodology (Henderson 1975), and genetic covariances are expressed in the pedigree-based relationship matrix (the A matrix). EBV accuracies are, therefore, highly dependent on the correctness of the documented pedigree. In practice, the A matrix does not provide information about all existing relationships, whereas the G matrix can reveal undocumented relationships, pedigree errors, and capture the variation arising from Mendelian sampling (Powell et al. 2010). Both hidden and incorrectly documented relationships may affect the accuracy of genetic parameters, biasing EBV and, by extension, decreasing GS accuracy if EBV (or dEBV) are used as pseudo-phenotypes in GS methodology. We show here that, in Altri's multigenerational breeding programme, the 11.2% parentage errors revealed by G matrix information had a significant impact on EBV for the three traits studied, highlighting the importance of pedigree checking before running GS models with pseudo-phenotypes. In *Pinus pinaster* polycross trials, EBV was shown to be improved by the use of a pedigree reconstructed from parentage analysis and allowing paternal identification (Vidal et al. 2015). In *Populus nigra*, pedigree-based BLUP models based on a corrected A matrix were found to be more accurate than models based on an uncorrected matrix (Pégard et al. 2020). The removal of pedigree errors had a negligible effect on additive genetic variance structure in *Picea rubens*, but the authors suggested that the population studied may have been too small for the assessment of variance components or that the magnitude of error was too small (Doerksen and Herbinger 2010). As Altri's breeding population remains largely ungenotyped, we can hypothesise that additional yet to be identified parentage errors may still bias *E. globulus* genetic estimates.

We took the incompleteness of genotypic data and the heterogeneity of phenotyping due to the use of data from many trials of different ages into account by using GS models based on pseudo-phenotypes as done in several reports on forest tree species (Resende et al. 2012; Bartholomé et al. 2016; Isik et al. 2016; Thistlethwaite et al. 2019). The pseudo-phenotypes came here from a meta-analysis of 31 trials located in the same breeding zone explaining that GxE interaction was not included in the model. When breeding value is used as a pseudo-phenotype, deregression can improve GS accuracy by reducing estimate shrinkage toward parental means, and by taking into account the heterogeneity of EBV reliability (Garrick et al. 2009). This method generates contrasting

results, according to the deregression process used. In maritime pine, the use of either EBV or dEBV had no effect on GS accuracy (Isik et al. 2016), whereas, in Douglas fir, the use of dEBV taking mean parental effect into account resulted in a much lower accuracy (Thistlethwaite et al. 2019). Despite these conflicting results, many GS studies in forest trees have used dEBV rather than EBV as the pseudo-phenotype in order to take into account phenotypic information from ungenotyped individuals, and is suitable in case of unbalanced data. For the three traits studied here, GS accuracy was estimated with dEBV, which was strongly correlated with EBV (>0.99). This strong correlation may be due to the use of a deregression process without the removal of parental average effect (as many crosses were of unknown paternity), as well as, to the high degree of relatedness in the POP_{GS} and the large number of clonal copies, both of which contributed to the high accuracy of EBV regardless of the generation considered. For $dEBV_I$ and $dEBV_C$, pedigree correction, principally applied to G1 genotypes, did not change GS accuracy (SV predictions in the S0 scenario) or increased it up to 39% for DBH in the $S1_a$ scenario. Munoz et al. (2014) reported higher predictive abilities (from 2% to 5%) for the use of $dEBV_C$ for various traits related to growth and tree architecture in loblolly pine. GS accuracy is commonly evaluated through the random allocation of individuals to either the validation or calibration population. With the S0 scenario encompassing the three generations indifferently, HT, DBH and SV GS accuracies (0.46, 0.60 and 0.48, respectively) were all consistent with published values in *E. globulus* (Durán et al. 2017; Ballesta et al. 2018). Even if the different scenarios implied a low number of genotypes (from 44 in the $S1_a$ scenario to 342 in the S3 scenario), the effective size of POP_{GS} estimated from status number N_S (Lindgren et al. 1996) was 34. This limited effective size as well as the high marker density (13.3 SNPs / cM in average) were both parameters impacting favourably accuracies of GS (Grattapaglia and Resende 2011). As suggested a study in maritime pine with N_S=25 (Bartholomé et al. 2016), GS accuracy resulted more from the high relatedness between the CP and the VP than from historical LD associations between markers and QTLs.

GS accuracy must be evaluated over generations to determine the value of GS for advanced breeding programmes. Through successive generations of a breeding process, genetic recombination between haplotypes may change the extent of LD, thereby limiting the efficacy of GS over several generations, as LD must be conserved between the CP and VP (Hayes et al. 2009a). The impact of such changes in LD over generations has mostly been investigated in simulation studies. For *Eucalyptus*, Denis and Bouvet (2013) found that GS accuracy decreased over successive breeding cycles for a breeding population with an effective size of N_e=100. As a means of coping with the loss of GS predictive ability across generations, predictive models should be refreshed by aggregating data from the two most recent breeding cycles, as in the simulations for oil palm performed by Cros et al. (2018). We investigated GS over generations with different scenarios (S1 and S2) in which the CP consisted of individuals from previous generations (G0, G1), used to make predictions for the most recent generations (G1 and G2). We obtained moderate-to-high prediction accuracies, with a value of 0.68 for DBH in the $S2_a$ scenario, indicating that it was possible to predict G2 genotypes from data for the parents (G1) and grandparents (G0) in the CP. Moreover growth trait prediction remained similar to that for conventional cross-validation (S0), suggesting that the predictive model was not altered over generations. This conclusion is consistent with other reports for a five consecutive progeny set in *Hordeum vulgare* (Sallam et al. 2015) and in *Avena sativa* L. (Asoro et al. 2011). Similarly, Bartholomé et al. (2016) showed, in a study on *Pinus pinaster,* that high accuracies (0.70 for height and 0.79 for circumference) could be obtained with only G0 and G1 genotypes for the CP, and G2 genotypes for the VP. In an F1 progeny test on Douglas fir, the accuracy of GS for juvenile height was evaluated at 0.9 in the F2

generation (Thistlethwaite et al. 2019). Here, we found that accuracy for the prediction of survival was slightly lower for the per-generation scenario (S1 and S2) than for S0. Moreover, scenarios $S2_c$ and $S2_d$, calibrated with $dEBV_{C\text{-}T01}$, also provided strongly negative mean accuracies for SV prediction in G2 (-0.47 and -0.56, respectively). We can hypothesise that abiotic and/or biotic constraints affected the first two generations (G0 and G1) differently from the last generation (G2), making it difficult to predict complex traits, such as survival, with a CP and VP encompassing different environmental conditions. Accuracy decrease was also observed for growth traits when pseudo-phenotypes included in CP were poorly estimated (scenarios $S2_c$ and $S2_d$ vs. scenarios $S2_a$ and $S2_b$). Interestingly, the addition of progeny genotypes (scenarios $S1_b$, $S3_a$ and $S3_b$) improved the prediction of both G1 and G2 genotypes for all the traits studied, to 0.65 for HT and 0.78 for DBH in the $S3_b$ scenario, and for SV, with an optimum of 0.59. This result was expected, as higher levels of relatedness between calibration and validation populations has been shown to improve GS accuracy in other species, such as *Picea glauca* (Beaulieu et al. 2014).

Conclusion

In conclusion, we report here encouraging results for applied GS in *Eucalyptus globulus*. Given the relatively small population size, we were able to predict the breeding value of the most recent generation reasonably accurately, by aggregating data from the first two generations. In addition, pedigree correction for identity and parentage errors increased the accuracy of GS for all traits. Including a few relatives from targeted families in GS models also improved accuracy for all traits. Further investigations in *E. globulus* are required, particularly as concerning optimisation of the calibration and validation populations, as proposed in GS approaches for other species (Ahmadi and Bartholomé 2022). As POP_{GS} was not representative from the breeding population of Altri Florestal (POP_{TOT}), this study must be considered as a proof-of-concept. Genotyping efforts in this breeding population will need to continue before implementing concretely GS. In addition, the genotyping of the base population could be compared with the 13 races and eight genetic groups defined in previous studies (Dutkowski and Potts 1999; Costa et al. 2017) to define meta-founders usable for GS (Legarra et al. 2015). An alternative to deregressed EBV for taking the performance of non-genotyped individuals into account would be so-called "single-step genomic BLUP" (Legarra et al. 2014). This methodology has recently been successfully tested in *E. globulus* (Callister et al. 2021; Quezada et al. 2022) as the training population size can be increased by including both genotyped and not genotyped individuals.

Funding : This work was funded by CIRAD (Centre de coopération internationale en recherche agronomique pour le développement) and by the European Union's Horizon 2020 research and innovation programme under grant agreement n°773383 (B4EST). In this context, G. Haristoy received a doctoral fellowship.

Acknowledgements: Thanks to J. Ventura (Altri Florestal) who carried out work related to phenotyping, leaves sampling and handling, as well as, sending all samples to France. DNA extractions were performed at the BIOGECO research unit (Cestas, France) with the valuable advice of C. Lalanne (INRAE) and C. Larue (INRAE). We also thank H. San Clemente (CNRS), for the additional bioinformatic analysis during the reviewing process, revealing putative genes with polymorphic SNP. DNA quantification was performed at the Genome Transcriptome Facility of Bordeaux (grants from the Conseil Regional d'Aquitaine: n°20030304002FA and 20040305003FA, the European Union: FEDER n°2003227 and ANR: n°ANR-10-EQPX-16 Xyloforest). Genotyping was performed by Thermo Fisher Scientific (Santa Clara, USA). We are also grateful to J. Bartholomé (CIRAD), D. Cros (CIRAD), C. Dubos (INRAE), D. Lopez (CIRAD), P. Rozenberg (INRAE), and D. This (INRAE, L'Institut Agro) for helpful discussions.

Conflicts of interest: The authors have no conflict of interest to declare.

Authors' contributions: Conceptualization and PhD supervision: LB and JMG. Management of the breeding programme and curation of field data: LF, LL, JPP. Laboratory work: GH. Data analysis: LB, JMG, GH. Drafting the manuscript: LB, JMG, GH. Critical revision of the manuscript: LF, LL, JAPP, JPP. All authors have approved the publication of the final version of the manuscript.

Data archiving statement: Genomic, pedigree, and pseudo-phenotypic data will be deposited to the Data INRAE portal: https://data.inrae.fr/. The supplemental data "Marker_data.csv" contains the molecular profiles of the 401 genotypes of POP_{GS} (individuals in row, SNP in columns). The pedigree of POP_{GS} genotypes is available in the supplemental data file "Pedigree_data.csv". The first column contains the identifier of each individual, the second and third columns refer to the mother and father documented in the initial pedigree, and the fourth and fifth columns contain the mother and father after pedigree corrections. The pseudo-phenotypic data are available in 12 different ".csv" files (4 files for each trait), containing either $dEBV_I$, $dEBV_C$, $dEBV_{C\text{-}T01}$, $dEBV_{C\text{-}T2}$ (refers to the file name). The first column is the identifier of the individuals, and the second column contains the deregressed estimated breeding value.

References

Adams WT, Neale DB, Loopstra CA (1988) Verifying controlled crosses in conifer tree-improvement programs. Silvae Genet 37:147–152

Ahmadi N, Bartholomé J (eds) (2022) Complex Trait Prediction: Methods and Protocols. Springer US, New York

Ahmar S, Ballesta P, Ali M, Mora-Poblete F (2021) Achievements and Challenges of Genomics-Assisted Breeding in Forest Trees: From Marker-Assisted Selection to Genome Editing. Int J Mol Sci 22:10583. https://doi.org/10.3390/ijms221910583

Amadeu RR, Cellon C, Olmstead JW, et al (2016) AGHmatrix: R Package to Construct Relationship Matrices for Autotetraploid and Diploid Species: A Blueberry Example. Plant Genome 9:1. https://doi.org/10.3835/plantgenome2016.01.0009

Asoro FG, Newell MA, Beavis WD, et al (2011) Accuracy and Training Population Design for Genomic Selection on Quantitative Traits in Elite North American Oats. Plant Genome 4:132. https://doi.org/10.3835/plantgenome2011.02.0007

Ballesta P, Serra N, Guerra F, et al (2018) Genomic Prediction of Growth and Stem Quality Traits in *Eucalyptus globulus* Labill. at Its Southernmost Distribution Limit in Chile. Forests 9:779. https://doi.org/10.3390/f9120779

Bartholomé J, Mabiala A, Burlett R, et al (2020) The pulse of the tree is under genetic control: eucalyptus as a case study. Plant J 103:338–356. https://doi.org/10.1111/tpj.14734

Bartholomé J, Mandrou E, Mabiala A, et al (2015) High- resolution genetic maps of *Eucalyptus* improve *Eucalyptus grandis* genome assembly. New Phytol 206:1283–1296. https://doi.org/10.1111/nph.13150

Bartholomé J, Salmon F, Vigneron P, et al (2013) Plasticity of primary and secondary growth dynamics in *Eucalyptus* hybrids: a quantitative genetics and QTL mapping perspective. BMC Plant Biol 13:120. https://doi.org/10.1186/1471-2229-13-120

Bartholomé J, Van Heerwaarden J, Isik F, et al (2016) Performance of genomic prediction within and across generations in maritime pine. BMC Genom 17:604. https://doi.org/10.1186/s12864-016-2879-8

Beaulieu J, Doerksen T, Clément S, et al (2014) Accuracy of genomic selection models in a large population of open-pollinated families in white spruce. Heredity 113:343–352. https://doi.org/10.1038/hdy.2014.36

Blouin MS (2003) DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. Trends Ecol Evol 18:503–511. https://doi.org/10.1016/S0169-5347(03)00225-8

Borralho NMG (1995) The impact of individual tree mixed models (BLUP) in tree breeding. In: BM Potts, NMG Borralho, JB Reid, RN Cromer, WN Tibbits and CA Raymond (eds) Proc CRCTHF-IUFRO Conference Eucalypt Plantation: Improving Fibre Yield And Quality, Hobart, Tasmania, pp 141-145

Borralho NMG, Almeida MH, Potts BM (2007) O melhoramento do eucalipto em Portugal. In: AM Alves and JS Pereira and JMN Silva (eds), Eucaliptal em Portugal: Impactes Ambientais e Investigação Científica, ISAPress, Lisbon, Portugal, pp 61-110

Borralho NMG, Pina JP, Leal L, Araujo J (2018) The gain achieved from *Eucalyptus globulus* tree improvement programs in Portugal, a joint analysis of RAIZ and ALTRI trials. In: Proc Tecnicelpa XXIV Internacional Forest, Pulp and Paper Conference, Porto, Portugal

Butler DG, Cullis BR, Gilmour AR, et al (2017) ASReml estimates variance components under a general linear. VSN International Ltd

Callister AN, Bradshaw BP, Elms S, et al (2021) Single-step genomic BLUP enables joint analysis of disconnected breeding programs: an example with *Eucalyptus globulus* Labill. G3: Genes, Genomes, Genet 11:jkab253. https://doi.org/10.1093/g3journal/jkab253

Chambers PGS, Borralho NMG, Potts BM (1996) Genetic Analysis of Survival in *Eucalyptus globulus ssp. globulus*. Silvae Genet 45:107–112

Costa e Silva F, Shvaleva A, Broetto F, et al (2008) Acclimation to short-term low temperatures in two *Eucalyptus globulus* clones with contrasting drought resistance. Tree Physiol 29:77–86. https://doi.org/10.1093/treephys/tpn002

Costa J, Vaillancourt RE, Steane DA, et al (2017) Microsatellite analysis of population structure in *Eucalyptus globulus*. Genome 60:770–777. https://doi.org/10.1139/gen-2016-0218

Cros D, Tchounke B, Nkague-Nkamba L (2018) Training genomic selection models across several breeding cycles increases genetic gain in oil palm in silico study. Molecular Breeding 38:89. https://doi.org/10.1007/s11032-018-0850-x

Denis M, Bouvet J-M (2013) Efficiency of genomic selection with models including dominance effect in the context of *Eucalyptus* breeding. Tree Genet Genomes 9:37–51. https://doi.org/10.1007/s11295-012-0528-1

Doerksen TK, Herbinger CM (2008) Male reproductive success and pedigree error in red spruce open-pollinated and polycross mating systems. Can J For Res 38:1742–1749. https://doi.org/10.1139/X08-025

Doerksen TK, Herbinger CM (2010) Impact of reconstructed pedigrees on progeny-test breeding values in red spruce. Tree Genet Genomes 6:591–600. https://doi.org/10.1007/s11295-010-0274-1

Doyle J (1991) DNA Protocols for Plants. In: Hewitt GM, Johnston AWB, Young JPW (eds) Molecular Techniques in Taxonomy. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 283–293

Durán R, Isik F, Zapata-Valenzuela J, et al (2017) Genomic predictions of breeding values in a cloned *Eucalyptus globulus* population in Chile. Tree Genet Genomes 13:74. https://doi.org/10.1007/s11295-017-1158-4

Dutkowski GW, Potts BM (1999) Geographic Patterns of Genetic Variation in *Eucalyptus globulus ssp. globulus* and a Revised Racial Classification. Aust J Bot 47:237. https://doi.org/10.1071/BT97114

Ericsson (1999) The effect of pedigree error by misidentification of individual trees on genetic evaluation of a full-sib experiment. Silvae Genet 48:239-242.

Freeman JS, Potts BM, Downes GM, et al (2013) Stability of quantitative trait loci for growth and wood properties across multiple pedigrees and environments in *Eucalyptus globulus*. New Phytol 198:1121–1134. https://doi.org/10.1111/nph.12237

Gamal El-Dien O, Ratcliffe B, Klápště J, et al (2016) Implementation of the Realized Genomic Relationship Matrix to Open-Pollinated White Spruce Family Testing for Disentangling Additive from Nonadditive Genetic Effects. G3: Genes, Genomes, Genet 6:743–753. https://doi.org/10.1534/g3.115.025957

Garrick DJ, Taylor JF, Fernando RL (2009) Deregressing estimated breeding values and weighting information for genomic regression analyses. Genet, Sel, Evol 41:55. https://doi.org/10.1186/1297-9686-41-55

Grattapaglia D (2017) Status and Perspectives of Genomic Selection in Forest Tree Breeding. In: Varshney RK, Roorkiwal M, Sorrells ME (eds) Genomic Selection for Crop Improvement. Springer International Publishing, Cham, pp 199–249

Grattapaglia D, Resende MDV (2011) Genomic selection in forest tree breeding. Tree Genet Genomes 7:241–255. https://doi.org/10.1007/s11295-010-0328-4

Gu Z, Eils R, Schlesner M (2016) Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics 32:2847–2849. https://doi.org/10.1093/bioinformatics/btw313

Hamilton MG, Acuna M, Wiedemann JC, et al (2015) Genetic control of *Eucalyptus globulus* harvest traits. Can J For Res 45:615–624. https://doi.org/10.1139/cjfr-2014-0428

Hamilton MG, Potts BM, Greaves BL, Dutkowski GW (2010) Genetic correlations between pulpwood and solid-wood selection and objective traits in *Eucalyptus globulus*. Ann For Sci 67:511–511. https://doi.org/10.1051/forest/2010013

Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME (2009a) Invited review: Genomic selection in dairy cattle: Progress and challenges. J Dairy Sci 92:433–443. https://doi.org/10.3168/jds.2008-1646

Hayes BJ, Visscher PM, Goddard ME (2009b) Increased accuracy of artificial selection by using the realized relationship matrix. Genet Res 91:47–60. https://doi.org/10.1017/S0016672308009981

Henderson C (1950) Estimation of Genetic Parameters. Ann Math Stat 309–310

Henderson CR (1975) Best Linear Unbiased Estimation and Prediction under a Selection Model. Biometrics 31:423. https://doi.org/10.2307/2529430

Hill WG, Weir BS (2011) Variation in actual relationship as a consequence of Mendelian sampling and linkage. Genet Res 93:47–64. https://doi.org/10.1017/S0016672310000480

Hudson CJ, Freeman JS, Kullan AR, et al (2012) A reference linkage map for *Eucalyptus*. BMC Genom 13:240. https://doi.org/10.1186/1471-2164-13-240

ICNF (2019) 6° Inventário Florestal Nacional - Relatório Final. Instituto da Conservação da Natureza e das Florestas Lisboa, Portugal

Isik F (2014) Genomic selection in forest tree breeding: the concept and an outlook to the future. New For 45:379–401. https://doi.org/10.1007/s11056-014-9422-z

Isik F, Bartholomé J, Farjat A, et al (2016) Genomic selection in maritime pine. Plant Sci 242:108–119. https://doi.org/10.1016/j.plantsci.2015.08.006

Isik F, Holland J, Maltecca C (2017) Genetic Data Analysis for Plant and Animal Breeding. Springer International Publishing, Cham

Jarvis SF, Borralho NMG, Potts BM (1995) Implementation of a Multivariate BLUP Model for Genetic Evaluation of *Eucalyptus globulus* in Australia. In: BM Potts, NMG Borralho, JB Reid, RN Cromer, WN Tibbits, and CA Raymond (eds) Proc CRCTHF–IUFRO Conference Eucalypt Plantation: Improving Fibre Yield And Quality, Hobart, Tasmania, pp 212-216

Jones TH, Steane DA, Jones RC, et al (2006) Effects of domestication on genetic diversity in *Eucalyptus globulus*. For Ecol Manage 234:78–84. https://doi.org/10.1016/j.foreco.2006.06.021

Keil M, Griffin AR (1994) Use of random amplified polymorphic DNA (RAPD) markers in the discrimination and verification of genotypes in Eucalyptus. Theor Appl Genet 89:442–450. https://doi.org/10.1007/BF00225379

Klápště J, Lstibůrek M, El-Kassaby YA (2014) Estimates of genetic parameters and breeding values from western larch open-pollinated families using marker-based relationship. Tree Genet Genomes 10:241–249. https://doi.org/10.1007/s11295-013-0673-1

Kumar S, Richardson TE (2005) Inferring relatedness and heritability using molecular markers in radiata pine. Mol Breed 15:55–64. https://doi.org/10.1007/s11032-004-2059-4

Lebedev VG, Lebedeva TN, Chernodubov AI, Shestibratov KA (2020) Genomic Selection for Forest Tree Improvement: Methods, Achievements and Perspectives. Forests 11:1190. https://doi.org/10.3390/f11111190

Legarra A, Christensen OF, Aguilar I, Misztal I (2014) Single Step, a general approach for genomic selection. Livest Sci 166:54–65. https://doi.org/10.1016/j.livsci.2014.04.029

Legarra A, Christensen OF, Vitezica ZG, et al (2015) Ancestral Relationships Using Metafounders: Finite Ancestral Populations and Across Population Relationships. Genetics 200:455–468. https://doi.org/10.1534/genetics.115.177014

Lenz PRN, Nadeau S, Azaiez A, et al (2020) Genomic prediction for hastening and improving efficiency of forward selection in conifer polycross mating designs: an example from white spruce. Heredity 124:562–578. https://doi.org/10.1038/s41437-019-0290-3

Li Y, Dungey HS (2018) Expected benefit of genomic selection over forward selection in conifer breeding and deployment. PLoS ONE 13:208–232. https://doi.org/10.1371/journal.pone.0208232

Lindgren D, Gea LD, Jefferson PA (1996) Loss of genetic diversity monitored by status number. Silvae Genet 45:52–59

Lopez GA, Potts BM, Dutkowski G, et al (2002) Genetic variation and inter-trait correlations in *Eucalyptus globulus* base population trials in Argentina. For Genet 9:217–231

MAPA (2019) Anuário de Estadística Forestal 2019. Ministerio de Agricultura, Pesca y Alimentación, Madrid, Spain

Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. Genetics 157:1819–1829. https://doi.org/10.1093/genetics/157.4.1819

Mora F, Serra N (2014) Bayesian estimation of genetic parameters for growth, stem straightness, and survival in *Eucalyptus globulus* on an Andean Foothill site. Tree Genet Genomes 10:711–719. https://doi.org/10.1007/s11295-014-0716-2

Mrode RA (2013) Linear models for the prediction of animal breeding values, 3rd ed. CABI, Boston

Munoz F, Rodriguez LS (2020) breedR: Statistical Methods for Forest Genetic Resources Analysts. R package version 0.12-5

Munoz PR, Resende MFR, Huber DA, et al (2014) Genomic Relationship Matrix for Correcting Pedigree Errors in Breeding Populations: Impact on Genetic Parameters and Genomic Selection Accuracy. Crop Sci 54:1115–1123. https://doi.org/10.2135/cropsci2012.12.0673

Myburg AA, Grattapaglia D, Tuskan GA, et al (2014) The genome of *Eucalyptus grandis*. Nature 510:356–362. https://doi.org/10.1038/nature13308

Pégard M, Segura V, Muñoz F, et al (2020) Favorable Conditions for Genomic Evaluation to Outperform Classical Pedigree Evaluation Highlighted by a Proof-of-Concept Study in Poplar. Front Plant Sci 11:581954. https://doi.org/10.3389/fpls.2020.581954

Potts BM, Vaillancourt RE, Jordan G, et al (2004) Exploration of the *Eucalyptus globulus* gene pool. In: Borralho NMG, Pereira JS, Marques CMP, Coutinho J, Madeira M, Tomé M (eds) *Eucalyptus* in a changing world Proc IUFRO Conference, Aveiro, Portugal, pp 46–61

Powell JE, Visscher PM, Goddard ME (2010) Reconciling the analysis of IBD and IBS in complex trait studies. Nat Rev Genet 11:800–805. https://doi.org/10.1038/nrg2865

Przybylski P, Kowalczyk J, Odrzykoski I, Matras J (2019) Identyfing alien genotypes and their consequences for genetic variation in clonal seed orchards of *Pinus sylvestris* L. Dendrobiology 81:40–46. https://doi.org/10.12657/denbio.081.005

Quezada M, Aguilar I, Balmelli G (2022) Genomic breeding values' prediction including populational selfing rate in an open-pollinated Eucalyptus globulus breeding population. Tree Genet Genomes 18:10. https://doi.org/10.1007/s11295-021-01534-7

Raymond CA (2002) Genetics of *Eucalyptus* wood properties. Ann For Sci 59:525–531. https://doi.org/10.1051/forest:2002037

Resende MDV, Resende MFR, Sansaloni CP, et al (2012) Genomic selection for growth and wood quality in *Eucalyptus*: capturing the missing heritability and accelerating breeding for complex traits in forest trees. New Phytol 194:116–128. https://doi.org/10.1111/j.1469-8137.2011.04038.x

Rezende GDSP, de Resende MDV, de Assis TF (2014) *Eucalyptus* Breeding for Clonal Forestry. In: Fenning T (ed) Challenges and Opportunities for the World's Forests in the 21st Century. Springer Netherlands, Dordrecht, pp 393–424

Rojas PV (2017) Breeding *Eucalyptus globulus* for lower rainfall sites in the Bío-Bío Region of Chile. Australian Forestry. https://doi.org/10.1080/00049158.2017.1319260

Rstudio Team (2021) RStudio: Integrated Development for R. RStudio, Inc., Boston

Salas M, Nieto V, Perafán L, et al (2014) Genetic parameters and comparison between native and local landraces of *Eucalyptus globulus Labill. ssp. globulus* growing in the central highlands of Colombia. Ann For Sci 71:405–414. https://doi.org/10.1007/s13595-013-0342-4

Sallam AH, Endelman JB, Jannink J-L, Smith KP (2015) Assessing Genomic Selection Prediction Accuracy in a Dynamic Barley Breeding Population. Plant Genome 8:plantgenome2014.05.0020. https://doi.org/10.3835/plantgenome2014.05.0020

Sinnwell JP, Therneau TM, Schaid DJ (2014) The kinship2 R Package for Pedigree Data. Human Heredity 78:91–93. https://doi.org/10.1159/000363105

Tan B, Grattapaglia D, Martins GS, et al (2017) Evaluating the accuracy of genomic prediction of growth and wood traits in two *Eucalyptus* species and their F1 hybrids. BMC Plant Biol 17:110. https://doi.org/10.1186/s12870-017-1059-6

Tan B, Grattapaglia D, Wu HX, Ingvarsson PK (2018) Genomic relationships reveal significant dominance effects for growth in hybrid *Eucalyptus*. Plant Science 267:84–93. https://doi.org/10.1016/j.plantsci.2017.11.011

Thavamanikumar S, Arnold RJ, Luo J, Thumma BR (2020) Genomic Studies Reveal Substantial Dominant Effects and Improved Genomic Predictions in an Open-Pollinated Breeding Population of *Eucalyptus pellita*. G3: Genes, Genomes, Genetics 10:3751–3763. https://doi.org/10.1534/g3.120.401601

Thistlethwaite FR, Ratcliffe B, Klápště J, et al (2019) Genomic selection of juvenile height across a single-generational gap in Douglas-fir. Heredity 122:848–863. https://doi.org/10.1038/s41437-018-0172-0

Thumma BR, Joyce KR, Jacobs A (2022) Genomic studies with preselected markers reveal dominance effects influencing growth traits in *Eucalyptus nitens*. G3: Genes, Genomes, Genetics 12:jkab363. https://doi.org/10.1093/g3journal/jkab363

VanRaden PM (2008) Efficient Methods to Compute Genomic Predictions. Journal of Dairy Science 91:4414–4423. https://doi.org/10.3168/jds.2007-0980

Vidal M, Plomion C, Harvengt L, et al (2015) Paternity recovery in two maritime pine polycross mating designs and consequences for breeding. Tree Genet Genomes 11:105. https://doi.org/10.1007/s11295-015-0932-4

Volker PW, Dean CA, Tibbits WN, Ravenwook IC (1998) Genetic parameters and gains expected from selection in *Eucalyptus globulus* in Tasmania. Silvae Genet 18–21

Wickham H (2016) ggplot2. Springer International Publishing, Cham

Figure captions:

Fig. 1 Initial pedigree of the 412 POP_{TOT} genotypes selected for genotyping (blue dots) (grey dots represent ancestors not available from clonal archives and lines indicate parent-progeny relationships documented in the initial pedigree)

Fig. 2 Heatmap of the A coefficients (under the red diagonal) *vs.* G coefficients (above the diagonal) of 401 individuals from POP_{GS}. Two A matrix coefficients are shown: A_I coefficients without the correction of pedigree errors (a), and A_C coefficients with pedigree correction (b). In both cases, the 401 genotypes were ordered by generation, from top to bottom, and left to right (G0, G1, G2)

Fig. 3 Distribution of G coefficients within two A classes: 0.25 (a and c) and 0.5 (b and d). The upper two histograms (a and b) show G distributions based on the initial pedigree (A_I matrix), and the lower two histograms (c and d) show G distributions after pedigree correction (A_C matrix)

Fig. 4 Distribution of non-concordant SNPs number in the parent-progeny relationships (P/P).
The threshold (115 non-concordant SNPs) below which the parent of an individual is considered "true" is represented by the red dotted line

Fig. 5 Distribution of the percentage of half-sib and full-sib relationships differing from more 0.2 between A and G matrices for genotypes involved in "undetermined" P/P. The threshold (40%) above which the pedigree was considered inconsistent is represented by the dotted line.
a) A_I was considered i.e. initial documented pedigree b) A_C was considered i.e. pedigree after corrections

Fig. 6 Genetic correlations between EBV_C for tree height (HT-EBV_C), diameter at breast height (DBH-EBV_C) and survival (SV-EBV_C) across the whole POP_{GS} sample (in black) and in the three generations (G0 in red, G1 in green, and G2 in blue) of POP_{GS}

Fig. 7 GS accuracy of the S0 scenario for height (HT), diameter at breast height (DBH) and survival (SV), with either $dEBV_I$ (in grey) or $dEBV_C$ (in red) used as a pseudo-phenotype (the means are indicated by coloured dots)

Fig. 8 Accuracy of GS models for height (HT), diameter at breast height (DBH) and survival (SV) according to the nine scenarios tested: S0 (in red), $S1_a$ and $S1_b$ (green), $S2_a$, $S2_b$, $S2_c$, and $S2_d$ (blue), and $S3_a$ and $S3_b$ (purple). In all scenarios, deregressed and corrected EBV ($dEBV_C$) were used as pseudo-phenotypes. Significance is shown for each assessment of accuracy in Table S4)

Supplementary material captions:

Table S1 Settings for SNP quality control analysis in Axiom Suite Analysis software

Table S2 Descriptive statistics for EBV_C, $EBV_{C\text{-}T01}$, $EBV_{C\text{-}T2}$ and their mean accuracy (r) for tree height (HT), diameter at breast height (DBH) and survival (SV) for each generation of POP_{TOT}

Table S3 Mean accuracy with $dEBV_I$ or $dEBV_C$, for height (HT), diameter at breast height (DBH), and survival (SV), for scenarios S0, $S1_a$ and $S2_a$. (1) In S0, accuracy is the mean of 100 per-iteration accuracies, and the corresponding significance threshold is that for at least 95% of the 100 iterations

Table S4 Significance of GS accuracies by scenario. In cases of iteration (*), the accuracy given is the mean of 100 per-iteration accuracies, and the corresponding significance threshold is the threshold for at least 95% of the 100 iterations. The most globally significant accuracies are shown in bold

Fig. S1 Distribution of G coefficients according to A_I or A_C coefficients in POP_{GS}

Fig. S2 Regression of EBV_I on EBV_C for tree height (A), diameter at breast height (B), and survival (C) for the 401 genotypes of POP_{GS}. The black dots referred to the 356 genotypes without pedigree correction, and the red dots, the 45 genotypes for which the pedigree was corrected. Pearson correlation of EBV_I with EBV_C for the 45 genotypes is indicated in red

Fig. S3 Genetic correlations between EBV_C for tree height (HT-EBV_C), diameter at breast height (DBH-EBV_C) and survival (SV-EBV_C) across the whole POP_{TOT} sample (in black) and in the three generations (G0 in red, G1 in green, and G2 in blue) of POP_{TOT}

Fig. S4 Correlation matrix for the EBV_C, $EBV_{C\text{-}T01}$, and $EBV_{C\text{-}T2}$ estimates for the three traits (HT, DBH and SV) based on the POP_{GS} sample. Correlation coefficients are indicated for the whole POP_{GS} population (in grey) and for each generation (G0 in red, G1 in green and G2 in blue). The significance threshold is indicated as follows: 5% (*), 1% (**) and 0.1% (***)

Figure 1

Click here to access/download;Figure;Fig1.jpg

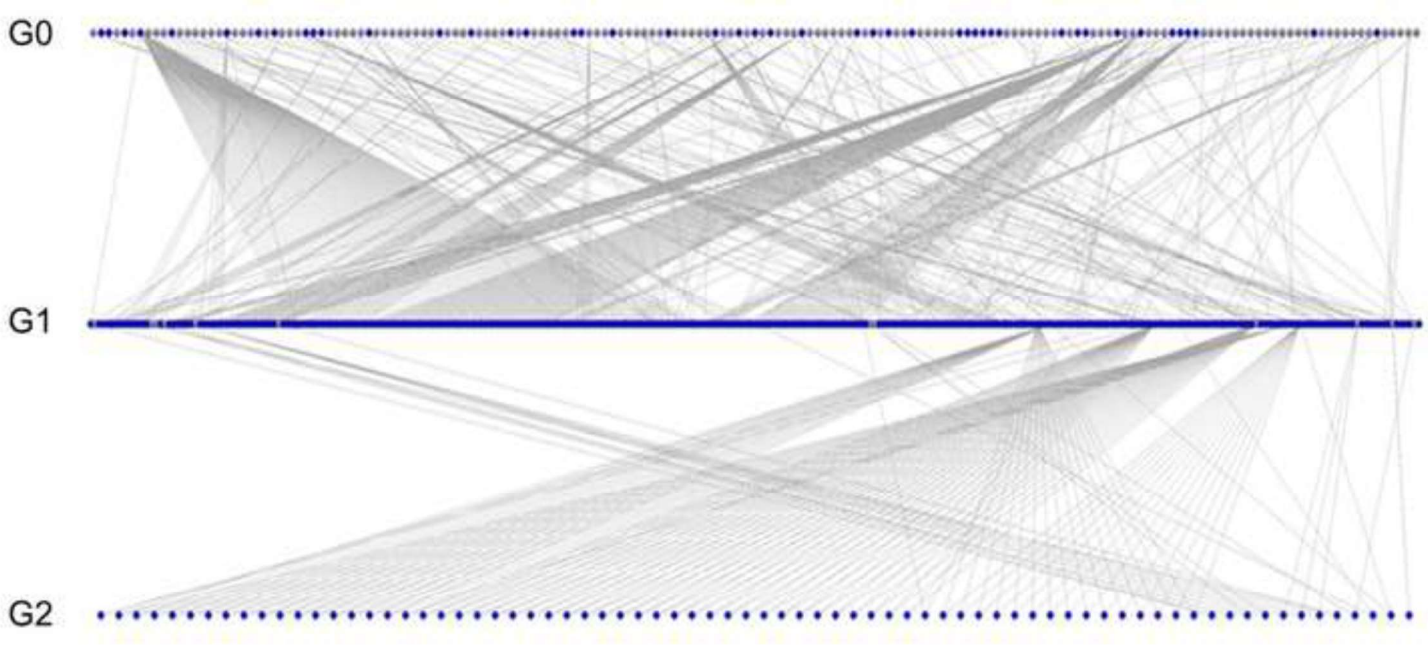

Figure 2

Click here to access/download;Figure;Fig2.jpg

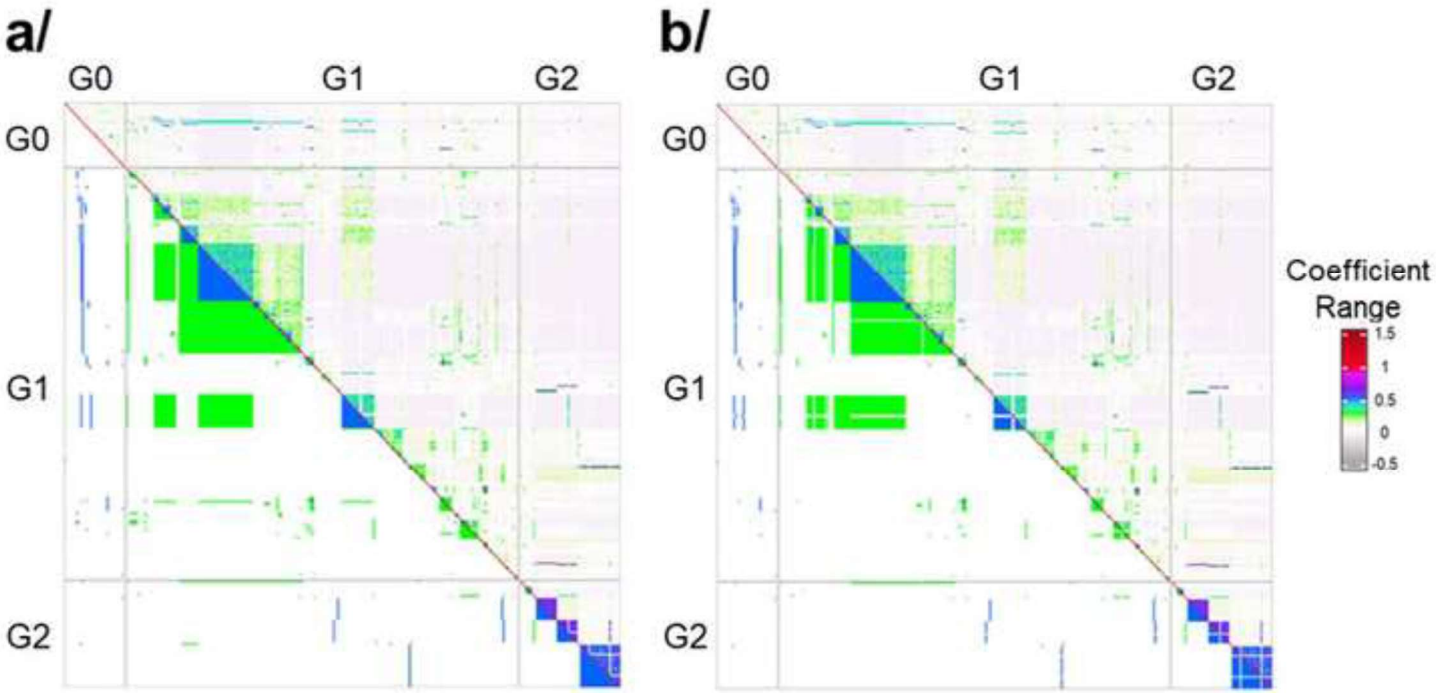

Figure 3

Click here to access/download;Figure;Fig3.jpg

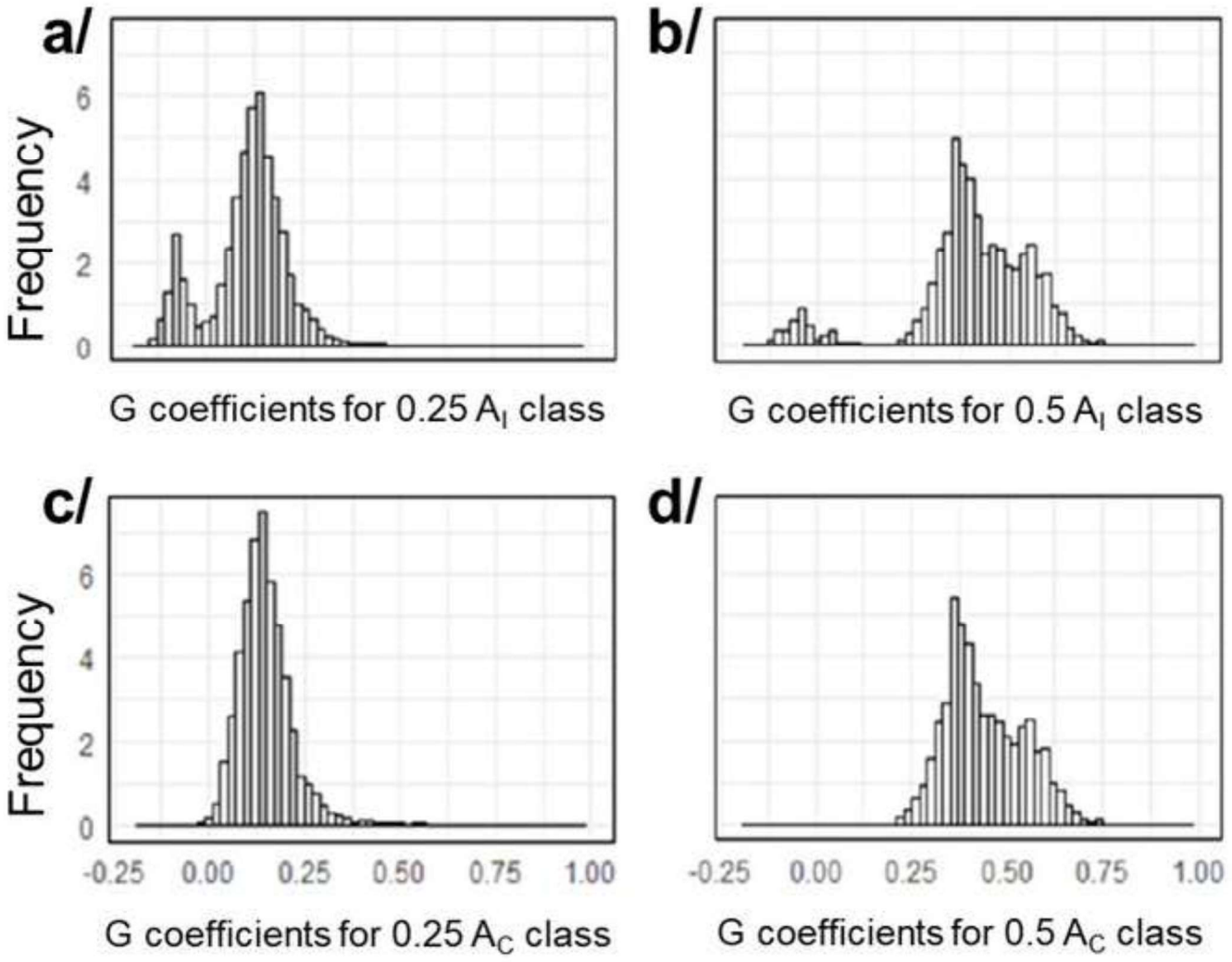

Figure 4

Click here to access/download;Figure;Fig4.jpg

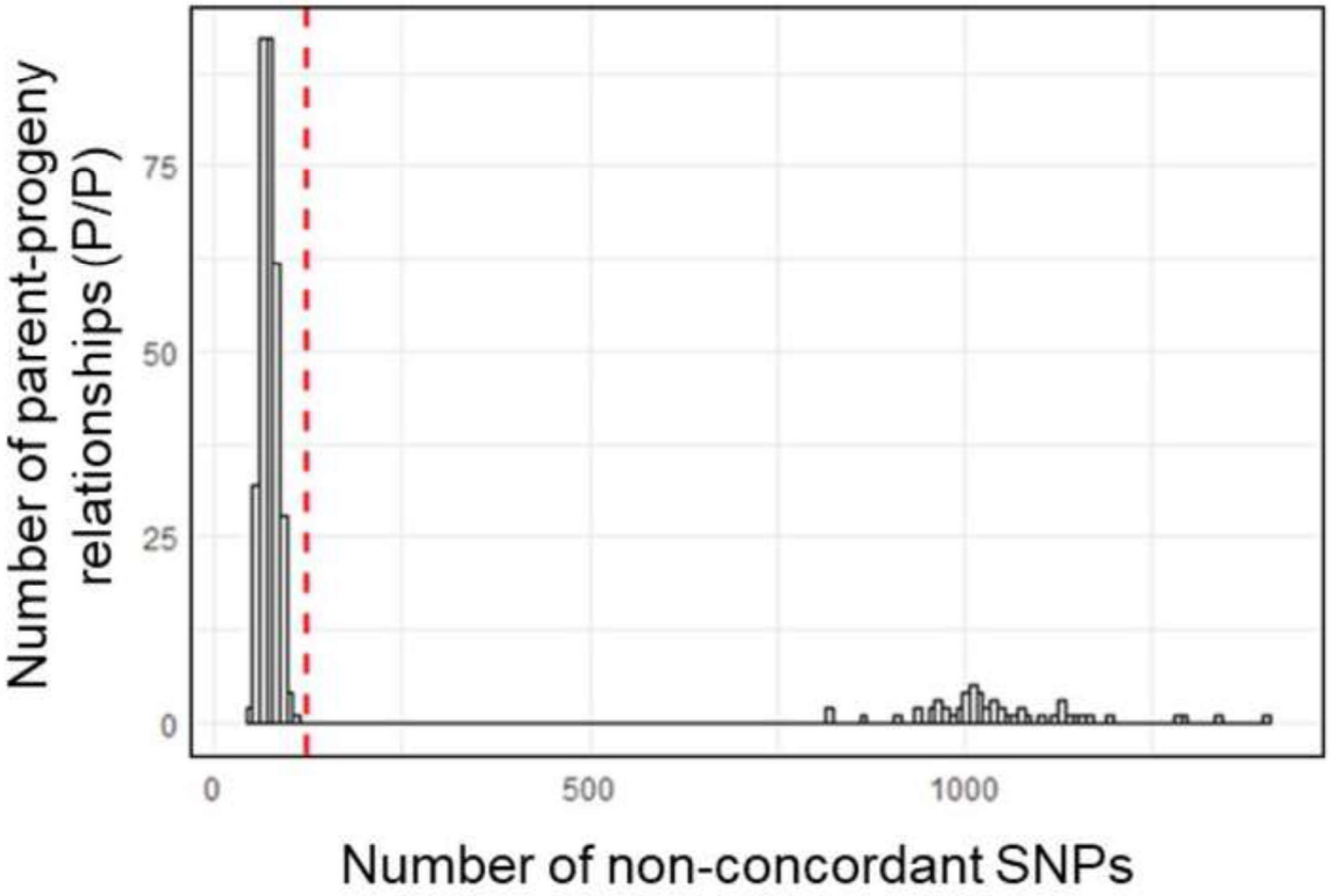

Figure 5

Click here to access/download;Figure;Fig5.jpg

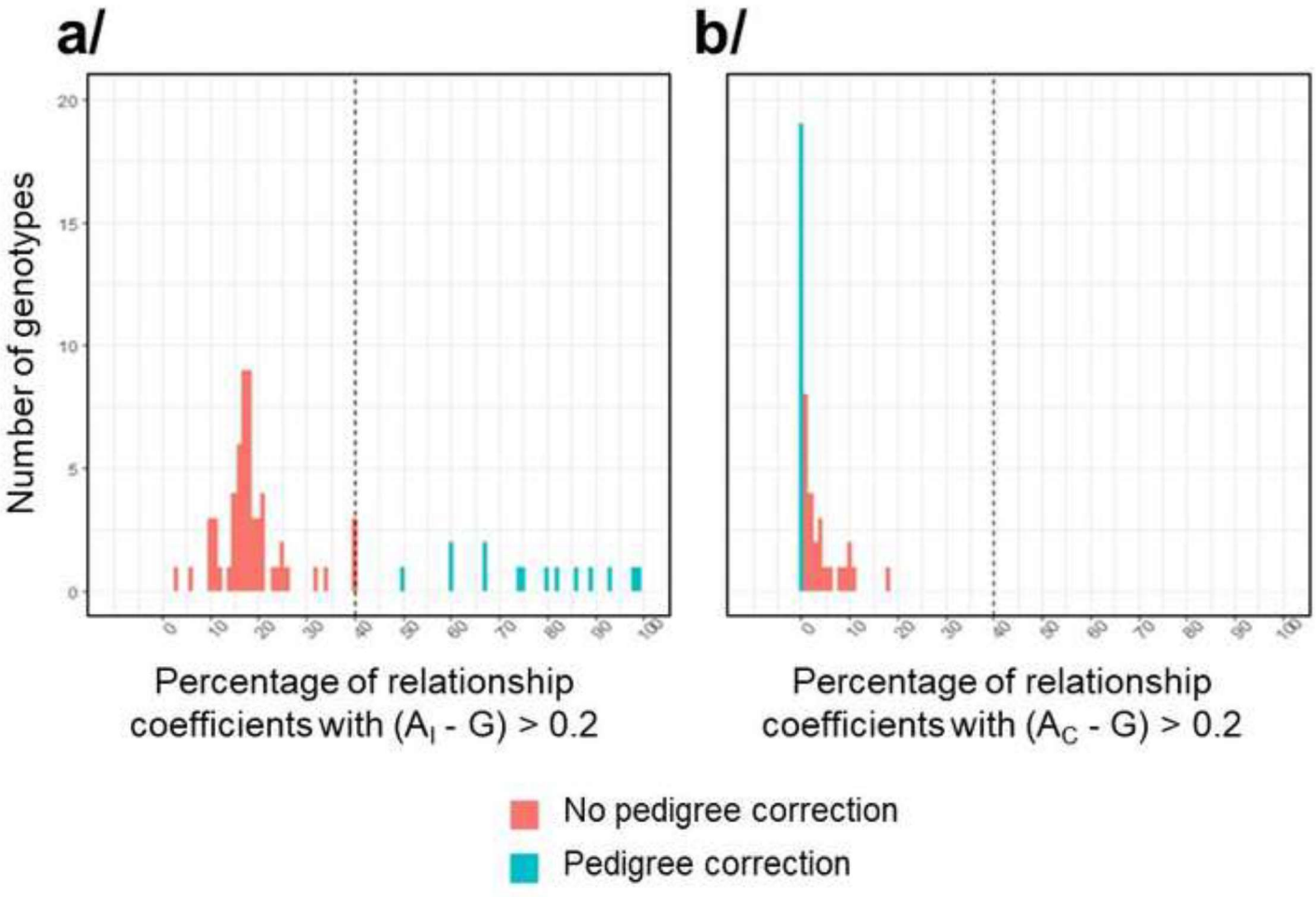

Figure 6

Click here to access/download;Figure;Fig6.jpg

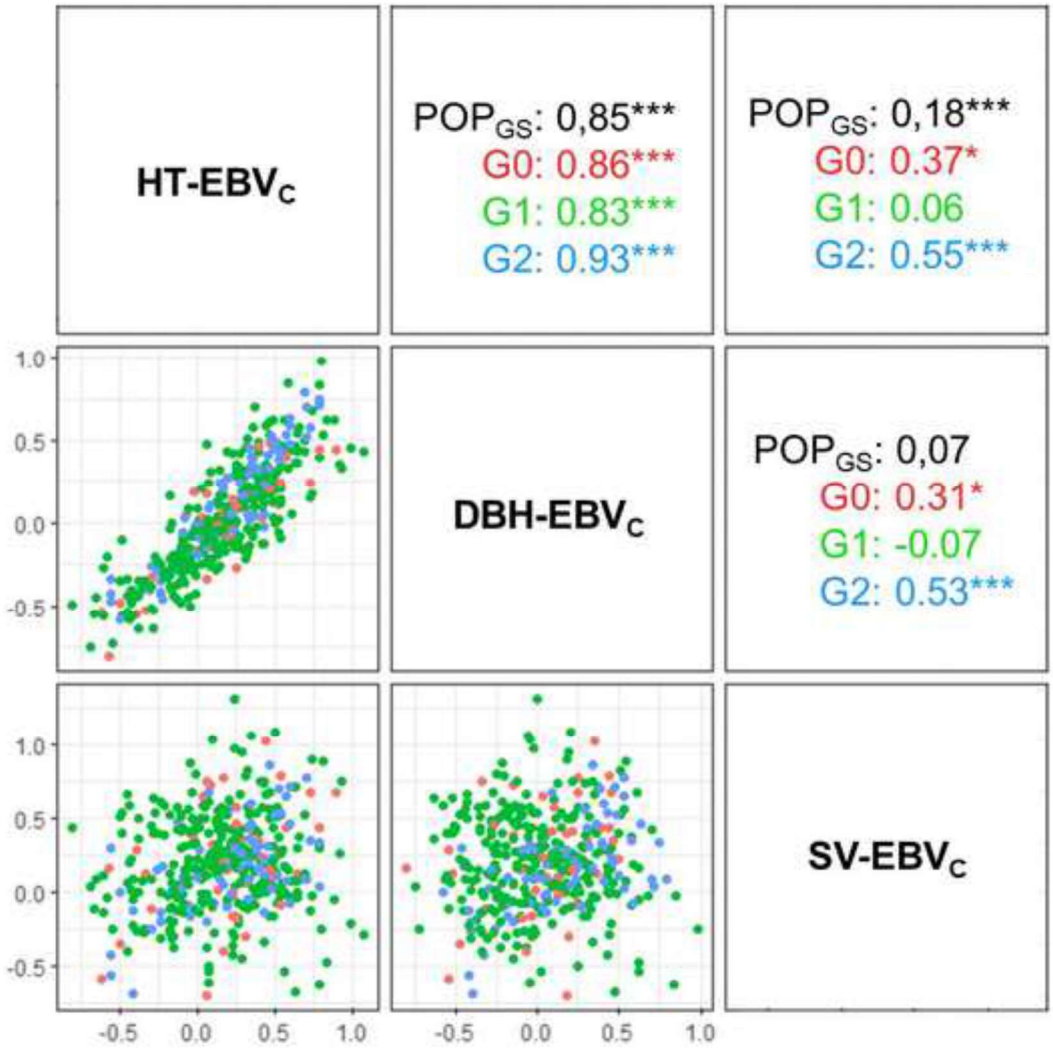

Figure 7

Click here to access/download;Figure;Fig7.jpg

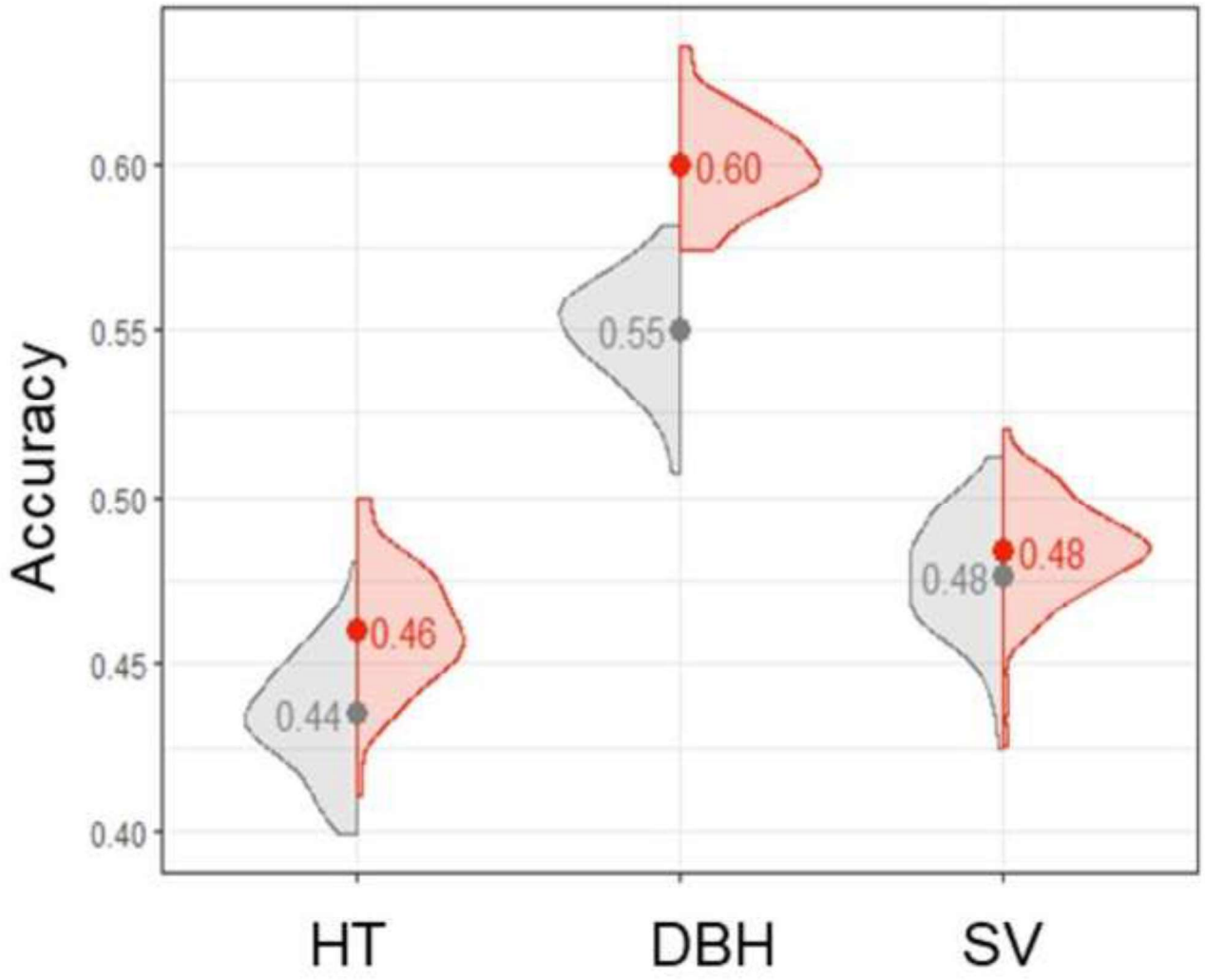

Figure 8

Click here to access/download;Figure;Fig8.jpg