Integrating genome-wide association mapping of additive and dominance genetic effects to improve genomic prediction accuracy in *Eucalyptus*

Biyue Tan[1,2], Pär K. Ingvarsson[3*]

[1] Umeå Plant Science Centre, Department of Ecology and Environmental Science, Umeå University, SE-90187, Umeå, Sweden

[2] Stora Enso AB, SE-131 04, Nacka, Sweden

[3] Department of Plant Biology, Uppsala BioCenter, Swedish University of Agricultural Sciences, SE-750 07 Uppsala, Sweden

* Correspondence: par.ingvarsson@slu.se

Word count: 6922

Number of figures: 4

Running head: GWAS and GS in Eucalyptus

Summary

Genome-wide association studies (GWAS) is a powerful and widely used approach to decipher the genetic control of complex traits. A major challenge for dissecting quantitative traits in forest trees is statistical power. In this study, we use a population consisting of 1123 samples from two successive generations that have been phenotyped for growth and wood property traits and genotyped using the EuChip60K chip, yielding 37,832 informative SNPs. We use multi-locus GWAS models to assess both additive and dominance effects to identify markers associated with growth and wood property traits in the eucalypt hybrids. Additive and dominance association models identified 78 and 82 significant SNPs across all traits, respectively, which captured between 39 and 86% of the genomic-based heritability. We also used SNPs identified from the GWAS and SNPs using less stringent significance thresholds to evaluate predictive abilities in a genomic selection framework. Genomic selection models based on the top 1% SNPs captured a substantially greater proportion of the genetic variance of traits compared to when all SNPs were used for model training. The prediction ability of estimated breeding values was significantly improved for all traits using either the top 1% SNPs or SNPs identified using a relaxed *p*-value threshold ($p<10^{-3}$). This study highlights the added value of also considering dominance effects for identifying genomic regions controlling growth traits in trees. Moreover, integrating GWAS results into genomic selection method provides enhanced power

40 relative to discrete associations for identifying genomic variation potentially useful
41 in tree breeding.

42 **Keywords**
43 *Eucalyptus*, dominance, FarmCPU, genome-wide association study, genomic
44 selection

45

Introduction

Deciphering the genetic basis of complex phenotypic traits is of fundamental importance for understanding biological processes and may ultimately provide information that can help enhance selection in plant breeding programs. Genome-wide association studies (GWAS) is a powerful way to identify putative causal genes or genomic segments that underlie phenotypic variation in plants, particularly for traits with complex genetic architectures (Ingvarsson and Street, 2011; Kruglyak, 2008). Dissection of complex traits have been undertaken in forest genetics to understand the genetic basis of adaptive phenotypes (Ingvarsson et al., 2008; Olson et al., 2013; Wang et al., 2018) or physiological or morphological traits, such as growth or wood properties. For example, Porth *et. al.* (Porth et al., 2013) and later Chhetri *et. al.* (Chhetri et al., 2019a) performed GWAS for wood traits, biomass, eco-physiological and phenology traits in *Populus trichocarpa* with genotyping based on 6.78 million single nucleotide polymorphisms (SNPs). Similarly, a study of *Salix viminalis* identified 29 SNPs that were associated with bud burst, leaf senescence, number of shoots or shoot diameter (Hallingback et al., 2016). In *Eucalyptus,* the earliest GWAS identified 16 markers that were associated with growth and two markers that were associated with lignin traits (Cappa et al., 2013). Recently, 26 quantitative trait loci (QTLs) were identified for productivity and disease resistance using a regional heritability mapping method that helps increase the genomic heritability to 5-15% from 4-6% when using SNPs individually(Resende et al., 2017a; Resende et al., 2017b).

GWAS studies can also provide tools for accelerating the long breeding cycles in tree breeding (reviewed in (Neale and Kremer, 2011)). For example, although many species of *Eucalyptus* display unusually fast growth, breeding cycles aimed at developing elite commercial genotypes still take between 12 to 16 years to complete, since identification of elite genotypes require progeny trials followed by two or more sequential clonal trials (Rezende et al., 2014). However, genomic selection based on genome-wide molecular makers is expected to reduce the time required for completing a cycle of developing elite clones to only 9 years mainly due to the shorter time needed for progeny tests when phenotypes can be predicted from the genomic selection models (Grattapaglia, 2017).

The rapid development in genomics has opened up opportunities to identify molecular markers that are associated with traits of interest and use these marker-trait associations to complement and extend traditional breeding programs. Despite the efforts to discover polymorphisms associated with economically relevant traits, much of the genetic contribution to complex traits in forest trees remains unexplained. One of the main reasons is that GWAS methods normally conduct tests on one marker at a time, for instance using a generalized linear model (GLM) or a mixed linear model (MLM). When dealing with complex traits such as growth and wood qualities, where the effect size of individual loci is likely small to moderate, these methods suffer from limited statistical power to detect loci of small effects (Muller et al., 2017). One potential approach to increase the power and to accurately identify more causal variants is so called ‘multi-locus mixed models’

(MLMM), which simultaneously test multiple markers by including them as covariates in a stepwise MLM to partially remove confounding between tested markers and kinship (Segura et al., 2012). One such method is the 'fixed and random model circulating probability unification' (FarmCPU) that performs marker tests using other associated markers as covariates in a fixed effect model (Liu et al., 2016). Optimization across the associated covariate markers using a random effect model is then performed separately. This approach has been reported to simultaneously reduce computational complexity, remove confounding between population structure, kinship and quantitative trait loci, prevent model over-fitting and control the number of false positives (Liu et al., 2016).

Most GWAS analyses to date have been undertaken by implicitly assuming a genetic architecture consisting of additive effects. However, non-additive effects, including dominance (Bruce, 1910), over-dominance (Crow, 1948) and epistasis (Hill, 1982) are known to also play important roles in controlling some traits. One trait where non-additive effects are likely to be pronounced is heterosis, or hybrid vigor, which is the near universally observed phenomenon of phenotypic superiority of hybrid progeny relative to their parents (Charlesworth and Willis, 2009). Not surprisingly, heterosis has been and continues to be of great importance in most plant breeding schemes (Duvick, 2001). To date, a limited number of studies have utilized GWAS methods to dissect the genetic basis of heterotic traits in *Arabidopsis thaliana* and rice. In the model plant *A.thaliana*, dominance and over-dominance of flowering time is a well-studied trait and significant loci from a

GWAS were shown to explain as much as 20% of the phenotypic variation in a hybrid population consisting of 435 individuals derived from inter-crossing 30 parents (Seymour et al., 2016). In rice, genome-wide dissection uncovered multiple non-additive effect loci for yield increase (Li et al., 2016; Zhen et al., 2017). For instance, a major QTL, rice heterosis 8 (*RH8*) was found to regulate grain-yield component traits (Li et al., 2016). In *Eucalyptus* hybrids dominance appears to be an important and widespread contributor to many growth-related traits (Bison et al., 2006; Bouvet and Vigneron, 1995; Volker et al., 2008) and ratios of dominance to additive variances exceeding 1.2 have been estimated for growth in *E. grandis* x *E. urophylla* hybrids (Bouvet et al., 2009; Makouanzi et al., 2014; Tan et al., 2017). Such results suggest that there should be ample opportunities to identify SNPs accounting for dominance and/or over-dominance effects in *Eucalyptus* hybrids.

Another genomic-based approach that has become widely used in plant and animal breeding in recent years is genomic selection (GS) or alternatively known as genomic prediction. Unlike GWAS, GS refers to marker-based selection where total genetic variance is captured using genome-wide markers without a prior step of identifying trait-associated markers. GS aims to predict the genetic potential (e.g. genome-estimated breeding values) of breeding individuals without locating genes or QTLs important for the trait(s) of interest. One of most important questions for GS is how to improve the prediction accuracy and methods for accuracy has long been a central research aim in genomic selection. Thus far progress on increasing prediction accuracies have been achieved through the development of new

34 statistical models, more efficient design of training populations, improved quality
35 of phenotypic measurements, a greater number of makers used for model building
36 and by also considering non-additive effects (Grattapaglia, 2017). In this paper we
37 assess methods for improving genomic prediction accuracy by integrating results
38 from GWAS studies into GS to predict the genetic potential of breeding targets. It
39 is well known that using only associated SNPs identified from a GWAS is usually
40 not sufficient for explaining a large fraction of the genetic variation in a trait of
41 interest (the so called "missing heritability" problem, (Makowsky et al., 2011)).
42 However, utilizing GWAS information in the form of associated SNPs, in
43 combination with other types of data has the potential to enhance prediction ability
44 in GS studies (Gowda et al., 2015).

45 In this study, we present results from a GWAS on growth and wood quality
46 traits, in a breeding population comprising two species of *Eucalyptus* and their
47 hybrids. We also integrate the GWAS results in a GS model with the goal of
48 assessing whether this can help increase prediction accuracies for the traits in
49 question. Specifically, our study has two objectives: first, we implement a state of
50 the art GWAS method that consider both additive and dominance effects for
51 dissecting the genetic architecture of growth and wood quality traits. We also
52 evaluate the proportion of phenotypic variation that explained by significant loci
53 for these two genetic effects. Second, we evaluate how different categories of
54 informative SNPs, selected based on the results from the GWAS, can be
55 implemented in a widely used model for genomic prediction, GBLUP, to estimate

variance components and to evaluate prediction accuracies of estimated breeding values.

Results

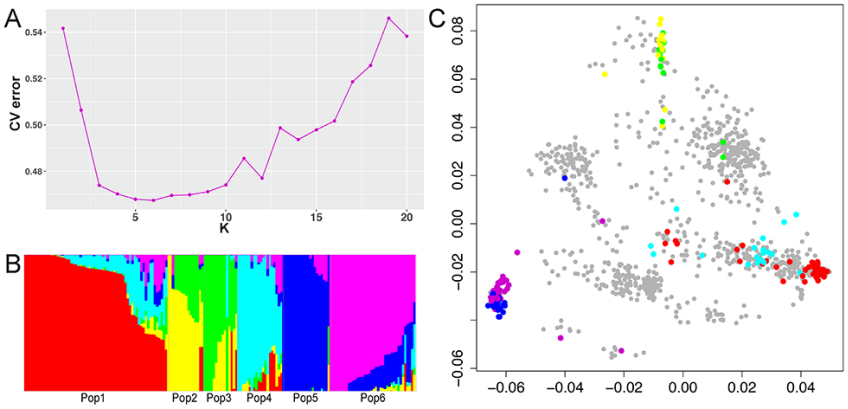
Characters of growth and wood traits

All growth traits were moderately variable at the different assessment ages (Table 1). We observed a lower phenotypic variation for height at 3 years of age, as judged by the coefficient of variation (Table 1). The F1 population underwent selection based on height in order to identify trees to use for genotyping and this selection process likely contributed to the lower phenotypic variation we see in height at 3 years of age. We also observed low phenotypic variation for basic density and pulp yield, which is commonly observed in many wood quality traits. Generally, variation in CBH was greater than in height but both mean and variance for both traits increased as the trees aged. Growth traits generally had low heritabilities ($h^2 < 0.2$) whereas wood quality traits showed moderate heritabilities (Table 1). Phenotypic correlations between growth traits were generally positive (0.24~0.74) whereas basic density was weakly negatively correlated with pulp yield (-0.28). The wood quality traits were generally independent from growth traits (correlations in the range -0.1 - 0.1) (Figure S1). The greatest positive phenotypic correlations were observed between CBH and height assessed at the same age (0.63 and 0.74 for 3 and 6 years, respectively).

78 Table 1. Statistical summary of phenotypes

Trait	Abbr.	No. obs.	Unit	Mean	CV(%)†	h^2
Circumference at breast height, age 3 years	CBH3	1123	cm	61.82	13.22	0.143
Height, age 3 years	Ht3	1094	m	22.43	9.81	0.162
Circumference at breast height, age 6 years	CBH6	1104	cm	83.80	18.67	0.186
Height, age 6 years	Ht6	985	m	28.40	13.09	0.182
Basic density	BD	1061	kg/m^3	532.78	6.83	0.381
Pulp yield	PY	1039	%	49.64	8.05	0.42

79 † CV: coefficient of variance.



80

81 **Figure 1.** Population structure of parents and F1 progenies. (A) Cross-validation
82 error in the admixture analysis for K varying from 1-20 for the 174 parents. (B)
83 Population structure of parents inferred using admixture for K=6. (C) PCA plot
84 based on genetic covariance among all individuals. Only the first two principle
85 components are shown. The colours used for the parents are in line with the
86 clustering shown in (B), with grey colour denoting all F1 progeny.

87 ***Population structure and model optimization***

88 To examine population structure in the breeding population including both
89 parents and their F1 progenies, we conducted both model-based admixture and
90 fastStructure analyses and principle component analysis (PCA) based on a set of
91 independent SNPs. The admixture analysis could not identify an optimal genetic
92 clustering on account of the minimization of the CV error even for K-values up to

93 K=100 (Figure S2). In contrast, fastStructure suggested an optimal genetic
94 clustering of K=1. Due to the inconsistences between the methods, we repeated the
95 population structure analyses using only the parents, given that the F1 individuals
96 were all obtained through crossings between these parents. Admixture analyses
97 based on the parents alone suggested K=6 minimized the CV error (Fig

98

99 ure 1A) and K=6 was also the optimal genetic clustering obtained from
100 fastStructure. The parents were assigned to the six subpopulations according to
101 individual ancestry proportions (Figure 1B). We also performed a PCA to
102 summarize genetic variation among parents and the first six components explained
103 21.53% of the total genetic variation. Notably, the eigenvalues beyond the first six
104 PCs were relatively small (Figure S3), consistent with the minimum K identified in
105 the admixture analyses. Based on first two principle components, parents can be
106 clearly separated into three clusters and two further sub-clusters can be identified
107 within in each major cluster. Progenies are inferred to be derived from crossing
108 parents either with the different major clusters or between them (Figure 1C) and
109 therefore we used the first six PCs in all subsequent analyses to correct for
110 population stratification.

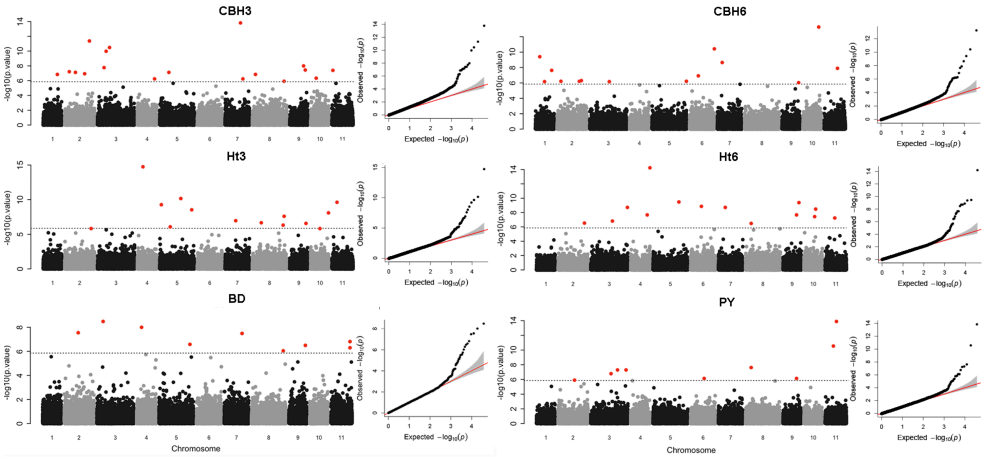


Figure 2. Manhattan plots and quantile-quantile (QQ) plots of the FarmCPU results using an additive effects model. The traits used are CBH and height at age 3 and 6 (Ht3 and Ht6, respectively) as well as basic density (BD) and pulp yield (PY). The Manhattan plots show -$\log_{10}$ *p*-values plotted against SNP positions on the 11 *Eucalyptus* chromosomes. Associations reaching genome-wide significance are displayed in red and the horizontal dotted line indicates a Bonferroni-corrected significant threshold of p <1.7E-06. The QQ plots for each of the six traits demonstrate the observed versus expected distribution of p-values. The solid red line represents the expected null distribution assuming no associations.

Genome-wide association study for additive effects

We first ran FarmCPU with an additive effect encoding to identify loci with significant additive effects on the different phenotypes. Quantile-quantile (QQ) plots suggest that population structure and kinship relationships were well controlled in the GWAS for the different traits (Figure 2). SNPs with *p*-values < 1.7E-06 threshold were declared statistically significant. Overall, we identified 78 significant SNPs across the six traits and these significant SNPs were distributed

across all 11 chromosomes (Figure 2). No significant SNPs were identified for more than one trait, even though both CBH and height show strong genetic correlations across ages. Comparing the number of significant SNPs found for the different traits, growth traits had more significant SNPs than wood traits, with height and CBH at the two different ages having between 14 and 18 significant SNPs whereas we only identified 9 significant SNPs for the two wood quality traits. We generally observe lower phenotypic variances explained by individual SNPs for CBH and height compared to pulp yield and basic density (Table S1). The maximum percentage of phenotypic variance explained by single associated SNP was 2.3% (for pulp yield) and the minimum percentage of phenotypic variance explained by a significantly associated SNP was 0.33% (CBH age 3 years).

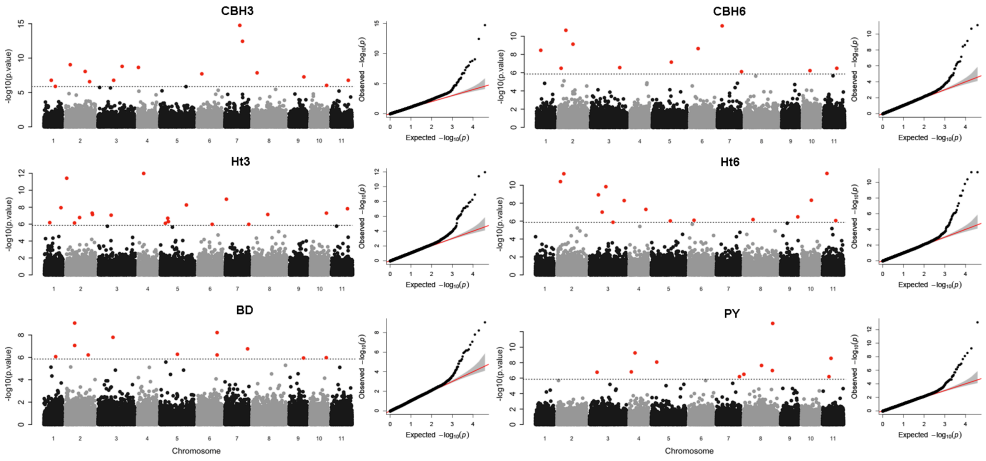


Figure 3. Manhattan plots and quantile-quantile (QQ) plots of the FarmCPU results for the dominance effects model. The traits used are CBH and height at age 3 and 6 (Ht3 and Ht6, respectively) as well as basic density (BD) and pulp yield (PY). The Manhattan plots show -$\log_{10}$ *p*-values plotted against SNP

positions on the 11 *Eucalyptus* chromosomes. Associations reaching genome-wide significance are displayed in red and the horizontal dotted line indicates a Bonferroni-corrected significant threshold of p <1.7E-06. The QQ plots for each of the six traits indicate the observed versus expected distribution of p-values. The solid red line represents the expected null distribution assuming no associations.

Genome-wide association study for dominance effects

FarmCPU efficiently controlled the false positive rates due to population structure and sample relationships also when identifying significant loci using dominance encoding (Figure 3, QQ plots). Under a dominance model we detected a total of 82 significant SNPs for the six traits. Height at 3 years old (Ht3) had the greatest number of associations with 19 SNPs displaying significant effects. Fewer associations were observed for the other traits, with between 11 and 15 significant SNPs identified (Figure 3, Manhattan plots). Two significant SNPs, Chr5.40663824 and Chr11.28479550, were found to overlap between CBH and height at age of 6 years. The maximum percentage of phenotypic variance explained by an associated SNP was 4%, a much higher value than found in the additive effect estimations. The smallest percentage of phenotypic variance explained by an associated SNP for the dominance model was of similar magnitude to that observed for additive effects model (Table S2). Comparing significant SNPs identified from the additive and dominance effects models, a total of 10 SNPs overlap between two models for different traits. This result suggest that the two genetic effects are not completely independent. Nine out of ten SNPs that overlap

between additive and dominance effects were identified for growth traits and with the remaining SNP observed for pulp yield.

Genome selection by using GWAS results

To confirm the utility of the SNPs identified from the GWAS and to further understand the performance of selecting SNPs for each trait based on the GWAS results, we conducted genomic prediction using four categories of SNPs by using both an additive genetic model (A) and an additive + dominance genetic model (AD). The four categories of SNPs used for the GBLUP analyses were selected from the GWAS results for each trait based on the following criteria: 1) '*associated SNPs*' were identified as significant from the GWAS using the threshold $p< 1.7E-6$; 2) '*putative SNPs*' were identified as significant from the GWAS using a more relaxed significance threshold $p<1E-3$ of each trait; 3) the '*top 1% SNPs*' included the top 1% SNPs for each trait, ranked according to GWAS significance. The rational of this category was to ensure that models utilised the same number of SNPs across all traits. Finally, 4) '*all SNPs*' used all 37,832 available SNPs when building the genomic selection models (Table S3).

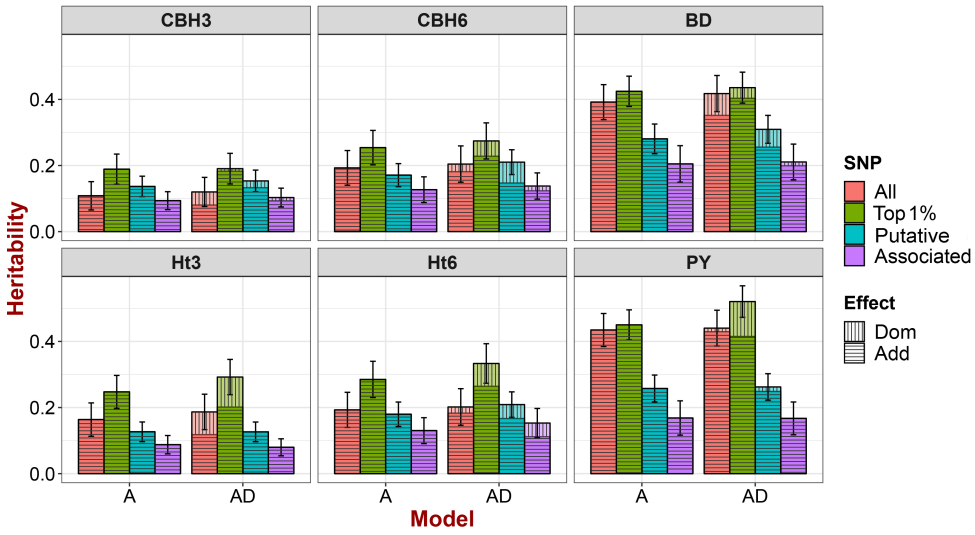


Figure 4. Genomic-based narrow- and broad-sense heritabilities based on an additive (A) or an additive + dominance (AD) model for the four different categories of SNPs used. The coloured bins represent the different categories of SNPs used, with red indicating 'all' SNPs (37,832), green indicates the 'top 1%' SNPs ranked according to GWAS p-value, cyan denotes 'putative' SNPs selected based on GWAS results with $p < 1E\text{-}3$ and purple denoted 'associated' SNPs selected based on GWAS results using $p < 1.7E\text{-}6$. The fill patterns represent different genetic effects. Vertical lines denote additive effects and horizontal lines denote dominance effects. Error bars indicate the standard error of total genetic variance.

The narrow-sense and broad-sense heritabilities were estimated using a modified GBLUP model based on different maker-based relationship matrices calculated using the four SNP categories. As expected, basic density and pulp yield had higher realised heritabilites than growth traits, independent of what category of SNPs that were used for the calculations. Broad-sense heritabilities were higher than narrow-sense heritabilities for most traits, demonstrating that dominance plays an important role in the expression of most traits (Figure 4) and in line with earlier

observations in this population (Tan et al., 2018). Comparing heritabilites (h^2 and H^2) for the different SNP categories suggest, perhaps surprisingly, that the 'top 1% SNPs' category explain more of the genetic variation than any of the other categories, including when all SNPs were used (Table S4). Furthermore, using the 'top 1% SNP' set yielded the largest estimates of dominance effects. As expected, using only SNPs that were significantly associated with a trait in the GWAS resulted in lower heritability estimates compared to using all SNPs. Comparing heritability estimates between the 'putative' and 'all' SNP categories showed that these yielded similar estimates for CBH and height, the 'putative' category of SNPs yielding significantly lower heritability estimates than the 'all' SNP category for the two wood quality traits (Figure 4).

We further estimated the prediction ability of breeding values for the *A* model and the prediction ability of genetic values for the *AD* model using a ten-fold cross-validation approach. The distribution of predictive abilities for each of the models and SNP, obtained using 100 replications, are displayed in Figure (5). Generally, we observe higher prediction abilities for wood quality traits, in line with the higher heritability values we observe for these traits. The AD model yielded slightly higher prediction abilities than the A model for most of traits. When comparing the different SNP categories, both the 'top 1%' and 'putative' SNP categories yielded substantial improvements in the predictive ability compared to both the 'all' and the 'significant' SNP categories. The 'associated' SNPs yielded more or less similar results to that obtained using 'all' SNPs. Moreover, the 'top

1%' SNP category yielded the highest average prediction ability for height, basic density and pulp yield with values being much higher than those calculated from the 'all' SNP category (Table S5).

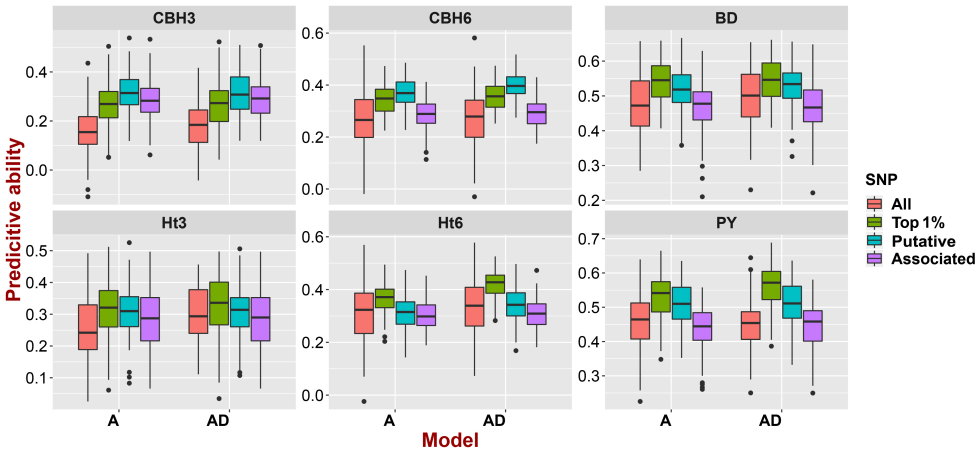


Figure 5. Predictive abilities of the additive (A) and the additive + dominance (AD) genomic prediction models for the four different categories of SNPs. The coloured boxplots display the distribution of predictive ability across 100 replicates of ten-fold cross-validation for the different categories of SNPs. Colours used as in Figure 4.

Discussion

We used GWAS with both additive and dominance effects to dissect the genetic architecture of growth and wood quality traits in hybrid *Eucalyptus*. The method we employed for GWAS, FarmCPU, was able to control the false positive rate induced by both the complex population structure and kinship that characterise our mapping population and efficiently identified significantly associated SNPs for both additive and dominance effects. Using top-ranking SNPs based on the GWAS

yielded higher genomic heritability estimates compared to using all available SNPs. We were also able to achieve more accurate genomic prediction results by filtering SNPs based on their associations in the GWAS and this help shed light on future directions in the application of genomic selection in *Eucalyptus* breeding.

FarmCPU perform superior in GWAS analysis

Most economic traits targeted for breeding in forestry, such as growth and wood properties, are quantitative traits and usually have a complex genetic architecture controlled by many loci of small effect. Here we utilised a recently developed method for the dissection of complex traits, FarmCPU, that has been proposed to efficiently address problems confounding between testing markers and covariates that often arise in GWAS (Liu et al., 2016). Several empirical studies have verified that FarmCPU offers enhanced power for GWAS of complex traits (Vanous et al., 2019; Ward et al., 2019; Zhu et al., 2018). In this study, we identified 78 and 82 significant associations having additive and dominant effects, respectively, with average of 13 SNPs identified per trait studied. These results are more efficient compared to another commonly used GWAS method, Genome-wide Efficient Mixed Model Association (GEMMA). In preliminary analyses we found that GEMMA was also able to control the false positive rate very well, but this came at the price of a relatively low statistical power. GEMMA consequently only identified two significant SNPs for additive effects (Figure S4) and a total of 13 significant SNPs for the dominance effects across traits (Figure S5). The FarmCPU methods therefore appears to be an attractive method that strike a good balance

between the identification of false positives and false negatives and thus have good power for the dissection complex traits.

Dominance effects play important roles in hybrid population

GWAS have traditionally assumed only additive effects of individual SNPs (Bush and Moore, 2012; Marjoram et al., 2014) but here we show the added value of also considering dominance effects for identifying genomic regions controlling growth and wood quality traits. By assessing also dominance effects, we identify an additional 72 associated SNPs across the traits, in addition to the 78 SNPs we identified using additive effects. Furthermore, a considerable proportion of the genetic variation in our hybrid population is attributable to non-additive effects and our results show that the alleles underlying this variation can be identified when dominance effects are explicitly considered in a GWAS setting. Several previous studies have used controlled crosses in crop species, particularly in maize and rice, to identified loci that exhibit dominance effects. For heterosis-related traits, data from maize is frequently cited as supporting the dominance model (Cui et al., 2017; Wallace et al., 2014), while rice has been proposed as a system that supports the over-dominance hypothesis (Li et al., 2016; Zhen et al., 2017). Our approach quantifies the contribution of dominance to the "missing heritability" in a *Eucalyptus* hybrid breeding population and we collectively show that up to 10% of the genomic-based heritability can be explained by associated SNPs that were identified using a dominance model (Table S4).

The benefit of integrating GWAS results on genomic prediction

385 Even if we capture a substantially larger number of associated SNPs by
386 considering both additive and dominance effects, a large fraction of the genomic-
387 based hertiabilies (14%-62%) cannot be explained by only considering
388 significantly associated SNPs (Table S4) and these observations are in line with
389 several earlier reports (Chhetri et al., 2019b; Tang et al., 2019; Zhao et al., 2019).
390 Also, when using significantly associated SNPs from the GWAS to investigate the
391 accuracy of genomic predication, we find that this yields no improvement in
392 accuracy, and sometimes even reduced accuracy, compared to predictions based on
393 all available SNPs which mirrors results seen in other similar studies (Gowda et al.,
394 2015; Wallace et al., 2016). Regions identified in a GWAS are consequently not
395 able to explain all of the genetic variation in the traits of interest and this problem
396 is greater for quantitative traits that are controlled by many genes of small effect.
397 These are the traits where current GWAS methods often suffers from insufficient
398 power to detect loci of small effect, unless sample sizes are substantially larger
399 than what is commonly used in most studies of plants.

400 In order to assess if the GWAS results could be used to enhance genomic
401 prediction in our breeding population, we also tried to identify possible 'candidate'
402 SNPs that were not detected as significant using the stringent significant threshold
403 we applied in our GWAS. The rational here is that, as outlined above, most GWAS
404 methods fail to detect loci of small effect but that the GWAS would nevertheless
405 serve as a useful 'filter' for ranking SNPs for their possible effects on the traits of
406 interest. We therefore selected two categories of SNPs using two different criteria

of relaxed significance and used these to estimate genomic heritabilities and perform genomic prediction. The first category, 'putative' SNPs include all SNPs that were found to be associated with the traits of interest based a more relaxed *p*-value (p< 1E-3). Using this more relaxed *p*-value we identify between 70 to184 SNPs for the different of trait when combined across the additive and dominance effect models. Using the 'putative' SNP category we observed large improvements in the heritability estimates for the growth traits, to the point where almost all of the genetic variation could be explained (Table S4). For wood quality traits, however, about 40% of the genetic variation remain unexplained compared to when using all SNPs for heritability estimation (Table S4). The second category of SNPs we considered consisted of the top 1% of SNPs, ranked by the *p*-value from the GWAS. Using this criterion ensures that the same number of SNPs are used for prediction across the different traits. Surprisingly we were able to explain a substantially greater proportion, up to 174%, of the genetic variation explained when using all SNPs (Table S4). When we performed genomic prediction using these two categories of SNPs we also observe a substantial increase in the prediction ability for all traits compared to predictions based on all available SNPs. This suggests that using all available SNPs introduce noise in the prediction models that negatively affects our prediction ability. Our method for analysing genomic selection and increasing prediction accuracy clearly benefited from integrating results from the GWAS analyses, but the number of associated SNPs that needs to be incorporated depends on the study trait in questions.

Detection of associations for complex traits in forest trees

Identifying candidate genes underlying growth and wood traits has long been an active area of research in forest trees, such as in *Eucalyptus* (Cappa et al., 2013; Müller et al., 2019; Muller et al., 2017; Resende et al., 2017a), *Populus* (Allwright et al., 2016; Du et al., 2016; Fahrenkrog et al., 2017; Porth et al., 2013) and *Pinus* (Bartholomé et al., 2016; Lu et al., 2017). To ensure good statistical power, both common and rare genetic variants needs to be considered to have a comprehensive understanding of the genetic regulation of complex traits, since many low-frequency variants were identified as associated with growth and wood composition traits (Fahrenkrog et al., 2017). For instance, regional heritability mapping (RHM), has previously been shown to successfully utilise information from both common and rare variants and can therefore capture a larger proportion of the genomic heritability in *Eucalyptus* (Müller et al., 2019; Resende et al., 2017a).

Furthermore, both additive and non-additive effects play important roles in association studies for many traits. Adding dominance effects to a GWAS analysis increase the possibility to identify additional variants that can help capture a greater fraction of the genetic variance (Du et al., 2016; Lu et al., 2017). Other methods, such increasing the sample size using meta-analysis (Müller et al., 2019) or using multi-locus GWAS approaches instead of single marker methods (Fahrenkrog et al., 2017) are methods that also can help increase statistical power in GWAS.

Putative genes for plant growth and stress response

Among the significantly associated SNPs we observe across additive and dominance effects estimations, we identified a total of 49 candidate genes that have known functions relevant for the traits in question. The details of these genes, including information on the position of associated SNPs and the putative functions of the genes, are summarised Table S6. In general, candidate genes can be separated into two groups, with one group containing genes that have direct functions associated to the morphological formation of different tissues or organs. The other group contain genes related to general responses to abiotic and biotic stress, which, more indirectly, influence plant growth and biomass.

Among the significant SNPs associated with morphology, a number of associations are linked to genes which are related to cell wall biosynthesis. For example, SNP Chr3.46653967 is associated with Ht6 using an additive effects model. This SNP is located on chromosome 3 and encodes a missense variant in a gene coding for a pectin lyase-like superfamily protein (PME). This gene is expressed in stamen and is involved in cell wall loosening and have previously been implicated in floral development (Francis et al., 2006). We also identified a significant SNP on chromosome 11 (Chr11.20479646) which is associated with Ht6 (Table S6). This SNP is located in the gene *Eucgr.K01691* which encodes a homolog to the Arabidopsis alpha-L-arabinofuranosidase 1 (*ARAF*1) gene. Expression of the *ARAF*1 gene is localized to several cell types in the vascular system of roots and stems and the protein is known to be involved in cell type-specific alterations of cell wall structure (Chávez Montes et al., 2008). Many other

studies have also identified cell wall biosynthesis related genes from GWAS performed using growth traits in forest trees. Du *et. al.* (2016) identified four significant SNPs that were located in genes involved in secondary cell wall biosynthesis when analysing growth traits in the *Populus* (Du et al., 2016). A SNP associated with volume in *E. pellita* is located in a gene whose function is known to be involved in cell wall cellulose biosynthesiss (Muller et al., 2017). Similarly, Muller *et. al.* (Müller et al., 2019) used a joint-GWAS approach in four *Eucalyptus* breeding populations and identified eight SNPs associated with growth traits that were all linked to genes which were related to cell wall biosynthesis.

Many of the candidate genes putatively related to abiotic and biotic stress show response to adverse conditions. It is perhaps not surprising that these genes show up in our GWAS, as the planting area of the study population alternates between extremely dry (from July to August) and wet (from August to October) conditions in most years, which often leads to stress-induced damage and high incidence of diseases. In line with this, we identified several candidate genes involved in stress response to adverse climate conditions. The SNP Chr2.1760161 is highly associated (*p*-value=3.72E-12) with height at 3 years age in the dominance model. This SNP is located upstream of the gene *Eucgr.B00092*, which encodes a putative *HVA*22 homologue E (*HVA*22*E*). *HVA*22*E* is upregulated to varying degrees in response to cold and salt stress, ABA treatment or dehydration (Chen et al., 2002; Shen et al., 2001). Another variant (Chr3.41941452), associated with CBH at age 6, is located in the upstream region of *High-affinity K+*

transporter 1 (HKT1) gene. *HKT*1 is expressed in root stelar cells and leaf cells (Hamamoto et al., 2015) and provides a key mechanism for protecting leaves from Na+ over-accumulation and salt stress (Berthomieu et al., 2003; Maser et al., 2002). The SNP Chr6.23066996 is associated with CBH at 6 years of age in the additive effects model, is located on chromosome 6 inside the gene *Eucgr.F01775* that encodes catalase 2 (*CAT*2). *CAT*2 controls levels and sensitivity to H_2O_2 (Bueso et al., 2007), photo-oxidative stress (Konert et al., 2015) and auxin levels (Gao et al., 2014).

Four of the candidate genes we identify in our GWAS have functions in both morphological formation and stress response. One common SNP (Chr11. 28479550), associated with CBH6 in both the additive and dominance models as well as with Ht6 for the dominance model is located in the vicinity of the gene *Eucgr.*K02133 which encodes a nucleotide-diphospho-sugar transferase (*QUA*1). This enzyme is expressed in vascular tissues and affects homogalacturonan, pectin and hemicellulose cell wall synthesis (Orfila et al., 2005). Recent studies have shown that *QUA*1 also functions in chloroplast-dependent calcium signalling under salt and drought stresses (Zheng et al., 2016). Finally, the SNP Chr4.11644680 is associated with Ht3 and is a synonymous variant located in the *SFR*6/*MED*16 gene which plays important roles in cold- and drought-inducible gene expression (Knight et al., 2009, defence gene expression {Wathugala, 2012 #739) as well as modulating iron uptake (Zhang et al., 2014) in response to cell wall defects (Sorek

et al., 2015). These findings suggest that stress resistance also plays an important role in affecting tree growth traits.

Conclusions

In this study, we have used a GWAS approach in a *Eucalyptus* hybrid population to dissect the genetic basis of growth and wood quality traits by accounting for both additive and dominance genetic effects. Altogether we identify 78 and 82 significant SNPs using additive and dominance models, respectively, with 10 SNPs showing an overlap between the two effect models, suggesting that additive and dominance effects are not completely independent. The associated genes could be grouped into two broad functional categories relating to how they influence tree growth and biomass. One group contain genes associated with morphological formation, such as cell wall biosynthesis, and the other group contain genes related to abiotic and biotic stress responses, such as oxidative, hormone-based and disease-induced stress. These results provide novel targets for possible transgenic or genome editing approaches in the future to directly improve growth and biomass related traits.

We also applied our results from the GWAS in a genomic selection analysis by using different categories of SNPs selected based on the GWAS results and used them to evaluate genomic-based heritabilities and predictive abilities. Our results show that prediction abilities of the estimated breeding values improved for all traits when using SNPs selected based on the GWAS results. Integrating GWAS

results into genomic selection thus appear to be a promising avenue to increase the efficiency of genomic selection in forest breeding.

Experimental procedures

Populations, phenotypic and genotypic data

A total of 1123 *Eucalyptus* individuals were used in this study, including 90 *E.grandis*, 84 *E.urophylla* parents and 949 F1 progenies derived from a random mating design that has previously been described (Tan et al., 2017). Briefly, F1 individuals were identified to be comprised of inter- and intra-crossing of the two parental species. Of the 949 F1 individuals, 57% were interspecific *E.grandis* × *E.urophylla* hybrids, 21% were intraspecific *E.grandis* × *E.grandis* progeny and 22% were intraspecific *E.urophylla* × *E.urophylla* progeny (Tan et al., 2018).

The phenotypic and genotypic data utilized in this study has been previously described in detail (Tan et al., 2017). The phenotypes include height and circumference at breast height (CBH), where F1 individuals were evaluated at ages three and six and the pure species parents were evaluated at age five. In addition, we obtained data on two wood quality traits, basic density and pulp yield, that were assessed at age five. Genotyping was performed using an Illumina Infinium EuCHIP60K SNP chip that contains probes for 60,904 unique SNPs (Silva-Junior et al., 2015). Across the 1123 individuals, 37,832 SNPs were retained after quality-control based on call rates (>0.7) for both SNPs and samples and following filtering based on minor allele frequencies (>0.01) and deviations from Hardy-

Weinberg equilibrium (>1e-7). Any missing data remaining in the 37,832 SNPs were subsequently imputed using BEAGLE 4.1 (Browning and Browning, 2007).

Phenotypic data analyses

Phenotype data for the parental and F1 population were adjusted separately to minimize environmental variation by fitting a mixed linear model for each trait:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Z_b r} + \boldsymbol{\varepsilon} \quad (1)$$

where $\mathbf{y}$ is the vector of phenotypic observation, $\boldsymbol{\beta}$ is the vector of overall mean as fixed effect, $\mathbf{r}$ is the vector of random replication effects following $r \sim N(0, I\sigma_r^2)$, where σ_r^2 is the replication variance, $\boldsymbol{\varepsilon}$ is the vector of random residual effects. $\mathbf{X}$ and $\mathbf{Z_b}$ is design matrix for $\boldsymbol{\beta}$ and $\mathbf{r}$, respectively. For the F1 population, the residual variance-covariance matrix is $R = I\sigma_e^2 + AR1(\rho_r) \otimes AR1(\rho_c)\sigma_\eta^2$, where $AR1(\rho_r)$ and $AR1(\rho_c)$ are autoregressive correlation matrices for the row model (autocorrelation parameter ρ_r) and column model (autocorrelation parameter ρ_c), respectively. σ_e^2 is the independent residual variance while σ_η^2 is the spatial variance. For the pure parental species, we fitted the model in Equation 1 by setting the residual matrix $R = I\sigma_\varepsilon^2$ since spatial coordinates and position information were not available for these individuals. All mixed-linear model analyses were performed in ASReml 4 (Gilmour et al., 2015). Phenotypes of individuals from the F1 and parental populations were adjusted for random block effects (r) and spatial effects (s), respectively.

The heritability (h^2) was estimated using a mixed model $\mathbf{y} = \mathbf{X\beta} + \mathbf{Zu} + \boldsymbol{\varepsilon}$, where $\mathbf{y}$ represents the adjusted phenotypes of single trait, $\boldsymbol{\beta}$ is the vector of fixed effects, including overall mean and age difference. $\mathbf{u}$ is a vector of random additive or dominance genetic effect of individuals with a normal distribution, $\mathbf{u}\sim N(0, \mathbf{A}\sigma_a^2)$, $\mathbf{A}$ being the revised, pedigree-based genetic relationships among individuals (Tan et al., 2018); and $\boldsymbol{\varepsilon}$ is a heterogeneous random residual effects represented different experimental sites. $\mathbf{X}$ and $\mathbf{Z}$ are incidence matrices for $\boldsymbol{\beta}$ and $\mathbf{u}$, respectively. We obtained restricted maximum likelihood (REML) estimates of σ_a^2 and $\boldsymbol{\varepsilon}$, and estimated $h^2 = \sigma_a^2/(\sigma_a^2 + \sum \varepsilon/n)$.

Population structure, kinship analysis and GWAS

In the association analyses, confounding effects of population structure and kinship between individuals need to be accounted for. Population structure (Q) was estimated using a model-based clustering and through principle component analysis (PCA) using 13,245 independent SNPs obtained by LD-pruning the original SNP data set and including only SNPs that have pairwise linkage disequilibrium (LD) values (r^2) less than 0.2. Model-based clustering was implemented using the software *admixture* v.1.3.0 which infer population structure by estimating individual admixture proportions using multi-locus SNP data through a maximum-likelihood method (Alexander et al., 2009). The number of ancestral populations (P) was varied from 1 to 100 when using admixture and *fastStructure* (Raj et al., 2014). Five-fold cross-validation (CV) was performed to choose the optimal P value in

admixture. A PCA was also performed using the *smartpca* program in *eigensoft* v6.0 to estimate individual ancestry proportions (Price et al., 2006).

GWAS was conducted using a recently developed method, FarmCPU, which explicitly takes into account the confounding that exists between covariates and test marker by using both a fixed effect model and a random effect model (Liu et al., 2016). The results from the PCA and the kinship matrix were used as covariates in FarmCPU to account for population structure and relatedness among samples, respectively. We ran the GWAS using the R package *FarmCPU*. False positive errors due to multiple testing were controlled by an adjusted Bonferroni method, *simpleM* (Gao et al., 2008). This method infers the number of independent SNPs by filtering on LD and performs a standard Bonferroni correction to correct for multiple testing based on the number of 'independent tests' performed. For the present data a p-value < 1.7E-06 was selected as a cut-off to indicate significant associations.

GWAS for additive and dominance models

We conducted GWA analyses in FarmCPU using either an additive or dominance encoding of genotypes. For the additive encoded data, the homozygous major allele was encoded with 0, the heterozygous genotype with 1 and the homozygous minor allele with 2. For the dominance encoding, both homozygous minor and major alleles were encoded as 0 whereas the heterozygous genotype was encoded as 1 (Seymour et al., 2016).

Genomic selection (GS) with different informative SNPs

Genomic selection (GS) models were constructed based on four different categories of SNPs using the Genomic Best Linear Unbiased Prediction (GBLUP) method. Details on how the genomic-based additive and dominance relationship matrices are estimated have been previously described in detail in Tan *et al.* (Tan et al., 2018). Here we focus on the details of how the four categories of SNPs we have used in all subsequent analyses were selected. The four categories of SNPs employed for estimating the additive and dominance relationship matrices were: 1) '*associated SNPs*' which contain only SNPs that were identified as significantly associated with the corresponding trait at the Bonferroni-adjusted *p*-value <1.7E-6; 2) '*putative SNPs*' are all SNPs that were significant in the GWAS for the corresponding trait using a more relaxed *p*-value threshold ($p<$ 1E-3) in order to capture also possible causal SNPs that do not reach significance using the more stringent criteria in the original GWAS; 3) The '*top 1% SNPs*' category use the top 378 SNPs for each trait in the GWAS ranked after *p*-value in the GWAS in order to evaluate the same number of SNPs across different traits when building genomic selection models; and finally 4) '*all SNPs*' which use all of the 37,832 SNPs available and is therefore identical to the models originally used in Tan et al (2017) and Tan el al (2018).

Two separate GBLUP models were evaluated that included i) either only additive (*A*) or ii) both additive and dominance (*AD*) genetic effects using the four SNP categories described above to create marker-based relationship matrices. The *A* and *AD* models have been well described earlier in Tan *et. al.*(Tan et al., 2018).

The genomic-based narrow- and broad-sense heritability (h^2 and H^2 respectively) were calculated after fitting each model across the different traits. Narrow-sense heritability in the A model was estimated as $h^2 = \sigma_a^2/\sigma_p^2$ and the broad-sense heritability of AD model was estimated as $h^2 = (\sigma_a^2 + \sigma_d^2)/\sigma_p^2$, where σ_a^2, σ_d^2 and σ_p^2 represented the estimated additive, dominance and phenotypic variance, respectively. The prediction ability was estimated for all models and relationship matrices using a ten-fold cross-validation scheme where 100 replications was implemented to evaluate the prediction accuracy of the different models. For each replication, the dataset was randomly divided into 10 subsets and nine out of the ten partitions were used as the training population to fit a model using both phenotypes and genotypes, while the remaining partition was used as the validation set where phenotypic data was removed and then used to predict breeding values or total genetic values for the model in question. The predictive ability of the model was evaluated by estimating the correlation between phenotypes and breeding/genetic values, $r(\hat{A}_{vali}, Y_{vali})$ or $r(\hat{G}_{vali}, Y_{vali})$.

Assigning significant SNPs to putative candidate genes

Genes within ±5kb away from a SNP that was significantly associated with a measured phenotypic trait were extracted from the *E.grandis* v2.0 reference genome (BRASUZ1) in Phytozome (www.phytozome.net) using SnpEff v4.2 (Cingolani et al., 2012). The 5kb window threshold used was based on the distance over which LD decays in this population (Tan et al., 2017). The putative functions

of these candidate genes were determined based on their homology to functionally characterized genes in *A. thaliana* (TAIR).

Acknowledgements

The study has partly been funded through grants from Vetenskapsrådet and Kempestiftelserna to PKI. BT gratefully acknowledges financial support from the Umeå Plant Science Centre (UPSC) "The Research School of Forest Genetics, Biotechnology and Breeding". The research was conducted using resources from the High Performance Computing Centre North, Umeå (HPC2N). The authors are grateful to Yousry El-Kassaby for comment on an earlier version of the manuscript.

References

Alexander, D.H., Novembre, J. and Lange, K. (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19**, 1655-1664.

Allwright, M.R., Payne, A., Emiliani, G., Milner, S., Viger, M., Rouse, F., Keurentjes, J.J.B., Bérard, A., Wildhagen, H., Faivre-Rampant, P., Polle, A., Morgante, M. and Taylor, G. (2016) Biomass traits and candidate genes for bioenergy revealed through association genetics in coppiced European Populus nigra (L.). *Biotechnology for Biofuels* **9**, 195.

Bartholomé, J., Bink, M.C.A.M., van Heerwaarden, J., Chancerel, E., Boury, C., Lesur, I., Isik, F., Bouffier, L. and 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**, e0165323.

Berthomieu, P., Conejero, G., Nublat, A., Brackenbury, W.J., Lambert, C., Savio, C., Uozumi, N., Oiki, S., Yamada, K., Cellier, F., Gosti, F., Simonneau, T., Essah, P.A., Tester, M., Very, A.A., Sentenac, H. and Casse, F. (2003)

Functional analysis of AtHKT1 in *Arabidopsis* shows that Na+ recirculation by the phloem is crucial for salt tolerance. *Embo J* **22**, 2004-2014.

Bison, O., Ramalho, M., Rezende, G., Aguiar, A. and De Resende, M. (2006) Comparison between open pollinated progenies and hybrids performance in *Eucalyptus grandis* and *Eucalyptus urophylla*. *Silvae Genet* **55**, 192-196.

Bouvet, J.M., Saya, A. and Vigneron, P. (2009) Trends in additive, dominance and environmental effects with age for growth traits in *Eucalyptus* hybrid populations. *Euphytica* **165**, 35-54.

Bouvet, J.M. and Vigneron, P. (1995) Age trends in variances and heritabilities in *Eucalyptus* factorial mating designs. *Silvae Genet* **44**, 206-216.

Browning, S.R. and Browning, B.L. (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet* **81**, 1084-1097.

Bruce, A.B. (1910) The mendelian theory of heredity and the agumentation of vigor. *Science* **32**, 627.

Bueso, E., Alejandro, S., Carbonell, P., Perez-Amador, M.A., Fayos, J., Belles, J.M., Rodriguez, P.L. and Serrano, R. (2007) The lithium tolerance of the Arabidopsis cat2 mutant reveals a cross-talk between oxidative stress and ethylene. *Plant J* **52**, 1052-1065.

Bush, W.S. and Moore, J.H. (2012) Chapter 11: Genome-wide association studies. *PLoS Comput Biol* **8**, e1002822.

Cappa, E.P., El-Kassaby, Y.A., Garcia, M.N., Acuna, C., Borralho, N.M., Grattapaglia, D. and Marcucci Poltri, S.N. (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**, e81267.

Charlesworth, D. and Willis, J.H. (2009) The genetics of inbreeding depression. *Nat Rev Genet* **10**, 783.

'16 Chávez Montes, R.A., Ranocha, P., Martinez, Y., Minic, Z., Jouanin, L., Marquis,
'17 M., Saulnier, L., Fulton, L.M., Cobbett, C.S., Bitton, F., Renou, J.-P., Jauneau,
'18 A. and Goffner, D. (2008) Cell wall modifications in Arabidopsis plants with
'19 altered alpha-L-arabinofuranosidase activity. *Plant Physiol* **147**, 63-77.

'20 Chen, C.N., Chu, C.C., Zentella, R., Pan, S.M. and Ho, T.H.D. (2002) AtHVA22
'21 gene family in *Arabidopsis*: phylogenetic relationship, ABA and stress
'22 regulation, and tissue-specific expression. *Plant Mol Biol* **49**, 633-644.

'23 Chhetri, H.B., Macaya-Sanz, D., Kainer, D., Biswal, A.K., Evans, L.M., Chen, J.-
'24 G., Collins, C., Hunt, K., Mohanty, S.S., Rosenstiel, T., Ryno, D., Winkeler,
'25 K., Yang, X., Jacobson, D., Mohnen, D., Muchero, W., Strauss, S.H.,
'26 Tschaplinski, T.J., Tuskan, G.A. and DiFazio, S.P. (2019a) Multitrait genome-
'27 wide association analysis of Populus trichocarpa identifies key polymorphisms
'28 controlling morphological and physiological traits. *New Phytol* **223**, 293-309.

'29 Chhetri, H.B., Macaya-Sanz, D., Kainer, D., Biswal, A.K., Evans, L.M., Chen, J.-
'30 G., Collins, C., Hunt, K., Mohanty, S.S., Rosenstiel, T., Ryno, D., Winkeler,
'31 K., Yang, X., Jacobson, D., Mohnen, D., Muchero, W., Strauss, S.H.,
'32 Tschaplinski, T.J., Tuskan, G.A. and DiFazio, S.P. (2019b) Multitrait genome-
'33 wide association analysis of Populus trichocarpa identifies key polymorphisms
'34 controlling morphological and physiological traits. *New Phytol* **0**.

'35 Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J.,
'36 Lu, X.Y. and Ruden, D.M. (2012) A program for annotating and predicting the
'37 effects of single nucleotide polymorphisms, SnpEff. *Fly* **6**, 80-92.

'38 Crow, J.F. (1948) Alternative hypotheses of hybrid vigor. *Genetics* **33**, 477-487.

'39 Cui, M., Jia, B., Liu, H.H., Kan, X., Zhang, Y., Zhou, R.H., Li, Z.P., Yang, L.,
'40 Deng, D.X. and Yin, Z.T. (2017) Genetic mapping of the leaf number above
'41 the primary rar and its relationship with plant height and flowering time in
'42 maize. *Front Plant Sci* **8**.

'43 Du, Q., Gong, C., Wang, Q., Zhou, D., Yang, H., Pan, W., Li, B. and Zhang, D.
'44 (2016) Genetic architecture of growth traits in Populus revealed by integrated

quantitative trait locus (QTL) analysis and association studies. *New Phytol* **209**, 1067-1082.

Duvick, D.N. (2001) Biotechnology in the 1930s: the development of hybrid maize. *Nat Rev Genet* **2**, 69-74.

Fahrenkrog, A.M., Neves, L.G., Resende Jr, M.F.R., Vazquez, A.I., de los Campos, G., Dervinis, C., Sykes, R., Davis, M., Davenport, R., Barbazuk, W.B. and Kirst, M. (2017) Genome-wide association study reveals putative regulators of bioenergy traits in Populus deltoides. *New Phytol* **213**, 799-811.

Francis, K.E., Lam, S.Y. and Copenhaver, G.P. (2006) Separation of Arabidopsis Pollen Tetrads Is Regulated by QUARTET1, a Pectin Methylesterase Gene. *Plant Physiol* **142**, 1004.

Gao, X., Starmer, J. and Martin, E.R. (2008) A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genet Epidemiol* **32**, 361-369.

Gao, X., Yuan, H.-M., Hu, Y.-Q., Li, J. and Lu, Y.-T. (2014) Mutation of Arabidopsis CATALASE2 results in hyponastic leaves by changes of auxin levels. *Plant, Cell & Environment* **37**, 175-188.

Gilmour, A.R., Gogel, B.J., Cullis, B.R., Welham, S.J. and Thompson, R. (2015) ASReml User Guide Release 4.1 Structural Specification. *VSN International Ltd, Hemel Hempstead, HP1 1ES, UK* *www.vsni.co.uk*.

Gowda, M., Das, B., Makumbi, D., Babu, R., Semagn, K., Mahuku, G., Olsen, M.S., Bright, J.M., Beyene, Y. and Prasanna, B.M. (2015) Genome-wide association and genomic prediction of resistance to maize lethal necrosis disease in tropical maize germplasm. *Theor Appl Genet* **128**, 1957-1968.

Grattapaglia, D. (2017) Status and perspectives of genomic selection in forest tree breeding. In: *Genomic Selection for Crop Improvement* (Varshney, R.K., Roorkiwal, M. and Sorrells, M.E. eds), pp. 199-250. Brazil: Springer International Publishing.

'73 Hallingback, H.R., Fogelqvist, J., Powers, S.J., Turrion-Gomez, J., Rossiter, R.,
'74 Amey, J., Martin, T., Weih, M., Gyllenstrand, N., Karp, A., Lagercrantz, U.,
'75 Hanley, S.J., Berlin, S. and Ronnberg-Wastljung, A.C. (2016) Association
'76 mapping in *Salix viminalis* L. (Salicaceae) - identification of candidate genes
'77 associated with growth and phenology. *Gcb Bioenergy* **8**, 670-685.

'78 Hamamoto, S., Horie, T., Hauser, F., Deinlein, U., Schroeder, J.I. and Uozumi, N.
'79 (2015) HKT transporters mediate salt stress resistance in plants: from structure
'80 and function to the field. *Curr Opin Biotech* **32**, 113-120.

'81 Hill, W.G. (1982) Dominance and rpistasis as components of heterosis. *J Anim*
'82 *Breed Genet* **99**, 161-168.

'83 Ingvarsson, P.K., Garcia, M.V., Luquez, V., Hall, D. and Jansson, S. (2008)
'84 Nucleotide Polymorphism and Phenotypic Associations Within and Around
'85 the *phytochrome B2* Locus in European Aspen (*Populus tremula*, Salicaceae).
'86 *Genetics* **178**, 2217.

'87 Ingvarsson, P.K. and Street, N.R. (2011) Association genetics of complex traits in
'88 plants. *New Phytol* **189**, 909-922.

'89 Knight, H., Mugford, S.G., Ulker, B., Gao, D.H., Thorlby, G. and Knight, M.R.
'90 (2009) Identification of SFR6, a key component in cold acclimation acting
'91 post-translationally on CBF function. *Plant J* **58**, 97-108.

'92 Konert, G., Rahikainen, M., Trotta, A., Durian, G., Salojärvi, J., Khorobrykh, S.,
'93 Tyystjärvi, E. and Kangasjärvi, S. (2015) Subunits B′γ and B′ζ of protein
'94 phosphatase 2A regulate photo-oxidative stress responses and growth in
'95 Arabidopsis thaliana. *Plant, Cell & Environment* **38**, 2641-2651.

'96 Kruglyak, L. (2008) The road to genome-wide association studies. *Nat Rev Genet* **9**,
'97 314-318.

'98 Li, D.Y., Huang, Z.Y., Song, S.H., Xin, Y.Y., Mao, D.H., Lv, Q.M., Zhou, M.,
'99 Tian, D.M., Tang, M.F., Wu, Q., Liu, X., Chen, T.T., Song, X.W., Fu, X.Q.,
300 Zhao, B.R., Liang, C.Z., Li, A.H., Liu, G.Z., Li, S.G., Hu, S.N., Cao, X.F., Yu,

J., Yuan, L.P., Chen, C.Y. and Zhu, L.H. (2016) Integrated analysis of phenome, genome, and transcriptome of hybrid rice uncovered multiple heterosis-related loci for yield increase. *P Natl Acad Sci USA* **113**, E6026-E6035.

Liu, X., Huang, M., Fan, B., Buckler, E.S. and Zhang, Z. (2016) Iterative Usage of Fixed and Random Effect Models for Powerful and Efficient Genome-Wide Association Studies. *PloS Genet* **12**, e1005767.

Lu, M., Krutovsky, K.V., Nelson, C.D., West, J.B., Reilly, N.A. and Loopstra, C.A. (2017) Association genetics of growth and adaptive traits in loblolly pine (Pinus taeda L.) using whole-exome-discovered polymorphisms. *Tree Genet Genomes* **13**, 57.

Makouanzi, G., Bouvet, J.M., Denis, M., Saya, A., Mankessi, F. and Vigneron, P. (2014) Assessing the additive and dominance genetic effects of vegetative propagation ability in Eucalyptus-influence of modeling on genetic gain. *Tree Genet Genomes* **10**, 1243-1256.

Makowsky, R., Pajewski, N.M., Klimentidis, Y.C., Vazquez, A.I., Duarte, C.W., Allison, D.B. and de los Campos, G. (2011) Beyond Missing Heritability: Prediction of Complex Traits. *PloS Genet* **7**, e1002051.

Marjoram, P., Zubair, A. and Nuzhdin, S.V. (2014) Post-GWAS: where next? More samples, more SNPs or more biology? *Heredity* **112**, 79-88.

Maser, P., Eckelman, B., Vaidyanathan, R., Horie, T., Fairbairn, D.J., Kubo, M., Yamagami, M., Yamaguchi, K., Nishimura, M., Uozumi, N., Robertson, W., Sussman, M.R. and Schroeder, J.I. (2002) Altered shoot/root Na+ distribution and bifurcating salt sensitivity in Arabidopsis by genetic disruption of the Na+ transporter AtHKTI1. *Febs Lett* **531**, 157-161.

McKown, A.D., Klapste, J., Guy, R.D., Geraldes, A., Porth, I., Hannemann, J., Friedmann, M., Muchero, W., Tuskan, G.A., Ehlting, J., Cronk, Q.C., El-Kassaby, Y.A., Mansfield, S.D. and Douglas, C.J. (2014) Genome-wide

association implicates numerous genes underlying ecological trait variation in natural populations of *Populus trichocarpa*. *New Phytol* **203**, 535-553.

Müller, B.S.F., de Almeida Filho, J.E., Lima, B.M., Garcia, C.C., Missiaggia, A., Aguiar, A.M., Takahashi, E., Kirst, M., Gezan, S.A., Silva-Junior, O.B., Neves, L.G. and Grattapaglia, D. (2019) Independent and Joint-GWAS for growth traits in Eucalyptus by assembling genome-wide data for 3373 individuals across four breeding populations. *New Phytol* **221**, 818-833.

Muller, B.S.F., Neves, L.G., de Almeida, J.E., Resende, M.F.R., Munoz, P.R., dos Santos, P.E.T., Paludzyszyn, E., Kirst, M. and 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**.

Neale, D.B. and Kremer, A. (2011) Forest tree genomics: growing resources and applications. *Nat Rev Genet* **12**, 111-122.

Olson, M.S., Levsen, N., Soolanayakanahally, R.Y., Guy, R.D., Schroeder, W.R., Keller, S.R. and Tiffin, P. (2013) The adaptive potential of *Populus balsamifera* L. to phenology requirements in a warmer global climate. *Mol Ecol* **22**, 1214-1230.

Orfila, C., Sørensen, S.O., Harholt, J., Geshi, N., Crombie, H., Truong, H.-N., Reid, J.S.G., Knox, J.P. and Scheller, H.V. (2005) QUASIMODO1 is expressed in vascular tissue of Arabidopsis thaliana inflorescence stems, and affects homogalacturonan and xylan biosynthesis. *Planta* **222**, 613-622.

Porth, I., Klapste, J., Skyba, O., Hannemann, J., McKown, A.D., Guy, R.D., DiFazio, S.P., Muchero, W., Ranjan, P., Tuskan, G.A., Friedmann, M.C., Ehlting, J., Cronk, Q.C.B., El-Kassaby, Y.A., Douglas, C.J. and Mansfield, S.D. (2013) Genome-wide association mapping for wood characteristics in *Populus* identifies an array of candidate single nucleotide polymorphisms. *New Phytol* **200**, 710-726.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A. and Reich, D. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904-909.

Raj, A., Stephens, M. and Pritchard, J.K. (2014) fastSTRUCTURE: Variational inference of population structure in large SNP data sets. *Genetics* **197**, 573-589.

Resende, R.T., Resende, M.D.V., Silva, F.F., Azevedo, C.F., Takahashi, E.K., Silva-Junior, O.B. and Grattapaglia, D. (2017a) Regional heritability mapping and genome-wide association identify loci for complex growth, wood and disease resistance traits in *Eucalyptus*. *New Phytol* **213**, 1287-1300.

Resende, R.T., Resende, M.D.V., Silva, F.F., Azevedo, C.F., Takahashi, E.K., Silva, O.B. and 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.

Rezende, G.D.S.P., de Resende, M.D.V. and de Assis, T.F. (2014) Eucalyptus breeding for clonal forestry. In: *Challenges and opportunities for the world's forests in the 21st Century* (Fenning, T. ed) pp. 393-424. Dordrecht: Springer Netherlands.

Segura, V., Vilhjálmsson, B.J., Platt, A., Korte, A., Seren, Ü., Long, Q. and Nordborg, M. (2012) An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat Genet* **44**, 825.

Seymour, D.K., Chae, E., Grimm, D.G., Pizarro, C.M., Habring-Muller, A., Vasseur, F., Rakitsch, B., Borgwardt, K.M., Koenig, D. and Weigel, D. (2016) Genetic architecture of nonadditive inheritance in *Arabidopsis thaliana* hybrids. *P Natl Acad Sci USA* **113**, E7317-E7326.

Shen, Q.X., Chen, C.N., Brands, A., Pan, S.M. and Ho, T.H.D. (2001) The stress- and abscisic acid-induced barley gene HVA22: developmental regulation and homologues in diverse organisms. *Plant Mol Biol* **45**, 327-340.

Silva-Junior, O.B., Faria, D.A. and 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 Phytol* **206**, 1527-1540.

Sorek, N., Szemenyei, H., Sorek, H., Landers, A., Knight, H., Bauer, S., Wemmer, D.E. and Somerville, C.R. (2015) Identification of MEDIATOR16 as the *Arabidopsis* COBRA suppressor MONGOOSE1. *P Natl Acad Sci USA* **112**, 16048-16053.

Tan, B., Grattapaglia, D., Wu, H.X. and Ingvarsson, P.K. (2018) Genomic relationships reveal significant dominance effects for growth in hybrid *Eucalyptus*. *Plant Sci* **267**, 84-93.

Tan, B.Y., Grattapaglia, D., Martins, G.S., Ferreira, K.Z., Sundberg, B. and Ingvarsson, P.K. (2017) Evaluating the accuracy of genomic prediction of growth and wood traits in two *Eucalyptus* species and their F1 hybrids. *Bmc Plant Biol* **17**.

Tang, Z., Xu, J., Yin, L., Yin, D., Zhu, M., Yu, M., Li, X., Zhao, S. and Liu, X. (2019) Genome-Wide Association Study Reveals Candidate Genes for Growth Relevant Traits in Pigs. *Front Genet* **10**.

Vanous, A., Gardner, C., Blanco, M., Martin-Schwarze, A., Wang, J., Li, X., Lipka, A.E., Flint-Garcia, S., Bohn, M., Edwards, J. and Lübberstedt, T. (2019) Stability Analysis of Kernel Quality Traits in Exotic-Derived Doubled Haploid Maize Lines. *Plant Genome* **12**.

Volker, P.W., Potts, B.M. and Borralho, N.M.G. (2008) Genetic parameters of intra- and inter-specific hybrids of *Eucalyptus globulus* and *E.nitens*. *Tree Genet Genomes* **4**, 445-460.

Wallace, J.G., Larsson, S.J. and Buckler, E.S. (2014) Entering the second century of maize quantitative genetics. *Heredity* **112**, 30-38.

Wallace, J.G., Zhang, X., Beyene, Y., Semagn, K., Olsen, M., Prasanna, B.M. and Buckler, E.S. (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 Sci* **56**, 2365-2378.

Wang, J., Ding, J., Tan, B., Robinson, K.M., Michelson, I.H., Johansson, A., Nystedt, B., Scofield, D.G., Nilsson, O., Jansson, S., Street, N.R. and Ingvarsson, P.K. (2018) A major locus controls local adaptation and adaptive life history variation in a perennial plant. *Genome Biol* **19**, 72.

Ward, B.P., Brown-Guedira, G., Kolb, F.L., Van Sanford, D.A., Tyagi, P., Sneller, C.H. and Griffey, C.A. (2019) Genome-wide association studies for yield-related traits in soft red winter wheat grown in Virginia. *PloS One* **14**, e0208217.

Zhang, Y., Wu, H.L., Wang, N., Fan, H.J., Chen, C.L., Cui, Y., Liu, H.F. and Ling, H.Q. (2014) Mediator subunit 16 functions in the regulation of iron uptake gene expression in *Arabidopsis*. *New Phytol* **203**, 770-783.

Zhao, Y., Wang, H., Bo, C., Dai, W., Zhang, X., Cai, R., Gu, L., Ma, Q., Jiang, H., Zhu, J. and Cheng, B. (2019) Genome-wide association study of maize plant architecture using F1 populations. *Plant Mol Biol* **99**, 1-15.

Zhen, G., Qin, P., Liu, K.Y., Nie, D.Y., Yang, Y.Z., Deng, X.W. and He, H. (2017) Genome-wide dissection of heterosis for yield traits in two-line hybrid rice populations. *Sci Rep* **7**.

Zheng, Y., Liao, C., Zhao, S., Wang, C. and Guo, Y. (2016) The Glycosyltransferase QUA1 Regulates Chloroplast-Associated Calcium Signaling During Salt and Drought Stress in Arabidopsis. *Plant and Cell Physiology* **58**, 329-341.

Zhu, X.-M., Shao, X.-Y., Pei, Y.-H., Guo, X.-M., Li, J., Song, X.-Y. and Zhao, M.-A. (2018) Genetic Diversity and Genome-Wide Association Study of Major Ear Quantitative Traits Using High-Density SNPs in Maize. *Front Plant Sci* **9**.