New Phytologist

Independent and Joint-GWAS for growth traits in *Eucalyptus* by assembling genome-wide data for 3373 individuals across four breeding populations

Bárbara S. F. Müller[1,2], Janeo E. de Almeida Filho[3], Bruno M. Lima[4], Carla C. Garcia[5], Alexandre Missiaggia[4], Aurelio M. Aguiar[4], Elizabete Takahashi[6], Matias Kirst[7], Salvador A. Gezan[7], Orzenil B. Silva-Junior[2,8], Leandro G. Neves[9] and Dario Grattapaglia[1,2,8]

[1]Molecular Biology Program, Cell Biology Department, Biological Sciences Institute, University of Brasília, Campus Darcy Ribeiro, Brasília, DF 70910-900, Brazil; [2]EMBRAPA Genetic Resources and Biotechnology – EPqB, Brasília, DF 70770-910, Brazil; [3]Plant Breeding Laboratory, State University of North Fluminense "Darcy Ribeiro", Campos dos Goytacazes, RJ 28013-602, Brazil; [4]FIBRIA S.A. Technology Center, Jacareí, SP 12340-010, Brazil; [5]International Paper of Brazil, Rodovia SP 340 KM 171, Mogi Guaçu, SP 13840-970, Brazil; [6]Celulose Nipo-Brasileira (CENIBRA) S.A., Belo Oriente, MG 35196-000, Brazil; [7]School of Forest Resources and Conservation, University of Florida, Gainesville, FL 32611, USA; [8]Genomic Sciences and Biotechnology Program, SGAN, Catholic University of Brasília, 916 modulo B, Brasília, DF 70790-160, Brazil; [9]RAPiD Genomics LLC, Gainesville, FL 32601, USA

Author for correspondence:
Dario Grattapaglia
Tel: +55 61 999712142
Email: dario.grattapaglia@embrapa.br

Received: *8 June 2018*
Accepted: *13 August 2018*

New Phytologist (2019) **221:** 818–833
doi: 10.1111/nph.15449

Key words: *Eucalyptus*, genome-wide association study (GWAS), high-throughput SNP genotyping, Joint-GWAS, meta-analysis, regional heritability mapping (RHM), relatedness.

Summary

- Genome-wide association studies (GWAS) in plants typically suffer from limited statistical power. An alternative to the logistical and cost challenge of increasing sample sizes is to gain power by meta-analysis using information from independent studies.
- We carried out GWAS for growth traits with six single-marker models and regional heritability mapping (RHM) in four *Eucalyptus* breeding populations independently and by Joint-GWAS, using gene and segment-based models, with data for 3373 individuals genotyped with a communal EUChip60KSNP platform.
- While single-single nucleotide polymorphism (SNP) GWAS hardly detected significant associations at high-stringency in each population, gene-based Joint-GWAS revealed nine genes significantly associated with tree height. Associations detected using single-SNP GWAS, RHM and Joint-GWAS set-based models explained on average 3–20% of the phenotypic variance. Whole-genome regression, conversely, captured 64–89% of the pedigree-based heritability in all populations. Several associations independently detected for the same SNPs in different populations provided unprecedented GWAS validation results in forest trees. Rare and common associations were discovered in eight genes involved in cell wall biosynthesis and lignification.
- With the increasing adoption of genomic prediction of complex phenotypes using shared SNPs and much larger tree breeding populations, Joint-GWAS approaches should provide increasing power to pinpoint discrete associations potentially useful toward tree breeding and molecular applications.

Introduction

Identifying the discrete genotype−phenotype associations underlying complex traits in forest trees, and plants in general, continues to be a challenge of far-reaching biological and economic importance. Given the high genetic diversity, low extent of linkage disequilibrium (LD) and lack of structure of forest tree populations, these were proposed as ideal systems for genetic association studies (Neale & Savolainen, 2004). Earlier reports, largely in species of *Populus* and *Pinus*, examined the variation in candidate genes in which a few associations explaining small proportions of the genetic variation were detected (Neale & Savolainen, 2004; Thumma *et al.*, 2005; Neale, 2007; Wegrzyn *et al.*, 2010; Khan & Korban, 2012; Guerra *et al.*, 2013; Thavamanikumar *et al.*, 2014; Jaramillo-Correa *et al.*, 2015). With the development of accessible high-density single nucleotide polymorphism (SNP) genotyping platforms, genome-wide association studies (GWAS) were performed using marker densities in the range of several thousand SNPs in collections of a few hundred individuals (Cappa *et al.*, 2013; Porth *et al.*, 2013; Evans *et al.*, 2014; McKown *et al.*, 2014; Allwright *et al.*, 2016; Du *et al.*, 2016; Fahrenkrog *et al.*, 2016). Relatively large effect associations were detected for phenology and wood properties traits. However, very few associations were found for complex growth traits, and the proportion of genetic variation explained by these individual associations was typically very small.

Many GWAS in forest trees have employed collections of trees directly sampled from the wild. This sampling strategy aimed to

minimize genetic structure and extent of LD to provide improved resolution for discovering potentially useful causal variants for marker-assisted selection (MAS). However, rare alleles detected in wild populations may not be segregating or have a negligible effect in an elite material background. Moreover, differently from crop breeding in which backcross introgression of high-value wild alleles into elite lines is commonplace, such a route is not an option in forest trees. An alternative strategy to detect high-value alleles by GWAS has been to develop more structured discovery populations, such as the nested association mapping (NAM) populations (Li *et al.*, 2016; Wu *et al.*, 2016). This approach puts the population through a one-generation bottleneck, raising some alleles to high and detectable frequency, while eliminating many others (Hamblin *et al.*, 2011). Although less genetic variation is in principle available in such structured populations, the associations detected in genetically improved material should be considerably more relevant to further breeding, as their effect would be relevant in an already elite background. Following an equivalent rationale, three studies have reported GWAS for growth and wood quality traits in breeding populations of *Eucalyptus* (Cappa *et al.*, 2013; Müller *et al.*, 2017; Resende *et al.*, 2017a). Interestingly, the results were on par with those described earlier in natural populations, in which few associations explaining small fractions of the genetic variation were detected. This finding corroborates the complexity of the traits and the insufficient detection power to detect small effects.

The statistical power to detect associations between DNA variants and a trait depends largely on the experimental sample size (Visscher *et al.*, 2017). Due to the inherent challenges of creating large populations for GWAS in plants, most studies utilized populations smaller than a few hundred individuals, with the exception of studies using NAM populations in which several thousand individuals have been used (reviewed by Xiao *et al.*, 2017). A potentially more viable alternative to gain statistical power in plants can be obtained by combining information from multiple populations using Meta-GWAS and Joint-GWAS (Mägi & Morris, 2010; Yang *et al.*, 2012; Bernal Rubio *et al.*, 2016; Li *et al.*, 2016; Wallace *et al.*, 2016; Wu *et al.*, 2016). Meta-GWAS combines the *P*-values from independent studies to increase the power to detect variants with small effect sizes and is a popular method for discovering new genetic risk variant in human datasets (Evangelou & Ioannidis, 2013). Joint-GWAS, alternatively, combines the populations before the association analysis, leading to more resolution and the detection of more associations for complex traits (Lin & Zeng, 2009). As each experiment is independently designed, both methods have to account for the heterogeneity created by population structure and phenotype measurements, among other potential sources of variability (Magosi *et al.*, 2017).

A second approach to increase the power of a GWAS is to capture the information of all genetic variants in a genomic region, including rare and low-frequency ones. Methods to exploit the combined effect of multiple SNPs in genomic segments using region or gene-based GWAS have been developed to account for rare and low-frequency variants (Wu *et al.*, 2011; Nagamine *et al.*, 2012; Bakshi *et al.*, 2016). The regional heritability mapping (RHM, Nagamine *et al.*, 2012) is a region-based GWAS approach with good potential for these cases, as it captures more of these underlying small genetic effects. This method provides heritability estimates for short genomic regions, using the genomic relationship matrix (GRM), and has the power to detect regions containing common and rare SNP variants that individually contribute too little variance to be detected by single-SNP GWAS. As many trait-associated genetic variants identified from GWAS tend to be enriched in genic regions (Schork *et al.*, 2013), it is more powerful to test the aggregated effect of a set of SNPs using a set-based association approach for the detection of complex trait genes (Bakshi *et al.*, 2016).

In this study, we performed a Joint-GWAS for growth traits by assembling a considerably large association population from individual *Eucalyptus* breeding populations. *Eucalyptus grandis*, *Eucalyptus urophylla* and their hybrids are among the most widely planted tree species for pulp and solid wood production in the tropics. Interspecific hybrids between *E. grandis* × *E. urophylla* make up almost the totality of large scale operational plantations and are the main target of breeding programs due to their combination of desirable traits, most notably fast growth from *E. grandis* and disease and abiotic stress resistance from *E. urophylla* (Myburg *et al.*, 2007). We assembled genome-wide SNP and growth trait data for 3373 trees across four unrelated *E. grandis* × *E. urophylla* breeding populations. We evaluated different GWAS models to correct for population stratification and relatedness, to detect associations within and across these different breeding populations. We also evaluated the performance of RHM in capturing larger fractions of the additive genetic variance. Association analyses by genes and segments were performed from summary data from Joint-GWAS to increase the power to detect trait associations. Several associations were independently detected for the same SNPs across the unrelated populations, providing validation results for specific loci. Associations were detected into genes related to cell wall biosynthesis and lignification processes suggesting potential pleiotropic effects. To the best of our knowledge, this is the first study to apply Joint-GWAS in a forest tree.

Materials and Methods

Populations and phenotypic data

This study was carried out using trees in progeny trials of four unrelated *E. grandis* × *E. urophylla* hybrid breeding populations (Pop1-IPB, Pop2-ARAB, Pop3-ARAC and Pop4-CNB), belonging to three Brazilian paper and pulp companies: International Paper of Brazil (IPB), Fibria Celulose (formerly Aracruz – ARA) and Cenibra Celulose (CNB). Details of the size of the trial, experimental design, number of families, age of measurement, location and sample sizes used in the GWAS are described in Table 1. Three of the four populations were used in previously published genomic prediction studies Pop1-IPB (Lima, 2014), and Pop3-ARAC and Pop4-CNB (Resende *et al.*, 2012, 2017b). While populations Pop1-IPB, Pop2-ARAB and Pop3-ARAC were largely composed of first generation hybrids, Pop4-CNB went through one additional selection cycle being equivalent to

an outbred F_2, as the parents were themselves hybrids (F_1) between *E. grandis* × *E. urophylla*. All trees were phenotyped at age 2–5 yr for diameter at breast height (DBH, cm) and total height (HT, m) (Supporting Information Fig. S1).

SNP genotyping and quality control

In total, 3417 trees were genotyped using the *Eucalyptus* Illumina Infinium EUChip60K (Silva-Junior *et al.*, 2015), of which 44 with >10% missing data were removed, therefore 3373 remaining for further analyses. A combined dataset was generated by merging the genotyping data of each population. The genotypic data for each population and for the combined data were filtered to remove SNPs with call rate (CR) < 90% and monomorphic SNPs, therefore maintaining rare SNPs with minor allele frequency (MAF) > 0 in the analyses (full marker dataset). Two alternative SNP datasets were also generated by retaining only SNPs with MAF ≥ 0.01 and MAF ≥ 0.05, respectively (Table S1). For the population stratification analyses, SNPs in intergenic regions (putatively neutral) were selected based on their localization outside of annotated gene models in the *Eucalyptus* genome (Myburg *et al.*, 2014). SNPs were then filtered using PLINK v1.9 (Purcell *et al.*, 2007) to generate a pruned subset of SNPs in approximate linkage equilibrium (LE).

Population stratification analyses

The underlying genetic structure of the four populations and the combined data was estimated based on a Bayesian clustering method implemented in STRUCTURE v.2.3.4 (Pritchard *et al.*, 2000), using the intergenic SNPs in approximate LE (Table S1). The admixture model was applied, with correlated allelic frequencies, using no previous population information. The number of tested clusters (K) ranged from 1 to 10, with 10 replications per K. The burn-in period and the number of Markov chain Monte Carlo (MCMC) iterations were 50 000 and 150 000, respectively. The number of genetic groups was determined based on the criteria proposed by Evanno *et al.* (2005). The POPHELPER R package (Francis, 2016) was used to generate the population structure bar plots by individuals. Principal component analysis (PCA) was performed using SNPRelate (Zheng *et al.*, 2012) to plot all individuals for the combined dataset. To correct for population stratification in the GWAS models, we performed a PCA using GCTA v1.26.0 (Yang *et al.*, 2011) in each population independently and for the combined dataset. The number of significant principal components for each population and combined data was determined by a broken stick model (Jackson, 1993) using evplot function (Borcard *et al.*, 2011). The pairwise genetic distances between populations (F_{ST}) were estimated according to Weir & Cockerham (1984) using SNPRelate.

Linkage disequilibrium and heritability estimation

Genome-wide pairwise estimates of LD were calculated by the classical measure of the squared correlation of allele frequencies at diallelic loci (r^2) for each chromosome separately and for all four populations independently using PLINK v1.9 (Purcell *et al.*, 2007). The LD decay of r^2 with distance in kbp was fitted by a non-linear regression model between adjacent sites. The

Table 1 Main characteristics of the four *Eucalyptus* association populations used in the study.

Phenotypic data	Pop1-IPB	Pop2-ARAB	Pop3-ARAC	Pop4-CNB
Company	International Paper Brazil	Fibria	Fibria	Cenibra
Site	Brotas, SP	Aracruz, ES	Aracruz, ES	Sabinópolis, Virginópolis, Antônio Dias, MG
Coordinates	22°S; 48°W	19°S; 40°W	19°S; 40°W	18°S; 42°W
Total number of parents	46	52	47	10
Total number of full-sibs (FS) families	58	68	75	43
Number of families remaining in the analyses	45	68	75	37
Number of individuals/FS family remaining in the analyses	22	13	10	21
Number of species involved in the population composition	3 (*E. grandis*, *E. urophylla*, *E. camaldulensis*)	5 (*E. grandis*, *E. urophylla*, *E. camaldulensis*, *E. saligna*, *E. globulus*)	4 (*E. grandis*, *E. urophylla*, *E. globulus*, *E. maidenii*)	2 (*E. grandis*, *E. urophylla*)
Number of blocks	8	30	40	36
Number of tree per plot	6	1	1	1
Experimental Design	RCBD	ALD-IB	ALD-IB	RCBD
Total number of trees in trial	2784	5280	9600	4900
Number of trees used in the GWAS analyses	979	875	758	761
Year when trees were planted	2006	2006	2006	2005
Year when trees were phenotyped	2011	2008	2008	2008

ALD-IB, alpha lattice design (incomplete block); RCBD, randomized complete block design.

drift-recombination model (Hill & Weir, 1988) was used to fit a nonlinear regression of the expectation of r^2, using the R script from Marroni *et al.* (2011) and the equation described by Remington *et al.* (2001). Finally, to visualize patterns of LD decay in the four *Eucalyptus* breeding populations, the LD estimated (r^2) were plotted in a 1 Mbp window. The variances components, genomic and pedigree-based heritabilities for DBH and HT were estimated for each population separately using the BGLR v.1.0.5 R package (Pérez & de los Campos, 2014) (see Methods S1).

GWAS models

For the following GWAS analyses, we used the adjusted phenotypic data from each population separately, and for the combined data we corrected these adjusted phenotypes for age and population as described in Methods S2. Three different GWAS approaches were used to detect associations: single SNP-based models, regional heritability mapping (RHM) and SNP set-based models. A brief description is presented below, and more details can be found in Methods S3.

Single SNP-based models Six distinct GWAS models were implemented in the EMMAX software (Kang *et al.*, 2010), using three linear model-based association (LMA) and three mixed linear model-based association (MLMA) models. The LMA models were fitted without any correction for population stratification and relatedness (Model 1: None); with only the *Q*-matrix from STRUCTURE (Model 2: *Q*); and with significant principal components from PCA (Model 3: *P*). The MLMA models is similar to LMA models, except for the inclusion of the polygenic effect captured by the GRM, which is evaluated with the kinship matrix only (Model 4: *K*); fitted with the kinship matrix and *Q*-matrix (Model 5: *K*+ *Q*); and fitted with the kinship matrix and significant principal components (Model 6: *K*+ *P*). All these single-SNP GWAS models were performed for each population independently and for the combined data (Joint-GWAS). The proportion of the phenotypic variation explained by each marker (h^2_m) were calculated as described in Methods S4.

Regional heritability mapping The RHM method was applied to each population independently using GCTA (Yang *et al.*, 2011). This method divides the genome into windows of predetermined numbers of SNPs (regions) for each chromosome, and the variance for each window is estimated. As described in the original methodology (Nagamine *et al.*, 2012), we used a window size of 100 adjacent SNPs to build a regional relationship matrix and the window was shifted every 50 SNPs. At the end of the chromosome, a minimum of 100 SNPs encompasses the last window. The whole-genomic and regional heritabilities were estimated as $h^2_g = \sigma^2_g/\sigma^2_y$ and $h^2_r = \sigma^2_r/\sigma^2_y$, respectively (see Methods S3).

SNP set-based models To increase the power of the Joint-GWAS (combined data), given that the effect sizes of individual genetic variants potentially detected by single SNP-based models could be very small, we tested the aggregated effect of sets of SNPs. Two set-based methods were used: (1) a gene-based model and (2) a segment-based GWAS model using segments of 100 kbp. Both approaches were performed using fastBAT (Bakshi *et al.*, 2016) in GCTA (Yang *et al.*, 2011). The proportion of the phenotypic variation explained by each window (h^2_w) for set-based models was estimated as described in Methods S4.

Assessments of the statistical significance in GWAS and RHM To select significant associations different multiple test corrections were applied to the *P*-values obtained in the GWAS and the RHM approaches. For the single SNP-based GWAS models, a genome-wide level using the Bonferroni procedure was implemented to control for type I error at $\alpha = 0.05$ and a suggestive level with the Benjamini & Hochberg (1995) procedure was used to control for false discovery rate (FDR) at 5%. A third less stringent *ad hoc* threshold of $-\log_{10}(P) \geq 4$, was also used to declare additional significant associations that were not detected under Bonferroni and FDR corrections. This *ad hoc* threshold was defined based on the threshold value established in a previous study in population Pop4-CNB, using a permutation test with Bonferroni correction for multiple tests (Resende *et al.*, 2017a). This threshold value is more stringent than the one ($-\log_{10}(P) \geq 3.5$) reported in a soybean study (Kaler *et al.*, 2017) using a comparable number of SNPs (31 260) and similar to the *ad hoc* threshold considered in a *Populus nigra* study (Allwright *et al.*, 2016). The thresholds considered for the set-based GWAS were the same as those used for the single SNP-based GWAS models, but instead of the number of SNPs tested, the number of genes was used for the gene-based GWAS and the number of regions created by the segment-based GWAS. For the RHM approach, to account for the overlapping windows for the RHM approach, half of the total number of windows tested were used in the Bonferroni and FDR at 5% multiple testing corrections. Additionally, an *ad hoc* threshold ($-\log_{10}(P) \geq 2$) was tested to declare significant RHM windows associated with growth traits. Manhattan plots were generated using the QQMAN (Turner, 2018) and the GGPLOT2 (Wickham, 2009) R packages.

Results

SNP genotyping and population stratification

In total, 59 222 SNPs were targeted for genotyping using the EUChip60K chip (Silva-Junior *et al.*, 2015). More than 46 000 SNPs were retained for all populations and 51 274 SNPs for the combined dataset following a filter for call rate (CR) ≥ 90-%. After removing monomorphic markers (MAF = 0), *c.* 30 000 SNPs were retained for the four populations and 41 320 for the combined set with an overall final rate of missing data of only 1%. The distribution of the number of filtered SNPs into MAF classes showed enrichment for low frequency alleles (MAF 0–0.1, Fig. S2). Two alternative SNP datasets with different MAF thresholds were also used to investigate whether removing lower frequency SNPs had an impact on GWAS results. Finally, sets of 7000 to 14 000 from the filtered SNPs were selected in intergenic regions and in approximate

LE for the population structure analyses (Table S1). The most likely numbers of subpopulations varied between $K = 2$ for Pop1-IPB and Pop2-ARAB to $K = 5$ for Pop4-CNB (Table S1). When all four populations were combined the most likely number of subpopulations was $K = 2$ (Fig. 1a), using no previous population information in the admixture model. The ancestry coefficient bar plots from STRUCTURE showed K ranging from two to four subpopulations in the combined dataset (Fig. 1a). For $K = 2$ Pop4-CNB was separated from the other three populations, consistent with the highest F_{ST} estimates observed between Pop4-CNB and the others (F_{ST} range from 0.0712 to 0.0937). For $K = 3$, Pop2-ARAB and Pop3-ARAC were grouped together, showing that individuals from these two populations are more closely related ($F_{ST} = 0.0370$) than the others, in agreement with the origin of these two populations that belong to the same breeding program and share some common parents. When $K = 4$ the combined dataset was subdivided into four populations, although some proportion of admixture was present. The numbers of significant principal components according to a broken stick model were used in the GWAS analyses to correct for population stratification based on the PCA results. The significant PCs defined in the four populations and the combined population (Table S1) cumulatively explained 8.6%, 3.3%, 6.6%, 11.2% and 7.6% of the variation for each data, respectively. The PCA for the combined dataset showed that all four populations have a similar genetic background, with the first two principal components explaining only 3.2% and 2.4% of the genetic variance (Fig. 1b).

LD, genomic and pedigree-estimated heritabilities

The pairwise estimates of LD (r^2) were calculated between all high-quality polymorphic SNPs (MAF > 0) on each chromosome separately for the four populations independently. The average genome-wide LD for pairs of SNPs within a 1 Mbp distance from each other ranged from 0.052 (Pop1-IPB) to 0.256 (Pop4-CNB). The genome-wide decay of LD to an r^2 below 0.2 were considerably faster for Pop1-IPB (34.8 kbp), Pop3-ARAC (42 kbp) and Pop2-ARAB (75.1 kbp) compared with that of Pop4-CNB (637.7 kbp) (Fig. 1c). The more extensive LD on Pop4-CNB may be explained by the more advanced selection state of this population when compared with the others. The estimated pedigree-based narrow-sense heritabilities (h^2) were moderate (0.374 for Pop4-CNB) to high (0.683 for Pop3-ARAC), with the lowest and highest values observed for DBH. Estimates

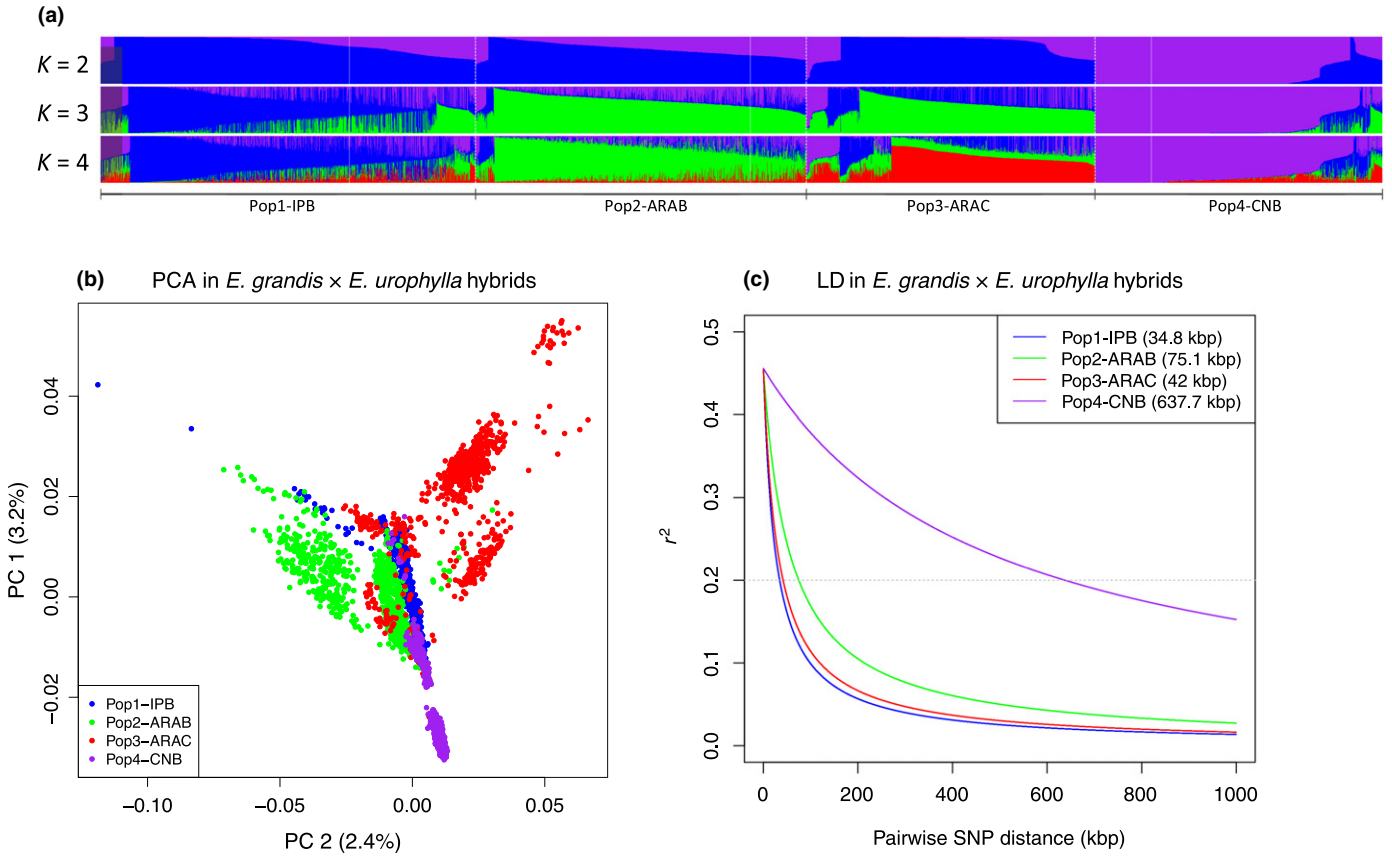

Fig. 1 Population structure, principal component analysis (PCA) and linkage disequilibrium (LD) decay for the four unrelated *Eucalyptus grandis* × *E. urophylla* hybrid breeding populations. (a) Bar plots from population structure for number of cluster (K) ranging from $K = 2$ to $K = 4$. (b) PCA with two eigenvectors (PC 1 and PC 2) and (c) genome-wide pattern of LD (correlation coefficient r^2) decay plotted up to 1 Mbp pairwise single nucleotide polymorphism (SNP) distances and a dashed line at $r^2 = 0.2$ indicates the frequently used threshold of usable LD.

of genomic heritabilities varied from 0.296 for HT in Pop4-CNB to 0.528 for DBH in Pop3-ARAC, corresponding to a large proportion (64–89%) of the pedigree-based heritabilities (Table S2). Estimates of variance components are also reported in Table S2.

Single-SNP GWAS

The LMA Models 1, 2 and 3 without the introduction of a GRM (*K* of kinship) resulted in the detection of a large number of associations. Most of these were deemed spurious due to the structured nature of these breeding populations (Table 2). For instance, in Model 1 (no correction) there were hundreds to thousands of SNPs associated with growth traits for all populations. When the population stratification covariate obtained either by STRUCTURE (*Q*) or PCA (*P*) was included in the LMA model (Models 2 and 3), the number of associations for each population reduced drastically, except for Pop1-IPB that showed a slight increase. The quantile−quantile (QQ) plots show the inadequacy of the LMA model without the kinship matrix for GWAS analyses, since the observed and expected *P*-values differed considerably for a large number of SNPs (Fig. S3).

When the random effects captured by the kinship matrix (*K*) and the fixed effects captured by population stratification (STRUCTURE or PCA) were included in the MLMA models, no associations were detected for DBH in each population separately after correction for multiple testing (Table 2; Figs S4–S8). The same was observed for total height, except for Pop4-CNB where several significant associations (Figs 2, S7b,d) were detected using a FDR (5%) threshold (Table 2), in agreement with its more extensive LD (> 600 kbp, Fig. 1c) and smaller effective population size (Table 1). All these significant associations detected by single-SNP GWAS for Pop4-CNB are for common SNPs, with allele frequencies ranging from 0.27 to 0.47, suggesting that this approach is suitable for the detection of common variants. Nevertheless, when a more stringent adjustment for multiple testing was used (Bonferroni at 5%), no significant association remained (Table 2). Most *P*-values were similar to the expected diagonal in the QQ plots in the MLMA models adjusted for GRM, which indicates better appropriateness of these GWAS models (Fig. S3). Furthermore, the models built with GRM produced a drastic reduction in the number of significant markers, highlighting the impact of relatedness on GWAS in these breeding populations. The two alternative marker datasets (MAF ≥ 0.01 and MAF ≥ 0.05) did not show any difference as far as results for the single-SNP GWAS because all SNPs found associated were common.

To increase the power of detection, a Joint-GWAS was performed combining the data for all populations. Using this approach, three associations were detected for DBH when the kinship matrix was included after multiple-test correction (Bonferroni at 5%) (Table 2; Fig. 3). For HT, no significant association was found after inclusion of the GRM in the model (Table 2; Fig. S8b,d). Although traditional multiple testing thresholds (FDR and Bonferroni) are important to control for type I error (false positive), they may be excessively stringent for GWAS, when several thousand markers are used and a minority are expected to be associated with a phenotypic response. A less stringent *ad hoc* threshold ($-\log_{10}(P) \geq 4$) was used to declare additional significant associations not detected before. With this threshold, eight variants putatively associated ($P \leq 0.00008$) with DBH were found in the Joint-GWAS analysis (Fig. 3, green dashed line). Collectively, out of these 11 SNPs associated with DBH from Joint-GWAS results (three associations and eight

Table 2 Number of significant single nucleotide polymorphism (SNP) associations for growth traits using linear model-based association (LMA; Models 1–3) and mixed linear model-based association (MLMA; Models 4–6) models for the four *Eucalyptus* breeding populations and for the joint genome-wide association study (Joint-GWAS; all) analyses.

Population	Trait	No. of SNPs	Model 1 None	Model 2 *Q*	Model 3 *P*	Model 4 *K*	Model 5 *K* + *Q*	Model 6 *K* + *P*
Pop1-IPB	DBH	32 110	3805 (260)	4212 (315)	4155 (318)	0 (0) | 7	0 (0) | 7	0 (0) | 7
Pop2-ARAB		34 859	11 147 (1783)	4373 (212)	2906 (109)	0 (0) | 3	0 (0) | 4	0 (0) | 2
Pop3-ARAC		30 979	17 954 (6729)	9464 (1668)	3655 (302)	0 (0) | 13	0 (0) | 6	0 (0) | 8
Pop4-CNB		28 795	12 542 (3411)	1149 (74)	1396 (34)	0 (0) | 4	0 (0) | 3	0 (0) | 2
All[a]		41 320	24 635 (11 871)	18 395 (6291)	18 406 (5693)	3 (3) | 10	3 (2) | 11	2 (2) | 7
Pop1-IPB	HT	32 110	2731 (119)	3201 (148)	3237 (163)	0 (0) | 7	0 (0) | 6	0 (0) | 7
Pop2-ARAB		34 859	5797 (350)	2854 (167)	2654 (145)	0 (0) | 3	0 (0) | 3	0 (0) | 3
Pop3-ARAC		30 979	13 815 (3338)	4927 (347)	2201 (80)	0 (0) | 6	0 (0) | 9	0 (0) | 9
Pop4-CNB		28 795	8303 (959)	1104 (242)	3263 (472)	27 (0) | 40	97 (0) | 78	12 (0) | 45
All[a]		41 320	17 383 (4385)	13 560 (2791)	12 259 (2606)	0 (0) | 3	0 (0) | 4	0 (0) | 1

Also reported the number of SNPs putatively associated with growth traits using MLMA models (Models 4–6). Number of significant associations using false discovery rate (FDR) of 5%. Numbers between parenthesis correspond to significant association using Bonferroni correction with an experimental type I error rate of $\alpha = 0.05$. Numbers after '|' are the number of SNPs putatively associated using an *ad hoc* threshold of $-\log_{10}(P) \geq 4$.
Model 1 (None), LMA without any correction for population stratification and relatedness; Model 2 (*Q*), LMA with *Q*-matrix from STRUCTURE; Model 3 (*P*), LMA with significant principal components (PCs); Model 4 (*K*), MLMA with EMMAX kinship matrix (GRM); Model 5 (*K* + *Q*), MLMA with GRM and *Q*-matrix; Model 6 (*K* + *P*), MLMA with GRM and significant PCs.
DBH, diameter at breast height; HT, total height.
[a]Combined dataset generated by merging the genotypic data of all four populations.

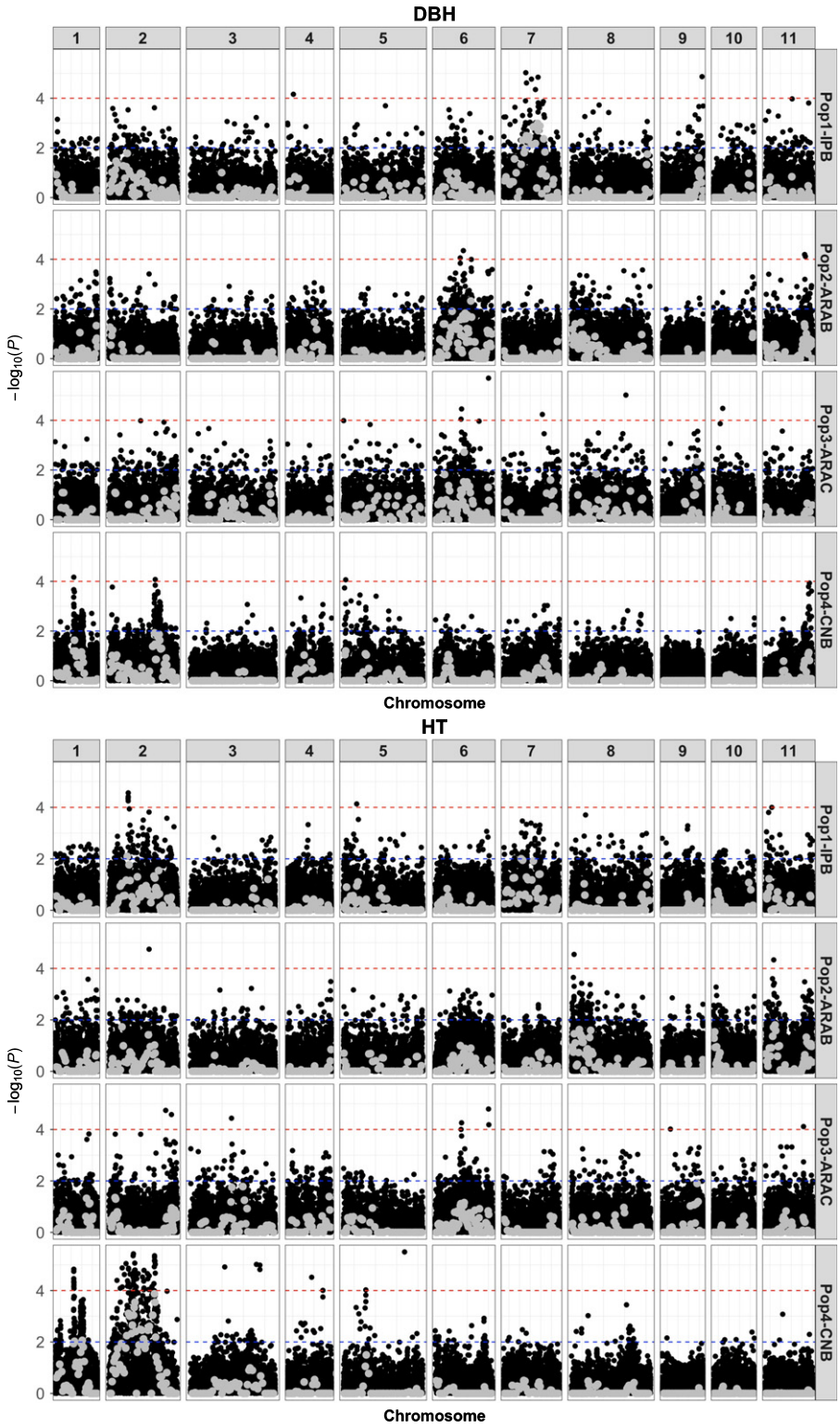

Fig. 2 Manhattan plots for growth traits (DBH, diameter at breast height; HT, total height) using single-single nucleotide polymorphism (SNP) genome-wide association study (GWAS, black points, Model 5: $K + Q$) and regional heritability mapping (RHM, grey points), corrected for population structure and the kinship matrix, for the four unrelated *Eucalyptus grandis* × *E. urophylla* hybrids breeding populations. Red and blue line indicate *ad hoc* thresholds adopted for the single-SNP GWAS and RHM analyses, respectively.

putative associations) six are located in genes, including the three most significant ones. Six of the 11 associations are common SNPs (MAF = 0.058–0.422) and the remaining five SNPs are rare (MAF = 0.001–0.015). When the *ad hoc* threshold was considered for HT, four associations were detected, in which the most significant SNP ($P = 0.000006$) is also the most significant one detected for DBH (EuBR07s38098526, see later Table 5) and located on chromosome 7. The third SNP (EuBR08 s48262720) associated with DBH (FDR at 5%) was also detected for HT on chromosome 8. These results are not unexpected given the high phenotypic correlation between these two growth traits ($r = 0.82$). For the four SNPs putatively associated

with HT, three are rare (MAF = 0.015–0.017) and one is common (MAF = 0.429).

When the *ad hoc* threshold was considered for the single-SNP GWAS corrected by kinship matrix and STRUCTURE (Model 5), putatively associated SNPs were detected for both traits in all populations (Table 2; Fig. 2, red line) and explained on average only 3% of the phenotypic variance (Table 3). However, significant GWAS hits found in each population after correcting for both family and population structure were generally not shared across populations. To investigate whether shared associations across populations could provide an independent way to validate the associations found, results from Model 2 (*Q*-matrix from STRUCTURE) were used to create a comparison dataset for all populations, leading to four and six associations shared for DBH and HT, respectively (Fig. 4). These results were comparable to those obtained using Model 3 (significant PCs), where the number of shared associations were three for DBH and seven for HT (data not shown). Amongst the shared associations from Model 2 and 3 for DBH, one association (EuBR10s19747657) was common between these two methods of correction for population stratification. For HT, four associations were found in common between Model 2 and 3 for all populations, one located on chromosome 1 (EuBR01s5300169) and the others on chromosome 2 within an interval of 13 kbp (EuBR02s42875938, EuBR02s42876352 and EuBR02s42888917). Interestingly, the number of common associations found among all populations increased considerably when Pop4-CNB was excluded from the analysis and comparisons were made only among Pop1-IPB, Pop2-ARAB and Pop3-ARAC. Under this scenario, 157 and 40 significant associations were shared for DBH and HT, respectively (Fig. 4). This considerable difference in results likely reflects the significant genetic differentiation and smaller effective population size of Pop4-CNB when compared with other populations in the structure analysis (Fig. 1a).

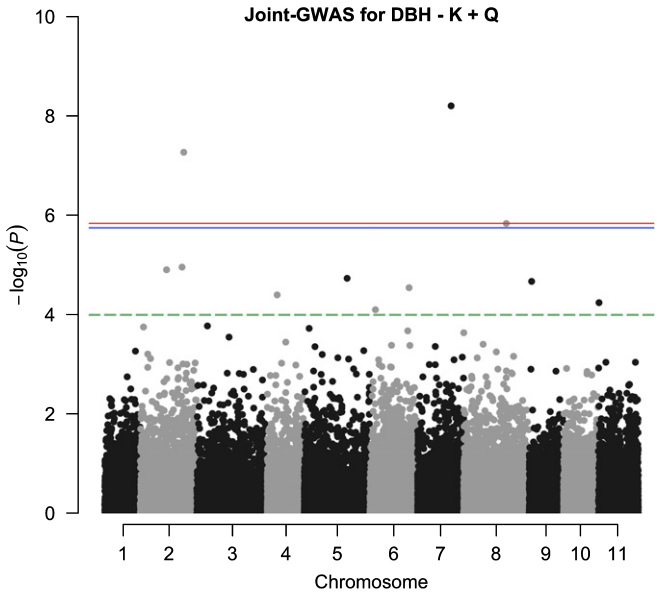

Fig. 3 Manhattan plot of the associations for diameter at breast height (DBH) for *Eucalyptus* using a single-single nucleotide polymorphism (SNP) joint genome-wide association study (Joint-GWAS: 41 320 SNPs), adjusted for kinship matrix and population structure (Model 5: $K + Q$), and age of measurements and population of origin for the combined dataset. Red line indicates Bonferroni-corrected threshold with an experimental type I error rate at $\alpha = 0.05$, blue line indicates a false discovery rate (FDR) at 5% and green dashed line represents the *ad hoc* threshold.

Regional heritability mapping

Regional heritability mapping (RHM) was performed for each population independently to evaluate whether additional associated variants could be detected. For Pop1-IPB, Pop2-ARAB and Pop3-ARAC no significant regions were declared significant with this approach using multiple testing correction. Alternatively, for Pop4-CNB, eight regions (each with 100 SNPs) were significantly associated with total height on chromosome 2 (Fig. 2; Table 4) at the suggestive level (FDR at 5%), with one of those reaching the genome-wide level (Bonferroni at 5%). This result is consistent with the single SNP-based GWAS, which detected 78 significant common variants clustered on chromosome 2 using the correction for multiple testing (FDR at 5%) for Model 5 (Fig. 2). The most significant window ($h^2_r = 0.07$) detected by RHM for HT in Pop4-CNB captured 24% of the genomic heritability ($h^2_g = 0.29$). Altogether, each of the eight significant

Table 3 Proportion of the genomic heritability explained by each marker (h^2_m) using single-single nucleotide polymorphism (SNP) genome-wide association studies (GWAS) and by each window (h^2_w) using Joint-GWAS gene and segment-based models.

Single-SNP GWAS	Trait	No. of significant SNPs[a]	Min.[b] of h^2_m	Max.[c] of h^2_m	Average of h^2_m
Pop1-IPB	DBH	7	0.015	0.059	0.042
Pop2-ARAB		4	0.008	0.045	0.032
Pop3-ARAC		6	0.009	0.025	0.019
Pop4-CNB		3	0.001	0.044	0.021
All[d]		11	0.0005	0.071	0.025
Pop1-IPB	HT	6	0.016	0.048	0.039
Pop2-ARAB		3	0.023	0.031	0.026
Pop3-ARAC		9	0.003	0.093	0.041
Pop4-CNB		78	0.007	0.066	0.041
All[d]		3	0.001	0.014	0.006

Joint-GWAS	Trait	No. of significant SNPs per window[a]	Min. of h^2_w	Max. of h^2_w	Average of h^2_w
Gene based	DBH	3–6	0.024	0.149	0.087
Gene based	HT	2–9	0.005	0.384	0.263
Segment based	DBH	5	0.149	0.149	0.149
Segment based	HT	2–9	0.178	0.385	0.299

Significant association obtained from results utilizing mixed linear model-based association (MLMA; Model 5) for the four *Eucalyptus* breeding populations and combined dataset (All). DBH, diameter at breast height; HT, total height.

[a]Number of significant associations declared using Bonferroni correction with an experimental type I error rate of $\alpha = 0.05$, or false discovery rate (FDR) of 5% or using an *ad hoc* threshold of $-\log_{10}(P) \geq 4$.

[b]Minimum values.

[c]Maximum values.

[d]Combined dataset generated by merging the genotypic data of all four populations.

windows declared by RHM explained 5–10% of the total genomic heritability captured by the whole-genome relationship matrix. In addition to these eight associations, 12 more were putatively associated considering the lower *ad hoc* threshold for RHM adopted ($-\log_{10}(P) \geq 2$), where two windows are located on chromosome 1 and the remaining 10 on chromosome 2. For DBH in Pop4-CNB, no significant regions were detected. Still under this lower *ad hoc* threshold, 12 windows in Pop1-IPB were putatively associated with DBH on chromosome 7, with the most significant one showing a regional heritability of 0.13, which alone captures 25% of the total genomic heritability ($h^2_g = 0.52$). Additionally, one association was declared significant for DBH on Pop2-ARAB (chromosome 6) and two for Pop3-ARAC (chromosome 6 and 9). For HT, two windows were putatively associated for Pop1-IPB, one on chromosome 2 and one on 7 (Table 4). The associations detected by RHM explained on average 6% of the phenotypic variance, slightly superior to the single-SNP GWAS (3%) for these complex traits.

Joint-GWAS from summary datasets

To further assess the power of combining all populations into a single analysis, we analyzed the summary data from Joint-GWAS into genic and segment-based SNP sets. Of the 36 349 total genes in the *E. grandis* genome v.2.0 (Myburg *et al.*, 2014), 31 770 genes were considered as gene sets in our analysis as they contained SNPs targeted by the EUChip60K in their sequence or vicinity (50 kbp). For the gene-based Joint-GWAS, nine genes with six contiguous SNPs were significantly associated with HT at the genome-wide level (Bonferroni at 5%) on chromosome 10, after adjusting for kinship and population structure (Fig. 5b).

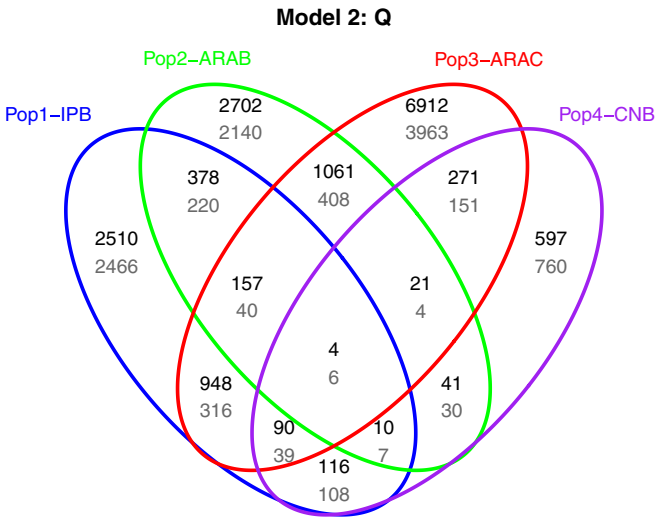

Fig. 4 Venn diagram of the number of significant associations identified for growth traits using single-single nucleotide polymorphism (SNP) genome-wide association study (GWAS) for the four unrelated *Eucalyptus grandis* × *E. urophylla* hybrid breeding populations. Comparison of the number of significant associations identified for diameter at breast height (DBH, black numbers) and total height (HT, grey numbers) by false discovery rate (FDR) threshold at 5%, using linear model-based association (LMA) model corrected for population structure (Model 2: Q).

When considering only the kinship matrix, without the correction for population structure, a peak on chromosome 9 containing 15 genes was also significant. Other significant signals were detected at the suggestive level (FDR at 5%), with one gene associated with two close SNPs on chromosome 3 and another locus with five SNPs on chromosome 7 (Fig. 5a). For the segment-based Joint-GWAS (Fig. 5c,d), 4766 segments of size 100 kbp were tested, with four of those regions being associated with HT (Fig. 5c). The most significant region (Bonferroni at 5%) contains three SNPs, located on chromosome 2 and near two genes, that had not been detected in the previous GWAS analyses performed for the trait. The second most significant region considering a genome-wide level is the same as the most significant one detected by the gene-based approach. The remaining two associated segments were the same regions detected by the gene-based method, showing an agreement between segment-based and gene-based Joint-GWAS. Despite the detection of three significant associations for DBH with single-SNPs Joint-GWAS, no association was detected using the summary datasets for this trait by multiple test correction. These results highlight the increased power of set-based Joint-GWAS with significant genomic regions explaining on average 20% of the phenotypic variance (Table 3).

Discussion

This study further advances the investigation of discrete genomic regions controlling growth traits in forest trees. Significant associations were detected for height and diameter with the increased power of Joint-GWAS experiments, which leveraged genome-wide data from 3373 individuals across four *Eucalyptus* breeding populations. Our study further corroborates the complex architecture of growth traits and suggests that combining data from multiple independent populations is a viable option to increase the sample size and increase the power to detect at least part of the slightly larger effects segregating in breeding populations.

Detection of associations for complex traits in forest trees

Various studies attempted GWAS for growth traits in forest trees, namely in *Populus* (Porth *et al.*, 2013; Allwright *et al.*, 2016; Du *et al.*, 2016; Fahrenkrog *et al.*, 2016), *Pinus* (Bartholomé *et al.*, 2016; Lu *et al.*, 2017) and *Eucalyptus* (Cappa *et al.*, 2013; Müller *et al.*, 2017; Resende *et al.*, 2017a). Despite the considerably large number of individuals used in our study for each population and for the combined dataset, our results suggested that much larger numbers will be necessary to identify discrete regions capturing larger fractions of the genetic variance of complex traits as indicated by simulations (Spencer *et al.*, 2009; Visscher *et al.*, 2017). Although the overall genomic heritabilities estimated using all markers (0.296–0.528) accounted for a large proportion (64–89%) of the pedigree-based heritabilities, the discrete GWAS results contributed little genetic variance given the relatively low number of associations identified for these complex traits. Using the RHM approach, we identified 37 windows, 15 for DBH and 22 for HT (Table 4), each one encompassing 100 SNPs likely to be containing rare and

common variants. This approach was more effective than single-SNP GWAS to capture rare variants that do not have large enough effect to be declared significant at the genome-wide level, as observed earlier (Nagamine *et al.*, 2012; Riggio *et al.*, 2013; Resende *et al.*, 2017a). Some genomic windows identified by RHM individually explained 3–13% of the genomic heritability, similar to a previous study by Resende *et al.* (2017a). Additional genomic regions were identified using a Joint-GWAS approach with a larger number of individuals (Figs 3, 5).

Accounting for population structure and family-based relatedness in the single-SNP GWAS analysis, 356 significant SNPs were detected for DBH and HT. These included 210 (59%) associations within genes (184 unique genes) for all populations independently as well as for the combined dataset (50% within 60 unique genes). The Joint-GWAS from summary data identified another 30 genes, out of which 28 were detected using gene-based and two genes using segment-based models. We performed functional annotation of these genes and altogether they encompass different functional categories related to cell wall construction of growing tissues, cell wall cellulose biosynthetic process, RNA/DNA-binding and ion-binding, transporter activity, transcription factor activity, response to stimulus and others. Similar results were obtained for growth traits in *Populus* (Du *et al.*, 2016), suggesting that tree growth is controlled by multiple factors affecting cell division and meristems expansion requiring regulation of complex metabolic pathways with indirect effects on wood formation (Grattapaglia *et al.*, 2009).

Genes underlying the most significant associations were classified using gene ontology (GO) enrichment analysis for *E. grandis* terms with agriGO v2.0 (Tian *et al.*, 2017). Significant GO terms (FDR ≤ 5%) were identified encompassing four terms for biological process (single-organisms process, signaling,

Table 4 Regional heritability mapping (RHM) results for windows significantly ($-\log_{10}(P) > 3.0$) and putatively ($-\log_{10}(P) > 2.0$) associated with growth traits in the four *Eucalyptus* breeding populations.

Trait	Population	Chr.	SNP start	Position start (bp)	SNP end	Position end (bp)	LRT	h_r^2	$-\text{Log}_{10}(P)$
DBH	Pop1-IPB (626)[a]	7	EuBR07s31328798	31 328 798	EuBR07s32568262	32 568 262	10.76	0.13	2.98
		7	EuBR07s33146968	33 146 968	EuBR07s35110610	35 110 610	10.34	0.07	2.89
		7	EuBR07s32581478	32 581 478	EuBR07s33957780	33 957 780	10.19	0.07	2.85
		7	EuBR07s25714206	25 714 206	EuBR07s30352510	30 352 510	10.18	0.06	2.85
		7	EuBR07s31961430	31 961 430	EuBR07s33145504	33 145 504	9.53	0.07	2.69
		7	EuBR07s20583725	20 583 725	EuBR07s22342031	22 342 031	8.67	0.06	2.49
		7	EuBR07s22342887	22 342 887	EuBR07s25713595	25 713 595	8.53	0.06	2.46
		7	EuBR07s36784209	36 784 209	EuBR07s38583112	38 583 112	8.15	0.04	2.36
		7	EuBR07s21332450	21 332 450	EuBR07s24532470	24 532 470	7.93	0.05	2.31
		7	EuBR07s16684296	16 684 296	EuBR07s19147010	19 147 010	7.91	0.05	2.31
		7	EuBR07s28499843	28 499 843	EuBR07s31328715	31 328 715	7.38	0.06	2.18
		7	EuBR07s24545298	24 545 298	EuBR07s28499722	28 499 722	7.18	0.05	2.13
	Pop2-ARAB (683)[a]	6	EuBR06s32562797	32 562 797	EuBR06s34328105	34 328 105	7.96	0.05	2.32
	Pop3-ARAC (660)[a]	6	EuBR06s26699092	26 699 092	EuBR06s27751254	27 751 254	9.62	0.12	2.72
		9	EuBR09s31933985	31 933 985	EuBR09s33975533	33 975 533	8.02	0.07	2.33
HT	Pop1-IPB (626)[a]	2	EuBR02s16102502	16 102 502	EuBR02s18417551	18 417 551	7.03	0.04	2.10
		7	EuBR07s14930135	14 930 135	EuBR07s18103200	18 103 200	6.65	0.04	2.00
	Pop4-CNB (560)[a]	2	EuBR02s42263455	42 263 455	EuBR02s43397707	43 397 707	14.59	0.07	3.87[b]
		2	EuBR02s23849815	23 849 815	EuBR02s25008423	25 008 423	13.16	0.10	3.54[c]
		2	EuBR02s42780242	42 780 242	EuBR02s43864353	43 864 353	13.15	0.06	3.54[c]
		2	EuBR02s23213141	23 213 141	EuBR02s24367225	24 367 225	12.29	0.07	3.34[c]
		2	EuBR02s17118873	17 118 873	EuBR02s19701439	19 701 439	11.30	0.06	3.11[c]
		2	EuBR02s39648070	39 648 070	EuBR02s42778399	42 778 399	11.01	0.06	3.04[c]
		2	EuBR02s22594380	22 594 380	EuBR02s23832192	23 832 192	11.01	0.06	3.04[c]
		2	EuBR02s18112848	18 112 848	EuBR02s20898091	20 898 091	10.84	0.05	3.00[c]
		2	EuBR02s20898153	20 898 153	EuBR02s23179227	23 179 227	9.37	0.06	2.66
		2	EuBR02s36468959	36 468 959	EuBR02s39639880	39 639 880	9.16	0.07	2.61
		2	EuBR02s27662134	27 662 134	EuBR02s31703058	31 703 058	8.66	0.07	2.49
		2	EuBR02s29793530	29 793 530	EuBR02s32366267	32 366 267	8.54	0.06	2.46
		2	EuBR02s35689626	35 689 626	EuBR02s37602857	37 602 857	8.54	0.06	2.46
		2	EuBR02s24391700	24 391 700	EuBR02s25950796	25 950 796	8.33	0.06	2.41
		2	EuBR02s19809602	19 809 602	EuBR02s22590567	22 590 567	7.99	0.04	2.33
		2	EuBR02s15507347	15 507 347	EuBR02s18088025	18 088 025	7.65	0.04	2.25
		2	EuBR02s43397812	43 397 812	EuBR02s44360147	44 360 147	7.41	0.05	2.19
		2	EuBR02s31704444	31 704 444	EuBR02s33368834	33 368 834	7.13	0.05	2.12
		1	EuBR01s24700230	24 700 230	EuBR01s26451182	26 451 182	6.89	0.04	2.06
		1	EuBR01s17433597	17 433 597	EuBR01s19071730	19 071 730	6.74	0.03	2.03

DBH, diameter at breast height; HT, total height; LRT, likelihood ratio test; h_r^2, regional heritability.
[a]Total number of windows.
[b]Bonferroni correction with an experimental type I error rate of $\alpha = 0.05$.
[c]False discovery rate (FDR) of 5%.

localization and cellular component organization or biogenesis), four for cellular component (macromolecular complex, cell, organelle and membrane part) and two for molecular function (binding and transporter activity). The binding category, including DNA, RNA, protein and ion-binding, was the most represented (56%), which can better explain the growth trait heritability as it has more associations. This was also noted by Boyle *et al.* (2017), who showed a strong linear relationship between the sizes of the functional categories and the proportion of heritability that they explained.

Associations for growth pinpoint genes involved in cell wall biosynthesis

Our study focused on commonly measured growth traits that, together with wood specific gravity, constitute the mainstay of

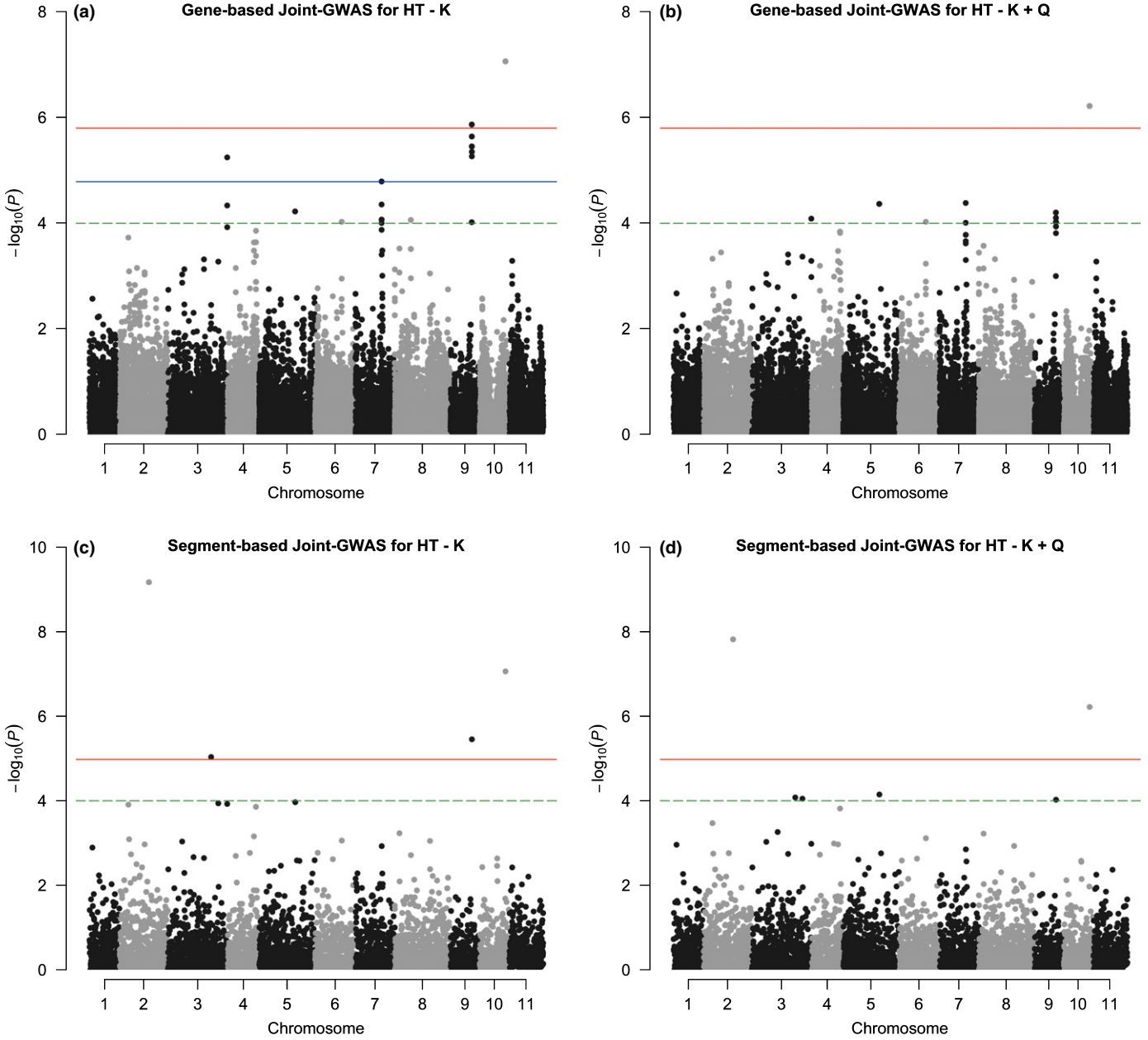

Fig. 5 Manhattan plots of the associations for total height (HT) using gene-based (31 770 genes) and segment-based (4766 windows) joint genome-wide association study (Joint-GWAS) for the combined dataset using four unrelated *Eucalyptus grandis* × *E. urophylla* hybrid breeding populations. (a) Gene-based Joint-GWAS adjusted for kinship matrix, age of measurements and population of origin. (b) Gene-based Joint-GWAS adjusted for all covariates mentioned before with the inclusion of population structure. (c) Segment-based Joint-GWAS adjusted for kinship matrix, age of measurements and population of origin. (d) Segment-based Joint-GWAS adjusted for all other covariates with the inclusion of population structure. Red line indicates Bonferroni-corrected threshold with an experimental type I error rate at $\alpha = 0.05$, blue line indicates a false discovery rate (FDR) at 5% and green dashed line represents an *ad hoc* threshold of $-\log_{10}(P) = 4.0$.

tree breeding and productivity in *Eucalyptus* and forest trees in general. Interestingly, some SNP associations map to genes were related to cell wall biosynthesis. In our Joint-GWAS analysis, the most significant SNP (Bonferroni at 5%) associated with DBH and HT (EuBR07s38098526, MAF = 0.015) was detected in the exon of the gene model (Eucgr.G02075/AT1G14720) encoding for xyloglucan endotransglucosylase/hydrolase 28 (XTH28) (Table 5). Another xyloglucan endotransglucosylase/hydrolase 5 (XTH5, Eucgr.G0190/AT5G13870), also located on chromosome 7, was detected using gene-based Joint-GWAS from summary data in HT with eight SNPs in the segment (Table 5), in which the top putatively associated SNP (lowest *P*-value) is a common variant (EuBR07s34941110, MAF = 0.123). Xyloglucan endotransglycosylase/hydrolase (XTH) enzymes act to remodel cell wall hemicelluloses, with various functions including wall strengthening and xylem formation (Bourquin, 2002; Cosgrove, 2005). Both XTH28 and XTH5 cleave and re-ligate xyloglucan polymers, a hemicellulose that is an essential constituent of the primary cell wall. Therefore, they participate in the cell wall development of growing tissues (Van Sandt *et al.*, 2007), with evident effect on root growth and cell wall extension (Maris *et al.*, 2009).

The Joint-GWAS approach also detected a common SNP (EuBR06s6100971, MAF = 0.423) that was putatively associated with DBH located on chromosome 6 in gene model Eucgr.F00486 (AT5G42100) and that encodes for a glucan endo-1,3-β-glucosidase (Table 5). This enzyme is a type of glycosyl hydrolase (GHs) whose function is the hydrolysis of any *O*-glycosyl bond (Lopez-Casado *et al.*, 2008). The hydrolysis of (1,3)-β-D-glucosidic linkages in (1,3)-β-D-glucans is important for carbohydrate metabolic process and cell wall organization (Lopez-Casado *et al.*, 2008). A GWAS in *Populus* (Du *et al.*, 2016) also detected an association in the glucan endo-1,3-β-glucosidase gene (Potri.018G000900). We also identified a significant SNP (EuBR04s17486529, FDR at 5%) in three of the four populations (Fig. 4) associated with DBH in the Eucgr.D00955 gene located on chromosome 4. This gene (Eucgr.D00955/AT4G17180) encodes an *O*-glycosyl hydrolases family 17 protein, another type of GHs. An additional significant SNP (EuBR05s70210869, FDR at 5%) shared between three populations was associated with total height on chromosome 5. This common variant in gene Eucgr.E04103 (AT1G61820), encoding a β-glucosidase 46 (BGLU46), which is also a type of GHs, may be involved in lignification by hydrolyzing monolignol glucosides (Escamilla-Treviño *et al.*, 2006).

The analysis of the larger number of shared associations among three of the four populations also showed a significant SNP (EuBR04s17531959, FDR at 5%) associated with DBH on chromosome 4 in a galacturonosyltransferase 4 (GAUT4) gene (Eucgr.D00963/AT5G47780). The GAUT4 is involved in pectin and xylans biosynthesis in cell walls, with a role in cell expansion and promoting growth (de Godoy *et al.*, 2013; Bryan *et al.*, 2016). Pectin is a structural heteropolysaccharide contained in the primary cell walls (Voragen *et al.*, 2009) and xylan is a type of hemicellulose (Studer *et al.*, 2011). Another common variant (EuBR10s8284185, FDR at 5%) identified in three populations was associated with DBH located on chromosome 10 in a xanthine dehydrogenase 1 (XDH1) gene (Eucgr.J00782/AT4G34890). The XDH1 is a key enzyme involved in purine catabolism and plays an important role during plant growth and development, senescence and response to stresses (Hesberg *et al.*, 2004; Yesbergenova *et al.*, 2005;

Table 5 Main associations detected in *Eucalyptus* breeding populations for growth traits (DBH, diameter at breast height; HT, total height) pinpointing genes involved in cell wall biosynthesis.

GWAS data	Trait	SNP	Chr.	Position (bp)	$-\text{Log}_{10}(P)$	REF/ALT	Eg – Gene	At – Gene	Annotation
Joint-GWAS Single-SNP	DBH	EuBR07s38098526	7	38 098 526	8.21[a]	G/A	Eucgr.G02075	AT1G14720	Xyloglucan endotransglucosylase/ hydrolase 28
Joint-GWAS Single-SNP	DBH	EuBR08s48262720	8	48 262 720	5.84[b]	A/G	Eucgr.H03281	AT3G06720	Armadillo/beta-catenin -like repeats-containing protein-related
Joint-GWAS Single-SNP	DBH	EuBR06s6100971	6	6100 971	4.10[c]	A/G	Eucgr.F00486	AT5G42100	Glucan 1, 3-beta-glucosidase A
Pop1-IPB/Pop2-ARAB/ Pop3-ARAC	DBH	EuBR04s17486529	4	17 486 529	–[b]	C/T	Eucgr.D00955	AT4G17180	O-Glycosyl hydrolases family 17 protein
Pop1-IPB/Pop2-ARAB/ Pop3-ARAC	DBH	EuBR04s17531959	4	17 531 959	–[b]	G/A	Eucgr.D00963	AT5G47780	Galacturonosyltransferase 4
Pop1-IPB/Pop2-ARAB/ Pop3-ARAC	DBH	EuBR10s8284185	10	8284 185	–[b]	G/A	Eucgr.J00782	AT4G34890	Xanthine dehydrogenase 1
Joint-GWAS Gene-based	HT	EuBR07s34941110	7	34 941 110	4.36[c]	G/A	Eucgr.G01909	AT5G13870	Xyloglucan endotransglucosylase/ hydrolase 5
Pop1-IPB/Pop2-ARAB/ Pop3-ARAC	HT	EuBR05s70210869	5	70 210 869	–[b]	C/T	Eucgr.E04103	AT1G61820	Beta glucosidase 46

There are more than one value (–).
[a]Bonferroni correction with an experimental type I error rate of $\alpha = 0.05$.
[b]False discovery rate (FDR) of 5%.
[c]*Ad hoc* threshold of $-\log_{10}(P) \geq 4$.

Nakagawa *et al.*, 2007). The simultaneous silencing of XDH1 and XDH2 showed reduced growth in *Arabidopsis* (Nakagawa *et al.*, 2007).

The significant association with DBH at 5% FDR threshold on chromosome 8 in the Joint-GWAS analysis (Fig. 3, blue line) is located inside gene model Eucgr.H03281, a gene encoding for an armadillo/beta-catenin-like repeats-containing protein-related, whose function is involved in the cellulose biosynthetic process. In a recent GWAS study in another *Eucalyptus* species, *E. pellita*, we also found a significant association for growth inside Eucgr.F03806, a gene that codes for another armadillo/beta-catenin-like repeat positioned on a different chromosome (6) (Müller *et al.*, 2017). This gene in *Arabidopsis thaliana* (AT1G77460) transcribes the protein cellulose synthase interactive 3 (CSI3), which regulates primary cell wall biosynthesis and cellulose microfibrils organization (Lei *et al.*, 2013).

Concluding remarks

In this study, we carried out a GWAS for growth traits by gathering a considerably larger association population than previous forest tree studies, with 3373 individuals across four breeding populations of *Eucalyptus* in an attempt to evaluate the impact of a large sample size on the ability to detect discrete associations. We tested several GWAS models with variable levels of correction for population stratification and relatedness and different segment-based approaches in an effort to capture a wider frequency spectrum of variants. Although the different associated genes identified in our study will require further validation, consistency with GWAS results from other studies provides valuable preliminary leads for further investigation. It is noteworthy that it was claimed that the first approved *Eucalyptus* transgenic (Nature News, 2015) produced between 4 and 20% more wood than the wild type (Ledford, 2014). This transgenic was engineered to contain an endo-1,4-β-glucanase (CEL1) from Arabidopsis that affects plant growth (Shani *et al.*, 2006), this gene is related to the cellulose synthase-like C family that encodes a β-1,4 glucan synthase (Cocuron *et al.*, 2007). The identification of several genes involved in cell wall biosynthesis in our study may therefore provide new targets for transgenic and genome editing approaches.

Overall, our results do not differ substantially from those reported in GWAS for growth traits in different forest trees to date. However, the access to different large populations allowed us to provide evidence of validation of marker-trait associations, a cornerstone for the scientific credibility of GWAS results. We found that some SNPs or genes were associated with growth across independent populations, to the best of our knowledge, the first such results in forest trees. Our results also further corroborate the evidence that growth is controlled by many variants of relatively small effect, such that the infinitesimal model fits the data well. While slightly more encouraging GWAS results have been reported for simpler wood properties and phenology traits, the large proportions of genomic heritability that we captured by whole-genome data for growth point to the fact that genomic prediction approaches could be considerably more efficient for tree breeding, at least for the time being (Grattapaglia, 2017). However, as more tree breeding programs start to adopt genomic data to predict phenotypes using a publicly shared SNP platform like the EUChip60K used in this work, the much larger collective datasets should provide increased power to dissect individual associations not only to accelerate breeding, but also to advance the mechanistic understanding of the complex relationships between sequence variation and phenotypes.

Acknowledgements

This work was supported by PRONEX-FAP-DF grant 2009/00106-8 and CNPq grant 400663/2012-0 to DG. BSFM had a doctoral fellowship and DG a research fellowship from CNPq. We acknowledge the managers of FIBRIA, CENIBRA and INTERNATIONAL PAPER of Brazil for providing logistic support in the field trials. The authors gratefully thank Dr Alexandre Coelho for useful comments and suggestions.

Author contributions

BSFM, LGN and DG planned and designed the experiment. BML, CCG, AM, AMA, and ET conducted the field experiments and collected the phenotypic data. OBS-J and DG generated the SNP data. BSFM performed the analysis and wrote the first version of the manuscript. MK contributed with computing infrastructure. SAG assisted with statistical analysis. LGN and JEAF made substantial contributions to the bioinformatics, GWAS analyses and interpretation of data. BSFM, LGN and DG wrote the final version of the manuscript. DG coordinated the study and edited the final version of the manuscript. All authors have read and approved the manuscript.

ORCID

Janeo E. de Almeida Filho http://orcid.org/0000-0002-3906-8884
Salvador A. Gezan http://orcid.org/0000-0002-5025-2658
Dario Grattapaglia http://orcid.org/0000-0002-0050-970X
Matias Kirst http://orcid.org/0000-0002-8186-3945
Bárbara S. F. Müller http://orcid.org/0000-0002-3494-139X
Leandro G. Neves http://orcid.org//0000-0002-8365-6827
Orzenil B. Silva-Junior http://orcid.org/0000-0002-2104-2010

References

Allwright MR, Payne A, Emiliani G, Milner S, Viger M, Rouse F, Keurentjes JJB, Bérard A, Wildhagen H, Faivre-Rampant P *et al.* 2016. Biomass traits and candidate genes for bioenergy revealed through association genetics in coppiced European *Populus nigra* (L.). *Biotechnology for Biofuels* **9**: 1–22.

Bakshi A, Zhu Z, Vinkhuyzen AAE, Hill WD, McRae AF, Visscher PM, Yang J. 2016. Fast set-based association analysis using summary data from GWAS identifies novel gene loci for human complex traits. *Scientific Reports* **6**: 1–9.

Bartholomé J, Bink MC, Van Heerwaarden J, Chancerel E, Boury C, Lesur I, Isik F, Bouffier L, Plomion C. 2016. Linkage and association mapping for two major traits used in the maritime pine breeding program: height growth and stem straightness. *PLoS ONE* **11**: 1–21.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Royal Statistical Society* **57**: 289–300.

Bernal Rubio YL, Gualdrõn Duarte JL, Bates RO, Ernst CW, Nonneman D, Rohrer GA, King A, Shackelford SD, Wheeler TL, Cantet RJC *et al.* 2016. Meta-analysis of genome-wide association from genomic prediction models. *Animal Genetics* **47**: 36–48.

Borcard D, Gillet F, Legendre P. 2011. *Numerical ecology with R.* New York, NY, USA: Springer-Verlag.

Bourquin V. 2002. Xyloglucan endotransglycosylases have a function during the formation of secondary cell walls of vascular tissues. *The Plant Cell* **14**: 3073–3088.

Boyle EA, Li YI, Pritchard JK. 2017. An expanded view of complex traits: from polygenic to omnigenic. *Cell* **169**: 1177–1186.

Bryan AC, Jawdy S, Gunter L, Gjersing E, Sykes R, Hinchee MAW, Winkeler KA, Collins CM, Engle N, Tschaplinski TJ *et al.* 2016. Knockdown of a laccase in *Populus deltoides* confers altered cell wall chemistry and increased sugar release. *Plant Biotechnology Journal* **14**: 2010–2020.

Cappa EP, El-Kassaby YA, Garcia MN, Acuña C, Borralho NMG, Grattapaglia D, Marcucci Poltri SN. 2013. Impacts of population structure and analytical models in genome-wide association studies of complex traits in forest trees: a case study in *Eucalyptus globulus*. *PLoS ONE* **8**: 1–16.

Cocuron J-C, Lerouxel O, Drakakaki G, Alonso AP, Liepman AH, Keegstra K, Raikhel N, Wilkerson CG. 2007. A gene from the cellulose synthase-like C family encodes a beta-1,4 glucan synthase. *Proceedings of the National Academy of Sciences, USA* **104**: 8550–8555.

Cosgrove DJ. 2005. Growth of the plant cell wall. *Nature Reviews Molecular Cell Biology* **6**: 850–861.

Du Q, Gong C, Wang Q, Zhou D, Yang H, Pan W, Li B, Zhang D. 2016. Genetic architecture of growth traits in *Populus* revealed by integrated quantitative trait locus (QTL) analysis and association studies. *New Phytologist* **209**: 1067–1082.

Escamilla-Treviño LL, Chen W, Card ML, Shih MC, Cheng CL, Poulton JE. 2006. *Arabidopsis thaliana* beta-glucosidases *BGLU45* and *BGLU46* hydrolyse monolignol glucosides. *Phytochemistry* **67**: 1651–1660.

Evangelou E, Ioannidis JPA. 2013. Meta-analysis methods for genome-wide association studies and beyond. *Nature Reviews Genetics* **14**: 379–389.

Evanno G, Regnaut S, Goudet J. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology* **14**: 2611–2620.

Evans LM, Slavov GT, Rodgers-Melnick E, Martin J, Ranjan P, Muchero W, Brunner AM, Schackwitz W, Gunter L, Chen J-G *et al.* 2014. Population genomics of *Populus trichocarpa* identifies signatures of selection and adaptive trait associations. *Nature Genetics* **46**: 1089–1096.

Fahrenkrog AM, Neves LG, Resende MFR, Vazquez AI, de los Campos G, Dervinis C, Sykes R, Davis M, Davenport R, Barbazuk WB *et al.* 2016. Genome-wide association study reveals putative regulators of bioenergy traits in *Populus deltoides*. *New Phytologist* **213**: 799–811.

Francis RM. 2016. POPHELPER: an R package and web app to analyse and visualise population structure. *Molecular Ecology Resources* **17**: 27–32.

de Godoy F, Bermúdez L, Lira BS, De Souza AP, Elbl P, Demarco D, Alseekh S, Insani M, Buckeridge M, Almeida J *et al.* 2013. Galacturonosyltransferase 4 silencing alters pectin composition and carbon partitioning in tomato. *Journal of Experimental Botany* **64**: 2449–2466.

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*. Cham, Switzerland: Springer International, 199–250.

Grattapaglia D, Plomion C, Kirst M, Sederoff RR. 2009. Genomics of growth traits in forest trees. *Current Opinion in Plant Biology* **12**: 148–156.

Guerra FP, Wegrzyn JL, Sykes R, Davis MF, Stanton BJ, Neale DB. 2013. Association genetics of chemical wood properties in black poplar (*Populus nigra*). *New Phytologist* **197**: 162–176.

Hamblin MT, Buckler ES, Jannink JL. 2011. Population genetics of genomics-based crop improvement methods. *Trends in Genetics* **27**: 98–106.

Hesberg C, Hänsch R, Mendel RR, Bittner F. 2004. Tandem orientation of duplicated xanthine dehydrogenase genes from *Arabidopsis thaliana*: differential gene expression and enzyme activities. *Journal of Biological Chemistry* **279**: 13547–13554.

Hill WG, Weir BS. 1988. Variances and covariances of squared linkage disequilibria in finite populations. *Theoretical Population Biology* **33**: 54–78.

Jackson DA. 1993. Stopping rules in principal components analysis: a comparison of Heuristical and statistical approaches. *Ecology* **74**: 2204–2214.

Jaramillo-Correa JP, Prunier J, Vázquez-Lobo A, Keller SR, Moreno-Letelier A. 2015. Molecular signatures of adaptation and selection in forest trees. *Advances in Botanical Research* **74**: 1–42.

Kaler AS, Ray JD, Schapaugh WT, King CA, Purcell LC. 2017. Genome-wide association mapping of canopy wilting in diverse soybean genotypes. *Theoretical and Applied Genetics* **130**: 2203–2217.

Kang HM, Sul JH, Service SK, Zaitlen NA, Kong S, Freimer NB, Sabatti C, Eskin E. 2010. Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics* **42**: 348–354.

Khan MA, Korban SS. 2012. Association mapping in forest trees and fruit crops. *Journal of Experimental Botany* **63**: 4045–4060.

Ledford H. 2014. Brazil considers transgenic trees. *Nature* **512**: 357.

Lei L, Li S, Du J, Bashline L, Gu Y. 2013. Cellulose synthase interactive 3 regulates cellulose biosynthesis in both a microtubule-dependent and microtubule-independent manner in *Arabidopsis*. *The Plant Cell* **25**: 4912–4923.

Li YX, Li C, Bradbury PJ, Liu X, Lu F, Romay CM, Glaubitz JC, Wu X, Peng B, Shi Y *et al.* 2016. Identification of genetic variants associated with maize flowering time using an extremely large multi-genetic background population. *Plant Journal* **86**: 391–402.

Lima BM. 2014. *Bridging genomics and quantitative genetics of Eucalyptus: genome-wide prediction and genetic parameter estimation for growth and wood properties using high-density SNP data.* PhD thesis, University of São Paulo, Piracicaba, SP, Brazil, 92.

Lin DY, Zeng D. 2009. Meta-analysis of genome-wide association studies: no efficiency gain in using individual participant data. *Genetic Epidemiology* **34**: 1–12.

Lopez-Casado G, Urbanowicz BR, Damasceno CM, Rose JK. 2008. Plant glycosyl hydrolases and biofuels: a natural marriage. *Current Opinion in Plant Biology* **11**: 329–337.

Lu M, Krutovsky KV, Nelson CD, West JB, Reilly NA, Loopstra CA. 2017. Association genetics of growth and adaptive traits in loblolly pine (*Pinus taeda* L.) using whole-exome-discovered polymorphisms. *Tree Genetics and Genomes* **13**: 1–18.

Mägi R, Morris AP. 2010. GWAMA: software for genome-wide association meta-analysis. *BMC Bioinformatics* **11**: 1–6.

Magosi LE, Goel A, Hopewell JC, Farrall M. 2017. Identifying systematic heterogeneity patterns in genetic association meta-analysis studies. *PLoS Genetics* **13**: 1–17.

Maris A, Suslov D, Fry SC, Verbelen JP, Vissenberg K. 2009. Enzymic characterization of two recombinant xyloglucan endotransglucosylase/hydrolase (*XTH*) proteins of *Arabidopsis* and their effect on root growth and cell wall extension. *Journal of Experimental Botany* **60**: 3959–3972.

Marroni F, Pinosio S, Zaina G, Fogolari F, Felice N, Cattonaro F, Morgante M. 2011. Nucleotide diversity and linkage disequilibrium in *Populus nigra* cinnamyl alcohol dehydrogenase (*CAD4*) gene. *Tree Genetics and Genomes* **7**: 1011–1023.

McKown AD, Klápště J, Guy RD, Geraldes A, Porth I, Hannemann J, Friedmann M, Muchero W, Tuskan GA, Ehlting J *et al.* 2014. Genome-wide association implicates numerous genes underlying ecological trait variation in natural populations of *Populus trichocarpa*. *New Phytologist* **203**: 535–553.

Müller BSF, Neves LG, de Almeida Filho JE, Resende MFR, Muñoz PR, dos Santos PET, Filho EP, Kirst M, Grattapaglia D. 2017. Genomic prediction in contrast to a genome-wide association study in explaining heritable variation of complex growth traits in breeding populations of *Eucalyptus*. *BMC Genomics* **18**: 524.

Myburg A, Grattapaglia D, Tuskan G, Hellsten U, Hayes RD, Grimwood J, Jenkins J, Lindquist E, Tice H, Bauer D *et al.* 2014. The genome of *Eucalyptus grandis*. *Nature* **509**: 356–362.

Myburg AA, Potts BM, Marques CM, Kirst M, Gion J-M, Grattapaglia D, Grima-Pettenatti J. 2007. Eucalypts. In: Kole C, ed. *Genome mapping and*

molecular breeding in plants, Volume 7 Forest Trees. Berlin/Heidelberg, Germany: Springer-Verlag, 115–160.

Nagamine Y, Pong-Wong R, Navarro P, Vitart V, Hayward C, Rudan I, Campbell H, Wilson J, Wild S, Hicks AA *et al.* 2012. Localising loci underlying complex trait variation using regional genomic relationship mapping. *PLoS ONE* **7**: 1–12.

Nakagawa A, Sakamoto S, Takahashi M, Morikawa H, Sakamoto A. 2007. The RNAi-mediated silencing of xanthine dehydrogenase impairs growth and fertility and accelerates leaf senescence in transgenic *Arabidopsis* plants. *Plant and Cell Physiology* **48**: 1484–1495.

Nature News. 2015. Brazil approves transgenic *Eucalyptus*. *Nature Biotechnology* **33**: 577.

Neale DB. 2007. Genomics to tree breeding and forest health. *Current Opinion in Genetics & Development* **17**: 539–544.

Neale DB, Savolainen O. 2004. Association genetics of complex traits in conifers. *Trends in Plant Science* **9**: 325–330.

Pérez P, de los Campos G. 2014. Genome-wide regression and prediction with the BGLR statistical package. *Genetics* **198**: 483–495.

Porth I, Klápště J, Skyba O, Hannemann J, McKown AD, Guy RD, Difazio SP, Muchero W, Ranjan P, Tuskan GA *et al.* 2013. Genome-wide association mapping for wood characteristics in *Populus* identifies an array of candidate single nucleotide polymorphisms. *New Phytologist* **200**: 710–726.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics* **155**: 945–959.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MARR, Bender D, Maller J, Sklar P, De Bakker PIW, Daly MJ *et al.* 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics* **81**: 559–575.

Remington DL, Thornsberry JM, Matsuoka Y, Wilson LM, Whitt SR, Doebley J, Kresovich S, Goodman MM, Buckler ES. 2001. Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proceedings of the National Academy of Sciences, USA* **98**: 11479–11484.

Resende MDV, Resende MFR Jr, Sansaloni CP, Petroli CD, Missiaggia AA, Aguiar AM, Abad JM, Takahashi EK, Rosado AM, Faria DA *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 Phytologist* **194**: 116–128.

Resende RT, Resende MDV, Silva FF, Azevedo CF, Takahashi EK, Silva-Junior OB, Grattapaglia D. 2017a. Regional heritability mapping and genome-wide association identify loci for complex growth, wood and disease resistance traits in Eucalyptus. *New Phytologist* **213**: 1287–1300.

Resende RT, Resende MDV, Silva FF, Azevedo CF, Takahashi EK, Silva-Junior OB, Grattapaglia D. 2017b. Assessing the expected response to genomic selection of individuals and families in *Eucalyptus* breeding with an additive-dominant model. *Heredity* **119**: 245–255.

Riggio V, Matika O, Pong-Wong R, Stear MJ, Bishop SC. 2013. Genome-wide association and regional heritability mapping to identify loci underlying variation in nematode resistance and body weight in Scottish Blackface lambs. *Heredity* **110**: 420–429.

Schork AJ, Thompson WK, Pham P, Torkamani A, Roddey JC, Sullivan PF, Kelsoe JR, O'Donovan MC, Furberg H, Schork NJ *et al.* 2013. All SNPs are not created equal: genome-wide association studies reveal a consistent pattern of enrichment among functionally annotated SNPs. *PLoS Genetics* **9**: 1–13.

Shani Z, Dekel M, Roiz L, Horowitz M, Kolosovski N, Lapidot S, Alkan S, Koltai H, Tsabary G, Goren R *et al.* 2006. Expression of endo-1,4-β-glucanase (*cel1*) in *Arabidopsis thaliana* is associated with plant growth, xylem development and cell wall thickening. *Plant Cell Reports* **25**: 1067–1074.

Silva-Junior OB, Faria DA, Grattapaglia D. 2015. A flexible multi-species genome-wide 60K SNP chip developed from pooled resequencing of 240 *Eucalyptus* tree genomes across 12 species. *New Phytologist* **206**: 1527–1540.

Spencer CCA, Su Z, Donnelly P, Marchini J. 2009. Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS Genetics* **5**: 1–13.

Studer MH, DeMartini JD, Davis MF, Sykes RW, Davison B, Keller M, Tuskan GA, Wyman CE. 2011. Lignin content in natural *Populus* variants affects sugar release. *Proceedings of the National Academy of Sciences, USA* **108**: 6300–6305.

Thavamanikumar S, McManus LJ, Ades PK, Bossinger G, Stackpole DJ, Kerr R, Hadjigol S, Freeman JS, Vaillancourt RE, Zhu P *et al.* 2014. Association mapping for wood quality and growth traits in *Eucalyptus globulus* ssp. *globulus* Labill identifies nine stable marker-trait associations for seven traits. *Tree Genetics and Genomes* **10**: 1661–1678.

Thumma BR, Nolan MF, Evans R, Moran GF. 2005. Polymorphisms in cinnamoyl CoA reductase (*CCR*) are associated with variation in microfibril angle in *Eucalyptus* spp. *Genetics* **171**: 1257–1265.

Tian T, Liu Y, Yan H, You Q, Yi X, Du Z, Xu W, Su Z. 2017. AgriGO v2.0: a GO analysis toolkit for the agricultural community, 2017 update. *Nucleic Acids Research* **45**: 122–129.

Turner SD. 2018. qqman: an R package for visualizing GWAS results using Q-Q and Manhattan plots. *Journal of Open Source Software* **3**: 731.

Van Sandt VST, Suslov D, Verbelen JP, Vissenberg K. 2007. Xyloglucan endotransglucosylase activity loosens a plant cell wall. *Annals of Botany* **100**: 1467–1473.

Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J. 2017. 10 years of GWAS discovery: biology, function, and translation. *American Journal of Human Genetics* **101**: 5–22.

Voragen AGJ, Coenen GJ, Verhoef RP, Schols HA. 2009. Pectin, a versatile polysaccharide present in plant cell walls. *Structural Chemistry* **20**: 263–275.

Wallace JG, Zhang X, Beyene Y, Semagn K, Olsen M, Prasanna BM, Buckler ES. 2016. Genome-wide association for plant height and flowering time across 15 tropical maize populations under managed drought stress and well-watered conditions in sub-Saharan Africa. *Crop Science* **56**: 2365–2378.

Wegrzyn JL, Eckert AJ, Choi M, Lee JM, Stanton BJ, Sykes R, Davis MF, Tsai CJ, Neale DB. 2010. Association genetics of traits controlling lignin and cellulose biosynthesis in black cottonwood (*Populus trichocarpa*, Salicaceae) secondary xylem. *New Phytologist* **188**: 515–532.

Weir BS, Cockerham CC. 1984. Estimating F-statistics for the analysis of population structure. *Evolution* **38**: 1358–1370.

Wickham H. 2009. *ggplot2: elegant graphics for data analysis*. New York, NY, USA: Springer-Verlag.

Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. 2011. Rare-variant association testing for sequencing data with the sequence kernel association test. *American Journal of Human Genetics* **89**: 82–93.

Wu X, Li Y, Shi Y, Song Y, Zhang D, Li C, Buckler ES, Li Y, Zhang Z, Wang T. 2016. Joint-linkage mapping and GWAS reveal extensive genetic loci that regulate male inflorescence size in maize. *Plant Biotechnology Journal* **14**: 1551–1562.

Xiao Y, Liu H, Wu L, Warburton M, Yan J. 2017. Genome-wide association studies in maize: praise and stargaze. *Molecular Plant* **10**: 359–374.

Yang J, Ferreira T, Morris AP, Medland SE, Madden PAF, Heath AC, Martin NG, Montgomery GW, Weedon MN, Loos RJ *et al.* 2012. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nature Genetics* **44**: 369–375.

Yang J, Lee SH, Goddard ME, Visscher PM. 2011. GCTA: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics* **88**: 76–82.

Yesbergenova Z, Yang G, Oron E, Soffer D, Fluhr R, Sagi M. 2005. The plant Mo-hydroxylases aldehyde oxidase and xanthine dehydrogenase have distinct reactive oxygen species signatures and are induced by drought and abscisic acid. *Plant Journal* **42**: 862–876.

Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. 2012. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**: 3326–3328.

Supporting Information

Additional Supporting Information may be found online in the Supporting Information section at the end of the article:

Fig. S1 Phenotypic distributions with density line of the growth traits measured in the four *Eucalyptus grandis* × *E. urophylla* hybrid breeding populations.

Fig. S2 Distribution of the number of SNPs into MAF classes for each *Eucalyptus* population and combined data (All) using CR ≥ 90% and MAF > 0.

Fig. S3 Quantile−quantile (QQ) plots for SNP-based models for diameter at breast height (DBH) and total height (HT), respectively, in *Eucalyptus*.

Fig. S4 Manhattan plots for SNP-based models in *Eucalyptus* Pop1-IPB.

Fig. S5 Manhattan plots for SNP-based models in *Eucalyptus* Pop2-ARAB.

Fig. S6 Manhattan plots for SNP-based models in *Eucalyptus* Pop3-ARAC.

Fig. S7 Manhattan plots for SNP-based models in *Eucalyptus* Pop4-CNB.

Fig. S8 Manhattan plots using SNP-based models for Joint-GWAS for the four *Eucalyptus* populations combined dataset.

Methods S1 Heritability estimation.

Methods S2 Statistical analysis of phenotypic data.

Methods S3 GWAS models.

Methods S4 Contribution of individual associations to the phenotypic variance.

Table S1 Genotypic data information and number of subpopulations determined using STRUCTURE and PCA for the four *Eucalyptus* populations.

Table S2 Estimates of additive genetic variances (σ2a), residual variances (σ2e), phenotypic variances (σ2p) and narrow-sense heritabilities (h2) in *Eucalyptus*.

Please note: Wiley Blackwell are not responsible for the content or functionality of any Supporting Information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.