UNIVERSITY OF BRASÍLIA
BIOLOGICAL SCIENCES INSTITUTE
CELL BIOLOGY DEPARTMENT
MOLECULAR BIOLOGY PROGRAM

THESIS

Genomic Selection and Genome-Wide Association Studies for growth traits in breeding populations of *Eucalyptus*

BÁRBARA MÜLLER SALOMÃO DE FARIA

THESIS PRESENTED TO THE
UNIVERSITY OF BRASÍLIA TO OBTAIN THE DEGREE OF
DOCTOR OF SCIENCE IN MOLECULAR BIOLOGY

BRASÍLIA, 2017

UNIVERSIDADE DE BRASÍLIA
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
DEPARTAMENTO DE BIOLOGIA CELULAR
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOLOGIA MOLECULAR

TESE DE DOUTORADO

Seleção Genômica e Estudos de Associação Genômica Ampla para características de crescimento em populações de melhoramento de *Eucalyptus*

BÁRBARA MÜLLER SALOMÃO DE FARIA

Orientador: Dario Grattapaglia

Tese apresentada à Universidade de Brasília, como parte das exigências do Programa de Pós-Graduação em Biologia Molecular, para obtenção do título de *Doctor Scientiae*.

BRASÍLIA, 2017

TERM OF APPROVAL

Thesis presented to the Department of Cell Biology at University of Brasília, as a requirement to obtain the degree of Doctor of Science in Molecular Biology.

Thesis presented, on 12/12/2017, to the Examining Committee:

Prof. Dr. Dario Grattapaglia – Advisor (Chair)
Embrapa Genetic Resources and Biotechnology (Cenargen)

Prof. Dr. Matias Kirst – External Advisor (Co-chair)
University of Florida (UF)

Prof. Dr. Robert Neil Gerard Miller – Internal Member
University of Brasília (UnB)

Prof. Dr. Alexandre Siqueira Guedes Coelho – External Member
Federal University of Goiás (UFG)

Dr. Orzenil Bonfim da Silva Junior – External Member
Embrapa Genetic Resources and Biotechnology (Cenargen)

To my true love, Leandro; and to my parents, Rosirene and Washington, who inspired and supported me.

"*It is not the strongest of the species that survives, nor the most intelligent that survives. It is the one that is most adaptable to change.*" - Charles Darwin

FINANCIAL SUPPORT

This work was supported by CNPq (National Council for Scientific and Technological Development) through the granting of fellowships in Brazil (CNPq 140846/2014-0) and in U.S.A. (CNPq 232580/2014-6), by FAP-DF (The Foundation Support Research DF) through the project NEXTREE-"Nucleus of excellence for applied forestry genomics" and by Embrapa Genetic Resources and Biotechnology (Cenargen / Brazilian Agricultural Research Corporation) through the project 03.11.01.007.00.00 "Genomic Selection in *Eucalyptus*: development of genomic predictions models for frost tolerant *E. benthamii*". The field trial logistics for *E. benthamii* was supported by GOLDEN TREE, the *E. pellita* by COMIGO, and the four *E. grandis x E. urophylla* hybrid breeding populations by FIBRIA, CENIBRA and INTERNATIONAL PAPER of Brazil.

ACKNOWLEDGMENTS

I am very thankful to my parents, Rosirene Müller Salomão and Washington Lylly de Faria, and my grandmother Rosmary Müller Salomão, for all affection, dedication, support and unconditional love. Especially, to my parents, for all the efforts destined for my education, for moral values and for discipline, during my personal and professional development.

I am extremely grateful to my husband, Dr. Leandro Neves, for all love, support and patience during these four years of PhD. Also for the luck of having an exemplary professional by my side, understanding and for helping me in the bioinformatics analysis, for the attention and dedication in all the stages of this work.

I am very grateful to my advisor, Dr. Dario Grattapaglia, for constant supervision, trust and opportunity. In addition, I am grateful for the example of ethics and professional commitment and for believing in my professional development.

I am grateful to Dr. Matias Kirst, Dr. Alexandre Coelho, Dr. Orzenil Silva-Junior, Dr. Márcio Resende Jr., Dr. Patricio Muñoz and Dr. Salvador Gezan, who collaborated in different aspects of the development and conclusion of this work. Especially to Dr. Matias for the supervision at University of Florida and for the support with the infrastructure to conduct my research.

I am grateful to Dr. Janeo de Almeida Filho for the contributions in the statistical analysis and in the discussion of my results, for sharing his knowledge and for the friendship.

I am grateful to my former advisors, Dr. Rosana Vianello and Dr. Everaldo de Barros, for always contributing to my personal and professional development.

Especially to Dr. Rosana for the friendship and the academic collaboration during my PhD.

I thank all the Forest Genomics Laboratory colleagues during my research component at University of Florida, mainly to Dr. Annette Fahrenkrog, Dr. Fernanda Gaiotto, Dr. Flora Bittencourt, Dr. Ananda Aguiar, Dr. Maria Teresa and to Msc. Christopher Dervinis. Thank you for your support and for always keeping alive the need to learn and share new knowledge.

I thank Dr. Robert Miller who collaborated during the qualification examination and assistance in all the administrative procedures of the University of Brasília to assure that everything went well during the developed of my research and conclusion of my PhD.

I thank Msc. Antonielle Monclaro for friendship and willingness to help me with all the documentation and requirements of the University of Brasília.

I am very thankful to my friends and family for the support and to understand my absence. Especially to my brother Alexandre Salomão, my sister Natália Müller and my brother Vinícius de Faria for the support.

I would like to acknowledge the University of Brasília (UnB), and particularly the Department of Cell Biology and the Molecular Biology Program (PPGBM) for accepting me to the PhD and opportunity to perform this work.

I would like to acknowledge the Embrapa Genetic Resources and Biotechnology (Cenargen), Federal University of Goiás, University of Florida and CNPq (National Council for Scientific and Technological Development), for the infrastructure, classes, technical and financial support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

RESUMO

Esta tese de doutorado apresenta resultados de pesquisa em seleção genômica ampla (GS) e estudos de associação genômica ampla (GWAS) em seis populações de melhoramento de *Eucalyptus*, com o objetivo de avaliar o potencial destas abordagens em explicar a herdabilidade, detectar associações significativas e predizer fenótipos de características de crescimento. No primeiro capítulo foram comparadas as abordagens de predição genômica e associação genômica ampla para características de crescimento em populações de melhoramento de *Eucalyptus benthamii* (n = 505) e *Eucalyptus pellita* (n = 732). Ambas as espécies são de crescente interesse comercial para o desenvolvimento de germoplasma adaptado a estresses ambientais. A capacidade preditiva atingiu 0,16 em *E. benthamii* e 0,44 em *E. pellita* para crescimento em diâmetro. As capacidades preditivas usando BLUP genômico ou diferentes métodos Bayesianos atingiram resultados semelhantes, indicando que as características de crescimento se ajustam adequadamente ao modelo infinitesimal. Nenhuma diferença foi detectada na capacidade preditiva quando diferentes conjuntos de SNPs foram utilizados, com base na posição (equidistantes no genoma, dentro de genes, podados considerando o desequilíbrio de ligação ou em cromossomos individuais), desde que o número total de SNPs utilizados fosse superior a 5.000. As capacidades preditivas obtidas pela remoção de parentesco entre as populações de treinamento e validação caíram para quase zero em *E. benthamii* e foram reduzidas pela metade em *E. pellita*. Esses resultados corroboram a visão atual de que o parentesco é o principal motor da predição genômica, embora algum desequilíbrio de ligação histórico provavelmente tenha sido capturado para *E. pellita*. Um estudo de associação genômica identificou apenas uma associação significativa para volume em *E. pellita*, ilustrando o fato de que, embora a predição genômica seja capaz de explicar grandes proporções da herdabilidade, muito pouco ou quase nada é capturado em associações significativas usando a abordagem de GWAS nas populações de melhoramento do tamanho avaliado neste estudo. Este estudo forneceu dados experimentais adicionais que indicam

perspectivas positivas de usar dados genômicos para capturar grandes proporções de herdabilidade e predizer características de crescimento em espécies florestais com acurácias iguais ou melhores do que aquelas capturadas pela seleção fenotípica convencional. Adicionalmente, esses resultados documentaram a superioridade da abordagem de GS na capacidade de capturar grandes proporções da variância genética para crescimento, em comparação com o valor limitado da abordagem de GWAS ao se considerar aplicações no melhoramento operacional. A maioria dos estudos de GWAS em plantas, no entanto, assim como este descrito acima, tem sofrido com um poder estatístico limitado, especialmente para características complexas. Tamanhos amostrais maiores são necessários no sentido de aumentar a capacidade de detecção de variantes, especialmente aquelas de baixa frequência e de pequeno efeito. Devido aos desafios logísticos e altos custos para aumentar o tamanho das populações em estudos de associação, uma alternativa tem sido a implementação de Joint-GWAS, utilizando informações combinadas de populações independentes. Joint-GWAS utiliza diferentes abordagens estatísticas para combinar os resultados de múltiplos estudos em um esforço para aumentar o poder de detecção em relação a estudos individuais, melhorar as estimativas do tamanho dos efeitos e/ou resolver a incerteza quando resultados dos estudos individuais não concordam. No segundo capítulo desta tese de doutorado foi realizado um estudo de associação genômica ampla utilizando dados de quatro populações independentes para características de crescimento, montando assim uma população de associação consideravelmente maior do que estudos anteriores em espécies florestais. Dados de um total de 3.373 árvores de quatro populações híbridas de *Eucalyptus grandis* x *Eucalyptus urophylla* não relacionadas, cada uma delas com 758 a 979 indivíduos, foram utilizados. Estas populações haviam sido genotipadas com uma plataforma de SNP comum desenvolvida para *Eucalyptus* permitindo assim a implementação de Joint-GWAS. O impacto da correção para estrutura de população e/ou parentesco sobre a capacidade de detecção de associações significativas foi explorado utilizando seis modelos estatísticos com base na análise de SNPs individuais. Uma redução drástica no

número de associações significativas foi observada ao se adotar correções mais rigorosas. Foi avaliado ainda o desempenho de diferentes abordagens de mapeamento de associação utilizando segmentos genômicos contendo vários SNPs em contrates com SNPs individuais. A abordagem de mapeamento de herdabilidades regionais, nas quatro populações analisadas de forma independente, identificou regiões genômicas que explicaram individualmente 3-13% da herdabilidade genômica. Variantes raras foram detectadas usando abordagens de Joint-GWAS baseadas em conjuntos de SNPs dentro de genes e em segmentos específicos. Associações foram detectadas em genes relacionados à biossíntese da parede celular e resistência à doença, sugerindo potenciais efeitos pleiotrópicos no crescimento da árvore. De maneira geral, o aumento do tamanho amostral e a aplicação de diferentes abordagens de análise combinada de populações ainda revelaram um número limitado de associações, corroborando a complexidade de características de crescimento e a provável participação de um grande número de variantes de pequeno efeito de difícil detecção no controle de crescimento. Entretanto, estes resultados indicam ainda que à medida que mais programas de melhoramento de *Eucalyptus* adotarem genômica para predizer fenótipos com base em uma plataforma SNP comum, conjuntos de dados cada vez maiores ficarão disponíveis e Joint-GWAS como descrito neste estudo de forma inédita em espécies florestais, será capaz de contribuir para a identificação de SNPs ou segmentos genômicos que controlam proporções relevantes da herdabilidade.

Palavras-chave: Predição genômica, análise de associação, genotipagem de alto desempenho, parentesco, melhoramento florestal, eucalipto.

ABSTRACT

This doctoral thesis presents results of research in genomic selection (GS) and genome-wide associations studies (GWAS) in six *Eucalyptus* breeding populations, with the objective of evaluating the potential of these approaches in explaining heritability, detecting significant associations and predicting phenotypes of growth traits. In the first chapter, genomic prediction approaches and association studies for growth traits in *Eucalyptus benthamii* (n = 505) and *Eucalyptus pellita* (n = 732) breeding populations were compared. Both species are of increasing commercial interest for the development of germplasm adapted to environmental stresses. Predictive ability reached 0.16 in *E. benthamii* and 0.44 in *E. pellita* for diameter growth. Predictive abilities using either Genomic BLUP or different Bayesian methods reached similar results, indicating that growth adequately fits the infinitesimal model. No difference was detected in predictive ability when different sets of SNPs were utilized, based on position (equidistantly genome-wide, inside genes, linkage disequilibrium pruned or on single chromosomes), as long as the total number of SNPs used was above ~5,000. Predictive abilities obtained by removing relatedness between training and validation sets fell near zero for *E. benthamii* and were halved for *E. pellita*. These results corroborate the current view that relatedness is the main driver of genomic prediction, although some historical linkage disequilibrium was likely captured for *E. pellita*. A genome-wide association study identified only one significant association for volume growth in *E. pellita*, illustrating the fact that while genome-wide regression is able to account for large proportions of the heritability, very little or none is captured into significant associations using GWAS in breeding populations of the size evaluated in this study. This study provided further experimental data supporting positive prospects of using genome-wide information to capture large proportions of trait heritability and predict growth traits in trees with accuracies equal or better than those attainable by phenotypic selection. Additionally, our results documented the superiority of the whole-genome regression approach in accounting for large proportions of the heritability of

complex traits, such as growth, in contrast to the limited value of the local GWAS approach toward breeding applications. Most GWAS in plants, like the one described above, have suffered from limited statistical power especially for complex traits. Larger sample sizes are needed to enhance the ability to identify variants, especially those of low-frequency and small effect. Due to the challenges and high costs of increasing sample size in GWAS, an alternative has been to implement Joint-GWAS, using the combined information from independent populations. Joint-GWAS use different statistical approaches to combine the results from multiple studies in an effort to increase detection power over individual studies, improve estimates of the size of the effect and/or to resolve uncertainty when reports from individual studies disagree. In the second chapter of this doctoral thesis, a GWAS for growth traits was performed by assembling a considerably larger association population than any previous GWAS in forest trees. Data for a total of 3,373 trees across four unrelated *Eucalyptus grandis* x *Eucalyptus urophylla* hybrid breeding populations, with samples sizes varying between 758 and 979 individuals, were used. These populations had been genotyped with a common SNP platform for *Eucalyptus* species, thus allowing a Joint-GWAS implementation. The impact of correcting for population structure and/or relatedness on the detection of significant associations was explored using six single-SNP GWAS models. A drastic reduction in the number of significant associations detected was observed when more stringent correction was adopted. We also evaluated the performance of different segment-based GWAS approaches involving several SNPs simultaneously in comparison to single-SNP analyses. Regional heritability mapping, in these four populations independently, pinpointed genomic regions that individually explained 3-13% of the genomic heritability. Rare variants were detected using gene and region-based Joint-GWAS approaches. Associations were detected into genes related to cell wall biosynthesis and disease resistance, suggesting potential pleiotropic effects on tree growth. In general, the increase in sample size and the application of different approaches of Joint-GWAS still revealed a limited number of associations, corroborating the complexity of growth traits and the likely participation of a large

number of variants of small effect of difficult detection in the control of growth traits. However, these results further indicate that as more *Eucalyptus* breeding programs adopt genomics to predict phenotypes based on a common SNP platform, increasing datasets will be available and Joint-GWAS as described in this study for the first time in forest trees, will be able to contribute to the identification of SNPs or genomic segments controlling relevant portions of trait heritability.

Keywords: Genomic prediction, association analysis, high-throughput genotyping, relatedness, tree breeding, eucalypts.

LITERATURE REVIEW

Eucalyptus genus

The *Eucalyptus* L'Hér. genus belongs to the Myrtaceae family and includes more than 800 species native to Australia and adjacent islands in Oceania (Brooker, 2000; Doughty, 2000). *Eucalyptus* has a total of ten subgenera described, being *Symphyomyrtus* the most important one with more than 470 species (Grattapaglia et al. 2012). This main subgenus has different sections, being the *Latoangulatae* section represented by species widely planted in tropical areas, such as *E. grandis*, *E. urophylla* and *E. pellita*. On the other hand, the *Maidenaria* section is represented by species adapted to temperate regions, such as *E. globulus*, *E. nitens* and *E. benthamii* (Brooker 2000). *Eucalyptus* is the genus including the most widely planted hardwood trees in the world, mainly due to their versatile applications (windbreak, landscape, bioenergy, pulp, paper and solid wood), superior adaptability and wood quality (Myburg et al. 2007; Grattapaglia and Kirst 2008).

Species of the *Eucalyptus* genus were quickly used for forest plantations, after their discovery by the Europeans in the XXVIII century (Eldridge et al. 1993). These species were introduced initially in several countries, such as India, France, Chile, Brazil, South Africa and Portugal, mainly due to their fast growth and high adaptability under different environmental conditions (Doughty 2000; Myburg et al. 2007; Grattapaglia and Kirst 2008). Currently, many countries have extensive germplasm collections of *Eucalyptus* selected for forest tree plantation (Grattapaglia and Kirst 2008). The species of *Eucalyptus* planted in Brazil are characterized by excellent adaptation and fast growth (Eldridge et al. 1993). *E. grandis* and *E. urophylla* are the most commercially important and broadly planted species, and together with their hybrids, are highly preferred for pulp and solid wood production in tropical areas (Henry 2011). The interspecific hybrids between *E. grandis* and *E. urophylla* have been grown in large operational scale in Brazil

due to their combination of desirable traits (e.g. superior wood properties and disease resistance), and represent most of the genetic background present in the breeding programs (Myburg et al. 2007).

Eucalyptus benthamii Maiden & Cambage (Camden white gum) is a rare species known for its restricted occurrence in its natural range, southwest of Sydney (Australia), and is now considered vulnerable to extinction (Benson 1985; Jovanovic and Booth 2002; Butcher et al. 2005). Agricultural activities and the expansion of the Sydney metropolitan area are the main factors reducing the *E. benthamii* natural population (Butcher *et al.*, 2005). However, *E. benthamii* is a species of growing commercial interest due to its cold tolerance combined with its rapid growth and high-quality pulpwood (Harwood 2011; Döll-Boscardin et al. 2012; Baccarin et al. 2015). Breeding programs in subtropical regions, such as southern Brazil, southeastern USA, Uruguay, Argentina, Chile and China, generally use *E. benthamii* as a pure species or in combinations with other species, such as *E. dunnii* (Butcher et al. 2005; Brondani et al. 2011; Booth 2012; Brondani et al. 2012a; Pirraglia et al. 2012; Brondani et al. 2012b; Arnold et al. 2015; Baccarin et al. 2015; Yu and Gallagher 2015). Several studies are currently underway to identify frost tolerant *Eucalyptus* germplasm adapted to temperate regions (Arnold et al. 2015; dos Santos et al. 2015; Yu and Gallagher 2015). The identification of these cold tolerant species and their use as pure species or in hybrids are strategies used to expand the production of *Eucalyptus* in previously unexplored areas, such as the extreme south of Brazil, the USA and China (Martins et al. 2013; Stanturf et al. 2013; Arnold et al. 2015; Baccarin et al. 2015; Yu and Gallagher 2015). *E. benthamii* is a species with great potential for this purpose, because the trees can survive in absolute minimum temperatures of -6º to -10º C (Lin et al. 2003; Ebling et al. 2012; Yu and Gallagher 2015).

Eucalyptus pellita F. Muell. (large-fruited red mahogany) occurs in tropical regions, more precisely in two disjointed natural forests, one in southern New Guinea (on the Papua island between Papua New Guinea and Indonesia) and the other in

northern Australia in the state of Queensland (House and Bell 1996; Harwood et al. 1997; Harwood 1998; Jovanovic and Booth 2002; Le et al. 2009; Hung et al. 2015). *E. pellita* has moderate to high levels of genetic diversity when compared to taxonomically related species, such as *E. urophylla* and *E. grandis* (House and Bell 1996; Le et al. 2009). *E. pellita* belongs to the group of the nine species of *Eucalyptus* most planted in the world, known as the "big nine". This group is represented by *E. camaldulensis*, *E. grandis*, *E. tereticornis*, *E. globulus*, *E. nitens*, *E. urophylla*, *E. saligna*, *E. dunnii* and *E. pellita*, which together with their hybrids represent 95% of the world's planted eucalypts (Harwood 2011; Stanturf et al. 2013). *E. pellita* has been recognized as a promising species, in relation to other *Eucalyptus* species for operational industrial plantation in tropical regions, because of fast growth and resistance to diseases and pests (Harwood 1998; Jovanovic and Booth 2002; Leksono et al. 2006; Leksono et al. 2008; Brawner et al. 2010; Mauro et al. 2010; Zauza et al. 2010; Agustini et al. 2014; Yuskianti et al. 2014). *E. pellita* displays high wood density being used for the production of charcoal, paper, pulp and solid wood for construction in general (bridges, poles, flooring, panels, etc.) and for furniture production (Harwood et al. 1997; Leksono et al. 2006; Oliveira et al. 2010; Redman and Mcgavin 2010; Sun et al. 2013; Hung et al. 2015). This species has been planted extensively in the last two decades in countries such as Brazil, India, Congo, Vietnam, southern China and many other tropical and subtropical countries (Harwood et al. 1997; Bernardo et al. 1998; Le et al. 2009; Brawner et al. 2010; Harwood 2011; Sun et al. 2013; Hung et al. 2015). More recently, *E. pellita* is being grown on an industrial scale in Indonesia, where there is a growing demand for genetically improved seeds to maximize plant productivity (Leksono et al. 2006; Leksono et al. 2008; Harwood 2011; Agustini et al. 2014; Sulichantini et al. 2014; Yuskianti et al. 2014).

Genomic and molecular tools applied to forest tree breeding

Brazilian transgenic *Eucalyptus* tree

Before discussing the main genomic technologies and approaches used towards the integration of genomic analysis into breeding, a very brief detour is made to mention the recent development of the first *Eucalyptus* transgenic tree developed by FuturaGene, a company of the Suzano Celulose group. This transgenic event was approved for commercial use in Brazil from the biosafety point of view by CTNBio (Brazilian National Technical Commission on Biosafety), although it is still subject to appeals at the level of the CNBS (Brazilian National Biosafety Council). The company claims that this transgenic eucalypts named H421, containing the overexpressed CEL1 gene (endo-1,4-β-glucanase) of *Arabidopsis thaliana*, shows a gain of about 20% in growth over the same non-transgenic genotype, clone SP530, which could reduce the harvest time to only five and a half years (Ledford 2014).

Experimental results following the company's first trials, reported in the application submitted to CTNBio, have indicated that this gain was considerably lower (between 4.0 to 8.5%) when the transgenic clone was planted in different environments and in larger scale plantations. In addition, even if the claimed growth gain occurs, the particular eucalypt wild type clone SP530, even transgenic (H421), has a productivity below the majority of public clones used commercially for large-scale plantation in Brazil. It will be interesting, therefore, to verify if there will be potential productivity gain by inserting this transgenic construction into the best current public clones, as well as those derived from breeding programs (D. Grattapaglia personal communication). The advantage of transgenic eucalypt in Brazil is that it is an exotic species (Ledford 2014; da Silva et al. 2017), such that the environmental risk of gene flow to wild relatives does not apply (Fort et al. 2004; Strauss et al. 2004; Brunner et al. 2007; da Silva et al. 2017). Unfortunately, however, the ultimate commercial plantation of this transgenic tree in Brazil is still

controversial, due to the commercial risk of compromising the large industry of organic honey production based on eucalypt plantation by the flow of transgenic pollen into honey. It is expected that transgenic technologies in *Eucalyptus* in Brazil will likely have a key role to solve the growing problems of pests.

Molecular markers and genotyping technologies

The increase in forest productivity and the refinement of physical and chemical wood traits through the application of genomic and molecular tools in genetic improvement is seen today as an important step to maintain the competitiveness and growth of the forest-based industry (Grattapaglia and Kirst 2008; Grattapaglia and Resende 2011; Ledford 2014). SNPs (Single Nucleotide Polymorphisms) are the molecular markers most used in genetic analysis because they are the most abundant genetic variation, occurring throughout the genome (Ching et al. 2002). These markers are based on the detection of polymorphisms resulting from the alteration of a single nucleotide and are the genetic basis of most allelic variations. Due to its abundance in the genome, its low mutation rate, and the possibility of automated detection, SNPs are increasingly being used as molecular markers (Mccouch et al. 2010; Yu et al. 2011; Gupta and Jr 2013). SNP markers, analyzed on a large scale, have been increasingly used for various genetic studies and in plant breeding, such as Genome-Wide Association Studies (GWAS) and Genomic Selection (GS) (Mammadov et al. 2012; Poland and Rife 2012; Thomson 2014), with the aim of identifying the complex relationships between genotypic and phenotypic variation. These markers have been widely used in a number of other areas, such as ecology and evolution, assisting in molecular phylogeny, population genetics, and conservation studies of natural populations (Allendorf et al. 2010; Narum et al. 2013).

Over the years, the costs of SNP genotyping technologies using arrays have become more accessible and the Illumina Infinium® and Affymetrix Axiom® chips were developed allowing the fast genotyping of thousands of previously identified

SNPs via re-sequencing on various tree species. The species of *Picea* (Pavy et al. 2008), *Pinus* (Eckert et al. 2010) and *Eucalyptus* (Grattapaglia et al. 2011) genus were the first to have SNP genotyping chips developed. Similar and larger scale resources were then developed for *Malus domestica* (apple) (Crowhurst et al. 2012; Bianco et al. 2016), *Prunus persica* (peach) (Verde et al. 2012), *Populus trichocarpa* (poplar) (Geraldes et al. 2013), *Prunus avium* and *P. cerasus* (cherry) (Peace et al. 2012), *Picea glauca* (white spruce) (Pavy et al. 2013), *Eucalyptus* (eucalypts) (Silva-Junior et al. 2015), *Populus nigra* (black poplar) (Faivre-Rampant et al. 2016) and *Pinus pinaster* (maritime pine) (Plomion et al. 2016). However, the initial cost (*set up cost*) of chip development is still relatively high, unless a couple of thousands of samples are genotyped, which may not be feasible for studies of species of limited commercial interest. In such cases, genomic strategies capable of reducing the complexity of the genome, discovering and genotyping SNPs at the same time, such as Genotyping by Sequencing (GbS, Elshire et al. 2011), RADseq (Davey et al. 2011; Davey et al. 2013) or selective sequence capture on arrays or in solution (Mamanova et al. 2010), have proven useful. In forest trees, genotyping by sequencing methods such as DArTseq were initially used in *Eucalyptus* (Sansaloni et al. 2011) and sequence capture methods have been used in *Populus* (Zhou and Holliday 2012) and *Pinus* (Neves et al. 2013; Neves et al. 2014). A number of studies then followed using different methods of genome complexity reduction in different forest tree species (Lu et al. 2016; Pavy et al. 2016; Plomion et al. 2016; Suren et al. 2016; Fitz-Gibbon et al. 2017).

Quantitative trait locus (QTL) mapping

Molecular tools that allow the identification of polymorphisms in DNA hold the promise to provide new opportunities for selection of growth traits, adaptability to climatic conditions and wood properties of cultivated trees (Grattapaglia et al. 2009). The development of elite trees via genetic improvement deals with the challenge of simultaneously advancing, through selection and recombination, several quantitative traits of silvicultural and industrial relevance (Grattapaglia et

al. 2009; Grattapaglia 2014). In addition, these phenotypic traits typically present complex genetic control, low heritability, strong environment interaction, low juvenile-adult correlation, and late expression (Grattapaglia 2014). Significant progress has been made in the last two decades in the development of genetic maps in trees with highlight to *Pinus* and *Eucalyptus* (Neale 2007; Grattapaglia and Kirst 2008). Hundreds of RFLPs, RAPDs, AFLPs, DArTs, SSRs and SNPs markers are available today for genetic analysis in forest tree species. Several papers have described the success in identifying QTLs (Quantitative Trait Loci). The QTL mapping has been important in dissecting complex traits, revealing from the phenotype the positions of genomic regions that affect the target traits (Grattapaglia et al. 2009). Different QTLs for components of productivity, wood quality, resistance to abiotic and biotic stresses have been reported in the forest trees literature (Kirst et al. 2004; Marques et al. 2005; Bundock et al. 2008; Freeman et al. 2008; Neale and Ingvarsson 2008; Grattapaglia et al. 2009; Alves et al. 2012; Bartholomé et al. 2013; Zarpelon et al. 2014; Butler et al. 2016).

Despite dozens or even hundreds of QTLs have been mapped to date, the information generated has not been immediately useful for Marker-Assisted Selection (MAS) in breeding with some exceptions (Bernardo 2008; Grattapaglia 2014). The MAS approach based on the standard logic of first identifying and validating the genes and/or QTLs that control the trait has been shown to be inefficient for the advancement of the breeding of multifactorial traits by a series of issues, for example: (i) only few QTLs are detected capturing limited fraction of variation given the allelic variation from biparental populations; (ii) the populations used are typically small, leading to overestimation of the magnitude of the QTL effects; and (iii) the unpredictable performance of the interaction between favorable QTL alleles and different genetic backgrounds, different locations and different ages (Bernardo 2008; Grattapaglia and Kirst 2008; Grattapaglia et al. 2009; Grattapaglia and Resende 2011; Korte and Farlow 2013; Grattapaglia 2014).

Genome-Wide Association Studies (GWAS)

Principles and applications of GWAS

Over the years, Genome-Wide Association Studies (GWAS) have been conducted for the detection and characterization of QTLs/genes that control different traits, promising their potential application in medicine and breeding (Korte et al. 2012; Korte and Farlow 2013; Morris et al. 2013; Fan et al. 2015; Spindel et al. 2015; Henrique et al. 2016; Zhu et al. 2016). Despite recent advances, association studies are a consolidated approach whose first empirical results were published more than a decade ago (Visscher et al. 2017) for human diseases (Klein et al. 2005; DeWan et al. 2006; Burton et al. 2007). The principle of association analysis is to explore the historical recombination events and the linkage disequilibrium (LD) structure resulting in the genome of natural populations or germplasm banks (Zhu et al. 2008; Khan and Korban 2012; Visscher et al. 2012; Korte and Farlow 2013). From the founders of these populations, it is expected that the recombination events in successive meiosis have randomized the LD blocks, generating a low LD structure, which allows better resolution in the detection of associations between markers and genes/QTLs of interest (Zhu et al. 2008). As a consequence of this high number of meiosis occurring in the population history, it is expected that a marker stated to be significantly associated with a phenotype will be physically close to the causal variant. This situation contrasts with what is observed in a biparental population used in the QTL mapping via linkage, which has only one recent recombination event and therefore presents a large extension of LD, which facilitates the detection of associations, but prevents a better resolution (Hamblin et al. 2011).

The success of GWAS, from the standpoint of resolution, depends on the ability to detect associations in a genome with low extension of LD between genotyped SNPs and causal variants, requiring a high-density of markers consistent with the extent of LD in the genome. At the same time, a large number of individuals,

usually hundreds, thousands or tens of thousands are needed to provide sufficient statistical power for the detection of small effects, often contributed by rare variants in the population (Zhu et al. 2008; Robinson et al. 2014). When compared to QTL mapping analysis with only hundreds of markers in more limited biparental populations in terms of sampling population variation, the GWAS analysis involves tens or hundreds of thousands of markers throughout the genome in a more diverse population, being potentially more powerful for gene discovery (Visscher et al. 2012; Korte and Farlow 2013; Visscher et al. 2014). In addition, GWAS does not require the generation of segregating (biparental) populations, but rather allows the direct use of natural populations, germplasm banks, landraces or breeding populations (Zhu et al. 2008; Khan and Korban 2012). As practical benefits, it is commonly proposed that markers found with a significant association with target traits may potentially be used in breeding programs via MAS (Gowda et al. 2014; Zhang et al. 2015).

The statistical robustness of the association found between alleles at two loci in the genome strongly depends on the allelic frequencies at these loci, such that a rare variant (e.g. with a frequency <0.01) will be at low LD with another common nearby variant, even if both variants are mapped to the same recombination interval (Wray 2005). On the other hand, from the practical point of view, the SNPs present in a genotyping chip in general are selected to be common (Yang et al. 2010; Yang et al. 2011b; Lee et al. 2012), that is, the majority having a Minor Allele Frequency (MAF) greater than 5%. Therefore, GWAS analyses are, for the most part, limited to detecting associations with relatively common causal variants in the population (Visscher et al. 2012; Korte and Farlow 2013). In recent years, by reducing the costs of sequencing and genotyping by sequencing, association studies that include low-frequency alleles have become possible for the identification of rare variants that contribute to the complex phenotypic traits variation (Brachi et al. 2011; Morris et al. 2013; Huang and Han 2014).

GWAS analyses have also been widely applied in human genetics studies to identify loci that influence complex diseases and traits, such as diabetes, schizophrenia, height, and body mass index (Klein et al. 2005; DeWan et al. 2006; Burton et al. 2007; Yang et al. 2010; Yang et al. 2011b; Wray et al. 2013; Yang et al. 2015; Gusev et al. 2016; Marouli et al. 2017). Association studies have been increasingly applied in animal breeding to discover loci related to economically important complex traits, which mainly influence beef and dairy cattle productivity (Goddard and Hayes 2009; Bolormaa et al. 2010; Fortes et al. 2010; Jiang et al. 2010; Bolormaa et al. 2011; Jiang et al. 2014; Raven et al. 2014; Fan et al. 2015; Henrique et al. 2016; Xia et al. 2016). In forest tree species, GWAS were carried out starting from an optimistic concept that praised forest tree populations as ideal for association studies, being proposed as a solution to the dilemma of QTL mapping (Neale and Savolainen 2004). The initial association studies, mainly in the *Pinus* and *Populus* species, were mostly focused on variants present in candidate genes, as there was a lack of genome-wide genotyping platforms and the assumption that individual genes would have a seminal role in the control of complex traits such as drought tolerance, growth and wood properties (Neale and Savolainen 2004; Thumma et al. 2005; Neale 2007; Wegrzyn et al. 2010; Khan and Korban 2012; Guerra et al. 2013; Thavamanikumar et al. 2014; Jaramillo-Correa et al. 2015). Despite these efforts, the results in general were limited with the few associations found explaining only a small proportion of the genetic variation (Grattapaglia et al. 2009). More recently, association studies were performed with genome-wide markers, although still with very limited statistical power due to the low number of individuals (<750) (Porth et al. 2013; Evans et al. 2014; Mckown et al. 2014; Allwright et al. 2016; Du et al. 2016; Fahrenkrog et al. 2016). To date, only three GWA studies have been conducted in *Eucalyptus* species (Cappa et al. 2013; Resende R.T. et al. 2016; Müller et al. 2017). These studies also detected few associations that combined explain small fractions of the genetic variation, with estimations inflated due to the “winner’s curse” (Goddard et al. 2009), and revealing the complexity of the target traits, especially growth traits, which are the pillars of forest tree breeding programs.

Population structure and relatedness, main factors that affect GWAS results

The assembly of individuals in a population for an association study should be done with caution, as the results can be strongly influenced by the population structure or stratification, relatedness and genetic drift (McCarthy et al. 2008; Wray et al. 2013; Jaramillo-Correa et al. 2015). Population structure is the presence of a difference of allelic frequencies between subpopulations, subgroups or families present in a population due to ancestry. Among other factors, it originates from founding events, processes of genetic drift and inbreeding. Relatedness represents the familial relationships between pairs of individuals, and the relative kinship matrix generated from this coefficient of relationship has an enormous impact on GWAS results when breeding populations are used (Müller et al. 2017). It is expected that most of these differences of genomic ancestry between subgroups present in the association population are not related to phenotypic variations and, therefore, if not controlled in the analytical model can be detected as false associations. Population structure and relatedness are the two main factors that can result in the detection of false positives (Yu et al. 2006). In addition, associations may be found to be specific to certain families or subpopulations, and will not be confirmed in other populations, thereby reducing the potential usefulness of their discovery (Korte et al. 2012; Bragg et al. 2015).

The presence of population structure, whether previously known or not, must be corrected using several statistical methods to avoid the detection of false positives (Huang and Han 2014). Generally, mixed linear models are employed for this purpose, dealing with population structure through the use of markers that measure the amount of phenotypic covariance that is due to the genetic relationship (Yu et al. 2006; Korte et al. 2012; Cappa et al. 2013; Eu-ahsunthornwattana et al. 2014; Evans et al. 2014; Huang and Han 2014; Thavamanikumar et al. 2014; Jaramillo-Correa et al. 2015; Du et al. 2016). To reduce these effects, therefore, the genetic relationship due to the population structure is included in the model (e.g. Qmatrix from STRUCTURE software and

significant principal components from Principal Component Analysis, PCA) as a fixed term and/or kinship (Genomic Relationship Matrix, GRM) as a random term (Korte et al. 2012; Korte and Farlow 2013). For example, Cappa et al. (2013) performed a GWA study in *E. globulus* and showed a considerable reduction in the number of associations for growth traits and wood properties after making such corrections to population structure and relatedness using a Unified Mixed Model (UMM). A multi-gene association mapping using 435 unrelated individuals in *Populus* detected more than 400 significant associations for growth traits without any correction for population structure (Du et al. 2016). Despite the fact that this latter study used a natural population, effects of population structure can still be present and need to be accounted to minimize bias due to past relatedness in the evolutionary history of the species. Most likely the vast majority of the 400 associations found are therefore spurious. In fact, another GWA study performed in a natural population with 714 individuals of *Populus nigra* showed a strong decrease of associations declared, especially for growth trait, when population structure and/or family-based correction were incorporated in the model (Allwright et al. 2016).

Challenges and limitations of GWAS

A major limitation of GWA studies carried out to date from the applied standpoint is that only a small number of associations are identified, typically characterized by common alleles in the population (Yang et al. 2010). This is a major limitation for the application of GWAS results in plant and animal breeding in general. Furthermore, generating large association populations to increase power of detection is technically challenging and expensive. In forest trees since the majority of the traits of interest are multifactorial and expected to be controlled by a large number of variants of small effect, considerably larger populations will have to be employed to identify a significant fraction of the relevant alleles (Robinson et al. 2014). As population size increases in association analysis, the power of detection increases and more QTLs/genes with small effects are detected,

however the inconsistency of these effects across different genetic backgrounds and environments can become a problem. In addition, even when rare variants are found associated with quantitative traits in natural populations, it is not directly clear how this information will be used in breeding programs via marker-assisted selection in advanced breeding populations (Collard et al. 2008; Desta and Ortiz 2014). These variants may, for example, already be fixed in these populations or have reduced allelic substitution effects than the segregating variants in the breeding populations and thus are irrelevant from the practical point of view. Finally, even if all loci and respective alleles, each controlling small fractions of the variation in the traits of interest, are found in association studies with high statistical power and well conducted experimental, the ability to use this information via MAS for several traits simultaneously will be difficult.

The major challenge of association studies arises essentially from the underlying hypothesis, now well supported by several studies in plants, animals and humans that complex traits are controlled by a large number of variants, possibly hundreds or thousands, of small effects distributed throughout the genome (Boyle et al. 2017). Thus, since association studies are typically very limited in terms of statistical power and genomic coverage, GWAS will capture only a small proportion of these variants, those with the greatest effect. This leaves much of the variation to be explained; which has been referred to as “missing heritability” (Maher 2008). The “missing heritability” can be comprehended as a proportion of the genetic variance not explained by the QTLs that the association study was able to detect (Manolio et al. 2009). The paradox of “missing heritability” is now a topic that has been extensively investigated in the area of quantitative genomics, both for complex diseases in humans, as well as for many economically important traits in cultivated plants (Eichler et al. 2010; Yang et al. 2010; Brachi et al. 2011; Visscher et al. 2014; Yang et al. 2015). There are numerous possible explanations for the “missing heritability”. One of them is that to fully explain the genetic variance of complex traits, it will be necessary to combine GWAS with other large-scale approaches (e.g. gene expression, methylation, Copy Number Variations (CNVs),

organellar DNA (cpDNA or mtDNA)), with the goal of identifying all the remaining genetic variants that affect the phenotype (Edwards et al. 2013; Robinson et al. 2014). Another possibility is the simple fact that most GWA studies focus on common variants, such that to access rare variants a major increase in the sample size of association studies is required (Agarwala et al. 2013; Cheng and Chen 2013; Robinson et al. 2014). Generally, rare variants are less likely to be in strong LD with the common variants present in the SNP genotyping chips (Robinson et al. 2014). In addition, the phenotypic data used may be imprecise, which reduces the ability to identify associations.

Among the several factors that have been cited in the literature, which include effects of interactions between genes (epistasis), epigenetic variation, *de novo* mutations and others, the increase in sample size of the experimental populations should have the greatest effect on detection power of a GWAS (Robinson et al. 2014). Even if high density SNP chips are replaced by Whole-Genome Sequencing (WGS) to detect low-frequency loci, this alone will not be enough to capture the effects of rare variants, unless a larger number of individuals are employed for its detection (Robinson et al. 2014). Therefore, instead of “missing heritability”, the most appropriate term appears to be “hidden heritability” (Vinkhuyzen 2013). Recently it has been proposed that complex traits are more likely under an “omnigenic” model in contrast to a polygenic model, where the association signals tend to be spread across the genome, including a large number of genes with no direct relevance to the trait (Boyle et al. 2017). As these limitations become more evident, the main change between the past and current GWA studies has been the increased number of individuals employed, usually thousands or hundreds of thousands, which seem to be needed to provide sufficient statistical power to detect the common and low-frequent variants in the population that contribute to the genetic variance (Marouli et al. 2017; Visscher et al. 2017).

A few approaches can be used to increase the power of association studies, namely increasing the number of samples using Meta-GWAS and Joint-GWAS

(Mägi and Morris 2010; Yang et al. 2012; Bernal Rubio et al. 2016) or exploiting multiples SNPs in a segment using region or gene-based GWAS to account for rare and low-frequency variants (Wu et al. 2011; Nagamine et al. 2012; Bakshi et al. 2016). Progress in identifying associated loci with complex traits has been accelerated by large-scale Meta and Joint-analyses through the combination of information coming from multiple populations. 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 and Ioannidis 2013). Joint-GWAS combine the populations prior to the association analysis, leading to more resolution and the detection of more associations for complex traits (Lin and 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). Although sharing individual-level datasets is logistically difficult and for human studies might have ethical restrictions, Joint-GWAS has become more common in plant research due to the ability to replicate genotypes (Li et al. 2016; Wallace et al. 2016; Wu et al. 2016). In some cases, even with medium to large sample size the statistical power to detect associations is usually very small for complex traits, because of their polygenic architecture characterized by the small effect sizes of each individual genetic variant (Visscher et al. 2012). The regional heritability mapping (RHM, Nagamine *et al.*, 2012) is an alternative approach for region-based GWAS 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 between individuals, and it 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 in enriched 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 associations in genes controlling complex trait (Bakshi et al. 2016).

In the case of association studies for forest tree species, although they have resulted in some progress in the identification of causal variants, they are far from providing sufficient information for the practical application in breeding. The proportions of variation explained are too small to impact forest tree breeding of complex traits and unless large fractions of the variation are captured by multiple associated markers, they will hardly have any impact on directional selection (Grattapaglia et al. 2009). The proportion of variation explained by each individual association has generally reached 1-6% (Neale 2007; Grattapaglia and Resende 2011; Korte et al. 2012; Korte and Farlow 2013; Mckown et al. 2014; Du et al. 2016; Nicolas et al. 2016; Resende R.T. et al. 2016), with a small increase of 5-15% when considering the RHM approach in *Eucalyptus* (Resende R.T. et al. 2016). Therefore, it is clear that there is a large number of additional variants of small effects that cannot be detected with the limited dimensions of the experiments and with the application of stringent significance tests (Visscher et al. 2012). GWAS analyses have been limited in explaining genetic variation, even for high heritability traits (Grattapaglia et al. 2009). In humans, a GWA study of approximately 250,000 individuals identified nearly 700 associations in 423 loci, which together accounted for about 20% of heritability for height, a trait that presents high heritability, estimated between 60 to 80% (Wood et al. 2014). For another GWA study for adult height, despite using more than 700,000 individuals, only 83 height-associated variants were detected with lower minor-allele frequencies (0.1-4.8%) (Marouli et al. 2017). Interestingly, the contribution to the phenotype of these rare variants were up to ten times those of the average common variants, representing up to 2 centimeters per allele. In this line, a recent study estimated that more than 100,000 SNPs influence height in human (Boyle et al. 2017), each one with a tiny impact (Callaway 2017). Robinson et al. (2014) when discussing the challenge of how to explain a greater amount of the genetic variation of complex traits propose that this challenge will be fulfilled with the technologies currently available by using much larger sample sizes, better

phenotyping, more focused designs on specific traits, and the integration of multiple sources of genetic and phenotypic information.

Genomic Selection (GS)

Principles and applications of genomic prediction

The use of methodologies involving the development of predictive models of complex phenotypes based on genome-wide genotyping has revolutionized the perspective of the application of genomic information in breeding practice. This methodology dispenses the necessity for prior identification of individual QTLs/genes, focusing exclusively on aspects of operational efficiency and genetic gain. This type of approach, called Genomic Selection (GS) or Genome-Wide Selection (GWS), was proposed almost two decades ago (Meuwissen et al. 2001) and has gained increasing interest and application as a new approach for breeding annual crops (Morrell et al. 2011; Poland et al. 2012; Massman et al. 2013; Crossa et al. 2014; Annicchiarico et al. 2015; Yabe et al. 2016; Acosta-Pech et al. 2017; Bernal-Vasquez et al. 2017), forest trees (Grattapaglia and Resende 2011; Iwata et al. 2011; Resende Jr et al. 2012a; Resende et al. 2012; Zapata-Valenzuela et al. 2012; Resende Jr et al. 2012b; Zapata-Valenzuela et al. 2013; Beaulieu et al. 2014a; Beaulieu et al. 2014b; de Almeida Filho et al. 2016; Lenz et al. 2017; Müller et al. 2017; Resende R.T. et al. 2017; Tan et al. 2017) and fruit plants (Kumar et al. 2012a; Kumar et al. 2012b; Kumar et al. 2015; Muranty et al. 2015; Duangjit et al. 2016; Iwata et al. 2016; Gezan et al. 2017; Migault et al. 2017). GS can be defined as being the simultaneous selection for hundreds or thousands of markers, depending on the organism and extent of linkage disequilibrium, covering the whole-genome. Consequently, it is anticipated that all alleles of interest will be in LD with at least one or more genotyped markers and, therefore, properly captured in predictive models.

Genomic selection, similar to GWAS, uses genotyping with large numbers of markers covering the whole-genome, but differs in that it is not based on the application of significance tests. Therefore, GS estimates simultaneously the effect of all markers on the phenotype of individuals from a representative population. Thus, unlike GWAS that focuses on the detection of individual associations, GS uses all or a large proportion of the markers to predict the phenotype through predictive models. Consequently, GS works on the principle that the LD provided by dense genotyping is sufficient to capture most of the QTLs relevant to the target trait. Avoiding the selection of markers and estimating the effects of markers in a broad and representative training population, GS tends to capture a greater genetic variance for the assessed trait. Therefore, GS mitigates the dilemma of how to capture the "missing heritability" of complex traits, explained by a large number of QTLs of small effects (Manolio et al. 2009; Makowsky et al. 2011).

Genomic selection opens a concrete perspective of significantly accelerating the progress of forest tree species breeding due to the long life cycle and traits that present complex genetic control and late expression (Grattapaglia 2014). The predictive methodologies of GS, dispensing the need to map and locate QTLs/genes, focusing exclusively on increasing efficiency with reduced breeding cycle and increased genetic gain, may have a greater probability of success (Grattapaglia et al. 2009; Grattapaglia and Resende 2011). Only in the last decade that large-scale genotyping technologies have enabled high-marker densities and whole-genome coverage to be achieved at very affordable costs, which rapidly renewed interest in "black box" methodologies for phenotype prediction based on genotype (Goddard and Hayes 2009; Habier et al. 2013). Recent results in the literature, mainly on genetics and animal breeding, are extremely encouraging (Goddard and Hayes 2009; Hayes et al. 2009a; Luan et al. 2009; Hayes et al. 2009b; Hayes and Goddard 2010; Thomasen et al. 2014; Casellas and Piedrafita 2015; Porto-Neto et al. 2015; van Binsbergen et al. 2015; Forneris et al. 2016; Hayes et al. 2016a; Meuwissen et al. 2016; Hayes et al. 2016b), since these studies indicate that this approach is particularly interesting for traits of low

heritability and for organisms of long life cycle (Schaeffer 2006; Lee et al. 2008; Legarra et al. 2008; Luan et al. 2009; Thomasen et al. 2014; Hayes et al. 2016a; Hayes et al. 2016b).

Genomic selection is now a reality for animal breeding, with several studies showing the genetic gains achieved by early selection and advantages over conventional breeding (Hayes et al. 2009a; Habier et al. 2010; Hayes and Goddard 2010; Legarra et al. 2014; Su et al. 2014; Thomasen et al. 2014; Casellas and Piedrafita 2015; Porto-Neto et al. 2015; van Binsbergen et al. 2015; Meuwissen et al. 2016; Hayes et al. 2016b; VanRaden et al. 2017; Wallén et al. 2017). Empirical studies have proven the excellent prospect of GS application in the breeding of annual plants, such as: maize (Crossa et al. 2010; Crossa et al. 2013; Massman et al. 2013; Crossa et al. 2014; Liu et al. 2015; Pace et al. 2015; Acosta-Pech et al. 2017; Bernardo 2017), wheat (Crossa et al. 2010; Poland et al. 2012; Crossa et al. 2014; Bassi et al. 2015; Thavamanikumar et al. 2015; Fiedler et al. 2017; Jarquín et al. 2017; Juliana et al. 2017), barley (Shengqiang et al. 2009; Schmidt et al. 2015; Li et al. 2017), rice (Grenier et al. 2015; Spindel et al. 2015; Onogi et al. 2016; Spindel et al. 2016), pea (Burstin et al. 2015; Tayeh et al. 2015; Annicchiarico et al. 2017) and soybean (Zhang et al. 2015; Chang et al. 2016; Duhnen et al. 2017). In forest tree species, GS began to be approached through some simulation studies (Grattapaglia and Resende 2011; Iwata et al. 2011) and soon afterwards in two pioneering empirical studies in *Pinus* (Resende Jr et al. 2012a) and *Eucalyptus* (Resende et al. 2012). Subsequently, a simulation study to test the GS efficiency, including dominance effect in the model, was published in *Eucalyptus* (Denis and Bouvet 2013). Afterwards, several other papers were published in forest tree species of different genus of conifers, such as: *Pinus* (Zapata-Valenzuela et al. 2012; Resende Jr et al. 2012b; Zapata-Valenzuela et al. 2013; Isik et al. 2015; de Almeida Filho et al. 2016), *Larix* (Klápště et al. 2014) and *Picea* (Beaulieu et al. 2014a; Beaulieu et al. 2014b; Gamal El-Dien et al. 2015; Ratcliffe et al. 2015; El-Dien et al. 2016; Fuentes-Utrilla et al. 2017; Lenz et al. 2017). More recently, several studies of GS in *Eucalyptus* were published (Cappa

et al. 2017; Durán et al. 2017; Müller et al. 2017; Resende R.T. et al. 2017; Tan et al. 2017). GS could represent a radical paradigm shift in forest tree breeding by allowing the ultra-early selection of elite trees still in the nursery stage for late expression traits, such as growth, wood quality and tolerance to abiotic and biotic stresses (Grattapaglia 2014). Individuals can be selected for the installation of clonal tests or their use as parents for the next-generation of breeding or both, as has been done in several companies today. This approach seeks to explore the combination of favorable traits and to identify exceptional individuals that consolidate several desirable traits. Breeding programs with this configuration are fully adequate for the implementation of GS.

Training and validation populations, and accuracy of the model

For GS implementation, a training population, also known as discovery population or estimation population, is genotyped with thousands of markers and phenotyped for the traits of interest (Jannink et al. 2010). From these datasets, predictive models are developed to estimate the Genomic Estimated Breeding Values (GEBV) for each trait individually. These models associate for each marker their predicted effect on the target trait. Thus, in the training population, the markers associated with the loci that control the traits are discovered through genotyping, as well as their effects are estimated (Grattapaglia 2014). Through the cross-validation, the predictive models of the GEBVs are tested to verify their prediction accuracy in a subset of individuals sampled randomly from the training population. These individuals of random sampling make up the validation population, being a subset of individuals, usually 10%, of the training population who did not participate in estimating the effects of the marker.

The GEBVs are predicted using the estimated effects in the training population and subjected to correlation analysis with observed phenotypic values to obtain the prediction accuracy. As the validation population was not involved in the prediction accuracy of the marker effects, the errors associated with the GEBVs

and the phenotypic values are independent. Therefore, the correlation between these values is predominantly of a genetic nature, equivalent to predictive ability (r_{gy}) of the GS in estimating the phenotypes. In general, the predictive ability is defined as the correlation between observed (y) and the genomic-estimated breeding values ($GEBV$) computed through cross-validation $\big(r(y, GEBV)\big)$. The accuracy ($r_{g\hat{g}}$) could be obtained by dividing the predictive ability by the square root of the individual heritability (h) (Legarra et al. 2008). In broad-spectrum, the narrow-sense heritability (h^2) is calculated as the ratio of the additive variance (σ_a^2) to the phenotypic variance (σ_y^2). In other words, narrow-sense heritability ($h^2 = \sigma_a^2/\sigma_y^2$) captures only that proportion of genetic variation that is due to additive genetic values. Currently, to apply GS in forest tree species it is more common to report the predictive ability (Beaulieu et al. 2014a; Beaulieu et al. 2014b; Bartholomé et al. 2016b; Müller et al. 2017; Resende R.T. et al. 2017) rather than the accuracy, due to the potential bias in estimating genomic heritability as discussed in de los Campos et al. (2015). The GEBV is calculated by multiplying the number of alleles in each of the markers by their estimated effect through Ridge Regression Best Linear Unbiased Prediction (RR-BLUP) or other Bayesian statistical method (LASSO, BayesA, BayesB, etc.). All these statistical methods aim to mitigate the problem of a small *n* and a large *p*, that is, the estimation of a large number of effects *p* (number of markers) from a limited number of observations *n* (sample size). The shrinkage estimators in regression coefficients avoid this problem by treating the effects of the markers as random variables and estimating them simultaneously (Crossa et al. 2010; Lorenz et al. 2011).

Factors affecting the prediction accuracy of GS

Initially, to estimate the prediction accuracy of the GS models, the correlation between the GEBV and the genetic value estimated by the observed phenotype is evaluated. There are several parameters that can affect the predictive ability of the model and generally these are dependent on each other. The main parameters are: (i) the distribution of QTL effects (number of loci and size of effects controlling

the trait); (ii) the heritability of the assessed trait; (iii) the number of genotyped and phenotyped individuals that compose the training population used to estimate the effect of markers; (iv) the effective population size (N_e); (v) the density of markers; (vi) the extent of LD between markers and QTLs; and the (vii) relatedness between the individuals in the training and validation populations (Habier et al. 2007; Legarra et al. 2008; Hayes et al. 2009b; Grattapaglia and Resende 2011; Lenz et al. 2017). The QTLs effects (i) and heritability (ii) are parameters dependent on the genetic architecture of the trait of interest in a specific environment and also on the genetic background of the study population. On the other hand, the number of individuals in the training population (iii), the N_e (iv); the density of markers (v); the extension of LD (vi) and the relatedness (vii) can be controlled experimentally by the breeder (Hayes et al. 2009b; Grattapaglia and Resende 2011; Grattapaglia 2014; Beaulieu et al. 2014b; Wallén et al. 2017).

Genetic architecture of the target trait

The number of QTLs that control traits of interest has an impact on the GS prediction accuracy. If there are few loci controlling large fractions of phenotypic variation, they are easily captured compared to a more complex genetic architecture involving a greater number of loci (Daetwyler et al. 2010). As expected, the reduction in the GS prediction accuracy with an increasing number of QTLs involved in the trait tends to be more pronounced in low-density marker panels or with a higher N_e (Habier et al. 2009; Grattapaglia and Resende 2011). Therefore, to achieve satisfactory prediction accuracy in the GS model (e.g. greater than 0.60) it would be necessary to use density panels with ≥ 5 markers/cM assuming a simpler genetic architecture, while 20 markers/cM would be required with a greater number of QTLs controlling the trait (Grattapaglia and Resende 2011). Heritability is a factor underlying the genetic architecture of a trait, being proportional to the accuracy reached in GS, that is, the greater the heritability, the greater will be the prediction accuracy. The study of Grattapaglia and Resende (2011) showed by simulation that a considerable increase in heritability (0.2 to 0.6)

results in a small increase in prediction accuracy (0.71 to 0.83). Empirical studies however show that heritability and prediction accuracy have a strong correlation (Resende Jr et al. 2012b; Muranty et al. 2015). Traits of lower heritability have less informative phenotypes from the genetic point of view and are therefore expected to be less predictable through GS (Resende Jr et al. 2012b). On the other hand, simulation studies have shown that the reduction of prediction accuracy with the reduction of heritability can be easily compensated by using a greater number of individuals in the training population (Meuwissen et al. 2001; Nielsen et al. 2009).

Number of individuals in the training population

The selection of a large number of individuals in a training population to accurately estimate the effects of markers is generally not a limitation on forest tree species (Grattapaglia 2014). The selection of a training population depends on the breeding strategy to be adopted and also on the structure and number of populations involved. The training population can be established by sampling trees in preexisting progeny tests. Generally, these progeny tests are derived from interbreeding (with free or controlled pollination) of a set of few elite parents, representative of the desired genetic variation with adequate size of N_e, to provide genetic gains that will be sustained by some future generations (Beaulieu et al. 2014b; Bartholomé et al. 2016b). In the study by Grattapaglia and Resende (2011), it was demonstrated by simulation that by increasing the number of individuals (n) up to 1,000, the prediction accuracy of the GS model increased rapidly, reaching satisfactory levels depending on the used N_e. On the other hand, using 2,000 individuals in the training population showed a small improvement of 6-10% of the estimated accuracy in relation to n = 1,000. In this same study, adopting a number of individuals above 2,000, the prediction accuracy reached a plateau, regardless of the density of markers in the genotyping and the N_e. Consequently, the number of individuals generally used to compose a training population for forest tree species has been 1,000 to 1,500. However, if the distribution of QTL violates the infinitesimal model of equal effects with common variance, not all genetic variation

would be explained and the GS prediction accuracy could decrease depending on the method used to estimate GEBV (Coster et al. 2010). Therefore, using a training population around n = 2,000 would be justified to protect against such model violations or in cases where several hundred QTLs control the trait variation (Grattapaglia and Resende 2011). In addition, larger training populations may mitigate the likelihood of losing favorable rare alleles as selection generations advance, although inevitably some of these alleles will be lost because they may be at low LD with any marker. The use of higher density of markers on genotyping, also will help in this regard by preserving rare alleles in these breeding populations, allowing greater long-term gains in selection (Grattapaglia 2014; Bartholomé et al. 2016b).

Effective population size (N_e) and marker density

For the GS application in plant breeding some requirements are necessary. In addition to having large-scale and low-cost genotyping platforms, appropriate populations for GS application should have properties that result in longer extensions of LD. In the case of *Eucalyptus* this situation is fully satisfied in most populations that tend to have an effective size between 10 and 100 parents who through cross-breeding generates large progenies in which individual selection is practiced (Grattapaglia et al. 2009; Grattapaglia and Resende 2011). The N_e corresponds to the number of individuals in an idealized population that would generate offspring, presenting the same amount of allele frequency dispersion under genetic drift or the same amount of inbreeding as the actual population under consideration (Wright 1931). The higher the N_e, a larger number of markers will be needed covering the genome to capture the LD between markers and QTLs in order to achieve and maintain a high GS prediction accuracy. Grattapaglia and Resende (2011) evaluated the impact of increasing effective population size on predictive accuracy. For a reduced N_e (N_e = 10 to 30), it would take between 2 to 5 markers per cM (centiMorgan) to achieve adequate accuracy. In contrast, for a higher value of N_e (N_e = 100 to 200), a density in the order of 10 to 30 markers per

cM would be required. For a total recombination rate of the *Eucalyptus* genome from 1,100 to 1,500 cM (Brondani et al. 2006; Hudson et al. 2012), 22,000 to 30,000 informative markers for GS practice would therefore be required. Currently, reaching this number of markers is no longer a limitation for GS application in forest tree. Silva-Junior et al. (2015) developed an Infinium *Eucalyptus* chip with 60,000 SNPs (EUChip60K). It is clear that the number of polymorphic SNPs will vary according to the species diversity and the specific population under analysis. In the work of Silva-Junior et al. (2015), 75% of the 60,000 SNPs were polymorphic in 42 individuals of *E. camaldulensis*, the most widely distributed *Eucalyptus* species in Australia (Butcher et al. 2009). On the other hand, despite genotyping 558 individuals of *E. benthamii*, only 23.5% of the SNPs were polymorphic in this species, revealing its restricted genetic base and suggesting a high genetic vulnerability (Butcher et al. 2005).

The density of markers in the genotyping is an important factor in maintaining GS predictive capacity with the advancement of selection in the next generations (Bartholomé et al. 2016b). Higher marker densities allow accuracy to persist over time due to slower LD decay between markers and loci (Grattapaglia 2014). High density genotyping is essential to support the accuracy of the model considering that selection together with recombination may alter the pattern of LD between markers and QTLs in the next generations (Long et al. 2011). The reduction in GS prediction accuracy over time can be mitigated by re-estimating the effects of the markers (Iwata et al. 2011). Due to the fast LD decay of the genomes of forest tree species, attempts to use low-density panels as an option to reduce the costs of genotyping, as proposed for domestic animals and annual crops (Habier et al. 2009; Vazquez et al. 2010; Zhao et al. 2012; Weller et al. 2014; Liu et al. 2015), should be viewed with caution. A lower marker density would make GS more susceptible to LD decay with recombination and selection. In addition, simultaneous selection for several traits would likely result in different sets of markers more informative to each trait. In such scenario, a useful high-density panel of markers among various breeding populations and with the aim of selecting

for several traits simultaneously will possibly be the best option (Grattapaglia 2014).

Linkage disequilibrium (LD) extension

The extent of Linkage Disequilibrium (LD) is one of the most determinant parameters in the GS prediction accuracy. Linkage disequilibrium is a measure of non-random association between alleles of different loci (Slatkin 2008), for example between marker alleles and QTLs. As N_e becomes smaller, the effect of genetic drift becomes stronger and more LD is generated because combinations between marker alleles and QTL alleles are unlikely to be sampled at a frequency corresponding to the product of their individual frequencies. Therefore, the resulting non-random association between alleles of marker loci and QTLs alleles allows the marker to predict the allelic status of nearby QTLs, and thus predict the phenotypes controlled by them. In equilibrium, the LD generated by genetic drift is balanced by recombination that occurs with advancement in breeding population generations. This causes LD dissipation in such a way that for the closest loci from the recombination point of view it is expected to have higher LD than the more distant ones. As a consequence, the relationship between N_e and LD impacts the marker density required to achieve and maintain the predictive ability in the GS model over generations. Therefore, the marker density required for GS practice depends on the LD level between markers and QTLs, which in turn is a function of N_e. The breeder can control both the effective size of the population using a greater or lesser number of parents and the density of markers depending on the financial resources available for genotyping (Grattapaglia 2014).

Relatedness between individuals in the training and validation populations

Generally, the forest tree breeder controls the N_e in order to reduce it aiming to increase the LD between the markers and QTL, being an efficient method to increase the GS predictions accuracy (Grattapaglia and Resende 2011). On the

other hand, when N_e is reduced in addition to increasing LD, it also increases the relatedness between individuals in the population. Recently, several experimental studies have shown that the relatedness between the training population and the selection candidates is mainly responsible for the GS predictions accuracy (Auinger et al. 2016; Michel et al. 2016; Müller et al. 2017). And these studies have evaluated the effect of relatedness on predictive ability by removing relatedness between training and validation populations using different approaches, either by removing relatedness between individuals belonging to the same families (Legarra et al. 2008; Albrecht et al. 2011; Makowsky et al. 2011; de los Campos et al. 2013a; Riedelsheimer et al. 2013; de los Campos et al. 2013b; Beaulieu et al. 2014a; Spiliopoulou et al. 2015; Spindel et al. 2015; Spindel et al. 2016; Lenz et al. 2017; Resende R.T. et al. 2017), populations (Hayes et al. 2009a; Habier et al. 2010; Clark et al. 2012; Riedelsheimer et al. 2013; Cros et al. 2015) and subpopulations (Saatchi et al. 2011; Windhausen et al. 2012; Ly et al. 2013; Beaulieu et al. 2014b; Arruda et al. 2015; Spindel et al. 2015; Spindel et al. 2016; Müller et al. 2017). GS predictions decreased markedly when unrelated individuals were used in the training population and in the validation population. With the dissipation of LD with recombination, it is expected, therefore, that in subsequent generations a lower predictive ability will be observed. The question of relatedness should therefore be carefully considered in the prospects of using a prediction model. The individuals on whom the models will be applied are candidates for selection, but the prediction accuracy of their phenotypes cannot be estimated because their phenotypes are not available. The models are tested by cross-validation, typically using a subsample of the training population. Since relatedness is an important component of predictive accuracy, the most important principle of selecting and assembling a training population is that it adequately mirrors the relationship between future candidates for selection and the training population (Daetwyler et al. 2013). If the validation population is more or less related to the training population than the candidates for selection, then the predictive accuracy will be overestimated or underestimated, respectively. The results shown in studies with *Eucalyptus* (Resende et al. 2012; Müller et al. 2017; Resende R.T. et al. 2017) and with *Picea*

(Beaulieu et al. 2014a; Beaulieu et al. 2014b; Lenz et al. 2017) corroborate to this point, that GS depends on the existence of relatedness between training population and candidates for selection.

Perspectives of the GS application in forest tree breeding

The first empirical study with GS in *Eucalyptus* was a proof of concept, showing that this approach reached similar and even higher accuracy than those obtained by conventional phenotypic selection (Resende et al. 2012). In addition, the GS methodology was important to capture large fractions of the heritabilities (75-97%) of the evaluated growth and wood quality traits. Recent experimental data presented promising perspectives of the GS application to increase the efficiency of *Eucalyptus* breeding programs (Cappa et al. 2017; Durán et al. 2017; Müller et al. 2017; Resende R.T. et al. 2017; Tan et al. 2017). This would be accomplished, fundamentally, by shortening the duration of the breeding cycle, excluding the stage of progeny testing and implementing the ultra-early selection of the phenotypes yet to be observed in the seedling stage in the greenhouse. In *Eucalyptus*, GS could not only eliminate progeny testing but also reduce the time and costs involved in the clonal testing phase by reducing the number of selected trees that are tested as large-scale clones. Thus, GS compared to conventional breeding can reduce the cycle from 18-10 to 9-5 years (Resende et al. 2012; Grattapaglia 2014). Currently, GS is a subject of great relevance in plant breeding (Morrell et al. 2011; Hickey et al. 2017) and forest tree species breeding programs have been conducting several experiments aimed at the operational implementation. How to incorporate GS into a forest genetic improvement program will vary from case to case after a detailed cost-benefit analysis. The gain in time reduction by replacing the progeny tests by GS clonal test will be unavoidable (Durán et al. 2017). In addition, the genetic gain achieved by allowing the simultaneous evaluation of all target traits in all progeny individuals at one time will be an additional great operational advantage of GS in forest trees (Grattapaglia 2014). Recently, genomic prediction models were evaluated across generations in

maritime pine (*Pinus pinaster*) (Isik et al. 2015; Bartholomé et al. 2016b), demonstrating even more encouraging perspectives of this novel approach to accelerate forest tree breeding programs.

Contribution to the field

This thesis makes contributions to the advancement in the understanding of the potential application of GS and GWAS for complex growth traits in breeding populations of *Eucalyptus*. The genomic prediction study described in chapter one provides experimental data supporting positive prospects of using genome-wide data to capture large proportions of trait heritability and predict growth traits, in species of *Eucalyptus* not contemplated before, with accuracies equal or better than those attainable by phenotypic selection. The study described in chapter two goes a step further in GWAS experiments, applying for the first time different approaches of Joint-GWAS in forest trees by assembling data for 3,373 individuals across four unrelated *Eucalyptus* breeding populations. Although large proportions of the heritability were explained by the genome-wide data in all populations, few associations were found, corroborating the high polygenic nature of growth traits. Still, interesting putative associations were detected in a range of candidate genes involved in cell wall biosynthesis processes and disease resistance. The studies described below suggest that it will be necessary to considerably increase the sample size by orders of magnitude to achieve sufficient power to detect a larger part of the variants segregating in the target *Eucalyptus* breeding populations. On the other hand, the results of genomic prediction further support that whole-genome regression should prove a useful approach to incorporate genomics into tree breeding.

CHAPTER 1: Genomic prediction and GWAS for growth traits in breeding populations of *Eucalyptus benthamii* and *E. pellita*

Published as:

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(1):524. doi: 10.1186/s12864-017-3920-2

INTRODUCTION

Species of *Eucalyptus* are the most planted hardwood trees worldwide due to their multipurpose applications (e.g. pulp, paper, solid wood and bioenergy), superior growth, high adaptability and wood quality (Myburg et al. 2007). Amongst the 800 catalogued species of *Eucalyptus* L'Hér. (Myrtaceae), the "big nine" species within subgenus *Symphyomyrtus* account for over 95% of the world's eucalypt plantations (Harwood 2011). Within this group, *Eucalyptus grandis* Hill ex Maiden*, E. urophylla* S.T. Blake, and *E. camaldulensis* Dehnh are the most economically prominent ones in tropical regions, whereas *E. globulus* Labill and *E. nitens* H.Deane & Maiden are notable in temperate regions (Myburg et al. 2007). The extensive intra- and interspecific diversity and sexual compatibility across species of *Symphyomyrtus* has been a major advantage to breeders, as it allows rapid blending of gene pools that evolved separately under contrasting environmental pressures (Grattapaglia and Kirst 2008). Nevertheless, there is still ample opportunities for expanding the use of some secondary species of *Symphyomyrtus* not included among the "big nine", to develop uniquely adapted genetic material that combine rapid growth, good wood quality and adaptation to environmental stresses such as frost, heat and drought.

Eucalyptus benthamii Maiden & Cambage (Camden white gum), a species of restricted occurrence in its natural range in Australia (Butcher et al. 2005), has showed great potential to expand eucalypt commercial plantations into subtropical regions subject to periodic frosts (Arnold et al. 2015). *Eucalyptus benthamii* planted as pure species or in hybrid combinations has received increasing attention in subtropical regions of southern Brazil and southeastern USA (Pirraglia et al. 2012; Costa et al. 2016). Another species of marginal importance until recently, *Eucalyptus pellita* F. Mueller (large-fruited red mahogany), is highly suitable for growth in year-round humid lowland equatorial climates under high temperatures, showing a particularly high resistance to pathogens. *Eucalyptus pellita* is endemic to tropical regions in two disjoint natural forests, in southern New Guinea and in northern Australia (Harwood et al. 1997). It has shown fast growth in hybrid combination with *E. grandis* providing resistance to a number of fungal diseases (Agustini et al. 2014).

Genomic selection (GS) was proposed by Meuwissen et al. 2001, and has gained increasing interest among forest tree breeders. This predictive methodology provides an alternative approach to using marker-assisted selection (MAS) that relies on previously detected discrete quantitative trait loci (QTL) in bi-parental mapping and association genetics experiments. In forest trees, genomic prediction began to be addressed by simulation studies (Grattapaglia and Resende 2011; Iwata et al. 2011) followed by experimental reports in *Pinus* (Resende Jr et al. 2012a) and *Eucalyptus* (Resende et al. 2012) demonstrating the positive prospects of this breeding method. Since then, a number of experimental genomic prediction studies have confirmed the potential of GS in conifer species, including *Pinus* (Resende Jr et al. 2012b; Zapata-Valenzuela et al. 2013; de Almeida Filho et al. 2016) and *Picea* (Beaulieu et al. 2014a; Beaulieu et al. 2014b; Gamal El-Dien et al. 2015; Ratcliffe et al. 2015). Recently, genomic prediction models were evaluated across generations in maritime pine (*Pinus pinaster*) (Isik et al. 2015; Bartholomé et al. 2016b), demonstrating even more encouraging perspectives of this novel approach to accelerate breeding of forest trees.

Several parameters were shown to affect GS prediction accuracy in simulation studies, such as the number of QTLs controlling the trait, trait heritability, the size of the training population, number of markers and the effective population size (N_e) of the target population (Grattapaglia and Resende 2011). If an adequate density of markers is provided for a given N_e, it is expected that most QTL will be in LD with at least one marker and will be captured in predictive models. Consequently, high-throughput and low-cost genotyping platforms constitute an essential tool to apply GS. The reduction of the effective population size leads to increased relatedness between individuals and more extensive LD in the population. Markers fitted in a GS model will capture not only LD but also relatedness between individuals in the training and validation sets. An increase in prediction ability with enhanced relatedness among the training and validation sets was shown early on from simulation studies (Habier et al. 2007), and underscored in all recent reviews on the perspectives GS in plant and domestic animals breeding (Van Eenennaam et al. 2014; Heslot et al. 2015). Phenotypes of individuals closely related to the training population will be better predicted over distantly related individuals.

In this study, we report the development of genomic prediction models for growth traits in two breeding populations of *E. benthamii* (n = 505) and *E. pellita* (n = 732) using SNP data generated with the multi-species *Eucalyptus* 60kSNP chip. Using a genomic relationship matrix (GRM) we compared the pedigree and genome-estimated breeding values and narrow-sense heritabilities in the two populations. Different Bayesian methods for predicting growth traits were compared. The impact of variable numbers of SNPs, different SNP sampling methods based on their position in the genome, and the impact of relatedness on genomic prediction were also evaluated. Finally, a genome-wide association analysis was carried out on the same datasets to evaluate what would be the ability to capture heritability and detect discrete associations for complex growth traits in operational breeding populations under selection.

MATERIAL AND METHODS

Populations and phenotypic data

This study was carried out on progeny trials of populations of *E. benthamii* and *E. pellita* that are part of the breeding program of EMBRAPA (Brazilian Agricultural Research Corporation). The *E. benthamii* progeny trial was composed of 40 seed sources, being 36 open-pollinated (OP) half-sib families from wild Australian populations and four bulked seed sources (two from Australian populations, one from a first generation breeding population established in Colombo, PR, Brazil and one from a second-generation breeding population planted in Candói, PR, Brazil). The complete *E. benthamii* trial involved 2000 trees planted in May 2007 in Candói, in a randomized complete block design with 50 blocks in single-tree plots (one progeny individual per block for each one of the 40 seed sources). The experiment was thinned three times (removing 600 trees in March 2009, 700 in March 2010 and approximately 200 in December 2010) to eliminate trees with poor growth, malformed stems and damaged plants. The population underwent 25 heavy frosts recorded (temperature varying from −3.4 to −12.6 °C) in 58 months, between planting (May 2007) and field evaluation (February 2012) that killed or affected the growth of many trees which were therefore culled. For *E. benthamii* 508 trees were ultimately phenotyped at age 56 months for the following growth traits: Diameter at Breast Height (DBH, cm), Total Height (HT, m) and Wood Volume (WV, m3) (Table 1-1). The *E. pellita* breeding trial was composed of 24 OP maternal families derived from a second-generation clonal seed orchard located in Mareeba, Queensland, Australia, established with selections from four provenances in the areas of Kiriwo, Serisa and Keru in the Morehead district of the Western Province of Papua New Guinea. The experimental design was a randomized complete block design with 24 families and 40 blocks in single-tree plots (960 trees total) planted in February 2010 in Rio Verde, GO, Brazil. For *E. pellita* phenotypic evaluations were made at age 42 months (September 2013) for DBH, HT and WV (Table 1-1).

Genotyping and filtering

A total of 552 *E. benthamii* trees and 771 *E. pellita* trees were genotyped using the *Eucalyptus* Illumina Infinium EUChip60K (Silva-Junior et al. 2015). The genotypic data were filtered to remove SNPs with call rate (CR) ≤ 90% and monomorphic SNPs, therefore keeping all SNPs with Minimum Allele Frequency (MAF) > 0 in the analysis. Because trees were genotyped before the final field measurements, some genotyped trees died, so that ultimately 505 individuals of *E. benthamii* and 732 of *E. pellita* had full genotypic and phenotypic data for further analyses. An alternative SNP dataset was also generated by keeping only SNPs MAF ≥0.05. With the objective of evaluating the effect of LD-pruning on predictions, polymorphic SNPs (CR ≥ 90% and MAF> 0 or MAF≥ 0.05) were pruned based on pairwise linkage disequilibrium (LD) estimates using PLINK v1.9 (Purcell et al. 2007), to generate a pruned subset of SNPs that are in approximate linkage equilibrium (LE). The LD based SNP pruning method was applied with a window size of 100 Kbp, shifting the window by one SNP at the end of each step and removing one SNP from a pair of SNPs if LD was greater than 0.2 (plink command: --indep-pairwise 100 kb 1 0.2).

Effective population size estimation, population structure and LD analyses

Effective population size (N_e) was estimated based on the linkage disequilibrium (LDN_e) method implemented in NeEstimator v2.01 (Do et al. 2014) or each species. A random mating model and MAF ≤ 0.05 was used for excluding rare alleles in LDN_e. Confidence intervals for these estimates were obtained using the parametric method in NeEstimator, where the number of independent alleles is used as the degree of freedom in a chi-square distribution. The genetic structure for both eucalypt populations estimated based on a Bayesian clustering method was determined with STRUCTURE v2.2.4 (Pritchard et al. 2000) using only the LD-pruned SNPs set. The individual structures were classified in *K* clusters according to genetic similarity. 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, and each *K* was replicated 10 times. The burn-in period and the number of Markov Chain Monte Carlo (MCMC) replications were 100,000 and 200,000, respectively. The number of genetic groups was determined based on the criteria proposed by Evanno et al. (2005) using the program STRUCTURE HARVESTER v0.6.93 (Earl and vonHoldt 2012). The software CLUMPP v1.1.2 (Jakobsson and Rosenberg 2007) was used to find consensus among the 10 most probable *K* interactions. Principal component analysis (PCA) was performed using SNPRelate R package (Zheng et al. 2012), with only the LD-pruned SNPs set. Analyses of linkage disequilibrium were performed using LDcorSV (Mangin et al. 2012). Pairwise estimates of LD were calculated by the classical measure of the squared correlation of allele frequencies at diallelic loci (r^2), as well as correcting for bias due to relatedness and population structure (r^2VS), and adjusting it independently for relatedness (r^2V) and for population structure (r^2S). To estimate the adjusted LD, the genomic relationship matrix (GRM) was computed using the Powell method (Powell et al. 2010) implemented in R. The population structure results were based on the most probable value of *K* (*K* = *2*). The drift-recombination model (Hill and Weir 1988) was used for nonlinear regression to fit the expectation of r^2, using the R script by Marroni et al. (2011) and the following equation based on Remington et al. (2001):

$$E(r^2) = \left[\frac{10 + C^1}{(2 + C)(11 + C)}\right]\left[1 + \frac{(3 + C)(12 + 12C + C^2)}{n(2 + C)(11 + C)}\right] \qquad (1)$$

where n is the sample size, and C is the population recombination parameter ($C = 4N_e c$; N_e is the effective population size and c is the recombination fraction between the pairwise SNPs). To visualize patterns of LD decay in the two eucalypts species, all the LD estimates (r^2, r^2V, r^2S, r^2VS) were plotted up to a 100 Kbp distance.

Genomic and pedigree-based breeding value predictions

Prediction of breeding values by best linear unbiased prediction (BLUP) (Henderson 1975) based on pedigree information (ABLUP) was calculated using the expected genetic relationship between individuals. For the genomic estimated breeding values the individual SNPs had their effects estimated by adjusting all the allelic effects simultaneously using Genomic BLUP (GBLUP) frequentist (VanRaden 2008). A 10-fold cross-validation approach was used, defined as a random subsampling partitioning of the data for each trait into two subsets. The first subset with 90% of the individuals was used as a training population to estimate the marker effects. The second subset with the remaining 10% was used as validation population, and had their phenotypes predicted based on the marker effects estimated in the training population. This process was repeated 10 times, randomly selecting in each fold a different set of samples as the validation population, until all individuals had their phenotypes predicted and validated. Analyses of each trait were carried out using the package rrBLUP (Endelman 2011) with the following mixed linear model:

$$y = Xb + Za + e \qquad (2)$$

where y is the phenotypic measure of the trait being analyzed; X and Z are incidence matrices for the vectors for parameters b and a, respectively; b is a vector of fixed block effects; *a* is a vector of random additive effects and e is the random residual effect. The variance structure of the model for pedigree-estimated breeding values or simply estimated breeding values (EBVs) was calculated with $a \sim N(0, A\sigma_a^2)$ and the genomic-estimated breeding values (GEBVs) with $a \sim N(0, G\sigma_a^2)$; where A is a matrix of additive genetic relationships among individuals and G is a GRM estimated using the method proposed by VanRaden (2008). The predictive ability (r_{gy}) was estimated as the correlation between the observed and the genomic-estimated breeding values $\big(r(y, GEBV)\big)$. The narrow-

sense heritability (h^2) was calculated as the ratio of the additive variance σ_a^2 to the phenotypic variance σ_y^2 ($h^2 = \sigma_a^2/\sigma_y^2$).

Bayesian methods

The SNP effects were estimated using six different Bayesian genome-wide regression models, being five additive genetic models: Bayesian Ridge-Regression (BRR), Bayes A, Bayes B, Bayes Cπ and Bayesian Lasso (BL); and one additive and non-additive genetic model: Bayesian Reproducing Kernel Hilbert Spaces Regressions (RKHS). These models were implemented in the BGLR package (Pérez and de los Campos 2014). For these methods, the genotypic information was fitted using the following base model:

$$y = Xb + g + e \qquad (3)$$

where y is the vector of observations representing the trait of interest; b is a vector with intercept and fixed block effects; g is a vector of random individual genetic merit (that change with the fitted model); X is incidence matrix for the vector for parameter b; e is a vector of the random error effects. In all Bayesian models, it was assumed that:

$$y|b, g, \sigma_e^2 \sim N(Xb + g, I\sigma_e^2)$$
$$b \sim N(0, 10^6 I)$$
$$e|\sigma_e^2 \sim N(0, I\sigma_e^2)$$
$$\sigma_e^2|S_e, \nu_e \sim \chi^{-2}(\nu_e, S_e)$$

The models BRR, Bayes A, Bayes B, Bayes Cπ and BL, predict the genetic merit (g) in the equation 3 as an average effect of the nucleotide substitution, thus these models consider:

$$g = Zm$$

m is a vector of random marker effects ($m = [m_1 \quad \cdots \quad m_k]^T$) and Z is the incidence matrix for the vector m. This matrix takes values 2, 1 or 0 if the genotype of the *i*th marker is AA, Aa and aa, respectively, where a is the least frequent allele. Missing genotypes were replaced by the mean of the genotype for the given SNP. The assumptions of the m vector depend on the prior adopted and the respective priors used in the linear regression coefficients for each model are described below.

The *Bayesian Ridge-Regression* (BRR) is the Bayesian version of RR-BLUP (Meuwissen et al. 2001) and assumes that all marker effects have the same variance component. Consequently, markers with the same allele frequency contribute equally to the genetic variance. For the BRR model it is assumed that:

$$m_i|\sigma_m^2 \sim N(0, \sigma_m^2)$$
$$\sigma_m^2|S_m, \nu_m \sim \chi^{-2}(S_m, \nu_m)$$

The *Bayes A* method was proposed by Meuwissen et al. (2001) and later modified by Pérez and de los Campos (2014) to reduce the influence of the hyperparameters and achieve better Bayesian learning. The Bayes A model is opposite to BRR, in that it assumes that all marker effects have heterogeneous variances. In the Bayes A model, it is assumed that:

$$m_i|\sigma_{m_i}^2 \sim N(0, \sigma_{m_i}^2)$$
$$\sigma_{m_i}^2|S_m, \nu_m \sim \chi^{-2}(S_m, \nu_m)$$
$$S_m|r, s \sim G(r, s)$$

The *Bayes B* method was also proposed by Meuwissen et al. (2001) and modified in Pérez and de los Campos (2014) to achieve better Bayesian learning and to estimate the proportion of markers with null effect. This model is similar to Bayes A, and assumes that the markers have heterogeneous variance component.

Additionally, it considers that a proportion of markers have non-null effects. This is in contrast to Bayes A, because the approach includes the selection of covariates (SNPs markers) that do not contribute to genetic variance. In the Bayes B model it is assumed that:

$$m_i|\sigma^2_{m_i}\begin{cases} i \sim N(0,\sigma^2_{m_i}) & \text{with probability equal 1-}\pi \\ i & \\ i = 0 & \text{with probability equal } \pi \end{cases}$$

$$\sigma^2_{m_i}|S_m,\nu_m \sim \chi^{-2}(S_m,\nu_m)$$
$$S_m|r,s \sim G(r,s)$$
$$\pi \sim Beta(p_0,\pi_0)$$

The *Bayes Cπ* method proposed by Habier et al. (2011) is derived from the Bayes C method and is similar to BRR. In this approach, it is assumed that the marker effects have a common variance. However, Bayes Cπ includes marker selection with parameter π, which is defined as the probability of a SNP marker having a null effect. For the Bayes Cπ method it is assumed that:

$$m_i|\sigma^2_m\begin{cases} \sim N(0,\sigma^2_m) & \text{with probability equal 1-}\pi \\ = 0 & \text{with probability equal } \pi \end{cases}$$

$$\sigma^2_m|S_m,\nu_m \sim \chi^{-2}(S_m,\nu_m)$$
$$\pi \sim Beta(p_0,\pi_0)$$

The *Bayesian Lasso* (BL) method was proposed by Park and Casella (2008) and was adapted for genomic prediction by de los Campos et al. (2009). Similar to Bayes A and Bayes B, the BL method assumes covariates with heterogeneous variance. The BL method does indirect marker selection, since the marginal distribution of the markers follows a double exponential distribution, providing strong shrinkage of the marker effects to close to zero for large number of markers. In the BL method, it is assumed that:

$$m_i|\sigma_e^2, \tau_i^2 \sim N(0, \sigma_e^2 \times \tau_i^2)$$
$$\tau_i^2|\lambda \sim Exp(0.5\lambda^2)$$
$$\lambda|r, s \sim G(r, s)$$

The Bayesian Semiparametric *Reproducing Kernel Hilbert Spaces Regressions* (RKHS) applied in genomic prediction was proposed by Gianola et al. 2006 and modified in Gianola and de los Campos 2008. This method inputs markers in the relationship matrix (K) and predicts individual genetic merit directly, without prediction of markers effects. The relationship in RKHS is controlled by a constant called "bandwidth" (φ). With small positive bandwidth values, the relationship between two individuals tend to 1 and this relationship tend to 0 with large positives bandwidth values. To better control this constant bandwidth, the RKHS used here considered three kernels (González-Camacho et al. 2012), following the kernel averaging approach proposed by de los Campos et al. 2010:

$$g = \sum_r^3 g_r;$$
$$g_r|K_r\sigma_{g_r}^2 \sim N(0, K_r\sigma_{g_r}^2);$$
$$\sigma_{g_r}^2|\nu, S_g \sim \chi^{-2}(\nu, S_g);$$
$$K_r = \exp(-\varphi_r D^2)$$

The matrix D^2 is a squared Euclidean distance computed from SNP covariates (Z matrix), and bandwidth values are 0.2/q, 1/q and 5/q respectively, where *q* is 5th percentile of D^2 leading to global, intermediate and local kernels, respectively (González-Camacho et al. 2012; Tusell et al. 2014).

To estimate the parameters of the models, a total 200,000 iterations of MCMC were used with a burn-in period of 50,000 cycles and every fifth sample was kept. For all these models, a 10-fold cross-validation approach was applied as described previously.

Genomic predictions using selected SNPs subsets

The Bayesian Ridge-Regression (BRR) model was fitted using different subsets of SNPs of various sizes and selected using different criteria as described below. Initially a random sampling of SNPs stratified by chromosome was tested using (i) a cumulative approach, such that from the smallest subset of SNPs tested, additional ones were added to the previous set and (ii) a non-cumulative fashion, where different final sets of SNPs were randomly selected from all available SNPs. Next, variable positions of SNPs were tested, including: (iii) evenly spaced SNPs across the genome; (iv) only SNPs within gene models annotated in the *Eucalyptus* reference genome (Myburg et al. 2014); (v) SNPs based on LD-pruning and (vi) SNPs from individual chromosomes. For each subset, we estimated the predictive ability and genomic heritability. First, we evaluated models using different SNP subsets (from all 13,787 and 19,506 SNPs available for *E. benthamii* and *E. pellita* respectively, down to 2000 in smaller increments of 1000 SNPs, 1500, 1250, 1000, 750, 500, 300, 250, 200, 150 and 100 SNPs) with either a cumulative (i) or non-cumulative (ii) sampling of SNPs. For each number of SNPs and sampling strategy, ten replicates were performed. The evenly spaced SNPs subsets (iii) were created using different target windows sizes, with 1 SNPs every 10, 50, 100, 250, 500 Kbp and 1 Mbp, resulting in variable average distances between SNPs (Table SM1-1). For the within-gene SNP subset (iv), all SNPs located within annotated gene models (genic regions) and SNPs located outside of annotated gene models (intergenic regions) in the *Eucalyptus* genome were evaluated. To create the subsets of SNPs selected based on LD pruning (v), SNPs in approximate LE ($r^2 \leq 0.2$) with each other were chosen using PLINK v1.9 (Purcell et al. 2007). Finally, in the chromosome-specific SNP subsets (vi) the prediction models were fitted independently using only SNPs on each chromosome separately.

Genomic prediction controlling for relatedness between training and validation sets

To assess the relative impact of relatedness versus historical LD on the predictive ability, BRR prediction models were fitted minimizing relatedness between training and validation populations. Individuals were split into training and validation sets based on a Principal Component Analysis (PCA) or STRUCTURE analysis (*K* = *2*). In *E. benthamii*, 21 outlier individuals were removed and the remaining individuals were split into two subpopulations based on maximum genetic distance, one with 310 trees and the other with 174. For *E. pellita*, 26 outliers were excluded and the remaining 706 individuals were split into two subpopulations with 192 and 514 trees. As a control, a 10-fold cross-validation in each direction, with the same numbers of individuals used in the split populations, was carried out by random allocation of the individuals to training and validation sets.

Genome-wide association analysis

A mixed linear model association (MLMA) analysis was performed using the GCTA software (Yang et al. 2011a). This association analysis was fitted using the following base model:

$$y = Xb + g + e \qquad (4)$$

where y is the phenotype; b is a vector of fixed effects including intercept, block, population structure and SNPs to be tested for association; X is the incidence matrix for the vectors for the parameters b; g is the polygenic effect (random effect) captured by the GRM calculated using all SNPs and e is the random residual effect. The covariate computed for population structure was based on the fact that the population had two subpopulations (*K* = 2). The variance structure of the MLMA model were $g \sim N(0, G\sigma_g^2)$; $e \sim N(0, I\sigma_e^2)$; $cov(g, e') = 0$, where G is the GRM between individuals calculated as described earlier (Yang et al. 2010) and I is the

identity matrix. For comparisons with the MLMA model, we also performed a linear model based association (LMA) analysis fitting each SNP independently. This single-SNP association analysis was carried out using PLINK (Purcell et al. 2007) with a similar model as MLMA described in the equation 4, except for the exclusion of the polygenic effect (g). The Bonferroni procedure was implemented to control for type I error at α = 0.05 and the Benjamini and Hochberg (1995) procedure was used to control for false discovery at a rate FDR = 5%. The quantile-quantile (Q-Q) and Manhattan plots were generated using the qqman R package (Turner 2014).

RESULTS

SNP genotyping

Of the 60,904 SNPs in the EUChip60K, 50,303 (82.6%) and 49,518 (81.3%) were genotyped for *E. benthamii* and *E. pellita* respectively (Fig. SM1-1A), by using the phylogenetically appropriate SNP clustering file for SNP calling (Silva-Junior et al. 2015) , and filtering for SNPs with CR ≥ 90%. After selecting polymorphic SNPs (MAF> 0) 13,787 and 19,506 SNPs were retained for further analyses with a final rate of missing data of 1.4% and 0.8% for *E. benthamii* and *E. pellita*, respectively. An alternative SNP dataset was also used by filtering out SNPs with MAF ≤ 0.05 to investigate whether removing lower frequency SNPs had an impact on genomic predictions. A total of 7,563 SNPs for *E. benthamii* and 12,483 SNPs for *E. pellita* were retained for this alternative set.

Linkage disequilibrium and estimated effective population sizes

Linkage disequilibrium (r^2) was calculated for all pairwise physical distances among all the polymorphic SNPs (MAF > 0) on each chromosome separately. The average, genome-wide LD for pair of SNPs within a 100 Kbp distance from each other was 0.141 and 0.271 for *E. benthamii* and *E. pellita*, respectively. When

correcting the LD for bias due to relatedness and population structure (r^2VS), the average estimates were reduced to 0.096 and 0.178 (Table SM1-2). The genome-wide LD decayed to an r^2 below 0.2 within 15.6 Kb and 70.6 Kb (red line), while r^2VS showed a slightly faster decay within 7.7 and 25.5 Kb (pink dots) for *E. benthamii* and *E. pellita*, respectively (Fig. 1-1A and 1-1C). Linkage disequilibrium decayed to <0.2 for r^2S within 14.7 and 66.2 Kb (green line), while r^2V showed a slightly faster decay within 7.7 and 25.6 Kb (blue line), very similar to r^2VS for *E. benthamii* and *E. pellita*, respectively (Fig. 1-1A and 1-1C, Table SM1-2). The faster LD decay for r^2V or r^2VS confirms the strong effect of genetic relationship in these breeding populations. Slightly different patterns of LD decay were observed when including the SNPs with MAF<0.05 (Fig. 1-1A and 1-1C, MAF > 0) or excluding those (Fig. 1-1B and 1-1D). Datasets without the SNPs with MAF ≥ 0.05 showed a slightly higher pairwise r^2, with LD corrected (r^2VS) decaying to $r^2 = 0.2$ at 14.5 Kb in *E. benthamii* and 35.8 Kb in *E. pellita* (Fig. 1-1B and 1-1D, Table SM1-2). Estimated effective populations sizes based on LD data were $N_e = 50$ and $N_e = 35$ for *E. benthamii* and *E. pellita*, respectively (Table 1-1).

Genomic and pedigree-estimated heritabilities

For *E. benthamii* the pedigree-based narrow-sense heritabilities (h^2) estimated for DBH and WV were 0.326 and 0.297, and considerably lower for HT (0.088). Estimates of genomic heritabilities varied depending on the method used, with GBLUP and BL yielding considerably lower heritabilities than the pedigree-based ones and those obtained using other Bayesian methods (Table 1-2). When using Bayes B and BRR, heritabilities were higher (0.155 and 0.190). Estimates of variance components are reported in Table SM1-3. In *E. benthamii*, the variance components had similar estimates with all methods used. The pedigree-based narrow-sense heritabilities estimated for *E. pellita* were zero for DBH and WV, and nearly zero for HT (0.019), while the genomic estimated heritabilities based on SNP data were considerably higher (e.g. 0.414 -0.527 for DBH using the different methods) (Table 1-2). This unexpected result strongly suggests that the informed

pedigrees for the *E. pellita* population do not match the true relationships that the SNP data correctly recovered. Differently from *E. benthamii*, for *E. pellita* the genomic heritabilities had similar estimates for all methods used. Average heritabilities for *E. pellita* considering all genomic methods (~0.47 for DBH; ~0.29 for HT; ~0.44 for WV) were higher for all traits, compared to those estimated for *E. benthamii* (~0.23 for DBH; ~0.09 for HT; ~0.20 for WV). Heritabilities estimated including or not lower frequency SNPs (MAF ≤ 0.05) in the genomic relationship matrix were equivalent for both species, varying within the standard error of the estimates (Table 1-2). Genomic heritabilites captured large proportions of the pedigree-based heritability in *E. benthamii*. The Bayesian methods on average captured 73% and 69% of the pedigree-heritability for DBH and WV, respectively. No assessment was possible for *E. pellita* due to the inconsistency of the pedigree data that provided no valid estimate of pedigree-based heritability.

Genomic predictions

Consistent with expectations, predictive abilities (r_{gy}) followed the same trend as the estimated genomic heritabilities (Table 1-2). Predictive abilities inferred by an additive model using Bayesian methods (BRR, BayesB, BayesA, BayesCπ, BL) produced very similar estimates to those obtained using GBLUP and pedigree-based (Fig. 1-2). For the *E. benthamii* population both pedigree and genomic predictive abilities were generally low, averaging 0.16 for DBH, 0.14 for WV and close to zero for HT across all methods. For *E. pellita*, genomic predictive abilities were considerably higher, averaging 0.44 for DBH, 0.34 for HT and 0.42 for WV, suggesting the presence of a larger amount of additive genetic variation for these traits in this breeding population (Table 1-2). No difference was observed in the predictive abilities when using SNP sets including or not lower frequency SNPs. An additive and non-additive model (RKHS, machine learning method) improved predictions by 9-18% for *E. pellita*, but no improvement was observed for *E. benthamii* (Fig. 1-2). During cross-validation of genomic predictions a considerable variation was observed in the predictive abilities estimated across the different

folds (Table SM1-4). This variation was larger for *E. benthamii*, where the predictive ability across folds ranged from a low -0.058 to 0.415 using BRR for DBH, with an average of 0.162 with a standard error (SE) of ±0.044. In *E. pellita*, the variation was smaller, with estimates ranging from 0.358 to 0.550 for DBH, with the ten-fold average equal to 0.441 ± 0.019 (Table SM1-4).

Impact of variable numbers of SNPs on genomic predictions

Based on results of the different prediction methods, we chose to use only BRR to evaluate the impact of different SNPs sampling schemes on the predictive abilities. Subsets with progressively increasing randomly selected numbers of SNPs stratified by chromosome were used to estimate genomic predictions. Estimates of predictive ability and heritability increased rapidly up to ~3,000 SNPs for all traits and in both populations, (Table 1-3, Fig. 1-3). Predictive abilities plateaued at approximately 5,000 SNPs although heritabilities and predictive abilities still increased by 5 to 10% after that. Additionally, when less than 5,000 SNPs were used, a much larger variation in predictive ability was seen across the validation folds. These results indicate that at least in these populations for cross-validation within the same generation, models with ~5,000 to 10,000 SNPs will provide predictive abilities equivalent to those obtainable by using all the available SNPs. The non-cumulative sampling approach yielded essentially the same results with a plateau at ~5,000 SNPs, but showed a more spiky pattern of increasing predictive ability as more SNPs were fit into the model (Fig. SM1-2).

Impact of variable position-based SNP sampling methods

Overall, no difference was seen in the estimates of heritabilities and predictive abilities when different position-based SNP sampling schemes were used, as long as the total number of SNPs was close to 5,000 (Table 1-3, Fig. 1-3). The predictive abilities estimated with a subset of SNPs evenly spaced every 1 Mbp windows (610 SNPs in *E. benthamii* and 609 SNPs in *E. pellita*), were slightly higher than

those using 500 randomly sampled SNPs (Table 1-3). Although these results indicate that the number, and not the position of SNPs, determines the accuracy of predictions, they also suggest that even distribution might provide a small-added advantage when compared to random sampling. No significant differences in predictions were seen for any trait in both species when SNPs located in genic versus intergenic regions were used, and the predictions were equivalent to those obtained by random sampling of equivalent numbers of SNPs. The same result was observed with the LD-pruning approach, where estimates of predictive ability were similar either using LD-pruned SNPs in LE or all polymorphic SNPs (Table 1-3). There was no difference observed in the estimates of variance components when different sets of SNPs sampled based on position in the genome were used (Table SM1-3).

When only SNPs located on single chromosomes were used, heritabilities dropped on average by 30-45% when compared to using all SNPs (e.g. for WV from 0.243 to 0.177 in *E. benthamii*; from 0.418 to 0.244 in *E. pellita*), indicating that genome-wide marker coverage is critical for capturing the additive genetic variance (Table 1-4). The predictive abilities using SNPs on single chromosomes were similar across chromosomes and also dropped on average by 15-30% when compared to using all SNPs (Table 1-4). However, when the heritabilities and predictive abilities provided by single chromosomes were compared to those obtained using equivalent numbers of randomly sampled SNPs subsets, no appreciable differences were seen. This result indicates that the drop in predictive ability is most likely due to the small number of SNPs per chromosome (average of 1,253 for *E. benthamii* and 1,773 for *E. pellita*) and not to the fact that they are located on a single chromosome. We did not have sufficient numbers of SNPs on a single chromosome to compare to the larger random subsets of 3,000 or 5,000 to see the effect on predictions.

Impact of relatedness between training and validation sets

To assess the relative contribution of relatedness to the predictive ability (as opposed to historical LD between SNPs and QTL), GS models were fitted trying to minimize relatedness between training and validation sets based on genetic differentiation determined by a PCA (Fig. SM1-3). Predictive ability obtained when minimizing relatedness was null for *E. benthamii* (Fig. 1-4A) (e.g. from 0.108 to -0.032 for DBH) and reduced approximately by half for *E. pellita* (e.g. from 0.348 to 0.154 for DBH) compared to the predictive abilities achieved when the same number of individuals were used to build the model without controlling for relatedness (Fig. 1-4B). These results suggest that predictions in the *E. benthamii* population were fully dependent on relatedness, while in *E. pellita* marker-QTL LD might be contributing to predictions, although relatedness also seems to be the main driver.

Association genetics models comparison

GWAS under an LMA model, i.e. without the introduction of a GRM, resulted in a large number of associations, most or all of them likely spurious. For example, with only block as a covariate in the model, the number of SNPs associated with wood volume (WV) in *E. pellita* was 249. When the population structure was included as covariate, the number of associated SNPs was reduced to 120 (Fig. 1-5A, red line). The quantile-quantile (Q-Q) plot exhibited in Figure 1-5B shows the inappropriateness of the LMA model without the GRM, as the observed and expected *P*-values differed considerably for a large number of SNPs. When the genomic relationship matrix, block and structure effects were included in the MLMA model, five significant associations (Fig. 1-5C, blue line) were detected using a FDR of 0.05 (Table SM1-5). All these five significant SNPs have low allele frequency (MAF < 0.005). Nevertheless, when a more stringent adjustment for multiple testing was used (Bonferroni 5%), only one significant association persisted for volume in *E. pellita* (Fig. 1-5C, red line). In the MLMA model adjusted for the GRM, population structure and block covariates, most *P*-values were consistent with the expected ones along the diagonal in the Q-Q plot, indicating

suitability of this GWAS model (Fig. 1-5D). Furthermore, the model built with GRM considerably reduced the number of significant associations, likely removing spurious associations. The single SNP associated with volume in *E. pellita* on chromosome 6 (Fig. 1-5C, red line) is located in an exonic region of a gene whose function is involved in a plant-type cell wall cellulose biosynthetic process (Table SM1-5). In *E. benthamii*, no significant associations were found when the GRM was included in the model.

DISCUSSION

This study makes a further step toward the experimental assessment of whole-genomic prediction of complex traits in species of *Eucalyptus*. Our results corroborate previous reports in major *Eucalyptus* species showing encouraging perspectives of using genome-wide SNP data to capture large proportions of trait heritability and predict traits such as height and diameter growth with accuracies as good as or better than those attainable by conventional phenotypic selection.

Patterns of LD in *Eucalyptus* natural and breeding populations

The extent of LD detected in these populations reflect their differences in evolutionary and breeding history. A faster genome-wide LD decay was observed in *E. benthamii* (7.7 Kb, Fig. 1-1A) than in *E. pellita* (25.6 Kbp, Fig. 1-1C), in agreement with expectations. The *E. benthamii* breeding population is largely derived from seeds collected in wild stands, but confined to a few remnant populations with a history of bottlenecks displaying limited genetic diversity in previous studies with microsatellites (Butcher et al. 2005). Therefore, although composing a selected breeding group, it resembles a sample of a wild population. In fact, the extent of LD was similar to that found for natural populations of *E. grandis* (≈4-6 Kb) (Silva-Junior and Grattapaglia 2015). On the other hand, the *E. pellita* population comes from a clonal seed orchard established with relatively advanced selections such that a smaller effective population size and more

extensive LD were expected. Our results for *E. benthamii* provide additional genome-wide estimates supporting a slower decay of LD in *Eucalyptus* (Silva-Junior and Grattapaglia 2015). However, differently from what was found earlier, a shorter genome-wide extent of LD decay was detected when lower frequency SNPs were included in the analysis (Fig. 1-1) in agreement with the observation that rarer SNPs tend to generate lower pairwise r^2 estimates (Pritchard and Przeworski 2001).

Genomic heritabilities estimated from SNP data

Genomic and pedigree heritabilities could be compared for *E. benthamii*. Genomic heritabilities, irrespective of the method used, were generally lower than the pedigree-based estimate. For example, for DBH, GBLUP captured 55% (0.181/0.326) of the pedigree heritability, while Bayes B captured a larger proportion (88%; 0.287/0.326) (Table 1-2). Similar results were reported when estimating genomic heritability based on open-pollinated progenies of spruce (Beaulieu et al. 2014a; Gamal El-Dien et al. 2015). These studies argued that genomic heritabilities better reflect the true genetic relationships among individuals such that more realistic estimates of breeding values and genetic gain are obtained. Pedigree-based heritability estimates from open-pollinated families could be inflated due to the presence of full-sibs or selfs and the inability of these estimates to disentangle the non-additive from the additive genetic components. In fact, in a study in *Pinus taeda* the use of genomic relationships in marker-based models yielded substantially more precise separation of additive and non-additive components of genetic variance when compared to pedigree-based estimates, improving breeding value prediction (Muñoz et al. 2014). For *E. pellita,* pedigree-based heritability and accuracies of breeding values could not be estimated due to errors found in the recorded pedigree for these families. However, this fortuitous episode prompted the use of the SNP data to ascertain the pedigree of this breeding trial and provided estimates of heritability that breeders would not otherwise have had access to. Genomic heritability averages for the three growth

traits in *E. pellita* varied between 0.29 and 0.47, within the same range of the only available estimate for DBH (0.32) for this species in a trial in Vietnam (Hung et al. 2015).

Genomic heritability corresponds to the proportion of phenotypic variance that can be explained by regression on molecular markers. A study from de los Campos et al. (2015) showed that genomic heritability and trait heritability are equal only when all causal variants are typed, and as such, caution is required when interpreting heritability estimates from genomic data. Nevertheless, that same study also concluded that when close relatives sharing long chromosome segments are analyzed, high prediction accuracy and very small bias in genomic heritability estimates are expected. Given the relatively long range LD and relatedness present in our populations and the genome-wide SNP coverage adopted, our estimates of genomic heritability for the two species should therefore closely reflect the amount of additive genetic variance for the traits in these breeding populations.

Genomic predictions: methods

Predictive abilities of growth traits using GBLUP and different Bayesian methods, which have inherently different assumptions, reached similar results for all traits, in line with previous reports that assessed similar traits in forest trees (Resende Jr et al. 2012b; Isik et al. 2015; Ratcliffe et al. 2015). These results provide further evidence that growth traits in *Eucalyptus,* and likely for all forest trees for that matter, are complex in architecture, controlled by a large number of small effect loci and fit adequately the infinitesimal model. The predictive ability estimates obtained for growth traits in *E. pellita* (0.34-0.44) using GBLUP were slightly lower than those reported for *E. grandis x E. urophylla* (0.46-0.55) (Resende et al. 2012). On the other hand, higher values of predictive abilities for *E. pellita* (0.37-0.52) were obtained using a non-additive (additive + dominance + epistatic effects) model, which may be explained by the more advanced selections that this population has been subjected to (more extensive LD = 25.6 Kbp). For *E.*

benthamii, predictive abilities were lower (~0.16), possibly the result of (i) the larger effective population size, (ii) the relatively limited number of individuals used for model training (only ~500), and (iii) the narrow genetic diversity available in this species and particularly so in this introduced population in Brazil. Still, what really matters from the applied breeding standpoint is that the predictive ability reached with genomic data was as good as or better than the predictive ability based on phenotypic data. Therefore, irrespective of the absolute estimates of genomic prediction ability intrinsic to each population, this result satisfactorily fulfills the promising perspective of adopting GS to accelerate breeding cycles by predicting breeding values of yet-to-be-phenotyped trees at early age.

Genomic predictions: number and positions of SNPs

Prediction models using ~5,000 SNPs provided predictive abilities almost equivalent to using all available SNPs for all traits with only a slight gain when moving to 10,000 SNPs. No difference was observed in predictive abilities when different sets of SNPs sampled based on genomic position were used, as long as the total number of SNPs reached such numbers. These results suggest that genomic prediction is largely driven by relatedness between training and validation and, once a certain number of randomly sampled SNPs across the genome are used, suitable predictive ability is reached. From a practical standpoint, this outcome indicates that low-density SNP chips could be contemplated as a way to reduce cost of GS and broaden its application to a larger number of breeding programs that operate on small budgets. Low-density SNP panels have been a standar practice in domestic animals, providing predictive abilities equivalent to those observed using the full set of SNPs, depending on the extent of LD, trait and other population parameters (Habier et al. 2009; Van Eenennaam et al. 2014). Despite the potential advantage of using smaller SNPs subsets to reduce costs, it is expected however that genomic predictions will decay over generations due to the combined effect of recombination and selection on the patterns of LD (Solberg et al. 2008) Habier et al. (2009) showed that GS predictions using low-density

panel decreased over generations, but it remained constant when high-density SNP panel was used to genotype the few individuals selected in each breeding generation. Nevertheless, if prediction accuracies are mainly driven by relationship, low-density marker panels could be suitable, provided that continuous model retraining strategies are adopted (Iwata et al. 2011). At this point, therefore, it is not clear whether the use of smaller SNP subsets is warranted for the long-term implementation of GS in *Eucalyptus*. A better assessment will be possible when predictions are carried out across breeding generations testing variable SNP densities. On the positive side, however, a recent study showed consistent prediction accuracies over two generations (parents and grandparents as training set and descendants as validation set) with a relatively modest panel of only 4,436 SNPs in *Pinus pinaster* (Bartholomé et al. 2016b).

Impact of relatedness on genomic prediction

We observed a major impact of relatedness on predictions, more so in *E. benthamii* than *E. pellita* (Fig. 1-4) consistent with theoretical expectations (Habier et al. 2007) and previous experimental results in forest trees (Resende et al. 2012; Beaulieu et al. 2014a; Beaulieu et al. 2014b). The relative contributions of historical LD and relatedness are however difficult to disentangle. Predictive ability can be high even in the absence of LD when markers capture genetic relationships, but it will be even greater if markers are in LD with causal loci (Habier et al. 2007). The presence of some level of historical LD could in part explain why predictions were still reasonable in *E. pellita* even after attempting to minimize relatedness between training and validation sets (Fig. 1-4B). However, another possibility is that our attempt to decrease relatedness was not completely efficient. A way to test this would be to compare the predictive abilities obtained using the same number of markers concentrated on a single chromosome (capturing largely the effect of relatedness), versus distributed genome-wide (capturing relatedness and LD). Assuming an infinitesimal model in which growth traits are controlled by many QTLs with small effects distributed genome-wide, the difference between these

two sets could be tentatively taken as the contribution of historical LD to predictions. In *E. pellita*, such a test was done when predictions carried out with 2,583 SNPs mapped only to chromosome 8 (Table 1-4) were compared to using a set of 2,297 evenly-spaced SNPs across the genome (Table 1-3). An increase of 22 to 35% in predictive ability was seen (e.g. 0.306 versus 0.414 for DBH) when genome-wide SNPs were used, suggesting that some LD between markers and causal loci could be accounted for in this population. Overall, however, our results corroborate previous reports on the major impact of relatedness on genomic prediction and further highlight the importance of properly planning the populations on which GS models will be trained and those where the models will be applied. If the training population is more or less related to the validation population than the future selection candidates, then the expected outcome of implementing genomic selection will be over- or underestimated, respectively.

GWAS versus GS in breeding populations

The genome-wide association studies carried out on the same datasets allowed us to assess the value of this approach in closed breeding populations under selection and compare it to genomic prediction from the standpoint of practical breeding for growth, the most important trait in all tree breeding programs irrespective of species. After controlling for population structure and experimental fixed effects, and applying experiment wise corrections for multiple tests, GWA identified only one significant association for volume growth in *E. pellita* (Fig. 1-5C). Clearly our populations had limited power to detect loci controlling complex traits such as growth, despite the relatively larger size of the *E. pellita* population (n = 732) when compared to populations used in published GWAS for growth related traits in forest trees, usually between 300 and 500 individuals (Cappa et al. 2013; Evans et al. 2014; Mckown et al. 2014; Fahrenkrog et al. 2016). Our results are in line with a recent GWAS study using a natural population of 714 *Populus nigra* trees, where, despite using over 10,000 SNPs specifically targeting genes and previously mapped QTL, only three significant associations for stem height

were detected (Allwright et al. 2016). The comparison between GS and GWAS corroborated the fact that while genome-wide regression is able to account for large proportions of the pedigree-heritability (e.g. 73% for DBH in *E. benthamii*) and provide useful phenotype predictions, very little or none of the heritability is captured into significant associations using the GWAS approach. Reasons for this major discrepancy have by now been widely discussed in the plant, animal and human literature (Lorenz et al. 2011; Robinson et al. 2014; Meuwissen et al. 2016). They derive essentially from the fact that GWAS by principle, relies on the application of stringent significance tests to declare an association. These very stringent tests typically result in only the largest effect QTLs being found, while the vast majority have too small an effect to be detectable in the limited power GWAS populations used and so are ignored. If no major effect exists, then no associations are found, which is most likely the case of our disappointing results for the growth traits targeted in our study.

A potential criticism to our GWAS assessment is that it was carried out in a breeding population with a relatively limited effective population size and limited diversity and not in a canonical GWAS population sampled from the wild where possibly more associations could be found. GWAS studies for growth traits in forest trees have mostly targeted collections of trees derived from natural populations in *Populus* (Evans et al. 2014; Mckown et al. 2014; Allwright et al. 2016; Fahrenkrog et al. 2016), *Eucalyptus* (Cappa et al. 2013) and *Pinus* (Cumbie et al. 2011). The driving goal of such studies has been to detect associations that would potentially allow gene discovery or even the identification of the elusive QTN (quantitative trait nucleotide) (Rockman 2012). However, notwithstanding the fact that very few associations were found for growth traits in these studies as well (e.g. Cumbie et al. 2011; Allwright et al. 2016), and that they explained overall negligible fractions of trait heritability, it still remains to be seen how such SNP-trait associations found in natural populations, far removed from selected breeding material, will be translated into useful information to breeding. Furthermore, targeted alleles found by GWAS in natural populations might already be fixed or

simply not be sampled in breeding populations (Hamblin et al. 2011). In our study, we were otherwise interested in assessing the value of genome-wide association in detecting discrete associations for complex growth traits in operational breeding populations under selection. Although less genetic variation is available in such closed breeding populations, associations found in such selected material should be considerably more useful to inform practical breeding decisions. Despite the disappointing, although not unexpected, results of our GWAS for growth traits in *Eucalyptus*, the availability of GWAS data could be valuable to improve genomic predictions accuracies. GWAS might provide information on traits' genetic architectures that can be used to assign locus- or trait-specific priors to genomic prediction models (Daetwyler et al. 2010). A recent study in rice showed that association mapping in breeding populations provided useful information for breeding decisions (Begum et al. 2015), which, integrated as fixed effects into genomic predictions, increased the accuracy (Spindel et al. 2016). Results of our study suggest, however, that such an approach in undomesticated outcrossed forest trees will require an increase of several order of magnitude in sample size, such that at least a portion of the larger effects segregating in the population may be uncovered by GWAS.

CONCLUSIONS

This study contributes further experimental data supporting the positive prospects of whole-genome regression methods to account for large proportions of trait heritability and predict traits such as height and diameter growth in forest trees with accuracies equivalent or superior to those achievable by phenotypic selection. We show that genetic relatedness captured by the SNP data between training and validation populations and, by extension, to future selection candidates, is what will most likely determine the successful use of genomic selection in *Eucalyptus* breeding. Finally, by evaluating different SNP sampling schemes across the genome we conclude that more important to GS than the number and position of the SNPs fitted in the model, is the extensive LD created in closed breeding

populations with small effective population sizes. Lower density SNP panels with ~5,000 to 10,000 SNPs, distributed across the genome, should provide a good compromise between genotyping costs and predictive ability in such standard breeding populations advanced by open pollinated breeding. Still, further experiments are needed to evaluate the performance of such SNP densities across generations of breeding. Finally, our results illustrate the superiority of the whole-genome regression approach in accounting for large proportions of the heritability in contrast to the limited value of the local GWAS approach for breeding applications. To provide useful GWAS data toward breeding for growth traits in *Eucalyptus* and likely in all forest trees, it will be necessary first to massively increase the sample size, such that sufficient power is reached to detect at least part of the slightly larger effects segregating in the target breeding population. In the meantime, the encouraging results of genomic prediction that we, and others, have shown in this and other studies should probably receive greater attention if the objective is to impact breeding practice.

TABLES

Table 1-1: General attributes of the trials studied for *E. benthamii* and *E. pellita*.

Phenotypic data	***E. benthamii***	***E. pellita***
Total number of trees in trial	2,000	960
Total number of open pollinated (OP) families	40	24
Number of blocks	50	40
Number of individuals/OP family	10	32
Number of trees measured	508	747
Number of trees used for analyses	505	732
Effective population size (N_e)	50	35
Age at phenotyping (*yr*)	4.6	3.5
Site	Candói, PR	Rio Verde, GO
Coordinates	25º43'0"S/52º11'0"W	17º44´42"S/50º55'00"W
Number of traits	3	3

Table 1-2: Estimates of narrow-sense heritabilities (h^2) and predictive abilities (r_{gy}), pedigree (ABLUP) and genome based (several methods), for the *E. benthamii* and *E. pellita* breeding populations.

Method	Filter	*E. benthamii*						*E. pellita*					
		DBH		HT		WV		DBH		HT		WV	
		h^2 (SE)	r_{gy} (SE)	h^2 (SE)	r_{gy} (SE)	h^2 (SE)	r_{gy} (SE)	h^2 (SE)	r_{gy} (SE)	h^2 (SE)	r_{gy} (SE)	h^2 (SE)	r_{gy} (SE)
ABLUP		0.326 (NA)	0.148 (0.045)	0.088 (NA)	0.090 (0.033)	0.297 (NA)	0.142 (0.039)	0.000 (NA)	- 0.030 (0.028)	0.019 (NA)	0.040 (0.028)	0.000 (NA)	- 0.009 (0.026)
RR-BLUP	MAF > 0	0.181 (NA)	0.157 (0.044)	0.000 (NA)	0.006 (0.044)	0.147 (NA)	0.141 (0.041)	0.466 (NA)	0.439 (0.019)	0.260 (NA)	0.342 (0.042)	0.424 (NA)	0.424 (0.028)
Bayes A		0.202 (0.017)	0.160 (0.045)	0.058 (0.016)	0.010 (0.040)	0.165 (0.020)	0.141 (0.041)	0.465 (0.008)	0.440 (0.019)	0.280 (0.011)	0.342 (0.042)	0.428 (0.008)	0.424 (0.028)
Bayes B		0.287 (0.032)	0.166 (0.045)	0.155 (0.052)	0.003 (0.041)	0.284 (0.028)	0.146 (0.038)	0.527 (0.020)	0.439 (0.019)	0.341 (0.017)	0.342 (0.042)	0.517 (0.025)	0.425 (0.028)
Bayes Cπ		0.267 (0.017)	0.158 (0.044)	0.109 (0.007)	0.016 (0.039)	0.237 (0.014)	0.148 (0.039)	0.480 (0.007)	0.439 (0.019)	0.303 (0.009)	0.342 (0.042)	0.453 (0.007)	0.423 (0.028)
BL		0.133 (0.019)	0.155 (0.045)	0.044 (0.004)	0.010 (0.042)	0.103 (0.011)	0.140 (0.041)	0.414 (0.021)	0.434 (0.021)	0.242 (0.014)	0.338 (0.043)	0.406 (0.007)	0.424 (0.028)
BRR		0.267 (0.008)	0.162 (0.044)	0.190 (0.005)	0.022 (0.036)	0.243 (0.008)	0.146 (0.039)	0.455 (0.005)	0.441 (0.019)	0.283 (0.008)	0.342 (0.042)	0.418 (0.005)	0.425 (0.028)
RR-BLUP	MAF ≥ 5%	0.179 (NA)	0.153 (0.044)	0.000 (NA)	0.009 (0.044)	0.144 (NA)	0.138 (0.041)	0.457 (NA)	0.437 (0.020)	0.254 (NA)	0.340 (0.042)	0.419 (NA)	0.422 (0.028)
Bayes A		0.214 (0.013)	0.158 (0.045)	0.073 (0.008)	0.020 (0.040)	0.190 (0.017)	0.144 (0.041)	0.463 (0.007)	0.438 (0.020)	0.279 (0.008)	0.340 (0.042)	0.437 (0.005)	0.422 (0.028)
Bayes B		0.354 (0.041)	0.162 (0.045)	0.110 (0.016)	0.019 (0.040)	0.269 (0.029)	0.146 (0.040)	0.551 (0.020)	0.438 (0.019)	0.393 (0.036)	0.339 (0.042)	0.501 (0.010)	0.423 (0.028)
Bayes Cπ		0.259 (0.011)	0.157 (0.046)	0.116 (0.006)	0.020 (0.039)	0.232 (0.008)	0.143 (0.040)	0.485 (0.005)	0.437 (0.020)	0.300 (0.008)	0.340 (0.042)	0.449 (0.007)	0.423 (0.028)
BL		0.143 (0.023)	0.153 (0.043)	0.045 (0.003)	0.020 (0.041)	0.101 (0.009)	0.134 (0.041)	0.408 (0.009)	0.427 (0.023)	0.244 (0.010)	0.339 (0.041)	0.403 (0.006)	0.422 (0.029)
BRR		0.260 (0.007)	0.158 (0.044)	0.184 (0.004)	0.025 (0.036)	0.239 (0.006)	0.143 (0.040)	0.443 (0.005)	0.437 (0.020)	0.280 (0.008)	0.341 (0.042)	0.415 (0.006)	0.422 (0.028)

Pedigree BLUP (ABLUP, Pedigree Best Linear Unbiased Predictor), Ridge Regression BLUP (RR-BLUP, Genomic Best Linear Unbiased Predictor), BL (Bayesian Lasso), BRR (Bayesian Ridge-Regression), Minimum Allele Frequency (MAF), Diameter at Breast Height (DBH, cm), Total Height (HT, m) and Wood Volume (WV, m^3). NA - The standard error of the heritability could not be estimated using rrBLUP (Endelman 2011).

Table 1-3: Genomic estimates of narrow-sense heritabilities (h^2) and predictive abilities (r_{gy}) for the *E. benthamii* and *E. pellita* breeding populations using different SNP sampling methods.

SNP sampling method	*E. benthamii*					*E. pellita*					
	Number of SNPs	DBH		WV		Number of SNPs	DBH		HT		WV	
		h^2 (SE)	r_{gy} (SE)	h^2 (SE)	r_{gy} (SE)		h^2 (SE)	r_{gy} (SE)	h^2 (SE)	r_{gy} (SE)	h^2 (SE)	r_{gy} (SE)
All SNPs	13,787	0.267 (0.008)	0.162 (0.044)	0.243 (0.008)	0.146 (0.039)	19,506	0.455 (0.005)	0.441 (0.019)	0.283 (0.008)	0.342 (0.042)	0.418 (0.005)	0.425 (0.028)
Randomly selected	5,000	0.250 (0.003)	0.163 (0.004)	0.234 (0.003)	0.148 (0.004)	5,000	0.410 (0.006)	0.427 (0.003)	0.269 (0.003)	0.336 (0.002)	0.390 (0.004)	0.416 (0.003)
Randomly selected	3,000	0.239 (0.005)	0.153 (0.008)	0.226 (0.004)	0.137 (0.008)	3,000	0.385 (0.006)	0.417 (0.003)	0.254 (0.005)	0.328 (0.003)	0.363 (0.006)	0.406 (0.003)
Randomly selected	1,500	0.229 (0.005)	0.153 (0.008)	0.217 (0.005)	0.137 (0.008)	1,500	0.334 (0.005)	0.397 (0.003)	0.232 (0.004)	0.313 (0.003)	0.322 (0.004)	0.389 (0.003)
Randomly selected	500	0.181 (0.006)	0.104 (0.017)	0.174 (0.005)	0.091 (0.015)	500	0.270 (0.008)	0.364 (0.006)	0.203 (0.003)	0.291 (0.006)	0.264 (0.008)	0.361 (0.006)
Evenly spaced 10 Kbp	10,837	0.264 (0.007)	0.159 (0.046)	0.235 (0.007)	0.141 (0.041)	13,946	0.452 (0.004)	0.436 (0.019)	0.272 (0.009)	0.340 (0.042)	0.415 (0.010)	0.421 (0.028)
Evenly spaced 50 Kbp	6,867	0.253 (0.007)	0.153 (0.041)	0.242 (0.006)	0.135 (0.035)	7,619	0.472 (0.008)	0.440 (0.021)	0.286 (0.008)	0.339 (0.043)	0.439 (0.008)	0.421 (0.031)
Evenly spaced 100 Kbp	4,634	0.252 (0.004)	0.146 (0.044)	0.241 (0.006)	0.141 (0.036)	4,846	0.460 (0.006)	0.442 (0.024)	0.287 (0.007)	0.339 (0.041)	0.452 (0.008)	0.432 (0.031)
Evenly spaced 250 Kbp	2,281	0.261 (0.004)	0.166 (0.039)	0.258 (0.005)	0.160 (0.029)	2,297	0.374 (0.007)	0.414 (0.026)	0.271 (0.005)	0.328 (0.042)	0.360 (0.004)	0.400 (0.030)
Evenly spaced 500 Kbp	1,203	0.212 (0.006)	0.131 (0.053)	0.199 (0.004)	0.116 (0.050)	1,204	0.326 (0.004)	0.388 (0.026)	0.226 (0.004)	0.306 (0.043)	0.307 (0.005)	0.378 (0.033)
Evenly spaced 1 Mbp	610	0.196 (0.002)	0.111 (0.031)	0.178 (0.003)	0.097 (0.022)	609	0.256 (0.004)	0.364 (0.027)	0.203 (0.004)	0.307 (0.041)	0.260 (0.004)	0.365 (0.029)
Genic regions	7,254	0.251 (0.008)	0.163 (0.045)	0.240 (0.006)	0.148 (0.037)	11,212	0.421 (0.007)	0.433 (0.020)	0.269 (0.008)	0.340 (0.042)	0.394 (0.005)	0.426 (0.028)
Intergenic regions	6,533	0.253 (0.008)	0.152 (0.046)	0.232 (0.005)	0.131 (0.046)	8,294	0.449 (0.007)	0.432 (0.021)	0.289 (0.009)	0.340 (0.041)	0.414 (0.006)	0.410 (0.030)
SNPs in LE (LD-pruning)	10,460	0.274 (0.011)	0.174 (0.043)	0.256 (0.010)	0.161 (0.039)	10,984	0.425 (0.010)	0.426 (0.024)	0.275 (0.007)	0.339 (0.041)	0.404 (0.006)	0.413 (0.031)

Pedigree BLUP (ABLUP, Pedigree Best Linear Unbiased Predictor), Ridge Regression BLUP (RR-BLUP, Genomic Best Linear Unbiased Predictor), BL (Bayesian Lasso), BRR (Bayesian Ridge-Regression), Minimum Allele Frequency (MAF), Diameter at Breast Height (DBH, cm), Total Height (HT, m) and Wood Volume (WV, m^3).

Table 1-4: Genomic estimates of narrow-sense heritabilities (h^2) and predictive abilities (r_{gy}) for the *E. benthamii* and *E. pellita* breeding populations using chromosome-specific SNP sets.

Chr.	*E. benthamii*					*E. pellita*						
	Number of SNPs	DBH		WV		Number of SNPs	DBH		HT		WV	
		h^2 (SE)	r_{gy} (SE)	h^2 (SE)	r_{gy} (SE)		h^2 (SE)	r_{gy} (SE)	h^2 (SE)	r_{gy} (SE)	h^2 (SE)	r_{gy} (SE)
1	848	0.162 (0.004)	0.070 (0.048)	0.161 (0.003)	0.071 (0.037)	1,329	0.240 (0.004)	0.336 (0.034)	0.223 (0.006)	0.327 (0.042)	0.241 (0.004)	0.337 (0.031)
2	1,672	0.186 (0.003)	0.085 (0.036)	0.183 (0.004)	0.071 (0.034)	2,245	0.228 (0.004)	0.313 (0.033)	0.188 (0.004)	0.272 (0.040)	0.218 (0.004)	0.303 (0.036)
3	1,544	0.195 (0.004)	0.170 (0.036)	0.207 (0.004)	0.185 (0.040)	2,026	0.282 (0.007)	0.363 (0.042)	0.172 (0.003)	0.282 (0.046)	0.267 (0.006)	0.355 (0.043)
4	886	0.180 (0.004)	0.134 (0.036)	0.171 (0.004)	0.104 (0.027)	1,303	0.256 (0.008)	0.315 (0.045)	0.203 (0.003)	0.271 (0.044)	0.251 (0.009)	0.294 (0.049)
5	1,356	0.195 (0.004)	0.123 (0.051)	0.190 (0.004)	0.123 (0.052)	1,872	0.303 (0.006)	0.379 (0.037)	0.227 (0.007)	0.325 (0.044)	0.277 (0.006)	0.353 (0.040)
6	1,440	0.166 (0.004)	0.090 (0.040)	0.157 (0.002)	0.063 (0.033)	2,012	0.277 (0.007)	0.375 (0.031)	0.197 (0.004)	0.294 (0.040)	0.274 (0.008)	0.369 (0.037)
7	1,207	0.219 (0.006)	0.187 (0.051)	0.210 (0.006)	0.158 (0.046)	1,594	0.226 (0.003)	0.337 (0.031)	0.168 (0.003)	0.241 (0.047)	0.217 (0.003)	0.323 (0.038)
8	1,771	0.183 (0.005)	0.082 (0.063)	0.168 (0.004)	0.059 (0.050)	2,583	0.212 (0.006)	0.306 (0.038)	0.185 (0.003)	0.267 (0.040)	0.222 (0.004)	0.316 (0.037)
9	940	0.170 (0.003)	0.121 (0.035)	0.164 (0.004)	0.100 (0.033)	1,381	0.228 (0.004)	0.330 (0.020)	0.182 (0.004)	0.285 (0.034)	0.233 (0.006)	0.332 (0.027)
10	1,034	0.152 (0.002)	0.059 (0.037)	0.150 (0.002)	0.047 (0.041)	1,448	0.218 (0.003)	0.339 (0.037)	0.184 (0.004)	0.292 (0.050)	0.224 (0.004)	0.353 (0.041)
11	1,089	0.195 (0.004)	0.138 (0.040)	0.193 (0.006)	0.143 (0.041)	1,713	0.250 (0.005)	0.338 (0.024)	0.201 (0.005)	0.304 (0.042)	0.258 (0.006)	0.350 (0.026)
All	13,787	0.267 (0.008)	0.162 (0.044)	0.243 (0.008)	0.146 (0.039)	19,506	0.455 (0.005)	0.441 (0.019)	0.283 (0.008)	0.342 (0.042)	0.418 (0.005)	0.425 (0.028)

Diameter at Breast Height (DBH, cm), Total Height (HT, m) and Wood Volume (WV, m^3).

FIGURES

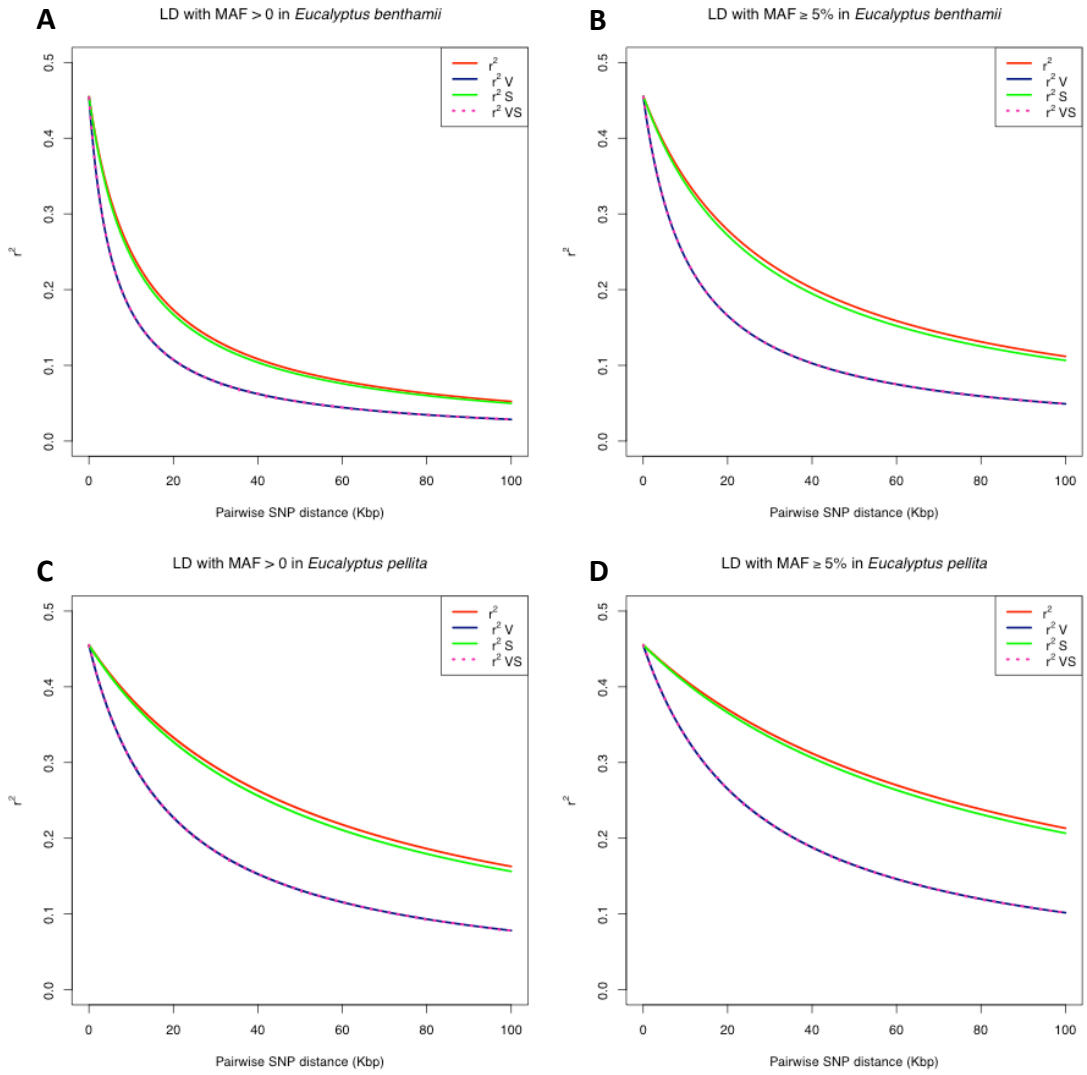


Figure 1-1: Genome-wide pattern of Linkage Disequilibrium (LD) decay up to 100 Kbp pairwise SNP distances. Decay curves of the classical measure of the squared correlation of allele frequencies at diallelic loci (r^2), adjusted for population structure (r^2S) and relatedness (r^2V), and adjusted for both (r^2VS). (A) Plot with SNPs filtered using MAF > 0 and (B) MAF ≥ 5% in *E. benthamii*. (C) Plot with SNPs filtered using MAF > 0 and (D) MAF ≥ 5% in *E. pellita*.

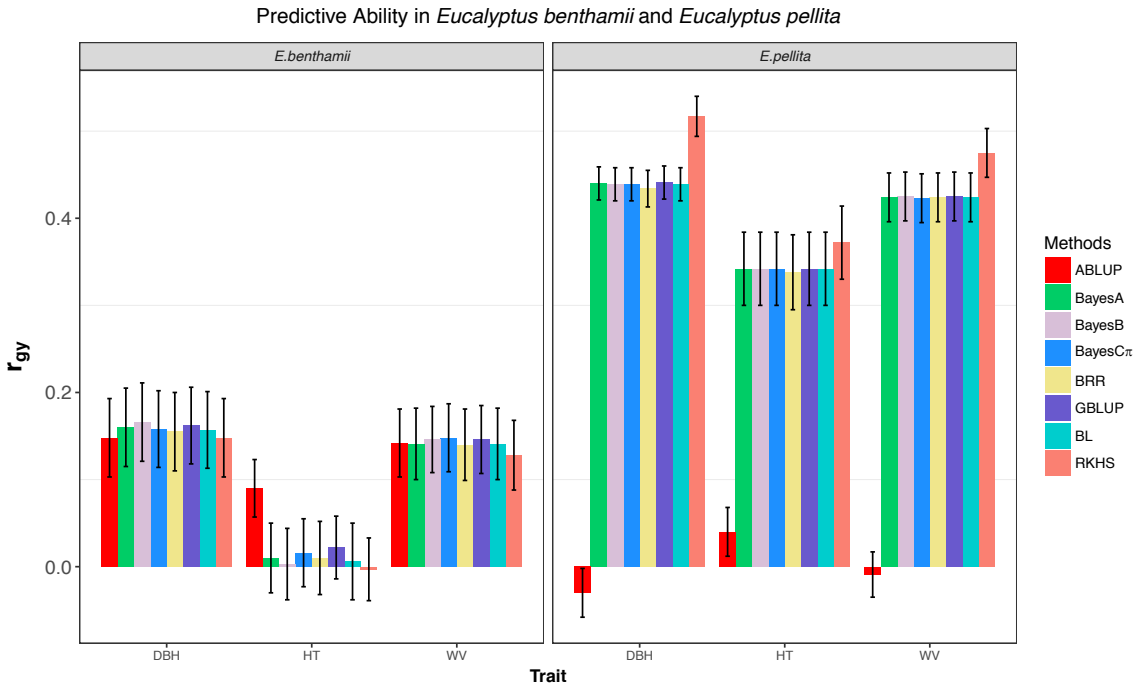


Figure 1-2: Estimates of predictive ability (r_{gy}) by ABLUP, GBLUP, Bayesian methods and RKHS. Estimates of r_{gy} for the *E. benthamii* and *E. pellita* breeding populations.

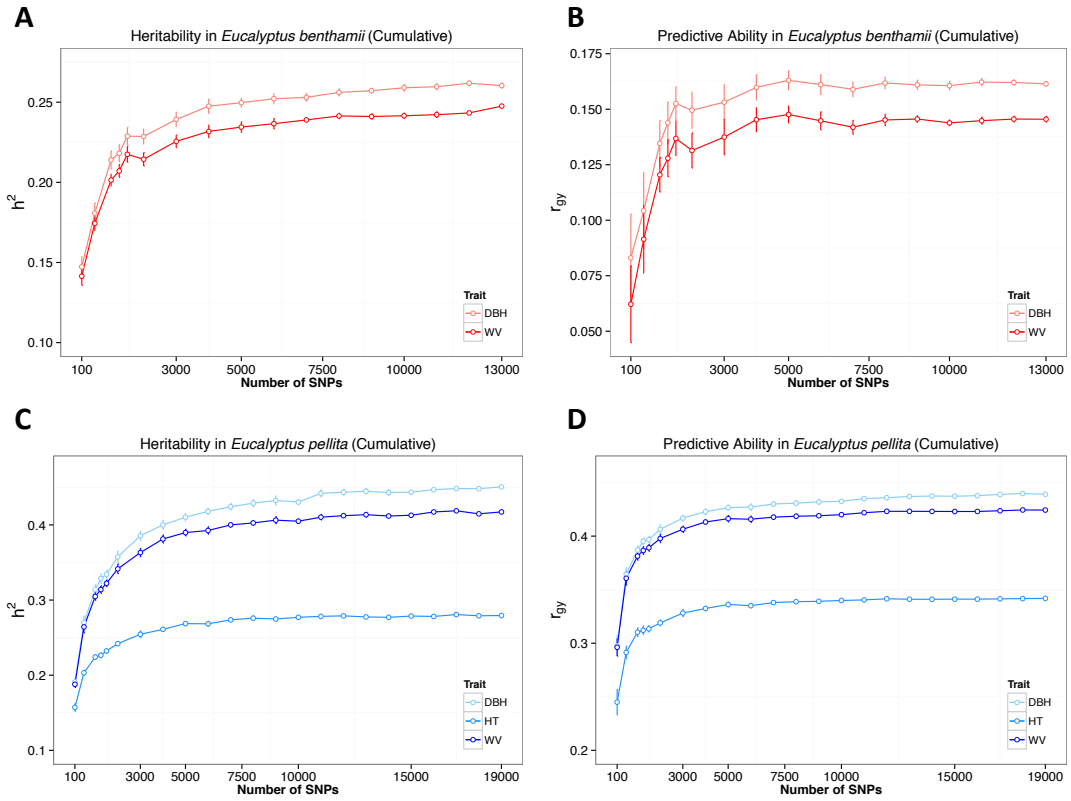


Figure 1-3: Estimates of heritability (h^2) and of predictive ability (r_{gy}) with increasing numbers of SNPs for different traits using a cumulative approach to SNP sampling. (A) and (B) estimates of h^2 and r_{gy} for *E. benthamii*, respectively. (C) and (D) estimates of h^2 and r_{gy} for *E. pellita*, respectively.

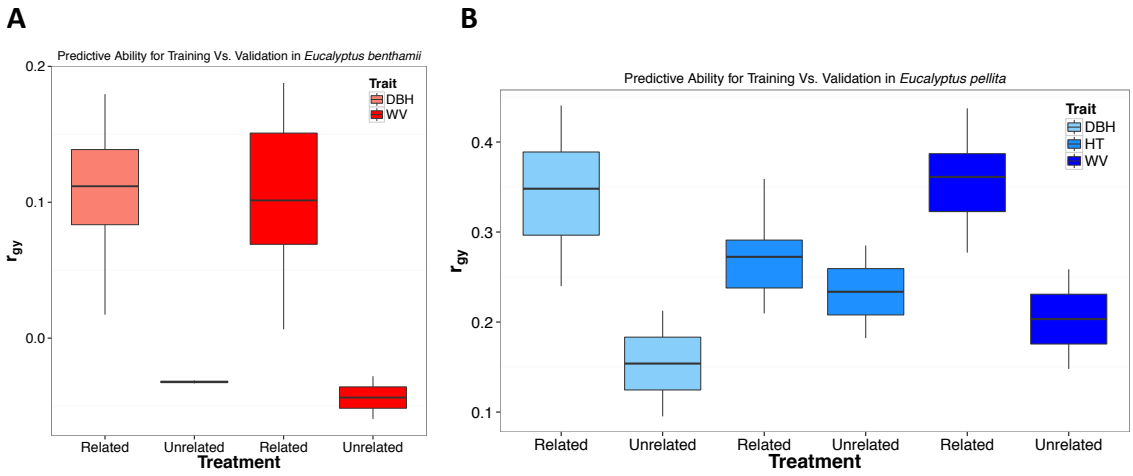


Figure 1-4: Estimates of predictive ability (r_{gy}) with different levels of relatedness between training and validation sets. Related: random allocation of individuals to training and validation sets; Unrelated: individuals were split into training and validation sets by minimizing relatedness between sets based on a principal component analysis. (A) *E. benthamii* and (B) *E. pellita*.

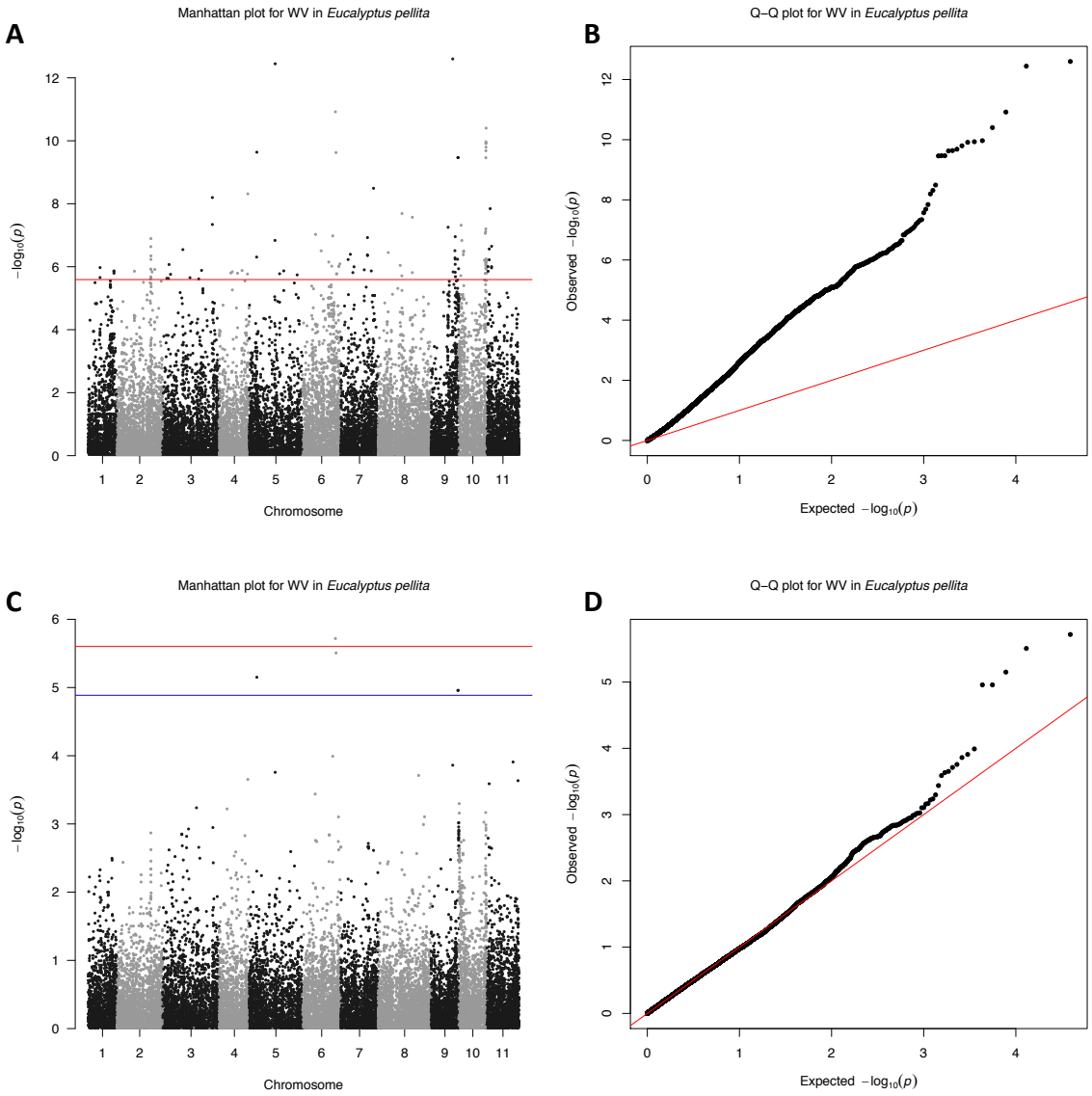


Figure 1-5: Manhattan and Quantile-quantile (Q-Q) plots for wood volume (WV) in *E. pellita*. (A) and (B) represent the Manhattan and the Q-Q plots, respectively, for LMA model adjusted for block and population structure covariates. (C) and (D) represent the Manhattan and the Q-Q plots, respectively, for MLMA model adjusted for block and population structure covariates, and also for the genomic relationship matrix. Red line indicates Bonferroni-corrected threshold with an experimental type I error rate at α = 0.05 and blue line indicates false discovery rate (FDR) at 5%.

SUPPLEMENTARY MATERIAL (SM1)

Table SM1-1: Numbers of SNPs and average distances between SNPs for the variable window sizes used to select evenly spaced SNP subsets for *E. benthamii* and *E. pellita*.

Selected window (Kbp)	*E. benthamii*		*E. pellita*	
	Average distance between markers (Kbp)	Number of SNPs	Average distance between markers (Kbp)	Number of SNPs
10	55.7	10,837	43.2	13,946
50	87.4	6,867	78.6	7,619
100	130	4,634	124	4,846
250	265	2,281	263	2,297
500	504	1,203	503	1,204
1,000	1,000	610	1,002	609

Table SM1-2: Linkage Disequilibrium (LD) estimates and genome-wide pattern of decay of LD up to pairwise SNP distance of 100 Kbp including rare alleles (MAF > 0) or not (MAF ≥ 5%) for the *E. benthamii* and *E. pellita* populations.

LD measurement	*E. benthamii*		*E. pellita*	
	MAF > 0	**MAF ≥ 5%**	**MAF > 0**	**MAF ≥ 5%**
Number of SNPs	13,787	7,563	19,506	12,483
		2,817,21	18,146,36	7,494,37
Number of SNPs pairwise	9,157,068	0	6	9
Mean r^2 all data	0.0169	0.0164	0.0167	0.0187
Mean r^2S all data	0.0161	0.0149	0.0158	0.0176
Mean r^2V all data	0.0117	0.0063	0.0067	0.0063
Mean r^2VS all data	0.0117	0.0063	0.0067	0.0063
Mean r^2 in 100 Kbp	0.1413	0.2284	0.2713	0.3173
Mean r^2S in 100 Kbp	0.1372	0.2225	0.2654	0.3123
Mean r^2V in 100 Kbp	0.0966	0.1443	0.1790	0.2142
Mean r^2VS in 100 Kbp	0.0966	0.1444	0.1788	0.2142
Mean r^2 in 50 Kbp	0.2015	0.2955	0.3331	0.3697
Mean r^2S in 50 Kbp	0.1966	0.2897	0.3280	0.3659
Mean r^2V in 50 Kbp	0.1451	0.2050	0.2437	0.2795
Mean r^2VS in 50 Kbp	0.1451	0.2051	0.2435	0.2794
r^2 <0.2 within (Kbp)	15.622	40.693	70.595	112.678
r^2S <0.2 within (Kbp)	14.755	38.096	66.237	106.198
r^2V <0.2 within (Kbp)	7.708	14.515	25.614	35.888
r^2VS <0.2 within (Kbp)	7.708	14.526	25.556	35.862
Half-decay distance (Kbp) r^2	12.217	31.729	55.106	87.923
Half-decay distance (Kbp) r^2S	11.543	29.701	51.706	82.865
Half-decay distance (Kbp) r^2V	6.059	11.348	20.033	28.042
Half-decay distance (Kbp) r^2VS	6.059	11.365	19.987	28.022
Average of SNPs by Chr.	1,253.4	687.5	1,773.3	1,134.8

LD estimates with classical measure of the squared correlation of allele frequencies at diallelic loci (r^2), adjusted for population structure (r^2S) and relatedness (r^2V), and adjusted for both (r^2VS).

Table SM1-3. Estimates of additive genetic variance (σ^2a) and residual variance (σ^2e) obtained with different prediction methods, different position-based SNP sampling methods and sampling related or unrelated individuals in the *E. benthamii* and *E. pellita* breeding populations.

Method	*E. benthamii*									*E. pellita*											
	# SNPs	DBH				WV				# SNPs	DBH				HT				WV			
		σ^2_a	SE (σ^2_a)	σ^2_e	SE (σ^2_e)	σ^2_a	SE (σ^2_a)	σ^2_e	SE (σ^2_e)		σ^2_a	SE (σ^2_a)	σ^2_e	SE (σ^2_e)	σ^2_a	SE (σ^2_a)	σ^2_e	SE (σ^2_e)	σ^2_a	SE (σ^2_a)	σ^2_e	SE (σ^2_e)
ABLUP	13787	2.4943	NA	5.1483	NA	0.0257	NA	0.0608	NA	19506	8.79E-09	NA	8.7930	NA	0.2114	NA	10.7807	NA	0.0000	NA	0.0045	NA
GBLUP	13787	1.3937	NA	6.3136	NA	0.0128	NA	0.0741	NA	19506	4.2236	NA	4.8406	NA	2.8275	NA	8.0514	NA	0.0019	NA	0.0026	NA
Bayes A	13787	1.5942	0.1418	6.2525	0.1177	0.0147	0.0019	0.0737	0.0015	19506	4.2697	0.1056	4.8943	0.0419	3.1044	0.1285	7.9792	0.1240	0.0020	4.62E-05	0.0026	2.50E-05
Bayes B	13787	2.6576	0.4473	6.2146	0.1251	0.0294	0.0038	0.0714	0.0013	19506	5.6134	0.4853	4.8823	0.0403	4.1862	0.3146	7.9754	0.0963	0.0030	0.00042 3952	0.0026	2.64E-05
Bayes Cπ	13787	2.2375	0.1565	6.1138	0.1125	0.0221	0.0015	0.0710	0.0011	19506	4.5744	0.1051	4.9500	0.0293	3.4967	0.1039	8.0294	0.1096	0.0022	4.23E-05	0.0026	2.43E-05
BL	13787	1.0156	0.1491	6.6508	0.1614	0.0090	0.0010	0.0779	0.0013	19506	3.6866	0.2344	5.1761	0.1236	2.6284	0.1696	8.2119	0.1295	0.0018	3.84E-05	0.0027	2.62E-05
BRR	13787	2.1556	0.0722	5.9168	0.0803	0.0228	0.0008	0.0688	0.0008	19506	4.1281	0.0646	4.9378	0.0279	3.1423	0.0908	7.9520	0.1014	0.0019	2.99E-05	0.0026	1.52E-05
SNP sampling method		σ^2_a	SE (σ^2_a)	σ^2_e	SE (σ^2_e)	σ^2_a	SE (σ^2_a)	σ^2_e	SE (σ^2_e)		σ^2_a	SE (σ^2_a)	σ^2_e	SE (σ^2_e)	σ^2_a	SE (σ^2_a)	σ^2_e	SE (σ^2_e)	σ^2_a	SE (σ^2_a)	σ^2_e	SE (σ^2_e)
Randomly selected SNPs (10 rep)	5000	2.0254	0.0560	6.0364	0.0842	0.0215	0.0006	0.0695	0.0009	5000	3.5780	0.0637	5.2892	0.0306	2.9728	0.0715	8.0812	0.0902	0.0017	3.2469E-05	0.0028	2.06256E-05
Randomly selected SNPs (10 rep)	3000	1.8909	0.0489	6.1565	0.0791	0.0203	0.0006	0.0707	0.0009	3000	3.4632	0.0617	5.4279	0.0276	2.8353	0.0593	8.1955	0.0890	0.0016	2.39674E-05	0.0029	2.07694E-05
Randomly selected SNPs (10 rep)	1500	1.8066	0.0464	6.2528	0.0775	0.0195	0.0005	0.0719	0.0009	1500	2.9479	0.0541	5.8354	0.0286	2.5600	0.0546	8.4846	0.0781	0.0014	1.90164E-05	0.0030	1.91402E-05
Randomly selected SNPs (10 rep)	500	1.4420	0.0313	6.6996	0.0771	0.0160	0.0004	0.0758	0.0009	500	2.3536	0.0386	6.4651	0.0285	2.2174	0.0409	8.8741	0.0698	0.0012	1.9618E-05	0.0033	1.95043E-05
Evenly spaced 10 Kb window	10837	2.1316	0.0623	5.9409	0.0826	0.0213	0.0007	0.0695	0.0009	13946	4.0942	0.0517	4.9681	0.0278	2.9999	0.1013	8.0311	0.1029	0.0019	5.91E-05	0.0027	3.16E-05
Evenly spaced 50 Kb window	6867	2.0388	0.0610	6.0309	0.0830	0.0222	0.0006	0.0696	0.0009	7619	4.3594	0.1068	4.8733	0.0482	3.1923	0.0922	7.9832	0.0956	0.0020	4.76E-05	0.0026	2.61E-05
Evenly spaced 100 Kb window	4634	2.0373	0.0383	6.0313	0.0677	0.0220	0.0006	0.0694	0.0009	4846	4.2056	0.0729	4.9322	0.0361	3.2086	0.0761	7.9658	0.0934	0.0021	5.09E-05	0.0026	2.43E-05
Evenly spaced 250 Kb window	2281	2.1190	0.0332	6.0096	0.0711	0.0238	0.0006	0.0683	0.0007	2297	3.3373	0.0748	5.5842	0.0374	3.0333	0.0644	8.1622	0.0767	0.0016	2.36E-05	0.0029	1.37E-05
Evenly spaced 500 Kb window	1203	1.7130	0.0531	6.3825	0.0752	0.0182	0.0004	0.0733	0.0008	1204	2.9044	0.0463	6.0104	0.0355	2.5294	0.0441	8.6552	0.0689	0.0014	2.57E-05	0.0031	2.31E-05
Evenly spaced 1000 Kb window	610	1.5849	0.0220	6.4843	0.0562	0.0163	0.0004	0.0751	0.0006	609	2.2280	0.0388	6.4598	0.0291	2.2512	0.0443	8.8326	0.0741	0.0012	1.99E-05	0.0033	1.62E-05
Genic	7254	2.0011	0.0659	5.9837	0.0819	0.0218	0.0006	0.0689	0.0008	11212	3.7492	0.0891	5.1470	0.0320	2.9664	0.0827	8.0692	0.0973	0.0018	2.66E-05	0.0027	1.74E-05

Intergenic	6533	2.0532	0.0677	6.0605	0.0879	0.0212	0.0005	0.0704	0.0009	8294	4.0981	0.0901	5.0289	0.0410	3.2237	0.0969	7.9490	0.1067	0.0019	3.67E-05	0.0027	2.23E-05
SNPs in LE (LD-pruning)	10460	2.2081	0.0853	5.8599	0.1018	0.0234	0.0010	0.0678	0.0010	10984	3.8135	0.1153	5.1519	0.0505	3.0485	0.0827	8.0305	0.0888	0.0018	3.19E-05	0.0027	2.05E-05
Chr1	848	1.3302	0.0282	6.8907	0.0769	0.0149	0.0002	0.0781	0.0008	1329	2.1694	0.0411	6.8599	0.0276	2.5436	0.0723	8.8585	0.0821	0.0011	1.90E-05	0.0035	1.94E-05
Chr2	1672	1.5457	0.0286	6.7510	0.0607	0.0171	0.0005	0.0765	0.0007	2245	2.0599	0.0437	6.9758	0.0370	2.1244	0.0518	9.1586	0.0736	0.0010	1.82E-05	0.0036	1.85E-05
Chr3	1544	1.6015	0.0381	6.6048	0.0546	0.0193	0.0005	0.0738	0.0007	2026	2.5603	0.0729	6.5060	0.0537	1.9307	0.0358	9.3269	0.0842	0.0012	2.81E-05	0.0034	2.75E-05
Chr4	886	1.4908	0.0307	6.7958	0.0813	0.0161	0.0003	0.0777	0.0009	1303	2.3990	0.0825	6.9470	0.0576	2.3523	0.0317	9.2179	0.0759	0.0012	5.09E-05	0.0036	3.07E-05
Chr5	1356	1.6043	0.0372	6.6333	0.0760	0.0176	0.0004	0.0753	0.0008	1872	2.8370	0.0624	6.5223	0.0474	2.5983	0.0757	8.8292	0.0925	0.0013	3.38E-05	0.0034	1.91E-05
Chr6	1440	1.3569	0.0368	6.8014	0.0578	0.0145	0.0002	0.0778	0.0007	2012	2.4812	0.0726	6.4731	0.0560	2.2279	0.0409	9.0616	0.0764	0.0013	4.01E-05	0.0033	3.06E-05
Chr7	1207	1.8093	0.0528	6.4344	0.0683	0.0196	0.0006	0.0737	0.0008	1594	2.0342	0.0272	6.9727	0.0304	1.9146	0.0342	9.4784	0.0785	0.0010	1.27E-05	0.0036	2.06E-05
Chr8	1771	1.4947	0.0425	6.6755	0.0772	0.0156	0.0003	0.0771	0.0009	2583	1.8972	0.0549	7.0279	0.0477	2.0875	0.0368	9.2127	0.0768	0.0010	2.25E-05	0.0035	2.00E-05
Chr9	940	1.4100	0.0330	6.9016	0.0698	0.0154	0.0003	0.0787	0.0009	1381	2.0620	0.0471	6.9944	0.0265	2.0567	0.0404	9.2584	0.0758	0.0011	2.96E-05	0.0035	1.76E-05
Chr10	1034	1.2504	0.0219	6.9511	0.0658	0.0140	0.0002	0.0790	0.0008	1448	1.9068	0.0299	6.8570	0.0322	2.0679	0.0428	9.1610	0.0793	0.0010	1.73E-05	0.0035	1.94E-05
Chr11	1089	1.6034	0.0323	6.6413	0.0765	0.0179	0.0005	0.0752	0.0009	1713	2.2804	0.0583	6.8153	0.0199	2.2902	0.0556	9.0925	0.0840	0.0012	2.98E-05	0.0035	2.53E-05

Diameter at Breast Height (DBH, cm), Total Height (HT, m) and Wood Volume (WV, m3), SE (Standard Error). NA - The standard error of the variance components could not be estimated.

Table SM1-4: Predictive ability of growth traits of different 10-fold cross-validation using Bayesian Ridge-Regression (BRR) models in *E. benthamii* and *E. pellita* populations.

BRR r_{gy}	*E. benthamii*		*E. pellita*		
Fold	**DBH**	**WV**	**DBH**	**HT**	**WV**
fold1	0.044	-0.015	0.482	0.356	0.506
fold2	0.251	0.294	0.431	0.448	0.444
fold3	0.415	0.335	0.550	0.540	0.536
fold4	0.180	0.195	0.358	0.236	0.309
fold5	0.205	0.142	0.416	0.421	0.395
fold6	0.199	0.196	0.502	0.355	0.493
fold7	0.012	0.046	0.373	0.070	0.263
fold8	0.268	0.156	0.397	0.302	0.382
fold9	0.104	0.164	0.459	0.427	0.432
fold10	-0.058	-0.056	0.439	0.261	0.488
Mean	0.162	0.146	0.441	0.342	0.425
SE	0.044	0.039	0.019	0.042	0.028

Table SM1-5: Significant SNP associations with wood volume trait in *E. pellita* using MLMA model adjusted for block, population structure covariates and genomic relationship matrix.

SNP*	MAF	*P*-value	Bonferroni[a]	FDR[b]	Annotation[c]	Description[c]	Function[d]
EuBR06s46273728	0.00410	1.91E-06	0.0373	0.0303	Eucgr.F03806.1 / AT1G77460.1	Armadillo/beta-catenin-like repeat; C2 calcium/lipid-binding domain (CaLB) protein.	Plant-type cell wall cellulose biosynthetic process and unidimensional cell growth.
EuBR06s47094282	0.00274	3.11E-06	0.0606	0.0303	Orysa|LOC_Os06g43560.1_GX6P	Phox domain-containing protein, putative, expressed LOC_Os06g43560.1.	Phosphatidylinositol binding.
EuBR05s10629849	0.00478	7.07E-06	0.1379	0.0430	Eucgr.E01008.1 / AT1G11050.1	Non-specific serine/threonine protein kinase / Threonine-specific protein kinase or Protein kinase superfamily protein.	Catalytic activity (ATP + a protein = ADP + a phosphoprotein).
EuBR09s37575166	0.00342	1.10E-05	0.2149	0.0430	Eucgr.I02627.1 / AT1G32750.1	Transcription initiation factor TFIID | HAC13 protein (HAC13).	Chromatin modification, DNA mediated transformation, regulation of transcription, DNA-templated.
EuBR09s38076481	0.00342	1.10E-05	0.2149	0.0430	Eucgr.I02694.1 / AT3G19720.1	Dynamin-like protein ARC5 | P-loop containing nucleoside triphosphate hydrolases superfamily protein.	Catalytic activity (GTP + H2O = GDP + phosphate).

*SNP name: e.g. EuBR06s46273728, SNP on chromosome 6 at position 46,273,728 bp

[a]Bonferroni-corrected threshold with an experimental type I error rate at α = 0.05.

[b]False Discovery Rate (FDR) threshold at 5%.

[c]Annotation information based on BLASTx in Phytozome for the *Eucalyptus grandis* genome (Myburg et al. 2014).

[d]Function information based on UniProt database.

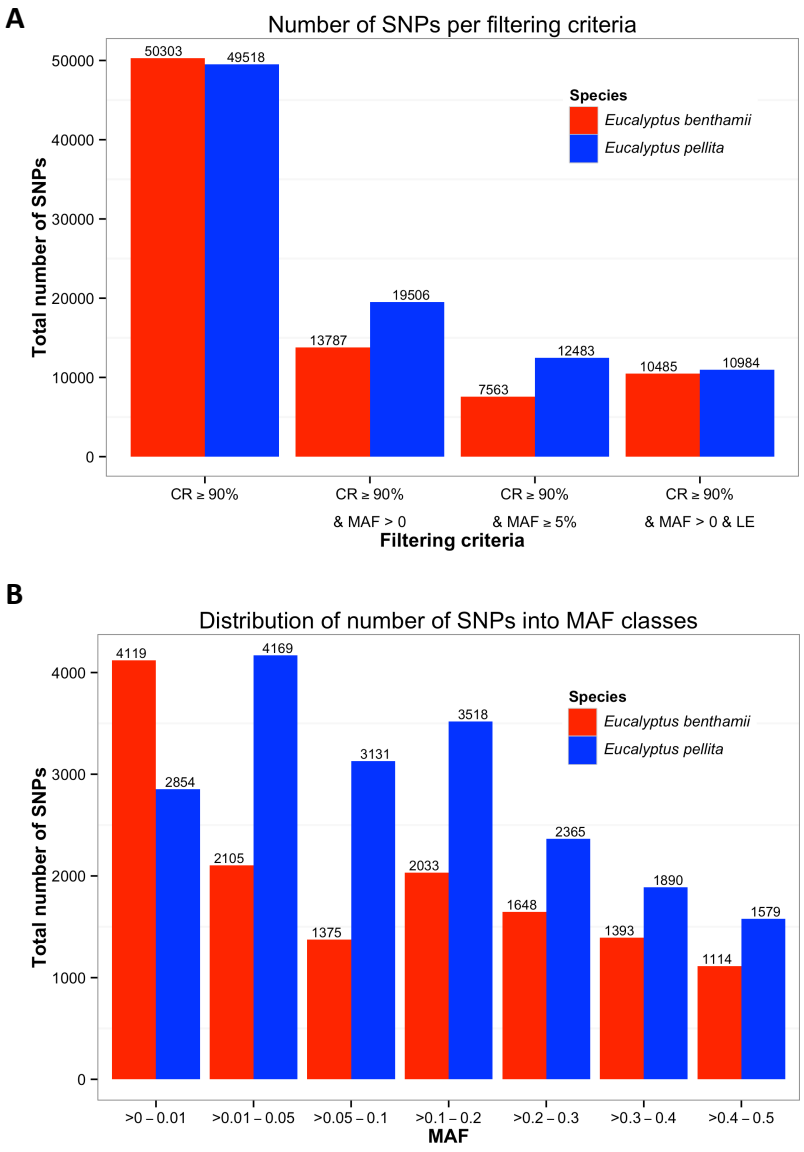


Figure SM1-1: Distribution of the numbers of SNPs for variable filtering criteria and MAFs classes. (A) Distribution of the number of SNPs retained in variable filtering criteria, and (B) Distribution of the number of SNPs into MAF classes for *E. benthamii* and *E. pellita* for CR ≥ 90% and MAF > 0 (CR, Call Rate; MAF, Minimum Allele Frequency; LE, Linkage Equilibrium).

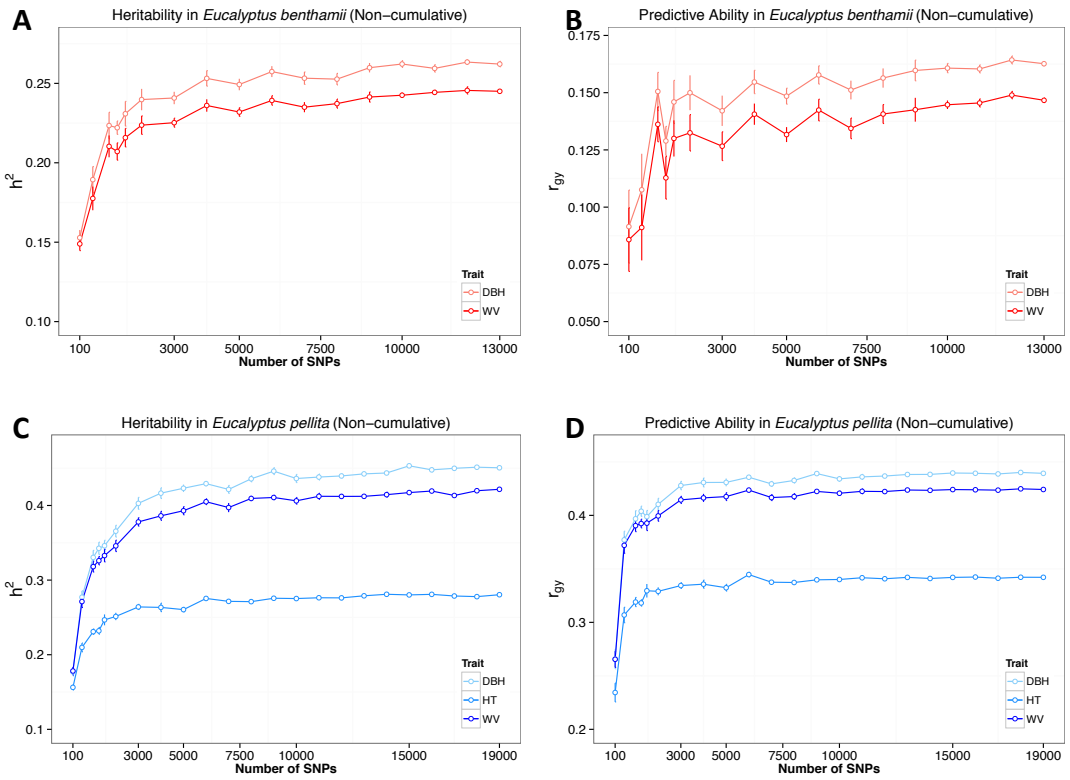


Figure SM1-2: Estimates of heritability (h^2) and of predictive ability (r_{gy}) with increasing numbers of SNPs for different traits using a non-cumulative approach to SNP sampling. (A) and (B) estimates of h^2 and r_{gy} for *E. benthamii*, respectively. (C) and (D) estimates of h^2 and r_{gy} for *E. pellita*, respectively.

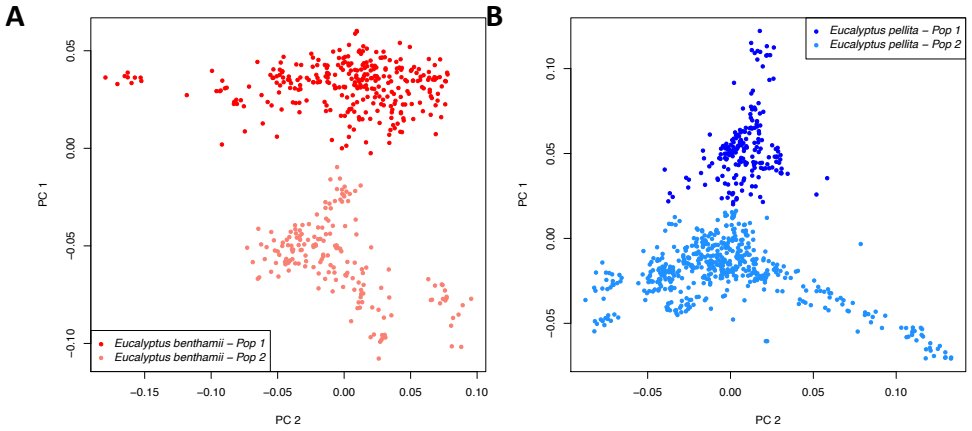


Figure SM1-3: Principal component analysis (PCA) of the 484 trees of *E. benthamii* and 706 trees of *E. pellita* used to split training and validation sets. (A) For *E. benthamii* 310 (red) and 174 (pink) individuals were used as training and validation sets. (B) In *E. pellita*, the number of individuals used in each set were 192 (dark blue) and 514 (light blue).

CHAPTER 2: A GWAS for growth traits in *Eucalyptus* by assembling genome-wide data for 3,373 individuals across four breeding populations

To be published with the following authors:

Bárbara S. F. Müller, Janeo E. de Almeida Filho, Bruno M. Lima, Carla C. Garcia, Alexandre Missiaggia, Aurelio M. Aguiar, Elizabete Takahashi, Matias Kirst, Salvador A. Gezan, Orzenil B. Silva-Junior, Leandro G. Neves, Dario Grattapaglia.

INTRODUCTION

Eucalyptus are the most widely planted species of hardwood trees in the world mainly due to its high adaptability to different environments, fast growth and superior wood quality for multiple applications (Myburg et al. 2007; Grattapaglia and Kirst 2008). The *Eucalyptus* L'Hér. (Myrtaceae) genus has over 800 species native to Australia and adjacent islands in Oceania. This genus has a total of ten subgenera described, being *Symphyomyrtus* the most important one with more than 470 species (Grattapaglia et al. 2012). *Eucalyptus grandis* and *Eucalyptus urophylla* are the most commercially important and broadly planted species, which belong to this subgenus in the *Latoangulatae* section and, together with their hybrids, are widely used for pulp and solid wood production in the tropics (Henry 2011). Interspecific hybrids between *E. grandis* x *E. urophylla* make up almost the totality of large scale operational plantations and are the main target of breeding programs in Brazil due to their combination of desirable traits, most notably fast growth from *E. grandis* and disease resistance from *E. urophylla* (Myburg et al. 2007).

In forest trees, following the disappointing translation of results coming from QTL mapping in biparental populations to breeding programs, genetic association studies were proposed based on an optimistic view that praised forest tree populations as ideal for such undertakings given their low extent of linkage disequilibrium (LD), lack of structure and high diversity (Neale and Savolainen 2004). Initial association studies, largely in species of *Populus* and *Pinus*, focused on variation in candidate genes, mainly due to a

lack of genome-wide genotyping platforms, but also motivated by the assumption that complex traits would be driven by some moderate-effect loci (Neale and Savolainen 2004; Thumma et al. 2005; Neale 2007; Wegrzyn et al. 2010; Khan and Korban 2012; Guerra et al. 2013; Thavamanikumar et al. 2014; Jaramillo-Correa et al. 2015). Besides the limited scope of candidate-gene association studies, results in general were limited to the detection of a few associations explaining small proportions of the genetic variation. In the last few years, however, with the development of accessible high-density SNP genotyping platforms, actual genome-wide association studies (GWAS) have been 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). Although several traits have been targeted by these studies, a common trend is that while better success in terms of the number of associations detected has been obtained for phenology and wood properties, very few associations have been found for complex growth traits. Still, the proportion of genetic variation explained has been quite limited.

A common feature of GWAS in forest trees has been the use of collections of trees directly sampled from the wild with the understanding that the rapid decay of LD would provide the necessary resolution to discover causal variants. Such variants, it is believed, could in turn be used in breeding programs via marker-assisted selection (MAS). Although such a rationale could pinpoint loci underlying complex traits, such loci or specific alleles detected in wild populations could perform differently or have relatively little or no value in an elite background. These loci could simply not be segregating or alleles could have negligible effect in comparison to existing allelic variation in selected breeding populations. In fact, GWAS can be carried out with natural populations, germplasm banks, landraces or breeding populations (Zhu et al. 2008; Khan and Korban 2012), as long as there is segregation for the relevant phenotypes under study. This issue has motivated for example the development of GWAS in nested association mapping (NAM) populations in maize (Li et al. 2016; Wu et al. 2016). This strategy in effect 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 at least in principle less genetic

variation is available in such structured populations, and the longer extent of LD limits resolution to pinpoint causal variants, associations detected in genetically improved material should be considerably more relevant to further breeding as they would be detected in an already elite background. Moreover, different from crop breeding where introgression of wild alleles into elite lines is commonplace, such a route is not an option in highly heterozygous forest trees.

Based on the reasoning that associations detected in breeding populations could be more useful to tree breeding, we have recently reported results from GWAS studies in breeding populations of *Eucalyptus* (Resende R.T. *et al.*, 2016; Müller *et al.*, 2017). These studies became possible with the development of a high-density SNP platform for species of *Eucalyptus* (Silva-Junior *et al.*, 2015) providing genome-wide coverage of one marker per 12-20 Kbp, and 47,069 SNPs located inside or within 10 Kbp of 30,444 out of the 36,349 predicted genes in the *E. grandis* genome (Myburg et al. 2014). Interestingly, such GWAS carried out in breeding populations reported essentially equivalent results to those described in natural populations for growth traits, when only a few associations were detected explaining small fractions of the genetic variation.

The statistical power to detect associations between DNA variants and a trait depends on the experimental sample size, the unknown distribution of effect sizes, the frequency of causal genetic variants segregating in the population, and the LD between genotyped DNA variants and the unknown causal variants (Visscher *et al.*, 2017). A clear movement in the direction of increasing sample sizes has taken place with GWAS for human traits to provide sufficient statistical power to detect the common and low-frequent variants (Marouli et al. 2017; Visscher et al. 2017). In plants, however, still rare are GWAS carried out in populations larger than a few hundred individuals with the exception of studies using the structured maize NAM populations where around 5,000 individuals have been used (reviewed by Xiao *et al.*, 2017). No such large experiments have yet been described in any forest tree. Besides increasing the specific sample size of a single GWAS experiment, statistical power can be increased by combining information coming from multiple populations using Meta-GWAS and Joint-GWAS (Mägi and Morris 2010; Yang

et al. 2012; Bernal Rubio 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 and Ioannidis 2013). Joint-GWAS combine the populations prior to the association analysis, leading to more resolution and the detection of more associations for complex traits (Lin and Zeng 2009). As each experiment is independently designed, both methods have to account for the heterogeneity created by population structure, phenotype measurements among other potential sources of variability (Magosi et al. 2017). Although sharing individual-level datasets is logistically difficult and for human studies might have ethical restrictions, Joint-GWAS have become more common in plant research due to the ability to replicate genotypes (Li et al. 2016; Wallace et al. 2016; Wu et al. 2016).

A second way to increase the power of a GWAS is to capture a wider frequency spectrum of variants. 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) between individuals, and it 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 in enriched genic regions (Schork et al. 2013), it would be 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 larger association population from individual *Eucalyptus* breeding populations. We leveraged the portability and power of the multi-species SNP genotyping platform for

Eucalyptus to assemble genome-wide SNP and growth trait data for 3,373 trees across four unrelated *E. grandis* x *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 regional heritability mapping in the four populations independently to pinpoint regions that would capture larger fractions of the additive genetic variance by considering common and rare variants at the same time. Association analyses by genes and regions were performed from summary data from Joint-GWAS to increase the power to detect trait associations. Several associations were detected for the same SNPs across the unrelated populations providing some initial validation. Associations were also detected into genes related to cell wall growth and disease resistance suggesting potential pleiotropic effects. To the best of our knowledge, this is the first study to apply Joint-GWAS in a forest tree.

MATERIAL AND METHODS

Populations and phenotypic data

This study was carried out using progeny trials established in four unrelated *E. grandis* x *E. urophylla* hybrid breeding populations (Pop1-IPB, Pop2-ARAB, Pop3-ARAC and Pop4-CNB), belonging to three different Brazilian paper and pulp companies, International Paper of Brazil, Fibria Celulose and Cenibra Celulose. Details of the size of the trial, experimental design, number of families, age of measurement, location and sample sizes used in the GWAS are listed in the Table 2-1. 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). While populations Pop1-IPB, Pop2-ARAB and Pop3-ARAC were largely composed of first generation hybrids, Pop4-CNB went through one more selection cycle being equivalent to an outbred F2, as the parents were themselves hybrids (F1) between *E. grandis* x *E. urophylla*. All trees were ultimately phenotyped at age two to five years for diameter at breast height (DBH, cm) and total height (HT, m) (Supplementary Material: Fig. SM2-1).

SNP genotyping and quality control

A total of 3,417 trees were genotyped using the *Eucalyptus* Illumina Infinium EUChip60K (Silva-Junior et al. 2015) and 44 individuals were removed with more than 10% missing data, remaining 3,373 samples for the further analyses. A combined dataset was generated by merging the genotyping datasets 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 keeping rare SNPs with minor allele frequency (MAF) > 0 in the analyses (full marker dataset). Two alternative SNP datasets were also generated by keeping only SNPs with MAF ≥ 0.01 and MAF ≥ 0.05 (Table 2-2). 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 (Turner 2014), for linkage disequilibrium to generate a pruned subset of SNPs in approximate linkage equilibrium (LE), subsequently used in the population stratification analyses. The LD-based SNP pruning method was applied with a window size of 100 Kbp, shifting the window by one SNP at the end of each step and removing one SNP from a pair of SNPs if LD was greater than 0.2 (plink command: --indep-pairwise 100 kb 1 0.2).

Population stratification analyses

The genetic structure for the four populations and combined data was estimated based on a Bayesian clustering method implemented by STRUCTURE v2.2.4 (Pritchard et al. 2000), using the intergenic SNPs of the full marker dataset in approximate LE. The individual structures were classified in *K* clusters according to their genetic similarity. 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 MCMC (Markov chain Monte Carlo) 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) using the program STRUCTURE HARVESTER v0.6.93 (Earl and vonHoldt 2012). The software

CLUMPP v1.1.2 (Jakobsson and Rosenberg 2007) was used to find consensus among the 10 most probable *K* interactions. The POPHELPER R package (Francis 2016) was used to generate the population structure barplots by individuals. Principal component analysis (PCA) was performed using SNPRelate R package (Zheng et al. 2012) to plot all individuals for the combined dataset. For the population stratification correction in the GWAS models, we performed the PCA using GCTA v1.26.0 (Yang et al. 2011a) 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 (F_{ST}) were estimated between populations (Weir and Cockerham 1984) using SNPRelate (Zheng et al. 2012).

Linkage disequilibrium and effective population size estimation

Genome-wide pairwise estimates of linkage disequilibrium 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 (Turner 2014). The LD decay of r^2 with distance in Kbp was fitted by a nonlinear regression model between adjacent sites. The drift-recombination model (Hill and Weir 1988) was used to fit a nonlinear regression of the expectation of r^2, using the R script by Marroni *et al.* (2011) and the equation based on 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. Effective population size (N_e) was estimated based on an updated version of the heterozygote excess method (Zhdanova and Pudovkin 2008) implemented in NeEstimator software v2.01 (Do et al. 2014) for each population using datasets including rare alleles (CR> 90% and MAF>0). Confidence intervals for these estimates were obtained using the parametric method in NeEstimator, where the number of independent alleles is used as the degree of freedom in a chi-square distribution.

Statistical analysis of phenotypic data

Growth data for height (HT) and diameter at breast height (DBH) for each individual genotyped were used in the analysis. The following mixed linear models were implemented in the "lmer" function of the R package lme4 v1.1-13 (Bates et al. 2015) to obtain an accurate trait estimate for each individual, while accounting for experimental effects, by population:

1. Pop1-IPB - Randomized complete block design (RCBD) with eight blocks using six trees per plot:

$$y_{ij} = \mu + B_i + p_{j(i)} + e_{ij} \qquad (1)$$

where y_{ij} is the phenotypic measure of the trait of the tree in the j^{th}plot within the i^{th} block; μ is the intercept; B_i is the fixed effect of blocks; $p_{j(i)}$ is the random effect of plot within block and e_{ij} is the random residual effect.

2. Pop2-ARAB - Alpha lattice design using incomplete block (ALD-IB) with 30 blocks in single-tree plots; and Pop3-ARAC - ALD-IB with 40 blocks in single-tree plots:

$$y_{ij} = \mu + r_i + b_{j(i)} + e_{ij} \qquad (2)$$

where y_{ij} is the phenotypic measure of the trait of the tree in the j^{th}block within the i^{th} repetition; μ is the intercept; r_i is the random effect of repetition; $b_{j(i)}$ is the random effect of blocks nested within repetition and e_{ij} is the random residual effect.

3. Pop4-CNB - Randomized complete block design (RCBD) with 36 blocks in single-tree plots:

$$y_i = \mu + B_i + e_i \qquad (3)$$

where y_i is the phenotypic measure of the trait of the tree in the i^{th} block; μ is the intercept; B_i is the fixed effect of blocks and e_i is the random residual effect.

4. Combined data - For the combined dataset, we used the phenotype adjusted for each population independently (as estimated above) and included a source of variation for population and for age of phenotyping directly in the GWAS models as covariates with the following mixed linear model:

$$y_{ij} = \mu + Pop_i + A_j + e_{ij} \qquad (4)$$

where y_{ij} is the phenotypic measure of the trait of the tree over the j^{th} age of measurement within the i^{th} each population; μ is the intercept; Pop_i is the fixed effect of population (discrete covariate); A_j is the fixed effect of age of phenotyping (quantitative covariate) within population and e_{ij} is the random residual effect.

Heritability estimation

The variances components, genomic and pedigree-based heritabilities for each trait were estimated using the adjusted phenotype data for each population separately using the BGLR v1.0.5 R package (Pérez and de los Campos 2014) with the following mixed linear model:

$$y = \mu + a + e \qquad (5)$$

where y is the vector of adjusted phenotypic values of the trait being analyzed; μ is the intercept; a is a vector of random additive effects and e is the random residual effect; $e \sim N(0, I\sigma_e^2)$; $cov(a, e') = \mathbf{0}$. The variance structure of the pedigree-based model was calculated with $a \sim N(0, A\sigma_a^2)$ and genomic-based model with $a \sim N(0, G\sigma_a^2)$; where A is a matrix of additive genetic relationships among the individuals calculated with SYNBREED v0.12-6 R package (Wimmer et al. 2012) and G is a genomic relationship matrix (GRM) estimated using Yang's method (Yang et al. 2010) with GCTA v1.26.0 (Yang et al. 2011a).

The narrow-sense heritability (h^2) estimates were calculated as the ratio of the additive variance σ_a^2 to the phenotypic variance σ_y^2 ($h^2 = \sigma_a^2/\sigma_y^2$).

GWAS models

For the following GWAS analyses, we used the adjusted phenotypic data for each population separately and for the combined data we corrected these adjusted phenotypes for age and population as described above. Three different GWAS approaches were tested to detect associations: single SNP-based models, regional heritability mapping (RHM) and SNP set-based models.

Single SNP-based models. Six distinct GWAS models were implemented using the EMMAX software (Kang et al. 2010) with the full marker dataset (CR> 90% and MAF>0): (1) A linear model based association (LMA) analysis was fitted independently for each SNP, without any correction for population stratification and relatedness:

$$y = Xb + e \qquad (6)$$

where y is the phenotype; b is a vector of fixed effects including intercept and the SNP candidate to be tested for association; X is incidence matrix for the vectors for the parameters b and e is the random residual effect.

(2) A LMA with the Q-matrix from STRUCTURE was fitted using the number of genetic groups determined based on the criteria proposed by Evanno *et al.*, (2005). In equation 6, the parameter b is a vector of fixed effects with addition of population structure corrected by STRUCTURE. Since each population had different subpopulations (*K* = 2 to 5, Table 2-2), the covariate was computed in the model.

(3) A LMA with significant principal components (PCs) determined by a broken stick model (Jackson 1993). In equation 6, the parameter b is a vector of fixed effects with addition of population stratification corrected by these significant PCs (PC = 1-3, Table 2-2).

(4) For comparisons with the LMA models, we also tested a mixed linear model based

association (MLMA) analysis. This association analysis was fitted using the following base model that uses a similar model as LMA (equation 6), except for the inclusion of the polygenic effect (g):

$$y = Xb + g + e \qquad (7)$$

where g is the polygenic effect (random effect) captured by the genomic relationship matrix (GRM) calculated using all SNPs. The covariate associated with the SNP take the values of the number of copies of the alternative allele (2, 1 or 0). The variance structure of the MLMA model were $g \sim N(0, G\sigma_g^2)$; $e \sim N(0, I\sigma_e^2)$; $cov(g, e') = \mathbf{0}$, where G is the GRM between individuals calculated using Balding-Nichols (BN) matrix (Kang et al. 2010) and I is the identity matrix.

(5) A MLMA including the GRM and Q-matrix. In the equation 7, the parameter b is a vector of fixed effects with addition of population structure corrected by STRUCTURE.

(6) A MLMA with GRM and significant PCs. In the equation 7, the parameter b is a vector of fixed effects with addition of the significant PCs.

All these GWAS models were performed for each population independently and for the combined data (Joint-GWAS). For the Joint-GWAS analysis, the fixed effects for the combined data were age of measurements (2-5 years) and population of origin (Pop1, Pop2, Pop3, Pop4) included as covariates, as described in equation 4.

Regional heritability mapping (RHM). The RHM method was applied to each population independently. This method divides the genome into windows of pre-determined 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 mixed linear model was fitted using GCTA (Yang et al. 2011a), that

uses an analogous model as MLMA (equation 7), except for the inclusion of the regional genomic additive effects (r):

$$y = Xb + g + r + e \qquad (8)$$

where b is a vector of fixed effects including intercept and population structure (from STRUCTURE); r is the vector of regional genomic additive effects (random effect) captured by the GRMs calculated using SNPs within each region (window). The variance structure of the RHM model were $g \sim N(0, G\sigma_g^2)$; $r \sim N(0, G_r\sigma_r^2)$; $e \sim N(0, I\sigma_e^2)$; $cov(g, r') = \mathbf{0}$; $cov(g, e') = \mathbf{0}$; $cov(r, e') = \mathbf{0}$, where G is the whole-genome relationship matrix between individuals calculated as described by Yang *et al.*, (2010) algorithm; G_r is the regional relationship matrix using the same algorithm for each window and I is the identity matrix. The whole-genomic and regional heritabilies were estimated as ${h_g}^2 = \sigma_g^2/\sigma_y^2$ and ${h_r}^2 = \sigma_r^2/\sigma_y^2$, respectively. To test for the presence of regional variance (σ_r^2) using the RHM method, a likelihood ratio test ($LRT = -2\ln(L_0/L_1)$) was used to compare a model fitting variance in a specific window (fitting both whole-genome and regional additive variance) against the null hypothesis of no variance in that window (whole-genome additive variance only); where L_0 and L_1 are the likelihood values for the hypothesis of the absence ($H_0: \sigma_r^2 = 0$) or presence ($H_1: \sigma_r^2 > 0$) of regional variance.

SNP Set-based models. Knowing that the effect sizes of individual genetic variants potentially detected by single SNP-based models are very small we tested the aggregated effect of a set of SNPs using a set-based association approach by gene or by segment with the goal to increase the power of the GWAS. Two set-based methods were used: (1) a gene-based model with fastBAT (Bakshi et al. 2016) in GCTA (Yang et al. 2011a) using summary data from the previous GWAS (SNP-based models) for the combined dataset (Joint-GWAS). The gene-based model was fitted using SNPs within 50 Kbp from the UTRs of an annotated gene in the *Eucalyptus* genome (Myburg et al. 2014); and (2) a second set-based GWAS model for the combined dataset (Joint-GWAS), called segment-based model (or region-based model). This segment-based model was implemented using the same list of association *P*-values (summary data) used in the gene-based

models, but was performed using fastBAT analysis (Yang et al. 2011a; Bakshi et al. 2016) based on segments of 100 Kbp.

To select significant associations, different multiple test corrections were applied to the *p*-values obtained. The Bonferroni procedure was implemented to control for type I error at $\alpha = 0.05$ and the Benjamini & Hochberg (1995) procedure was used to control for false discovery rate (FDR) at 5% on SNP-based GWAS models. A third less stringent *ad hoc* threshold of $-\text{Log}_{10}(P) \geq 4$, was also used to declare additional significant associations that did not survive using 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 R.T. et al. 2016). This threshold value is more stringent than the one ($-\text{Log}_{10} P \geq 3.5$) reported in a recent soybean study (Kaler et al. 2017) using a comparable number of SNPs (31,260) and similar to the *ad hoc* threshold (α = 1/effective marker number) considered in a *Populus nigra* study (Allwright et al. 2016). For the RHM approach, to account for the overlapping windows 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 ($-\text{Log}_{10} P \geq 2$) was tested to declare significant RHM windows associated with growth traits. 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. The Manhattan plots were generated using the qqman (Turner 2014) and the ggplot2 R packages (Wickham 2009).

RESULTS

SNP genotyping and population stratification

A total of 59,222 SNPs originally converted in the EUChip60K chip (Silva-Junior et al. 2015) were targeted for genotyping. More than 46,000 SNPs were retained for all populations following a filter for call rate (CR) ≥ 90%, and 51,274 SNPs for the combined

dataset. After removing monomorphic markers (MAF > 0), around 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%. After filtering for call rate and MAF, the distribution of the number of SNPs into MAF classes showed enrichment for low frequency alleles (MAF 0–0.1, Fig. SM2-2). 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 7,000 to 14,000 SNPs in intergenic regions and in approximate linkage equilibrium were generated for the population structure analyses (Table 2-2). Population structure analyses with these SNP sets revealed that the most likely numbers of subpopulations varied between $K = 2$ for Pop1-IPB and Pop2-ARAB to $K = 5$ for Pop4-CNB (Table 2-2). When all four populations were combined the most likely number of subpopulations was $K = 2$ (Fig. 2-1A), using no previous population information in the admixture model. The ancestry coefficient barplots from STRUCTURE showed K ranging from two to four subpopulations in the combined dataset (Fig. 2-1A). For $K = 2$ only Pop4-CNB was clearly different 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 company (Fibria) and have therefore some parents in common. When $K = 4$ the combined dataset was subdivided in 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 2-2) 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% of the genetic variance (Fig. 2-1B and 2-1C).

Linkage disequilibrium, effective population size, genomic and pedigree-estimated heritabilities

The pairwise estimates of LD (r^2) were calculated for all pairwise distances among the high-quality polymorphic SNPs (MAF > 0) on each chromosome separately for the four populations independently. The genome-wide LD average 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 (75.1Kbp) compared to that of Pop4-CNB (637.7 Kbp) (Fig. 2-2). The more extensive LD on Pop4-CNB may be explained by the more advanced selection state of this population when compared with the others. Estimated effective populations sizes based on the heterozygote excess method (N_e) were 6.2 for Pop1-IPB, 19.3 for Pop2-ARAB, 11 for Pop3-ARAC and 12 for Pop4-CNB (Table 2-1), suggesting that although a larger number of parents originated the individuals of these *Eucalyptus* breeding groups some more ancestral level of relatedness between such parents exists. 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 of genomic heritabilities varied from 0.296 for HT in Pop4-CNB to 0.528 for DBH in Pop3-ARAC, accounting for a large proportion (64–89%) of the pedigree-based heritabilities (Table SM2-1). Estimates of variance components are also reported in Table SM2-1.

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 them likely spurious given the structured nature of these breeding populations (Table 2-3). For instance, in Model 1 (no correction) there were hundreds to thousands of SNPs associated with growth traits for all comparisons. When the population stratification covariate obtained either by STRUCTURE or PCA 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 likely inappropriateness 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. SM2-3).

When the random effects captured by the kinship matrix (GRM) 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 following the correction for multiple testing (Table 2-3 and Fig. SM2-4). The same was observed for total height, except for Pop4-CNB where several significant associations (Fig. 2-3 and Fig. SM2-4D: B/D) were detected using a FDR (p <0.05) threshold (Table 2-3). All these significant associations detected by single-SNP GWAS for Pop4-CNB are 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 2-3). 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. SM2-3). Furthermore, the models built with GRM produced a drastic reduction in the number of significant markers, showing 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 2-3). For HT, no significant association was found after inclusion of the GRM in the model (Table 2-3 and Fig. SM2-4E: B/D). Although traditional multiple testing thresholds (FDR and Bonferroni) are important to control for type I error (false positive), they tend to be moderate to excessively stringent for GWAS, where several thousand markers are used, and a minority are expected to be associated with phenotype. To address this potential problem, a less stringent *ad hoc* threshold ($-\text{Log}_{10} P \geq 4$) was used to declare additional

significant associations not detected before. With this threshold, eight variants putatively associated (*p*-value ≤ 0.00008) with DBH were found in the Joint-GWAS analysis (Fig. 2-4, green dashed line). Collectively, of these 11 SNPs associated with DBH using Joint-GWAS (three associations and eight putative associations) six are located into 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, where the most significant SNP (*p*-value 0.000006) is also the most significant one detected for DBH (EuBR07s38098526, Table 2-5) and located on chromosome 7. The third SNP (EuBR08s48262720) 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 2-3 and Fig. 2-3, red line). However, significant GWAS hits found in each population after correcting for both family and population structure were generally not shared across populations. To investigate further the sharing of associations across populations as a way to provide some independent validation for the associations found, results from Model 2 (Q-matrix from STRUCTURE) were used to create a comparison dataset. When the results for all populations were compared, four and six shared associations were detected for DBH and HT, respectively (Fig. 2-5). These results are 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). The small number of common associations

found among all populations, increases considerably, however, when Pop4-CNB is excluded from the analysis and comparisons are made only among Pop1-IPB, Pop2-ARAB and Pop3-ARAC. Under this scenario, 157 and 40 significant associations are shared for DBH and HT, respectively (Fig. 2-5). This considerable difference in results likely reflects the significant genetic differentiation found between Pop4-CNB and the other three populations in the structure analysis (Fig. 2-1A).

Regional heritability mapping (RHM)

RHM was performed for all populations to evaluate whether additional variants associated with the growth traits could be detected. For Pop1-IPB, Pop2-ARAB and Pop3-ARAC, no significant regions were declared significant with this approach using multiple testing correction. On the other hand, for Pop4-CNB, eight regions (each with 100 SNPs) were significantly associated with total height on chromosome 2 (Fig. 2-3 and Table 2-4) 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. 2-3). The most significant window (h_r^2 = 0.07) detected by RHM for HT in Pop4-CNB captured 24% of the genomic heritability (h_g^2 = 0.29). Altogether, each of the eight significant 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, twelve more are putatively associated considering the lower *ad hoc* threshold for RHM adopted ($-\text{Log}_{10}\ P \geq 2$), 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 a lower *ad hoc* threshold in Pop1-IPB, 12 widows 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_g^2 = 0.52). Still under this more liberal threshold, one association was declared 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 2-4).

Joint-GWAS from summary datasets

To assess further 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 as they contain SNPs located 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. 2-6B). When considering only the kinship matrix, without the correction for population structure, a peak on chromosome 9 was also considered significant, involving 15 genes. Other significant signals were detected at the suggestive level, with one gene associated with two close SNPs on chromosome 3 and another locus with five SNPs on chromosome 7 (Fig. 2-6A). The gain of power observed for this methodology is due to multiple small independent association signals at these loci analyzed, including rare and common alleles. For the segment-based Joint-GWAS (Fig. 2-6C-D), 4,766 segments of size 100 Kbp were tested, with four of those regions being associated with HT (Fig. 2-6C). 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 region-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.

DISCUSSION

This study further advances the investigation of discrete genomic regions controlling growth traits in forest trees in general and of *Eucalyptus* in particular. Significant

associations were detected for height and diameter growth with the increased power of Joint-GWAS experiments, which leveraged genome-wide data for 3,373 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 the target breeding populations. The single-SNP GWAS and RHM identified genomic regions associated with growth traits, especially for total height in Pop4-CNB, which was the population with more extensive LD (637.7 Kbp). Both approaches had a comparable profile identifying similar regions associated with growth traits (Fig. 2-3) and performed well when there was a strong evidence of association arising from alleles present at high frequency in the population.

Impact of population structure, LD and relatedness on GWAS

Over the years, different single-SNP MLMA models have been proposed for GWAS to account for population and family-based structure, involving stringent multiple comparisons, and using either population structure (Korte et al. 2012), PCA (Price et al. 2006) or relatedness (Yu et al. 2006). Our results correcting for either population structure (Model 2) and PCA (Model 3) yielded different results and differed more when the kinship matrix was added to the models (Models 5 and 6) (Fig. SM2-3 and SM2-4). We compared five different kinship matrices to account for relatedness, using GRMs built by different methodologies: (i) VanRaden 2008; (ii) IBS (identity by state) and (iii) Balding-Nichols (BN) using Kang *et al.* 2010; (iv) Powell *et al.* 2010 and (v) Yang *et al.* 2010. We found strong concordance between all these methods with no differences detected for the single-SNP GWAS analysis (results not shown). Similar results were also obtained in a comparative study (Eu-ahsunthornwattana et al. 2014) using different kinship matrices and software, concluding that the choice of MLMA model implementation cannot be based on power/type I error considerations, but must instead be based on user-friendliness and speed. The estimated effective population sizes of these breeding populations are much smaller than the number of parents used to generate the individuals

studied (N_e = 6.2-19.3). Notwithstanding the fact that such estimates of N_e might be downward biased, these results suggest the existence of considerable cryptic ancestral relatedness among the parent trees possibly due to the fact that they derive from common selected families and seed sources from the wild. This complex family-based structure present in these elite breeding populations is not surprising and it likely inflates the rate of false-positive associations, further challenging the detection of true associations. An order of magnitude higher extent of LD was observed for Pop4-CNB (637.7 Kbp, Fig. 2-2) when compared to the other three populations. This was not unexpected as Pop4-CNB comes from a second-generation hybrid breeding. Consistent with expectations, Pop4-CNB with a higher LD was the only population where associations were detected for HT at FDR = 5% (Table 2-3). In a recent GWAS performed in two breeding populations of different species of *Eucalyptus* (Müller et al. 2017), we also found more associations in the population displaying longer-range LD. Nevertheless, such an extent of LD at 637 Kbp while favorable for detection power will not allow any resolution to arrive to single genes and much less to causal variants.

Although the use of GRM to control for relatedness is important to remove false associations (Astle and Balding 2009; Speed and Balding 2014), it can also be considered a very stringent process that might exclude some bona fide associations (Table 2-3). Several studies can be found in the literature where naïve association models were employed that did not account for family-based or even for population structure factors. A multi-gene association mapping using 435 unrelated individuals of *Populus* detected more than 400 significant associations for growth traits without any correction for population structure (Du et al. 2016). Despite the fact that this latter study used a sample from the wild, effects of population structure can still be present and need to be accounted for to minimize bias due to past relatedness in the evolutionary history of the species or even more recent one due to family structure. A GWA study performed in *Populus nigra* showed a strong decrease in the number of associations declared, especially for growth trait, when population structure and/or family-based correction were incorporated in the model, even for a natural population of 714 individuals (Allwright et al. 2016). In our experiment, the large number of SNPs detected without correction of GRM suggests that

association signals are confounded with family-based structure (Table 2-3). However, it can be expected that some of the associations are indeed true and overcorrection of type I error may result in considerable type II errors. Since complex growth traits in *Eucalyptus* have been shown to most likely follow an infinitesimal model (Müller et al. 2017), it is expected that a large number of SNPs will be needed to explain a large proportion of trait heritability. In humans, a recent study estimated that more than 100,000 SNPs influence height (Boyle et al. 2017). We believe that total height in *Eucalyptus* will be no different. To apply some level of biological stringency over the thousands of associations obtained in Model 2 (controlling only for population structure) in each population, we carried out a comparative analysis and identified associations that are shared among multiple populations (Fig. 2-5). Despite the fact that these results were obtained with no control for relatedness with the GRM, we consider these results particularly important and novel as they constitute a form of independent validation of associations across populations. These four populations and their experimental settings are truly independent in time and space and moreover subject to genotype by environment interactions that could not be accounted for in the model as we had no common check trees across the experiments. Still we did find a large number of associations shared across populations, particularly when the distinct population Pop4-CNB was left out of the analysis. Following this validation approach, several interesting genes were identified with strong indication of underlying the phenotype (see below).

Associations for growth traits in forest trees

Various studies attempted a GWAS for growth traits in forest trees, mainly in *Populus* (Porth et al. 2013; Allwright et al. 2016; Du et al. 2016; Fahrenkrog et al. 2016), *Pinus* (Bartholomé et al. 2016a; Lu et al. 2017) and *Eucalyptus* (Cappa et al. 2013; Müller et al. 2017). All of them, however, had low detection power due to the small numbers of individuals used. Despite the considerably larger number of individuals used in our study for each population independently (n = 758-979) and for the combined dataset (n = 3,373), our results suggest that an even greater number of individuals will be necessary to identify regions that would capture larger fractions of the genetic effects for such

complex growth traits. Although the overall genomic heritabilities estimated using all markers (0.296-0.528) account for a large proportion (64–89%) of the pedigree-based heritabilities, the GWAS results contributed too little for genetic variance given the relatively low number of associations identified for these complex traits. Using the RHM approach, we identified a total of 37 windows, 15 for DBH and 22 for HT (Table 2-4), each one encompassing 100 SNPs and providing heritability estimations for genomic regions 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 in other studies (Nagamine et al. 2012; Riggio et al. 2013; Resende R.T. et al. 2016). Some genomic windows identified by RHM individually explained 3 to 13% of the genomic heritability, similarly to what was obtained by Resende R.T. *et al.* 2016. Additional genomic regions were identified using a Joint-GWAS approach with a large number of individuals (Fig. 2-4 and 2-6). We performed a region-based Joint-GWAS with a window of 100 Kbp, because the typical extent of LD was around 35-75 Kbp for three populations and larger (~600 Kbp) for Pop4-CNB, indicating that most 100 Kbp windows in the genome may include variants that affect growth traits. We also applied Joint-GWAS at a gene level, a powerful approach that detected important genes related to the growth traits analyzed (Fig. 2-6 and Table 2-5), being the first study in *Eucalyptus* to attempt gene-based GWAS.

Considering the single-SNP GWAS analysis accounting for population structure and relatedness, 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, where 28 were detected using gene-based and two genes using region-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,

meristems expansion and requires regulation of complex metabolic pathways with indirect effects on wood formation (Grattapaglia et al. 2009). This view is in line with the "omnigenic" model proposed by Boyle *et al.*, (2017), for human traits suggesting that association signals for complex traits tend to be spread across the genome, including core genes directly affecting the phenotype (common variants with large effects) vastly outnumbered by many peripheral genes without any obvious connection to the trait. Since these core genes only constitute a small fraction of all genes, most heritability comes from genes with indirect effects (Boyle et al. 2017), a view that also fits with Fisher's infinitesimal model (Fisher 1918). For this attempted functional description of the associations found, SNPs associated using RHM were not included, because each significant region has at least 100 SNPs and the concept of this approach is to identify regions with common and low-frequent variants rather than specific-genes (Nagamine et al. 2012; Riggio et al. 2013).

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 ≤ 0.05) were identified encompassing four significant terms for biological process (single-organisms process, signaling, 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 since 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 contributed. Furthermore, this suggests that broad functional categories contribute more to total trait heritability than genes in apparently specific-relevant functional categories related to the complex trait evaluated, as the largest contributor to heritability was simply the largest category. Although GWAS peaks might be peripheral to complex traits (Callaway 2017), identifying more associations might enable the identification of the biological networks implicated in growth and understand their interactions.

Associations for growth pinpoint genes involved in cell wall biosynthesis

Our study was limited to the most commonly measured growth traits, which together with wood specific gravity constitute the mainstay of tree breeding and forest productivity in *Eucalyptus*. No specific phenotype related to wood formation could be obtained in these populations. However, it is well known that growth is determined largely by cell wall biosynthesis, involving carbohydrate metabolism and lignification. Interestingly, a number of associations found are localized into genes related to cell wall biosynthesis. In our Joint-GWAS analysis, the most significant SNP (Bonferroni at 5%) associated with DBH and HT (EuBR07s38098526, MAF = 0.01468) was detected in the exon of gene model (Eucgr.G02075/AT1G14720) encoding for xyloglucan endotransglucosylase/hydrolase 28 (XTH28) (Table 2-5). 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 2-5), where the top putatively associated SNP (lowest *p*-value) is a common variant (EuBR07s34941110, MAF = 0.123). Xyloglucan endotransglycosylase/hydrolase (XTH) enzymes act remodeling 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. Hence, they participate in cell wall construction of growing tissues (Van Sandt et al. 2007), with evident effect on root growth and cell wall extension (Maris et al. 2009). A genome-wide study in natural populations of *P. trichocarpa* (Mckown et al. 2014) also detected significant association for phenology traits in a gene encoding XTH28 (Potri.008G138400 / AT1G14720).

The Joint-GWAS approach also detected a common SNP (EuBR06s6100971, MAF = 0.4226) putatively associated with DBH located on chromosome 6 in gene model Eucgr.F00486 (AT5G42100) that encodes for a glucan endo-1,3-β-glucosidase (Table 2-5), 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. 2-5) 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 role in stretching cells 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; 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. 2-4, 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 *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), that regulates primary cell wall biosynthesis and cellulose microfibrils organization (Lei et al. 2013). A GWAS in *Populus* identified SNPs associated with biomass, ecophysiology and phenology traits in different cellulose synthase genes (Mckown et al. 2014): CESA2 (cellulose synthase A2: Potri.007G076500 / AT4G39350), CESA4 (cellulose synthase A4: Potri.002G257900 / AT5G44030) and CSLA9 (cellulose synthase like A9: Potri.006G116900 / AT5G03760). The dissection of cellulose synthase complexes (CSCs), including cellulose synthase interactive proteins (CSIs) and cellulose synthase genes (CESAs), is important to understand the molecular mechanism underlying the intimate relationship between cellulose microfibrils and microtubules (Lei et al. 2012; Lei et al. 2014). Although the different genes identified in our study will require further validation, consistency with GWAS results from studies in poplar provide valuable preliminary leads for further investigation. It is noteworthy that the first *Eucalyptus* transgenic approved in Brazil (Nature 2015) with a claimed potential to produce between 4 and 20% more wood than the wild type (Ledford 2014), was engineered for an endo-1,4-β-glucanase (CEL1) from *Arabidopsis* that affects plant growth (Shani et al. 2006), a gene related to the cellulose synthase-like C family that encodes a β-1,4 glucan synthase (Cocuron et al. 2007).

Associations for growth pinpoint genes involved in disease resistance

In addition to the associations detected for growth revealing genes involved in the regulation of cell wall biosynthesis, we also identified associations with SNPs into disease resistance genes (Table 2-5). In plants, most of the disease resistance genes encode nucleotide-binding site leucine-rich repeat (NBS-LRR) proteins, that are subdivided into two functionally different domains: TIR (toll/interleukin-1 receptor) and CC (coiled-coil) subfamilies (McHale et al. 2006). The single-SNP analysis in Pop4-CNB detected an association with HT at SNP EuBR02s18469492 (FDR at 5%) located into gene model

Eucgr.B01164 (AT3G14470, RPPL1, putative disease resistance RPP13-like protein 1), encoding for NB-ARC domain-containing disease resistance protein. The Joint-GWAS gene-based from the summary data for HT detected associations in four different genes encoding for NBS-LRR resistance genes, encompassing four to six SNPs spanning each gene. The most significant SNP (EuBR10s34334357, Bonferroni at 5%) belongs to the CC-NBS-LRR (Eucgr.J02727 / AT5G48620, RPP8L4, probable disease resistance RPP8-like protein 4) and the other three genes belong to the TIR-NBS-LRR subfamily (Eucgr.E02914, Eucgr.H01749 and Eucgr.H01750). These four genes are located on chromosomes 5, 8 and 10, in agreement with the localization of NBS-LRR cluster and supercluster of disease resistance genes in *Eucalyptus* (Christie et al. 2016) and also with the GWAS carried out using RHM analysis by Resende R.T. *et al.* (2016). For the latter example, although the authors did not find significance in those regions, in our study they reached significance using Joint-GWAS after combining this population with the other three, further highlighting the gain of power obtained when multiple populations are combined.

CONCLUSIONS

In this study, we carried out a GWAS for growth traits by gathering a considerably larger association population of 3,373 individuals across four breeding populations of *Eucalyptus* in an attempt to evaluate the impact of a larger sample size on the ability to detect discrete associations. Because these trees were genotyped with a common SNP platform we were able to carry out Joint-GWAS analyses, highlighting the value of such public SNP resources – *Eucalyptus* surprisingly still a unique case among forest trees – to advance the investigation of the complex relationships between sequence variation and complex phenotypes. In parallel, we tested several GWAS models with variable levels of correction for population structure and relatedness and different segment-based approaches in an attempt to capture a wider frequency spectrum of variants. Under the most stringent corrections for population structure and relatedness, significant associations were found for height only in one of the populations, where the extent of LD was on the order of ~600 Kbp. When the combined set of 3,373 trees was used, either

as a single-SNP GWAS panel or by Joint-GWAS methods only a few associations were found for diameter growth and none for height. Regional heritability mapping was not able to improve detection and essentially revealed the same associations found by the single-SNP approach in the same population. At lower stringency thresholds or correcting only for population structure, several tens of associations were found, and more importantly, these associations were shared across three of the four populations. Although these associations could in principle be spurious given the no correction for relatedness, the fact that they were independently detected in three populations adds some credibility to them. Significant and putative associations were found in a number of genes related to cell wall biosynthesis and disease resistance, suggesting potential pleiotropic effect of these loci.

Overall, however, our results do not differ substantially from most GWAS for growth traits carried out in forest trees to date. Despite the fact that they were obtained in breeding populations and not in collections of wild trees, this further corroborates that growth is controlled by many variants of relatively small effect such that the infinitesimal model fits the data well. Consistent with this hypothesis and with results of previous GWAS in *Eucalyptus*, we also observed that genomic heritabilities accounted for large proportions (64–89%) of the pedigree-based heritabilities, suggesting that considerably larger samples will be necessary if one intends to capture single variants that explain relevant portions of the genetic variation for growth. While some more encouraging GWAS results have been reported for wood properties and phenology traits mostly in tree species other than *Eucalyptus*, for growth our results point to the fact that genomic prediction approaches shall be more productive when it comes to tree breeding applications (Grattapaglia 2017). Still, as pointed out earlier (Resende R.T. et al. 2016), GWAS data should be useful to enhance the predictive ability of genomic selection, especially from segment or gene-based approaches capturing a combination of common and rare variants contributing comparatively larger portions of the heritability than single-SNP. As more *Eucalyptus* breeding programs adopt genomics to predict phenotypes based on a common SNP platform, increasingly larger datasets will become available and Joint-

analyses, such as the one reported here for the first time in forest trees, should provide the necessary power to pinpoint such genomic segments.

TABLES

Table 2-1: Main characteristics of the four association populations of *Eucalyptus* used in the study.

Phenotypic data	**Pop1-IPB**	**Pop2-ARAB**	**Pop3-ARAC**	**Pop4-CNB**
Company	International Paper Brazil	Fibria	Fibria	Cenibra
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
Effective population size (Ne) estimated by heterozygote excess method (95% CIs)[a]	6.2 (6-6.3)	19.3 (17.9-20.8)	11 (10.6-11.5)	12 (11.3-12.7)
Year when trees were planted	2006	2006	2006	2005
Year when trees were phenotyped	2011	2008	2008	2008
Age at phenotyping (yr)	5	2	2	3
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

*RCBD, Randomized complete block design.

**ALD-IB, Alpha lattice design (incomplete block).

[a]95% Confidence Intervals are shown in parentheses.

Table 2-2: Genotypic data information and number of subpopulations determined using STRUCTURE and PCA for the four association populations.

Attribute	Pop1-IPB	Pop2-ARAB	Pop3-ARAC	Pop4-CNB	All
Number of trees used in the GWAS analyses	979	875	758	761	3373
Total number of SNPs genotyped	60904	60904	60904	60904	60904
Number of SNPs retained with call rate > 90%	46436	47606	46368	46795	51274
Number of SNPs retained with MAF>0*	32110	34859	33800	28795	41320
Number of SNPs retained with MAF>0.01	27366	31819	30979	22728	33700
Number of SNPs retained with MAF>0.05	24092	26484	26438	18075	25533
Number of SNPs retained with MAF>0 in intergenic regions	13853	14918	14503	12813	19038
Number of SNPs retained with MAF>0 in Intergenic Regions and in LE (STRUCTURE)	10042	10105	10389	7790	14705
Number of clusters (K) determined based on Evanno et al., (2005)	2	2	3	5	2
Number of significant principal components determined by a broken stick model (Jackson, 1993)	2	1	2	2	3

* This set of SNPs was used for GWAS analysis.

Table 2-3: Number of significant SNP associations for growth traits using LMA (Model 1 to 3) and MLMA (Model 4 to 6) models for the four breeding populations and for the Joint-GWAS (All) analyses. Also reported the number of SNPs putatively associated with growth traits using MLMA models (Model 4 to 6).

Population	Trait	Number 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	32110	3805(260*)	4212(315*)	4155(318*)	0 | 7	0 | 7	0 | 7
Pop2-ARAB		34859	11147(1783*)	4373(212*)	2906(109*)	0 | 3	0 | 4	0 | 2
Pop3-ARAC		30979	17954(6729*)	9464(1668*)	3655(302*)	0 | 13	0 | 6	0 | 8
Pop4-CNB		28795	12542(3411*)	1149(74*)	1396(34*)	0 | 4	0 | 3	0 | 2
All		41320	24635(11871*)	18395(6291*)	18406(5693*)	3(3*) | 10	3(2*) | 11	2(2*) | 7
Pop1-IPB	HT	32110	2731(119*)	3201(148*)	3237(163*)	0 | 7	0 | 6	0 | 7
Pop2-ARAB		34859	5797(350*)	2854(167*)	2654(145*)	0 | 3	0 | 3	0 | 3
Pop3-ARAC		30979	13815(3338*)	4927(347*)	2201(80*)	0 | 6	0 | 9	0 | 9
Pop4-CNB		28795	8303(959*)	1104(242*)	3263(472*)	27(0*) | 40	97(0*) | 78	12(0*) | 45
All		41320	17383(4385*)	13560(2791*)	12259(2606*)	0 | 3	0 | 4	0 | 1

False discovery rate (FDR) of 5%.

*Bonferroni-correction with an experimental type I error rate of $\alpha = 0.05$.

After "|" is the number of SNPs putatively associated using an *ad hoc* threshold of $-\text{Log}_{10}(P) \geq 4$.

Table 2-4: Regional heritability mapping for windows significantly ($-\log_{10} > 3.0$) and putatively ($-\log_{10} > 2.0$) associated with growth traits for the four breeding populations. Diameter at Breast Height (DBH), Total Height (HT), Likelihood Ratio Test (LRT), Regional heritability (h_r^2).

Trait	Population	Chr.	SNP Start	Position Start (bp)	SNP End	Position End (bp)	LRT	h_r^2	$-\log_{10}$
DBH	Pop1-IPB (626)*	7	EuBR07s31328798	31328798	EuBR07s32568262	32568262	10.76	0.13	2.98
		7	EuBR07s33146968	33146968	EuBR07s35110610	35110610	10.34	0.07	2.89
		7	EuBR07s32581478	32581478	EuBR07s33957780	33957780	10.19	0.07	2.85
		7	EuBR07s25714206	25714206	EuBR07s30352510	30352510	10.18	0.06	2.85
		7	EuBR07s31961430	31961430	EuBR07s33145504	33145504	9.53	0.07	2.69
		7	EuBR07s20583725	20583725	EuBR07s22342031	22342031	8.67	0.06	2.49
		7	EuBR07s22342887	22342887	EuBR07s25713595	25713595	8.53	0.06	2.46
		7	EuBR07s36784209	36784209	EuBR07s38583112	38583112	8.15	0.04	2.36
		7	EuBR07s21332450	21332450	EuBR07s24532470	24532470	7.93	0.05	2.31
		7	EuBR07s16684296	16684296	EuBR07s19147010	19147010	7.91	0.05	2.31
		7	EuBR07s28499843	28499843	EuBR07s31328715	31328715	7.38	0.06	2.18
		7	EuBR07s24545298	24545298	EuBR07s28499722	28499722	7.18	0.05	2.13
	Pop2-ARAB (683)*	6	EuBR06s32562797	32562797	EuBR06s34328105	34328105	7.96	0.05	2.32
	Pop3-ARAC (660)*	6	EuBR06s26699092	26699092	EuBR06s27751254	27751254	9.62	0.12	2.72
		9	EuBR09s31933985	31933985	EuBR09s33975533	33975533	8.02	0.07	2.33
HT	Pop1-IPB (626)*	2	EuBR02s16102502	16102502	EuBR02s18417551	18417551	7.03	0.04	2.10
		7	EuBR07s14930135	14930135	EuBR07s18103200	18103200	6.65	0.04	2.00
	Pop4-CNB (560)*	2	EuBR02s42263455	42263455	EuBR02s43397707	43397707	14.59	0.07	3.87[b]
		2	EuBR02s23849815	23849815	EuBR02s25008423	25008423	13.16	0.10	3.54[f]
		2	EuBR02s42780242	42780242	EuBR02s43864353	43864353	13.15	0.06	3.54[f]
		2	EuBR02s23213141	23213141	EuBR02s24367225	24367225	12.29	0.07	3.34[f]
		2	EuBR02s17118873	17118873	EuBR02s19701439	19701439	11.30	0.06	3.11[f]
		2	EuBR02s39648070	39648070	EuBR02s42778399	42778399	11.01	0.06	3.04[f]
		2	EuBR02s22594380	22594380	EuBR02s23832192	23832192	11.01	0.06	3.04[f]
		2	EuBR02s18112848	18112848	EuBR02s20898091	20898091	10.84	0.05	3.00[f]
		2	EuBR02s20898153	20898153	EuBR02s23179227	23179227	9.37	0.06	2.66
		2	EuBR02s36468959	36468959	EuBR02s39639880	39639880	9.16	0.07	2.61
		2	EuBR02s27662134	27662134	EuBR02s31703058	31703058	8.66	0.07	2.49
		2	EuBR02s29793530	29793530	EuBR02s32366267	32366267	8.54	0.06	2.46
		2	EuBR02s35689626	35689626	EuBR02s37602857	37602857	8.54	0.06	2.46
		2	EuBR02s24391700	24391700	EuBR02s25950796	25950796	8.33	0.06	2.41
		2	EuBR02s19809602	19809602	EuBR02s22590567	22590567	7.99	0.04	2.33
		2	EuBR02s15507347	15507347	EuBR02s18088025	18088025	7.65	0.04	2.25
		2	EuBR02s43397812	43397812	EuBR02s44360147	44360147	7.41	0.05	2.19
		2	EuBR02s31704444	31704444	EuBR02s33368834	33368834	7.13	0.05	2.12
		1	EuBR01s24700230	24700230	EuBR01s26451182	26451182	6.89	0.04	2.06
		1	EuBR01s17433597	17433597	EuBR01s19071730	19071730	6.74	0.03	2.03

*Total number of windows, [b]Bonferroni-correction with an experimental type I error rate of $\alpha = 0.05$, [f]False discovery rate (FDR) of 5%.

Table 2-5: Important associations for growth traits (DBH and HT) pinpoint genes involved in cell wall biosynthesis and in disease resistance.

GWAS Data	Trait	SNP	Chr.	Position (bp)	-log$_{10}$	REF/ ALT	Eg - Gene	At - Gene	Annotation
Joint-GWAS Single-SNP	DBH	EuBR07s38098526	7	38098526	8.21[b]	G/A	Eucgr.G02075	AT1G14720	Xyloglucan endotransglucosylase/hydrolase 28
Joint-GWAS Single-SNP	DBH	EuBR08s48262720	8	48262720	5.84[f]	A/G	Eucgr.H03281	AT3G06720	Armadillo/beta-catenin-like repeats-containing protein-related
Joint-GWAS Single-SNP	DBH	EuBR06s6100971	6	6100971	4.10[a]	A/G	Eucgr.F00486	AT5G42100	Glucan 1,3-beta-glucosidase A
Pop1-IPB/Pop2-ARAB/Pop3-ARAC	DBH	EuBR04s17486529	4	17486529	-[f]	C/T	Eucgr.D00955	AT4G17180	O-Glycosyl hydrolases family 17 protein
Pop1-IPB/Pop2-ARAB/Pop3-ARAC	DBH	EuBR04s17531959	4	17531959	-[f]	G/A	Eucgr.D00963	AT5G47780	Galacturonosyltransferase 4
Pop1-IPB/Pop2-ARAB/Pop3-ARAC	DBH	EuBR10s8284185	10	8284185	-[f]	G/A	Eucgr.J00782	AT4G34890	Xanthine dehydrogenase 1
Joint-GWAS Gene-based	HT	EuBR10s34334357	10	34334357	9.05[b]	C/T	Eucgr.J02727	AT5G48620	Disease resistance protein (CC-NBS-LRR class) family
Joint-GWAS Gene-based	HT	EuBR05s48058595	5	48058595	7.45[f]	A/G	Eucgr.E02914	AT5G17680	Disease resistance protein (TIR-NBS-LRR class), putative
Joint-GWAS Gene-based	HT	EuBR08s21538562	8	21538562	5.58[a]	G/T	Eucgr.H01749	AT5G36930	Disease resistance protein (TIR-NBS-LRR class) family
Joint-GWAS Gene-based	HT	EuBR08s21538562	8	21538562	5.58[a]	G/T	Eucgr.H01750	AT5G36930	Disease resistance protein (TIR-NBS-LRR class) family
Pop4-CNB	HT	EuBR02s18469492	2	18469492	4.87[f]	T/G	Eucgr.B01164	AT3G14470	NB-ARC domain-containing disease resistance protein
Joint-GWAS Gene-based	HT	EuBR07s34941110	7	34941110	4.36[a]	G/A	Eucgr.G01909	AT5G13870	Xyloglucan endotransglucosylase/hydrolase 5
Pop1-IPB/Pop2-ARAB/Pop3-ARAC	HT	EuBR05s70210869	5	70210869	-[f]	C/T	Eucgr.E04103	AT1G61820	Beta glucosidase 46

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

FIGURES

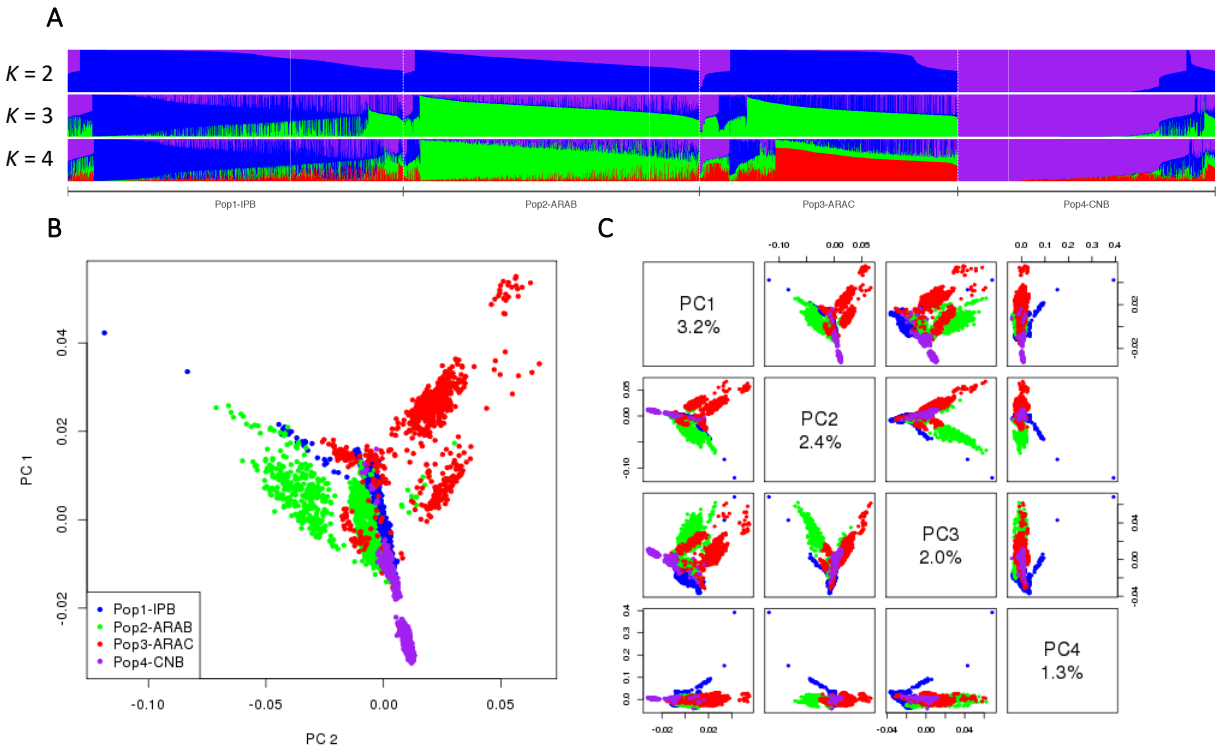


Figure 2-1: Population Structure and Principal Component Analysis (PCA) for the four unrelated *E. grandis* x *E. urophylla* hybrid breeding populations. (A) Barplots from STRUCTURE for number of cluster ranging from *K*=2 to *K*=4. (B) PCA with two eigenvectors (PC1 and PC2) and (C) PCA with PC1 to PC4.

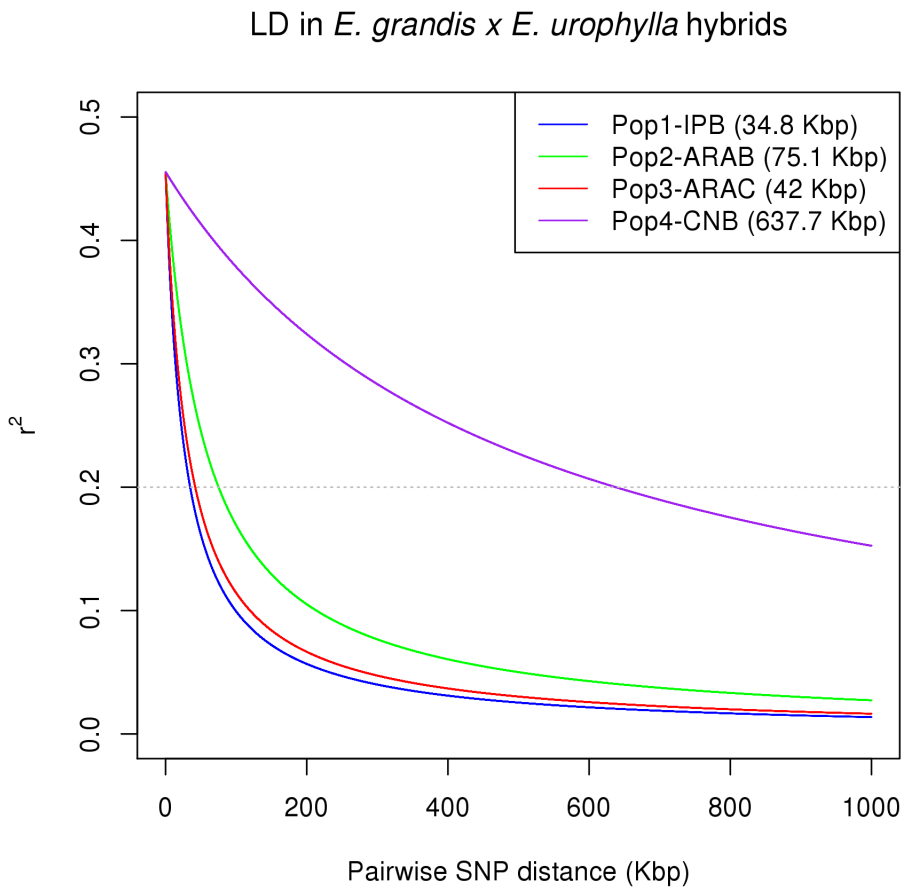


Figure 2-2: Genome-wide pattern of Linkage Disequilibrium (LD) decay plotted up to 1 Mbp pairwise SNP distances, considering rare alleles (MAF > 0). Decay curves of the classical measure of the squared correlation of allele frequencies at diallelic loci (r^2) for each population individually and a dashed line at r^2 = 0.2 indicates the frequently used threshold of usable LD.

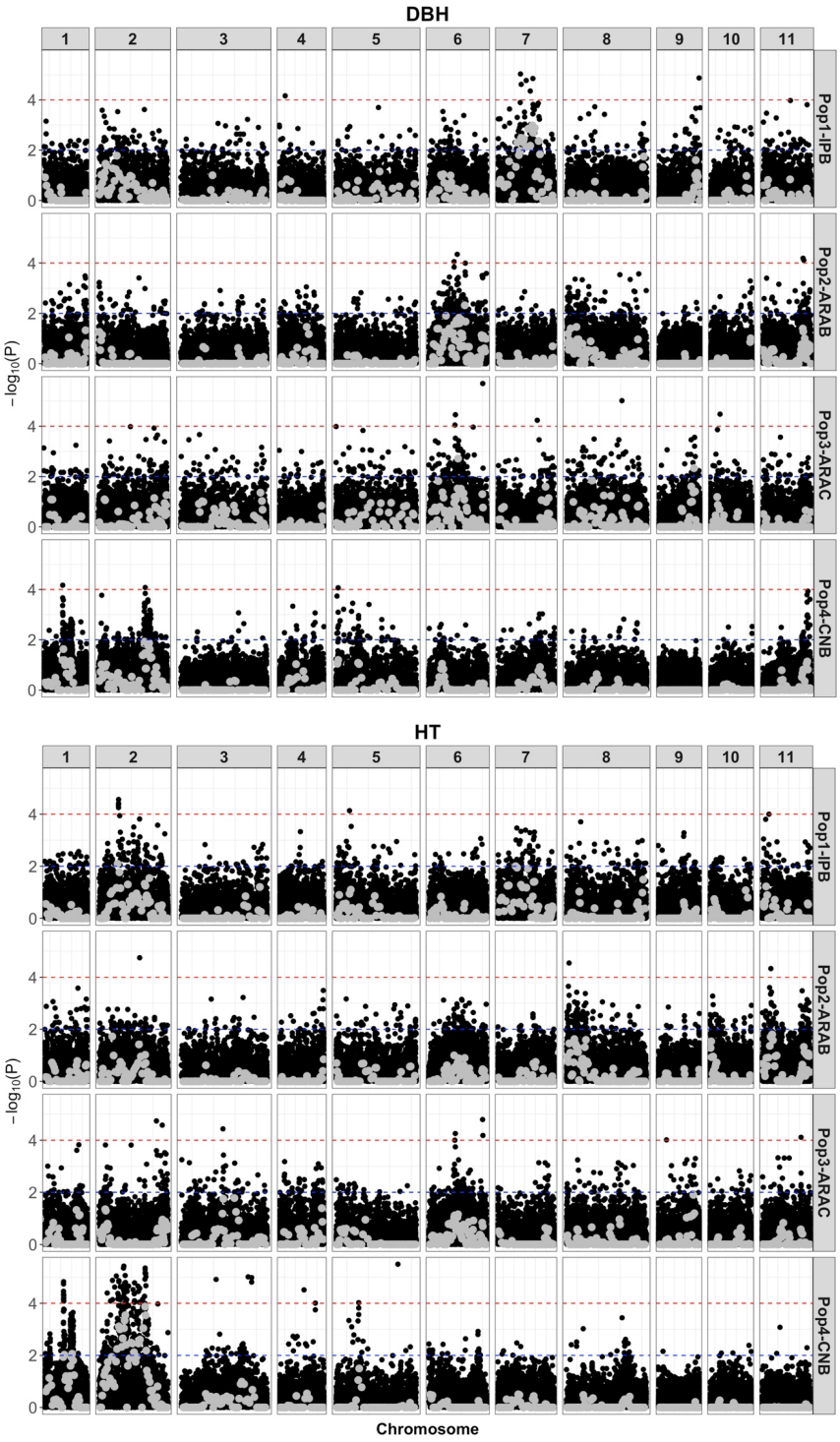


Figure 2-3: Manhattan plots for growth traits (DBH and HT) using single-SNP GWAS (black points) and RHM (grey points), corrected for population structure and the kinship matrix, for the four unrelated *E. grandis* x *E. urophylla* hybrids breeding populations. Red and blue line indicate *ad hoc* thresholds adopted for the single-SNP GWAS and RHM analyses, respectively.

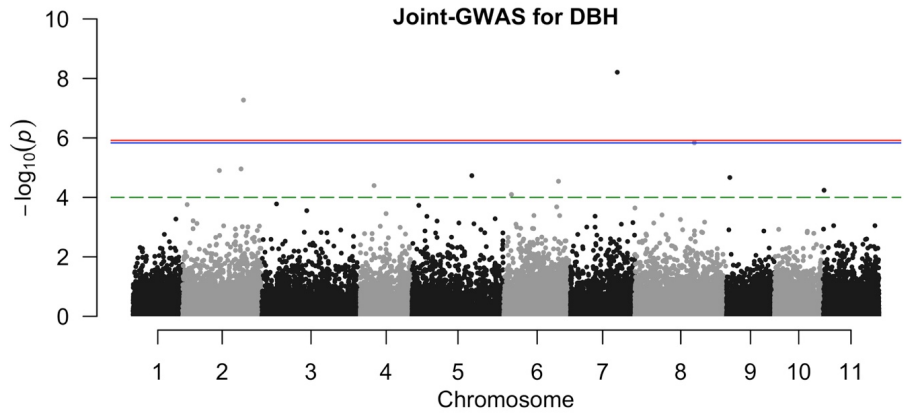


Figure 2-4: Manhattan plot of the associations for diameter at breast height (DBH) using single-SNP Joint-GWAS (41,320 SNPs), adjusted for STRUCTURE, the GRM, age of measurements and population of origin for the combined dataset. Red line indicates Bonferroni-corrected threshold with an experimental type I error rate at α =0.05, blue line indicates a false discovery rate (FDR) at 5% and green dashed line represents the *ad hoc* threshold.

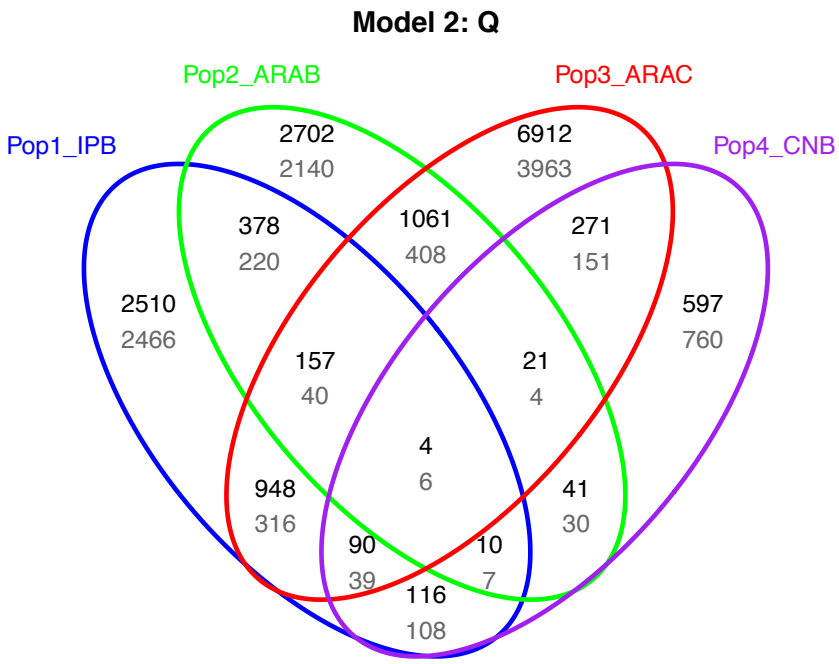


Figure 2-5: Venn Diagram of the number of significant associations identified for growth traits using single-SNP GWAS for the four unrelated *E. grandis* x *E. urophylla* hybrids breeding populations. Comparison of the number of significant associations identified for DBH (black numbers) and HT (grey numbers) by false discovery rate (FDR) threshold at 5%, using LMA model corrected for STRUCTURE (Model 2: Q). Diameter at Breast Height (DBH, cm), Total Height (HT, m).

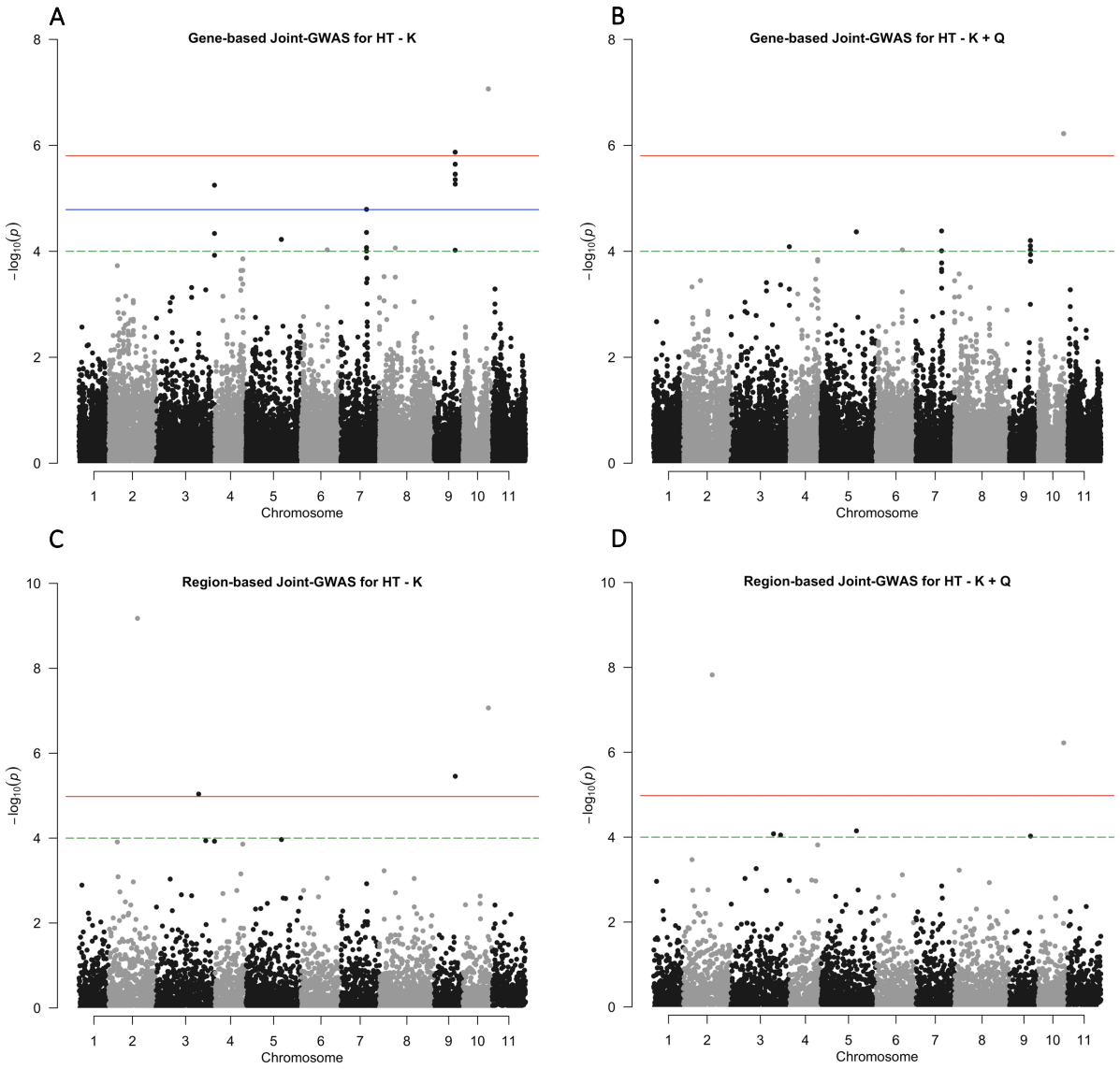


Figure 2-6: Manhattan plots of the associations for HT using gene-based (31,770 genes) and region-based (4,766 windows) Joint-GWAS for the combined dataset. (A) Gene-based Joint-GWAS adjusted for GRM, age of measurements and population of origin. (B) Gene-based Joint-GWAS adjusted for all covariates mentioned before with the inclusion of STRUCTURE. (C) Region-based Joint-GWAS adjusted for GRM, age of measurements and population of origin. (D) Region-based Joint-GWAS adjusted for all other covariates with the inclusion of STRUCTURE. Red line indicates Bonferroni-corrected threshold with an experimental type I error rate at α =0.05, blue line indicates a false discovery rate (FDR) at 5% and green dashed line represents an *ad hoc* threshold of $-\log_{10}= 4.0$

SUPPLEMENTARY MATERIAL (SM2)

Table SM2-1: Estimates of additive genetic variances (σ^2_a), residual variances (σ^2_e), phenotypic variances (σ^2_p) and narrow-sense heritabilities (h^2) for the four unrelated *E. grandis* x *E. urophylla* hybrids breeding populations. Standard Deviation (SD); Standard Error (SE).

Population	Trait	Method	σ^2_a	SD (σ^2_a)	SE (σ^2_a)	σ^2_e	SD (σ^2_e)	SE (σ^2_e)	σ^2_p	SD (σ^2_p)	SE (σ^2_p)	h^2	SD (h^2)	SE (h^2)
Pop1-IPB	DBH	Pedigree-based	1.8621	0.4485	0.0026	1.6813	0.2536	0.0015	3.5434	0.2546	0.0015	0.5204	0.0927	0.0005
	HT		0.6466	0.1808	0.0010	0.9202	0.1080	0.0006	1.5669	0.1059	0.0006	0.4082	0.0895	0.0005
Pop2-ARAB	DBH		3.5663	0.8665	0.0050	3.2070	0.4999	0.0029	6.7733	0.4953	0.0029	0.5214	0.0942	0.0005
	HT		2.9137	0.8013	0.0046	4.1260	0.4990	0.0029	7.0398	0.4717	0.0027	0.4097	0.0897	0.0005
Pop3-ARAC	DBH		7.1232	1.2246	0.0071	3.2523	0.6525	0.0038	10.3756	0.7559	0.0044	0.6826	0.0774	0.0004
	HT		5.1810	1.1265	0.0065	3.0535	0.6159	0.0036	8.2345	0.6541	0.0038	0.6235	0.0944	0.0005
Pop4-CNB	DBH		1.6006	0.6966	0.0040	2.5747	0.3839	0.0022	4.1754	0.3936	0.0023	0.3735	0.1255	0.0007
	HT		2.3014	0.8963	0.0052	2.9436	0.4878	0.0028	5.2451	0.5020	0.0029	0.4286	0.1269	0.0007
Pop1-IPB	DBH	Genomic-based	1.6183	0.2557	0.0015	1.9913	0.1295	0.0007	3.6096	0.2234	0.0013	0.4463	0.0486	0.0003
	HT		0.5492	0.1037	0.0006	1.0153	0.0627	0.0004	1.5645	0.0907	0.0005	0.3492	0.0509	0.0003
Pop2-ARAB	DBH		2.2276	0.3976	0.0023	3.9500	0.2626	0.0015	6.1775	0.3572	0.0021	0.3590	0.0497	0.0003
	HT		2.4659	0.4415	0.0025	4.2381	0.2859	0.0017	6.7039	0.3888	0.0022	0.3661	0.0507	0.0003
Pop3-ARAC	DBH		4.2881	0.6390	0.0037	3.8051	0.3231	0.0019	8.0932	0.5370	0.0031	0.5277	0.0513	0.0003
	HT		2.7555	0.4816	0.0028	4.1163	0.3055	0.0018	6.8717	0.4253	0.0025	0.3992	0.0524	0.0003
Pop4-CNB	DBH		1.1648	0.2199	0.0013	2.3316	0.1426	0.0008	3.4964	0.2238	0.0013	0.3313	0.0471	0.0003
	HT		1.3760	0.2654	0.0015	3.2499	0.1956	0.0011	4.6259	0.2841	0.0016	0.2960	0.0447	0.0003

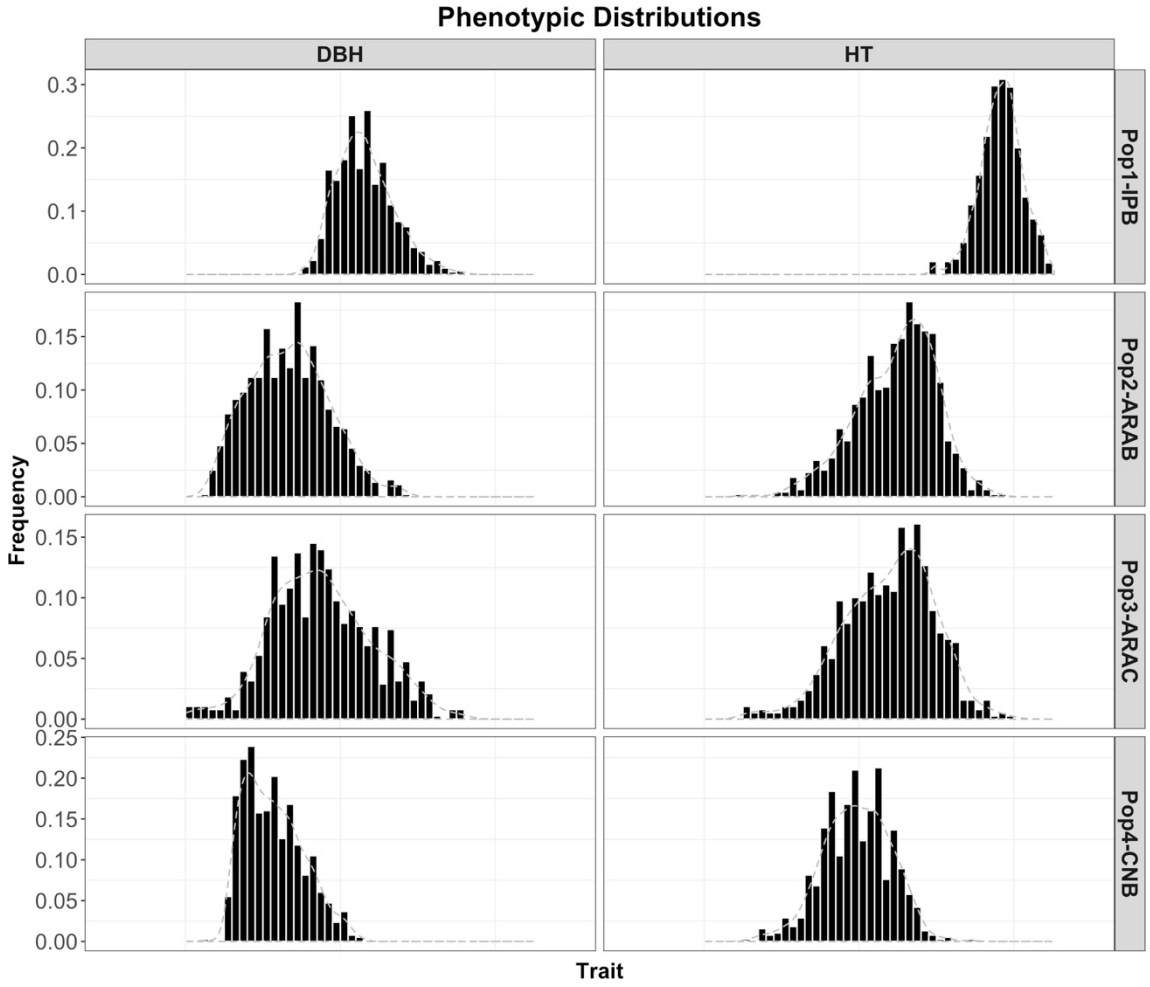


Figure SM2-1: Phenotypic distributions with density line of the growth traits measured in the four *Eucalyptus grandis* x *E. urophylla* hybrids breeding populations. Diameter at Breast Height (DBH, cm), Total Height (HT, m).

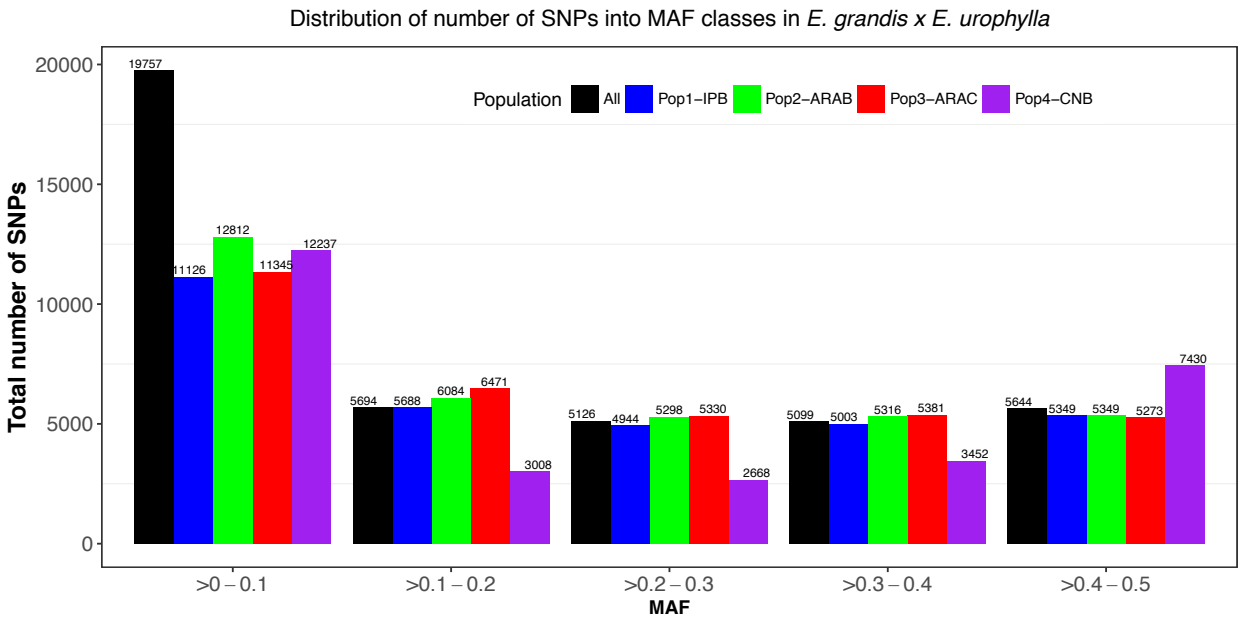
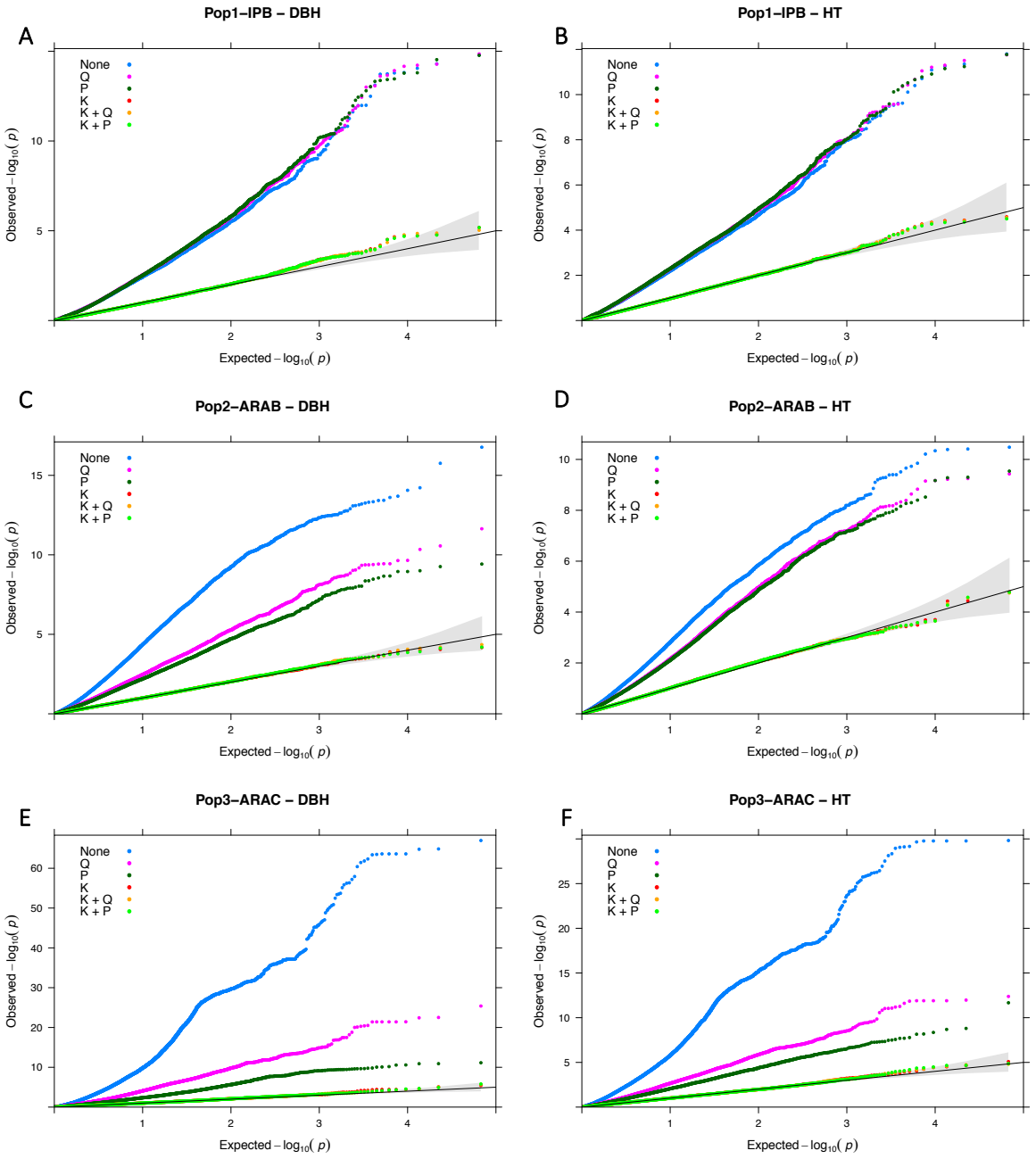


Figure SM2-2: Distribution of the number of SNPs into MAF classes for each population and combined data (All) using CR ≥ 90% and MAF > 0. Call Rate (CR); Minimum Allele Frequency (MAF).


A
Pop1–IPB – DBH
B
Pop1–IPB – HT
C
Pop2–ARAB – DBH
D
Pop2–ARAB – HT
E
Pop3–ARAC – DBH
F
Pop3–ARAC – HT
None
Q
P
K
K + Q
K + P
Observed – log10(p)
Expected – log10(p)

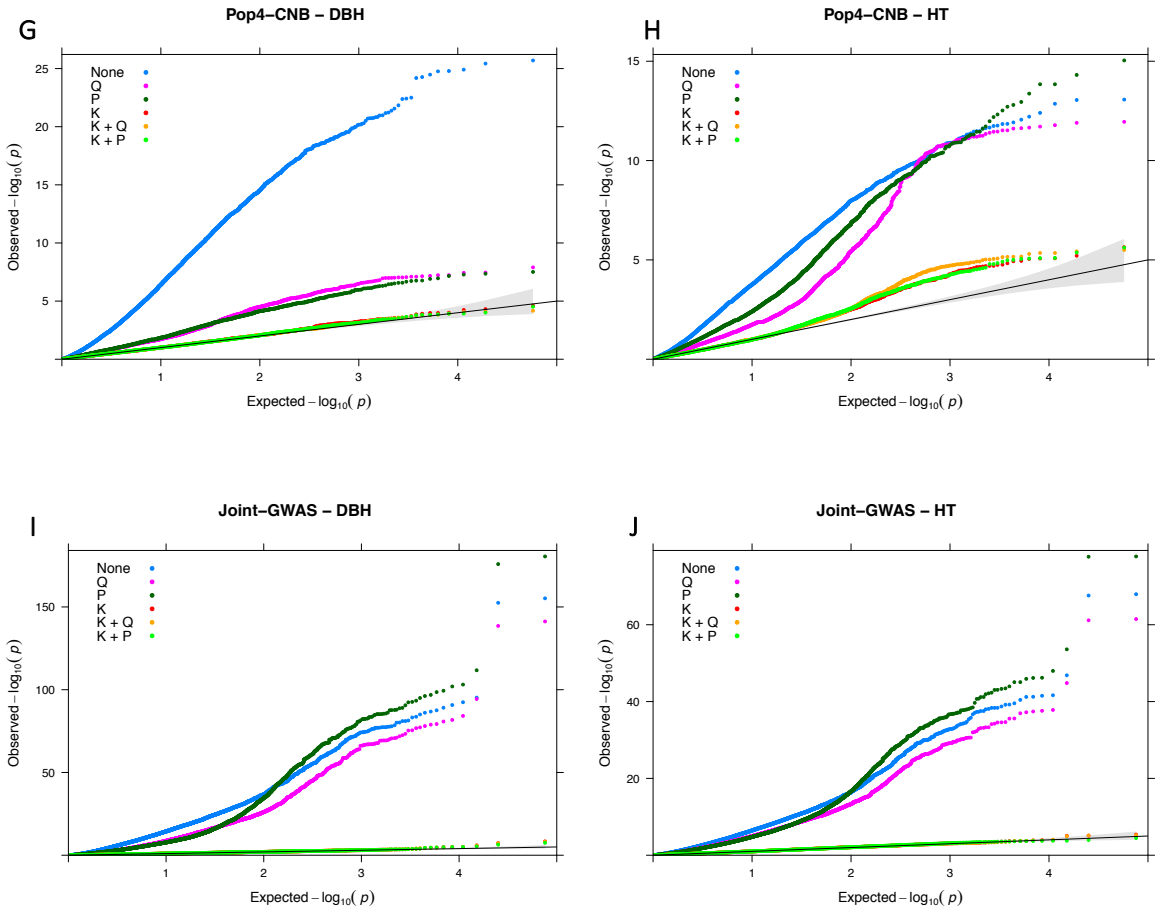


Figure SM2-3: Quantile-quantile (QQ) plots for SNP-based models for diameter at breast height (DBH) and total height (HT), respectively. (A) and (B) represent QQ plots for Pop1-IPB. (C) and (D) for Pop2-ARAB. (E) and (F) for Pop3-ARAC. (G) and (H) for Pop4-CNB. (I) and (J) for the Joint-GWAS (combined dataset).

Figure SM2-4: Manhattan plots for SNP-based models for all four populations independently and combined dataset (Joint-GWAS).

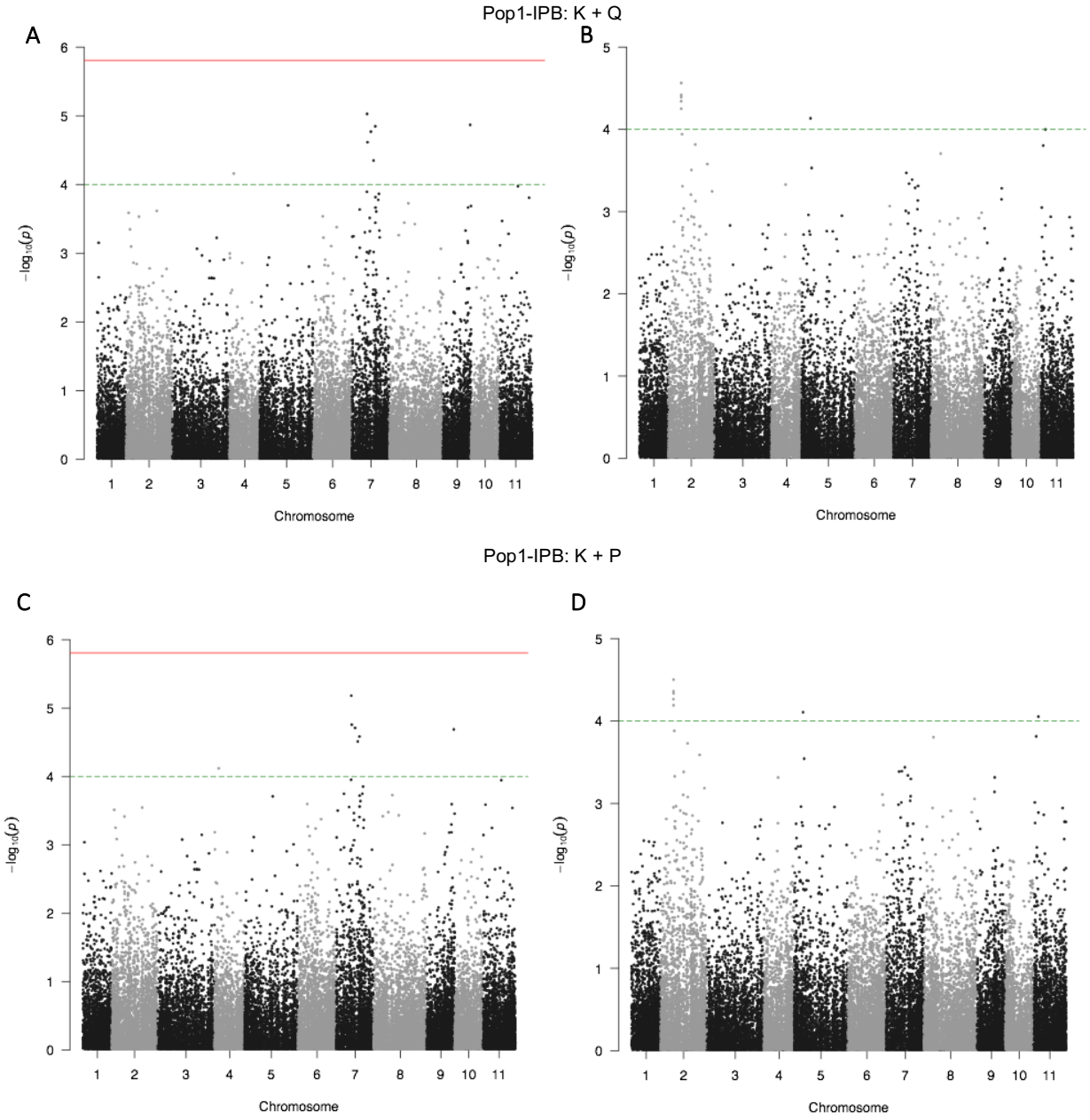


Figure SM2-4A: Manhattan plots for SNP-based models in Pop1-IPB. A and B represent Manhattan plots for DBH and HT, respectively, using MLMA model adjusted for GRM and STRUCTURE. C and D represent Manhattan plots for DBH and HT, respectively, using MLMA model adjusted for GRM and PCA. Red line indicates Bonferroni-corrected threshold with an experimental type I error rate at α =0.05 and green dashed line represents the *ad hoc* threshold. Diameter at Breast Height (DBH), Total Height (HT).

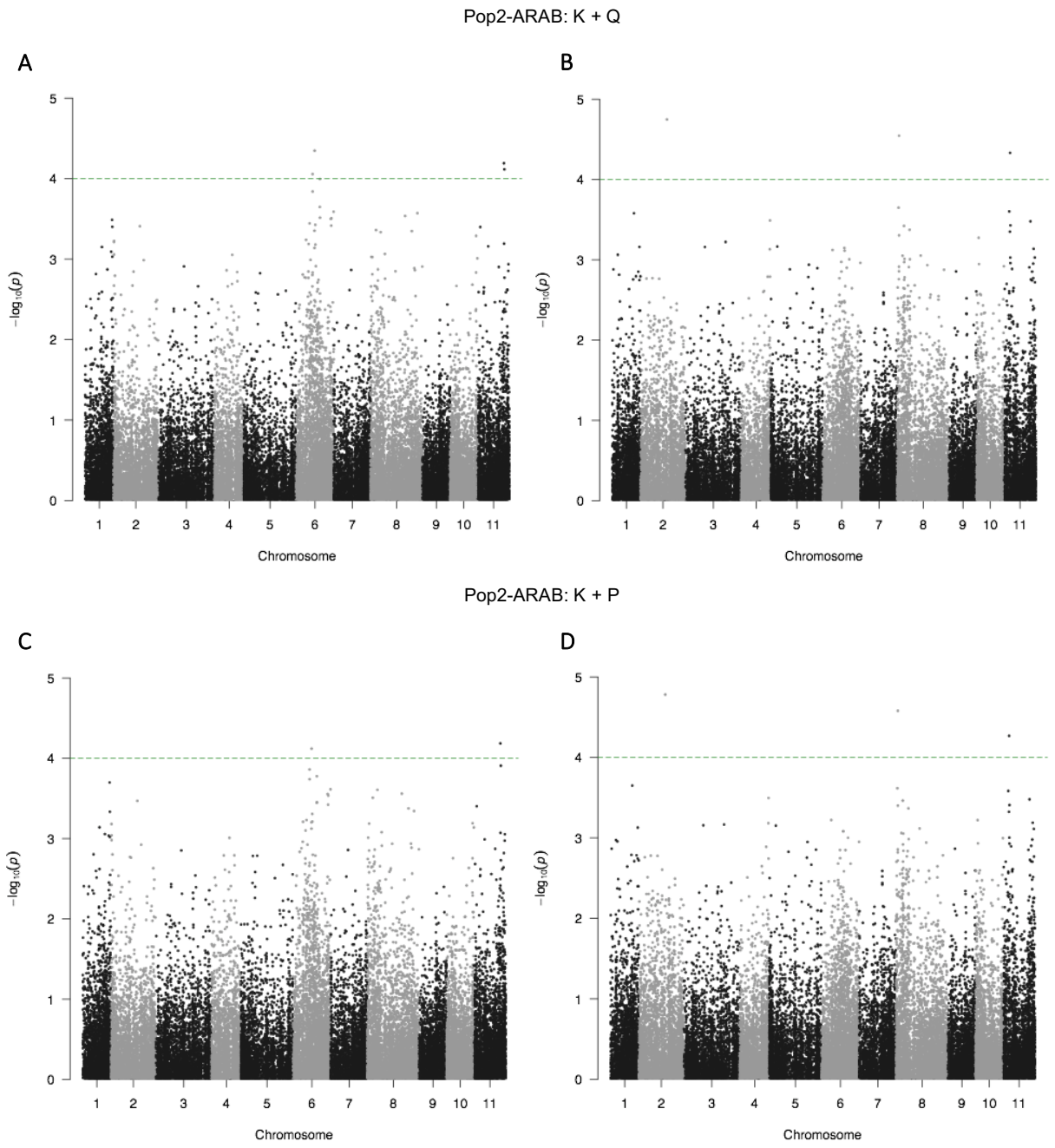


Figure SM2-4B: Manhattan plots for SNP-based models in Pop2-ARAB. A and B represent Manhattan plots for DBH and HT, respectively, using MLMA model adjusted for GRM and STRUCTURE. C and D represent Manhattan plots for DBH and HT, respectively, using MLMA model adjusted for GRM and PCA. Green dashed line represents the *ad hoc* threshold. Diameter at Breast Height (DBH), Total Height (HT).

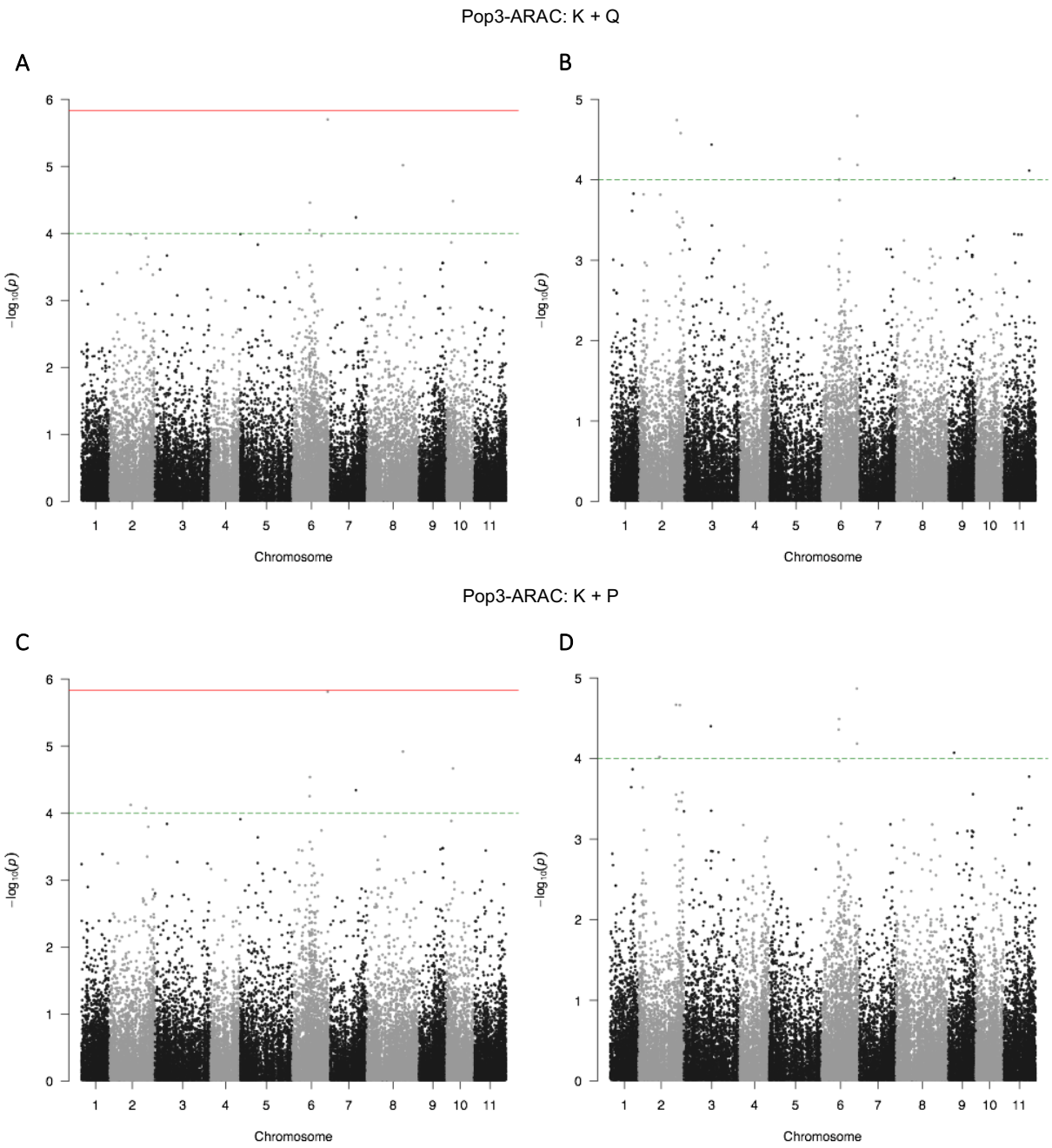


Figure SM2-4C: Manhattan plots for SNP-based models in Pop3-ARAC. A and B represent Manhattan plots for DBH and HT, respectively, using MLMA model adjusted for GRM and STRUCTURE. C and D represent Manhattan plots for DBH and HT, respectively, using MLMA model adjusted for GRM and PCA. Red line indicates Bonferroni-corrected threshold with an experimental type I error rate at α =0.05 and green dashed line represents the *ad hoc* threshold. Diameter at Breast Height (DBH), Total Height (HT).

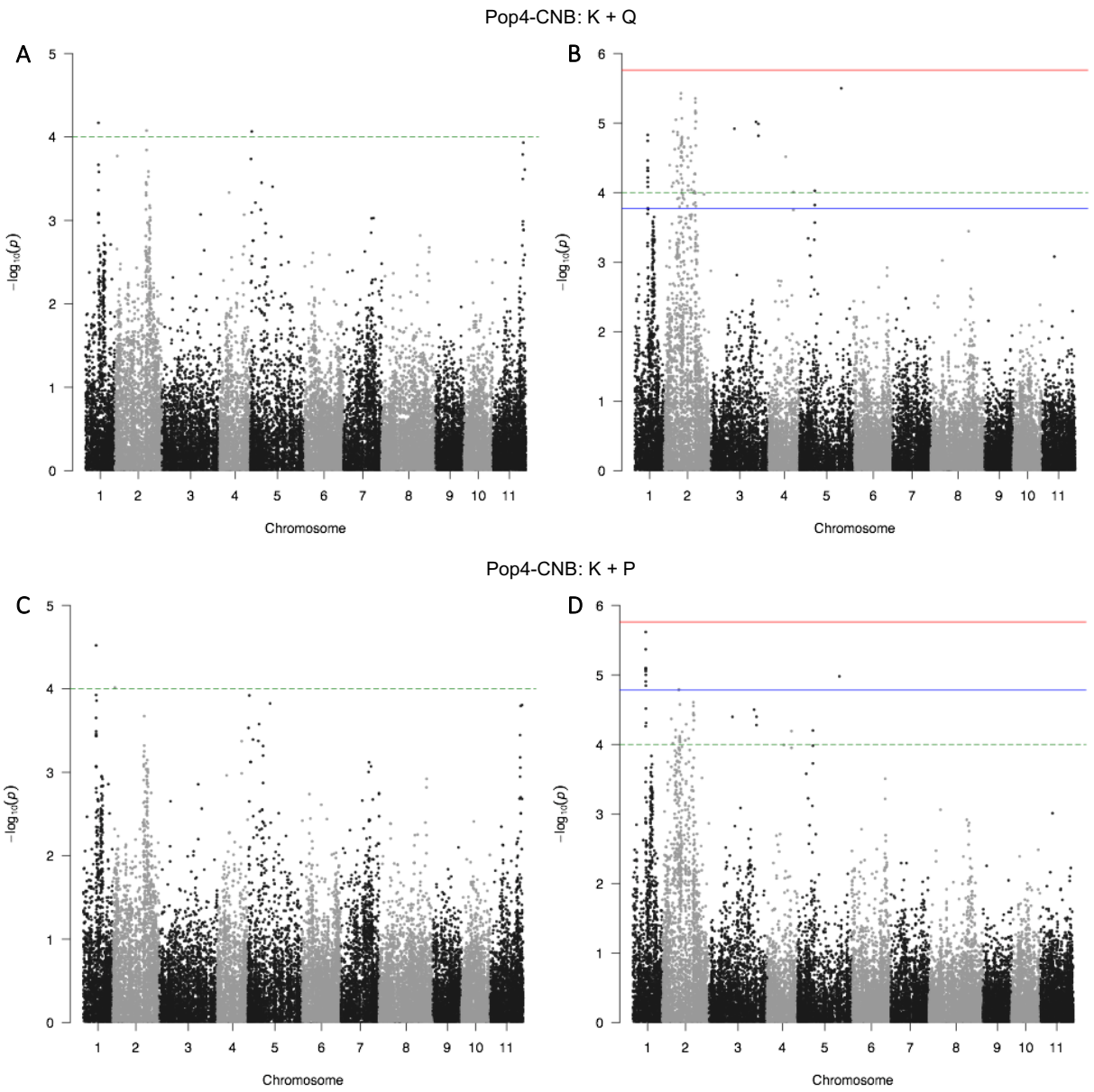


Figure SM2-4D: Manhattan plots for SNP-based models in Pop4-CNB. A and B represent Manhattan plots for DBH and HT, respectively, using MLMA model adjusted for GRM and STRUCTURE. C and D represent Manhattan plots for DBH and HT, respectively, using MLMA model adjusted for GRM and PCA. Red line indicates Bonferroni-corrected threshold with an experimental type I error rate at α =0.05, blue line indicates false discovery rate (FDR) at 5% and green dashed line represents the *ad hoc* threshold. Diameter at Breast Height (DBH), Total Height (HT).

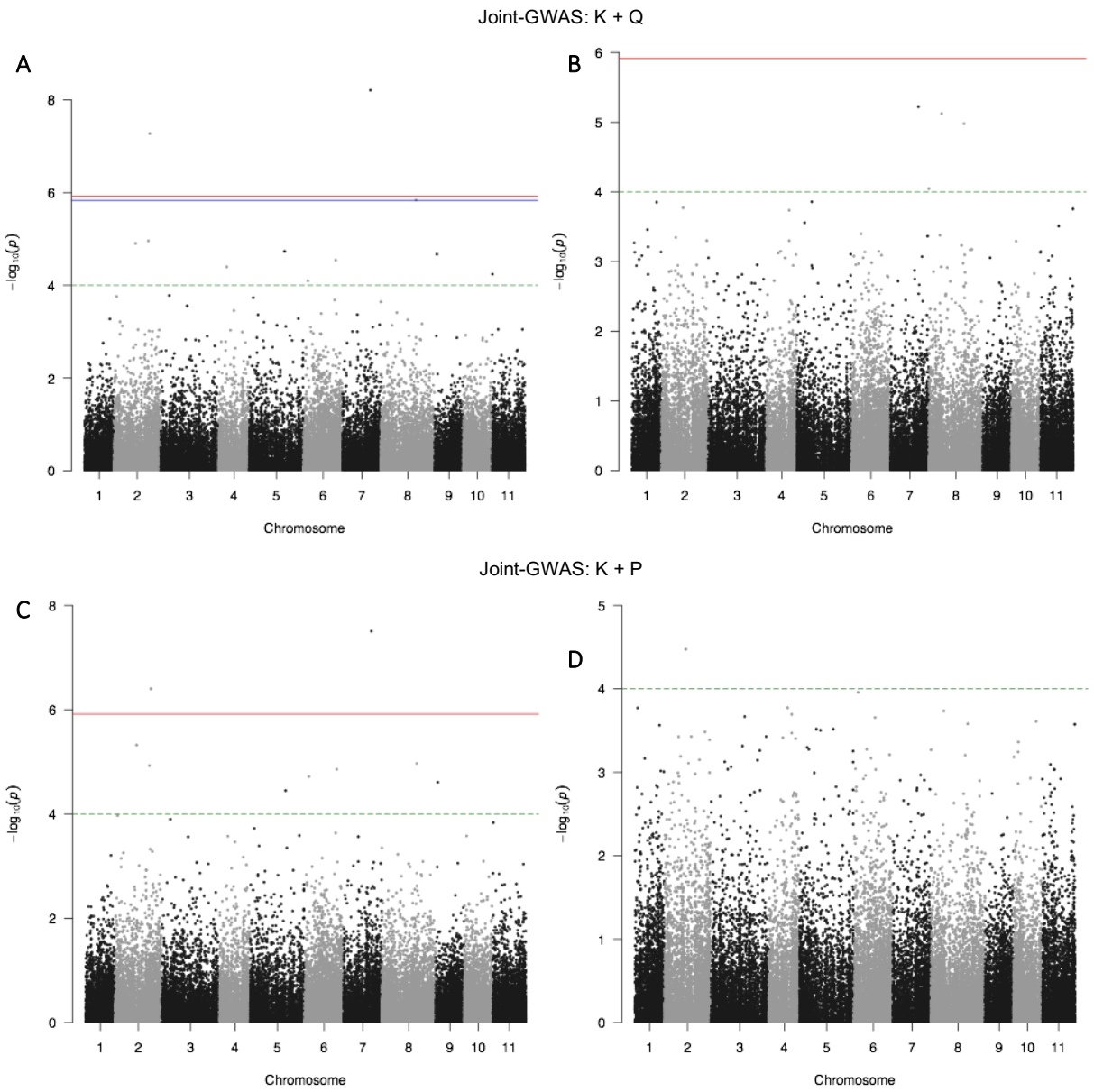


Figure SM2-4E: Manhattan plots using SNP-based models for Joint-GWAS for the four populations combined dataset. A and B represent Manhattan plots for DBH and HT, respectively, using MLMA model adjusted for GRM and STRUCTURE. C and D represent Manhattan plots for DBH and HT, respectively, using MLMA model adjusted for GRM and PCA. Red line indicates Bonferroni-corrected threshold with an experimental type I error rate at α =0.05, blue line indicates false discovery rate (FDR) at 5% and green dashed line represents the *ad hoc* threshold. Diameter at Breast Height (DBH), Total Height (HT).

REFERENCES

Acosta-Pech R, Crossa J, de los Campos G, Teyssèdre S, Claustres B, Pérez-Elizalde S, Pérez-Rodríguez P (2017) Genomic models with genotype × environment interaction for predicting hybrid performance: an application in maize hybrids. Theor Appl Genet 130:1431–1440. doi: 10.1007/s00122-017-2898-0

Agarwala V, Flannick J, Sunyaev S, Altshuler D (2013) Evaluating empirical bounds on complex disease genetic architecture. Nat Genet 45:1418–1427. doi: 10.1038/ng.2804

Agustini BL, Francis A, Glen M, Indrayadi H, Mohammed CL (2014) Signs and identification of fungal root-rot pathogens in tropical Eucalyptus pellita plantations. For Pathol 44:486–495. doi: 10.1111/efp.12145

Albrecht T, Wimmer V, Auinger HJ, Erbe M, Knaak C, Ouzunova M, Simianer H, Schön CC (2011) Genome-based prediction of testcross values in maize. Theor Appl Genet 123:339–350. doi: 10.1007/s00122-011-1587-7

Allendorf FW, Hohenlohe PA, Luikart G (2010) Genomics and the future of conservation genetics. Nat Rev Genet 11:697–709. doi: 10.1038/nrg2844

Allwright MR, Payne A, Emiliani G, Milner S, Viger M, Rouse F, Keurentjes JJB, Bérard A, Wildhagen H, Faivre-Rampant P, Polle A, Morgante M, Taylor G (2016) Biomass traits and candidate genes for bioenergy revealed through association genetics in coppiced European Populus nigra (L.). Biotechnol Biofuels 9:195. doi: 10.1186/s13068-016-0603-1

Alves AA, Rosado CCG, Faria DA, da Guimarães LMS, Lau D, Brommonschenkel SH, Grattapaglia D, Alfenas AC (2012) Genetic mapping provides evidence for the role of additive and non-additive QTLs in the response of inter-specific hybrids of Eucalyptus to Puccinia psidii rust infection. Euphytica 183:27–38. doi: 10.1007/s10681-011-0455-5

Annicchiarico P, Nazzicari N, Li X, Wei Y, Pecetti L, Brummer EC (2015) Accuracy of genomic selection for alfalfa biomass yield in different reference populations. BMC Genomics 16:1020. doi: 10.1186/s12864-015-2212-y

Annicchiarico P, Nazzicari N, Pecetti L, Romani M, Ferrari B, Wei Y, Brummer EC (2017) GBS-Based Genomic Selection for Pea Grain Yield under Severe Terminal Drought. Plant Genome 10:0. doi: 10.3835/plantgenome2016.07.0072

Arnold R, Li B, Luo J, Bai F, Baker T (2015) Selection of cold-tolerant *Eucalyptus* species and provenances for inland frost-susceptible, humid subtropical regions of southern China. Aust For 9158:1–14. doi: 10.1080/00049158.2015.1063471

Arruda MP, Brown PJ, Lipka AE, Krill AM, Thurber C, Kolb FL (2015) Genomic Selection for Predicting Head Blight Resistance in a Wheat Breeding Program. Plant Genome 8:0. doi: 10.3835/plantgenome2015.01.0003

Astle W, Balding D (2009) Population Structure and Cryptic Relatedness in Genetic Association Studies. Stat Sci 24:451–471. doi: 10.1214/09-STS307

Auinger H-J, Schönleben M, Lehermeier C, Schmidt M, Korzun V, Geiger HH, Piepho H-P, Gordillo A, Wilde P, Bauer E, Schön C-C (2016) Model training across multiple breeding cycles significantly improves genomic prediction accuracy in rye (*Secale cereale* L.). Theor Appl Genet 1–11. doi: 10.1007/s00122-016-2756-5

Baccarin FJB, Brondani GE, de Almeida LV, Vieira IG, de Oliveira LS, de Almeida M (2015) Vegetative rescue and cloning of *Eucalyptus benthamii* selected adult trees. New For 46:465–483. doi: 10.1007/s11056-015-9472-x

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. Sci Rep 6:1–9. doi: 10.1038/srep32894

Bartholomé J, Bink MC, Van Heerwaarden J, Chancerel E, Boury C, Lesur I, Isik F, Bouffier L, Plomion C (2016a) 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. doi: 10.1371/journal.pone.0165323

Bartholomé J, Salmon F, Vigneron P, Bouvet JM, Plomion C, Gion JM (2013) Plasticity of primary and secondary growth dynamics in Eucalyptus hybrids: a quantitative genetics and QTL mapping perspective. BMC Plant Biol 13:120. doi: 10.1186/1471-2229-13-120

Bartholomé J, Van Heerwaarden J, Isik F, Boury C, Vidal M, Plomion C, Bouffier L (2016b) Performance of genomic prediction within and across generations in maritime pine. BMC Genomics 17:604. doi: 10.1186/s12864-016-2879-8

Bassi FM, Bentley AR, Charmet G, Ortiz R, Crossa J (2015) Breeding schemes for the implementation of genomic selection in wheat (*Triticum* spp.). Plant Sci 242:23–36. doi: 10.1016/j.plantsci.2015.08.021

Bates D, Mächler M, Bolker B, Walker S (2015) Fitting Linear Mixed-Effects Models using lme4. J Stat Softw 67:1–48. doi: 10.18637/jss.v067.i01

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

Beaulieu J, Doerksen TK, MacKay J, Rainville A, Bousquet J (2014b) Genomic selection accuracies within and between environments and small breeding groups in white spruce. BMC Genomics 15:1048. doi: 10.1186/1471-2164-15-1048

Begum H, Spindel JE, Lalusin A, Borromeo T, Gregorio G, Hernandez J, Virk P, Collard B, McCouch SR (2015) Genome-wide association mapping for yield and other agronomic traits in an elite breeding population of tropical rice (Oryza sativa). PLoS One 10:1–19. doi: 10.1371/journal.pone.0119873

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. R Stat Soc 57:289–300.

Benson D (1985) Aspects of the ecology of a rare tree species, *Eucalyptus benthamii*, at Bents Basin, Wallacia. Cunninghamia 1:371–383.

Bernal-Vasquez A-M, Gordillo A, Schmidt M, Piepho H-P (2017) Genomic prediction in early selection stages using multi-year data in a hybrid rye breeding program. BMC Genet 18:51. doi: 10.1186/s12863-017-0512-8

Bernal Rubio YL, Gualdrõn Duarte JL, Bates RO, Ernst CW, Nonneman D, Rohrer GA, King A, Shackelford SD, Wheeler TL, Cantet RJC, Steibel JP (2016) Meta-analysis of genome-wide association from genomic prediction models. Anim Genet 47:36–48. doi: 10.1111/age.12378

Bernardo AL, Reis MGF, Reis GG, Harrison RB, Firme DJ (1998) Effect of spacing on growth and biomass distribution in Eucalyptus camaldulensis , E . pellita and E . urophylla plantations in southeastern Brazil.

Bernardo R (2008) Molecular markers and selection for complex traits in plants: Learning from the last 20 years. Crop Sci 48:1649–1664. doi: 10.2135/cropsci2008.03.0131
Bernardo R (2017) Prospective Targeted Recombination and Genetic Gains for Quantitative Traits in Maize. Plant Genome 10:0. doi: 10.3835/plantgenome2016.11.0118
Bianco L, Cestaro A, Linsmith G, Muranty H, Denancé C, Théron A, Poncet C, Micheletti D, Kerschbamer E, Di Pierro EA, Larger S, Pindo M, Van De Weg E, Davassi A, Laurens F, Velasco R, Durel CE, Troggio M (2016) Development and validation of the Axiom®Apple480K SNP genotyping array. Plant J 86:62–74. doi: 10.1111/tpj.13145
Bolormaa S, Porto Neto LR, Zhang YD, Bunch RJ, Harrison BE, Goddard ME, Barendse W (2011) A genome-wide association study of meat and carcass traits in australian cattle. J Anim Sci 89:2297–2309. doi: 10.2527/jas.2010-3138
Bolormaa S, Pryce JE, Hayes BJ, Goddard ME (2010) Multivariate analysis of a genome-wide association study in dairy cattle. J Dairy Sci 93:3818–3833. doi: 10.3168/jds.2009-2980
Booth TH (2012) Eucalypts and Their Potential for Invasiveness Particularly in Frost-Prone Regions. doi: 10.1155/2012/837165
Borcard D, Gillet F, Legendre P (2011) Numerical Ecology with R, 1st edn. Springer-Verlag, New York
Bourquin V (2002) Xyloglucan Endotransglycosylases Have a Function during the Formation of Secondary Cell Walls of Vascular Tissues. Plant Cell Online 14:3073–3088. doi: 10.1105/tpc.007773
Boyle EA, Li YI, Pritchard JK (2017) An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell 169:1177–1186. doi: 10.1016/j.cell.2017.05.038
Brachi B, Morris GP, Borevitz JO (2011) Genome-wide association studies in plants: the missing heritability is in the field. Genome Biol 12:232. doi: 10.1186/gb-2011-12-10-232
Bragg JG, Supple MA, Andrew RL, Borevitz JO (2015) Genomic variation across landscapes: insights and applications. New Phytol 207:953–967.
Brawner JT, Bush DJ, Macdonell PF, Warburton PM, Clegg PA (2010) Genetic parameters of red mahogany breeding populations grown in the tropics. Aust For 73:177–183. doi: 10.1080/00049158.2010.10676324
Brondani GE, de Wit Ondas HW, Baccarin FJB, Gonçalves AN, de Almeida M (2012a) Micropropagation of *Eucalyptus benthamii* to form a clonal micro-garden. Vitr Cell Dev Biol - Plant 48:478–487. doi: 10.1007/s11627-012-9449-9
Brondani GE, Dutra LF, Wendling I, Grossi F, Hansel FA, Araujo MA (2011) Micropropagation of an *Eucalyptus* hybrid (*Eucalyptus benthamii* x *Eucalyptus dunnii*). Acta Sci Agron 33:655–663. doi: 10.4025/actasciagron.v33i4.8317
Brondani GE, Wendling I, Brondani AE, Araujo MA, Silva ALL Da, Gonçalves AN (2012b) Dynamics of adventitious rooting in mini-cuttings of *Eucalyptus benthamii* x *Eucalyptus dunnii*. Acta Sci Agron 34:169–178. doi: 10.4025/actasciagron.v34i2.13059
Brondani RP, Williams ER, Brondani C, Grattapaglia D (2006) A microsatellite-based consensus linkage map for species of *Eucalyptus* and a novel set of 230 microsatellite markers for the genus. BMC Plant Biol 6:20. doi: 10.1186/1471-2229-

6-20
Brooker MIH (2000) A new classification of the genus *Eucalyptus* L'Her. (Myrtaceae). Aust Syst Bot 13:79–148. doi: 10.1071/SB98008
Brunner AM, Li J, Difazio SP, Shevchenko O, Montgomery BE, Mohamed R, Wei H, Ma C, Elias AA, Vanwormer K, Strauss SH (2007) Genetic containment of forest plantations. 527:75–100. doi: 10.1007/s11295-006-0067-8
Bryan AC, Jawdy S, Gunter L, Gjersing E, Sykes R, Hinchee MAW, Winkeler KA, Collins CM, Engle N, Tschaplinski TJ, Yang X, Tuskan GA, Muchero W, Chen JG (2016) Knockdown of a laccase in Populus deltoides confers altered cell wall chemistry and increased sugar release. Plant Biotechnol J 14:2010–2020. doi: 10.1111/pbi.12560
Bundock PC, Potts BM, Vaillancourt RE (2008) Detection and stability of quantitative trait loci (QTL) in *Eucalyptus globulus*. Tree Genet Genomes 4:85–95. doi: 10.1007/s11295-007-0090-4
Burstin J, Salloignon P, Chabert-martinello M, Siol M, Jacquin F, Chauveau A, Pont C, Aubert G, Delaitre C, Truntzer C, Duc G (2015) Genetic diversity and trait genomic prediction in a pea diversity panel. 1–17. doi: 10.1186/s12864-015-1266-1
Burton P, Clayton D, Cardon L, WTCCC (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661–78. doi: 10.1038/nature05911
Butcher PA, McDonald MW, Bell JC (2009) Congruence between environmental parameters, morphology and genetic structure in Australia's most widely distributed eucalypt, *Eucalyptus camaldulensis*. Tree Genet Genomes 5:189–210. doi: 10.1007/s11295-008-0169-6
Butcher PA, Skinner AK, Gardiner CA (2005) Increased inbreeding and inter-species gene flow in remnant populations of the rare Eucalyptus benthamii. Conserv Genet 6:213–226. doi: 10.1007/s10592-004-7830-x
Butler JB, Freeman JS, Vaillancourt RE, Potts BM, Glen M, Lee DJ, Pegg GS (2016) Evidence for different QTL underlying the immune and hypersensitive responses of *Eucalyptus globulus* to the rust pathogen *Puccinia psidii*. Tree Genet Genomes 12:39. doi: 10.1007/s11295-016-0987-x
Callaway E (2017) Genome studies attract criticism. Nature 546:463–463. doi: 10.1038/nature.2017.22152
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. doi: 10.1371/journal.pone.0081267
Cappa EP, El-Kassaby YA, Muñoz F, Garcia MN, Villalba P V., Klápště J, Marcucci Poltri SN (2017) Improving accuracy of breeding values by incorporating genomic information in spatial-competition mixed models. Mol Breed 37:125. doi: 10.1007/s11032-017-0725-6
Casellas J, Piedrafita J (2015) Accuracy and expected genetic gain under genetic or genomic evaluation in sheep flocks with different amounts of pedigree, genomic and phenotypic data. Livest Sci 182:58–63. doi: 10.1016/j.livsci.2015.10.014
Chang H-X, Brown PJ, Lipka AE, Domier LL, Hartman GL (2016) Genome-wide association and genomic prediction identifies associated loci and predicts the sensitivity of Tobacco ringspot virus in soybean plant introductions. BMC Genomics

17:153. doi: 10.1186/s12864-016-2487-7
Cheng KF, Chen JH (2013) Detecting Rare Variants in Case-Parents Association Studies. PLoS One. doi: 10.1371/journal.pone.0074310
Ching A, Caldwell KS, Jung M, Dolan M, Oscar S, Smith H, Tingey S, Morgante M, Rafalski AJ (2002) SNP frequency , haplotype structure and linkage disequilibrium in elite maize inbred lines. 14:1–14.
Christie N, Tobias PA, Naidoo S, Külheim C (2016) The Eucalyptus grandis NBS-LRR Gene Family: Physical Clustering and Expression Hotspots. Front Plant Sci 6:1–16. doi: 10.3389/fpls.2015.01238
Clark SA, Hickey JM, Daetwyler HD, Werf JHJ Van Der (2012) The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. 1–9.
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. Proc Natl Acad Sci U S A 104:8550–5. doi: 10.1073/pnas.0703133104
Collard BCY, Mackill DJ, B PTRS (2008) Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. Phil Trans R Soc B 363:557–572. doi: 10.1098/rstb.2007.2170
Cosgrove DJ (2005) Growth of the plant cell wall. Nat Rev Mol Cell Biol 6:850–861. doi: 10.1038/nrm1746
Costa R, Estopa R, Biernaski F, Mori E (2016) Prediction of genetics gains in *Eucalyptus benthamii* Maiden & Cambage progenies by different selection methods. Sci For 44:105–113.
Coster A, Bastiaansen JWM, Calus MPL, van Arendonk JAM, Bovenhuis H (2010) Sensitivity of methods for estimating breeding values using genetic markers to the number of QTL and distribution of QTL variance. Genet Sel Evol 42:9. doi: 10.1186/1297-9686-42-9
Cros D, Denis M, Sánchez L, Cochard B, Flori A, Durand-Gasselin T, Nouy B, Omoré A, Pomiès V, Riou V, Suryana E, Bouvet JM (2015) Genomic selection prediction accuracy in a perennial crop: case study of oil palm (Elaeis guineensis Jacq.). Theor Appl Genet 128:397–410. doi: 10.1007/s00122-014-2439-z
Crossa J, Beyene Y, Kassa S, Pérez P, Hickey JM, Chen C, Campos GDL, Burgueño J, Windhausen VS, Buckler E, Jannink J, Cruz MAL, Babu R (2013) Genomic Prediction in Maize Breeding Populations with Genotyping-by-Sequencing. G3 Genes, Genomes, Genet 3:1903–1926. doi: 10.1534/g3.113.008227
Crossa J, De Los Campos G, Pérez P, Gianola D, Burgueño J, Araus JL, Makumbi D, Singh RP, Dreisigacker S, Yan J, Arief V, Banziger M, Braun HJ (2010) Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186:713–724. doi: 10.1534/genetics.110.118521
Crossa J, Pérez P, Hickey J, Burgueño J, Ornella L, Cerón-Rojas J, Zhang X, Dreisigacker S, Babu R, Li Y, Bonnett D, Mathews K (2014) Genomic prediction in CIMMYT maize and wheat breeding programs. Heredity (Edinb) 112:48–60. doi: 10.1038/hdy.2013.16
Crowhurst RN, Troggio M, Davey MW, Gilmore B, Chagne D, Vanderzande S, Hellens

RP, Kumar S, Cestaro A, Velasco R, Main D, Rees JD, Iezzoni A, Mockler T, Wilhelm L, Weg E Van De, Gardiner SE, Bassil N, Peace C (2012) Genome-Wide SNP Detection , Validation , and Development of an 8K SNP Array for Apple. doi: 10.1371/journal.pone.0031745

Cumbie WP, Eckert A, Wegrzyn J, Whetten R, Neale D, Goldfarb B (2011) Association genetics of carbon isotope discrimination, height and foliar nitrogen in a natural population of Pinus taeda L. Heredity (Edinb) 107:105–114. doi: 10.1038/hdy.2010.168

da Silva PHM, Sebbenn AM, Grattapaglia D, Conti Jr LJF (2017) Forest Ecology and Management Realized pollen flow and wildling establishment from a genetically modified eucalypt field trial in Southeastern Brazil. For Ecol Manage 385:161–166. doi: 10.1016/j.foreco.2016.11.043

Daetwyler H, Calus M, Pong-Wong R, de los Campos G, Hickey J (2013) Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. Genetics 193:347–365. doi: 10.1534/genetics.112.147983

Daetwyler HD, Pong-Wong R, Villanueva B, Woolliams JA (2010) The impact of genetic architecture on genome-wide evaluation methods. Genetics 185:1021–1031. doi: 10.1534/genetics.110.116855

Davey J, Cezard T, Fuentes-Utrilla P, Eland C, Gharbi K, Blaxter M (2013) Special features of RAD Sequencing data : implications for genotyping. Mol Ecol 22:3151–3164. doi: 10.1111/mec.12084

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat Rev Genet 12:499–510. doi: 10.1038/nrg3012

de Almeida Filho JE, Guimarães JFR, e Silva FF, de Resende MD V, Muñoz P, Kirst M, Resende MFR (2016) The contribution of dominance to phenotype prediction in a pine breeding and simulated population. Heredity (Edinb) 117:33–41. doi: 10.1038/hdy.2016.23

de Godoy F, Bermúdez L, Lira BS, De Souza AP, Elbl P, Demarco D, Alseekh S, Insani M, Buckeridge M, Almeida J, Grigioni G, Fernie AR, Carrari F, Rossi M (2013) Galacturonosyltransferase 4 silencing alters pectin composition and carbon partitioning in tomato. J Exp Bot 64:2449–2466. doi: 10.1093/jxb/ert106

de los Campos G, Gianola D, Rosa GJM, Weigel KA, Crossa J (2010) Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. Genet Res (Camb) 92:295–308. doi: 10.1017/S0016672310000285

de los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MPL (2013a) Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics 193:327–345. doi: 10.1534/genetics.112.143313

de los Campos G, Naya H, Gianola D, Crossa J, Legarra A, Manfredi E, Weigel K, Cotes JM (2009) Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics 182:375–385. doi: 10.1534/genetics.109.101501

de los Campos G, Sorensen D, Gianola D (2015) Genomic heritability: what is it? PLoS Genet 11:1–21. doi: 10.1371/journal.pgen.1005048

de los Campos G, Vazquez AI, Fernando R, Klimentidis YC, Sorensen D (2013b) Prediction of Complex Human Traits Using the Genomic Best Linear Unbiased

Predictor. PLoS Genet. doi: 10.1371/journal.pgen.1003608
Denis M, Bouvet JM (2013) Efficiency of genomic selection with models including dominance effect in the context of Eucalyptus breeding. Tree Genet Genomes 9:37–51. doi: 10.1007/s11295-012-0528-1
Desta ZA, Ortiz R (2014) Genomic selection: Genome-wide prediction in plant improvement. Trends Plant Sci 19:592–601. doi: 10.1016/j.tplants.2014.05.006
DeWan A, Liu M, Hartman S, Zhang S, Liu D, Zhao C, Tam P, Chan W, Lam D, Snyder M, Barnstable C, Pang C, Hoh J (2006) HTRA1 promoter polymorphism in wet age-related macular degeneration. Science (80-) 314:989–992.
Do C, Waples RS, Peel D, Macbeth GM, Tillett BJ, Ovenden JR (2014) NeEstimator v2: Re-implementation of software for the estimation of contemporary effective population size (Ne) from genetic data. Mol Ecol Resour 14:209–214. doi: 10.1111/1755-0998.12157
Döll-Boscardin PM, Sartoratto A, De Noronha Sales Maia BHL, Padilha De Paula J, Nakashima T, Farago PV, Kanunfre CC (2012) In vitro cytotoxic potential of essential oils of *Eucalyptus benthamii* and its related terpenes on tumor cell lines. Evidence-based Complement Altern Med 39:1–8. doi: 10.1155/2012/342652
dos Santos PET, Filho EP, da Silva LT de M, Vandresen PB (2015) Genetic variation for growth and selection in adult plants of *Eucalyptus badjensis*. Genet Mol Biol 38:457–464. doi: 10.1590/S1415-475738420150041
Doughty R (2000) The *Eucalyptus*: a natural and commercial history of the gum tree, Johns Hopk. Johns Hopkins University Press, London, UK
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 Phytol 209:1067–1082. doi: 10.1111/nph.13695
Duangjit J, Causse M, Sauvage C (2016) Efficiency of genomic selection for tomato fruit quality. Mol Breed. doi: 10.1007/s11032-016-0453-3
Duhnen A, Gras A, Teyssèdre S, Romestant M, Claustres B, Daydé J, Mangin B (2017) Genomic selection for yield and seed protein content in Soybean: A study of breeding program data and assessment of prediction accuracy. Crop Sci 57:1325–1337. doi: 10.2135/cropsci2016.06.0496
Durán R, Isik F, Zapata-Valenzuela J, Balocchi C, Valenzuela S (2017) Genomic predictions of breeding values in a cloned Eucalyptus globulus population in Chile. Tree Genet Genomes. doi: 10.1007/s11295-017-1158-4
Earl DA, vonHoldt BM (2012) STRUCTURE HARVESTER: A website and program for visualizing STRUCTURE output and implementing the Evanno method. Conserv Genet Resour 4:359–361. doi: 10.1007/s12686-011-9548-7
Ebling G, Francisco B, Benedini J, Heron B, Wit W De, Luiz J, Antonio S, Gonçalves N, Almeida M De (2012) Low temperature , IBA concentrations and optimal time for adventitious rooting of Eucalyptus benthamii mini-cuttings. doi: 10.1007/s11676-012-0298-5
Eckert AJ, Bower AD, GonzÁlez-MartÍnez SC, Wegrzyn JL, Coop G, Neale DB (2010) Back to nature: Ecological genomics of loblolly pine (Pinus taeda, Pinaceae). Mol Ecol 19:3789–3805. doi: 10.1111/j.1365-294X.2010.04698.x
Edwards SL, Beesley J, French JD, Dunning M (2013) Beyond GWASs: Illuminating the

dark road from association to function. Am J Hum Genet 93:779–797. doi: 10.1016/j.ajhg.2013.10.012
Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH (2010) Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet 11:446–450. doi: 10.1038/nrg2809
El-Dien OG, Ratcliffe B, Klapste J, Porth I, Chen C, El-Kassaby YA, Gamal El-Dien O, Ratcliffe B, Klap t J, Porth I, Chen C, El-Kassaby YA, Klap t J, Porth I, Chen C, El-Kassaby YA (2016) Implementation of the realized genomic relationship matrix to open-pollinated white spruce family testing for disentangling additive from non-additive genetic effects. G3 Genes, Genomes, Genet 6:743–753. doi: 10.1534/g3.115.025957
Eldridge K, Davidson J, Harwood C, Wyk G van (1993) Eucalypt domestication and breeding. Oxford University Press, Oxford, UK
Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6:1–10. doi: 10.1371/journal.pone.0019379
Endelman JB (2011) Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome J 4:250. doi: 10.3835/plantgenome2011.08.0024
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. doi: 10.1016/j.phytochem.2006.05.022
Eu-ahsunthornwattana J, Miller EN, Fakiola M, Jeronimo SMB, Blackwell JM, Cordell HJ (2014) Comparison of Methods to Account for Relatedness in Genome-Wide Association Studies with Family-Based Data. PLoS Genet. doi: 10.1371/journal.pgen.1004445
Evangelou E, Ioannidis JPA (2013) Meta-analysis methods for genome-wide association studies and beyond. Nat Rev Genet 14:379–389. doi: 10.1038/nrg3472
Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. Mol Ecol 14:2611–2620. doi: 10.1111/j.1365-294X.2005.02553.x
Evans LM, Slavov GT, Rodgers-Melnick E, Martin J, Ranjan P, Muchero W, Brunner AM, Schackwitz W, Gunter L, Chen J-G, Tuskan GA, DiFazio SP (2014) Population genomics of Populus trichocarpa identifies signatures of selection and adaptive trait associations. Nat Genet 46:1089–1096. doi: 10.1038/ng.3075
Fahrenkrog AM, Neves LG, Resende MFR, Vazquez AI, de los Campos G, Dervinis C, Sykes R, Davis M, Davenport R, Barbazuk WB, Kirst M (2016) Genome-wide association study reveals putative regulators of bioenergy traits in Populus deltoides. New Phytol 213:799–811. doi: 10.1111/nph.14154
Faivre-Rampant P, Zaina G, Jorge V, Giacomello S, Segura V, Scalabrin S, Guérin V, De Paoli E, Aluome C, Viger M, Cattonaro F, Payne A, PaulStephenRaj P, Le Paslier M, Berard A, Allwright MR, Villar M, Taylor G, Bastien C, Morgante M (2016) New resources for genetic studies in *Populus nigra* : genome wide SNP discovery and development of a 12k Infinium array. Mol Ecol Resour n/a-n/a. doi: 10.1111/1755-0998.12513
Fan H, Wu Y, Zhou X, Xia J, Zhang W, Song Y, Liu F, Chen Y, Zhang L, Gao X, Gao H,

Li J (2015) Pathway-based genome-wide association studies for two meat production traits in simmental cattle. Sci Rep 5:18389. doi: 10.1038/srep18389
Fiedler JD, Salsman E, Liu Y, Michalak de Jiménez M, Hegstad JB, Chen B, Manthey FA, Chao S, Xu S, Elias EM, Li X (2017) Genome-Wide Association and Prediction of Grain and Semolina Quality Traits in Durum Wheat Breeding Populations. Plant Genome 0:0. doi: 10.3835/plantgenome2017.05.0038
Fisher RA (1918) The correlation between relatives on the supposition of Mendelian inheritance. Trans R Soc Edinburgh 52:399–433. doi: 10.1074/jbc.M107031200
Fitz-Gibbon S, Hipp AL, Pham KK, Manos PS, Sork VL (2017) Phylogenomic inferences from reference-mapped and de novo assembled short-read sequence data using RADseq sequencing of California white oaks (Quercus section Quercus). Genome 60:743–755. doi: 10.1139/gen-2016-0202
Forneris NS, Steibel JP, Legarra A, Vitezica ZG, Bates RO, Ernst CW, Basso AL (2016) A comparison of methods to estimate genomic relationships using pedigree and markers in livestock populations. 1–11. doi: 10.1111/jbg.12217
Fort PO, Carson M, Difazio SP, Slavov GT, Burczyk J, Leonardi S, Strauss SH, Division ES, Ridge O, Ridge O (2004) Gene flow from tree plantations and implications for transgenic risk assessment. 661:405–422.
Fortes MRS, Reverter A, Zhang Y, Collis E, Nagaraj SH, Jonsson NN (2010) Association weight matrix for the genetic dissection of puberty in beef cattle. Proc Natl Acad Sci 107:13642–13647. doi: 10.1073/pnas.1002044107/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1002044107
Francis RM (2016) POPHELPER: An R package and web app to analyse and visualise population structure. Mol Ecol Resour 17:27–32. doi: 10.1111/1755-0998.12509
Freeman JS, Potts BM, Vaillancourt RE (2008) Few Mendelian genes underlie the quantitative response of a forest tree, eucalyptus globulus, to a natural fungal epidemic. Genetics 178:563–571. doi: 10.1534/genetics.107.081414
Fuentes-Utrilla P, Goswami C, Cottrell JE, Pong-Wong R, Law A, A'Hara SW, Lee SJ, Woolliams JA (2017) QTL analysis and genomic selection using RADseq derived markers in Sitka spruce: the potential utility of within family data. Tree Genet Genomes. doi: 10.1007/s11295-017-1118-z
Gamal El-Dien O, Ratcliffe B, Klápště J, Chen C, Porth I, El-Kassaby YA (2015) Prediction accuracies for growth and wood attributes of interior spruce in space using genotyping-by-sequencing. BMC Genomics 16:370. doi: 10.1186/s12864-015-1597-y
Geraldes A, Difazio SP, Slavov GT, Ranjan P, Muchero W, Hannemann J, Gunter LE, Wymore AM, Grassa CJ, Farzaneh N, Porth I, McKown AD, Skyba O, Li E, Fujita M, Klápště J, Martin J, Schackwitz W, Pennacchio C, Rokhsar D, Friedmann MC, Wasteneys GO, Guy RD, El-Kassaby YA, Mansfield SD, Cronk QCB, Ehlting J, Douglas CJ, Tuskan GA (2013) A 34K SNP genotyping array for *Populus trichocarpa*: design, application to the study of natural populations and transferability to other *Populus* species. Mol Ecol Resour 13:306–23. doi: 10.1111/1755-0998.12056
Gezan SA, Osorio LF, Verma S, Whitaker VM (2017) An experimental validation of genomic selection in octoploid strawberry. Hortic Res 4:16070. doi: 10.1038/hortres.2016.70
Gianola D, de los Campos G (2008) Inferring genetic values for quantitative traits non-

parametrically. Genet Res (Camb) 90:525. doi: 10.1017/S0016672308009890
Gianola D, Fernando RL, Stella A (2006) Genomic-assisted prediction of genetic value with semiparametric procedures. Genetics 173:1761–1776. doi: 10.1534/genetics.105.049510
Goddard ME, Hayes BJ (2009) Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nat Rev Genet 10:381–391. doi: nrg2575 [pii]\r10.1038/nrg2575
Goddard ME, Wray NR, Verbyla K, Visscher PM (2009) Estimating effects and making predictions from genome-wide marker data. Stat Sci 24:517–529. doi: 10.1214/09-STS306
González-Camacho JM, de los Campos G, Pérez P, Gianola D, Cairns JE, Mahuku G, Babu R, Crossa J (2012) Genome-enabled prediction of genetic values using radial basis function neural networks. Theor Appl Genet 125:759–771. doi: 10.1007/s00122-012-1868-9
Gowda M, Zhao Y, Würschum T, Longin CF, Miedaner T, Ebmeyer E, Schachschneider R, Kazman E, Schacht J, Martinant J-P, Mette MF, Reif JC (2014) Relatedness severely impacts accuracy of marker-assisted selection for disease resistance in hybrid wheat. Heredity (Edinb) 112:552–561. doi: 10.1038/hdy.2013.139
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, 1st edn. Springer International Publishing, pp 199–250
Grattapaglia D (2014) Breeding forest trees by genomic selection: current progress and the way forward. In: Tuberosa R, Graner A, Frison E (eds) Genomics of Plant Genetic Resources, 1st edn. Springer New York Heidelberg Dordrecht London, New York, pp 651–682
Grattapaglia D, Kirst M (2008) Eucalyptus applied genomics: from gene sequences to breeding tools. New Phytol 179:911–929. doi: 10.1111/j.1469-8137.2008.02503.x
Grattapaglia D, Plomion C, Kirst M, Sederoff RR (2009) Genomics of growth traits in forest trees. Curr Opin Plant Biol 12:148–56. doi: 10.1016/j.pbi.2008.12.008
Grattapaglia D, Resende MD V (2011) Genomic selection in forest tree breeding. Tree Genet Genomes 7:241–255. doi: 10.1007/s11295-010-0328-4
Grattapaglia D, Silva-Junior OB, Kirst M, de Lima B, Faria DA, Pappas GJ (2011) High-throughput SNP genotyping in the highly heterozygous genome of Eucalyptus: assay success, polymorphism and transferability across species. BMC Plant Biol 11:65. doi: 10.1186/1471-2229-11-65
Grattapaglia D, Vaillancourt RE, Myburg AA (2012) Progress in Myrtaceae genetics and genomics: Eucalyptus as the pivotal genus. Tree Genet Genomes 8:463–508. doi: 10.1007/s11295-012-0491-x
Grenier C, Cao T, Ospina Y, Quintero C (2015) Accuracy of genomic selection in a rice synthetic population developed for recurrent selection breeding. 1–25. doi: 10.1371/journal.pone.0136594
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 Phytol 197:162–176. doi: 10.1111/nph.12003
Gupta RKSBB, Jr CNS (2013) From genomics to functional markers in the era of next-generation sequencing. doi: 10.1007/s10529-013-1377-1

Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx BWJH, Jansen R, Geus EJC De, Boomsma DI, Wright FA, Sullivan PF, Nikkola E (2016) Integrative approaches for large-scale transcriptome-wide association studies. Nat Publ Gr 48:245–252. doi: 10.1038/ng.3506

Habier D, Fernando RL, Dekkers JCM (2009) Genomic selection using low-density marker panels. Genetics 182:343–353. doi: 10.1534/genetics.108.100289

Habier D, Fernando RL, Dekkers JCM (2007) The Impact of Genetic Relationship Information on Genome-Assisted Breeding Values. Genetics 177:2389–2397. doi: 10.1534/genetics.107.081190

Habier D, Fernando RL, Garrick DJ (2013) Genomic BLUP decoded: A look into the black box of genomic prediction. Genetics 194:597–607. doi: 10.1534/genetics.113.152207

Habier D, Fernando RL, Kizilkaya K, Garrick DJ (2011) Extension of the bayesian alphabet for genomic selection. BMC Bioinformatics 12:186. doi: 10.1186/1471-2105-12-186

Habier D, Tetens J, Seefried F, Lichtner P, Thaller G (2010) The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genet Sel Evol 42:1–12.

Hamblin MT, Buckler ES, Jannink JL (2011) Population genetics of genomics-based crop improvement methods. Trends Genet 27:98–106. doi: 10.1016/j.tig.2010.12.003

Harwood C (2011) New introductions – doing it right. In: Walker J (ed) Developing a eucalypt resource: learning from Australia and elsewhere, Wood Techn. University of Canterbury, Christchurch, New Zealand, pp 43–54

Harwood C (1998) *Eucalyptus pellita*: an annotated bibliography, Internatio. Collingwood CSIRO Publishing 1998, Nairobi, Kenya

Harwood CE, Alloysius D, Pomroy P, Robson KW, Haines MW (1997) Early growth and survival of *Eucalyptus pellita* provenances in a range of tropical environments, compared with *E. grandis* , *E. urophylla* and *Acacia mangium*. New For 14:203–219. doi: 10.1023/A:1006524405455

Hayes B, Goddard M (2010) Genome-wide association and genomic selection in animal breeding. Genome 53:876–883. doi: 10.1139/G10-076

Hayes BJ, Bowman PJ, Chamberlain AC, Verbyla K, Goddard ME (2009a) Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genet Sel Evol 41:51. doi: 10.1186/1297-9686-41-51

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

Hayes BJ, Daetwyler HD, Goddard ME (2016a) Models for genome × environment interaction: examples in livestock. Crop Sci 0:0. doi: 10.2135/cropsci2015.07.0451

Hayes BJ, Donoghue KA, Reich CM, Mason BA, Bird-Gardiner T, Herd RM, Arthur PF (2016b) Genomic heritabilities and genomic estimated breeding values for methane traits in Angus cattle. J Anim Sci 94:902. doi: 10.2527/jas.2015-0078

Henderson R (1975) Best linear unbiased estimation and prediction under a selection model. Biometrics 31:423–447. doi: 10.2307/2529430

Henrique M, Santana DA, Antônio G, Junior O, Silva A, Cesar M, Freua MC, Gomes C, Leme PR, Fukumasu H, Carvalho ME, Ventura RV, Coutinho LL (2016) Copy

number variations and genome-wide associations reveal putative genes and metabolic pathways involved with the feed conversion ratio in beef cattle. doi: 10.1007/s13353-016-0344-7
Henry RJ (2011) Eucalyptus. In: Kole C (ed) Wild Crop Relatives: Genomic and Breeding Resources, Forest Trees. Springer-Verlag, Berlin Heidelberg,
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. J Biol Chem 279:13547–13554. doi: 10.1074/jbc.M312929200
Heslot N, Jannink J-L, Sorrells ME (2015) Perspectives for Genomic Selection Applications and Research in Plants. Crop Sci 55:1. doi: 10.2135/cropsci2014.03.0249
Hickey JM, Chiurugwi T, Mackay I, Powell W, Participants IGS in CBPW (2017) Genomic prediction unifies animal and plant breeding programs to form platforms for biological discovery. Natures Genet 49:1297–1303. doi: 10.1038/ng.3920
Hill WG, Weir BS (1988) Variances and covariances of squared linkage disequilibria in finite populations. Theor Popul Biol 33:54–78. doi: 10.1016/0040-5809(88)90004-4
House APN, Bell JC (1996) Genetic diversity, mating system and systematic relationships in two red mahoganies, *Eucalyptus pellita* and *E. scias*. Aust J Bot 44:157–174.
Huang X, Han B (2014) Natural Variations and Genome-Wide Association Studies in Crop Plants. Annu Rev Plant Biol 65:531–551. doi: 10.1146/annurev-arplant-050213-035715
Hudson CJ, Freeman JS, Kullan ARK, Petroli CD, Sansaloni CP, Kilian A, Detering F, Grattapaglia D, Potts BM, Myburg AA, Vaillancourt RE (2012) A reference linkage map for *Eucalyptus*. BMC Genomics 13:240. doi: 10.1186/1471-2164-13-240
Hung TD, Brawner JT, Meder R, Lee DJ, Southerton S, Thinh HH, Dieters MJ (2015) Estimates of genetic parameters for growth and wood properties in Eucalyptus pellita F . Muell. to support tree breeding in Vietnam. Ann For Sci 72:205–217. doi: 10.1007/s13595-014-0426-9
Isik F, Bartholomé J, Farjat A, Chancerel E, Raffin A, Sanchez L, Plomion C, Bouffier L (2015) Genomic selection in maritime pine. Plant Sci 242:108–119. doi: 10.1016/j.plantsci.2015.08.006
Iwata H, Hayashi T, Tsumura Y (2011) Prospects for genomic selection in conifer breeding: A simulation study of Cryptomeria japonica. Tree Genet Genomes 7:747–758. doi: 10.1007/s11295-011-0371-9
Iwata H, Minamikawa MF, Kajiya-Kanegae H, Ishimori M, Hayashi T (2016) Genomics-assisted breeding in fruit trees. Breed Sci 66:100–115. doi: 10.1270/jsbbs.66.100
Jackson DA (1993) Stopping rules in principal components analysis: a comparison of Heuristical and statistical approaches. Ecology 74:2204–2214.
Jakobsson M, Rosenberg NA (2007) CLUMPP: A cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics 23:1801–1806. doi: 10.1093/bioinformatics/btm233
Jannink J, Lorenz AJ, Iwata H (2010) Genomic selection in plant breeding: from theory to practice. Brief Funct Genomics 9:166–77. doi: 10.1093/bfgp/elq001
Jaramillo-Correa JP, Prunier J, Vázquez-Lobo A, Keller SR, Moreno-Letelier A (2015) Molecular Signatures of Adaptation and Selection in Forest Trees. Adv Bot Res

74:265–306. doi: 10.1016/bs.abr.2015.04.003
Jarquín D, Lemes da Silva C, Gaynor RC, Poland J, Fritz A, Howard R, Battenfield S, Crossa J (2017) Increasing Genomic-Enabled Prediction Accuracy by Modeling Genotype × Environment Interactions in Kansas Wheat. Plant Genome 10:0. doi: 10.3835/plantgenome2016.12.0130
Jiang L, Liu J, Sun D, Ma P, Ding X, Yu Y, Zhang Q (2010) Genome wide association studies for milk production traits in Chinese Holstein population. PLoS One. doi: 10.1371/journal.pone.0013661
Jiang L, Liu X, Yang J, Wang H, Jiang J, Liu L, He S, Ding X, Liu J, Zhang Q (2014) Targeted resequencing of GWAS loci reveals novel genetic variants for milk production traits. BMC Genomics 15:1–9. doi: 10.1186/1471-2164-15-1105
Jovanovic T, Booth T (2002) Improved species climatic profiles. A Rep RIRDC/L&W Aust MDBC Jt Ventur Agrofor Progr 74.
Juliana P, Singh RP, Singh PK, Crossa J, Rutkoski JE, Poland JA, Bergstrom GC, Sorrells ME (2017) Comparison of Models and Whole-Genome Profiling Approaches for Genomic-Enabled Prediction of Septoria Tritici Blotch, Stagonospora Nodorum Blotch, and Tan Spot Resistance in Wheat. Plant Genome 10:0. doi: 10.3835/plantgenome2016.08.0082
Kaler AS, Ray JD, Schapaugh WT, King CA, Purcell LC (2017) Genome-wide association mapping of canopy wilting in diverse soybean genotypes. Theor Appl Genet 1–15. doi: 10.1007/s00122-017-2951-z
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. Nat Genet 42:348–354. doi: 10.1038/ng.548
Khan MA, Korban SS (2012) Association mapping in forest trees and fruit crops. J Exp Bot 63:4045–4060. doi: 10.1093/jxb/ers105
Kirst M, Myburg AA, De León JP, Kirst ME, Scott J, Sederoff R (2004) Coordinated genetic regulation of growth and lignin revealed by quantitative trait locus analysis of cDNA microarray data in an interspecific backcross of Eucalyptus. Plant Physiol 135:2368–2378. doi: 10.1104/pp.103.037960
Klápště J, Lstibůrek M, El-Kassaby YA (2014) Estimates of genetic parameters and breeding values from western larch open-pollinated families using marker-based relationship. Tree Genet Genomes 10:241–249. doi: 10.1007/s11295-013-0673-1
Klein R, Zeiss C, Chew E, Tsai J, Sackler R, Haynes C, Henning A, SanGiovanni J, Mane S, Mayne S, Bracken M, Ferris F, Ott J, Barnstable C, Hoh J (2005) Complement factor H polymorphism in age-related macular degeneration. Science (80-) 308:385–389. doi: 10.1126/science.1109557
Korte A, Farlow A (2013) The advantages and limitations of trait analysis with GWAS: a review. Plant Methods 9:29. doi: 10.1186/1746-4811-9-29
Korte A, Vilhjálmsson BJ, Segura V, Platt A, Long Q, Nordborg M (2012) A mixed-model approach for genome-wide association studies of correlated traits in structured populations. Nat Genet 44:1066–1071. doi: 10.1038/ng.2376
Kumar S, Bink MCAM, Volz RK, Bus VGM, Chagné D (2012a) Towards genomic selection in apple (*Malus* × *domestica* Borkh.) breeding programmes: Prospects, challenges and strategies. Tree Genet Genomes 8:1–14. doi: 10.1007/s11295-011-0425-z

Kumar S, Chagné D, Bink MC a M, Volz RK, Whitworth C, Carlisle C (2012b) Genomic selection for fruit quality traits in apple (*Malus*×*domestica* Borkh.). PLoS One 7:e36674. doi: 10.1371/journal.pone.0036674

Kumar S, Molloy C, Munoz P, Daetwyler H, Chagne D, Volz R (2015) Genome-enabled estimates of additive and nonadditive genetic variances and prediction of apple phenotypes across environments. G3 Genes, Genomes, Genet 5:2711–2718. doi: 10.1534/g3.115.021105

Le S, Nock C, Henson M, Shepherd M (2009) Genetic differentiation among and within three red mahoganies (series *Annulares*), *Eucalyptus pellita*, *E. resinifera* and *E. scias* (Myrtaceae). Aust Syst Bot 22:332–343. doi: 10.1071/SB09004

Ledford H (2014) Brazil considers transgenic trees. Nature 2014.

Lee SH, DeCandia TR, Ripke S, Yang J, Sullivan PF, Goddard ME, Keller MC, Visscher PM, Wray NR (2012) Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. Nat Genet 44:247–250. doi: 10.1038/ng.1108

Lee SH, Van Der Werf JHJ, Hayes BJ, Goddard ME, Visscher PM (2008) Predicting unobserved phenotypes for complex traits from whole-genome SNP data. PLoS Genet. doi: 10.1371/journal.pgen.1000231

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

Legarra A, Robert-Granié C, Manfredi E, Elsen JM (2008) Performance of genomic selection in mice. Genetics 180:611–618. doi: 10.1534/genetics.108.088575

Lei L, Li S, Bashline L, Gu Y (2014) Dissecting the molecular mechanism underlying the intimate relationship between cellulose microfibrils and cortical microtubules. Front Plant Sci 5:1–8. doi: 10.3389/fpls.2014.00090

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. Plant Cell 25:4912–4923. doi: 10.1105/tpc.113.116715

Lei L, Li S, Gu Y (2012) Cellulose Synthase Complexes: Composition and Regulation. Front Plant Sci 3:1–6. doi: 10.3389/fpls.2012.00075

Leksono B, Kurinobu ÆS, Ide ÆY (2008) Realized genetic gains observed in second generation seedling seed orchards of Eucalyptus pellita in Indonesia. 110–116. doi: 10.1007/s10310-008-0061-0

Leksono B, Kurinobu S, Ide Y (2006) Optimum age for selection based on a time trend of genetic parameters related to diameter growth in seedling seed orchards of *Eucalyptus pellita* in Indonesia. J For Res 11:359–364. doi: 10.1007/s10310-006-0223-x

Lenz PRN, Beaulieu J, Mansfield SD, Clément S, Desponts M, Bousquet J (2017) Factors affecting the accuracy of genomic selection for growth and wood quality traits in an advanced-breeding population of black spruce (Picea mariana). BMC Genomics 18:335. doi: 10.1186/s12864-017-3715-5

Li YX, Li C, Bradbury PJ, Liu X, Lu F, Romay CM, Glaubitz JC, Wu X, Peng B, Shi Y, Song Y, Zhang D, Buckler ES, Zhang Z, Li Y, Wang T (2016) Identification of genetic variants associated with maize flowering time using an extremely large multi-genetic background population. Plant J 86:391–402. doi: 10.1111/tpj.13174

Li Z, Philipp N, Spiller M, Stiewe G, Reif JC (2017) Genome-Wide Prediction of the Performance of Three-Way Hybrids in Barley. Plant Genome 10:1–35. doi: 10.3835/plantgenome2016.05.0046

Lima BM De (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. University of São Paulo

Lin DY, Zeng D (2009) Meta-analysis of genome-wide association studies: no efficiency gain in using individual participant data. Genet Epidemiol 34:n/a-n/a. doi: 10.1002/gepi.20435

Lin M, Arnold R, Li B, Yang M (2003) Selection of cold-tolerant eucalypts for Hunan Province. In: Turnbull J (ed) Eucalypts in Asia: proceedings of a international conference held in Zhanjiang, Guangdong, people's Republic of China. ACIAR Proceedings, Zhanjiang, Guangdong, China, pp 107–116

Liu H, Zhou H, Wu Y, Li X, Zhao J, Zuo T, Zhang X, Zhang Y, Liu S, Shen Y, Lin H, Zhang Z, Huang K, Lübberstedt T, Pan G (2015) The Impact of Genetic Relationship and Linkage Disequilibrium on Genomic Selection. PLoS One 10:1–13. doi: 10.1371/journal.pone.0132379

Long N, Gianola D, Rosa GJM, Weigel KA (2011) Long-term impacts of genome-enabled selection. J Appl Genet 52:467–480. doi: 10.1007/s13353-011-0053-1

Lopez-Casado G, Urbanowicz BR, Damasceno CM, Rose JK (2008) Plant glycosyl hydrolases and biofuels: a natural marriage. Curr Opin Plant Biol 11:329–337. doi: 10.1016/j.pbi.2008.02.010

Lorenz AJ, Chao S, Asoro FG, Heffner EL, Hayashi T, Iwata H, Smith KP, Sorrells ME, Jannink JL (2011) Genomic selection in plant breeding. Knowledge and prospects.

Lu M, Krutovsky K V., Nelson CD, Koralewski TE, Byram TD, Loopstra CA (2016) Exome genotyping, linkage disequilibrium and population structure in loblolly pine (Pinus taeda L.). BMC Genomics 17:730. doi: 10.1186/s12864-016-3081-8

Lu M, Krutovsky K V., 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 Genet Genomes. doi: 10.1007/s11295-017-1140-1

Luan T, Woolliams JA, Lien S, Kent M, Svendsen M, Meuwissen THE (2009) The Accuracy of Genomic Selection in Norwegian Red Cattle Assessed by Cross-Validation. 1126:1119–1126. doi: 10.1534/genetics.109.107391

Ly D, Hamblin M, Rabbi I, Melaku G, Bakare M, Gauch HG, Okechukwu R, Dixon AGO, Kulakow P, Jannink JL (2013) Relatedness and genotype ?? environment interaction affect prediction accuracies in genomic selection: A study in cassava. Crop Sci 53:1312–1325. doi: 10.2135/cropsci2012.11.0653

Mägi R, Morris AP (2010) GWAMA: software for genome-wide association meta-analysis. BMC Bioinformatics 11:288. doi: 10.1186/1471-2105-11-288

Magosi LE, Goel A, Hopewell JC, Farrall M (2017) Identifying systematic heterogeneity patterns in genetic association meta-analysis studies. PLoS Genet 13:1–17. doi: 10.1371/journal.pgen.1006755

Maher B (2008) The case of the missing heritability. Nature 456:18–21. doi: doi:10.1038/456018a

Makowsky R, Pajewski NM, Klimentidis YC, Vazquez AI, Duarte CW, Allison DB, de los

Campos G (2011) Beyond missing heritability: Prediction of complex traits. PLoS Genet. doi: 10.1371/journal.pgen.1002051

Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E, Shendure J, Turner DJ (2010) Target-enrichment strategies for next-generation sequencing. Nat Methods 7:111–118. doi: 10.1038/nmeth.1419

Mammadov J, Aggarwal R, Buyyarapu R, Kumpatla S (2012) SNP Markers and Their Impact on Plant Breeding. doi: 10.1155/2012/728398

Mangin B, Siberchicot A, Nicolas S, Doligez A, This P, Cierco-Ayrolles C (2012) Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. Heredity (Edinb) 108:285–291. doi: 10.1038/hdy.2011.73

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TFC, McCarroll SA, Visscher PM (2009) Finding the missing heritability of complex diseases. Nature 461:747–753. doi: 10.1038/nature08494

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. J Exp Bot 60:3959–3972. doi: 10.1093/jxb/erp229

Marouli E, Graff M, Medina-Gomez C, Lo KS, Wood AR, Kjaer TR, Fine RS, Lu Y, Schurmann C, Highland HM, Rüeger S, Thorleifsson G, Justice AE, Lamparter D, Stirrups KE, Turcot V, et al (2017) Rare and low-frequency coding variants alter human adult height. Nature 542:186–190. doi: 10.1038/nature21039

Marques CM, Carocha VJ, Pereira De Sá AR, Oliveira MR, Pires AM, Sederoff R, Borralho NMG (2005) Verification of QTL linked markers for propagation traits in Eucalyptus. Tree Genet Genomes 1:103–108. doi: 10.1007/s11295-005-0013-1

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 Genet Genomes 7:1011–1023. doi: 10.1007/s11295-011-0391-5

Martins SA, Menezzi CHSD, Ferraz JM, Souza MR (2013) Bonding behavior of *Eucalyptus benthamii* wood to manufacture edge glued panels. Maderas Cienc y Tecnol 15:79–92. doi: 10.4067/S0718-221X2013

Massman JM, Jung HJG, Bernardo R (2013) Genomewide selection versus marker-assisted recurrent selection to improve grain yield and stover-quality traits for cellulosic ethanol in maize. Crop Sci 53:58–66. doi: 10.2135/cropsci2012.02.0112

Mauro L, Titon M, Lau D, Rosse LN, Samo L, Oliveira S, Cristina C, Rosado G, Gegenheiner G, Christo O, Alfenas AC (2010) Eucalyptus pellita as a source of resistance to rust , ceratocystis wilt and leaf blight. 124–131.

McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, LITTLE J, Ioannidis JPA, Hirschhorn JN (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. NatRevGenet 9:356–369. doi: 10.1038/nrg2344

Mccouch SR, Zhao K, Wright M, Tung C, Ebana K, Thomson M, Reynolds A, Wang D, Declerck G, Ali L, Mcclung A, Eizenga G, Bustamante C (2010) Development of

genome-wide SNP assays for rice. 535:524–535. doi: 10.1270/jsbbs.60.524
McHale L, Tan X, Koehl P, Michelmore RW (2006) Plant NBS-LRR proteins: adaptable guards. Genome Biol 7:212. doi: 10.1186/gb-2006-7-4-212
Mckown AD, Klápště J, Guy RD, Geraldes A, Porth I, Hannemann J, Friedmann M, Muchero W, Tuskan GA, Ehlting J, Cronk QCB, El-Kassaby YA, Mansfield SD, Douglas CJ (2014) Genome-wide association implicates numerous genes underlying ecological trait variation in natural populations of Populus trichocarpa. New Phytol 203:535–553. doi: 10.1111/nph.12815
Meuwissen T, Hayes B, Goddard M (2016) Genomic selection: a paradigm shift in animal breeding. Anim Front 6:6. doi: 10.2527/af.2016-0002
Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense markers maps. Genetics 157:1819–1829. doi: 11290733
Michel S, Ametz C, Gungor H, Epure D, Grausgruber H, LÃ¶schenberger F, Buerstmayr H (2016) Genomic selection across multiple breeding cycles in applied bread wheat breeding. Theor Appl Genet 129:1179–1189. doi: 10.1007/s00122-016-2694-2
Migault V, Pallas B, Costes E (2017) Combining Genome-Wide Information with a Functional Structural Plant Model to Simulate 1-Year-Old Apple Tree Architecture. Front Plant Sci 7:1–14. doi: 10.3389/fpls.2016.02065
Morrell PL, Buckler ES, Ross-Ibarra J (2011) Crop genomics: advances and applications. Nat Rev Genet 13:85–96. doi: 10.1038/nrg3097
Morris GP, Ramu P, Deshpande SP, Hash CT, Shah T, Upadhyaya HD, Riera-Lizarazu O, Brown PJ, Acharya CB, Mitchell SE, Harriman J, Glaubitz JC, Buckler ES, Kresovich S (2013) Population genomic and genome-wide association studies of agroclimatic traits in sorghum. Proc Natl Acad Sci U S A 110:453–8. doi: 10.1073/pnas.1215985110
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. doi: 10.1186/s12864-017-3920-2
Muñoz PR, Resende MFR, Gezan SA, Deon M, Resende V (2014) Unraveling additive from nonadditive effects using genomic relationship matrices. Genetics 198:1759–1768. doi: 10.1534/genetics.114.171322
Muranty H, Troggio M, Sadok I Ben, Rifaï M Al, Auwerkerken A, Banchi E, Velasco R, Stevanato P, van de Weg WE, Di Guardo M, Kumar S, Laurens F, Bink MCAM (2015) Accuracy and responses of genomic selection on key traits in apple breeding. Hortic Res 2:15060. doi: 10.1038/hortres.2015.60
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. Springer-Verlag, Berlin Heidelberg, pp 115–160
Myburg A, Grattapaglia D, Tuskan G, Hellsten U, Hayes RD, Grimwood J, Jenkins J, Lindquist E, Tice H, Bauer D, Goodstein DM, Dubchak I, Poliakov A, Mizrachi E, Kullan ARK, et al (2014) The genome of Eucalyptus grandis. Nature 509:356–362. doi: 10.1038/nature13308
Nagamine Y, Pong-Wong R, Navarro P, Vitart V, Hayward C, Rudan I, Campbell H, Wilson J, Wild S, Hicks AA, Pramstaller PP, Hastie N, Wright AF, Haley CS (2012)

Localising Loci underlying Complex Trait Variation Using Regional Genomic Relationship Mapping. PLoS One. doi: 10.1371/journal.pone.0046501
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 Cell Physiol 48:1484–1495. doi: 10.1093/pcp/pcm119
Narum SR, Buerkle CA, Davey JW, Miller MR (2013) Genotyping-by-sequencing in ecological and conservation genomics. 2841–2847. doi: 10.1111/mec.12350
Nature (2015) Brazil approves transgenic eucalyptus. Nat Biotechnol 33:577. doi: 10.1038/nbt0615-577c
Neale D, Savolainen O (2004) Association genetics of complex traits in conifers. Trends Plant Sci 9:325–330. doi: 10.1111/j.1469-8137.2010.03593.x
Neale DB (2007) Genomics to tree breeding and forest health. 539–544. doi: 10.1016/j.gde.2007.10.002
Neale DB, Ingvarsson PK (2008) Population, quantitative and comparative genomics of adaptation in forest trees. Curr Opin Plant Biol 11:149–155. doi: 10.1016/j.pbi.2007.12.004
Neves LG, Davis JM, Barbazuk WB, Kirst M (2013) Whole-exome targeted sequencing of the uncharacterized pine genome. 146–156. doi: 10.1111/tpj.12193
Neves LG, Davis JM, Barbazuk WB, Kirst M (2014) A High-Density Gene Map of Loblolly Pine (Pinus taeda L .) Based on Exome Sequence Capture Genotyping. G3 Genes, Genomes, Genet 4:29–37. doi: 10.1534/g3.113.008714
Nicolas SD, Péros J-P, Lacombe T, Launay A, Le Paslier M-C, Bérard A, Mangin B, Valière S, Martins F, Le Cunff L, Laucou V, Bacilieri R, Dereeper A, Chatelet P, This P, Doligez A (2016) Genetic diversity, linkage disequilibrium and power of a large grapevine (Vitis vinifera L) diversity panel newly designed for association studies. BMC Plant Biol 16:74. doi: 10.1186/s12870-016-0754-z
Nielsen HM, Sonesson AK, Yazdi H, Meuwissen THE (2009) Comparison of accuracy of genome-wide and BLUP breeding value estimates in sib based aquaculture breeding schemes. Aquaculture 289:259–264. doi: 10.1016/j.aquaculture.2009.01.027
Oliveira AC, De Carneiro ACO, Vital BR, Almeida W, Pereira BLC, Cardoso MT (2010) Quality parameters of *Eucalyptus pellita* F. Muell. wood and charcoal. Sci For Sci 38:431–439.
Onogi A, Watanabe M, Mochizuki T, Hayashi T, Nakagawa H, Hasegawa T, Iwata H (2016) Toward integration of genomic selection with crop modelling: the development of an integrated approach to predicting rice heading dates. Theor Appl Genet 129:805–817. doi: 10.1007/s00122-016-2667-5
Pace J, Yu X, Lübberstedt T (2015) Genomic prediction of seedling root length in maize (*Zea mays* L.). Plant J 83:903–912. doi: 10.1111/tpj.12937
Park T, Casella G (2008) The Bayesian Lasso. J Am Stat Assoc 103:681–686. doi: 10.1198/016214508000000337
Pavy N, Gagnon F, Deschênes A, Boyle B, Beaulieu J, Bousquet J (2016) Development of highly reliable in silico SNP resource and genotyping assay from exome capture and sequencing: An example from black spruce (Picea mariana). Mol Ecol Resour 16:588–598. doi: 10.1111/1755-0998.12468
Pavy N, Gagnon F, Rigault P, Blais S, Deschênes A, Boyle B, Pelgas B, Deslauriers M,

Clément S, Lavigne P, Lamothe M, Cooke JEK, Jaramillo-Correa JP, Beaulieu J, Isabel N, Mackay J, Bousquet J (2013) Development of high-density SNP genotyping arrays for white spruce (*Picea glauca*) and transferability to subtropical and nordic congeners. Mol Ecol Resour 13:324–36. doi: 10.1111/1755-0998.12062

Pavy N, Pelgas B, Beauseigle S, Blais S, Gagnon F, Gosselin I, Lamothe M, Isabel N, Bousquet J (2008) Enhancing genetic mapping of complex genomes through the design of highly-multiplexed SNP arrays: application to the large and unsequenced genomes of white spruce and black spruce. BMC Genomics 9:21. doi: 10.1186/1471-2164-9-21

Peace C, Bassil N, Main D, Ficklin S, Rosyara UR, Stegmeir T, Sebolt A, Gilmore B, Lawley C, Mockler TC, Bryant DW, Wilhelm L, Iezzoni A (2012) Development and Evaluation of a Genome-Wide 6K SNP Array for Diploid Sweet Cherry and Tetraploid Sour Cherry. doi: 10.1371/journal.pone.0048305

Pérez P, de los Campos G (2014) Genome-wide regression and prediction with the BGLR statistical package. Genetics 198:483–495. doi: 10.1534/genetics.114.164442

Pirraglia A, Gonzalez R, Saloni D, Wright J, Denig J (2012) Fuel properties and suitability of *Eucalyptus benthamii* and *Eucalyptus macarthurii* for torrefied wood and pellets. BioResources 7:217–235.

Plomion C, Bartholomé J, Lesur I, Boury C, Rodríguez-Quilón I, Lagraulet H, Ehrenmann F, Bouffier L, Gion JM, Grivet D, de Miguel M, de María N, Cervera MT, Bagnoli F, Isik F, Vendramin GG, González-Martínez SC (2016) High-density SNP assay development for genetic analysis in maritime pine (Pinus pinaster). Mol Ecol Resour 16:574–587. doi: 10.1111/1755-0998.12464

Poland JA, Rife TW (2012) Genotyping-by-Sequencing for Plant Breeding and Genetics. Plant Genome 5:92–102. doi: 10.3835/plantgenome2012.05.0005

Poland J, Endelman J, Dawson J, Rutkoski J, Wu S, Manes Y, Dreisigacker S, Crossa J, Sánchez-villeda H, Sorrells M, Jannink J (2012) Genomic Selection in Wheat Breeding using Genotyping-by-Sequencing. Plant Genome 5:103–113. doi: 10.3835/plantgenome2012.06.0006

Porth I, Klapšte J, Skyba O, Hannemann J, Mckown AD, Guy RD, Difazio SP, Muchero W, Ranjan P, Tuskan GA, Friedmann MC, Ehlting J, Cronk QCB, El-Kassaby YA, Douglas CJ, Mansfield SD (2013) Genome-wide association mapping for wood characteristics in Populus identifies an array of candidate single nucleotide polymorphisms. New Phytol 200:710–726. doi: 10.1111/nph.12422

Porto-Neto LR, Barendse W, Henshall JM, McWilliam SM, Lehnert SA, Reverter A (2015) Genomic correlation: harnessing the benefit of combining two unrelated populations for genomic selection. Genet Sel Evol 47:84. doi: 10.1186/s12711-015-0162-0

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

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38:904–909. doi: 10.1038/ng1847

Pritchard JK, Przeworski M (2001) Linkage disequilibrium in humans: models and data. Am J Hum Genet 69:1–14. doi: 10.1086/321275

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959. doi: 10.1111/j.1471-

8286.2007.01758.x
Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MARR, Bender D, Maller J, Sklar P, Bakker PIW De, Daly MJ, Sham PC, de Bakker PIW, Daly MJ, Sham PC (2007) PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum Genet 81:559–575. doi: 10.1086/519795
Ratcliffe B, El-Dien OG, Klápště J, Porth I, Chen C, Jaquish B, El-Kassaby YA (2015) A comparison of genomic selection models across time in interior spruce (*Picea engelmannii* × *glauca*) using unordered SNP imputation methods. Heredity (Edinb) 1–9. doi: 10.1038/hdy.2015.57
Raven L-A, Cocks BG, Hayes BJ (2014) Multibreed genome wide association can improve precision of mapping causative variants underlying milk production in dairy cattle. BMC Genomics 15:62. doi: 10.1186/1471-2164-15-62
Redman AL, Mcgavin RL (2010) Accelerated Drying of Plantation Grown Eucalyptus cloeziana and Eucalyptus pellita Sawn Timber. 60:339–345.
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. Proc Natl Acad Sci 98:11479–11484. doi: 10.1073/pnas.201394398
Resende Jr MFR, Munoz P, Acosta JJ, Peter GF, Davis JM, Grattapaglia D, Jr MFRR, Mun P, Resende MD V, Kirst M (2012a) Accelerating the domestication of trees using genomic selection : accuracy of prediction models across ages and environments. New Phytol 193:617–624. doi: 10.1111/j.1469-8137.2011.03895.x
Resende Jr MFR, Munoz P, Resende MD V., Garrick DJ, Fernando RL, Davis JM, Jokela EJ, Martin TA, Peter GF, Kirst M (2012b) Accuracy of Genomic Selection Methods in a Standard Data Set of Loblolly Pine (Pinus taeda L.). Genetics 190:1503–1510. doi: 10.1534/genetics.111.137026
Resende R.T., Resende MDV, Silva FF, Azevedo CF, Takahashi EK, Silva-Junior OB, Grattapaglia D (2016) Regional heritability mapping and genome-wide association identify loci for complex growth, wood and disease resistance traits in Eucalyptus. New Phytol 213:1287–1300. doi: 10.1111/nph.14266
Resende R.T., Resende MD V, Silva FF, Azevedo CF, Takahashi EK, Silva-Junior OB, Grattapaglia D (2017) Assessing the expected response to genomic selection of individuals and families in Eucalyptus breeding with an additive-dominant model. Heredity (Edinb) 1–11. doi: 10.1038/hdy.2017.37
Resende MD V, Resende Jr MFR, Sansaloni CP, Petroli CD, Missiaggia AA, Aguiar AM, Abad JM, Takahashi EK, Rosado AM, Faria DA, Pappas Jr GJ, Kilian A, Grattapaglia D (2012) Genomic selection for growth and wood quality in Eucalyptus: capturing the missing heritability and accelerating breeding for complex traits in forest trees. New Phytol 194:116–128.
Riedelsheimer C, Endelman JB, Stange M, Sorrells ME, Jannink JL, Melchinger AE (2013) Genomic predictability of interconnected biparental maize populations. Genetics 194:493–503. doi: 10.1534/genetics.113.150227
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 (Edinb) 110:420–429. doi: 10.1038/hdy.2012.90

Robinson MR, Wray NR, Visscher PM (2014) Explaining additional genetic variation in complex traits. Trends Genet 30:124–132. doi: 10.1016/j.tig.2014.02.003

Rockman M V. (2012) The QTN program and the alleles that matter for evolution: all that's gold does not glitter. Evolution (N Y) 66:1–17. doi: 10.1111/j.1558-5646.2011.01486.x

Saatchi M, Mcclure MC, Mckay SD, Rolf MM, Kim J, Decker JE, Taxis TM, Chapple RH, Ramey HR, Northcutt SL, Bauck S, Woodward B, Dekkers JCM, Fernando RL, Schnabel RD, Garrick DJ, Taylor JF (2011) Accuracies of genomic breeding values in American Angus beef cattle using K-means clustering for cross-validation. Genet Sel Evol 43:1–16.

Sansaloni C, Petroli C, Jaccoud D, Carling J, Detering F, Grattapaglia D, Kilian A (2011) Diversity Arrays Technology (DArT) and next- generation sequencing combined : genome-wide , high throughput , highly informative genotyping for molecular breeding of Eucalyptus. BMC Proc 5:P54. doi: 10.1186/1753-6561-5-S7-P54

Schaeffer LR (2006) Strategy for applying genome wide selection in dairy cattle. J Anim Breed Genet 123:218–223.

Schmidt M, Kollers S, Maasberg-Prelle A, Großer J, Schinkel B, Tomerius A, Graner A, Korzun V (2015) Prediction of malting quality traits in barley based on genome-wide marker data to assess the potential of genomic selection. Theor Appl Genet 129:203–213. doi: 10.1007/s00122-015-2639-1

Schork AJ, Thompson WK, Pham P, Torkamani A, Roddey JC, Sullivan PF, Kelsoe JR, O'Donovan MC, Furberg H, Schork NJ, Andreassen OA, Dale AM (2013) All SNPs Are Not Created Equal: Genome-Wide Association Studies Reveal a Consistent Pattern of Enrichment among Functionally Annotated SNPs. PLoS Genet 9:1–13. doi: 10.1371/journal.pgen.1003449

Shani Z, Dekel M, Roiz L, Horowitz M, Kolosovski N, Lapidot S, Alkan S, Koltai H, Tsabary G, Goren R, Shoseyov O (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 Rep 25:1067–1074. doi: 10.1007/s00299-006-0167-9

Shengqiang Z, Dekkers JCM, Fernando RL, Jannink JL (2009) Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: A barley case study. Genetics 182:355–364. doi: 10.1534/genetics.108.098277

Silva-Junior OB, Grattapaglia D (2015) Genome-wide patterns of recombination , linkage disequilibrium and nucleotide diversity from pooled resequencing and single nucleotide polymorphism genotyping unlock the evolutionary history of Eucalyptus grandis. New Phytol 208:830–845. doi: 10.1111/nph.13505

Silva-Junior OBB, Faria DAA, 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. doi: 10.1111/nph.13322

Slatkin M (2008) Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. Nat Rev Genet 9:477–485. doi: 10.1038/nrg2361

Solberg TR, Sonesson AK, Woolliams JA, Meuwissen THE (2008) Genomic selection using different marker types and densities. J Anim Sci 86:2447–2454. doi: 10.2527/jas.2007-0010

Speed D, Balding DJ (2014) Relatedness in the post-genomic era: is it still useful? Nat Rev Genet 16:33–44. doi: 10.1038/nrg3821

Spiliopoulou A, Nagy R, Bermingham ML, Huffman JE, Hayward C, Vitart V, Rudan I, Campbell H, Wright AF, Wilson JF, Pong-Wong R, Agakov F, Navarro P, Haley CS (2015) Genomic prediction of complex human traits: Relatedness, trait architecture and predictive meta-models. Hum Mol Genet 24:4167–4182. doi: 10.1093/hmg/ddv145

Spindel J, Begum H, Akdemir D, Virk P, Collard B (2015) Genomic selection and association mapping in rice (*Oryza sativa*): effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding line. PLoS Genet 1–25. doi: 10.1371/journal.pgen.1004982

Spindel JE, Begum H, Akdemir D, Collard B, Redoña E, Jannink J-L, McCouch S (2016) Genome-wide prediction models that incorporate de novo GWAS are a powerful new tool for tropical rice improvement. Heredity (Edinb) 116:395–408. doi: 10.1038/hdy.2015.113

Stanturf JA, Vance ED, Fox TR, Kirst M (2013) Eucalyptus beyond Its Native Range : Environmental Issues in Exotic Bioenergy Plantations.

Strauss S, Brunner A, Busov V, Ma C, Meilan R (2004) Ten lessons from 15 years of transgenic *Populus* research.

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. Proc Natl Acad Sci 108:6300–6305. doi: 10.1073/pnas.1009252108

Su G, Christensen OF, Janss L, Lund MS (2014) Comparison of genomic predictions using genomic relationship matrices built with different weighting factors to account for locus-specific variances. J Dairy Sci 1–13. doi: 10.3168/jds.2014-8210

Sulichantini E, Sutisna M, Sukartiningsih, Rusdiansyah (2014) Clonal propagation of two clones *Eucalyptus pellita* F. Muell by mini-cutting. Int J Sci Eng 6:117–121. doi: 10.12777/ijse.6.2.117-121

Sun B, Wang X, Liu J (2013) Changes in dimensional stability and mechanical properties of *Eucalyptus pellita* by melamine-urea-formaldehyde resin impregnation and heat treatment. Eur J Wood Wood Prod 71:557–562. doi: 10.1007/s00107-013-0700-9

Suren H, Hodgins KA, Yeaman S, Nurkowski KA, Smets P, Rieseberg LH, Aitken SN, Holliday JA (2016) Exome capture from the spruce and pine giga-genomes. Mol Ecol Resour 16:1136–1146. doi: 10.1111/1755-0998.12570

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

Tayeh N, Klein A, Le Paslier M-C, Jacquin F, Houtin H, Rond C, Chabert-Martinello M, Magnin-Robert J-B, Marget P, Aubert G, Burstin J (2015) Genomic Prediction in Pea: Effect of Marker Density and Training Population Size and Composition on Prediction Accuracy. Front Plant Sci 6:1–11. doi: 10.3389/fpls.2015.00941

Thavamanikumar S, Dolferus R, Thumma BR (2015) Comparison of Genomic Selection Models to Predict Flowering Time and Spike Grain Number in Two Hexaploid Wheat Doubled Haploid Populations. G3: Genes|Genomes|Genetics 5:1991–1998. doi: 10.1534/g3.115.019745

Thavamanikumar S, McManus LJ, Ades PK, Bossinger G, Stackpole DJ, Kerr R, Hadjigol

S, Freeman JS, Vaillancourt RE, Zhu P, Tibbits JFG (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 Genet Genomes 10:1661–1678. doi: 10.1007/s11295-014-0787-0

Thomasen JR, Egger-Danner C, Willam A, Guldbrandtsen B, Lund MS, Sørensen AC (2014) Genomic selection strategies in a small dairy cattle population evaluated for genetic gain and profit. J Dairy Sci 97:458–470. doi: 10.3168/jds.2013-6599

Thomson MJ (2014) High-Throughput SNP Genotyping to Accelerate Crop Improvement. 2014:195–212.

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. doi: 10.1534/genetics.105.042028

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 Res 45:W122–W129. doi: 10.1093/nar/gkx382

Turner SD (2014) qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. bioRxiv 1–2. doi: 10.1101/005165

Tusell L, Pérez-Rodríguez P, Forni S, Gianola D (2014) Model averaging for genome-enabled prediction with reproducing kernel Hilbert spaces: a case study with pig litter size and wheat yield. J Anim Breed Genet 131:105–115. doi: 10.1111/jbg.12070

van Binsbergen R, Calus MPL, Bink MCAM, van Eeuwijk FA, Schrooten C, Veerkamp RF (2015) Genomic prediction using imputed whole-genome sequence data in Holstein Friesian cattle. Genet Sel Evol 47:71. doi: 10.1186/s12711-015-0149-x

Van Eenennaam AL, Weigel KA, Young AE, Cleveland MA, Dekkers JCM (2014) Applied Animal Genomics: Results from the Field. Annu Rev Anim Biosci 2:105–139. doi: 10.1146/annurev-animal-022513-114119

Van Sandt VST, Suslov D, Verbelen JP, Vissenberg K (2007) Xyloglucan endotransglucosylase activity loosens a plant cell wall. Ann Bot 100:1467–1473. doi: 10.1093/aob/mcm248

VanRaden PM (2008) Efficient Methods to Compute Genomic Predictions. J Dairy Sci 91:4414–4423. doi: 10.3168/jds.2007-0980

VanRaden PM, Tooker ME, O'Connell JR, Cole JB, Bickhart DM (2017) Selecting sequence variants to improve genomic predictions for dairy cattle. Genet Sel Evol 49:32. doi: 10.1186/s12711-017-0307-4

Vazquez AI, Rosa GJM, Weigel KA, de los Campos G, Gianola D, Allison DB (2010) Predictive ability of subsets of single nucleotide polymorphisms with and without parent average in US Holsteins. J Dairy Sci 93:5942–5949. doi: 10.3168/jds.2010-3335

Verde I, Bassil N, Scalabrin S, Gilmore B, Lawley CT, Gasic K, Micheletti D, Rosyara UR, Cattonaro F, Vendramin E, Main D, Aramini V, Blas AL, Mockler TC, Bryant DW, Peace C, Morgante M (2012) Development and Evaluation of a 9K SNP Array for Peach by Internationally Coordinated SNP Detection and Validation in Breeding Germplasm. doi: 10.1371/journal.pone.0035668

Vinkhuyzen AA (2013) Estimation and Partitioning of Heritability in Human Populations using Whole Genome Analysis Methods. Annu Rev Genet 47:75–95. doi: 10.1146/annurev-genet-111212-133258.Estimation

Visscher PM, Brown MA, McCarthy MI, Yang J (2012) Five years of GWAS discovery. Am J Hum Genet 90:7–24. doi: 10.1016/j.ajhg.2011.11.029
Visscher PM, Hemani G, Vinkhuyzen AAE, Chen GB, Lee SH, Wray NR, Goddard ME, Yang J (2014) Statistical Power to Detect Genetic (Co)Variance of Complex Traits Using SNP Data in Unrelated Samples. PLoS Genet. doi: 10.1371/journal.pgen.1004269
Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J (2017) 10 Years of GWAS Discovery: Biology, Function, and Translation. Am J Hum Genet 101:5–22. doi: 10.1016/j.ajhg.2017.06.005
Voragen AGJ, Coenen GJ, Verhoef RP, Schols HA (2009) Pectin, a versatile polysaccharide present in plant cell walls. Struct Chem 20:263–275. doi: 10.1007/s11224-009-9442-z
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 Sci 56:2365–2378. doi: 10.2135/cropsci2015.10.0632
Wallén SE, Lillehammer M, Meuwissen THE (2017) Strategies for implementing genomic selection for feed efficiency in dairy cattle breeding schemes. J Dairy Sci 100:6327–6336. doi: 10.3168/jds.2016-11458
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 Phytol 188:515–532. doi: 10.1111/j.1469-8137.2010.03415.x
Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. Evolution (N Y) 38:1358–1370.
Weller JI, Glick G, Shirak A, Ezra E, Seroussi E, Shemesh M, Zeron Y, Ron M (2014) Predictive ability of selected subsets of single nucleotide polymorphisms (SNPs) in a moderately sized dairy cattle population. Animal 8:208–16. doi: 10.1017/S1751731113002188
Wickham H (2009) ggplot2: Elegant graphics for data analysis, 1st edn. Springer- Verlag, New York
Wimmer V, Albrecht T, Auinger H, Schön C, Breeding P, München TU (2012) synbreed : a framework for the analysis of genomic prediction data using R. 28:2086–2087. doi: 10.1093/bioinformatics/bts335
Windhausen V, Atlin G, Hickey J, Crossa J, Jannink J, Sorrells M, Raman B, Cairns J, Tarekegne A, Semagn K, Beyene Y, Grudloyma P, Technow F, Riedelsheimer C, Melchinger A (2012) Effectiveness of genomic prediction of maize hybrid performance in different breeding populations and environments effectiveness of genomic prediction of maize hybrid performance in different breeding populations and environments. G3 Genes, Genomes, Genet 2:1427–1436. doi: 10.1534/g3.112.003699
Wood A, Esko T, Yang J, Vedantam S, Pers T, Gustafsson S, Chu A, Estrada K, Luan J, Kutalik Z, Amin N, Buchkovich M, Croteau-Chonka D, Day F, Duan, Y, et al. (2014) Defining the role of common variation in the genomic and biological architecture of adult human height. Nat Genet 46:1173–1186. doi: 10.1038/ng.3097
Wray NR (2005) Allele frequencies and the r2 measure of linkage disequilibrium: impact

on design and interpretation of association studies. Twin Res Hum Genet 8:87–94. doi: 10.1375/twin.8.2.87

Wray NR, Yang J, Hayes BJ, Price AL, Goddard ME, Visscher PM (2013) Pitfalls of predicting complex traits from SNPs. Nat Rev Genet 14:507–515. doi: 10.1038/nrg3457-c2

Wright S (1931) Evolution in Mendelian populations. Genetics 16:97–159.

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. Am J Hum Genet 89:82–93. doi: 10.1016/j.ajhg.2011.05.029

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 Biotechnol J 14:1551–1562. doi: 10.1111/pbi.12519

Xia J, Qi X, Wu Y, Zhu B, Xu L, Zhang L, Gao X, Chen Y, Li J, Gao H (2016) Genome-wide association study identifies loci and candidate genes for meat quality traits in Simmental beef cattle. Mamm Genome 27:246–255. doi: 10.1007/s00335-016-9635-x

Xiao Y, Liu H, Wu L, Warburton M, Yan J (2017) Genome-wide Association Studies in Maize: Praise and Stargaze. Mol Plant 10:359–374. doi: 10.1016/j.molp.2016.12.008

Yabe S, Yamasaki M, Ebana K, Hayashi T, Iwata H (2016) Island-model genomic selection for long-term genetic improvement of autogamous crops. PLoS One 11:1–21. doi: 10.1371/journal.pone.0153945

Yang J, Bakshi A, Zhu Z, Hemani G, Vinkhuyzen AAE, Lee SH, Robinson MR, Perry JRB, Nolte IM, van Vliet-Ostaptchouk J V, Snieder H, Esko T, Milani L, Mägi R, Metspalu A, Hamsten A, Magnusson PKE, Pedersen NL, Ingelsson E, Soranzo N, Keller MC, Wray NR, Goddard ME, Visscher PM (2015) Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. Nat Genet 47:1114–1120. doi: 10.1038/ng.3390

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, Goddard ME, Visscher PM (2010) Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42:565–569. doi: 10.1038/ng.608

Yang J, Ferreira T, Morris AP, Medland SE, Madden PAF, Heath AC, Martin NG, Montgomery GW, Weedon MN, Loos RJ, Frayling TM, McCarthy MI, Hirschhorn JN, Goddard ME, Visscher PM (2012) Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat Genet 44:369–375. doi: 10.1038/ng.2213

Yang J, Lee SH, Goddard ME, Visscher PM (2011a) GCTA: A Tool for Genome-wide Complex Trait Analysis. Am J Hum Genet 88:76–82. doi: 10.1016/j.ajhg.2010.11.011

Yang J, Manolio TA, Pasquale LR, Boerwinkle E, Caporaso N, Cunningham JM, de Andrade M, Feenstra B, Feingold E, Hayes MG, Hill WG, Landi MT, Alonso A, Lettre G, Lin P, Ling H, Lowe W, Mathias RA, Melbye M, Pugh E, Cornelis MC, Weir BS, Goddard ME, Visscher PM (2011b) Genome partitioning of genetic variation for complex traits using common SNPs. Nat Genet 43:519–525. doi: 10.1038/ng.823

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 J 42:862–876. doi: 10.1111/j.1365-313X.2005.02422.x

Yu A, Gallagher T (2015) Analysis on the Growth Rhythm and Cold Tolerance of Five-Year Old Eucalyptus benthamii Plantation for Bioenergy. 585–592.

Yu H, Xie W, Wang J, Xing Y, Xu C, Li X, Xiao J, Zhang Q (2011) Gains in QTL detection using an ultra-high density SNP map based on population sequencing relative to traditional RFLP/SSR markers. PLoS One. doi: 10.1371/journal.pone.0017595

Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38:203–208. doi: 10.1038/ng1702

Yuskianti V, Glen M, Puspitasari D, Francis A, Rimbawanto A, Gafur A, Indrayadi H, Mohammed CL (2014) Species-specific PCR for rapid identification of *Ganoderma philippii* and *Ganoderma mastoporum* from *Acacia mangium* and *Eucalyptus pellita* plantations in Indonesia. For Pathol 44:477–485. doi: 10.1111/efp.12144

Zapata-Valenzuela J, Isik F, Maltecca C, Wegrzyn J, Neale D, McKeand S, Whetten R (2012) SNP markers trace familial linkages in a cloned population of *Pinus taeda* - prospects for genomic selection. Tree Genet Genomes 8:1307–1318. doi: 10.1007/s11295-012-0516-5

Zapata-Valenzuela J, Whetten RW, Neale D, Mckeand S, Isik F (2013) Genomic Estimated Breeding Values Using Genomic Relationship Matrices in a Cloned Population of Loblolly Pine. G3 Genes, Genomes, Genet 3:909–916. doi: 10.1534/g3.113.005975

Zarpelon TG, da Silva Guimarães LM, Faria DA, Coutinho MM, Cápua Neto B, Teixeira RU, Grattapaglia D, Alfenas AC (2014) Genetic mapping and validation of QTLs associated with resistance to *Calonectria* leaf blight caused by *Calonectria pteridis* in *Eucalyptus*. Tree Genet Genomes. doi: 10.1007/s11295-014-0803-4

Zauza EA V, Alfenas AC, Old K, Couto MMF, Graça RN, Maffia LA (2010) Myrtaceae species resistance to rust caused by *Puccinia psidii*. Australas Plant Pathol 39:406–411. doi: 10.1071/AP10077

Zhang J, Song Q, Cregan PB, Jiang G-L (2015) Genome-wide association study, genomic prediction and marker-assisted selection for seed weight in soybean (*Glycine max*). Theor Appl Genet. doi: 10.1007/s00122-015-2614-x

Zhao Y, Gowda M, Liu W, Würschum T, Maurer HP, Longin FH, Ranc N, Reif JC (2012) Accuracy of genomic selection in European maize elite breeding populations. Theor Appl Genet 124:769–776. doi: 10.1007/s00122-011-1745-y

Zhdanova O, Pudovkin A (2008) Nb_HetEx: a program to estimate the effective number of breeders. J Hered 99:694–695. doi: 10.1093/jhered/esn061

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. 28:3326–3328. doi: 10.1093/bioinformatics/bts606

Zhou L, Holliday JA (2012) Targeted enrichment of the black cottonwood (Populus trichocarpa) gene space using sequence capture. BMC Genomics 13:1. doi: 10.1186/1471-2164-13-703

Zhu C, Gore M, Buckler ES, Yu J (2008) Status and Prospects of Association Mapping in Plants. Plant Genome J 1:5. doi: 10.3835/plantgenome2008.02.0089

Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, Montgomery GW, Goddard ME, Wray NR, Visscher PM, Yang J (2016) Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nat Genet 48:481–487. doi: 10.1038/ng.3538