

LUCAS FERNANDES ROCHA

**GENOME-WIDE ASSOCIATION AND OPTIMIZATION OF PREDICTION
ACCURACIES ON GENOMIC SELECTION MODELS FOR *Eucalyptus grandis* W.
Hill**

Botucatu

2022

LUCAS FERNANDES ROCHA

**GENOME-WIDE ASSOCIATION AND OPTIMIZATION OF PREDICTION
ACCURACIES ON GENOMIC SELECTION MODELS FOR *Eucalyptus grandis* W.
Hill**

**Thesis presented to the School of
Agriculture - Botucatu campus, São
Paulo State University, to obtain the title
of Ph.D. in Forest Science.**

**Advisor: Evandro Vagner Tambarussi
Co-advisors: Juan Jose Acosta
Jamarillo; Roberto Fritsche-Neto;
Thiago Romanos Benatti**

Botucatu

2022

R672g Rocha, Lucas Fernandes
Genome-wide association and optimization of prediction accuracies on genomic selection models for *Eucalyptus grandis*
W. Hill / Lucas Fernandes Rocha. -- Botucatu, 2022
125 p. : il., tabs.

Tese (doutorado) - Universidade Estadual Paulista (Unesp),
Faculdade de Ciências Agrônômicas, Botucatu
Orientador: Evandro Vagner Tambarussi

1. Melhoramento florestal. 2. Genética vegetal. 3.
Associação genômica ampla. 4. Seleção genômica. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca da Faculdade de Ciências Agrônômicas, Botucatu. Dados fornecidos pelo autor(a).

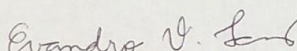
Essa ficha não pode ser modificada.

CERTIFICADO DE APROVAÇÃO

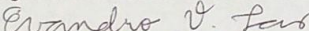
TÍTULO DA TESE: GENOME-WIDE ASSOCIATION AND OPTIMIZATION OF PREDICTION ACCURACIES
ON GENOMIC SELECTION MODELS FOR *Eucalyptus grandis* W.Hill

AUTOR: LUCAS FERNANDES ROCHA
ORIENTADOR: EVANDRO VAGNER TAMBARUSSI
COORIENTADOR: JUAN JOSE ACOSTA JARAMILLO
COORIENTADOR: ROBERTO FRITSCH NETO
COORIENTADOR: THIAGO ROMANOS BENATTI

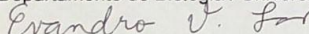
Aprovado como parte das exigências para obtenção do Título de Doutor em CIÊNCIA FLORESTAL,
pela Comissão Examinadora:



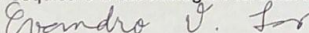
Prof. Dr. EVANDRO VAGNER TAMBARUSSI (Participação Virtual)
Engenharia Florestal / Universidade Estadual do CentroOeste



Prof. Dr. EVANDRO NOVAES (Participação Virtual)
Departamento de Biologia / Universidade Federal de Lavras - Instituto de Ciências Naturais



Pesquisadora Dr.^a IZABEL CHRISTINA GAVA DE SOUZA (Participação Virtual)
Pesquisa e Biotecnologia Florestal / Suzano Papel e Celulose



Prof. Dr. FREDDY MORA POBLETE (Participação Virtual)
Instituto de Ciencias Biológicas / Universidad de Talca



Prof. Dr. HUMBERTO FANELLI CARVALHO (Participação Virtual)
Centro de Biotecnología y Genómica de Plantas / Universidad Politecnica de Madrid

Botucatu, 29 de março de 2022

*I dedicate this thesis to my mom and my dad,
who always taught me to fight for my education,
and for their endless love, encouragement, and support.*

ACKNOWLEDGMENTS

To God, the primordial energy that command the universe, for giving me the strength to always be a better human being.

To Professor Evandro V. Tambarussi, for his guidance, support, and suggestions to accomplish this study. I will always be thankful for having such a good advisor like you.

To the SUZANO S.A. for sharing the data used in this study. The success of this study was only possible because of so many competent people behind this company.

To Professor Juan Jose Acosta (NCSU), for giving me fundamental guidance during my doctorate.

To Dr. Roberto Fritsche-Neto (IRRI), for having welcomed me in his lab, and giving all necessary support for elaboration of this work.

To Dr. Thiago R. Benatti (SUZANO), for believing in my work, in addition to giving brilliant ideas to develop this study.

To Leandro de Siqueira and Izabel Christina Souza (SUZANO) for supporting me with indispensable information about the genotypic and phenotypic data.

To the Allogamous Plant Breeding Lab (ESALQ-USP) for all support. I am profusely thankful to Rafael, Humberto, Karina, Gabriela, Melina, and Alisson for helping me with data analysis and also for the coffee times.

To the Forest Breeding Lab (UNICENTRO) for the indispensable discussions for my training to be a tree breeder. In special, I would like to thank Dandara Yasmim and João Paludeto for always being available to help me and all the discussions about forest breeding.

To the Thünen Institute of Forest Genetics (Germany) and the genome research group, in special Dr. Niels Müller for the guidance during my part-time scholarship and for not hesitating to help me during a world pandemic.

To my mother and father for all love and fundamental education given during my childhood. You both are the main reason of all good things that I accomplished in my life.

To my sisters, Danielly, Brunna, Maria Eduarda, Maria Vitória, and my nephews Wágner Júnior and Cecília, for being the cornerstone of my life.

To Paul, for being by my side during hard times.

To my dear friend Professor Dulcinéia de Carvalho (UFLA), the person responsible to make me deeply admire the genetics field.

To my friend Allan de Amorim, for always helping me when I need it most.

To the friends I have made in Botucatu-SP (Marina Sbardella and Fernando Cessel), Piracicaba-SP (Francisco, Isa, Taiane, and Yara) and Três Lagoas-MS (Sofia, Amanda, and Vágna), for the friendship during my doctorate.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) - Finance Code 001.

To the Deutscher Akademischer Austauschdienst (DAAD) for the short-time doctoral scholarship.

Immeasurable appreciation and deepest gratitude for all those that supported me and contributed to make this study possible.

“Thus, from the war of nature, from famine and death, the most exalted object which we are capable of conceiving, namely, the production of the higher animals, directly follows. There is a grandeur in this view of life, with its several powers, having been originally breathed into a few forms or into one; and that whilst this planet has gone cycling on according to the fixed law of gravity, from so simple a beginning endless forms most beautiful and most wonderful have been, and are being, evolved.”

DARWIN, C. **On the Origin of Species**. 6. ed. introduction by W. R. Thompson. New York: Everyman's Library, 1872. p. 488.

ABSTRACT

Molecular markers that are widely distributed throughout the genome offer a fundamental tool to optimize forest tree breeding programs. This study aimed to evaluate the genetic architecture of quantitative genes and optimize genomic selection models for growth and wood-quality traits of *Eucalyptus grandis*. We evaluated an open-pollinated breeding population with 1,772 genotypes, composed of 27 different families, that was established using complete randomized block design with 20 replicates each. Individuals were genotyped using the Illumina Infinium EuCHIP60K chip and 12 different phenotypic variables were evaluated for growth traits (diameter at breast height, height, and volume evaluated at 3 and 6 years after planting) and wood-quality traits (pure cellulose yield, basic wood density, syringyl/guayacil, soluble lignin, total solids, and total extractives). First, we performed a genome-wide association study (GWAS) using the single-trait model (farmCPU) and multi-trait (MTMM) mixed models. Next, we searched for quantitative trait loci (QTLs) and their predicted functional effects using a database for *Eucalyptus*. Subsequently, the accuracy of the prediction ability, coincidence of selection, and selection gains of genomic selection models were analyzed based on the Genomic Best Linear Unbiased Prediction (GBLUP) method. We tested different approaches considering the additive variance, additive-dominant variance, optimization of training set, and multi-trait models. Finally, we analyzed the efficiency of using growth traits to increase the prediction ability of wood-quality traits considering a multi-trait model with optimization of training set methodology. After quality control, a total of 21,254 informative SNPs were found that have a wide distribution and a high linkage disequilibrium decay across the 11 chromosomes. For the GWAS analysis, the farmCPU model identified 43 and 38 small effect markers that are significantly associated with growth and wood quality traits, respectively. Similarly, pleiotropic SNPs were also discovered between growth (24) and wood quality traits (6) using the MTMM model. Through gene ontology analysis, we identified genes responsible for plant growth and related with hydric stress. For the genomic selection analysis, growth traits appeared to be more influenced by dominance than wood quality traits, meanwhile GBLUP models were effective in predicting wood quality traits. Although the results for CS appear to be low, SG values were relatively high. The optimization of the training set analysis effectively selected the best genotypes to be used as the training set. Additionally, the multi-trait and multi-trait with optimization of the training set were also able to increase the prediction ability of the GBLUP models. Thus, information from growth traits can be used to effectively increase the prediction ability of wood quality traits. Our study demonstrates the complex nature of quantitative traits, provides new evidence for the architecture of genes related to trait expression, and highlights the efficiency of genomic selection models to predict phenotypic expression in *E. grandis*.

Keywords: genome-wide association; genomic selection; multi-trait analysis; forest breeding; eucalypt.

RESUMO

O uso de marcadores moleculares amplamente distribuídos ao longo do genoma são uma ferramenta fundamental para otimizar programas de melhoramento florestal. Esta pesquisa teve como objetivos avaliar a arquitetura genética de caracteres quantitativos e otimizar modelos de seleção genômica de *Eucalyptus grandis* para variáveis de crescimento e qualidade da madeira. Dessa forma, foi avaliada uma população de polinização aberta de *E. grandis* composta por 1.772 genótipos e provenientes de 27 famílias estabelecidas usando um delineamento de blocos ao acaso com 20 plantas/parcela. Os indivíduos foram genotipados usando o chip Illumina Infinium EuCHIP60K e 12 caracteres fenotípicos foram avaliados e classificados em caracteres de crescimento (diâmetro à altura do peito, altura e volume aos três e seis anos após o plantio) e qualidade da madeira (produção de celulose pura, densidade básica da madeira, relação siringil/guayacil, lignina solúvel, sólidos totais e extrativos totais). Primeiramente, foi realizada uma análise de associação genômica ampla (GWAS) usando modelos mistos de single-trait (farmCPU) e multi-trait (MTMM). Em seguida, foram identificados loci de caracteres quantitativos (QTLs) utilizando o banco de anotações para *Eucalyptus*. Posteriormente, a habilidade de predição, a coincidência de seleção e os ganhos de seleção de modelos de seleção genômica foram analisados utilizando a metodologia *Genomic Best Linear Unbiased Prediction* (GBLUP). Foram testadas diferentes abordagens considerando apenas a variância aditiva, as variâncias aditivas-dominantes, modelos de otimização da população de treinamento e modelos *multi-trait*. Finalmente, foi avaliada a efetividade da utilização de caracteres fenotípicos de crescimento para aumentar a habilidade de predição de caracteres de qualidade da madeira usando uma metodologia conjunta entre *multi-trait* e otimização da população de treinamento. Após o controle de qualidade, um total de 21.254 SNPs informativos foram encontrados com ampla distribuição e alto decaimento de desequilíbrio de ligação nos 11 cromossomos. Considerando a análise GWAS, o modelo farmCPU identificou 43 e 38 marcadores de pequeno efeito significativamente associados às variáveis nas classes crescimento e de qualidade da madeira, respectivamente. Semelhantemente, marcadores pleiotrópicos também foram identificados entre caracteres crescimento (24) e de qualidade da madeira (6) usando o modelo MTMM. A análise da ontologia genética identificou diversos genes responsáveis pelo crescimento celular e associados ao stress hídrico. Considerando a análise de seleção genômica, os caracteres de crescimento foram mais influenciados pela dominância. Por outro lado, os modelos GBLUP foram eficientes para prever caracteres de qualidade da madeira. Embora a coincidência de seleção pareça ter valores baixos, os valores de ganhos de seleção encontrados foram relativamente altos. A análise de otimização da população de treinamento foi eficiente para selecionar os melhores genótipos a serem utilizados como conjunto de treinamento. Adicionalmente, as análises *multi-trait* e *multi-trait* com otimização da população de treinamento também foram eficientes para aumentar a habilidade de predição dos modelos GBLUP. Dessa forma, o uso de informações do crescimento pode ser usado de forma eficiente para aumentar a habilidade de predição dos caracteres de qualidade da madeira. Nosso estudo demonstra que a natureza dos caracteres quantitativos fornece novas evidências para a arquitetura de genes relacionados à expressão de caracteres, bem como a eficiência de modelos de seleção genômica para prever a expressão fenotípica em *E. grandis*.

Palavras-chave: associação genômica ampla; seleção genômica; análise *multi-trait*; melhoramento florestal; eucalipto.

SUMMARY

GENERAL INTRODUCTION	17
CHAPTER 1 - QUANTITATIVE TRAIT LOCI RELATED TO GROWTH AND WOOD QUALITY TRAITS IN <i>Eucalyptus grandis</i> W. Hill IDENTIFIED THROUGH SINGLE AND MULTI-TRAIT GENOME-WIDE ASSOCIATION STUDIES	21
ABSTRACT	21
1.1 INTRODUCTION	22
1.2 MATERIAL AND METHODS	24
1.2.1 Plant material and phenotypes	24
1.2.2 DNA extraction and quality control	26
1.2.3. SNP repositioning.....	26
1.2.4 Genetic parameters and population structure.....	27
1.2.5 Linkage disequilibrium (LD) decay	28
1.2.6 Genome-wide association study.....	28
1.2.7 Gene ontology	29
1.3 RESULTS.....	30
1.3.1 Phenotypic data.....	30
1.3.2 Population structure and genetic diversity parameters	31
1.3.3 SNP repositioning and quality control.....	32
1.3.4 Genome-wide association study.....	34
1.3.4.1 Broad- and narrow-sense heritability and gene annotation	34
1.3.4.2 Multi-trait genome-wide association.....	38
1.4 DISCUSSION	40
1.5 CONCLUSION	44
1.6 ACKNOWLEDGMENTS	45
REFERENCES	46
APPENDIX A - SUPPLEMENTARY MATERIAL FOR CHAPTER 1	52
CHAPTER 2 - IMPROVEMENT OF GENOMIC PREDICTION MODELS AND TRAINING SETS FOR <i>Eucalyptus grandis</i> W. Hill BREEDING	81
2.1 INTRODUCTION.....	82
2.2 MATERIAL AND METHODS	84
2.2.1 Phenotypic data.....	84

2.2.2	Quality control and effective population size	85
2.2.3	Phenotypic analysis	86
2.2.4	Genomic prediction	87
2.2.4.1	Additive and additive-dominant models	87
2.2.4.2	Predictive accuracies and genetic parameters	87
2.2.4.3	Optimization of the training set with a genetic algorithm	89
2.2.4.4	Multi-trait analysis.....	89
2.2.4.5	Multi-trait with optimization of the training set model.....	91
2.3	RESULTS	91
2.3.1	Genotypic and phenotypic data.....	91
2.3.2	Heritabilities	94
2.3.3	Training set size.....	95
2.3.4	Predictive ability	96
2.3.4.1	Additive and additive-dominant model.....	96
2.3.4.2	Optimization of the training set.....	97
2.3.4.3	Multi-trait genomic selection.....	98
2.3.4.4	Selection coincidence and predicted selection gains.....	100
2.3.5.	Workflow analysis	103
2.4	DISCUSSION	104
2.5	CONCLUSION	109
2.6	ACKNOWLEDGMENTS	110
	REFERENCES	111
	APPENDIX A - SUPPLEMENTARY MATERIAL FOR CHAPTER 2	118
	FINAL CONSIDERATIONS	119
	REFERENCES	121

GENERAL INTRODUCTION

Worldwide, Brazil is the top producer and exporter of eucalypt round wood and cellulose pulp (GONCALVES et al., 2013) and the second largest producer of cellulose pulp in the world (IBÁ, 2019). The genus *Eucalyptus* L'Heritier consists of approximately 900 species and subspecies, and *Eucalyptus grandis* W.Hill is the most commonly cultivated tree species (CAMPOE et al., 2013). The main uses of the species are related to the production of cellulose and wood (ACOSTA; MASTRANDREA; LIMA, 2008). In addition, the species presents great importance due to intense hybrid vigor generated by crossings with species of the same subgenus, with an outstanding relevance regarding the hybrids between *E. grandis* and *Eucalyptus urophylla* ST Blake, which generates intense vigor, allowing a high productivity as well as the possibility of planting in remote areas (POTTS; DUNGEY, 2004). However, its high adaptability in subtropical regions, the species genome has generated introgression into the genome in plantations in regions with arid climate (MOSTERT-O'NEILL et al., 2021).

Currently, there is an urgent need to implement genomic approaches to optimize time, cost, and accuracy of *Eucalyptus* breeding strategies (GRATTAPAGLIA et al., 2018). Several methods have been proposed to reduce the time needed for breeding based on early selection strategies (BURDON, 1989; WU et al., 1998). However, marker-assisted selection methods have emerged as important tools that can improve both the accuracy and reduce the length of the breeding cycle (O'MALLEY; MCKEAND, 1994; WU et al., 1998; GUIMARÃES, 2007; MURANTY et al., 2014; AHMAR et al. 2021). Additionally, because of the complex, polygenic nature of quantitative traits, the development of high throughput molecular technology has become key not only to breeding program improvement (DENIS; BOUVET, 2013; THAVAMANIKUMAR et al., 2013), but also to discovering the genetic architecture behind the expression of important traits in tree species (MACKAY, 2001).

Long breeding cycles together with late flowering are major problems in forest tree breeding (NAMKOONG; BARNES; BURLEY, 1980; GRATTAPAGLIA, 2017). Thus, the development of new methods that can shorten selection cycles is necessary to increase tree breeding efficiency and reduce the time required to obtain superior genotypes (CROSSA et al., 2017). Several methods have been developed to understand the simultaneous genetic control of quantitative traits (VARSHNEY;

ROORKIWAL; SORRELLS, 2017), while studies aimed at improving the selection cycle of forest species have also been undertaken (GRATTAPAGLIA, 2014). Thus, considering breeding strategies as well as the key species analyzed in the study, new methodologies must be developed that address problems and optimize breeding programs.

Eucalyptus grandis is a diploid species that has 11 chromosomes ($2n = 22$). The *Eucalyptus* genome consists of about 640 megabases, with homozygous regions covering 24% of the genome (MYBURG et al., 2014). The first studies involving the sequencing of single polymorphism nucleotides (SNPs) in *E. grandis* were performed using expressed genes (ETS) by Novaes et al. (2008). Recently, a large-scale SNP-array has also been developed for *Eucalyptus* species, with 64K SNPs (EUChip60K) (SILVA-JUNIOR; FARIA; GRATTAPAGLIA, 2015). Therefore, several methodologies have been developed to optimize eucalypt breeding, such as genomic selection (GS) (CAPPA et al., 2018; ARCURI et al., 2020; MPHABLELE et al., 2020) and genome-wide association studies (GWAS) (MÜLLER et al., 2017; KAINER et al., 2019).

The identification of genomic associations for economically relevant traits remains a challenge for forest species (GRATTAPAGLIA et al., 2018). Single nucleotide polymorphisms (SNPs) are molecular markers that occur from a single base variation in the genome (MAMMADOV et al., 2012) which can be used to help identify important regions in the genome of the species under study. The accuracy of the applicability of SNP markers is generally related to their abundance in the genome, in addition to their high reproducibility in species (AGARWAL; SHRIVASTAVA; PADH, 2008). Thus, from the development of SNPs markers, it is possible to detect quantitative trait loci (QTLs) associated with a gene by linkage disequilibrium (LD) (RAFALSKI, 2002), thus enabling the development of efficient methods of genome association (WANG et al., 2005).

GWAS methodologies analyze the associations between quantitative traits and genetic variants that are widely distributed in the genome of individuals (BUSH; MOORE, 2012). Studies using GWAS seek to identify not only associations between regions of the genome with traits of interest, but also regions with a significant effect on phenotypes (HIRSCHHORN; DALY, 2005). From this, it is possible to study the biological functions involved with QTLs and, consequently, understand the genetic influence on the phenotypic expression of a genotype (WANG et al., 2005). Although GWAS were developed to assess the genetic interaction of diseases in humans, the

methodology has been applied in several other areas, such as animal breeding (DUIJVESTIEN et al., 2010; BOLORMAA et al., 2011; WU et al., 2013), improvement of agricultural species (WANG et al., 2012; JABBARI et al., 2018; TAO et al., 2020), in addition to forest improvement (MÜLLER et al., 2017; KAINER et al., 2019; BALLESTA et al., 2020). The associations in GWAS are based on the theory of linkage disequilibrium (LD) between SNP markers and the phenotypic traits of interest, and have different methods associated with its application.

Meuwissen et al. (2001) proposed the concept of GS using genome-wide single nucleotide polymorphism (SNPs). Soon after, using the information from pedigree-based best linear unbiased prediction (BLUP) using mixed models (HENDERSON, 1975), and more recently the genomic best linear unbiased prediction (GBLUP) (HAYES et al., 2009) and the random regression of the best linear unbiased prediction (RR-BLUP) (WHITTAKER; THOMPSON; DENHAM, 2000; MEUWISSEN; HAYES; GODDARD, 2001), offered quick and accurate predictions for traits of interest. Since then, several extensions of these statistical methods have been developed, using penalized regression models (WALDRON et al., 2011; LI; SILLANPÄÄ, 2012), Bayesian modeling (CROSSA et al., 2010; HABIER et al., 2011; LEGARRA et al., 2011; XU, 2007), and nonparametric and semiparametric regression models (HOWARD; CARRIQUIRY; BEAVIS, 2014; BUDHLAKOTI et al., 2020). From this, studies have compared the statistical power of GS studies (CHANG et al., 2018; LEBEDEV et al., 2020; MISZTAL; STEIN; LOURENCO, 2022); yet, depending on the trait and population under study, applying different models may provide better results for accuracy (BERNARDO, 2021). For most studies, GBLUP has shown to have more robust predictions, achieving the highest correlations between predicted and observed genotypes (WANG; YANG; XU, 2015; XU; XU; XU, 2017).

The implementation of GS in *Eucalyptus* breeding has several advantages, of these a reduction in selection cycle length stands out as it is the most prominent obstacle encountered in tree breeding programs (REZENDE; DE RESENDE; DE ASSIS, 2014). Genomic selection enables the prediction of genotype behavior from models trained for genotypes of the same species and based on genetic composition, considering the occurrence of SNPs and their influence on phenotypic traits (GODDARD; HAYES, 2007). The superiority of GS models is due to the use of molecular markers that cover a large part of the genome, ensuring that all QTLs are in LD with at least one marker (MEUWISSEN; HAYES; GODDARD, 2001). The use of

methodologies such as accounting for dominance variance (PALUDETO et al., 2021), optimizing the training populations (CERICOLA et al., 2017; TAYEH et al., 2015; BERRO et al., 2019), and using multi-trait methodologies (BASTIAANSEN et al., 2012; JIA; JANNINK, 2012), can lead to a reduction in selection costs, while also increasing the efficiency of genomic selection models. Although using simple GS methodologies can achieve good prediction capabilities, optimized models can increase the accuracy of the GS models.

Thus, the thesis structure is divided into two chapters. The main objective of our study is to perform a GWAS to identify single and pleiotropic markers associated with growth and wood quality traits and optimize the predictive ability of genomic selection models of *E. grandis*. For chapter one, we specifically aimed to *i*) run single and multi-trait GWAS to identify significant markers associated with phenotypic expression, and *ii*) identify single and pleiotropic genes related to trait expression. On the other hand, for chapter two, we aimed to *i*) evaluate the influence of dominance on genomic selection models; *ii*) optimize the training set of genomic selection models; *iii*) perform multi-trait analysis to increase the predictive ability of genomic selection models, and *iv*) use optimization of the training set and multi-trait analysis to increase the predictive ability.

CHAPTER 1
QUANTITATIVE TRAIT LOCI RELATED TO GROWTH AND WOOD QUALITY
TRAITS IN *Eucalyptus grandis* W. Hill IDENTIFIED THROUGH SINGLE AND
MULTI-TRAIT GENOME-WIDE ASSOCIATION STUDIES

ABSTRACT

The genetic improvement of forest tree species is a constant challenge mainly due to the complex genetic nature of quantitative traits. Nevertheless, understanding the pleiotropic effects of genes and polygenic inheritance is crucial for tree breeding programs. Thus, the aim of this study was to conduct single- and multi-trait genome wide association studies (GWAS) and identify quantitative trait loci (QTLs) for the expression of phenotypic traits in *Eucalyptus grandis*. We evaluated an open-pollinated breeding population with 1,772 genotypes composed of 27 different families established using a randomized complete block design. We performed single-trait GWAS using the fixed and random model circulating probability unification (FarmCPU), and multi-trait GWAS for genetically correlated phenotypic traits using the multi-trait mixed model (MTMM). Then, gene annotation was identified through the Phytozome database. The FarmCPU model identified 43 and 38 QTLs that are significantly associated with growth and wood quality traits, respectively. Similarly, 40 pleiotropic QTLs were discovered using the MTMM model. Gene ontology for single-trait analysis identified genes responsible for regulating several important biological processes in different tissues and at different stages of maturation. On the other hand, the multi-trait model identified genes associated with gibberellin signaling, which regulates several aspects of plant growth and development, as well as genes related to the reinforcement of cell wall composition. Our study demonstrates the complex nature of *E. grandis* quantitative traits and provides new evidence for the architecture of genes associated with the expression of important phenotypic traits.

Key-words: association mapping, genome-wide association study, *Eucalyptus*, gene ontology

1

¹ Formatted according to the “Tree Genetics and Genomes” submission guidelines.

1.1 INTRODUCTION

Genome wide association studies (GWAS) are used to identify significant associations among quantitative traits and genetic loci in plant and animal genomes (Bush and Moore 2012). GWAS have been used extensively to understand the genetic complexity of economically important traits in tree species (Korte and Farlow 2013). Within the *Eucalyptus* genus, the species *Eucalyptus grandis* stands out because of its fast growth, high adaptability, and superior wood quality (Malan 1993). It is the most commonly planted hardwood tree globally, with a diverse range of applications in cellulose, paper, timber, and charcoal production (Malan and Gerischer, 1987; Grattapaglia 2008; Carocha et al. 2015).

Cellulose in particular is a key wood product that meets a wide variety of primary human needs, such as paper (Hollertz et al. 2017; Jin et al. 2021), pharmaceuticals (Beyger and Nairn 1986; Giri et al. 2020), biofuels (Carere et al. 2008; Rubin 2008; Carroll and Somerville, 2009), and food (Lavanya et al. 2011; Shi et al. 2014). Therefore, tree breeding strategies should focus on selecting genotypes considering not only growth characteristics, but also wood quality traits (Byram et al. 2005; Apiolaza et al. 2013). To improve the quality of wood production, several studies have emphasized the importance of finding the genetic basis of wood quality traits such as lignin (Li et al. 2008; Hisano et al. 2009), syringyl/guaiacyl ratio (Stackpole et al. 2011; Denis and Bouvet, 2013), wood density (Osorio et al. 2001; Stackpole et al. 2010), total extractives (Gallo et al. 2018; Makouanzi et al. 2018), and cellulose yield (Schimleck et al. 2004; Kien et al. 2009). In this context, developing new GWAS strategies are essential for identifying associations between genomic regions of the traits of interest and those significantly associated with the phenotype (Hirschhorn and Daly 2005).

Several studies have identified genes related to the expression of growth and wood quality traits in *Eucalyptus* (Resende et al. 2017; Müller et al. 2017, 2019). Generally, growth traits tend to be more correlated with moderate levels of heritability, while wood quality traits are less correlated, but commonly present higher levels of heritability (Mphahlele et al. 2020). For *Eucalyptus*, Kainer et al. (2019) examined the genetic effects on oil yield, while Resende et al. (2017) conducted regional heritability mapping for growth and wood quality traits to identify quantitative trait loci (QTLs). Nevertheless, few studies have sought to understand the genetic effect of pleiotropic

genes in *Eucalyptus* by comparing single-trait and multi-traits GWAS (Tan and Ingvarsson, 2018; Rambolarimanana et al. 2018).

Pleiotropic effects occur when genetic loci have an influence on more than one trait (Solovieff et al. 2013). The application of pleiotropy in breeding means a movement away from selecting for one trait at the genetic level to selecting for multiple traits at a phenotypic level (Paaby and Rockman 2013). Although single-trait GWAS has identified the polygenic inheritance effect of markers, several efforts have been made to understand the pleiotropism between quantitative traits (Liu and Yan 2019), such as multiple trait selection assisted by genetic markers. Among single-trait GWAS algorithms, the fixed and random model circulating probability unification procedure (FarmCPU) performs a multi-locus linear mixed model (MLMM) to effectively control for spurious associations (Liu et al. 2016). On the other hand, multi-trait mixed models (MTMM) were developed by Korte et al. (2012) to perform multi-trait GWAS and examine the common genetic effects that act in pleiotropy on two correlated phenotypic traits.

The MTMM algorithm performs three different analyses, categorized as full, common, and interaction. While the full model considers both common and interaction effects, the common and interaction models separate these effects individually. Thus, the common model performs a statistical analysis that demonstrates the coincident effects on two traits. Meanwhile, the interaction model identifies interacting genetic effects that act in the opposite direction between two traits (Korte et al. 2012). In the presence of pleiotropy, the power of the multi-trait GWAS is superior to single-trait GWAS because of the additional accuracy obtained when data for two traits are considered together (Korte et al. 2012; Korte and Farlow 2013; Oladzad et al. 2019).

The present study focused on using GWAS to assess the genetic architecture of growth and wood quality traits of an open-pollinated *E. grandis* seed orchard. The specific objectives of the present study were to: (1) develop and compare the performance of single- and multi-trait GWAS models in the identification of significant SNP markers related to growth and wood quality traits; (2) identify QTLs significantly associated with the expression of phenotypic traits; (3) and understand the pleiotropic effects and the genetic architecture of important traits.

1.2 MATERIAL AND METHODS

1.2.1 Plant material and phenotypes

The study population was an open-pollinated seed orchard of *E. grandis* located in the municipality of São Miguel Arcanjo, São Paulo, Brazil (-23.890188, -47.937138). The population was established in September 2012 by the SUZANO company's breeding team. The experiment consisted of a randomized complete block design, with four blocks, each containing 27 families (treatments) and one clonal control test (commercial clone), with four plots of 20 individuals each (five plants per plot). The spacing between plants was 3 m × 2 m, resulting in a planted area of 1.344 ha with 2,240 trees. The open-pollinated seeds used to establish the experiment were collected from seven different locations across Brazil (Rio Claro - São Paulo (SP); Teixeira de Freitas - Bahia; Biritiba Mirim - SP; Salto - SP; Sarapui - SP; Mogi Guaçu - SP; and São Simão - SP) and one from Zimbabwe, Africa. The 27 families are originally from Coff's Harbour (New South Wales, NSW) and Atherton (Queensland, QLD), Australia.

For the analysis, we considered the genomic and phenotypic information from 1,772 individuals. The control genotype was an *E. grandis* commercial clone used by the SUZANO company. The phenotypic information was subdivided into growth traits (GWT) and wood quality traits (WQT). Growth traits were measured at two different ages (three and six years after planting) and were classified as height (HEI3 and HEI6) in meters and diameter at breast height (DBH3 and DBH6) in centimeters. The DBH (DBH3/DBH6) and height (HEI3/HEI6) were used to estimate tree volume at three and six years of age (VOL3 and VOL6, respectively) in cubic meters according to formula described by Schumacher and Hall (1933):

$$VOL = DBH^2 \times \frac{\pi}{40000} \times HEI \times f \quad (1)$$

Further, we analyzed six wood quality traits related to cellulose production. To do so, an increment borer was used at breast height to collect wood cores of 12 mm at 6.5 years after planting. Then, wood material was sent to the laboratory for processing to obtain spectral information using near-infrared spectroscopy (NIRS).

Sawdust samples from 69 genotypes were retained in a mesh sieve and placed in circular cells. The NIR reflectance spectra were obtained using scans of wavelength ranges. Curve calibration was based on samples from five different species (*Eucalyptus grandis*, *Eucalyptus urophylla*, *Eucalyptus brassiana*, *Eucalyptus tereticornis*, and *Eucalyptus pellita*) collected in three different regions of Brazil (Maranhão, São Paulo, and Bahia) at six years after planting. An internal company calibration (SUZANO S. A.) model was developed using the Bruker FT-NIR spectrophotometer MPA II. The resulting calibration database containing NIR wood spectra was obtained through following methods outlined in the reference literature. The prediction of constituent values based on existing calibration curves were used to estimate the following wood quality traits: pure cellulose yield (PCY) in percentage; basic wood density (WBD) in cubic meters; syringyl/guaiacyl ratio (SGR); soluble lignin (SOL) in percentage; total solid content (TSC); and total extractives (TEX) in percentage.

1.2.2 Phenotypic data analysis

Each of the 1,772 samples was evaluated using the Bonferroni outlier test to find the mean-shift outlier with studentized residuals in linear mixed models. Thus, outliers were removed by deleting observations based on standard deviation with the car package in the R software environment (Fox et al. 2012). Then, the normal distribution of phenotypic data was verified using the Shapiro-Wilk test, and data normalization was performed using the bestNormalize package in R (Peterson 2021). Finally, with the normalized dataset, the best linear unbiased predictions (BLUPs) (Rodriguez et al. 2020) were estimated for each trait with the breedR package in R (Muñoz and Sanchez, 2015) using the following mixed model:

$$Y_{ijk} = \mu + Xb_j + Zt_i + Zp_k + \varepsilon_{ij} \quad (2)$$

where, μ is the average mean; b_j is the fixed effect of the j^{th} block; t_j is the fixed effect of the j^{th} family effect (progeny); P_k is the random effect of the j^{th} plot with $p \sim N(0, \sigma_p^2)$; and ε_{ij} is the residual error that represents the nongenetic effects. The matrices X and Z are the incidence matrices for the fixed and random effects, respectively.

Deregressed best linear unbiased prediction/predictor (dBLUP) were then estimated to avoid shrinkage properties (Henderson 1975) according to the formula $\frac{\hat{g}}{r^2}$ (Garrick et al. 2009), where: \hat{g} is the genomic BLUP; and r^2 is the reliability, estimated as $1 - (PEV/\sigma_g^2)$, where PEV is the prediction error variance and σ_g^2 is the genotypic variance. Pearson's genetic correlation tests were then performed using the BLUPs to verify the correlation between the 12 growth and wood quality traits. Correlation distributions were plotted using the ggcorrplot package in R (Kassambara 2019).

1.2.2 DNA extraction and quality control

Cambium tissue was collected individually from 1,772 trees and processed using the CTAB Lysis Buffer. DNA was extracted using the CTAB method (Doyle and Doyle 1987). DNA integrity was confirmed in 1% agarose gel electrophoresis and quantified by the Nanodrop spectrophotometer (Thermo Fisher, Waltham, MA, USA). DNA genotyping was performed using the EUChip60K high-density Illumina Infinium SNPchip for *Eucalyptus* species (Silva-Junior et al. 2015). Duplicate SNPs were eliminated from the raw dataset based on markers with the lowest call rate. Quality control was conducted using the R package snpReady (Granato et al. 2018). Markers were removed if they were monomorphic or had a call rate lower than 95%. Alleles with minor allele frequency (MAF) lower than or equal to 0.01 were also excluded. The genotypes were coded as "0" and "2" for homozygotes and "1" for heterozygotes. The remaining genotypic data was imputed using the R package snpReady considering Wright's equilibrium of the probability of occurrence considering the combination of allelic frequency and heterozygosity observed from the markers (Granato et al 2018). Later, the filtered markers were submitted to linkage disequilibrium (LD) pruning, removing markers with a pairwise r^2 higher than 0.99. This step was performed using the SNPRelate package in R (Zheng et al. 2012). After quality control, high-quality SNPs were selected for association mapping.

1.2.3 SNP repositioning

We repositioned the markers using the information from the SNP probes in Illumina. Probe sequences were used to align with the second version of the *Eucalyptus grandis* reference genome (v2.0) (<https://data.jgi.doe.gov/refine->

download/phytozome?genome_id=297) with the bowtie 2 aligner (Langmead and Salzberg 2012) and sensitive global alignment settings. The SNP position from version 2.0 was used in the GWAS analysis. We removed all scaffolds from the Brasuz v2.0 that were not in the linkage groups (from chromosome 1 to 11). The success of the repositioning was analyzed using a comparison map, and the dotplot coincidence graphs of the positioning of the two reference genomes (v1.0 and v2.0) were plotted using the R packages RIdeogram (Hao et al. 2020) and ggplot2 (Wickham 2011), respectively.

1.2.4 Genetic parameters and population structure

The effective population size (N_e) was estimated using the molecular linkage disequilibrium method (Waples and Do 2008) as implemented in NeEstimator V2.1 (Do et al. 2014). Population genetic parameters were estimated using the popgen function in the R package SNPReady (Granato et al. 2018), and include: Nei's genetic diversity, as $GD = 1 - p_j^2 - q_j^2$; polymorphic information content, where $PIC = 1 - (p_j^2 + q_j^2) - (2p_j^2q_j^2)$; and minor allele frequency using the formula $MAF = \min(p_j, q_j)$. The observed heterozygosity (H_o) was obtained with the formula: $H_o = nH_j/N$, where H_j is the number of heterozygous individuals and N is the number of individuals. For each trait, we estimated the narrow-sense ($h_a^2 = \sigma_a^2 / \sigma_p^2$) and broad-sense ($h_g^2 = \sigma_a^2 + \sigma_d^2 / \sigma_p^2$) heritability, where σ_a^2 represents the additive variance and σ_p^2 is the phenotypic variance. Then, the degree of differentiation between the two origin populations (F_{ST}) was estimated using the formula $F_{ST} = 1 - H_S/H_T$, where: H_S is the average expected heterozygosity for each population (two different origins); and H_T is the expected heterozygosity in the total population.

The population structure was first analyzed by a principal component analysis (PCA) using genotypic data, where the first two principal components (PC1 and PC2) were used to determine the extent of population structuration. The two different origins were represented by different colors. We subsequently used the ADMIXTURE software to identify different genetic clusters with a fixed number of populations (K) ranging from 1 to 40. Genetic correlation between phenotypes was estimated using the BreedR package in R (Munoz and Rodriguez 2014). Correlation was estimated in pairs considering the same model used to estimate BLUPs (Item 2.2). The genomic

kinship matrix (G_a) was obtained using the SNPReady package in R (Granato et al. 2018), following VanRaden (2008), with the following equation:

$$G_a = \frac{Z_A Z_A^T}{2 \sum_1^{m_i} p_i (1 - p_i)} \quad (3)$$

where Z_A is a matrix coded as 0 for homozygote A_1A_1 , 1 for heterozygote A_1A_2 , and 2 for homozygote A_2A_2 ; p_i is the frequency of an allele from locus i ; and Z is an $n \times m$ matrix of marker incidence (n is the number of genotypes and m is the number of markers).

1.2.5 Linkage disequilibrium (LD) decay

Genome-wide pairwise linkage disequilibrium (LD) was estimated for each chromosome using the function *LD.decay* from the sommer package v 2.9 in R v 4.0.2 (Covarrubias-Pazarán 2016). LD was estimated by the squared allele frequency correlation r^2 between marker pairs, and the decay was plotted considering the first distance classes based on the marker matrix and a map with distances between SNPs on a loess curve. To investigate the average LD decay in the whole genome and within chromosomes, significant intra-chromosomal r^2 values were plotted against the genetic distance between markers using the ggplot2 package in R (Wickham 2011).

1.2.6 Genome-wide association study

We performed single-trait GWAS using the fixed and random model circulating probability unification (FarmCPU) (Liu et al. 2016) and multi-trait GWAS using the multi-trait mixed model (MTMM) (Korte et al. 2012) to identify genetic factors associated with the expression of phenotypic traits. The corrected phenotypic data (BLUP) and the genotypic information were used for single- and multi-trait GWAS. The single-trait association was performed using the genome association and prediction integrated tool (GAPIT) (Lipka et al. 2012; Tang et al. 2016). The population structure based on PCA matrix (Q) and kinship (K) were automatically generated (VanRaden 2008; Lipka et al. 2012) using genotypic data and the default GAPIT parameters. Using

the GWAS results, we estimated the phenotypic variance explained by a significant marker (PVE), described as follows:

$$PVE = 2 * (\beta^2) * MAF * (1 - MAF) \quad (4)$$

where, β is the effect of allele substitution and MAF is the minor allele frequency of markers. The pleiotropic effect among phenotypic traits, which is a SNP marker having an effect on two or more traits, was estimated using the multi-trait mixed model (MTMM) (Korte et al. 2012). We performed multi-trait GWAS in pairs for the significantly associated growth and wood quality phenotypic traits. The R scripts provided by Korte et al. (2012) partition the interaction effects into three different analysis models: interaction, common, and full. Thus, considering two traits using a single marker model, the MTMM model can be written as (Korte, 2012):

$$y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = s_1\mu_1 + s_2\mu_2 + x\beta + (x \times s_1)\alpha + v \quad (5)$$

where y_1 and y_2 are phenotypic values for genotype interactions of two traits. The y value is estimated as $X\beta + u_G + u_{G \times E} + e$, considering u_G and $u_{G \times E}$ are the genotype and genotype-by-environment interaction values; s_1 and s_2 are vectors of 1 or 0 for all values of the trait in question; μ_1 and μ_2 are the means; x is the marker effect; β represents the effect size of fixed effects; and v is the prediction error. The interaction and common models identify markers that act differentially or in the same direction for two traits. On the other hand, the full model identifies SNPs with either an interaction or common effect. The significance threshold used for the p -values estimated by single- and multi-trait GWAS was calculated using the Bonferroni method ($\alpha = 0.05$). The p -values ($-\log_{10} P$) for each evaluated SNP and model was used to generate Manhattan and QQ (quantile-quantile) plots using the R package CMPlot (Yin, 2018).

1.2.7 Gene ontology

The significant SNPs for growth and wood quality traits were used to conduct a gene ontology analysis according to the physical distance within the GWAS peak regions. Since there were no strong LD blocks along the genome, which is probably related to LD-pruning, the downstream and upstream distance to search for candidate

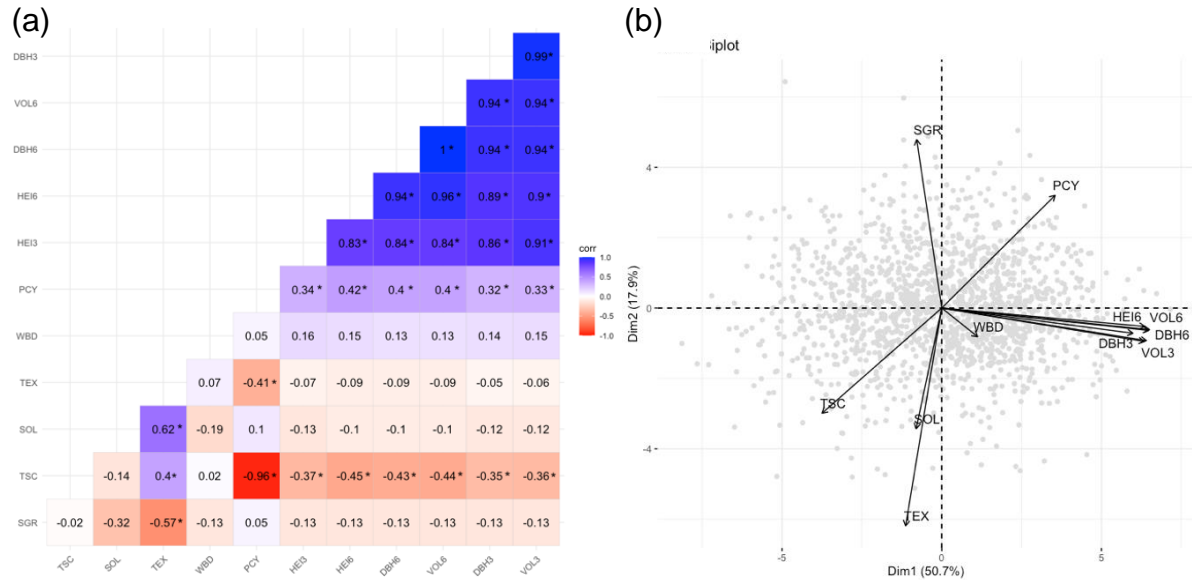
genes were estimated considering the distance of the two nearest flanking markers to the significant SNP. The genetic annotation and predicted functional effect of each gene were obtained by searching the database for version 2.0. of *E. grandis* from Phytozome v11.0 (Egrandis_297_v2.0.gene.gff3.gz). Ven diagrams were developed using the jvenn plot (Bardou et al. 2014).

1.3 RESULTS

1.3.1 Phenotypic data

The number of outliers removed varied among the 12 phenotypic traits (DBH3: 63; HEI3: 111; VOL3: 2; DBH6: 0; HEI6: 6; VOL6: 1; PCY: 22; WBD: 21; SGR: 111; TSC: 23; SOL: 78; and TEX: 83). The genetic correlation among phenotypic traits ranged from -0.96 (PCY/TSC) to 1 (DBH6/VOL6) (Figure 1a). Similarly, the highest correlation between wood quality traits was found between TEX and SOL (0.62). The PCA biplot represents the first two components for the full set of 12 traits (six growth and six wood quality). The first two axes account for 50.7% and 17.9% of the variation in the phenotypic data (Figure 1b).

Figure 1 - (a) Genotypic correlation between phenotypic traits for growth and wood quality categories across the 1,772 *Eucalyptus grandis* genotypes; (b) Principal component analysis for wood quality and growth traits

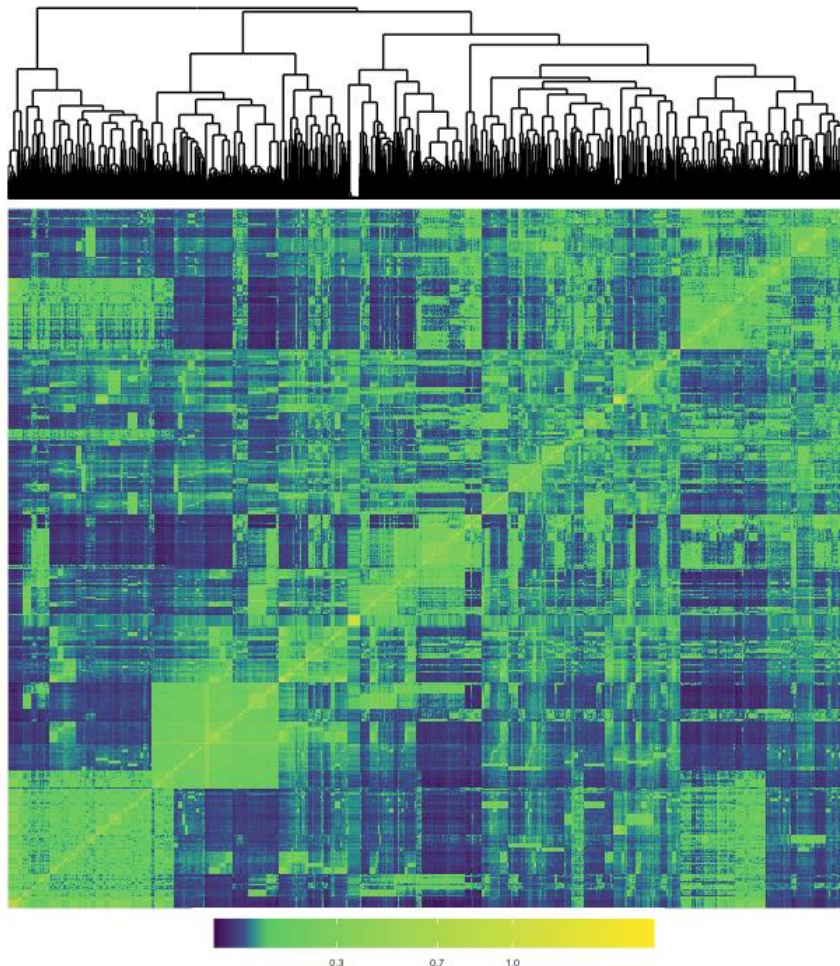


DBH3 = Diameter at breast height at 3 years; DBH6 = DBH at 6 years; VOL3 = Volume at 3 years, VOL6 = Volume at 6 years; HEI3 = Height at 3 years; HEI6 = Height at 6 years; PCY = Pure cellulose yield; WBD = Basic wood density; SGR = Syringyl/guaiacyl ratio; SOL = Soluble lignin; TSC = Total solid content; and TEX = Total extractives.

1.3.2 Population structure and genetic diversity parameters

The PCA using genotypic data revealed that the first component was mainly responsible for the genetic variation (55%) (Figure S1). Although there was a slight grouping of genotypes according to their origin by PCA, the ADMIXTURE analysis showed an absence of population genetic structure (Figure S2). Accordingly, the genetic differentiation (F_{ST}) between individuals from two different origins presented a value of 0.036, indicating limited genetic divergence between them. A similar pattern was found for the kinship matrix (VanRaden 2008), where different subpopulations were identified but with no evidence of a strong population structure. Although the genotypes evaluated are originally from two native populations, the seeds which were used to establish the breeding population are from open-pollinated trials installed in eight different locations. Thus, we believe that crossings among genotypes from different origins may have generated stratification in the population.

Figure 2 - Kinship relationships for the *Eucalyptus grandis* breeding population of 1,772 individuals using the 21,254 SNPs based on the VanRaden method



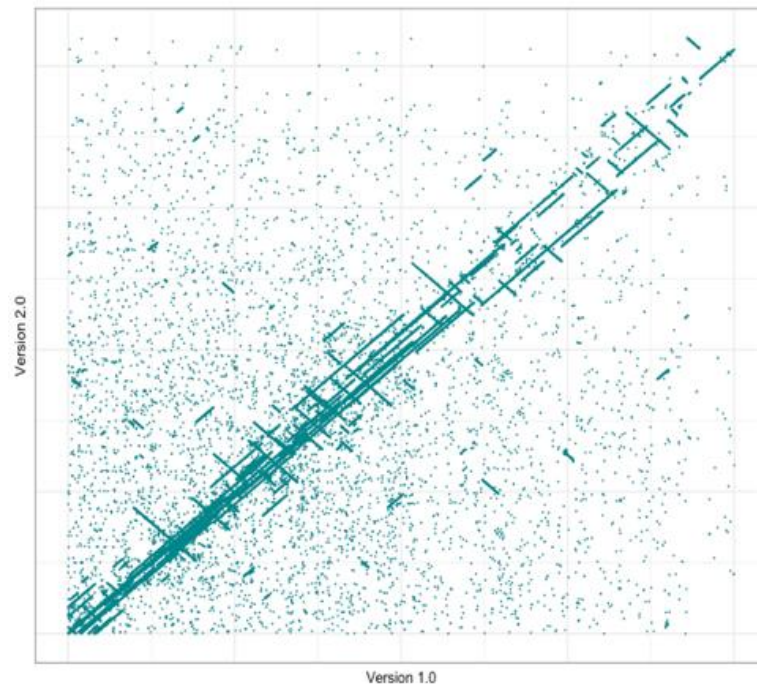
In general, genetic diversity parameters showed moderate values. Nei's genetic diversity of the whole population ranged from 0.07 to 0.50, with an average of 0.35. The marker polymorphic information content (PIC) ranged from 0.07 to 0.38, with an average of 0.28. The minimum allele frequency (MAF) showed a mean value of 0.26, ranging from 0.04 to 0.50. The observed heterozygosity (H_o) had an average of 0.40, ranging from 0.24 to 0.47. Similarly, the inbreeding coefficient ranged from 0.04 to 0.50, with a mean value of 0.26. We found an effective population size (N_e) of 31.5 considering linkage disequilibrium between markers (LDN_e).

1.3.3 SNP repositioning and quality control

In general, several SNPs changed their original relative position between the first and second version of the *E. grandis* reference genome, and some even changed

chromosomes. However, the genome-scale SNP collinearity (Figure 3) between the two versions showed that most SNPs maintained similar positions. We did note a high collinearity pattern and more reliable linkage maps with version 2.0 (Bartholomé et al. 2015). Thus, we chose SNP positions estimated using the second version to perform GWAS analysis and identify QTLs and candidate genes related to trait expression.

Figure 3 - Synteny dotplot for SNP positions across the two versions of the *Eucalyptus grandis* genome. The x-axis represents version 1.0 and the y-axis represents version 2.0

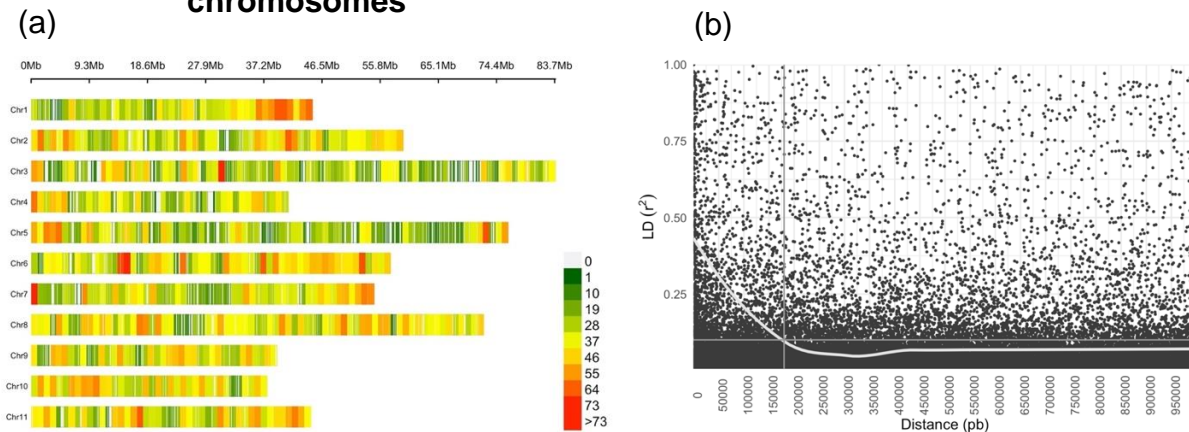


For quality control, from the initial total of 64,639 markers, 3,425 duplicate SNPs were removed considering the call rate, leaving 61,214 markers. After SNP repositioning, 1,946 markers were removed as they were located in small scaffolds. A total of 28,957 markers were removed due to MAF (0.05), and 8,229 markers were removed due to the call rate (0.9), leaving 22,082 markers. Furthermore, 1.08% missing points were imputed. Finally, after LD pruning, 828 SNPs with high linkage disequilibrium were removed, leaving a final total of 21,254 markers for the analysis.

The informative SNPs selected were uniformly distributed across the 11 chromosomes of the *E. grandis* genome. Figure 4a shows the occurrence of SNPs along the *E. grandis* chromosomes, where the number of SNPs is summed within adjacent 1 Mb windows. LD showed a quick and similar decay pattern across the 11

E. grandis chromosomes (Figure 4b). The *ad-hoc* value of r^2 (0.10) indicated an average LD across chromosomes ranging from 150 kb to 200 kb (Figure S4).

Figure 4 - (a) SNP density plot across each chromosome representing the number of SNPs after quality control within a 1 Mb window size; (b) Pairwise LD-decay across the 11 chromosomes of the 1,772 individuals genotyped using the EUChip60K. Different colors represent different SNP density and “Chr” represents the *E. grandis* chromosomes



1.3.4 Genome-wide association study

1.3.4.1 Broad- and narrow-sense heritability and gene annotation

For growth traits, we found moderate values of narrow-sense heritability, ranging from 0.4299 (HEI3) to 0.5816 (DBH6) (Table 1). Three wood quality traits (SGR, SOL, and TEX) presented relatively low narrow-sense heritability (0.1599, 0.1845, and 0.1515, respectively). On the other hand, pure cellulose yield (PCY) presented the highest broad-sense heritability (0.7107) among all growth and wood quality traits.

The FarmCPU model successfully performed single trait GWAS, indicating significant associations between growth and wood quality traits in *E. grandis*. After Bonferroni correction, a total of 81 SNPs with a significant association were identified for six growth traits (43 SNPs) and five wood quality traits (38 SNPs). Only the wood quality trait total extractives (TEX) showed no significant associations with markers (Table 1). The number of significant markers associated with phenotypic traits ranged from 2 (DBH6) to 14 (WBD) (Figure 5a and Figure 5b, respectively).

The average minor allele frequency (MAF) for the significant markers ranged from 0.0946 (DBH6) to 0.3164 (TSC). For all significant SNPs, the total phenotypic

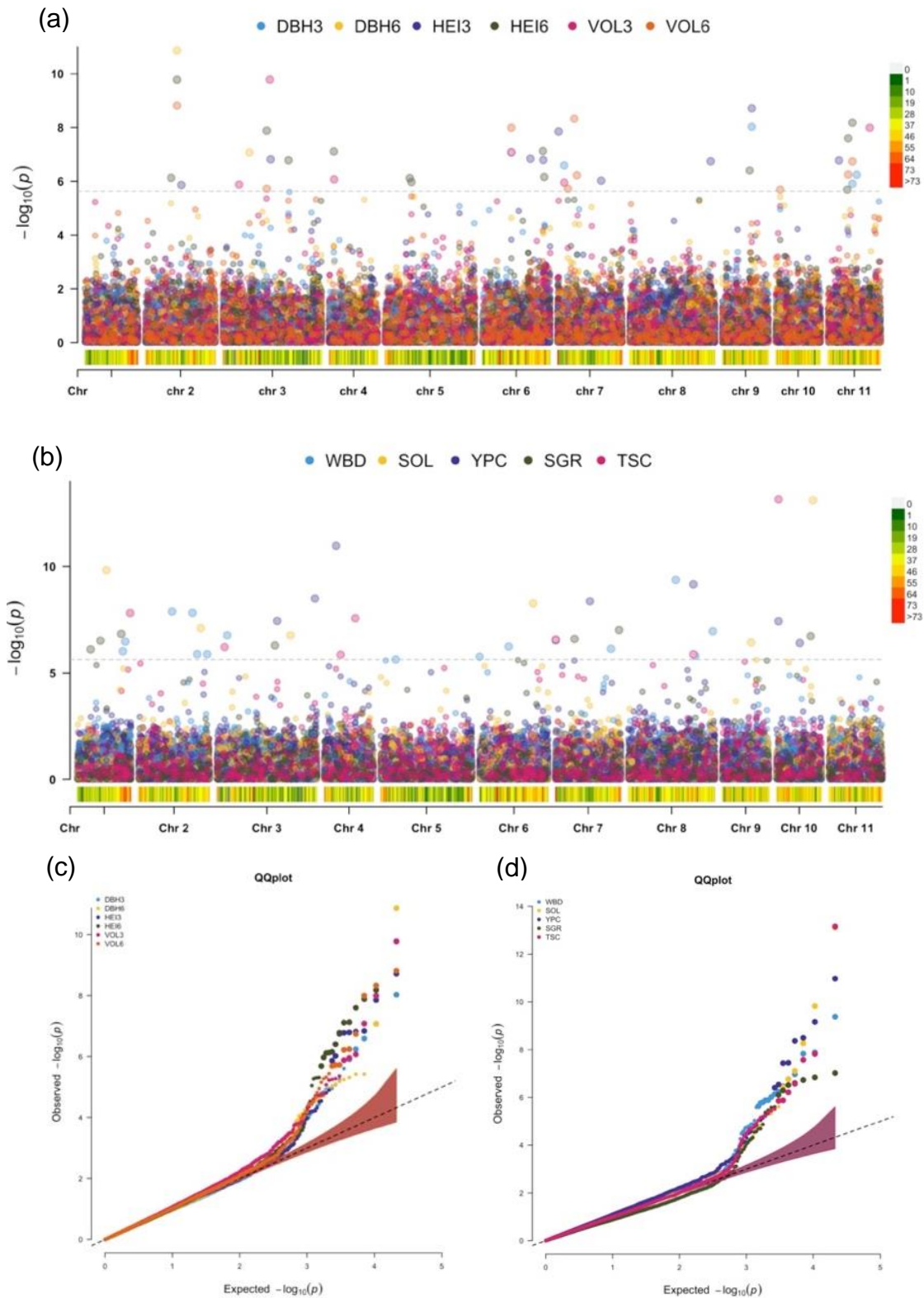
variance explained by a given SNP (PVE) was low, ranging from 0.0529 (SOL) to 0.2110 (PCY). Marker EuBR04s9558885 (PCY) showed the highest phenotypic variance (0.1014), suggesting a strong influence of this marker on phenotypic expression. Several markers were found associated with multiple phenotypic traits for trait expression and candidate gene annotation (Figure 5).

Table 1 - Significant associations for growth and wood quality traits using the single-trait model (FarmCPU) for a *Eucalyptus grandis* breeding population. Traits are divided into growth (GWT) and wood quality (WQT). The number of SNPs, MAF, PVE, and number of genes are related to the significant number of associations found by the FarmCPU model.

Type	Trait	h_a^2	h_g^2	SNPs	MAF	PVE	Genes
GWT	DBH3	0.5657	0.6764	5	0.1868	0.0976	10
	HEI3	0.4299	0.5543	8	0.2549	0.2108	28
	VOL3	0.5614	0.6537	6	0.1615	0.1039	15
	DBH6	0.5816	0.6775	2	0.0946	0.0551	0
	HEI6	0.5640	0.6837	13	0.2285	0.1502	44
	VOL6	0.5504	0.6322	9	0.1789	0.1910	21
WQT	PCY	0.6056	0.7107	8	0.3146	0.2110	22
	WBD	0.5702	0.5931	14	0.2404	0.1991	45
	SGR	0.1599	0.1774	7	0.2363	0.0867	10
	SOL	0.1845	0.1845	6	0.2423	0.0529	28
	TSC	0.5355	0.5934	3	0.3164	0.1197	14
	TEX	0.1068	0.1515	0	-	-	0

h_a^2 = Narrow-sense heritability; h_g^2 = Broad-sense heritability; SNPs = Number of significant SNPs; MAF = Average minor allele frequency; PVE = Sum of phenotypic variance explained by the significant SNPs; Genes = Number of genes found; GWT = Growth-traits; WQT = Wood quality traits; DBH3: Diameter at breast height at 3 years; DBH6: DBH at 6 years; VOL3: Volume at 3 years, VOL6: Volume at 6 years; HEI3: Height at 3 years; HEI6: Height at 6 years; Pure cellulose yield (PCY); Basic wood density (WBD); Syringyl/guaiacyl ratio (SGR); Soluble lignin (SOL); Total solid content (TSC); and Total extractives (TEX).

Figure 5 - Manhattan and QQ-plots of GWAS for growth traits (a,c) and wood quality traits (b,d), respectively, using the FarmCPU model for an *Eucalyptus grandis* breeding populations with 21,254 markers. Different colors represent different tested traits. Dashed line indicates the Bonferroni threshold ($\alpha = 0.05$)



The number of annotated genes associated with the expression of phenotypic traits ranged from 0, with no gene annotation for the significant SNP (DBH6), to 46 (WBD). We found QTLs significantly related to more than one trait for both categories (Table S1; Figure S3). In general, functional gene annotation presented several categories and descriptions associated with tissue growth on cell walls, cellulose biosynthetic process, transporter activity, DNA, ion and protein binding, oxidation-reduction process and catalytic activity, among others. The function and description of all candidate genes for both growth and wood quality traits are shown in Table S1.

The pleiotropic effect among genes for growth traits was first seen for the SNP marker EuBR09s24960947, which presented the most significant association for traits DBH3 and HEI3, with p -values of 9.34×10^{-9} and 1.91×10^{-9} , respectively. This marker tags seven different genes (Eucgr.I01459, Eucgr.I01460, Eucgr.I01461, Eucgr.I01462, Eucgr.I01463, Eucgr.I01464, and Eucgr.I01465). In general, the single-trait GWAS revealed 13 candidate genes significantly associated with DBH3. Similarly, most genes found for HEI3 also showed comparable genome locations and molecular functions. Considering the trait HEI3, we found no annotation for candidate genes located near three significant SNPs (EuBR07s925067, EuBR06s38139098, and EuBR08s70063929) (Table S1). On the other hand, for HEI6, we found 44 annotated genes related to trait expression with different descriptions and gene ontology terms.

The SNP EuBR11s17004419 (HEI6) showed four different flanking genes (Eucgr.K01383, Eucgr.K01384, Eucgr.K01385, and Eucgr.K01386). The marker EuBR06s23565060 was identified for both ages for volume (VOL3 and VOL6) and for DBH3 (p -values 8.29×10^{-8} ; 1.01×10^{-8} ; 8.44×10^{-8} , respectively). We found a similar pattern of significant SNPs correlated with more than one phenotypic trait for wood quality in single-trait GWAS. For PCY and TSC, three different SNPs (EuBR07s252985, EuBR08s57640594, and EuBR10s1696823) were significantly correlated with the expression of these traits. These two phenotypic traits (PCY and TSC) presented the highest negative correlation (-0.96), indicating that negative correlations can be effective in identifying pleiotropic genes.

Regarding pure cellulose yield, SNP marker EuBR04s9558885 had the highest significance (p -values 1.06×10^{-11}), presenting five different flanking genes related to trait expression (Eucgr.D00522, Eucgr.D00523, Eucgr.D00525, Eucgr.D00526, and Eucgr.D00527). The gene ontology terms suggested association with a lipid metabolic process. Additionally, SNP EuBR10s1696823 was related to PCY, with the presence

of gene Eucgr.J00155, a wound-induced protein that plays a role in reinforcing cell wall composition. Similarly, for TSC, we found 12 annotated genes significantly associated with trait expression. The marker EuBR01s39512949 presented six different genes related to its expression (Eucgr.A02909, Eucgr.A02910, Eucgr.A02911, and Eucgr.A02912). Gene Eucgr.A02909 is a sugar and monosaccharide transmembrane transporter.

1.3.4.2 Multi-trait genome-wide association

The multi-trait GWAS showed good performance for all significant combinations among traits (Figure S3). We found significant marker-phenotype associations for growth and wood quality traits that were not identified with the single-trait GWAS. Considering the 33 phenotypic correlations among the 12 traits, 22 combinations showed significant associations (Table 1; Figure S3). The MT models (full, common, and/or interaction) for some models were unable to properly perform the GWAS since the p -values seemed to be deflated, and the QQ-plot showed more noise (e.g., Fig. S4d). These models were also unable to find significant associations considering the Bonferroni correction, which suggests no influence of possible false positives on the results.

The combinations among growth traits in multi-trait GWAS resulted in the highest number of significant SNPs with pleiotropic effects (24). Further, the multi-trait GWAS analysis among wood quality (6) and between the two categories (GWT and WQT) (10) tended to express less significant markers. The multi-trait GWAS revealed 40 SNPs influencing the expression of multiple phenotypic traits (Table S6). Not surprisingly, the multi-trait methodology showed greater power to identify associations considering that most single-trait analyses using the MTMM methodology (Korte et al. 2012) could not identify associations due to the strict Bonferroni cutoff ($\alpha = 0.05$; p -value = 1.63×10^{-5}).

Significant associations between the same trait in different years of data sampling (e.g., EuBR03s72654230 for DBH3 and DBH6; EuBR04s246324 for HEI3 and HEI6; EuBR02s2712998 for VOL3 and VOL6) indicate a strong pleiotropic effect on trait association. The SNP EuBR06s39120397 presented a strong p -value, and this marker was also statistically significant in the expression of HEI3, DBH6, and VOL6. This pattern may be related to the strong genetic correlation among these traits.

Similarly, SNP EuBR01s28498846 was also found for the combination of traits TSC, DBH6, VOL6, and VOL3, indicating evidence of a pleiotropic effect of the marker on growth and wood quality traits.

We identified one SNP that was significantly associated with two trait combinations (HEI3 and DBH6; HEI3 and VOL6) (EuBR06s39120397; p -values 7.57×10^{-6} ; 1.14×10^{-5} , respectively), with nine candidate genes related to its expression (Eucgr.F02939 ~ Eucgr.F02947). Among several different functions, the genes Eucgr.F02941 and Eucgr.F02943 are described as being associated with cellulase activity. One SNP was found to be significant by the full model between the combinations of traits DBH3 and DBH6 and DBH3 and VOL6 (EuBR03s72654230; p -values 1.59×10^{-5} ; 1.02×10^{-5} , respectively). In addition, two genes were related to trait expression (Eucgr.C03882 and Eucgr.C03884) that act as amino acid transmembrane transport.

The SNP EuBR03s43394028 was significant for three combinations of traits (HEI6 and VOL6; VOL3 and HEI3; and DBH3 and HEI6) (p -values 9.33×10^{-7} ; 1.01×10^{-5} and 1.54×10^{-6}). However, between traits HEI3 and HEI6, although SNP marker EuBR03s22449999 (p -value 8.41×10^{-6}) was identified as significant by the common model, there were no annotations for candidate genes. On the other hand, SNP EuBR04s246324 was significant for the common and full models with a high p -value (9.05×10^{-7}). Considering that both traits HEI3 and HEI6 represent plant height, the power of multi-trait GWAS to detect significant candidate genes proved to be effective even for the same trait, considering different developmental stages.

One SNP marker was detected as significant for traits DBH3 and VOL3 (EuBR05s62102817; p -value 2.19×10^{-6}) (Table S1). Similarly, between traits VOL3 and DBH6, the marker EuBR07s16969079 showed significant association (p -value 1.38×10^{-5}). We found two significant markers for the first WQT combination SOL and TEX (EuBR06s19529730 and EuBR06s52964694; p -values 2.66×10^{-7} and 5.06×10^{-6} , respectively). The multi-trait GWAS combination between the traits SGR and TEX identified three significant markers (EuBR11s43922247, EuBR11s44284539, and EuBR03s16484895; p -values 7.26×10^{-6} , 1.16×10^{-5} , and 9.83×10^{-6}) through the full and interaction models. Similarly, the genomic regions for marker EuBR11s44284539 revealed two flanking candidate genes (Eucgr.K03516 and Eucgr.K03517). GO analysis between gene Eucgr.K03517 and AT3G62650 from *Arabidopsis thaliana* classified this gene as a response to light intensity and to red or far-red light.

1.4 DISCUSSION

Single and multi-trait GWAS were effective in properly identifying QTLs as well as annotated genes related to phenotypic expression in the studied *E. grandis* breeding population. Additionally, the quality control process was able to remove uninformative markers, leaving a total of 21,254 highly informative markers that were used in the GWAS analysis. In general, most of the markers removed (28,957) during quality control were due to a low minor allele frequency (< 5%), which is the frequency of the second most common allele in the population. Also, although there were some rearrangements during SNP reposition, using new SNP positions for v2.0 of the genome was effective in finding QTLs and annotated genes.

In relation to the rearrangement of the *E. grandis* genome assembly, Bartholomé et al. (2015) identified 43 non-collinear and 13 non-synthetic regions. Thus, although there are modifications in marker collinearity found by the linear trend between the two versions of the genome, the new arrangement may be related to modifications in genome assembly. We reinforce that as far as we know, this is the first GWAS study developed using repositioned SNP probes that compares the positions of the two genome versions. Furthermore, considering that gene annotation is based on the second version of the *Eucalyptus* genome (Bartholomé et al. 2015), we believe that the possibility of errors was reduced.

Another important point to consider is related to population structure. Herein, we found no clear structuration of the population between individuals, which may be related to the population's breeding history. Although there are two origins and it is likely that there would have been population structure, the breeding population was established from eight different provenances, which might have promoted outcrossing between individuals from different origins. According to Hayes (2013), not considering population structure in GWAS can cause false positive associations. Thus, both models (single- and multi-trait GWAS) were tested against population structure, and we believe that this effect did not have an impact on our results as they were considered in the analysis.

Several genetic mapping through association studies have been used to assess the complexity of the genetic architecture of growth (Müller et al. 2017, 2019), wood quality traits (Cappa et al. 2013; Resende et al. 2017; Dasgupta et al. 2021), and non-wood traits (Resende et al. 2017; Kainer et al. 2019; Mhoswa et al. 2020) of

Eucalyptus. Using the second version of the *Eucalyptus* genome, it was possible to more accurately identify QTLs. Many studies have also developed single-trait GWAS for growth, wood quality, and disease resistance in *Eucalyptus* spp. (Resende et al. 2017; Kainer et al. 2019; Müller et al. 2019; Ballesta et al. 2020; Mhoswa et al. 2020; Valenzuela et al. 2021). However, few studies have evaluated the multi-trait association models for growth and even fewer for wood quality in eucalypts (Rambolarimanana et al. 2018; Tan and Ingvarsson 2018). As expected, although several markers were found to be significant, the results from the single- and multi-trait GWAS indicate limited genetic variance, which can explain the relatively low number of associations. This pattern might be related to the polygenic nature of quantitative traits (Grattapaglia et al. 2018), indicating that there are many genes related to trait expression, as predicted by Fisher's infinitesimal model (Fisher 1918).

Although the complexity of multiple genes influences the expression of quantitative traits, the number of significant SNPs identified herein, and consequently the number of QTLs for both single and multi-trait GWAS, was similar to previous studies (Müller et al. 2019; Ballesta et al. 2020). Additionally, besides the reliable accuracy achieved by the single- and multi-trait GWAS, the phenotypic information used in the present study was obtained from a single environment, which may have limited the phenotypic precision of each individual. Thus, our study reinforces the importance of using multi-trait models combined with single-trait models for highly complex quantitative traits. According to Liu et al. (2016), the FarmCPU model offers the best trade-off between predictive power and false positives. On the other hand, the power of the MTMM approach considering the correlation between two traits (multi-trait GWAS) can improve the identification of more evident pleiotropic effects than those found using a single marginal trait analysis (Korte et al. 2012).

The implementation of GWAS using phenotypic information from different traits can lead to the discovery of effects stronger than those identified by single trait analysis (Korte et al. 2012). To increase the statistical power of GWAS, several studies have used multi-trait analysis to identify significant genetic-phenotypic associations (Jaiswal et al. 2016; Thoen et al. 2017; Yoshida and Yáñez 2021). Thus, multi-trait GWAS can increase the power of single-trait GWAS using different measures or multiple traits with a high pattern of genetic correlation (Porter and O'Reilly 2017). Regarding Pearson's genetic correlations between phenotypic traits, the strongest associations between growth variables found herein are expected because diameter, height, and volume are

directly related. On the other hand, wood quality traits did not show strong patterns of association, except for PCY which presented several significant and positive associations with growth traits. This finding suggests that selection for growth traits might lead to a large increase in cellulose yield, which for example, could have a further effect of reducing the total solid content production. Thus, pleiotropic QTLs are important when using marker assisted selection for multiple traits.

Generally, our results show that multi-trait GWAS was able to increase the power of single-trait GWAS (FarmCPU) to identify genes that directly affect mutual traits, thus increasing the capacity to identify markers with minor effects. Further, compared to the multi-trait GWAS (MTMM), the FarmCPU showed a lack of power to identify pleiotropic markers and correlated traits with low phenotypic correlation, as shown in previous studies (Korte et al. 2012). The joint association analysis which considered the full, common, and interaction models, suggested genetic factors acting in the same direction, differentially, or with an interaction or common effect for the expression of the growth and wood quality traits.

Considering single trait GWAS, several studies identified that FarmCPU increased the power of GWAS for complex traits (Tang et al. 2016; Kusmec and Schnable 2018; Miao et al. 2019). Our study corroborated this finding, identifying 81 significant markers for growth (43) and wood quality (38) traits. Furthermore, FarmCPU was able to control for false positives caused by population structure and kinship because of the distribution of quantile-quantile (QQ) plots. On the other hand, the MTMM model performed in the multi-trait GWAS identified a smaller number of significant markers (31) among all significant trait combinations (Table S6). The importance of finding pleiotropic QTLs is related to marker assisted selection, which can be used together to select multiple regions related to the expression of both growth and wood quality traits (Gupta et al. 2010). Regarding genomic heritability, low/moderate heritability levels were found for growth traits. The high/moderate heritability for the wood quality traits PCY and WBD indicates that they are less influenced by environment. However, three wood quality traits (SGR, SOL, and TEX) showed a critically low heritability, making GWAS not appropriate for these traits.

Herein, the pleiotropic effect of genes influencing the expression of phenotypic traits was primarily found through in single-trait GWAS analysis. A similar tendency was found by Ward et al. (2019) comparing yield traits in soft red winter wheat, where several markers presenting genes with pleiotropic effects were identified by the

FarmCPU model. Here, pleiotropy was identified for both growth and wood quality traits in *E. grandis*. However, the markers with a pleiotropic effect identified for different traits by single-trait GWAS were not identified when using the multi-trait GWAS. The difference of significant SNPs found in these analyses might result from the different statistical methodologies that explore GWAS associations (Hayes 2013).

Some significant QTLs identified in this study were located close to several strong functional candidate genes associated with growth and plant development. For instance, the candidate gene *Eucgr.G01887* (EuBR07s34761317) was identified as related to cytokinin expression for pure cellulose yield. According to Chakraborty and Akhtar (2021), cytokinins (CKs) are hormones that influence plant growth, development, and physiology. Several plant processes are involved with CKs, such as seed germination, apical dominance, flowering, fruit development, leaf senescence, and plant-pathogen interaction. Further, CKs can promote cell division or cytokinesis in plant roots and shoots. Additionally, according to Li et al. (2021), CKs are a class of phytohormones that regulate plant growth, development, and stress response.

We also found the candidate genes *Eucgr.K01383* and *Eucgr.K01384* (SNP EuBR11s17004419; trait HEI6) from the GRAS domain family. These play an important role in gibberellin signaling, regulating several aspects of plant growth and development (Hirsch and Oldroyd 2009). Similarly, the gene *Eucgr.J00155* is near to the marker EuBR10s1696823 and plays a role in reinforcing cell wall composition after wounding and during plant development. Considering basic wood density, SNP EuBR08s73877790 presented the flanking gene *Eucgr.H05146*, which is related to the pentatricopeptide repeat superfamily protein that acts in the cell and performs physiological functions during plant growth and development (Barkan and Small 2014).

Several markers indicated an effect on lipid metabolism (e.g., EuBR04s9558885, EuBR02s52286175, and EuBR06s42411988). According to Laskin et al. (2002), since plant cells assimilate more carbon than they can store, the excess carbon is converted into lipids, responsible for storing energy, signaling, and acting as structural components of cell membranes. Similarly, we found several candidate genes related to cell wall composition (e.g., EuBR10s1696823, gene *Eucgr.J00155*), as well as the expression of cellulase (*Eucgr.F02941* and *Eucgr.F02943*), which are enzymes responsible for breaking down the cellulose of plant cell walls into simple sugars (Thapa et al. 2020). According to MacMillan et al. (2010), wood tissue synthesis in plants is associated with the development of strong and flexible plant structures and

facilitates the transport of water and nutrients. Because of the large number of genes related to quantitative traits, the functional relationship between the identified QTLs and the phenotypic variation in growth and wood quality traits in *E. grandis* is still unclear. Thus, the identified significant genomic regions and their potential relationship with phenotypic traits must be further analyzed.

1.5 CONCLUSION

Our study highlights the importance of examining associations between markers and phenotypes for eucalypt species. Herein, we identified markers that act individually on each trait using the single-trait GWAS and markers that have pleiotropic effects and influence several traits using multi-trait GWAS. The results corroborate previously published data for eucalypt species using moderate size populations along with high-density SNP data sets. As far as we know, most of the markers identified herein have never been described in previous GWAS for eucalypt species. This result is consistent with previous theories indicating that phenotypic expression is linked to both a large number of genes as well as the effects of the environment. The results discussed herein provide a better understanding of gene expression and offer important information to inform marker assisted selection.

In terms of identifying QTLs using single and multi-trait GWAS, we were able to find clear results related to gene interaction. Gene ontology analysis of GWAS was also important in identifying the biological context of genes. The different GWAS methodologies applied involved the scanning of the whole genome from different trees and identifying genetic markers that can be used to predict phenotypic traits. As a result, GWAS effectively identified candidate genes related to the expression of phenotypic traits. We believe that the results can be used in genetic selection to increase the productivity of eucalypt plantations and improve future breeding programs. Nevertheless, further studies should be conducted to identify significant associations with multiple environmental conditions. Thus, it is essential to continue evaluating the genetic effects and the complexity of the genetic architecture of economically important traits to continue to accumulate genetic gains in each breeding cycle.

1.6 ACKNOWLEDGMENTS

We acknowledge Suzano S.A. for providing phenotypic and genotypic data. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. We also thank the German Academic Exchange Service (DAAD) which co-financed a short-term research grant (ref. no.: 91781916). Evandro V. Tambarussi is supported by a research productivity fellowship from CNPq (grant number 304899/2019-4).

REFERENCES

- Ballesta P, Bush D, Silva FF, Mora F (2020) Genomic Predictions Using Low-Density SNP Markers, Pedigree and GWAS Information: A Case Study with the Non-Model Species *Eucalyptus cladocalyx*. *Plants* 9:99
- Bardou P, Mariette J, Escudié F, et al (2014) jvenn: an interactive Venn diagram viewer. *BMC Bioinformatics* 15:1–7
- Barkan A, Small I (2014) Pentatricopeptide repeat proteins in plants. *Annu Rev Plant Biol* 65:415–442
- Bartholomé J, Mandrou E, Mabilia A, et al (2015) High-resolution genetic maps of *Eucalyptus* improve *Eucalyptus grandis* genome assembly. *New Phytol* 206:1283–1296
- Beyger JW, Nairn JG (1986) Some factors affecting the microencapsulation of pharmaceuticals with cellulose acetate phthalate. *J Pharm Sci* 75:573–578
- Bush WS, Moore JH (2012) Genome-wide association studies. *PLoS Comput Biol* 8:e1002822
- Cappa EP, El-Kassaby YA, Garcia MN, et al (2013) Impacts of population structure and analytical models in genome-wide association studies of complex traits in forest trees: a case study in *Eucalyptus globulus*. *PLoS One* 8:e81267
- Carere CR, Sparling R, Cicek N, Levin DB (2008) Third generation biofuels via direct cellulose fermentation. *Int J Mol Sci* 9:1342–1360
- Carocha V, Soler M, Hefer C, et al (2015) Genome-wide analysis of the lignin toolbox of *Eucalyptus grandis*. *New Phytol* 206:1297–1313
- Carroll A, Somerville C (2009) Cellulosic biofuels. *Annu Rev Plant Biol* 60:165–182
- Chakraborty T, Akhtar N (2021) Biofertilizers: Prospects and challenges for future. *Biofertilizers: Study and Impact* 575–590
- Covarrubias-Pazarán G (2016) Genome-assisted prediction of quantitative traits using the R package sommer. *PLoS One* 11:e0156744
- Dasgupta M, Parveen ABM, Shanmugavel S, et al (2021) Targeted re-sequencing and genome-wide association analysis for wood property traits in breeding population of *Eucalyptus tereticornis* × *E. grandis*. *Genomics*
- Denis M, Bouvet J-M (2013) Efficiency of genomic selection with models including dominance effect in the context of *Eucalyptus breeding*. *Tree Genet Genomes* 9:37–51
- Do C, Waples RS, Peel D, et al (2014) NeEstimator v2: re-implementation of software for the estimation of contemporary effective population size (N_e) from genetic data. *Mol Ecol Resour* 14:209–214

- Doyle J, Doyle JL (1987) Genomic plant DNA preparation from fresh tissue-CTAB method. *Phytochem Bull* 19:11–15
- Fisher RA (1918) XV.—The correlation between relatives on the supposition of Mendelian inheritance. *Earth Environ Sci Trans R Soc Edinburgh* 52:399–433
- Fox J, Weisberg S, Adler D, et al (2012) Package ‘car.’ Vienna R Found Stat Comput.
- Gallo R, Pantuza IB, dos Santos GA, et al (2018) Growth and wood quality traits in the genetic selection of potential *Eucalyptus dunnii* Maiden clones for pulp production. *Ind Crops Prod* 123:434–441
- Garrick DJ, Taylor JF, Fernando RL (2009) Deregressing estimated breeding values and weighting information for genomic regression analyses. *Genet Sel Evol* 41:1–8
- Giri BR, Poudel S, Kim DW (2020) Cellulose and its derivatives for application in 3D printing of pharmaceuticals. *J Pharm Investig* 1–22
- Granato I, Galli G, de Oliveira Couto E, et al (2018) snpReady: a tool to assist breeders in genomic analysis
- Grattapaglia D (2008) Genomics of *Eucalyptus*, a global tree for energy, paper, and wood. In: *Genomics of tropical crop plants*. Springer, pp 259–298
- Grattapaglia D, Silva-Junior OB, Resende RT, et al (2018) Quantitative genetics and genomics converge to accelerate forest tree breeding. *Front Plant Sci* 9:1-10
- Hao Z, Lv D, Ge Y, et al (2020) RIdiogram: drawing SVG graphics to visualize and map genome-wide data on the idiograms. *PeerJ Comput Sci* 6:e251
- Hayes B (2013) Overview of statistical methods for genome-wide association studies (GWAS). *Genome-wide Assoc Stud genomic Predict* 149–169
- Henderson CR (1975) Best linear unbiased estimation and prediction under a selection model. *Biometrics* 423–447
- Hirsch S, Oldroyd GED (2009) GRAS-domain transcription factors that regulate plant development. *Plant Signal Behav* 4:698–700
- Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* 6:95–108
- Hisano H, Nandakumar R, Wang Z-Y (2009) Genetic modification of lignin biosynthesis for improved biofuel production. *Vitr Cell Dev Biol* 45:306–313
- Hollertz R, Durán VL, Larsson PA, Wågberg L (2017) Chemically modified cellulose micro-and nanofibrils as paper-strength additives. *Cellulose* 24:3883–3899

- Jaiswal V, Gahlaut V, Meher PK, et al (2016) Genome wide single locus single trait, multi-locus and multi-trait association mapping for some important agronomic traits in common wheat (*T. aestivum* L.). *PLoS One* 11:e0159343
- Jin K, Tang Y, Liu J, et al (2021) Nanofibrillated cellulose as coating agent for food packaging paper. *Int J Biol Macromol* 168:331–338
- Kainer D, Padovan A, Degenhardt J, et al (2019) High marker density GWAS provides novel insights into the genomic architecture of terpene oil yield in *Eucalyptus*. *New Phytol* 223:1489–1504
- Kassambara MA (2019) Package 'ggcorrplot.' R Packag version 01 3:
- Kien ND, Quang TH, Jansson G, et al (2009) Cellulose content as a selection trait in breeding for kraft pulp yield in *Eucalyptus urophylla*. *Ann For Sci* 66:1–8
- Korte A, Farlow A (2013) The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods* 9:29
- Korte A, Vilhjálmsson BJ, Segura V, et al (2012) A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nat Genet* 44:1066–1071
- Kusmec A, Schnable PS (2018) Farm CPU pp: efficient large-scale genome-wide association studies. *Plant Direct* 2:e00053
- Laskin JD, Heck DE, Laskin DL (2002) The ribotoxic stress response as a potential mechanism for MAP kinase activation in xenobiotic toxicity. *Toxicol Sci* 69:289–291
- Lavanya D, Kulkarni PK, Dixit M, et al (2011) Sources of cellulose and their applications—A review. *Int J Drug Formul Res* 2:19–38
- Li S-M, Zheng H-X, Zhang X-S, Sui N (2021) Cytokinins as central regulators during plant growth and stress response. *Plant Cell Rep* 40:271–282
- Li X, Weng J, Chapple C (2008) Improvement of biomass through lignin modification. *Plant J* 54:569–581
- Lipka AE, Tian F, Wang Q, et al (2012) GAPIT: genome association and prediction integrated tool. *Bioinformatics* 28:2397–2399
- Liu H, Yan J (2019) Crop genome-wide association study: a harvest of biological relevance. *Plant J* 97:8–18
- Liu X, Huang M, Fan B, et al (2016) Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. *PLoS Genet* 12:e1005767

- MacMillan CP, Mansfield SD, Stachurski ZH, et al (2010) Fasciclin-like arabinogalactan proteins: specialization for stem biomechanics and cell wall architecture in *Arabidopsis* and *Eucalyptus*. *Plant J* 62:689–703
- Makouanzi G, Chaix G, Nourissier S, Vigneron P (2018) Genetic variability of growth and wood chemical properties in a clonal population of *Eucalyptus urophylla* × *Eucalyptus grandis* in the Congo. *South For a J For Sci* 80:151–158
- Malan FS (1993) The wood properties and qualities of three South African-grown eucalypt hybrids. *South African For J* 167:35–44
- Malan FS, Gerischer GFR (1987) Wood property differences in South African grown *Eucalyptus grandis* trees of different growth stress intensity. *Holzforschung* 41:331–335
- Mhoswa L, O'Neill MM, Mphahlele MM, et al (2020) A Genome-Wide Association Study For Resistance To The Insect Pest *Leptocybe invasa* In *Eucalyptus grandis* Reveals Genomic Regions And Positional Candidate Defence Genes. *Plant Cell Physiol* 61(7):1285-1296
- Miao C, Yang J, Schnable JC (2019) Optimising the identification of causal variants across varying genetic architectures in crops. *Plant Biotechnol J* 17:893–905
- Mphahlele MM, Isik F, Mostert-O'Neill MM, et al (2020) Expected benefits of genomic selection for growth and wood quality traits in *Eucalyptus grandis*. *Tree Genet Genomes* 16:1–12
- Müller BSF, de Almeida Filho JE, Lima BM, et al (2019) Estimating additive and non-additive genetic variances and predicting genetic merits using genome-wide dense single nucleotide polymorphism markers. *New Phytol* 221:235–254
- Müller BSF, Neves LG, de Almeida Filho JE, et al (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
- Munoz F, Rodriguez LS (2014) breedR: Statistical methods for forest genetic resources analysis. In: *Trees for the future: plant material in a changing climate*. pp 13-p
- Oladzad A, Porch T, Rosas JC, et al (2019) Single and multi-trait GWAS identify genetic factors associated with production traits in common bean under abiotic stress environments. *G3 Genes, Genomes, Genet* 9:1881–1892
- Osorio LF, White TL, Huber DA (2001) Age trends of heritabilities and genotype-by-environment interactions for growth traits and wood density from clonal trials of *Eucalyptus grandis* Hill ex Maiden. *Silvae Genet* 50:108–116
- Paaby AB, Rockman M V (2013) The many faces of pleiotropy. *Trends Genet* 29:66–73

- Peterson RA (2021) Finding Optimal Normalizing Transformations via bestNormalize. *R J* 13(1):310-329
- Porter HF, O'Reilly PF (2017) Multivariate simulation framework reveals performance of multi-trait GWAS methods. *Sci Rep* 7:1–12
- Rambolarimanana T, Ramamonjisoa L, Verhaegen D, et al (2018) Performance of multi-trait genomic selection for *Eucalyptus robusta* breeding program. *Tree Genet Genomes* 14:1–13
- Resende RT, Resende MDV, Silva FF, et al (2017) Regional heritability mapping and genome-wide association identify loci for complex growth, wood and disease resistance traits in *Eucalyptus*. *New Phytol* 213:1287–1300
- Rodriguez M, Scintu A, Posadinu CM, et al (2020) GWAS Based on RNA-Seq SNPs and High-Throughput Phenotyping Combined with Climatic Data Highlights the Reservoir of Valuable Genetic Diversity in Regional Tomato Landraces. *Genes (Basel)* 11:1387
- Rubin EM (2008) Genomics of cellulosic biofuels. *Nature* 454:841–845
- Schimleck LR, Kube PD, Raymond CA (2004) Genetic improvement of kraft pulp yield in *Eucalyptus nitens* using cellulose content determined by near infrared spectroscopy. *Can J For Res* 34:2363–2370
- Schumacher FX (1933) Logarithmic expression of timber-tree volume. *J Agric Res* 47:719–734
- Shi Z, Zhang Y, Phillips GO, Yang G (2014) Utilization of bacterial cellulose in food. *Food Hydrocoll* 35:539–545
- Silva-Junior OB, Faria DA, Grattapaglia D (2015) A flexible multi-species genome-wide 60K SNP chip developed from pooled resequencing of 240 *Eucalyptus* tree genomes across 12 species. *New Phytol* 206:1527–1540
- Solovieff N, Cotsapas C, Lee PH, et al (2013) Pleiotropy in complex traits: challenges and strategies. *Nat Rev Genet* 14:483–495
- Stackpole DJ, Vaillancourt RE, Alves A, et al (2011) Genetic variation in the chemical components of *Eucalyptus globulus* wood. *G3 Genes| Genomes| Genet* 1:151–159
- Stackpole DJ, Vaillancourt RE, de Aguilar M, Potts BM (2010) Age trends in genetic parameters for growth and wood density in *Eucalyptus globulus*. *Tree Genet Genomes* 6:179–193
- Tan B, Ingvarsson PK (2018) Multivariate genome-wide association identify loci for complex growth traits by considering additive and over-dominance effects in hybrid *Eucalyptus*. *The Plant Gen* e20208: 1-16

- Tang Y, Liu X, Wang J, et al (2016) GAPIT version 2: an enhanced integrated tool for genomic association and prediction. *Plant Genome* 9:plantgenome2015-11
- Thoen MPM, Davila Olivas NH, Kloth KJ, et al (2017) Genetic architecture of plant stress resistance: multi-trait genome-wide association mapping. *New Phytol* 213:1346–1362
- Valenzuela CE, Ballesta P, Ahmar S, et al (2021) Haplotype-and SNP-Based GWAS for Growth and Wood Quality Traits in *Eucalyptus cladocalyx* Trees under Arid Conditions. *Plants* 10:148
- VanRaden PM (2008) Efficient methods to compute genomic predictions. *J Dairy Sci* 91:4414–4423
- Waples RS, Do CHI (2008) LDNE: a program for estimating effective population size from data on linkage disequilibrium. *Mol Ecol Resour* 8:753–756
- Ward BP, Brown-Guedira G, Kolb FL, et al (2019) Genome-wide association studies for yield-related traits in soft red winter wheat grown in Virginia. *PLoS One* 14:e0208217
- Wickham H (2011) ggplot2. *Wiley Interdiscip Rev Comput Stat* 3:180–185
- Yin L (2018) CMplot: circle manhattan plot. <https://cran.r-project.org/package=CMplot>
- Yoshida GM, Yáñez JM (2021) Multi-trait GWAS using imputed high-density genotypes from whole-genome sequencing identifies genes associated with body traits in Nile tilapia. *BMC Genomics* 22:1–13
- Zheng X, Levine D, Shen J, et al (2012) A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* 28:3326–3328

APPENDIX A - SUPPLEMENTARY MATERIAL FOR CHAPTER 1

Figure S1 - Principal component analysis for genotypic data for 1,772 *Eucalyptus grandis* genotypes from a breeding population located in São Miguel Arcanjo, Brazil, genotyped using the EuChip60 (Silva-Júnior et al. 2015)

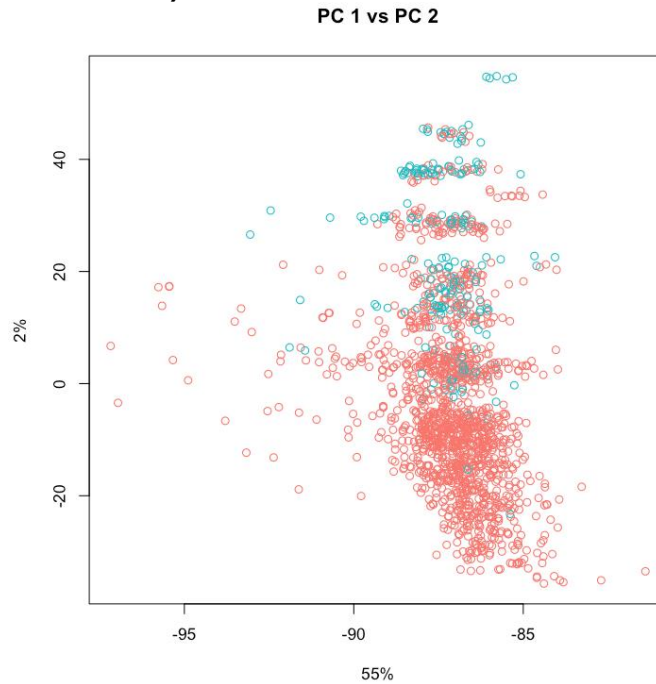


Figure S2. Number of clusters found considering the standard error from the ADMIXTURE analysis for the 1,772 *Eucalyptus grandis* genotypes from a breeding population located in São Miguel Arcanjo, Brazil

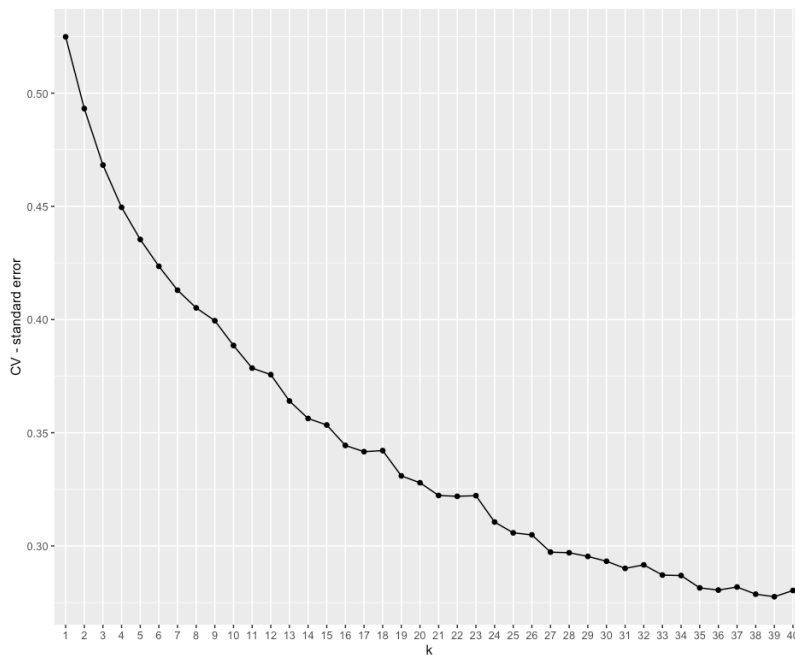
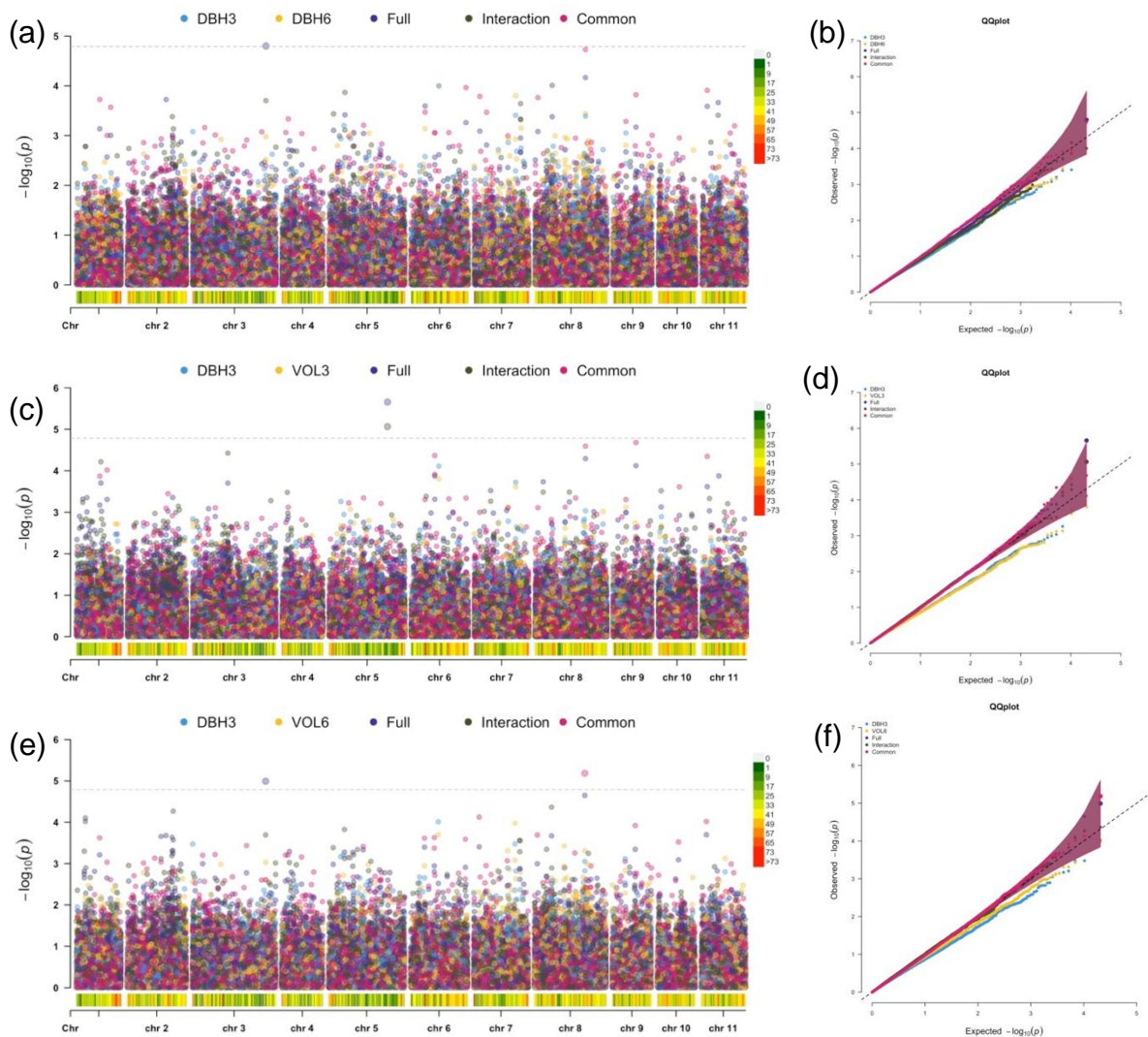
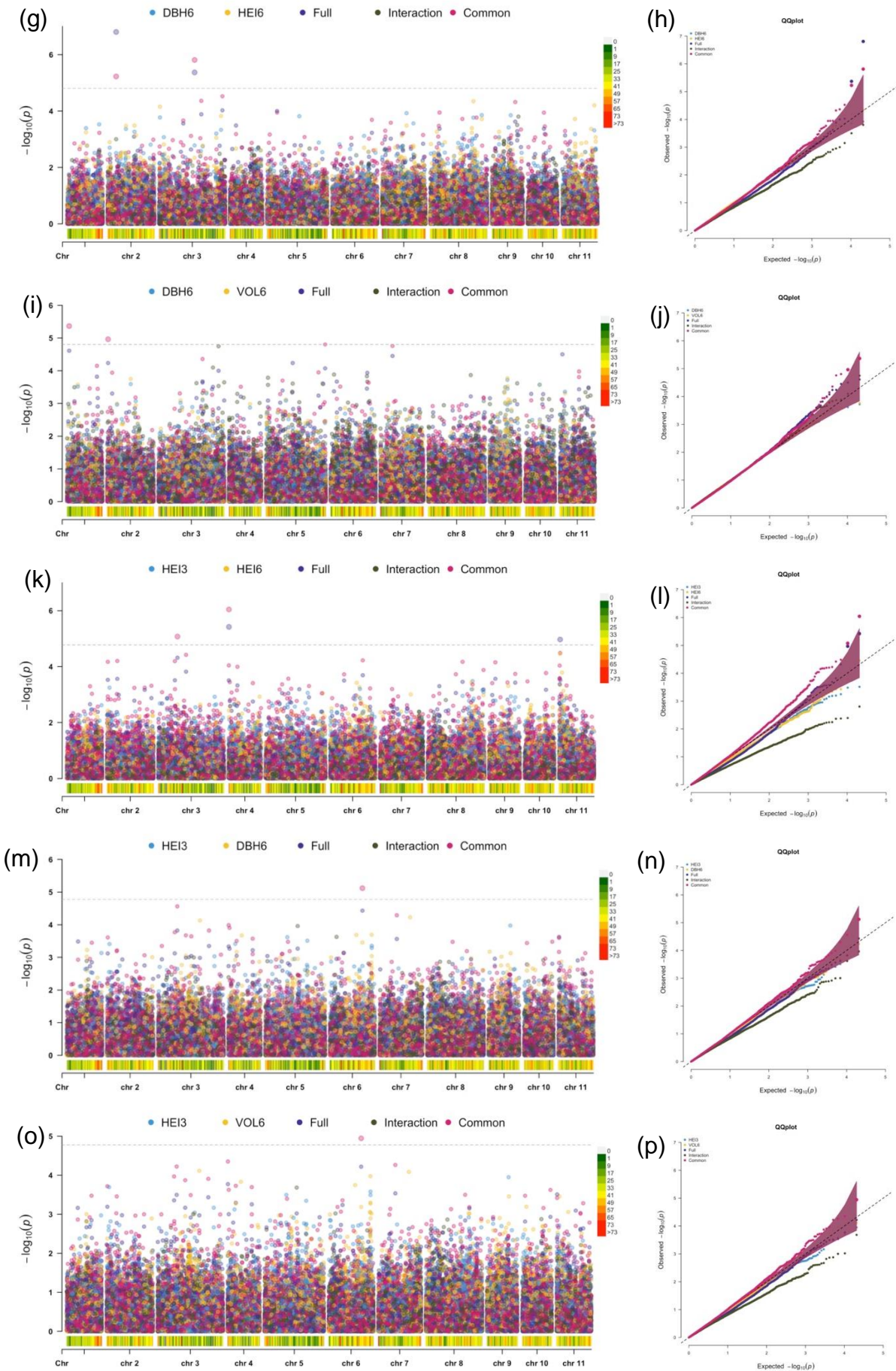


Figure S3. Manhattan and QQ-plots using the multi-trait model (MTMM) between growth traits for the 1,772 *Eucalyptus grandis* individuals genotyped using the *Eucalyptus* chip (EUChip60K). a, b Diameter at breast height at three years and Diameter at breast height at 6 years; c, d Diameter at breast height at three years and volume at three years; e, f Diameter at breast height at three years and volume at six years; g, h Diameter at breast height at six years and height at six years; i, j height at three years and height at six years; k, l; Height at three years and diameter at breast height at six years; m, n Height at three years and volume at six years; o, p Height at six years and volume at six years; q, r Volume at three years and height at six years; s, t Volume at three years and diameter at breast height at six years; u, v Volume at three years and volume at six years.





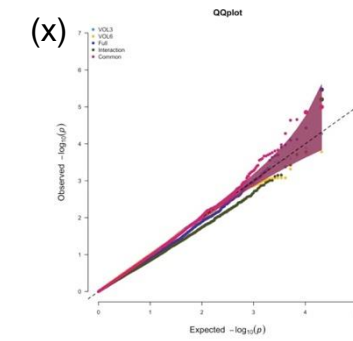
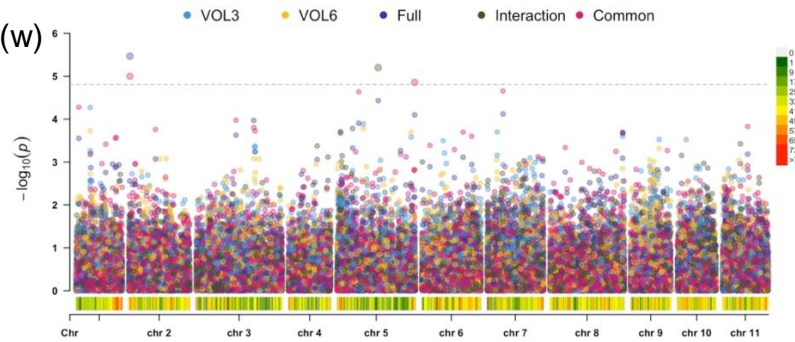
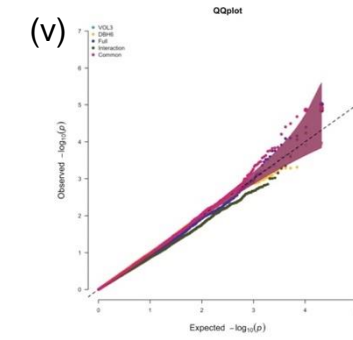
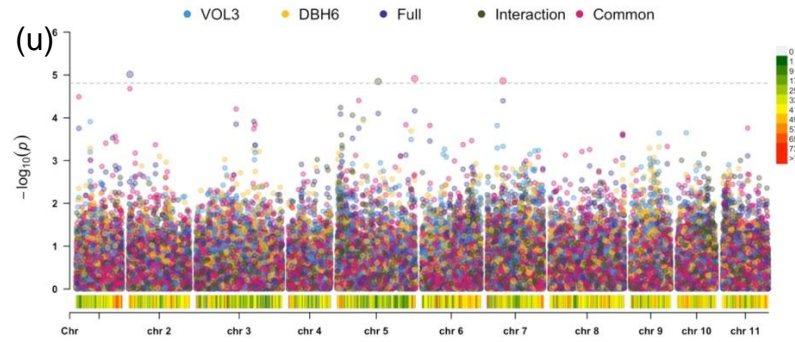
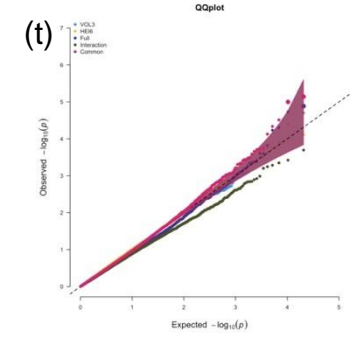
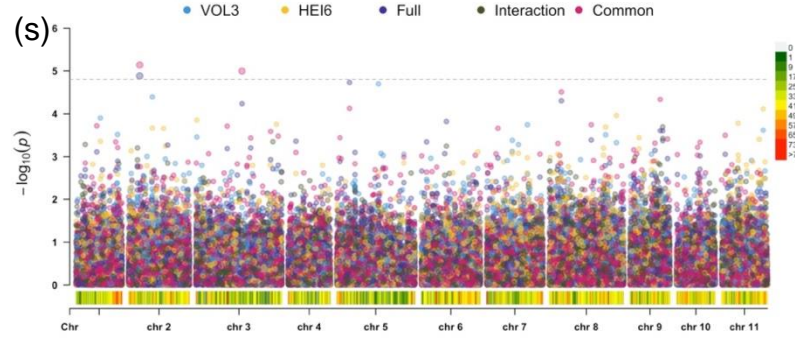
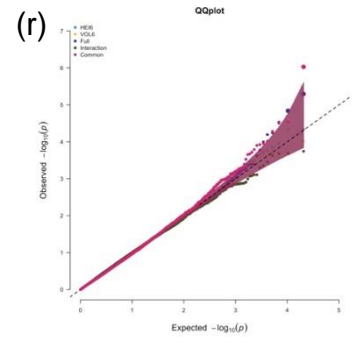
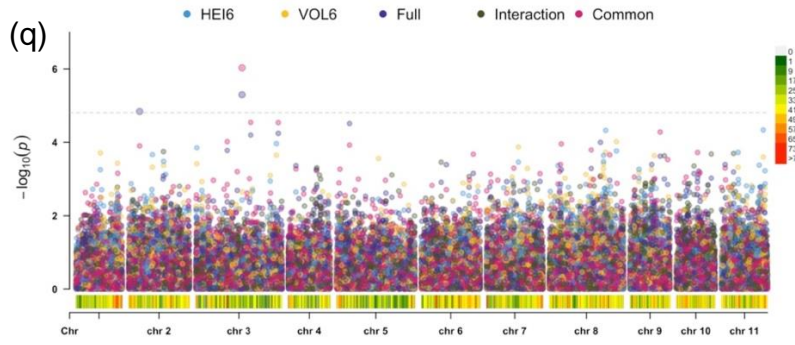


Figure S4. Manhattan and QQ-plots using the multi-trait model (MTMM) between wood-quality traits for the 1,772 *Eucalyptus grandis* individuals genotyped using the *Eucalyptus* chip (EUChip60K). a, b Soluble lignin and total extractives; c, d Syringyl/guaiacyl ratio and total extractives; e, f Syringyl/guaiacyl ratio and soluble lignin.

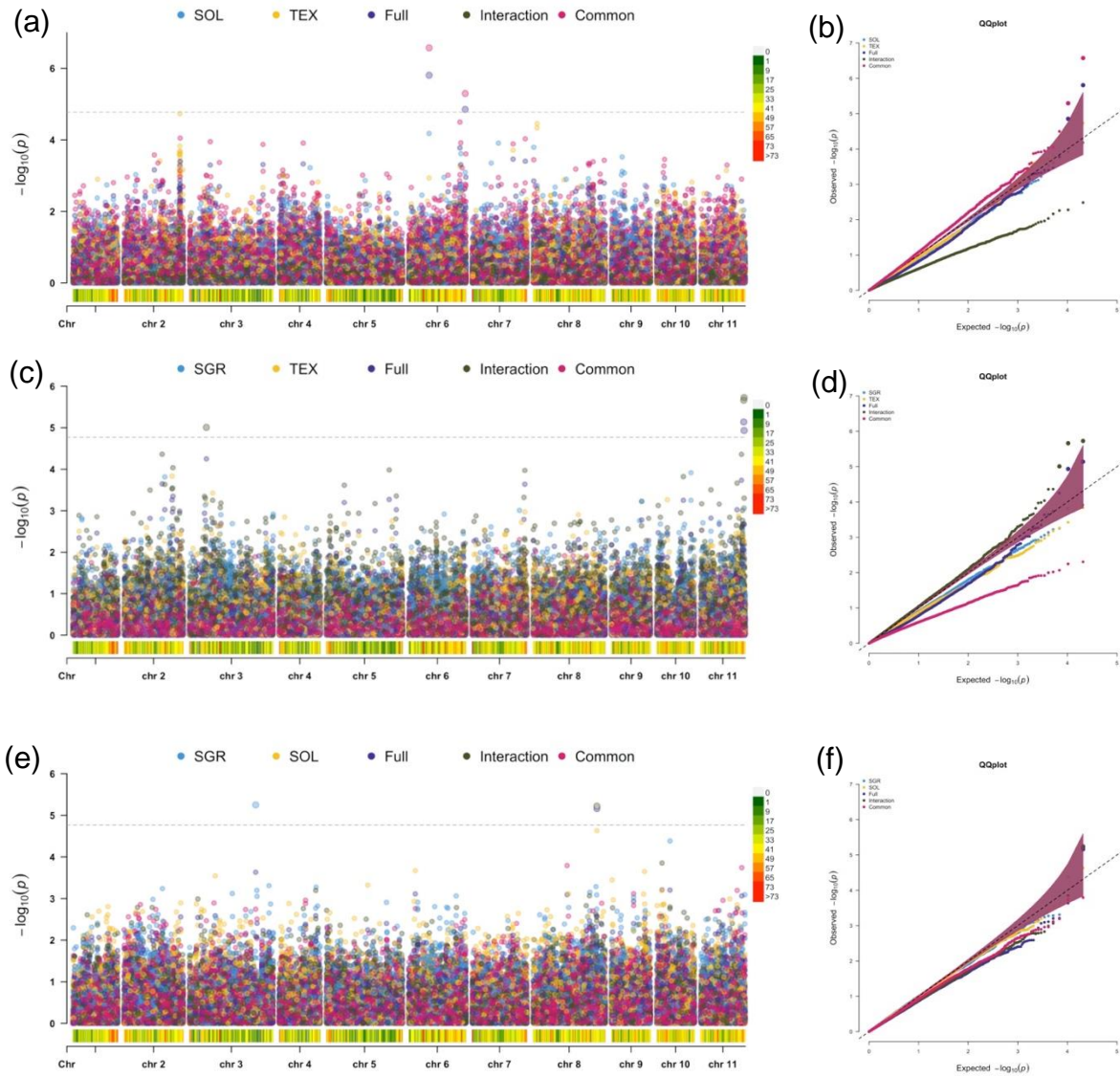
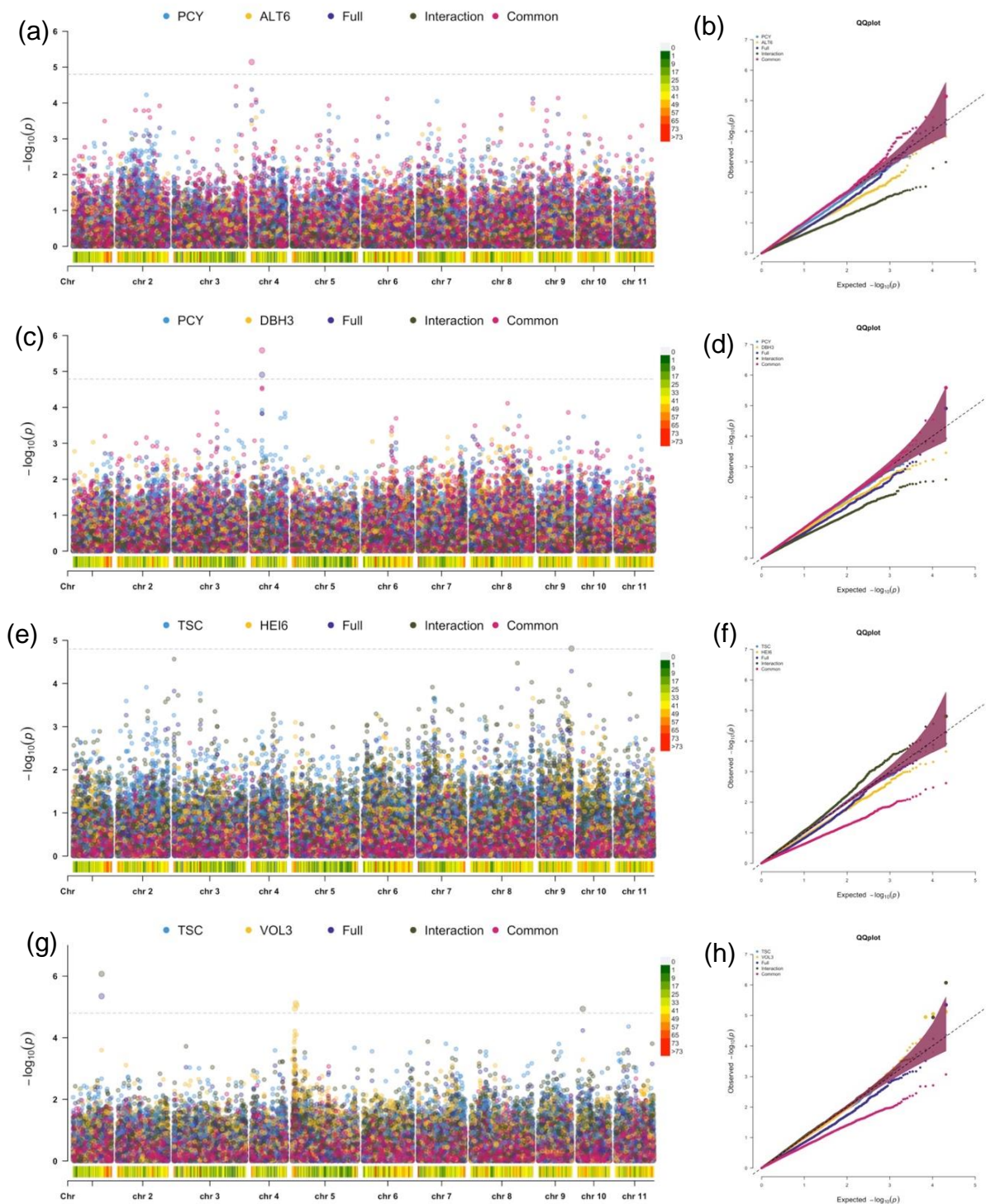


Figure S5. Manhattan and QQ-plots using the multi-trait model (MTMM) between growth traits and wood-quality traits for the 1,772 *Eucalyptus grandis* individuals genotyped using the *Eucalyptus* chip (EUChip60K). a, b Pure cellulose yield and height at six years; c, d Pure cellulose yield and diameter at breast height at three years; e, f, Total solid content and height at six years; g, h Total solid content and volume at three years; i, j Total solid content and diameter at breast height at three years; k, l Total solid content and diameter at breast height at six years; m, n Total solid content and volume at three years; o, p Total solid content and volume at six years.



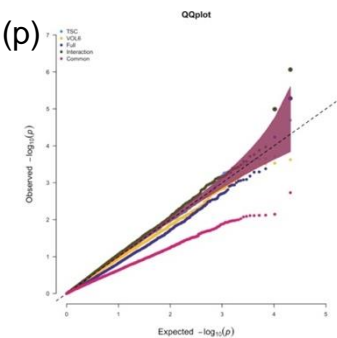
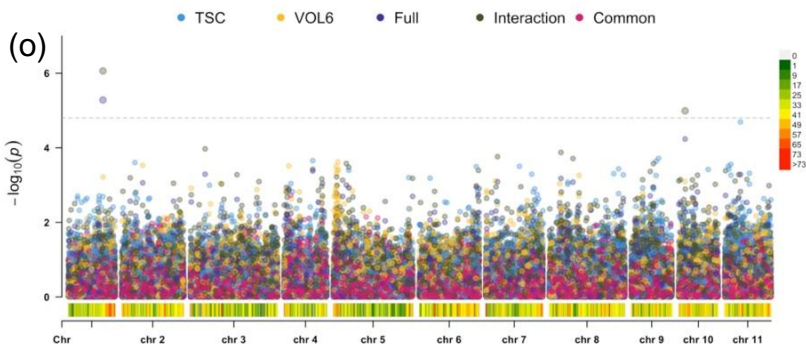
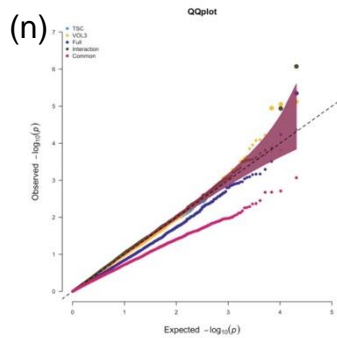
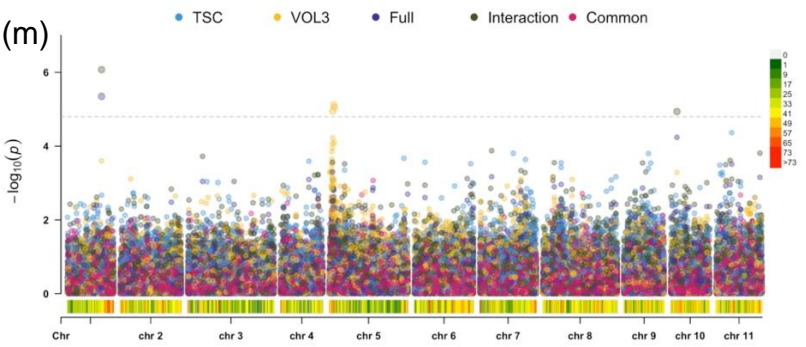
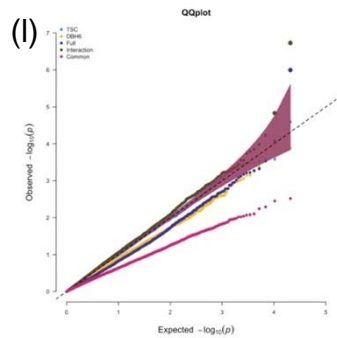
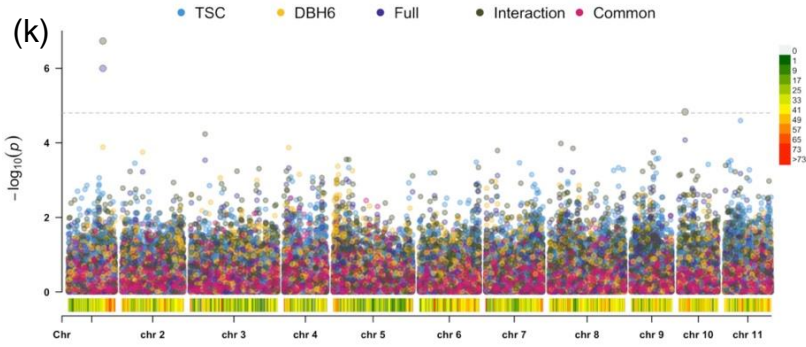
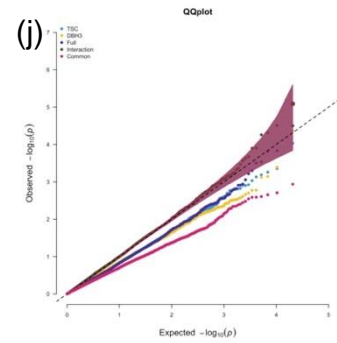
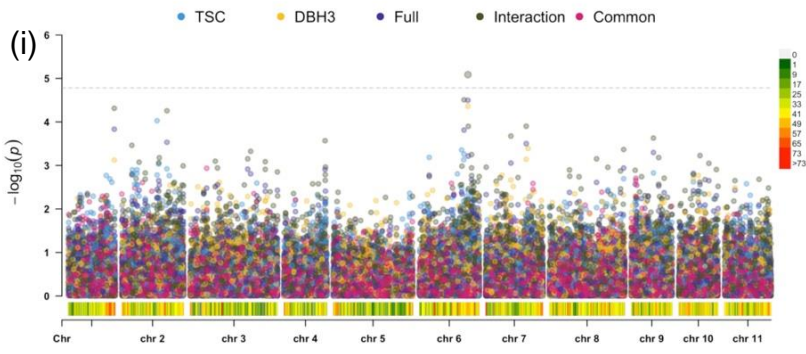


Figure S6. Vem diagram for the number of candidate genes for (a) growth and (b) wood-quality traits in a breeding population of *Eucalyptus grandis* located in São Miguel Arcanjo, Brazil

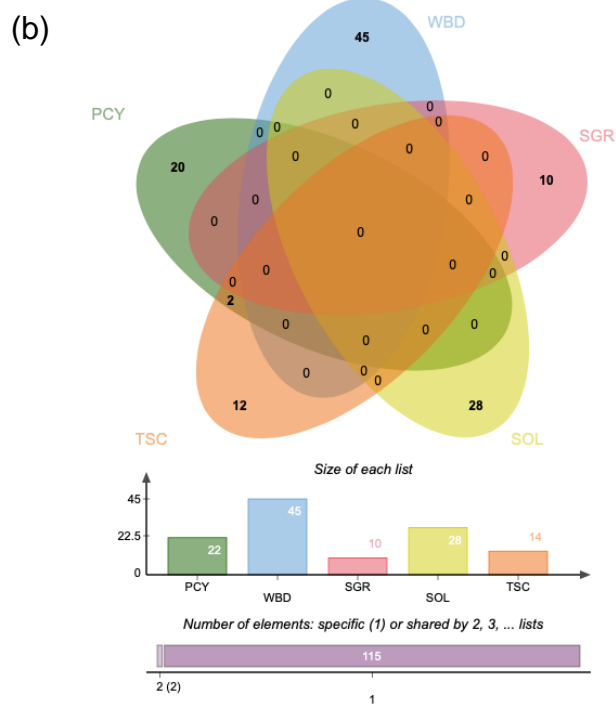
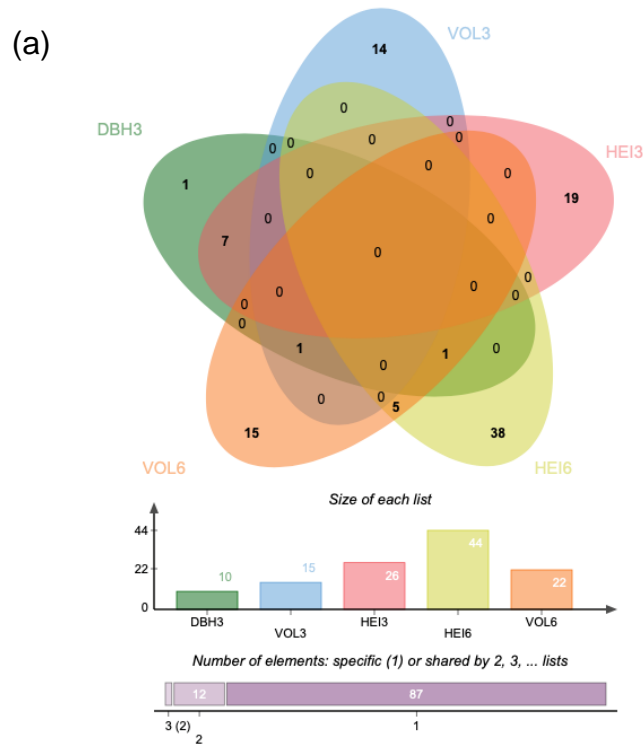


Table S1 - Significant SNP markers and candidate genes for growth traits in *Eucalyptus grandis* using farmCPU model and single-trait analysis

Trait	SNP	Chr	Position	P.value	Gene name	Description
					Eucgr.I01459	(M=4) PF05180 - DNL zinc finger;AT5G27280-NA Zim17-type zinc finger protein nucleus
					Eucgr.I01460	(M=2) K11804 - WD repeat-containing protein 42A;AT4G35140-NA TransducinWD40 repeat-like superfamily protein cytosol
					Eucgr.I01461	(M=1) PTHR24015:SF87 - PENTATRICOPEPTIDE (PPR) REPEAT-CONTAINING PROTEIN;AT2G17033-NA pentatricopeptide (PPR) repeat-containing protein plastid
DBH3	EuBR09s24960947	9	24925225	9.34×10^{-9}	Eucgr.I01462	(M=1) K06640 - serine/threonine-protein kinase ATR;AT5G40820-ATATR,ATR,ATRAD3 Ataxia telangiectasia-mutated and RAD3-related nucleus
					Eucgr.I01463	(M=21) KOG4629 - Predicted mechanosensitive ion channel;AT1G78610-MSL6 mechanosensitive channel of small conductance-like 6 cytosol
					Eucgr.I01464	(M=7) PTHR23257:SF82 - PROTEIN KINASE ATMRK1;AT4G38470-NA ACT-like protein tyrosine kinase family protein cytosol
					Eucgr.I01465	NA
DBH3	EuBR06s23565060	6	24478814	8.44×10^{-8}	Eucgr.F01827	NA
DBH3	EuBR07s6053942	7	5644975	2.55×10^{-7}	NA	NA
DBH3	EuBR11s23336060	11	24494998	5.75×10^{-7}	Eucgr.K01867	(M=3) K11594 - ATP-dependent RNA helicase DDX3X;AT2G42520-NA P-loop containing nucleoside triphosphate hydrolases superfamily protein nucleus;AT2G42520-NA P-loop containing nucleoside triphosphate hydrolases superfamily protein nucleus
					Eucgr.K01867	;AT2G42520-NA P-loop containing nucleoside triphosphate hydrolases superfamily protein nucleus
DBH3	EuBR11s19104457	11	20793674	1.23×10^{-6}	Eucgr.K01579	(M=1) K01142 - exodeoxyribonuclease III;AT2G41460-ARP apurinic endonuclease-redox protein nucleus
					Eucgr.I01459	(M=4) PF05180 - DNL zinc finger;AT5G27280-NA Zim17-type zinc finger protein nucleus
					Eucgr.I01460	(M=2) K11804 - WD repeat-containing protein 42A;AT4G35140-NA TransducinWD40 repeat-like superfamily protein cytosol
					Eucgr.I01461	
HEI3	EuBR09s24960947	9	24925225	1.91×10^{-9}	Eucgr.I01462	(M=1) K06640 - serine/threonine-protein kinase ATR;AT5G40820-ATATR,ATR,ATRAD3 Ataxia telangiectasia-mutated and RAD3-related nucleus
					Eucgr.I01463	(M=21) KOG4629 - Predicted mechanosensitive ion channel;AT1G78610-MSL6 mechanosensitive channel of small conductance-like 6 cytosol
					Eucgr.I01464	(M=7) PTHR23257:SF82 - PROTEIN KINASE ATMRK1;AT4G38470-NA ACT-like protein tyrosine kinase family protein cytosol
					Eucgr.I01465	NA
HEI3	EuBR07s925067	7	925007	1.39×10^{-8}	NA	NA
HEI3	EuBR06s38139098	6	40637919	1.44×10^{-7}	NA	NA
HEI3	EuBR03s39418342	3	40444329	1.52×10^{-7}	Eucgr.C02176	(M=2) PTHR10516:SF15 - PEPTIDYLPROLYL ISOMERASE;AT2G43560-NA FKBP-like peptidyl-prolyl cis-trans isomerase family protein plastid;AT2G43560-NA FKBP-like peptidyl-prolyl cis-trans isomerase family protein plastid
					Eucgr.C02176	;AT2G43560-NA FKBP-like peptidyl-prolyl cis-trans isomerase family protein plastid
					Eucgr.C02176	;AT2G43560-NA FKBP-like peptidyl-prolyl cis-trans isomerase family protein plastid
					Eucgr.C02177	(M=1) K01961 - acetyl-CoA carboxylase, biotin carboxylase subunit;AT5G35360-CAC2 acetyl Co-enzyme a carboxylase biotin carboxylase subunit plastid;AT5G35360-CAC2 acetyl Co-enzyme a carboxylase biotin carboxylase subunit plastid

Continue....

Trait	SNP	Chr	Position	P.value	Gene name	Description
HEI3	EuBR03s39418342	3	40444329	1.52×10^{-7}	Eucgr.C02177	;AT5G35360-CAC2 acetyl Co-enzyme a carboxylase biotin carboxylase subunit plastid
					Eucgr.F04077	(M=281) PTHR24420//PTHR24420:SF474 - LEUCINE-RICH REPEAT RECEPTOR-LIKE PROTEIN KINASE // SUBFAMILY NOT NAMED;AT3G47570-NA Leucine-rich repeat protein kinase family protein plasma membrane
HEI3	EuBR06s49043638	6	51542500	1.60×10^{-7}	Eucgr.F04078	;AT1G33500-NA NA nucleus
					Eucgr.F04079	;AT1G33500-NA NA nucleus
					Eucgr.F04080	NA
					Eucgr.K00803	(M=172) PTHR10641 - MYB-LIKE DNA-BINDING PROTEIN MYB;AT2G46410-CPC Homeodomain-like superfamily protein nucleus
					Eucgr.K00804	(M=4) KOG1770 - Translation initiation factor 1 (eIF-1/SUI1);AT4G27130-NA Translation initiation factor SUI1 family protein cytosol;AT4G27130-NA Translation initiation factor SUI1 family protein cytosol
HEI3	EuBR11s9232784	11	9489911	1.66×10^{-7}	Eucgr.K00804	;AT4G27130-NA Translation initiation factor SUI1 family protein cytosol
					Eucgr.K00805	NA
					Eucgr.K00806	(M=92) 1.11.1.7 - Peroxidase.;AT5G15180-NA Peroxidase superfamily protein extracellular
					Eucgr.K00808	(M=92) 1.11.1.7 - Peroxidase.;AT3G01190-NA Peroxidase superfamily protein extracellular
					Eucgr.K00809	(M=92) 1.11.1.7 - Peroxidase.;AT1G05260-RCI3,RCI3A Peroxidase superfamily protein extracellular
					Eucgr.K00810	(M=3) PTHR21530 - PHEROMONE SHUTDOWN PROTEIN;AT1G05270-NA TraB family protein cytosol;AT1G05270-NA TraB family protein cytosol
					Eucgr.K00810	;AT1G05270-NA TraB family protein cytosol
HEI3	EuBR08s70063929	8	68658658	1.78×10^{-7}	NA	NA
HEI3	EuBR08s71999611	7	36903328	9.50×10^{-7}	Eucgr.H05019	;AT5G24280-GMI1 gamma-irradiation and mitomycin c induced 1 nucleus
					Eucgr.B01857	(M=7) PTHR23155//PTHR23155:SF304 - LEUCINE-RICH REPEAT-CONTAINING PROTEIN;AT1G69550-NA disease resistance protein (TIR-NBS-LRR class) plasma membrane
					Eucgr.B01858	NA
HEI3	EuBR02s34899535	2	30476873	1.37×10^{-6}	Eucgr.B01859	(M=214) PF00646 - F-box domain;AT1G55000-NA peptidoglycan-binding LysM domain-containing protein plasma membrane
					Eucgr.B01863	(M=7) PTHR23155//PTHR23155:SF304 - LEUCINE-RICH REPEAT-CONTAINING PROTEIN;AT1G69550-NA disease resistance protein (TIR-NBS-LRR class) plasma membrane
					Eucgr.B01864	NA
					Eucgr.C02137	(M=6) PTHR24420//PTHR24420:SF203 - LEUCINE-RICH REPEAT RECEPTOR-LIKE PROTEIN KINASE // SUBFAMILY NOT NAMED;AT2G24370-NA Protein kinase protein with adenine nucleotide alpha hydrolases-like domain nucleus
VOL3	EuBR03s38709900	3	39735887	1.65×10^{-10}	Eucgr.C02138	(M=29) PF03080 - Domain of unknown function (DUF239);AT2G44220-NA Protein of Unknown Function (DUF239) extracellular
					Eucgr.K02808	;AT5G63100-NA S-adenosyl-L-methionine-dependent methyltransferases superfamily protein mitochondrion
					Eucgr.K02809	(M=37) PF03195 - Protein of unknown function DUF260;AT5G63090-LOB Lateral organ boundaries (LOB) domain family protein nucleus
					Eucgr.K02810	;AT1G21280-NA NA nucleus
VOL3	EuBR11s36177647	11	35503789	1.01×10^{-8}	Eucgr.K02812	(M=5) PF07223 - Protein of unknown function (DUF1421);AT5G14540-NA Protein of unknown function (DUF1421) nucleus
					Eucgr.K02813	(M=6) 2.4.1.82 - Galactinol--sucrose galactosyltransferase.;AT5G40390-SIP1 Raffinose synthase family protein cytosol
					Eucgr.K02814	(M=1) PTHR13948//PTHR13948:SF20 - RNA-BINDING PROTEIN;AT4G34140-NA D111/G-patch domain-containing protein nucleus;AT4G34140-NA D111/G-patch domain-containing protein nucleus

Continue...

Trait	SNP	Chr	Position	P.value	Gene name	Description
					Eucgr.K02814	;AT4G34140-NA D111/G-patch domain-containing protein nucleus
					Eucgr.K02814	;AT4G34140-NA D111/G-patch domain-containing protein nucleus
					Eucgr.K02815	(M=29) KOG1208 - Dehydrogenases with different specificities (related to short-chain alcohol dehydrogenases);;AT5G50130-NA NAD(P)-binding Rossmann-fold superfamily protein mitochondrion
					Eucgr.K02816	NA
					Eucgr.K02817	(M=1) K11550 - kinetochore protein Spc25, animal type;;AT3G48210-NA NA nucleus;AT3G48210-NA NA nucleus
					Eucgr.K02817	;AT3G48210-NA NA nucleus
					Eucgr.K02818	(M=29) PF03080 - Domain of unknown function (DUF239);;AT5G50150-NA Protein of Unknown Function (DUF239) extracellular
VOL3	EuBR06s23565060	6	24478814	8.29×10^{-8}	Eucgr.F01827	NA
VOL3	EuBR04s3954175	4	4182167	8.47×10^{-7}	NA	NA
VOL3	EuBR07s6053942	7	5644975	1.12×10^{-6}	NA	NA
					Eucgr.C00825	(M=1) PTHR12465 - UBIQUITIN SPECIFIC PROTEASE HOMOLOG 49;AT2G28230-NA TATA-binding related factor (TRF) of subunit 20 of Mediator complex cytosol;AT2G28230-NA TATA-binding related factor (TRF) of subunit 20 of Mediator complex cytosol
VOL3	EuBR03s13584769	3	13747344	1.32×10^{-6}	Eucgr.C00825	;AT2G28230-NA TATA-binding related factor (TRF) of subunit 20 of Mediator complex cytosol
					Eucgr.C00825	;AT2G28230-NA TATA-binding related factor (TRF) of subunit 20 of Mediator complex cytosol
					Eucgr.C00825	;AT2G28230-NA TATA-binding related factor (TRF) of subunit 20 of Mediator complex cytosol
					Eucgr.C00825	;AT2G28230-NA TATA-binding related factor (TRF) of subunit 20 of Mediator complex cytosol
					Eucgr.C00826	(M=172) K09422 - myb proto-oncogene protein, plant;AT5G49330-ATMYB111,MYB111,PGF3 myb domain protein 111 nucleus
DBH6	EuBR02s22040507	2	26814555	1.35×10^{-11}	NA	NA
DBH6	EuBR03s23262715	3	22481832	8.44×10^{-8}	NA	NA
HEI6	EuBR02s22040507	2	26814555	1.67×10^{-10}	NA	NA
HEI6	EuBR11s19104457	11	20793674	6.58×10^{-9}	Eucgr.K01579	(M=1) K01142 - exodeoxyribonuclease III;AT2G41460-ARP apurinic endonuclease-redox protein nucleus
HEI6	EuBR03s36174125	3	37200112	1.30×10^{-8}	Eucgr.C02023	(M=3) PF01424 - R3H domain
					Eucgr.C02023	NA
					Eucgr.K01383	(M=85) PF03514 - GRAS domain family;AT2G37650-NA GRAS family transcription factor nucleus
					Eucgr.K01384	(M=85) PF03514 - GRAS domain family;AT2G37650-NA GRAS family transcription factor nucleus
HEI6	EuBR11s17004419	11	17261546	2.50×10^{-8}	Eucgr.K01385	(M=9) PF05910 - Plant protein of unknown function (DUF868);;AT2G04220-NA Plant protein of unknown function (DUF868) multiple
					Eucgr.K01386	;AT3G07640-NA NA cytosol
HEI6	EuBR06s48724759	6	51223621	7.47×10^{-8}	NA	NA
					Eucgr.D00230	(M=31) KOG0054 - Multidrug resistance-associated protein/mitoxantrone resistance protein, ABC superfamily;AT2G34660-ATMRP2,EST4,MRP2 multidrug resistance-associated protein 2 vacuole
					Eucgr.D00231	;AT1G30410-ATMRP13,MRP13 multidrug resistance-associated protein 13 plasma membrane
					Eucgr.D00232	
HEI6	EuBR04s3546467	4	3940694	7.74×10^{-8}	Eucgr.C02697	(M=651) PTHR23155 - LEUCINE-RICH REPEAT-CONTAINING PROTEIN;AT1G69550-NA disease resistance protein (TIR-NBS-LRR class) plasma membrane
					Eucgr.C02699	NA
					Eucgr.C02701	(M=411) PF01582 - TIR domain;AT1G27170-NA transmembrane receptors;ATP binding cytosol
					Eucgr.C02703	NA

Continue...

Trait	SNP	Chr	Position	P.value	Gene name	Description
HEI6	EuBR04s3546467	4	3940694	7.74×10^{-8}	Eucgr.C02704	(M=425) KOG0472 - Leucine-rich repeat protein;AT1G69550-NA disease resistance protein (TIR-NBS-LRR class) plasma membrane
					Eucgr.C02705	(M=425) KOG0472 - Leucine-rich repeat protein;AT1G69550-NA disease resistance protein (TIR-NBS-LRR class) plasma membrane
					Eucgr.C02706	NA
					Eucgr.C02708	(M=651) PTHR23155 - LEUCINE-RICH REPEAT-CONTAINING PROTEIN;AT1G69550-NA disease resistance protein (TIR-NBS-LRR class) plasma membrane
					Eucgr.C02709	(M=1421) PF00560 - Leucine Rich Repeat;AT4G16940-NA Disease resistance protein (TIR-NBS-LRR class) family nucleus
					Eucgr.C02710	NA
					Eucgr.C02711	NA
HEI6	EuBR09s23175321	9	23139599	3.93×10^{-7}	Eucgr.I01229	(M=1) K11876 - proteasome assembly chaperone 2;AT3G18940-NA clast3-related cytosol
					Eucgr.I01230	
					Eucgr.I01231	(M=7) PF05562 - Cold acclimation protein WCOR413;AT2G15970-ATCOR413-PM1,ATCYP19,COR413-PM1,FL3-5A3,WCOR413,WCOR413-LIKE cold regulated 413 plasma membrane 1 plasma membrane
HEI6	EuBR06s49882819	6	52381681	6.93×10^{-7}	Eucgr.F04163	(M=84) PTHR11709:SF9 - LACCASE;AT5G09360-LAC14 laccase 14 extracellular
					Eucgr.F04164	NA
					Eucgr.F04165	;AT1G71780-NA NA cytosol
					Eucgr.F04166	(M=7) KOG0394 - Ras-related GTPase;AT4G09720-ATRABG3A,RABG3A RAB GTPase homolog G3A multiple;AT4G09720-ATRABG3A,RABG3A RAB GTPase homolog G3A multiple
					Eucgr.F04166	;AT4G09720-ATRABG3A,RABG3A RAB GTPase homolog G3A multiple
					Eucgr.F04168	(M=15) PF00280 - Potato inhibitor I family;AT3G46860-NA Serine protease inhibitor, potato inhibitor I-type family protein extracellular
					Eucgr.F04169	(M=29) KOG1208 - Dehydrogenases with different specificities (related to short-chain alcohol dehydrogenases);AT4G09750-NA NAD(P)-binding Rossmann-fold superfamily protein mitochondrion
HEI6	EuBR02s27050204	2	21804858	7.44×10^{-7}	Eucgr.F04170	(M=17) K13460 - disease resistance protein RPS5;AT1G12220-RPS5 Disease resistance protein (CC-NBS-LRR class) family cytosol
					Eucgr.F04171	(M=29) KOG1208 - Dehydrogenases with different specificities (related to short-chain alcohol dehydrogenases);AT4G09750-NA NAD(P)-binding Rossmann-fold superfamily protein mitochondrion
					Eucgr.F04172	(M=17) K13460 - disease resistance protein RPS5;AT1G12220-RPS5 Disease resistance protein (CC-NBS-LRR class) family cytosol
HEI6	EuBR05s22005045	5	20942275	7.63×10^{-7}	NA	NA
HEI6	EuBR05s23402100	5	22339330	1.07×10^{-6}	Eucgr.E01694	(M=204) KOG0157 - Cytochrome P450 CYP4/CYP19/CYP26 subfamilies;AT2G26710-BAS1,CYP72B1,CYP734A1 Cytochrome P450 superfamily protein endoplasmic reticulum
					NA	NA
					Eucgr.K01308	(M=2) PTHR10891:SF43 - CALCIUM-BINDING PROTEIN;AT4G20780-CML42 calmodulin like 42 multiple
					Eucgr.K01309	;AT2G28330-NA NA nucleus
					Eucgr.K01310	(M=88) PF03168 - Late embryogenesis abundant protein;AT3G05975-NA Late embryogenesis abundant (LEA) hydroxyproline-rich glycoprotein family plasma membrane
Eucgr.K01311	(M=88) PF03168 - Late embryogenesis abundant protein;AT3G54200-NA Late embryogenesis abundant (LEA) hydroxyproline-rich glycoprotein family plasma membrane					
Eucgr.K01312	(M=88) PF03168 - Late embryogenesis abundant protein;AT3G54200-NA Late embryogenesis abundant (LEA) hydroxyproline-rich glycoprotein family plasma membrane					

Continue...

Trait	SNP	Chr	Position	P.value	Gene name	Description
					Eucgr.K01312	(M=88) PF03168 - Late embryogenesis abundant protein;AT3G54200-NA Late embryogenesis abundant (LEA) hydroxyproline-rich glycoprotein family plasma membrane
					Eucgr.K01313	(M=88) PF03168 - Late embryogenesis abundant protein;AT3G54200-NA Late embryogenesis abundant (LEA) hydroxyproline-rich glycoprotein family plasma membrane
HEI6	EuBR05s23402100	5	22339330	1.07×10^{-6}	Eucgr.K01314	(M=3) KOG4513 - Phosphoglycerate mutase;AT3G08590-NA Phosphoglycerate mutase, 2,3-bisphosphoglycerate-independent cytosol;AT3G08590-NA Phosphoglycerate mutase, 2,3-bisphosphoglycerate-independent cytosol
					Eucgr.K01315	;AT5G01015-NA NA extracellular
					Eucgr.K01316	(M=13) PF04525 - Tubby C 2;AT3G15810-NA Protein of unknown function (DUF567) cytosol
					Eucgr.K01317	(M=2) 2.7.7.14 - Ethanolamine-phosphate cytidyltransferase.;AT2G38670-PECT1 phosphorylethanolamine cytidyltransferase 1 mitochondrion
					Eucgr.K01318	(M=1) PTHR23324//PTHR23324:SF29 - SEC14 RELATED PROTEIN // SUBFAMILY NOT NAMED;AT5G01010-NA NA cytosol
VOL6	EuBR02s22040507	2	26814555	1.53×10^{-9}	NA	NA
					Eucgr.G00894	;AT3G08880-NA NA nucleus
VOL6	EuBR07s15774740	7	14235600	4.68×10^{-9}	Eucgr.G00895	(M=35) KOG1493 - Anaphase-promoting complex (APC), subunit 11;AT1G72220-NA RING/U-box superfamily protein nucleus
VOL6	EuBR06s23565060	6	24478814	1.01×10^{-8}	Eucgr.F01827	NA
VOL6	EuBR11s19104457	11	20793674	1.80×10^{-7}	Eucgr.K01579	(M=1) K01142 - exodeoxyribonuclease III;AT2G41460-ARP apurinic endonuclease-redox protein nucleus
					Eucgr.K01383	(M=85) PF03514 - GRAS domain family;AT2G37650-NA GRAS family transcription factor nucleus
					Eucgr.K01384	(M=85) PF03514 - GRAS domain family;AT2G37650-NA GRAS family transcription factor nucleus
VOL6	EuBR11s17004419	11	17261546	5.68×10^{-7}	Eucgr.K01385	(M=9) PF05910 - Plant protein of unknown function (DUF868);AT2G04220-NA Plant protein of unknown function (DUF868) multiple
					Eucgr.K01386	;AT3G07640-NA NA cytosol
					Eucgr.G01047	(M=164) PF02365 - No apical meristem (NAM) protein;AT1G33060-ANAC014,NAC014 NAC 014 nucleus;AT1G33060-ANAC014,NAC014 NAC 014 nucleus
					Eucgr.G01047	;AT1G33060-ANAC014,NAC014 NAC 014 nucleus
					Eucgr.G01048	NA
VOL6	EuBR07s18186900	7	16413125	6.00×10^{-7}	Eucgr.G01049	(M=164) PF02365 - No apical meristem (NAM) protein;AT4G35580-NTL9 NAC transcription factor-like 9 nucleus
					Eucgr.G01050	NA
					Eucgr.G01051	NA
					Eucgr.G01051	NA
VOL6	EuBR07s10484525	7	8835665	1.90×10^{-6}	Eucgr.G00596	(M=12) K08511 - vesicle-associated membrane protein 72;AT1G04760-ATVAMP726,VAMP726 vesicle-associated membrane protein 726 plasma membrane
					Eucgr.G00598	NA
					Eucgr.G00599	(M=12) K08511 - vesicle-associated membrane protein 72;AT2G32670-ATVAMP725,VAMP725 vesicle-associated membrane protein 725 plasma membrane
					Eucgr.G00600	(M=281) PTHR24420//PTHR24420:SF474 - LEUCINE-RICH REPEAT RECEPTOR-LIKE PROTEIN KINASE // SUBFAMILY NOT NAMED;AT3G47570-NA Leucine-rich repeat protein kinase family protein plasma membrane
					Eucgr.G00600	NA
					Eucgr.G00601	(M=12) K08511 - vesicle-associated membrane protein 72;AT1G04760-ATVAMP726,VAMP726 vesicle-associated membrane protein 726 plasma membrane

Continue...

Trait	SNP	Chr	Position	P.value	Gene name	Description
					Eucgr.G00601	;AT1G04760-ATVAMP726,VAMP726 vesicle-associated membrane protein 726 plasma membrane
					Eucgr.G00602	(M=12) K08511 - vesicle-associated membrane protein 72;AT2G32670-ATVAMP725,VAMP725 vesicle-associated membrane protein 725 plasma membrane
VOL6	EuBR03s36174125	3	37200112	1.90×10^{-6}	Eucgr.C02023	(M=3) PF01424 - R3H domain
					Eucgr.C02023	NA
VOL6	EuBR10s3594677	10	3594617	2.03×10^{-6}	Eucgr.J00366	(M=180) PF02458 - Transferase family;AT5G41040-NA HXXXD-type acyl-transferase family protein cytosol
					Eucgr.J00367	(M=4) PTHR10030 - ALPHA-L-FUCOSIDASE;AT2G28100-ATFUC1,FUC1 alpha-L-fucosidase 1 extracellular

Table S2 - Significant SNP markers and candidate genes using the single trait analysis for wood quality traits in a *Eucalyptus grandis* breeding population located in São Miguel Arcanjo, São Paulo, Brazil.

Trait	SNP	Chr	Position	P.value	Gene name	Description
PCY	EuBR04s9558885	4	9786877	1.06×10^{-11}	Eucgr.D00522	(M=28) 3.1.1.3 - Triacylglycerol lipase.;AT4G18550-NA alpha/beta-Hydrolases superfamily protein cytosol
					Eucgr.D00523	(M=123) PTHR22835:SF39 - LATERAL SIGNALING TARGET PROTEIN 2;AT5G45910-NA GDSL-like Lipase/Acylhydrolase superfamily protein extracellular
					Eucgr.D00525	(M=123) PTHR22835:SF39 - LATERAL SIGNALING TARGET PROTEIN 2;AT5G45910-NA GDSL-like Lipase/Acylhydrolase superfamily protein extracellular
					Eucgr.D00526	(M=123) PTHR22835:SF39 - LATERAL SIGNALING TARGET PROTEIN 2;AT5G45910-NA GDSL-like Lipase/Acylhydrolase superfamily protein extracellular
					Eucgr.D00527	(M=123) PTHR22835:SF39 - LATERAL SIGNALING TARGET PROTEIN 2;AT5G45910-NA GDSL-like Lipase/Acylhydrolase superfamily protein extracellular
PCY	EuBR08s57640594	8	54276404	6.86×10^{-10}	NA	NA
PCY	EuBR03s79713635	3	82024066	3.17×10^{-9}	NA	NA
PCY	EuBR07s34761317	7	28804454	4.26×10^{-9}	Eucgr.G01887	(M=15) K10760 - adenylate isopentenyltransferase (cytokinin synthase); AT5G19040-ATIPT5,IPT5 isopentenyltransferase 5 mitochondrion
					Eucgr.G01888	(M=2) PF03048 - UL92 family;AT5G45360-NA F-box family protein nucleus
					Eucgr.G01889	;AT2G23690-NA NA cytosol
					Eucgr.G01890	(M=590) PF01535 - PPR repeat; Tetratricopeptide repeat (TPR)-like superfamily protein mitochondrion
					Eucgr.G01891	(M=184) PTHR24420//PTHR24420:SF441 - LEUCINE-RICH REPEAT RECEPTOR-LIKE PROTEIN KINASE // SUBFAMILY NOT NAMED; receptor serine/threonine kinase, putative plasma membrane
					Eucgr.G01892	;AT1G49350-NA pfkB-like carbohydrate kinase family protein peroxisome
					Eucgr.G01893	(M=1) K12670 - oligosaccharyltransferase complex subunit beta; AT5G66680-DGL1 dolichyl-diphosphooligosaccharide-protein glycosyltransferase 48k Da subunit family protein endoplasmic reticulum
					Eucgr.G01894	(M=11) KOG0324 - Uncharacterized conserved protein;AT1G47740-NA PPPDE putative thiol peptidase family protein nucleus
					Eucgr.G01894	;AT1G47740-NA PPPDE putative thiol peptidase family protein nucleus
					Eucgr.G01895	NA
					Eucgr.G01896	(M=590) PF01535 - PPR repeat;AT3G18840-NA Tetratricopeptide repeat (TPR)-like superfamily protein mitochondrion
					Eucgr.G01897	(M=184) PTHR24420//PTHR24420:SF441 - LEUCINE-RICH REPEAT RECEPTOR-LIKE PROTEIN KINASE // SUBFAMILY NOT NAMED; AT4G18250-NA receptor serine/threonine kinase, putative plasma membrane
					Eucgr.G01898	(M=184) PTHR24420//PTHR24420:SF441 - LEUCINE-RICH REPEAT RECEPTOR-LIKE PROTEIN KINASE // SUBFAMILY NOT NAMED; receptor serine/threonine kinase, putative plasma membrane
Eucgr.G01899	(M=31) PF05055 - Protein of unknown function (DUF677);AT4G34320-NA Protein of unknown function (DUF677) cytosol					
PCY	EuBR03s47117919	3	50465764	3.58×10^{-8}	NA	NA
PCY	EuBR10s1696823	10	1696763	3.70×10^{-8}	Eucgr.J00155	(M=7) PF07107 - Wound-induced protein WI12;AT3G10985-ATWI-12,SAG20,WI12 senescence associated gene 20 plasma membrane
PCY	EuBR07s252985	7	252925	2.89×10^{-7}	Eucgr.G00025	(M=11) PTHR13943 - HRAS-LIKE SUPPRESSOR - RELATED;AT3G02700-NA NC domain-containing protein-related multiple

Continue...

Trait	SNP	Chr	Position	P.value	Gene name	Description
PCY	EuBR10s20170300	10	19395285	3.93×10^{-7}	Eucgr.J01562	(M=2) K02958 - small subunit ribosomal protein S15e;AT5G09500-NA Ribosomal protein S19 family protein cytosol
					Eucgr.J01563	;AT5G43650-BHLH92 basic helix-loop-helix (bHLH) DNA-binding superfamily protein nucleus
WBD	EuBR08s41044396	8	39604466	4.21×10^{-10}	Eucgr.H02844	(M=18) K13428 - LRR receptor-like serine/threonine-protein kinase EFR;AT3G47090-NA Leucine-rich repeat protein kinase family protein plasma membrane
					Eucgr.H02845	;AT4G13690-NA NA mitochondrion
WBD	EuBR02s28457850	2	28158241	1.30×10^{-8}	Eucgr.B01645	(M=651) PTHR23155 - LEUCINE-RICH REPEAT-CONTAINING PROTEIN;AT1G69550-NA disease resistance protein (TIR-NBS-LRR class) plasma membrane
WBD	EuBR02s47319408	2	45237800	1.49×10^{-8}	Eucgr.B02518	(M=1) KOG0383 - Predicted helicase;AT1G08060-MOM,MOM1 ATP-dependent helicase family protein nucleus;AT1G08060-MOM,MOM1 ATP-dependent helicase family protein nucleus
					Eucgr.B02518	;AT1G08060-MOM,MOM1 ATP-dependent helicase family protein nucleus
					Eucgr.B02519	(M=6) PTHR24420//PTHR24420:SF481 - LEUCINE-RICH REPEAT RECEPTOR-LIKE PROTEIN KINASE // SUBFAMILY NOT NAMED;AT2G23770-NA protein kinase family protein / peptidoglycan-binding LysM domain-containing protein plasma membrane
					Eucgr.B02520	(M=14) K01674 - carbonic anhydrase;AT1G08080-ACA7,ATACA7 alpha carbonic anhydrase 7 extracellular
					Eucgr.B02521	NA
WBD	EuBR08s73877790	8	70566605	1.08×10^{-7}	Eucgr.H05143	;AT1G48720-NA NA cytosol
					Eucgr.H05144	(M=157) KOG4412 - 26S proteasome regulatory complex, subunit PSMD10;AT4G10720-NA Ankyrin repeat family protein multiple
					Eucgr.H05145	;AT4G10720-NA Ankyrin repeat family protein multiple
					Eucgr.H05146	(M=590) PF01535 - PPR repeat;AT1G52640-NA Pentatricopeptide repeat (PPR) superfamily protein mitochondrion
					Eucgr.H05147	(M=15) PF00280 - Potato inhibitor I family;AT3G46860-NA Serine protease inhibitor, potato inhibitor I-type family protein extracellular
WBD	EuBR03s9730959	3	8701931	1.72×10^{-7}	Eucgr.C00545	(M=9) K11159 - carotenoid cleavage dioxygenase;AT3G63520-ATCCD1,ATNCED1,CCD1,NCED1 carotenoid cleavage dioxygenase 1 cytosol
WBD	EuBR01s35497723	1	40180587	3.37×10^{-7}	Eucgr.A02484	(M=11) PTHR21495//PTHR21495:SF10 - NUCLEOPORIN-RELATED // SUBFAMILY NOT NAMED;AT1G58170-NA Disease resistance-responsive (dirigent-like protein) family protein extracellular
					Eucgr.A02485	(M=51) PTHR21495 - NUCLEOPORIN-RELATED;AT1G65870-NA Disease resistance-responsive (dirigent-like protein) family protein extracellular
					Eucgr.A02486	(M=51) PTHR21495 - NUCLEOPORIN-RELATED;AT1G65870-NA Disease resistance-responsive (dirigent-like protein) family protein extracellular
					Eucgr.A02487	(M=51) PTHR21495 - NUCLEOPORIN-RELATED;AT1G65870-NA Disease resistance-responsive (dirigent-like protein) family protein extracellular
WBD	EuBR06s23537926	6	24451680	5.75×10^{-7}	NA	NA
WBD	EuBR07s44074670	7	46392648	7.29×10^{-7}	Eucgr.G02554	(M=8) PTHR11654:SF21 - NITRATE TRANSPORTER (NTL1);AT1G69850-ATNRT1:2,NRT1:2,NTL1 nitrate transporter 1:2 plasma membrane
					Eucgr.G02555	NA
					Eucgr.G02556	NA
					Eucgr.G02557	(M=1) PTHR11224//PTHR11224:SF27 - MAKORIN-RELATED;AT2G02160-NA CCCH-type zinc finger family protein nucleus;AT2G02160-NA CCCH-type zinc finger family protein nucleus
					Eucgr.G02557	;AT2G02160-NA CCCH-type zinc finger family protein nucleus
					Eucgr.G02558	NA

Continue...

Trait	SNP	Chr	Position	P.value	Gene name	Description
					Eucgr.G02559	(M=5) PF06454 - Protein of unknown function (DUF1084);AT2G02180-TOM3 tobamovirus multiplication protein 3 plasma membrane;AT2G02180-TOM3 tobamovirus multiplication protein 3 plasma membrane
					Eucgr.G02559	;AT2G02180-TOM3 tobamovirus multiplication protein 3 plasma membrane
					Eucgr.G02559	;AT2G02180-TOM3 tobamovirus multiplication protein 3 plasma membrane
					Eucgr.G02559	;AT2G02180-TOM3 tobamovirus multiplication protein 3 plasma membrane
					Eucgr.G02560	(M=3) 1.13.99.1 - Inositol oxygenase.;AT1G14520-MIOX1 myo-inositol oxygenase 1 cytosol;AT1G14520-MIOX1 myo-inositol oxygenase 1 cytosol
					Eucgr.G02560	;AT1G14520-MIOX1 myo-inositol oxygenase 1 cytosol
					Eucgr.G02560	;AT1G14520-MIOX1 myo-inositol oxygenase 1 cytosol
					Eucgr.G02561	(M=38) PTHR11732:SF34 - ALDO-KETO REDUCTASE;AT2G37770-NA NAD(P)-linked oxidoreductase superfamily protein multiple
					Eucgr.G02562	NA
					Eucgr.G02563	(M=1) PTHR24015:SF155 - PENTATRICOPEPTIDE (PPR) REPEAT-CONTAINING PROTEIN;AT5G55840-NA Pentatricopeptide repeat (PPR) superfamily protein plastid
					Eucgr.G02564	(M=1) PTHR12999:SF1 - GB DEF: AT1G50300/F14I3_23 (TAF15);AT1G50300-TAF15 TBP-associated factor 15 nucleus
WBD	EuBR01s33473818	1	38156682	9.60 × 10 ⁻⁷	NA	NA
					Eucgr.B02853	(M=39) PF02309 - AUX/IAA family;AT2G01200-IAA32,MEE10 indole-3-acetic acid inducible 32 nucleus
					Eucgr.B02854	(M=31) PF00564 - PB1 domain;AT2G01190-NA Octicosapeptide/Phox/Bem1p family protein nucleus
					Eucgr.B02854	;AT2G01190-NA Octicosapeptide/Phox/Bem1p family protein nucleus
					Eucgr.B02855	(M=8) 3.1.3.4 - Phosphatidate phosphatase.;AT1G15080-ATLPP2,ATPAP2,LPP2 lipid phosphate phosphatase 2 plasma membrane
					Eucgr.B03812	(M=3) PTHR11668:SF17 - BSU-PROTEIN PHOSPHATASE;AT1G08420-BSL2 BRI1 suppressor 1 (BSU1)-like 2 nucleus
					Eucgr.B03813	(M=1) K11108 - RNA 3'-terminal phosphate cyclase-like protein;AT5G22100-NA RNA cyclase family protein mitochondrion
					Eucgr.B03814	(M=25) PTHR22849 - WDSAM1 PROTEIN;AT2G35930-PUB23 plant U-box 23 cytosol
					Eucgr.B03815	(M=1) PTHR21678 - GROWTH INHIBITION AND DIFFERENTIATION RELATED PROTEIN 88;AT5G22120-NA NA nucleus
					Eucgr.B03816	(M=6) PTHR13902:SF5 - GB DEF: BHLH TRANSCRIPTION FACTOR;AT2G27230-LHW transcription factor-related nucleus
					Eucgr.B03817	(M=10) PTHR10797 - CCR4-NOT TRANSCRIPTION COMPLEX SUBUNIT;AT5G22250-NA Polynucleotidyl transferase, ribonuclease H-like superfamily protein nucleus
					Eucgr.B03818	(M=15) PF04844 - Transcriptional repressor, ovate;AT2G36026-NA Ovate family protein nucleus
					Eucgr.B03819	(M=88) PF03168 - Late embryogenesis abundant protein;AT3G52460-NA hydroxyproline-rich glycoprotein family protein plasma membrane
					Eucgr.H04137	;AT1G08530-NA NA plastid
					Eucgr.H04137	;AT1G08530-NA NA plastid
					Eucgr.H04138	(M=15) 3.1.2.2 - Palmitoyl-CoA hydrolase.;AT1G48320-NA Thioesterase superfamily protein peroxisome
					Eucgr.H04139	(M=15) 3.1.2.2 - Palmitoyl-CoA hydrolase.;AT1G48320-NA Thioesterase superfamily protein peroxisome;AT1G48320-NA Thioesterase superfamily protein peroxisome
					Eucgr.F00028	(M=6) PTHR11731//PTHR11731:SF54 - PROTEASE FAMILY S9B,C DIPEPTIDYL-PEPTIDASE IV-RELATED // SUBFAMILY NOT NAMED;AT5G36210-NA alpha/beta-Hydrolases superfamily protein plastid
					Eucgr.F00029	NA

Continue...

Trait	SNP	Chr	Position	P.value	Gene name	Description
WBD	EuBR06s779311	6	147085	1.73×10^{-6}	Eucgr.F00030	(M=7) PF04759 - Protein of unknown function, DUF617;AT5G42680-NA Protein of unknown function, DUF617 plastid
WBD	EuBR05s13465317	5	12396851	2.32×10^{-6}	NA	NA
SGR	EuBR07s50574163	7	53140783	9.54×10^{-8}	Eucgr.G03236	(M=1) PTHR26402//PTHR26402:SF52 - RESPONSE REGULATOR OF TWO-COMPONENT SYSTEM;AT3G57040-ARR9,ATRR4 response regulator 9 nucleus
SGR	EuBR01s32182181	1	36865045	1.45×10^{-7}	NA	NA
					Eucgr.J02335	NA
					Eucgr.J02336	(M=1) PTHR23074:SF17 - FIDGETIN LIKE-1;AT3G27120-NA P-loop containing nucleoside triphosphate hydrolases superfamily protein nucleus
SGR	EuBR10s29287132	10	28512117	1.87×10^{-7}	Eucgr.J02337	(M=204) KOG0157 - Cytochrome P450 CYP4/CYP19/CYP26 subfamilies;AT3G48520-CYP94B3 cytochrome P450, family 94, subfamily B, polypeptide 3 endoplasmic reticulum
					Eucgr.J02338	(M=1) 1.8.3.1 - Sulfite oxidase.;AT3G01910-AT-SO,AtSO,SOX sulfite oxidase multiple
					Eucgr.J02339	NA
SGR	EuBR07s17699641	7	15925866	2.51×10^{-7}	NA	NA
					Eucgr.B02072	(M=7) PTHR11260:SF31 - OS04G0435500 PROTEIN;AT1G77290-NA Glutathione S-transferase family protein cytosol
SGR	EuBR02s39973654	1	19363017	3.02×10^{-7}	Eucgr.B02073	NA
					Eucgr.B02818	;AT5G01470-NA S-adenosyl-L-methionine-dependent methyltransferases superfamily protein cytosol
SGR	EuBR03s19349239	3	48609424	5.11×10^{-7}	Eucgr.C01233	(M=1) KOG1081 - Transcription factor NSD1 and related SET domain proteins;AT4G30860-ASHR3,SDG4 SET domain group 4 nucleus
SGR	EuBR01s14426556	1	11173430	7.81×10^{-7}	NA	NA
SOL	EuBR01s1482744	10	30533938	7.68×10^{-14}	Eucgr.A00114	;AT5G24360-ATIRE1-1,IRE1-1 inositol requiring 1-1 multiple
SOL	EuBR01s18411052	1	24353087	1.49×10^{-10}	Eucgr.J02542	NA
					Eucgr.A01183	NA
					Eucgr.F03346	(M=13) PF06830 - Root cap;AT5G54370-NA Late embryogenesis abundant (LEA) protein-related extracellular
					Eucgr.F03347	(M=13) PF06830 - Root cap;AT5G54370-NA Late embryogenesis abundant (LEA) protein-related extracellular
					Eucgr.F03348	(M=18) K14496 - abscisic acid receptor PYR/PYL family;AT1G70880-NA Polyketide cyclase/dehydrase and lipid transport superfamily protein cytosol
					Eucgr.F03349	(M=18) K14496 - abscisic acid receptor PYR/PYL family;AT1G70880-NA Polyketide cyclase/dehydrase and lipid transport superfamily protein cytosol
SOL	EuBR06s42411988	6	44910850	5.37×10^{-9}	Eucgr.F03351	;AT2G20825-ULT2 Developmental regulator, ULTRAPETALA nucleus
					Eucgr.F03352	;AT4G27390-NA NA plastid
					Eucgr.F03352	;AT4G27390-NA NA plastid
					Eucgr.F03352	;AT4G27390-NA NA plastid
					Eucgr.F03352	;AT4G27390-NA NA plastid
					Eucgr.F03353	(M=3) K12486 - stromal membrane-associated protein;AT5G54310-AGD5,NEV ARF-GAP domain 5 nucleus
					Eucgr.F03354	NA
					Eucgr.B03164	;AT1G26610-NA C2H2-like zinc finger protein nucleus
					Eucgr.B03165	(M=79) PF03106 - WRKY DNA -binding domain;AT1G69310-ATWRKY57,WRKY57 WRKY DNA-binding protein 57 nucleus
SOL	EuBR02s56837794	2	52149729	7.79×10^{-8}	Eucgr.B03166	(M=3) 2.4.1.1 - Glycogen phosphorylase.;AT3G29320-NA Glycosyl transferase, family 35 plastid;AT3G29320-NA Glycosyl transferase, family 35 plastid
					Eucgr.B03166	;AT3G29320-NA Glycosyl transferase, family 35 plastid

Continue...

Trait	SNP	Chr	Position	P.value	Gene name	Description
					Eucgr.B03166	;AT3G29320-NA Glycosyl transferase, family 35 plastid
					Eucgr.B03166	;AT3G29320-NA Glycosyl transferase, family 35 plastid
					Eucgr.B03166	;AT3G29320-NA Glycosyl transferase, family 35 plastid
					Eucgr.B03167	(M=4) K08509 - synaptosomal-associated protein, 29kDa;AT5G61210-ATSNAP33,ATSNAP33B,SNAP33,SNP33 soluble N-ethylmaleimide-sensitive factor adaptor protein 33 nucleus
					Eucgr.B03168	(M=34) PTHR22953 - ACID PHOSPHATASE RELATED;AT1G13900-NA Purple acid phosphatases superfamily protein vacuole
					Eucgr.B03169	(M=4) K08509 - synaptosomal-associated protein, 29kDa;AT5G61210-ATSNAP33,ATSNAP33B,SNAP33,SNP33 soluble N-ethylmaleimide-sensitive factor adaptor protein 33 nucleus
					Eucgr.B03170	(M=34) PTHR22953 - ACID PHOSPHATASE RELATED;AT1G13900-NA Purple acid phosphatases superfamily protein vacuole
					Eucgr.B03171	(M=104) PF11721 - Di-glucose binding within endoplasmic reticulum;AT1G25570-NA Di-glucose binding protein with Leucine-rich repeat domain vacuole
SOL	EuBR02s56837794	2	52149729	7.79 × 10 ⁻⁸	Eucgr.B03172	(M=1) PTHR12993 - N-ACETYLGLUCOSAMINYL-PHOSPHATIDYLINOSITOL DE-N-ACETYLASE-RELATED;AT3G58130-NA N-acetylglucosaminylphosphatidylinositol de-N-acetylase family protein plasma membrane;AT3G58130-NA N-acetylglucosaminylphosphatidylinositol de-N-acetylase family protein plasma membrane
					Eucgr.B03172	;AT3G58130-NA N-acetylglucosaminylphosphatidylinositol de-N-acetylase family protein plasma membrane
					Eucgr.B03172	;AT3G58130-NA N-acetylglucosaminylphosphatidylinositol de-N-acetylase family protein plasma membrane
					Eucgr.B03173	(M=1) PTHR24420//PTHR24420:SF518 - LEUCINE-RICH REPEAT RECEPTOR-LIKE PROTEIN KINASE;AT5G61240-NA Leucine-rich repeat (LRR) family protein multiple;AT5G61240-NA Leucine-rich repeat (LRR) family protein multiple
					Eucgr.B03173	;AT5G61240-NA Leucine-rich repeat (LRR) family protein multiple
					Eucgr.B03174	(M=8) PTHR14363 - HEPARANASE-RELATED;AT5G61250-AtGUS1,GUS1 glucuronidase 1 plasma membrane
					Eucgr.B03175	(M=18) PF03763 - Remorin, C-terminal region;AT1G13920-NA Remorin family protein nucleus
					Eucgr.B03176	(M=2) PTHR13139 - RING FINGER AND CCCH-TYPE ZINC FINGER DOMAIN-CONTAINING PROTEIN;AT1G69330-NA RING/U-box superfamily protein nucleus
					Eucgr.C03075	(M=5) PF01578 - Cytochrome C assembly protein;ATMG00960-NA Cytochrome C assembly protein NA
					Eucgr.C03076	(M=46) KOG0617 - Ras suppressor protein (contains leucine-rich repeats);AT1G69550-NA disease resistance protein (TIR-NBS-LRR class) plasma membrane;AT1G69550-NA disease resistance protein (TIR-NBS-LRR class) plasma membrane
SOL	EuBR03s57703759	3	61561917	1.73 × 10 ⁻⁷	Eucgr.C03076	;AT1G69550-NA disease resistance protein (TIR-NBS-LRR class) plasma membrane
					Eucgr.C03076	;AT1G69550-NA disease resistance protein (TIR-NBS-LRR class) plasma membrane
					Eucgr.C03076	;AT1G69550-NA disease resistance protein (TIR-NBS-LRR class) plasma membrane
					Eucgr.C03076	;AT1G69550-NA disease resistance protein (TIR-NBS-LRR class) plasma membrane
					Eucgr.C03076	;AT1G69550-NA disease resistance protein (TIR-NBS-LRR class) plasma membrane
					Eucgr.C03077	NA
SOL	EuBR09s24197228	9	24161506	3.69 × 10 ⁻⁷	Eucgr.I01375	(M=1) K13119 - protein FAM50;AT2G21150-XCT XAP5 family protein nucleus

Continue...

Trait	SNP	Chr	Position	P.value	Gene name	Description
TSC	EuBR10s1696823	10	1696763	6.93×10^{-14}	Eucgr.J00155	(M=7) PF071107 - Wound-induced protein W112;AT3G10985-ATWI-12,SAG20,W112 senescence associated gene 20 plasma membrane
					Eucgr.A02909	(M=140) PF00083 - Sugar (and other) transporter;AT5G61520-NA Major facilitator superfamily protein plasma membrane
					Eucgr.A02909	;AT5G61520-NA Major facilitator superfamily protein plasma membrane
TSC	EuBR01s39512949	1	44180804	1.52×10^{-8}	Eucgr.A02910	(M=3) 1.6.5.5 - NADPH:quinone reductase.;AT5G61510-NA GroES-like zinc-binding alcohol dehydrogenase family protein cytosol
					Eucgr.A02910	;AT5G61510-NA GroES-like zinc-binding alcohol dehydrogenase family protein cytosol
					Eucgr.A02911	(M=84) KOG0254 - Predicted transporter (major facilitator superfamily);AT5G61520-NA Major facilitator superfamily protein plasma membrane
					Eucgr.A02912	(M=4) PTHR23176 - RHO/RAC/CDC GTPASE-ACTIVATING PROTEIN;AT5G61530-NA small G protein family protein / RhoGAP family protein multiple
					Eucgr.D01237	(M=2) K12180 - COP9 signalosome complex subunit 7;AT1G02090-ATCSN7,COP15,CSN7,FUS5 Proteasome component (PCI) domain protein multiple
					Eucgr.D01238	(M=2) K13127 - RING finger protein 113A;AT1G01350-NA Zinc finger (CCCH-type/C3HC4-type RING finger) family protein nucleus
					Eucgr.D01238	NA
					Eucgr.D01238	NA
TSC	EuBR04s22295129	4	25830558	2.70×10^{-8}	Eucgr.D01239	(M=3) 3.1.2.22 - Palmitoyl-protein hydrolase.;AT3G60340-NA alpha/beta-Hydrolases superfamily protein extracellular;AT3G60340-NA alpha/beta-Hydrolases superfamily protein extracellular
					Eucgr.D01239	NA
					Eucgr.D01240	(M=157) KOG4412 - 26S proteasome regulatory complex, subunit PSMD10;AT1G03670-NA ankyrin repeat family protein multiple
					Eucgr.D01241	NA
					Eucgr.D01243	;AT4G03460-NA Ankyrin repeat family protein cytosol
					Eucgr.D01244	(M=157) KOG4412 - 26S proteasome regulatory complex, subunit PSMD10;AT1G03670-NA ankyrin repeat family protein multiple
TSC	EuBR07s252985	7	252925	2.69×10^{-7}	Eucgr.G00025	(M=11) PTHR13943 - HRAS-LIKE SUPPRESSOR - RELATED;AT3G02700-NA NC domain-containing protein-related multiple
TSC	EuBR03s7498452	3	6469424	6.10×10^{-7}	NA	NA
TSC	EuBR08s57640594	8	54276404	1.33×10^{-6}	NA	NA
TSC	EuBR04s14035937	4	13563332	1.39×10^{-6}	Eucgr.D00751	(M=172) K09422 - myb proto-oncogene protein, plant;AT4G21440-ATM4,ATMYB102,MYB102 MYB-like 102 nucleus

Table S3 - Significant SNP markers and candidate genes for growth traits for the 1,772 *Eucalyptus grandis* individuals using the multi-trait analysis.

Traits	Model	SNP name	Chr	Position	P.value	Gene name	Description
HEI3-HEI6	Common	EuBR03s22449999	3	24421549	8.41×10^{-6}	NA	
HEI3-HEI6	Common and Full	EuBR04s246324	4	246264	9.05×10^{-7}	Eucgr.D00028	(M=1) PTHR10438:SF14 - THIOREDOXIN-RELATED;AT2G35010-ATO1,TO1 thioredoxin O1 mitochondrion
						Eucgr.K00190	(M=55) PTHR11945:SF19 - MADS BOX PROTEIN;AT5G51870-AGL71 AGAMOUS-like 71 nucleus
						Eucgr.K00191	(M=55) PTHR11945:SF19 - MADS BOX PROTEIN;AT5G62165-AGL42 AGAMOUS-like 42 multiple
						Eucgr.K00192	(M=55) PTHR11945:SF19 - MADS BOX PROTEIN;AT5G62165-AGL42 AGAMOUS-like 42 multiple
						Eucgr.K00193	(M=55) PTHR11945:SF19 - MADS BOX PROTEIN;AT5G62165-AGL42 AGAMOUS-like 42 multiple;AT5G62165-AGL42 AGAMOUS-like 42 multiple
HEI3-HEI6	Full	EuBR11s2322411	11	1333920	1.07×10^{-5}	Eucgr.K00193	;AT5G62165-AGL42 AGAMOUS-like 42 multiple
						Eucgr.K00193	(M=55) PTHR11945:SF19 - MADS BOX PROTEIN;AT5G62165-AGL42 AGAMOUS-like 42 multiple
						Eucgr.K00194	(M=55) PTHR11945:SF19 - MADS BOX PROTEIN;AT5G62165-AGL42 AGAMOUS-like 42 multiple
						Eucgr.K00194	(M=55) PTHR11945:SF19 - MADS BOX PROTEIN;AT5G62165-AGL42 AGAMOUS-like 42 multiple
						Eucgr.K00194	(M=55) PTHR11945:SF19 - MADS BOX PROTEIN;AT5G62165-AGL42 AGAMOUS-like 42 multiple
						Eucgr.K00195	NA
						Eucgr.F02939	NA
						Eucgr.F02940	(M=14) KOG0131 - Splicing factor 3b, subunit 4;AT4G10110-NA RNA-binding (RRM/RBD/RNP motifs) family protein nucleus;AT4G10110-NA RNA-binding (RRM/RBD/RNP motifs) family protein nucleus
						Eucgr.F02940	;AT4G10110-NA RNA-binding (RRM/RBD/RNP motifs) family protein nucleus
						Eucgr.F02940	;AT4G10110-NA RNA-binding (RRM/RBD/RNP motifs) family protein nucleus
						Eucgr.F02940	;AT4G10110-NA RNA-binding (RRM/RBD/RNP motifs) family protein nucleus
HEI3-DBH6	Common	EuBR06s39120397	6	41619218	7.57×10^{-6}	Eucgr.F02941	(M=25) 3.2.1.4 - Cellulase.;AT1G71380-ATCEL3,ATGH9B3,CEL3 cellulase 3 extracellular
						Eucgr.F02942	;AT5G66460-NA Glycosyl hydrolase superfamily protein extracellular
						Eucgr.F02943	(M=27) PF00150 - Cellulase (glycosyl hydrolase family 5);AT5G66460-NA Glycosyl hydrolase superfamily protein extracellular
						Eucgr.F02944	(M=2) K03635 - molybdopterin synthase catalytic subunit;AT4G10100-CNX7,SIR5 co-factor for nitrate, reductase and xanthine dehydrogenase 7 cytosol
						Eucgr.F02945	;AT4G13530-NA NA nucleus
						Eucgr.F02945	;AT4G13530-NA NA nucleus
						Eucgr.F02945	;AT4G13530-NA NA nucleus

Continue...

Traits	Model	SNP name	Chr	Position	P.value	Gene name	Description
HEI3-DBH6	Common	EuBR06s39120397	6	41619218	7.57×10^{-6}	Eucgr.F02946	(M=4) 6.1.1.21 - Histidine--tRNA ligase.;AT3G02760-NA Class II aaRS and biotin synthetases superfamily protein cytosol
						Eucgr.F02946	;AT3G02760-NA Class II aaRS and biotin synthetases superfamily protein cytosol
						Eucgr.F02947	NA
HEI3-VOL6	Common	EuBR06s39120397	6	41619218	1.14×10^{-5}	Eucgr.F02939	NA
						Eucgr.F02940	(M=14) KOG0131 - Splicing factor 3b, subunit 4;AT4G10110-NA RNA-binding (RRM/RBD/RNP motifs) family protein nucleus
						Eucgr.F02940	;AT4G10110-NA RNA-binding (RRM/RBD/RNP motifs) family protein nucleus
						Eucgr.F02940	;AT4G10110-NA RNA-binding (RRM/RBD/RNP motifs) family protein nucleus
						Eucgr.F02940	;AT4G10110-NA RNA-binding (RRM/RBD/RNP motifs) family protein nucleus
						Eucgr.F02941	(M=25) 3.2.1.4 - Cellulase.;AT1G71380-ATCEL3,ATGH9B3,CEL3 cellulase 3 extracellular
						Eucgr.F02942	;AT5G66460-NA Glycosyl hydrolase superfamily protein extracellular
						Eucgr.F02943	(M=27) PF00150 - Cellulase (glycosyl hydrolase family 5);AT5G66460-NA Glycosyl hydrolase superfamily protein extracellular
						Eucgr.F02944	(M=2) K03635 - molybdopterin synthase catalytic subunit;AT4G10100-CN7,SIR5 co-factor for nitrate, reductase and xanthine dehydrogenase 7 cytosol
						Eucgr.F02945	;AT4G13530-NA NA nucleus
						Eucgr.F02945	;AT4G13530-NA NA nucleus
Eucgr.F02945	;AT4G13530-NA NA nucleus						
Eucgr.F02946	(M=4) 6.1.1.21 - Histidine--tRNA ligase.;AT3G02760-NA Class II aaRS and biotin synthetases superfamily protein cytosol						
Eucgr.F02946	;AT3G02760-NA Class II aaRS and biotin synthetases superfamily protein cytosol						
Eucgr.F02947	NA						
HEI6-VOL6	Common and Full	EuBR03s43394028	3	44908926	9.33×10^{-7}	Eucgr.C02324	(M=60) PTHR23070 - BCS1 AAA-TYPE ATPASE;AT5G57480-NA P-loop containing nucleoside triphosphate hydrolases superfamily protein endoplasmic reticulum
						Eucgr.C02326	(M=12) KOG0251 - Clathrin assembly protein AP180 and related proteins, contain ENTH domain;AT4G25940-NA ENTH/ANTH/VHS superfamily protein nucleus
						Eucgr.C02327	(M=18) K13428 - LRR receptor-like serine/threonine-protein kinase EFR;AT5G20480-EFR EF-TU receptor plasma membrane
						Eucgr.C02328	(M=55) PF00560//PF07714//PF08263 - Leucine Rich Repeat // Protein tyrosine kinase // Leucine rich repeat N-terminal domain;AT3G47570-NA Leucine-rich repeat protein kinase family protein plasma membrane
						Eucgr.C02329	(M=39) PF02309 - AUX/IAA family;AT5G57420-IAA33 indole-3-acetic acid inducible 33 nucleus
HEI6-VOL6	Full	EuBR02s12845338	2	10920732	1.45×10^{-5}	NA	NA

Continue...

Traits	Model	SNP name	Chr	Position	P.value	Gene name	Description
DBH3-DBH6	Full	EuBR03s72654230	3	73723975	1.59×10^{-5}	Eucgr.C03882	(M=90) PTHR22950 - AMINO ACID TRANSPORTER;AT1G77380-AAP3,ATAAP3 amino acid permease 3 plasma membrane
						Eucgr.C03884	(M=90) PTHR22950 - AMINO ACID TRANSPORTER;AT1G77380-AAP3,ATAAP3 amino acid permease 3 plasma membrane
DBH3-VOL3	Full and Interaction	EuBR05s62102817	5	58629530	2.19×10^{-6}	Eucgr.E03571	(M=74) PF01397//PF03936 - Terpene synthase, N-terminal domain // Terpene synthase family, metal binding domain;AT4G16740-ATTPS03,TPS03 terpene synthase 03 cytosol
						Eucgr.E03572	(M=74) PF01397//PF03936 - Terpene synthase, N-terminal domain // Terpene synthase family, metal binding domain;AT1G61680-ATTPS14,TPS14 terpene synthase 14 plastid
						Eucgr.E03573	
DBH3-VOL6	Common	EuBR08s53916832	8	50764014	6.55×10^{-6}	Eucgr.H03708 Eucgr.H03709	;AT3G62450-NA NA extracellular
DBH3-VOL6	Full	EuBR03s72654230	3	73723975	1.02×10^{-5}	Eucgr.C03882	(M=90) PTHR22950 - AMINO ACID TRANSPORTER;AT1G77380-AAP3,ATAAP3 amino acid permease 3 plasma membrane
						Eucgr.C03884	(M=90) PTHR22950 - AMINO ACID TRANSPORTER;AT1G77380-AAP3,ATAAP3 amino acid permease 3 plasma membrane
VOL3-HEI6	Common and Full	EuBR02s12845338	2	10920732	7.23×10^{-6}	NA	NA
VOL3-HEI6	Common	EuBR03s43394028	3	44908926	1.01×10^{-5}	Eucgr.C02324	(M=60) PTHR23070 - BCS1 AAA-TYPE ATPASE;AT5G57480-NA P-loop containing nucleoside triphosphate hydrolases superfamily protein endoplasmic reticulum
						Eucgr.C02326	(M=12) KOG0251 - Clathrin assembly protein AP180 and related proteins, contain ENTH domain;AT4G25940-NA ENTH/ANTH/VHS superfamily protein nucleus
						Eucgr.C02327	(M=18) K13428 - LRR receptor-like serine/threonine-protein kinase EFR;AT5G20480-EFR EF-TU receptor plasma membrane
						Eucgr.C02328	(M=55) PF00560//PF07714//PF08263 - Leucine Rich Repeat // Protein tyrosine kinase // Leucine rich repeat N-terminal domain;AT3G47570-NA Leucine-rich repeat protein kinase family protein plasma membrane
						Eucgr.C02329	(M=39) PF02309 - AUX/IAA family;AT5G57420-IAA33 indole-3-acetic acid inducible 33 nucleus
VOL3-DBH6	Common	EuBR05s73512379	5	75396072	1.22×10^{-5}	Eucgr.E04289	(M=194) PTHR24420//PTHR24420:SF473 - LEUCINE-RICH REPEAT RECEPTOR-LIKE PROTEIN KINASE // SUBFAMILY NOT NAMED;AT1G58190-AtRLP9,RLP9 receptor like protein 9 multiple NA
						Eucgr.E04290	(M=194) PTHR24420//PTHR24420:SF473 - LEUCINE-RICH REPEAT RECEPTOR-LIKE PROTEIN KINASE // SUBFAMILY NOT NAMED;AT1G74190-AtRLP15,RLP15 receptor like protein 15 plasma membrane
						Eucgr.E04291	
VOL3-DBH6	Common	EuBR07s16969079	7	15429939	1.38×10^{-5}	Eucgr.G00982 Eucgr.G00985	(M=101) PF00226 - DnaJ domain;AT5G64360-NA Chaperone DnaJ-domain superfamily protein cytosol NA

Continue...

Traits	Model	SNP name	Chr	Position	P.value	Gene name	Description
VOL3-DBH6	Full and Common	EuBR02s2712998	2	1260315	9.68×10^{-6}	Eucgr.B00184	(M=261) KOG0156 - Cytochrome P450 CYP2 subfamily;AT5G25120-CYP71B11 ytochrome p450, family 71, subfamily B, polypeptide 11 endoplasmic reticulum
VOL3-DBH6	Interaction	EuBR05s40135536	5	39939948	1.42×10^{-5}	Eucgr.E02580	(M=52) PTHR11695:SF285 - L-THREONINE 3-DEHYDROGENASE;AT4G37980-ATCAD7,CAD7,ELI3,ELI3-1 elicitor-activated gene 3-1 cytosol
VOL3-VOL6	Common	EuBR05s73512379	5	75396072	1.40×10^{-5}	Eucgr.E04289	(M=194) PTHR24420//PTHR24420:SF473 - LEUCINE-RICH REPEAT RECEPTOR-LIKE PROTEIN KINASE // SUBFAMILY NOT NAMED;AT1G58190-AtRLP9,RLP9 receptor like protein 9 multiple
						Eucgr.E04290	(M=52) PTHR11695:SF285 - L-THREONINE 3-DEHYDROGENASE;AT4G37980-ATCAD7,CAD7,ELI3,ELI3-1 elicitor-activated gene 3-1 cytosol
						Eucgr.E04291	(M=194) PTHR24420//PTHR24420:SF473 - LEUCINE-RICH REPEAT RECEPTOR-LIKE PROTEIN KINASE // SUBFAMILY NOT NAMED;AT1G74190-AtRLP15,RLP15 receptor like protein 15 plasma membrane
VOL3-VOL6	Full	EuBR02s2712998	2	1260315	3.41×10^{-6}	Eucgr.B00184	(M=261) KOG0156 - Cytochrome P450 CYP2 subfamily;AT5G25120-CYP71B11 ytochrome p450, family 71, subfamily B, polypeptide 11 endoplasmic reticulum
VOL3-VOL6	Interaction	EuBR05s40135536	5	39939948	6.29×10^{-6}	Eucgr.E02580	(M=52) PTHR11695:SF285 - L-THREONINE 3-DEHYDROGENASE;AT4G37980-ATCAD7,CAD7,ELI3,ELI3-1 elicitor-activated gene 3-1 cytosol
DBH6-HEI6	Common and Full	EuBR02s12845338	2	10920732	5.92×10^{-6}	NA	NA
DBH6-HEI6	Common and Full	EuBR03s43394028	3	44908926	1.54×10^{-6}	Eucgr.C02324	(M=60) PTHR23070 - BCS1 AAA-TYPE ATPASE;AT5G57480-NA P-loop containing nucleoside triphosphate hydrolases superfamily protein endoplasmic reticulum
						Eucgr.C02326	(M=12) KOG0251 - Clathrin assembly protein AP180 and related proteins, contain ENTH domain;AT4G25940-NA ENTH/ANTH/VHS superfamily protein nucleus
						Eucgr.C02327	(M=18) K13428 - LRR receptor-like serine/threonine-protein kinase EFR;AT5G20480-EFR EF-TU receptor plasma membrane
						Eucgr.C02328	(M=55) PF00560//PF07714//PF08263 - Leucine Rich Repeat // Protein tyrosine kinase // Leucine rich repeat N-terminal domain;AT3G47570-NA Leucine-rich repeat protein kinase family protein plasma membrane
						Eucgr.C02329	(M=39) PF02309 - AUX/IAA family;AT5G57420-IAA33 indole-3-acetic acid inducible 33 nucleus
DBH6-VOL6	Common	EuBR01s116241	1	2688124	4.31×10^{-6}	Eucgr.A00011	;AT2G30695-NA NA plastid
DBH6-VOL6	Common	EuBR02s2712998	2	1260315	1.09×10^{-5}	Eucgr.B00184	(M=261) KOG0156 - Cytochrome P450 CYP2 subfamily;AT5G25120-CYP71B11 ytochrome p450, family 71, subfamily B, polypeptide 11 endoplasmic reticulum

Table S4 - Significant SNP markers and candidate genes for wood-quality traits for the 1,772 *Eucalyptus grandis* individuals using the multi-trait analysis.

Traits	Model	SNP name	Chr	Position	P.value	Gene name	Description
SOL-TEX	Common and Full	EuBR06s19529730	6	20443484	2.66×10^{-7}	Eucgr.F01519	NA
						Eucgr.F01520	NA
						Eucgr.F01521	NA
						Eucgr.F01522	NA
						Eucgr.F01523	NA
SOL-TEX	Common and Full	EuBR06s52964694	6	56543212	5.06×10^{-6}	Eucgr.F04390	(M=281) PTHR24420//PTHR24420:SF474 - LEUCINE-RICH REPEAT RECEPTOR-LIKE PROTEIN KINASE // SUBFAMILY NOT NAMED;AT3G47570-NA Leucine-rich repeat protein kinase family protein plasma membrane
						Eucgr.F04391	(M=281) PTHR24420//PTHR24420:SF474 - LEUCINE-RICH REPEAT RECEPTOR-LIKE PROTEIN KINASE // SUBFAMILY NOT NAMED;AT3G47570-NA Leucine-rich repeat protein kinase family protein plasma membrane
						Eucgr.F04392	NA
						Eucgr.F04393	(M=281) PTHR24420//PTHR24420:SF474 - LEUCINE-RICH REPEAT RECEPTOR-LIKE PROTEIN KINASE // SUBFAMILY NOT NAMED;AT3G47570-NA Leucine-rich repeat protein kinase family protein plasma membrane
						Eucgr.F04394	NA
						Eucgr.F04395	;AT1G72500-NA NA cytosol
						Eucgr.F04396	(M=281) PTHR24420//PTHR24420:SF474 - LEUCINE-RICH REPEAT RECEPTOR-LIKE PROTEIN KINASE // SUBFAMILY NOT NAMED;AT3G47570-NA Leucine-rich repeat protein kinase family protein plasma membrane
						Eucgr.F04397	(M=2) PTHR10338 - VON WILLEBRAND FACTOR, TYPE A DOMAIN CONTAINING;AT1G72500-NA NA cytosol;AT1G72500-NA NA cytosol
						Eucgr.F04397	NA
						Eucgr.F04397	NA
Eucgr.F04397	NA						
Eucgr.F04397	NA						

Continue...

Traits	Model	SNP name	Chr	Position	P.value	Gene name	Description
SOL-TEX	Common and Full	EuBR06s52964694	6	56543212	5.06×10^{-6}	Eucgr.F04398	(M=1) K02730 - 20S proteasome subunit alpha 1;AT2G05840-PAA2 20S proteasome subunit PAA2 multiple
						Eucgr.F04400	NA
						Eucgr.F04401	(M=15) PF07911 - Protein of unknown function (DUF1677);AT1G72510-NA Protein of unknown function (DUF1677) plasma membrane;AT1G72510-NA Protein of unknown function (DUF1677) plasma membrane
						Eucgr.F04401	NA
SGR-TEX	Full and Interaction	EuBR11s43922247	11	43248389	7.26×10^{-6}	Eucgr.K03483	(M=2) PF00271//PF00642 - Helicase conserved C-terminal domain // Zinc finger C-x8-C-x5-C-x3-H type (and similar);AT2G47680-NA zinc finger (CCCH type) helicase family protein nucleus
						Eucgr.K03516	NA
SGR-TEX	Full and Interaction	EuBR11s44284539	11	43610681	1.16×10^{-5}	Eucgr.K03517	;AT3G62650-NA NA cytosol
						Eucgr.C01030	;AT3G24630-NA NA nucleus
SGR-TEX	Interaction	EuBR03s16484895	3	16899795	9.83×10^{-6}	Eucgr.C01031	(M=4) PTHR11886//PTHR11886:SF22 - DYNEIN LIGHT CHAIN // SUBFAMILY NOT NAMED;AT1G52240-ATROPGEF11,PIRF1,ROPGEF11 RHO guanyl-nucleotide exchange factor 11 multiple
						Eucgr.C01032	(M=4) PF04367 - Protein of unknown function (DUF502);AT2G20120-COV1 Protein of unknown function (DUF502) multiple
						Eucgr.K03517	;AT3G62650-NA NA cytosol
SGR-SOL	Full and Interaction	EuBR08s49045121	8	63920596	6.88×10^{-6}	Eucgr.H03343	(M=49) PF03492 - SAM dependent carboxyl methyltransferase;AT5G66430-NA S-adenosyl-L-methionine-dependent methyltransferases superfamily protein cytosol
						Eucgr.H03343	;AT5G66430-NA S-adenosyl-L-methionine-dependent methyltransferases superfamily protein cytosol
						Eucgr.H03343	;AT5G66430-NA S-adenosyl-L-methionine-dependent methyltransferases superfamily protein cytosol
						Eucgr.H03343	;AT5G66430-NA S-adenosyl-L-methionine-dependent methyltransferases superfamily protein cytosol
						Eucgr.H03343	;AT5G66430-NA S-adenosyl-L-methionine-dependent methyltransferases superfamily protein cytosol

Table S5 - Significant SNP markers and candidate genes for growth traits and wood-quality traits for the 1,772 *Eucalyptus grandis* individuals using the multi-trait analysis.

Traits	Model	SNP name	Chr	position	P.value	Gene name	description
PCY/DBH3	Common/Full	EuBR04s12292131	4	12520123	2.60×10^{-6}	Eucgr.D00674	(M=26) 3.4.16.5 - Carboxypeptidase C.;AT1G11080-scpl31 serine carboxypeptidase-like 31 extracellular
PCY/HEI6	Common	EuBR04s510275	4	510215	7.22×10^{-6}	NA	NA
TSC/HEI6	Interaction	EuBR09s38638406	9	38502871	1.54×10^{-5}	Eucgr.I02760	(M=3) PF06241 - Protein of unknown function (DUF1012);AT5G49960-NA NA mitochondrion;AT5G49960-NA NA mitochondrion
						Eucgr.I02761	(M=17) PF05078 - Protein of unknown function (DUF679);AT3G02430-NA Protein of unknown function (DUF679) plasma membrane
						Eucgr.I02762	(M=19) PTHR22880 - FALZ-RELATED BROMODOMAIN-CONTAINING PROTEINS;AT1G58025-NA DNA-binding bromodomain-containing protein nucleus;AT1G58025-NA DNA-binding bromodomain-containing protein nucleus
						Eucgr.I02764	(M=3) K06173 - tRNA pseudouridine synthase A [EC:5.4.99.12];AT1G09800-NA Pseudouridine synthase family protein cytosol
TSC/DBH6	Full/interaction	EuBR01s28498846	1	33181710	1.00×10^{-6}	NA	NA
TSC/DBH6	Interaction	EuBR10s6097228	10	6097168	1.48×10^{-5}	Eucgr.J00554	(M=1) PF00954//PF07714//PF11883 - S-locus glycoprotein family // Protein tyrosine kinase // Domain of unknown function (DUF3403);AT4G21380-ARK3,RK3 receptor kinase 3 plasma membrane
TSC/VOL6	Full	EuBR01s28498846	1	33181710	5.21×10^{-6}	NA	NA
TSC/VOL6	Full/Interaction	EuBR10s6097228	10	6097168	8.66×10^{-7}	Eucgr.J00554	(M=1) PF00954//PF07714//PF11883 - S-locus glycoprotein family // Protein tyrosine kinase // Domain of unknown function (DUF3403);AT4G21380-ARK3,RK3 receptor kinase 3 plasma membrane
TSC/DBH3	Interaction	EuBR06s44029558	6	46528420	8.17×10^{-6}	Eucgr.F03545	(M=2) K01164 - ribonuclease P/MRP protein subunit POP1;AT2G47300-NA ribonuclease Ps multiple
						Eucgr.F03546	(M=11) PF04783 - Protein of unknown function (DUF630);AT4G35240-NA Protein of unknown function (DUF630 and DUF632) nucleus
TSC/VOL3	Interaction	EuBR01s28498846	1	33181710	8.46×10^{-7}	Eucgr.J00554	(M=1) PF00954//PF07714//PF11883 - S-locus glycoprotein family // Protein tyrosine kinase // Domain of unknown function (DUF3403);AT4G21380-ARK3,RK3 receptor kinase 3 plasma membrane
TSC_VOL3	Interaction	EuBR10s6097228	10	6097168	1.16×10^{-5}	NA	NA

Table S6 - Number of significant single nucleotide polymorphism (SNP) associations considering Bonferroni correction (p -value = 1.63×10^{-5}) for growth and wood-quality traits using multi-trait mixed model (MTMM) for the *Eucalyptus grandis* breeding population

Traits 1 and 2	SNP(s)	Position	Ch.	p-values				
				Full	Interaction	Common	Trait 1	Trait 2
DBH3 and DBH6	EuBR03s72654230	73723975	3	1.59×10^{-5}	1.98×10^{-4}	4.07×10^{-3}	0.079	1.05×10^{-3}
DBH3 and VOL3	EuBR05s62102817	58629530	5	2.18×10^{-6}	8.60×10^{-6}	0.012	0.23	0.050
DBH3 and VOL6	EuBR08s53916832	50764014	8	2.24×10^{-5}	0.29	6.55×10^{-6}	0.0014	0.0051
	EuBR03s72654230	73723975	3	1.02×10^{-5}	2.01×10^{-4}	0.0025	0.077	9.32×10^{-4}
HEI3 and HEI6	EuBR03s22449999	24421549	3	8.46×10^{-6}	0.83	8.41×10^{-6}	0.0034	0.0023
	EuBR04s246324	246264	4	9.05×10^{-7}	0.36	3.79×10^{-6}	0.14	0.0065
	EuBR11s2322411	1333920	11	1.07×10^{-5}	0.018	3.25×10^{-5}	0.028	3.48×10^{-5}
HEI3 and DBH6	EuBR06s39120397	41619218	06	3.69×10^{-5}	0.54	7.57×10^{-6}	0.0018	0.00035
HEI3 and VOL6	EuBR06s39120397	41619218	06	6.04×10^{-5}	0.69	1.14×10^{-5}	0.0018	0.00065
HEI6 and VOL6	EuBR03s43394028	44908926	03	5.07×10^{-6}	0.57	9.33×10^{-7}	0.0038	0.0023
	EuBR02s12845338	10920732	02	1.45×10^{-5}	0.0059	0.00012	0.14	0.025
VOL3 and HEI6	EuBR02s12845338	10920732	2	1.30×10^{-5}	0.12	7.23×10^{-6}	0.0060	0.046
	EuBR03s43394028	44908926	3	5.83×10^{-5}	0.98	1.01×10^{-5}	0.015	6.29×10^{-3}
VOL3 and DBH6	EuBR05s73512379	75396072	5	6.95×10^{-5}	0.87	1.22×10^{-5}	0.039	0.051
	EuBR07s16969079	15429939	7	4.01×10^{-5}	0.24	1.38×10^{-5}	4.66×10^{-4}	2.77×10^{-3}
	EuBR02s2712998	1260315	2	9.68×10^{-6}	0.025	2.09×10^{-5}	0.50	0.11
	EuBR05s40135536	39939948	5	0.84	1.42×10^{-5}	7.95×10^{-5}	0.66	0.27
VOL3 and VOL6	EuBR05s73512379	75396072	5	1.40×10^{-5}	0.9743	1.39×10^{-5}	0.0352	0.0376
	EuBR02s2712998	1260315	2	3.41×10^{-6}	0.0172	1.00×10^{-5}	0.4773	0.0849
	EuBR05s40135536	39939948	5	3.70×10^{-5}	6.29×10^{-6}	0.9303	0.6377	0.2492
DBH6 and HEI6	EuBR02s12845338	10920732	2	1.56×10^{-7}	9.97×10^{-4}	5.92×10^{-6}	0.0031	0.066
	EuBR03s43394028	44908926	3	4.26×10^{-6}	0.20053	1.54×10^{-6}	0.0016	0.0045
DBH6 and VOL6	EuBR01s116241	2688124	1	2.44×10^{-5}	0.7304	4.31×10^{-6}	0.0012	0.0014
	EuBR02s2712998	1260315	2	5.81×10^{-5}	0.6922	1.09×10^{-5}	0.1061	0.0966
SOL and TEX	EuBR06s19529730	20443484	6	1.56×10^{-6}	0.6076	2.66×10^{-7}	6.61×10^{-5}	0.0011
	EuBR06s52964694	56543212	6	1.40×10^{-5}	0.21	5.06×10^{-6}	0.00080	0.12
SGR and TEX	EuBR11s43922247	43248389	11	7.26×10^{-6}	2.17×10^{-6}	0.27	0.11	0.00085
	EuBR11s44284539	43610681	11	1.16×10^{-5}	1.88×10^{-6}	0.97	0.00087	0.0055
	EuBR03s16484895	16899795	3	1.26×10^{-5}	9.84×10^{-6}	0.86	5.60×10^{-5}	0.0023
SGR and SOL	EuBR08s49045121	63920596	8	6.88×10^{-6}	5.97×10^{-6}	0.070	0.014	2.33×10^{-5}
PCY and DBH3	EuBR04s12292131	12520123	4	1.23×10^{-5}	0.47	2.59×10^{-6}	0.00012	0.0037
PCY and ALT6	EuBR04s510275	510215	4	4.22×10^{-5}	0.90	7.21×10^{-6}	0.030	0.089
TSC and HEI6	EuBR09s38638406	38502871	9	5.16×10^{-5}	1.54×10^{-5}	0.30	0.11	0.0025
TSC and DBH6	EuBR01s28498846	33181710	1	1.00×10^{-6}	1.86×10^{-7}	0.51	0.0023	0.00013
	EuBR10s6097228	6097168	10	8.40×10^{-5}	1.48×10^{-5}	0.98	0.0036	0.0025
TSC and VOL6	EuBR01s28498846	33181710	1	5.21×10^{-6}	8.66×10^{-7}	0.58	0.90	0.60
	EuBR10s6097228	6097168	10	5.81×10^{-5}	1.02×10^{-5}	0.86	0.003	0.0008

Continue...

Traits 1 and 2	SNP(s)	Position	Ch.	p-values				
				Full	Interaction	Common	Trait 1	Trait 2
TSC and DBH3	EuBR06s44029558	46528420	6	3.16×10^{-5}	8.17×10^{-6}	0.36	0.0038	4.33×10^{-5}
TSC and VOL3	EuBR01s28498846	33181710	1	4.48×10^{-6}	8.46×10^{-7}	0.53	0.0021	0.00025
	EuBR10s6097228	6097168	10	1.55×10^{-5}	1.15×10^{-5}	0.60	0.0058	0.0051

Where: SNP: Single nucleotide polymorphism; Ch.: Chromosome; Diameter at breast height at 3 years; DBH6: Diameter at breast height at 6 years; VOL3: Volume at 3 years, VOL6: Volume at 6 years; HEI3: Height at 3 years; HEI6: Height at 6 years; Pure cellulose yield (PCY); Wood basic density (WBD); Syringyl/guaiacyl ratio (SGR); Soluble lignin (SOL); Total solid content (TSC); and Total extractives (TEX). *P-values* below the Bonferroni correction cutoff (5%) of 1.63×10^{-5} are highlighted in bold.

CHAPTER 2
IMPROVEMENT OF GENOMIC PREDICTION MODELS AND TRAINING SETS
FOR *Eucalyptus grandis* W. Hill BREEDING

ABSTRACT

Few studies have discussed methods to optimize genomic selection models for tree species. Using the genomic best linear unbiased prediction (GBLUP) model, we evaluated five different methods and tested seven different training set sizes (30, 40, 50, 60, 70, 80, and 90% of the total population). The study population was an open-pollinated seed orchard of *Eucalyptus grandis* consisting of 1,772 individuals genotyped using the Illumina Infinium EuCHIP60K chip. Individuals were evaluated for 12 different growth and wood quality traits. A sequential analysis was developed based on the best predictive ability of the previous model. The highest accuracy was observed when 80% of the population was used for the training set. Meanwhile, the inclusion of dominance variance was effective in increasing the predictive ability of GS models. Additionally, the mean of the prediction error variance statistic (PEV_{mean}) for the optimization of the training set methodology was able to select the best genotypes with high accuracy. Further, multi-trait analysis showed that using the entire set of information from growth traits offered the highest predictive ability values, indicating that growth traits can be used in GS models to increase the predictive ability for wood quality traits. The multi-trait method with optimization of the training set improved predictive accuracy of the models, further indicating the efficiency of both strategies in GS models for *E. grandis*. Our results offer tree breeding companies and research institutes optimization strategies to reduce phenotyping costs and increase the accuracy of GS models.

Key words: eucalypt; breeding; growth traits; wood quality traits; dominance effects; genomic selection

2.1 INTRODUCTION

Genomic selection (GS) is a genetic prediction method proposed by Meuwissen et al. (2001), which uses a phenotyped and a genotyped training set to predict the genomic estimated breeding value (GEBV) of a non-phenotyped validation set. The methodology considers information across the complete genome using the vast array of data acquired through the genotyping of individuals. Several studies have proven the superior performance of GS over traditional breeding methods to accelerate the breeding cycle, maintain genetic diversity, and increase selection gains for each cycle (CROSSA et al., 2017; FRISTICHE-NETO; AKDEMIR; JANNINK, 2018; CROS et al., 2019). The performance of GS models is commonly assessed by estimating predictive ability (PA), which is the correlation between the predicted GEBVs and the true breeding values. Nevertheless, it is important to identify new approaches to accelerate genetic gains, improve costs per unit of time (HESLOT; JANNINK; SORRELLS, 2015), and enhance predictive accuracy (DE LOS CAMPOS et al., 2013).

On the other hand, the application of GS models with a high selection intensity may result in low-performance genotypes, despite a high predictive ability (MENDONÇA et al., 2020). Considering that GS models can present different rankings of the best performing genotypes, the selection coincidence (SC) represents the efficiency of the GS model to identify the best genotypes using different selection intensities (SABADIN et al., 2021). Therefore, a well-designed training population may result in higher predictive ability values (BERRO et al., 2019; OU; LIAO, 2019). As such, it is necessary to develop optimized genomic selection methods to improve resource allocation and efficiently achieve selection gains (RIEDELSCHEIMER; MELCHINGER, 2013).

Studying dominance, which is the mean value of the comparison between homozygous and heterozygous genotypes, has become relevant in genomic selection studies for several species (VITEZICA; VARONA; LEGARRA, 2013). Not surprisingly, including the effects of dominance on GS models can increase the accuracy of the predictions and genetic responses (SUN et al., 2014; VARONA et al., 2018). For eucalypts, several studies have shown a strong influence of dominance on growth but a weak impact on wood quality traits (DENIS; BOUVET, 2013; RESENDE et al., 2017; PALUDETO et al., 2021). The effects of dominance and epistasis are important to consider for species with vegetative propagation, such as species of the *Eucalyptus*

genus, since they can be fixed in the population (MAKOUANZI et al., 2014). According to Varona et al. (2018), non-additive effects can also contribute to mate allocation in breeding programs and enhance non-additive variation in breeding schemes.

Several authors have found that the accuracy of genomic selection models is strongly correlated with the accurate definition of the training set (ISIDRO et al., 2015; NORMAN et al., 2018; ZHU et al., 2021). In general, large and randomly selected training set populations tend to increase the predictive ability since the models are trained using the majority of the variation in the population. On the other hand, studies have shown that an optimized small training set can be effective in maintaining predictive accuracy, with the potential to reduce costs associated with phenotyping in plant breeding programs (ISIDRO et al., 2015; AKDEMIR et al., 2019; BERRO et al., 2019; CROSSA et al., 2021; HE et al., 2022). In this sense, Akdemir et al. (2015) developed a genomic algorithm to select an optimized training set from a group of candidate individuals. These authors used the mean of the prediction error variance (PEVmean) (HENDERSON, 1975) and the mean of the coefficient of determination (CDmean) (LALOË; PHOCAS; MENISSIER, 1996) to assess the predictive reliability of GEBVs for individuals in the validation set.

The predicted error variance (PEV) is defined as the error variance of the estimated genotypic value of individuals (HENDERSON, 1975). However, a limitation in the use of the estimated PEV for training set optimization studies is the possibility of sampling individuals that are related, as it does not consider genetic variance within the population. Meanwhile, the CDmean is an estimate based on the square correlation between the observed and predicted genetic values (ISIDRO et al., 2015). Thus, CDmean may increase selection accuracy with different sample sizes, which can be represented as a relationship between the PEV and the genetic variance of individuals (RINCENT et al., 2012). The performance of CDmean in relation to PEVmean is related to the population structure and the covariance between genotypes of the training population (ISIDRO et al., 2015).

Genomic selection commonly uses single trait models; however, for correlated traits it is possible to increase the predictive accuracy through multi-trait analysis (FERNANDES et al., 2021). Furthermore, one of the main limitations in using univariate analyses is related to the evaluation of the complex interactions among phenotypic characteristics (FLORES; MORENO; CUBERO, 1998). Multi-trait (MT) analyses allow us to consider the association among several genes in phenotypic traits of interest by

using strongly correlated traits to improve the model's predictive capacity (ZHOU; STEPHENS, 2014). Overall, multi-trait methods are applied when the primary trait presents low heritability values, thus taking advantage of the heritability of secondary traits (WARD et al., 2019). Several studies have shown that multi-trait GS are superior to single trait methods when using highly correlated traits (BERRO et al., 2019; BUDHLAKOTI et al., 2019; LOZADA; CARTER, 2019). However, no information is available in the literature on analyzing the optimization of training sets (OTS) alongside a multi-trait strategy. Nevertheless, it is possible to use the optimized training set genotypes to improve the predictive accuracy of genomic selection.

To optimize genomic best linear unbiased prediction (GBLUP) models for *Eucalyptus grandis*, the most commonly planted tree species in the world, we studied the effectiveness of multiple methodologies to increase predictive ability and the accuracy of GS models. Thus, this study aims to: *i*) analyze the influence of additive and additive-dominance models on growth and wood quality traits; *ii*) compare the efficiency of two training set optimization strategies; *iii*) evaluate the influence of two cross-validation schemes on multi-trait GS models; and *iv*) test the most efficient methods from steps *i*, *ii*, and *iii* to generate an optimized prediction accuracy multi-trait model with training set optimization.

2.2 MATERIAL AND METHODS

2.2.1 Phenotypic data

Genomic prediction models were performed using 1,772 individuals from an open-pollinated seed orchard of *E. grandis* located in the municipality of São Miguel Arcanjo, São Paulo state, Brazil. The study population was established in September 2012 using a randomized complete block design with four blocks of 27 different families (treatments). Each treatment consisted of four plots with 20 genotypes each (5 plants per plot) and a spacing between plants of 3 × 2 m, in four rows of 5 plants each. The seed orchard contains selected individuals that originate from two regions in Australia (Coff's Harbour and Atherton).

Phenotypic information was collected for six growth and six wood quality traits at three and six years after planting. The growth traits were diameter at breast height in centimeters (DBH3 and DBH6), height in meters (HEI3 and HEI6), and volume in

cubic meters (VOL3 and VOL6). DBH and height at three and six years were used to estimate tree volume (VOL3 and VOL6) in cubic meters according to the formula (Equation 1) described by Schumacher and Hall (1933):

$$VOL = DBH^2 \times \frac{\pi}{40000} \times HEI \times f \quad (1)$$

where *DBH* is the diameter at breast height at three or six years (DBH3 or DBH6); *f* is the taper factor (assumed to be 0.45); and π is the ratio between the circumference and the diameter; *HEI* is the total height at three or six years (HEI3 or HEI6).

Wood quality traits were estimated at six years of age using near-infrared (NIR) spectroscopy. Non-destructive wood samples were collected at breast height using a 12 mm increment borer. The samples were stored at room temperature in a controlled atmosphere (50% moisture and temperature of 23 °C ± 2 °C). Subsequently, samples were ground using a Wiley mill to achieve a uniform particle size. The sawdust was then placed in the spectrophotometer sample chamber, calibrated with the equipment's SUZANO's internal reference standards. The estimated wood quality traits were percentage of pure cellulose yield (PCY), basic wood density in kilograms per cubic meter (BWD), Syrigil/Guayacil ratio (SGR), percentage of soluble lignin (SOL), total solid content (TSC), and percentage of total extractives (TEX).

2.2.2 Quality control and effective population size

SNP genotypes were obtained using the Illumina Infinium EuCHIP60K *Eucalyptus* SNP chip (SILVA-JUNIOR; FARIA; GRATTAPAGLIA, 2015). The first step of the quality control process was removing duplicate SNP markers. Then, genotypic data were filtered considering a call rate lower than 95% and monomorphic markers. The markers with a minor allele frequency (MAF) lower than or equal to 0.01, and those that showed high deviation from Hardy-Weinberg equilibrium (p-value < 1x10⁻⁶) were also excluded (GRANATO et al., 2018). The genotypes were coded as "0" and "2" for homozygotes and "1" for heterozygotes. These processes were performed using the SNPReady package in the R environment (GRANATO et al., 2018). Finally, the remaining filtered markers were subject to linkage disequilibrium (LD) pruning, removing markers with a pairwise *r*² greater than 0.99. This step was performed using

the SNPRelate package in R (ZHENG et al., 2012). After the quality control process, we estimated the effective population size (N_e) using the molecular linkage disequilibrium method (WAPLES; DO, 2008) as implemented in NeEstimator V2.1 (DO et al., 2014).

2.2.3 Phenotypic analysis

We estimated the restricted maximum likelihood/best linear unbiased prediction (REML/BLUP) for the 12 phenotypic traits using the breedR package in R (MUNOZ; RODRIGUEZ, 2014). For each trait, BLUPs were predicted separately according to the following mixed model (Equation 2):

$$Y_{ijk} = \mu + Xb_j + Zt_j + Zp_k + \varepsilon_{ij} \quad (2)$$

where μ is the average mean; b_j is the fixed effect of the j^{th} block; t_j is the fixed effect of the j^{th} progeny; P_k is the random effect of the j^{th} plot with $p \sim N(0, \sigma_p^2)$; and ε_{ij} is the residual error that represents the nongenetic effects. The matrices X and Z are the incidence matrices for the fixed and random effects, respectively. We then estimated the deregressed best linear unbiased prediction/predictor (dBLUP) to avoid shrinkage properties (HENDERSON, 1975) using the formula $\frac{\hat{g}}{r^2}$ (GARRICK; TAYLOR; FERNANDO, 2009), where: \hat{g} is the BLUP values and r^2 is the reliability, estimated as $1 - (PEV/\sigma_g^2)$, where PEV is the prediction error variance and σ_g^2 is the genotypic variance. Pearson correlation tests were then estimated using the BLUPs to verify the correlation between the 12 growth and wood quality traits. Correlation distributions were plotted using the metan package in R (OLIVOTO; LÚCIO, 2020), as follows (Equation 3):

$$r_{g(x,y)} = \frac{Cov_{g(x,y)}}{\sqrt{\sigma^2(x), \sigma^2(y)}} \quad (3)$$

where, $r_{g(x,y)}$ is the correlation between primary (x) and secondary traits (y); $Cov_{g(x,y)}$ is the covariance between trait x and y ; $\sigma^2(x)$ is the variance associated with trait x ; and $\sigma^2(y)$ is the variance associated with trait y .

2.2.4 Genomic prediction

2.2.4.1 Additive and additive-dominant models

The additive (G_a) and dominance (G_d) genomic kinship matrices were obtained according to VanRaden (2008): $G_a = \frac{Z_A Z_A^T}{2 \sum_1^{m_i} p_i(1-p_i)}$ and $G_d = \frac{Z_D Z_D^T}{4 \sum_1^{m_i} \{p_i(1-p_j)\}^2}$, where p_i is the frequency of an allele from locus i ; Z is an $n \times m$ matrix of marker incidence (n is the number of genotypes and m is the number of markers); and Z_A is a matrix coded as 0 for homozygote A_1A_1 , 1 for heterozygote A_1A_2 , and 2 for homozygote A_2A_2 . For the Z_D matrix, 0 was allocated to both homozygotes (A_1A_1 and A_2A_2) and 1 to the heterozygote (A_1A_2). The genomic kinship matrices (G_a and G_d) were estimated using the SNPReady package in R (GRANATO et al., 2018).

The first step of this study was to compare the efficiency of the additive (A) and additive-dominant (AD) genomic best linear unbiased prediction (GBLUP) models for growth and wood quality traits individually ($j = 1, \dots, 12$). The additive and additive-dominant GBLUP were used by fitting the following models (Equations 4 and 5):

$$y = X\beta + Z_a a + e \quad (4)$$

$$y = X\beta + Z_a a + Z_d d + e \quad (5)$$

where y is a vector of BLUP values of genotypes obtained from the single-trait model; β is a vector of fixed effects; a is a vector of additive genetic effects of the markers, where $a \sim N(0, \sigma_a^2 G_a)$ and G_a is the additive genomic kinship matrix; d is the vector of dominance effect, where $d \sim N(0, \sigma_d^2 G_d)$ and G_d is the dominance genomic kinship matrix; Z_a and Z_d are incidence matrices for a and d . Finally, e is the model residual, where $e \sim N(0, \sigma_e^2)$.

2.2.4.2 Predictive accuracies and genetic parameters

For the additive (A) and additive-dominant (AD) GBLUP models, we applied a repeated random sub-sampling cross-validation to investigate GS prediction accuracies with varying training set sizes. Individuals were randomly assigned to either the training or validation set. The GBLUP model was used as described above with the

training size defined at 30% (530), 40% (710), 50% (885), 60% (1,065), 70% (1,240), 80% (1,420), and 90% (1,595) of the total number of individuals. Variance components and genetic parameters were estimated using an average of 50 runs of repeated random sub-sampling cross validation for each training size. The variance components were estimated using the ASReml-R package (BUTLER et al., 2017).

Predictive abilities (PA) were calculated considering the Pearson correlation coefficient (ρ) between the genomic estimated breeding values (GEBV) and observed adjusted means from the validation set. The coincidence of selection (CS) was also computed for each TS size. The CS refers to the Pearson correlation coefficient (or percentage) of the shared set of individuals that would have been selected considering their adjusted means from the phenotypic analysis and their GEBV from the genomic prediction model under the same selection intensity. Furthermore, to verify the efficiency of GS methods to select the best genotypes, we considered five different selection intensities. The estimated efficiency of the models to select the best 20 (top 20), 30 (top 30), 50 (top 50), 100 (top 100), and 150 (top 150) individuals of the 1,772 genotypes that the GS model selected based on the ranking of adjusted means. To confirm genetic gains through GS, we estimated the selection gain (SG) as the percentage gain of the selected individuals beyond the population average for the different training sizes and methods, as (Equation 6):

$$SG = (\bar{X}_s - \bar{X}_0)h_a^2 \quad (6)$$

where, \bar{X}_s is the average of genotypes selected by the genomic selection model; \bar{X}_0 is the average of the original population; and h_a^2 is narrow-sense heritability. According to Jia and Jannink (2012), trait heritability of the predictive model is influenced by the genetic structure of the training set. Thus, after fitting each model, for each training size (30, 40, 50, 60, 70, 80, and 90%) and for each of the five different methods (additive, additive-dominant, optimization of the training set, multi-trait, and multi-trait with optimization of the training set), we estimated the narrow-sense heritability ($h_a^2 = \sigma_a^2 / \sigma_p^2$), where σ_a^2 represents the additive variance and σ_p^2 is the phenotypic variance. Broad-sense heritability ($h_g^2 = (\sigma_a^2 + \sigma_d^2) / \sigma_p^2$) was also estimated for the five different models, where σ_a^2 represents the additive variance, σ_d^2 represents the dominance variance, and σ_p^2 is the phenotypic variance.

2.2.4.3 Optimization of the training set with a genetic algorithm

The main goal of the optimizing the training set (OTS) analysis was to assess whether the OTS models can increase the predictive accuracy of the GBLUP model using smaller training set sizes. Thus, we compared the multi-trait results for two OTS statistics (coefficient of determination and prediction error variance) with repeated random sub-sampling validation. Defined by Laloë (1993), the coefficient of determination (CDmean) is the squared correlation between the true and predicted genetic values. The prediction error variance (PEVmean) is defined as the contrast between all selected genotypes and the mean of the population (RINCENT et al., 2012).

To select the most representative individuals and optimize the size of the training set, we compared the efficiency of the GS model considering the same training set sizes used in the previous analysis (30, 40, 50, 60, 70, 80, and 90%). We implemented two optimization algorithms (PEVmean and CDmean) using the STPGA package in R (AKDEMIR, 2017), repeated 10 times to minimize stochastic variation. The best training population for each of the seven different training set sizes was selected after 20 iterations of the genetic algorithm; other parameters in the function were set to default values. Then, using the predictive ability results from the random repeated subsampling, the best GBLUP method (A or AD) was selected to run the OTS model for all growth and wood quality traits using the two OTS algorithms (PEVmean and CDmean). Genetic parameters, predictive ability (PA), coincidence of selection (CS), and selection gain (SG) were then estimated for the two OTS algorithms. Afterward, the OTS analysis results were compared with a randomly selected training set (A or AD GBLUP), and the best OTS statistic was identified for use in the subsequent optimization process.

2.2.4.4 Multi-trait analysis

An additive-dominance GBLUP model was used for multi-trait (MT) genomic selection models ($j = 1, \dots, n$ traits) (Figure 1). The main goal of the MT analysis was to use growth traits to increase the accuracy of GS models for wood quality traits. The MT-GS and additive-dominant GBLUP models were used to estimate the genomic

estimated breeding values (GEBVs) considering a bivariate linear mixed model of two correlated traits (Equation 7), as follows:

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} X\beta_1 \\ X\beta_2 \end{bmatrix} + \begin{bmatrix} Z_a a & 0 \\ 0 & Z_a a \end{bmatrix} + \begin{bmatrix} Z_d d & 0 \\ 0 & Z_d d \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} \quad (7)$$

where y_1 and y_2 are the BLUP values for traits 1 (growth) and 2 (wood-quality), β_1 and β_2 are the vector of fixed effects; a is a vector of additive genetic effects of the individuals, where $a \sim N(0, \sigma_a^2 G_a)$; d is the vector of dominance effect, where $d \sim N(0, \sigma_d^2 G_d)$; Z_a and Z_d are incidence matrices for a and d . Further, e is the model residual, and it was considered that $e \sim N(0, \sigma_e^2)$.

Figure 1 - Illustration of the two multi-trait (MT) cross-validation schemes (MT-CV1 and MT-CV2) in genomic selection models for *Eucalyptus grandis*. Genotypic and phenotypic data were divided into training and validation sets. The green boxes represent the existing phenotypic information, white boxes represent absent phenotypic information, and red boxes represent the predicted phenotypic information.

Trait	Type	Training set	Validation set	CV-scheme
1 th	GWT	80%	-	MT-CV1
2 th	WQT	80%	20%	
1 th	GWT	80%	20%	MT-CV2
2 th	WQT	80%	20%	

MT-CV-1 = multi-trait cross-validation scheme 1; MT-CV-2 = multi-trait cross-validation scheme 2; 1 and 2 are the primary and secondary traits, respectively; GWT = Growth-trait; WQT = Wood quality trait.

For the multi-trait analysis, we used a 5-fold, 5-replicate cross-validation scheme to randomly select the training and validation sets using two different cross-validation strategies (MT-CV1 and MT-CV2) (Figure 1). In the first scheme (MT-CV1), we used 80% of the phenotypic information from the training set for primary (growth) and secondary (wood quality) traits and the remaining 20% of the validation set had no phenotypic information. For the second cross-validation scheme (MT-CV2), we used 100% of the information for the primary trait for both training and validation sets and 80% of phenotypic information of the secondary trait for the training set.

2.2.4.5 Multi-trait with optimization of the training set model

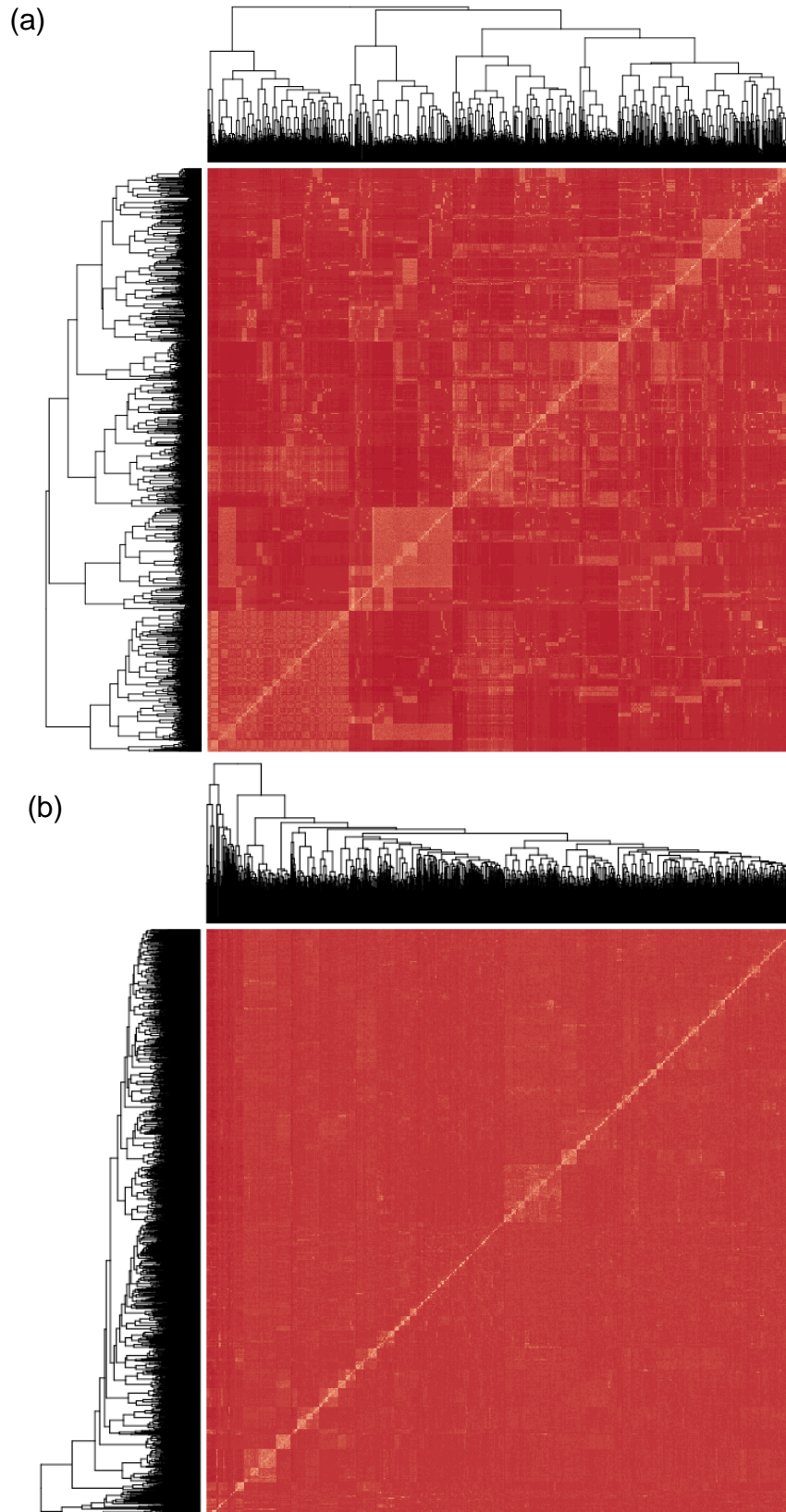
Subsequently, we performed a multi-trait analysis using the individuals selected by the training set optimization model (MT-OTS) to compare the performance of the optimization of the training set in a multi-trait genomic prediction scenario. The algorithm used for the MT-OTS analysis was selected considering the best predictive accuracy achieved in the previous step (multi-trait) as well as the best training set size based on predictive accuracy. The genetic parameters (PA, CS, SG, h_a^2 , and h_g^2) were also estimated for the MT-OTS analysis and compared with previous results (single-trait, OTS, and multi-trait).

2.3 RESULTS

2.3.1 Genotypic and phenotypic data

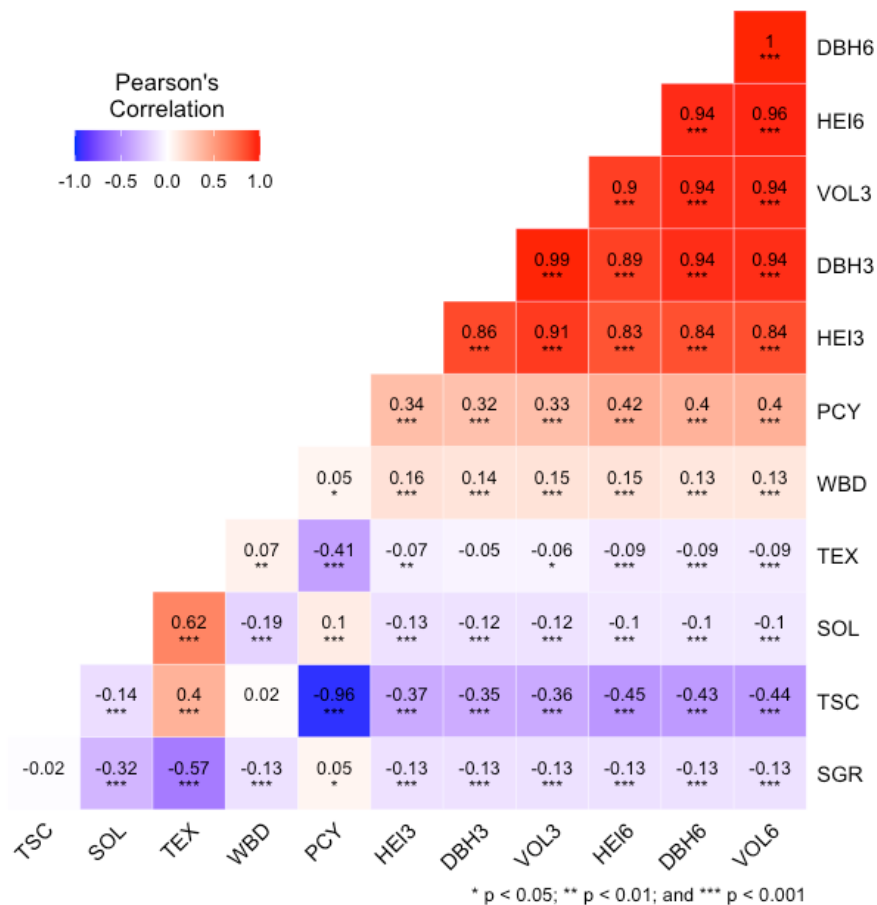
A total of 21,254 SNPs were used for the genomic selection analysis after quality control. We found an effective population size (N_e) of 31.5 considering linkage disequilibrium between markers (LDN_e). The heatmap showed that the strength of genomic relationships between the different genotypes is relatively low. Three large groups with genetically related individuals were identified, along with small clusters of more closely related individuals. Only a few genotypes showed kinship comparable to parent-offspring or full-sibs (Figure 2a). As expected for mixed-mating species, the dominance variance indicated a small contribution for phenotypic variation but with some clusters among genotypes. However, several blocks of genotypes were identified with high dominance patterns (Figure 2b).

Figure 2 - Heatmaps for the (a) additive and (b) dominance genetic variances of the 1,772 *Eucalyptus grandis* genotypes obtained following VanRaden (2008). Red indicates a limited relationship and yellow represents a high additive/dominance relationship between individuals



The genetic correlation indicates that growth traits are highly correlated, with values ranging from 0.83 (HEI3 and HEI6) to 1 (DBH6 and VOL6). Wood quality traits did not show a strong correlation, although there was a significant association for most traits (Figure 3). The highest positive association between wood quality traits was found between SOL and TEX (0.62). Further, a strong and negative correlation was found between traits PCY and TSC (-0.96). Only two trait combinations showed no significant Pearson’s correlation coefficient TEX and DBH3 (-0.05) and WBD and TSC (0.02).

Figure 3 - Pearson’s correlation coefficient of genetic values for 12 growth and wood quality traits evaluated for 1,772 *Eucalyptus grandis* genotypes from a open-pollinated seed orchard located in São Miguel Arcanjo, São Paulo, Brazil

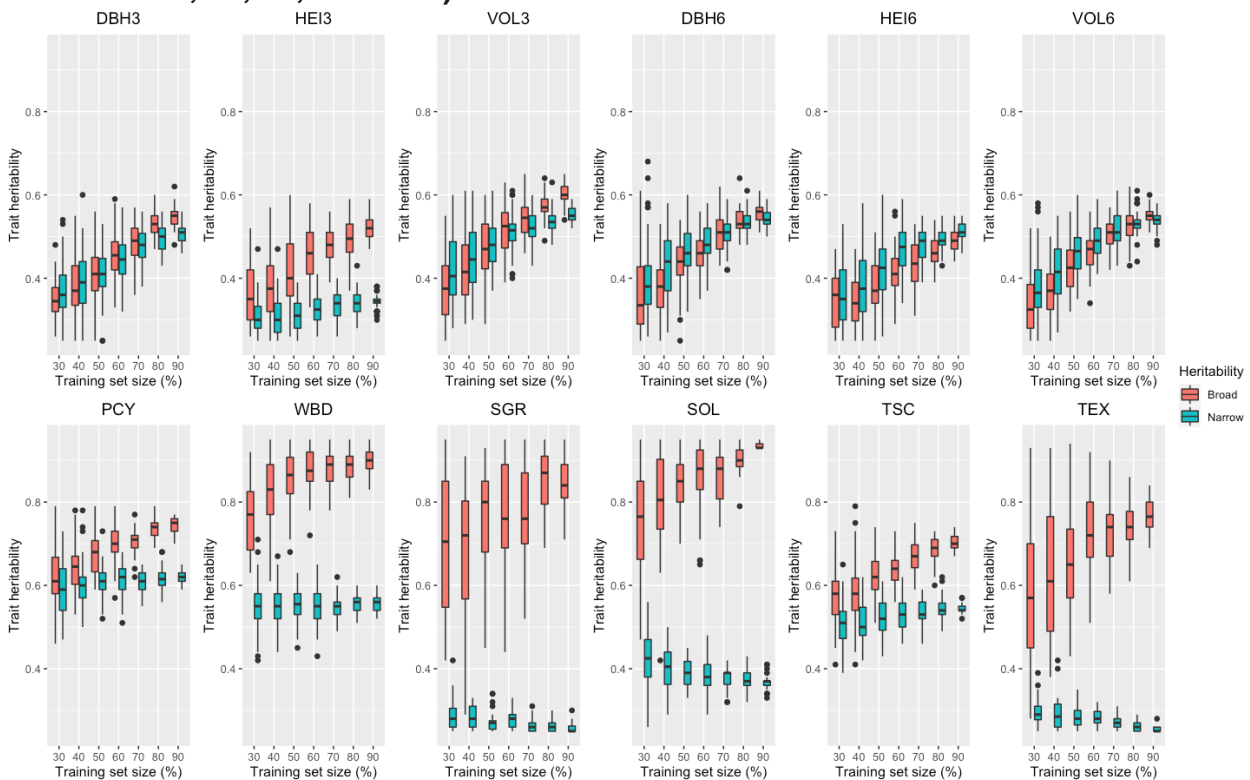


DBH3: Diameter at breast height (DBH) at 3 years; DBH6: DBH at 6 years; VOL3: Volume at 3 years, VOL6: Volume at 6 years; HEI3: Height at 3 years; HEI6: Height at 6 years; PCY: Pure cellulose yield; WBD: Basic wood density; SGR: Syringyl/guaiacyl ratio; SOL: Soluble lignin; TSC: Total solid content; and TEX :Total extractives.

2.3.2 Heritabilities

As expected, the narrow-sense heritability (h_a^2) tended to increase for all studied growth traits (DBH3, HEI3, VOL3, DBH6, HEI6, and VOL6) with an increase in the training set size (Figure 4). The same pattern was found for two wood quality traits (PCY and TSC) but was less pronounced than the others. On the other hand, three wood quality traits (WBD, SGR, and TEX) did not show the same trend of increased h_a^2 when the training set size increased. Only soluble lignin (SOL) decreased the narrow-sense heritability when the training set size increased considering the additive model. In general, traits that increased the h_a^2 also showed a reduction in standard deviation with an increase in the training set size. The same pattern was identified for narrow-sense heritability considering both the additive and additive-dominant models. This indicates that the genomic selection performed better with a larger training set.

Figure 4 - Broad-sense (pink) and narrow-sense (blue) heritabilities for the six growth (DBH3, HEI3, VOL3, DBH6, HEI6, VOL6) and six wood quality traits (PCY, WBD, SGR, SOL, TSC, and TEX) considering the additive-dominant models for the seven different training set sizes (30, 40, 50, 60, 70, 80, and 90%)



DBH3: DBH at 3 years; DBH6: DBH at 6 years; VOL3: Volume at 3 years, VOL6: Volume at 6 years; HEI3: Height at 3 years; HEI6: Height at 6 years; PCY: Pure cellulose yield; WBD: Basic wood density; SGR: Syringyl/guaiacyl ratio; SOL: Soluble lignin; TSC: Total solid content; and TEX :Total extractives.

Similarly, dominance did not significantly increase broad-sense heritability (h_g^2) for growth traits compared with narrow-sense heritability. Considering the height at three years (HEI3), dominance increased the trait heritability for GS models, which were more evident using larger training set sizes. Furthermore, for DBH at three years of age (DBH3), a high broad-sense heritability was found using the two largest training set sizes (80 and 90%). The results also showed a higher h_g^2 for the additive-dominant model for most wood quality traits, indicating that dominance has a greater influence on these traits than on growth traits. Comparing the narrow-sense and broad-sense heritabilities, the traits WBD, SGR, SOL, and TEX showed a strong influence of dominance. However, a high standard deviation for those traits was also found, indicating that h_g^2 is not representative, especially for smaller training set sizes.

In general, the standard deviation decreased when the size of the training set increased (Figure 5). We found a more evident increase in standard deviation for broad-sense heritability considering wood quality traits, which can be seen by the interquartile range of the boxplots, especially for WBD, SGR, SOL, and TEX. Also, traits PCY and TSC showed an increase in heritability related to dominance. However, the result was not as high as the others, and the effect on standard deviation was less evident. The remaining models (OTS, MT, and MT-OTS) presented similar broad- and narrow-sense heritabilities (Table S1).

2.3.3 Training set size

The predictive ability increased with an increase in the size of the training set (Figure 5). Although there was an effect on PA after increasing the size of the training set for all evaluated traits and categories (growth and wood quality), growth traits generally responded better to an increase in training set size compared to wood quality traits. This pattern may be related to the lower heritability levels for growth traits, which benefit more from an increase in training set size. The highest predictive ability among all growth and wood quality traits was achieved for VOL3 (0.52) and PCY (0.63), respectively. HEI3 (0.41) and TEX (0.32) showed the lowest predictive ability values for all analyzed traits. For DBH6, the predictive ability increased from 0.33 (30%) to 0.46 (90%) for the additive GBLUP model and from 0.42 (30%) to 0.51 (90%) for the additive dominant GBLUP model. Similarly, the predictive ability for TSC ranged from 0.48 (30%) to 0.55 (90%) and 0.51 (30%) to 0.57 (90%) for the additive and additive-

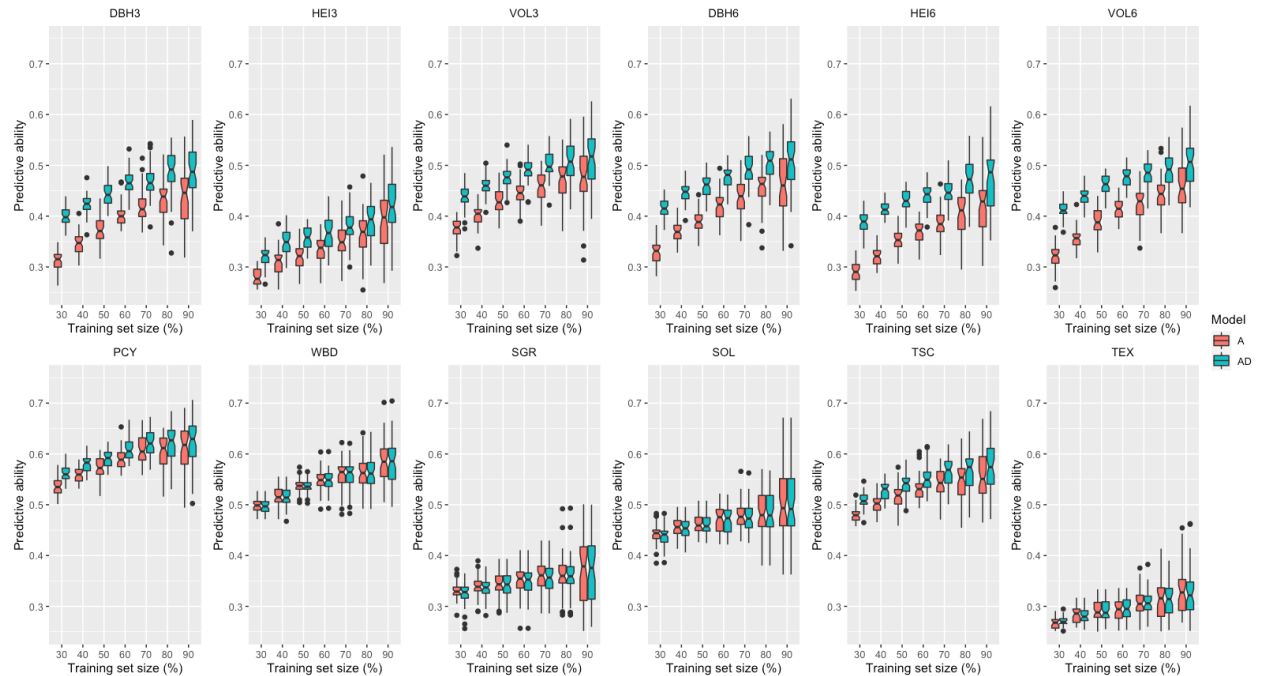
dominant models, respectively. Since we found no significant difference in predictive ability between training set sizes of 80% and 90% for most traits, we chose 80% as the best size to perform GS analyses to obtain the best predictive accuracies.

2.3.4 Predictive ability

2.3.4.1 Additive and additive-dominant model

The PA for the additive and additive-dominant GBLUP models indicated that the additive-dominant model performed better, suggesting that dominance increased the ability to predict phenotypes. Specifically, we observed that growth traits were more heavily influenced by dominance than wood quality traits (Figure 5), while the additive-dominant model did not increase PA for four wood quality traits (WBD, SGR, SOL, and TEX). For DBH6, considering the selected training set size (80%), the predictive ability increased from 0.46 to 0.51 when comparing the additive (Model 3) and additive-dominant (Model 4) models, respectively. Similarly, for TSC, considering a training size of 80%, the PA increased from 0.51 (additive) to 0.56 (additive-dominant). Since most traits presented an increase in predictive ability using dominance, we developed the subsequent procedures (OTS, MT, and MT-OTS) using the additive-dominant model.

Figure 5 - Predictive ability of GBLUP model considering additive (pink) and additive-dominant (blue) variances. The x-axis represents different training set sizes, which were determined as 30, 40, 50, 60, 70, 80, and 90% of the total number of individuals.



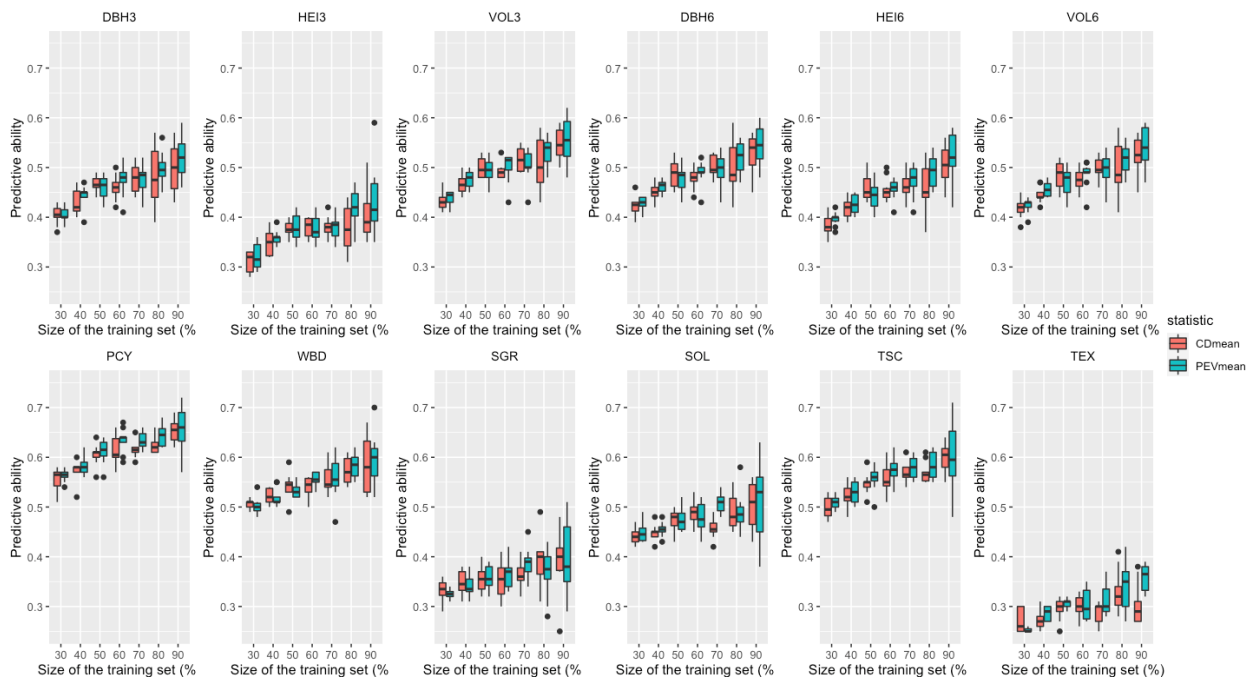
DBH3: DBH at 3 years; DBH6: DBH at 6 years; VOL3: Volume at 3 years, VOL6: Volume at 6 years; HEI3: Height at 3 years; HEI6: Height at 6 years; PCY: Pure cellulose yield; WBD: Basic wood density; SGR: Syringyl/guaiacyl ratio; SOL: Soluble lignin; TSC: Total solid content; and TEX :Total extractives.

2.3.4.2 Optimization of the training set

In general, the OTS methods CDmean and PEVmean performed well by selecting similar individuals between training and target sets, even when the training sizes were smaller (Figure 6). The PEVmean statistic increased the predictive ability of the open-pollinated eucalypt population when compared with random sampling. However, the CDmean statistic did not enhance genomic prediction since most PA values were lower than the values for the AD model. As a result, an increase in the predictive ability was only found for soluble lignin, which increased from 0.4829 (AD) to 0.5340 (CDmean). A comparison between the two statistics across the seven different training set sizes indicate that, for most traits, the predictive accuracy using PEVmean was generally higher than CDmean. Similarly, the PEVmean statistic was effective in predicting genetic values since the PA values increased for most traits. Only traits SGR (0.3530) and TEX (0.2999) showed slight decreases in PA when

comparing the PEVmean with the AD method (0.3604 and 0.3053, for SGR and TEX, respectively).

Figure 6 - Predictive ability for optimizing the training set model for *Eucalyptus grandis* using the CDmean (pink) and PEVmean (blue) metrics. The x-axis represents different sizes of the training set



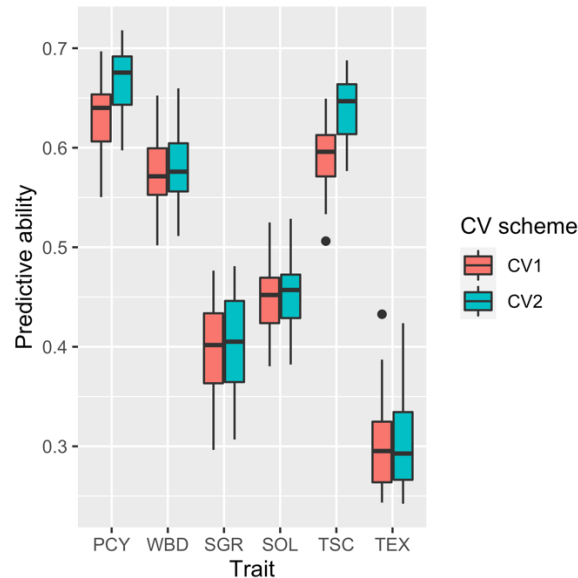
DBH3: DBH at 3 years; DBH6: DBH at 6 years; VOL3: Volume at 3 years, VOL6: Volume at 6 years; HEI3: Height at 3 years; HEI6: Height at 6 years; PCY: Pure cellulose yield; WBD: Basic wood density; SGR: Syringyl/guaiacyl ratio; SOL: Soluble lignin; TSC: Total solid content; and TEX: Total extractives.

2.3.4.3 Multi-trait genomic selection

The predictive ability using multi-trait models for wood quality traits was similar to the single trait model considering cross-validation scheme 1 (MT-CV1). Regarding MT-CV2, we found that an increase in PA was related to the intensity of the correlation (positive or negative) between traits. Traits PCY and TSC had the highest positive (0.32) and negative correlation (-0.37), respectively, with growth trait DBH3. The highest PA values for MT analysis were found with cross-validation scheme 2 (MT-CV2) (Figure 7). Traits PCY (0.66; MT-CV2) and TSC (0.64; MT-CV2) showed the greatest increase in predictive ability when comparing the two cross-validation schemes, while the predictive ability for three traits (SGR, SOL, and TEX) remained the same for the two schemes. However, the predictive ability for traits SGR (0.400) and TEX (0.450) increased when compared to both the additive-dominant method

(SGR: 0.360; TEX: 0.303) and the PEVmean (SGR: 0.351; TEX: 0.299). Both SGR and TEX showed the lowest predictive ability for wood quality traits. These results indicate the effectiveness of using growth traits to predict wood quality traits.

Figure 7 - Predictive ability for 1,772 *Eucalyptus grandis* genotypes using the GBLUP models with multi-trait analysis considering the two cross-validation schemes (CV1 and CV2)



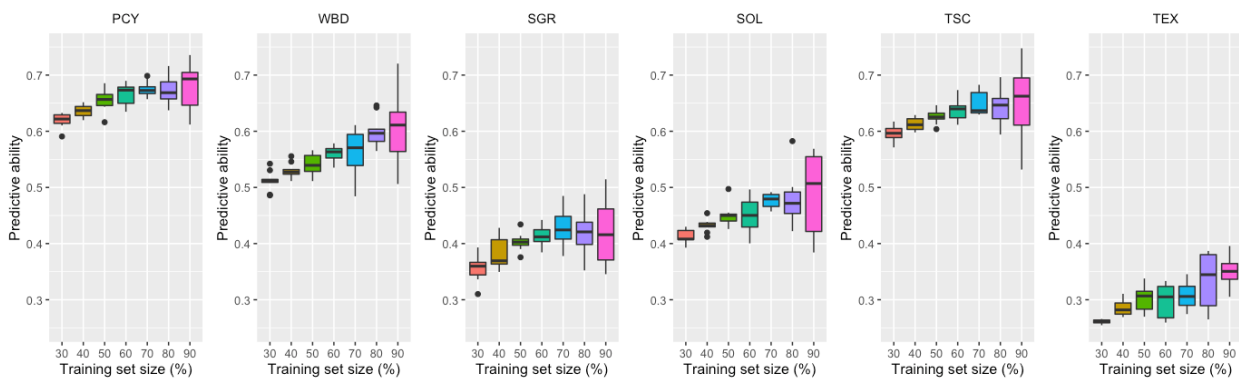
PCY: Pure cellulose yield; WBD: Basic wood density; SGR: Syringyl/guaiacyl ratio; SOL: Soluble lignin; TSC: Total solid content; and TEX: Total extractives.

Generally, the predictive ability for the MT-OTS method increased for three traits that showed the highest heritabilities (PCY, WBD, and TSC) compared to the multi-trait method (MT). Furthermore, these traits show the highest positive and negative Pearson correlation coefficients (0.32, 0.14, and -0.35, respectively) with the growth trait DBH3. As a result, the MT-OTS method did not increase the PA for traits with low heritabilities or weak Pearson correlation coefficients. In terms of SGR, we found a slight increase when compared with the MT method. However, compared with the additive-dominant method, the PA for SGR values showed an increase from 0.3604 (AD) to 0.4193 (MT-OTS). The traits soluble lignin and total extractives showed no increase in PA when compared with the MT method.

Thus, the MT-OTS method also indicates that it is possible to optimize the number of phenotyped individuals and that growth traits can be used to increase the predictive ability of wood quality traits. However, this process is more efficient with traits that have high heritability levels. Additionally, we found that some traits (PCY,

SGR, and TSC) achieve their highest predictive abilities when a training size of around 60% of the total population is used. These results indicate that the optimization of the training set using multi-trait analysis may also be effective in optimizing training set size, suggesting that resources in tree breeding programs can be efficiently allocated.

Figure 8 - Predictive ability of GBLUP model using a multi-trait model considering the optimization of the training set (PEVmean) for a *Eucalyptus grandis* breeding population. The x-axis represents different sizes of the training set



PCY: Pure cellulose yield; WBD: Basic wood density; SGR: Syringyl/guaiacyl ratio; SOL: Soluble lignin; TSC: Total solid content; and TEX: Total extractives.

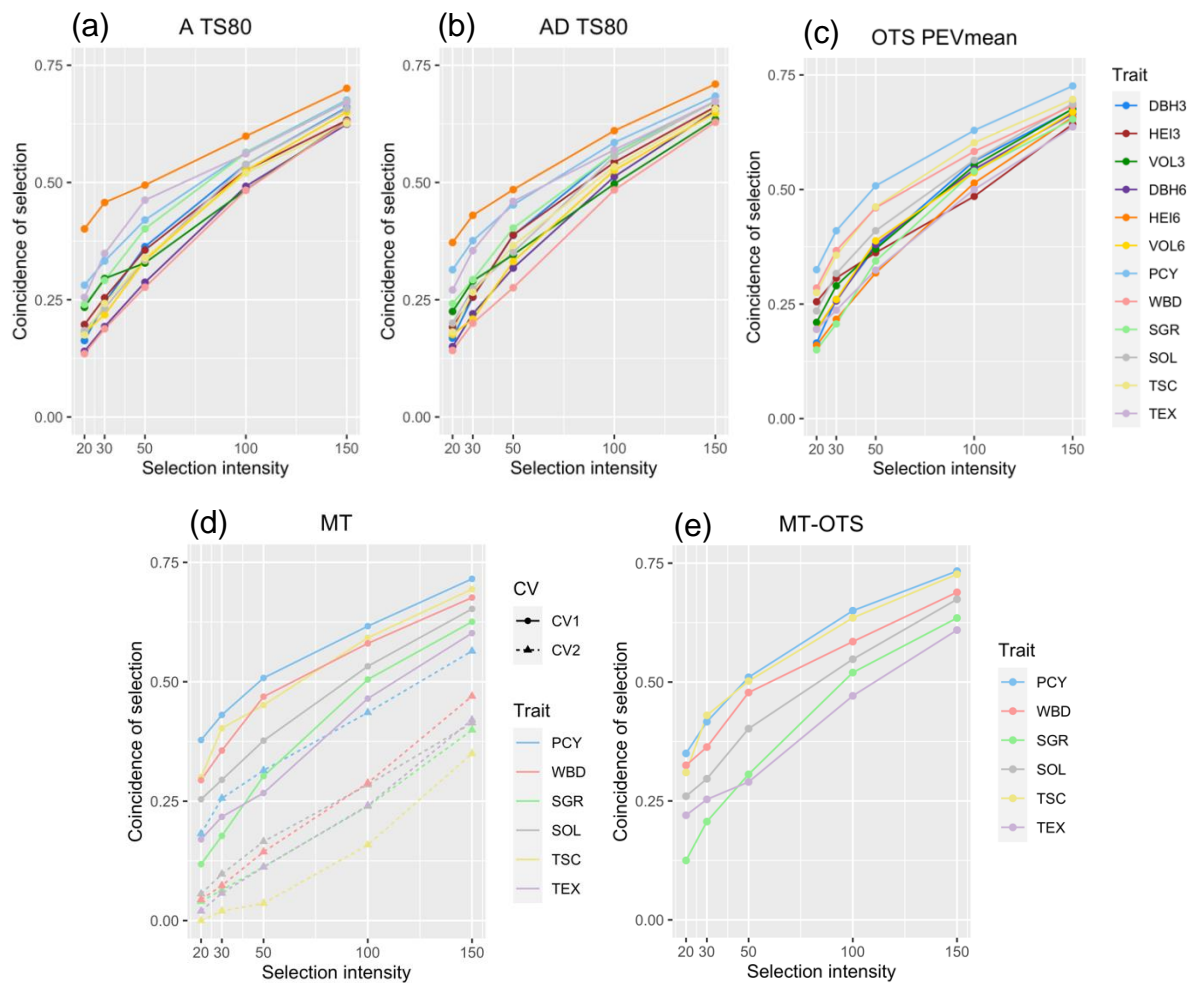
2.3.4.4 Selection coincidence and predicted selection gains

The selection coincidence increased substantially with an increase in selection intensity (Figure 9). Following from the results among all traits analyzed, we found that GS models have a weak ability to select the best individuals (e.g., top 20) with selection coincidence values ranging from 0 (TSC; MT-CV2) to 0.40 (HEI6; Additive). Nevertheless, considering all methods analyzed, GS was efficient in discarding progenies with low potential when using high selection intensity (e.g., top 150), with values ranging from 0.35 (TSC; MT-CV2) to 0.73 (PCY; MT-OTS). We assessed the performance of all traits and training set sizes for all studied methods.

The inability of GS to select the best genotypes using a high intensity was evident in MT-CV2 since almost all traits obtained a selection coincidence of only about 5% for the top 20 (1.13% of the total population). In general, we found was minimal difference in selection coincidence values when comparing the additive and additive-dominant models. Similarly, coincidence values were very similar between MT and MT-OTS. Regarding the A and AD methods (Figures 9a and 9b), the highest coincidence

of selection was found for trait HEI6, and the lowest was found for wood basic density. For the OTS PEVmean method, the trait DBH3 obtained the highest coincidence values, and trait HEI6 presented some of the lowest values (Figure 9c). Conversely, considering the two cross-validation schemes (CV1 and CV2) for MT, CV1 achieved a higher selection coincidence when compared with CV2 (Figure 9a). For multi-trait and multi-trait with OTS (Figure 9d), PCY and TSC presented the highest selection coincidence values. Additionally, trait TSC showed one of the highest coincidence considering CV1, but the lowest values for all training set sizes using CV2.

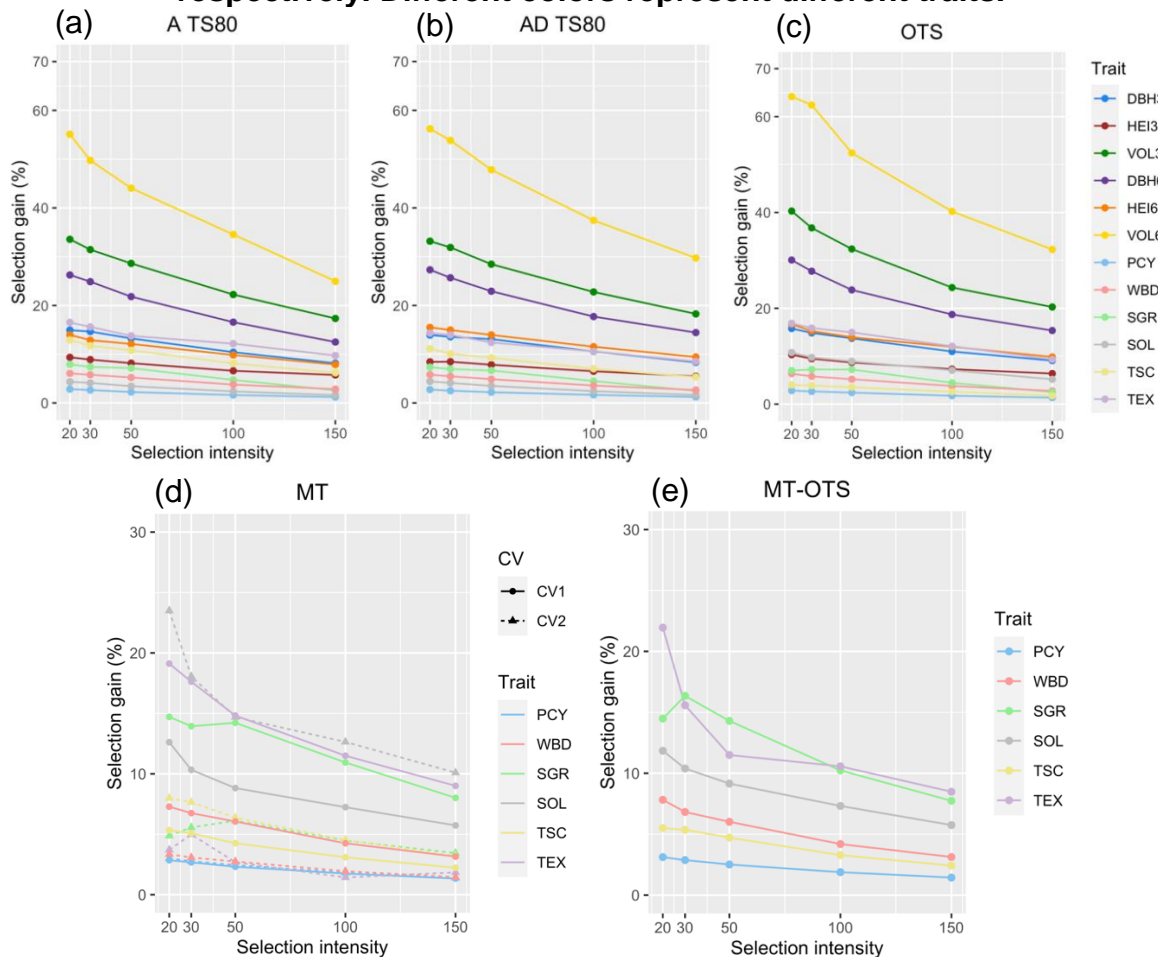
Figure 9 - The coincidence of selection for traits considering the (a) additive, (b) additive-dominant, (c) OTS, (d) multi-trait, and (e) multi-trait with OTS models for growth and wood quality traits in *Eucalyptus grandis*. For the multi-trait analysis, circles and triangles represent cross-validation schemes 1 (CV1) and 2 (CV2), respectively. Different colors represent different traits.



DBH3: DBH at 3 years; DBH6: DBH at 6 years; VOL3: Volume at 3 years, VOL6: Volume at 6 years; HEI3: Height at 3 years; HEI6: Height at 6 years; PCY: Pure cellulose yield; WBD: Basic wood density; SGR: Syringyl/guaiacyl ratio; SOL: Soluble lignin; TSC: Total solid content; and TEX: Total extractives.

Although a low coincidence was found for high selection intensity scenarios, most of the selection gains were positive (greater than 2%), indicating clear gains achieved by the GS models (Figure 10). Generally, restrictive scenarios presented the highest values, and low-intensity selection scenarios presented the lowest SG. For A (training set of 80%), AD (training set of 80%), and OTS models, growth traits VOL6, VOL3, and DBH3 showed the greatest gains, while PCY presented the lowest selection gains (Figure 10a, 10b, and 10c).

Figure 10 - Selection gains for traits considering the (a) additive, (b) additive-dominant, (c) Optimization of the Training Set (OTS), (d) multi-trait, and, (e) multi-trait with OTS models for growth and wood quality traits in *Eucalyptus grandis*. For the multi-trait models, circles and triangles represent cross-validation schemes 1 (CV1) and 2 (CV2), respectively. Different colors represent different traits.



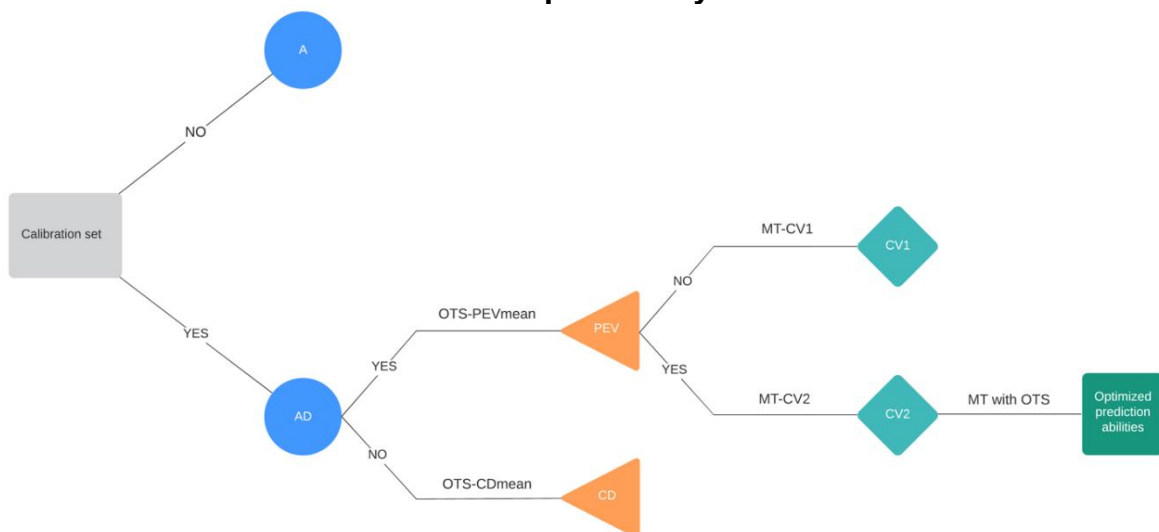
DBH3: DBH at 3 years; DBH6: DBH at 6 years; VOL3: Volume at 3 years, VOL6: Volume at 6 years; HEI3: Height at 3 years; HEI6: Height at 6 years; PCY: Pure cellulose yield; WBD: Basic wood density; SGR: Syringyl/guaiacyl ratio; SOL: Soluble lignin; TSC: Total solid content; and TEX: Total extractives.

For MT-CV2, TEX did not show a positive selection gain across different TS sizes. Regarding multi-trait and MT-OTS, the traits TEX and SGR showed a greater selection gain, and trait PCY had the lowest selection gain. Specifically, the MT method showed high selection gains for traits SGR, SOL, and TEX. There was no clear differentiation between cross-validation schemes.

2.3.5 Workflow analysis

Considering our results, we propose a flowchart based on increases in predictive ability (Figure 11). All growth and several wood quality traits showed an increase in PA due to dominance. The PA for several wood quality traits also increased, but none showed a reduction in predictive ability using dominance variance. Although the two OTS statistics (PEVmean and CDmean) were effective in predicting the genotypic value, some traits had higher PA values when using PEVmean. In addition, we found that some traits showed that MT-CV2 was more efficient than MT-CV1 in increasing PA. MT-OTS can be considered the best method to improve predictive ability. Therefore, based on the model's predictive ability, we suggest the following workflow: Additive-dominant, OTS-PEVmean, and MT-CV2.

Figure 11 - Flowchart of optimization of genomic prediction methods for *Eucalyptus grandis*. In green, the methods with the highest predictive abilities; orange represents methods that were tested but not selected for subsequent analyses



A = Additive GBLUP model; AD = Additive-dominant GBLUP model; PEV = Mean prediction error variance (PEVmean); CD = Mean coefficient of determination (CDmean); CV1 = Cross-validation scheme 1; CV2 = Cross-validation scheme 2.

2.4 DISCUSSION

Due to the time required for breeding cycles and the extensive costs associated with phenotypic and genotypic data acquisition, it is desirable to optimize genomic selection models to reduce costs and breeding program cycles (RIEDELSEIMER; MELCHINGER, 2013; HESLOT; JANNINK; SORRELLS, 2015). Furthermore, increases in the predictive accuracy of GS models can be enhanced by developing methodologies that consider non-additive and other effects (VITEZICA; VARONA; LEGARRA, 2013; SUN et al., 2014; VARONA et al., 2018), size optimization, selection of genotypes to be used in the training set (AKDEMIR; SANCHEZ; JANNINK, 2015; BERRO et al., 2019; ISIDRO et al., 2015), and multi-trait analyses (JIA; JANNINK, 2012; RAMBOLARIMANANA et al., 2018; LENZ et al., 2020).

According to Tan et al. (2017), the power of predictive accuracy in GS is mainly related to the composition and size of the training set. Here, we found that *Eucalyptus* breeding programs can be optimized by using GS models that have previously analyzed genetic parameters, such as predictive ability, selection coincidence, and selection gains. As a result, to optimize the resources used for phenotyping, it is recommended to estimate the accuracy of a prediction scenario before collecting the phenotypic information since it is possible to optimize the training set (RINCENT; CHARCOSSET; MOREAU, 2017). The efficiency of GS is mainly influenced by factors related to trait heritability (CALUS et al., 2008; RESENDE et al., 2012), population structure (ISIDRO et al., 2015; NORMAN et al., 2018; WERNER et al., 2020), kinship among genotypes (HESLOT et al., 2012; SCUTARI; MACKAY; BALDING, 2013), the method applied (HESLOT et al., 2012; WANG et al., 2018), and training set composition (AKDEMIR; SANCHEZ; JANNINK, 2015; ISIDRO et al., 2015; RIEDELSEIMER; MELCHINGER, 2013). Of these, the composition and size of the training set and the methodology and statistical model used can help to optimize GS predictive accuracy.

According to Grattapaglia et al. (2018), the training set size has a distinct effect on GS accuracy and should be large enough to guarantee good predictive abilities. However, our results suggest that most growth traits have a less significant impact on the predictive accuracy of a population in GS (Figure 5). Similar to the results of previous studies (ZHAO et al., 2012; NORMAN et al., 2018; OU; LIAO, 2019), a training set size of 80% of the total population was sufficient to enhance the most

powerful predictive accuracies of GS models for all methods analyzed. When selecting genotypes using dominance and the OTS algorithm, the predictive accuracy of smaller training set sizes increased, offering a possible opportunity to optimize the genotyping step of a breeding program. Indeed, GS is influenced by kinship and population structure (BASTIAANSEN et al., 2012; ISIDRO et al., 2015; WERNER et al., 2020).

Our results indicate that the AD model is the most effective for predicting growth traits. Considering the results found by Thumma et al. (2022) for a progeny test of *Eucalyptus nitens* (H. Deane & Maiden) Maiden, significant inbreeding depression was found for the traits DBH and kraft pulp yield due to dominance and epistasis, respectively. Thus, these authors conclude that it is possible to use the dominance effect to select the best parents and improve selection gains. Additionally, it is important to include dominance in GS models for *Eucalyptus* breeding since all variance (additive and dominance) is captured by vegetative propagation. Furthermore, dominance increased the heritability of wood quality traits, such as the SGR and TEX, but we found no evident increase in the PA of these traits, which may be related to an overestimation of the variances in GS models (JIA, 2017).

According to Rezende et al. (2014), the genetic control of growth and wood quality traits is mainly additive, but there is an influence of dominance on growth traits. Previous studies have found a more pronounced dominance effect on growth variables (DENIS; BOUVET, 2013; TAN et al., 2018), but less so for wood quality traits (PALUDETO et al., 2021; RESENDE et al., 2017). On the other hand, among all traits evaluated in our study (growth and wood quality), predictive abilities were generally higher for two wood quality traits (PCY and TSC). As expected, these traits present both the most positive and negative genetic correlations (Figure 3) and were also the only two wood quality traits that showed an increase in predictive ability when adding dominance (Figure 5).

Our study shows that identifying a suitable subset of individuals for phenotyping improved the predictive accuracy of GS models for a *Eucalyptus* breeding program. Defining the training set and its size is a trade-off between model accuracy and the resources available for phenotyping (OU; LIAO, 2019). Herein, the optimization of the training set demonstrated that both CDmean and PEVmean statistics were effective in selecting the best genotypes to be used as the training set in different training populations. However, consistent with previous studies, PEVmean achieved better predictive abilities than CDmean and random sampling (RINCENT et al., 2012;

KADAM; RODRIGUEZ; LORENZ, 2021). Nevertheless, several studies have found that CDmean outperformed PEVmean (ISIDRO et al., 2015; ZHANG et al., 2021). Thus, the efficiency of each OTS statistic should be evaluated for each scenario.

The CDmean and PEVmean have been successfully tested in a range of different species including maize, wheat, and oat, with populations presenting several levels of kinship and population structures (ASORO et al., 2011; KADAM; RODRIGUEZ; LORENZ, 2021; RINCENT et al., 2012; SARINELLI et al., 2019; ZHANG et al., 2021). However, to the best of our knowledge, no similar study has compared the effectiveness of using the two criteria for eucalypts. Isidro et al. (2015) reported that CDmean showed weak performance in highly structured populations. Here, the impact of the genotypic information and the population structure was not considered when applying the optimization criteria, which may explain the inferior performance of CDmean compared to PEVmean.

Further studies should be done to evaluate the efficiency of OTS in highly structured populations for complex traits to predict the effectiveness of GS and optimize the training set according to the targeted breeding population. Thus, the application of this methodology is indicated for use in GS models, especially for wood quality traits, which are expensive and difficult to collect. Further, our testing of OTS with smaller training sizes indicates that, for traits that are costly to analyze, we could obtain similar predictive accuracies by randomly selecting the TS, which may be an interesting option for breeding programs. Indeed, we have shown that the models select the most representative genotypes to be used as the training set. However, even when using OTS, for very small training set sizes (e.g., 30%), all models (A, AD, MT, and MT-OTS) were inefficient in obtaining high accuracy since a small number of genotypes cannot effectively represent the whole population. On the other hand, considering medium training set sizes (e.g., 60 and 70%), the OTS, MT, and MT-OTS methods effectively increased and optimized the predictive ability and selection gains in GS.

Understanding the correlation among traits is important to develop accurate multi-trait predictions (MONTESINOS-LÓPEZ et al., 2018; WARD et al., 2019). Apart from the increase in DBH, a significant increase in height and volume are determined by plant growth. Previous studies reported positive and negative correlations between growth (OSORIO; WHITE; HUBER, 2003; WEI; BORRALHO, 1998) and wood quality traits (GALLO et al., 2018; MPHAHLELE et al., 2020), respectively, in *Eucalyptus*

species. Meanwhile, research has also indicated positive correlations between growth traits and cellulose yield (PCY) and basic wood density (WBD) (WU et al., 2011). However, a negative correlation between growth and wood quality traits indicates possible negative effects when selecting for fast-growth and ideal cellulose pulp production. Strong negative correlation patterns between total extractives and pulp production with total solid content are expected since they are inversely proportional (RAYMOND, 2002). According to Silva et al. (2020), the strong negative correlation between PCY and TSC is expected because when cellulose production is increased, the total solid content decreases.

Multi-trait analysis showed the effectiveness of using growth traits to increase the predictive ability of wood quality traits compared to a single trait model. However, the cross-validation schemes showed a weak improvement in predictive ability for wood quality traits when only phenotypic information was included in the training set (MT-CV1). A similar result was found by Arojju et al. (2020) when analyzing the efficiency of multi-trait models in perennial ryegrass, where the predictive ability using MT-CV1 was comparable to the single-trait model. Similarly, Gill et al. (2021) found no improvement in predictive ability for MT-CV1 in relation to single trait analysis considering five wheat traits in the different growing seasons evaluated.

The increase in predictive ability resulting from MT-CV2 might be related to the high correlation between traits and the amount of information used to predict the wood quality traits. According to Rambolarimanana et al. (2018), the higher the correlation between traits, the greater the efficiency of the multi-trait approach. Additionally, the efficiency of multi-trait models is related to high levels of heritability between primary and secondary traits. Moreover, according to Jia and Jannink (2012), the predictive ability of GS models using multi-trait analysis can be increased when the heritability of the primary trait is high. Thus, using all information from growth traits can increase the predictive ability for wood quality traits.

Including the phenotypic information of the secondary trait in both training and validation sets considerably increased the predictive ability for two important wood quality traits (PCY and TSC). Not surprisingly, we found that these traits are most negatively correlated (-0.96), indicating that they are inversely proportional, as noted above. In addition, the narrow-sense heritabilities for PCY and TSC are some of the highest for wood quality traits, even though WBD also presents high narrow-sense heritability. However, we believe that the high levels of correlation between TSC (-

0.35), PCY (0.33), and WBD (0.14) with growth trait DBH3 could explain the limited differentiation in predictive ability between CV1 and CV2 for this trait. Consequently, the weak improvement in predictive ability using multi-trait analysis for SGR, SOL, and TEX are likely related to the weak genetic correlation with DBH3 (-0.13, -0.12, and 0.05, respectively), which presented the lowest correlations among all analyzed traits. Another explanation for the limited increase in predictive ability might be related to the methods used to analyze wood quality traits. According to Schwanninger et al. (2011), near-infrared (NIR) spectroscopy models might perform poorly for complex molecules such as wood, since these models represent the correlation with atom vibration, or chemical properties, which may be altered from its natural state.

Cellulose pulp yield is an important wood quality trait for tree species since it is directly correlated with the manufacturing process for cellulose production (RAYMOND, 2002; KIEN et al., 2009). Similarly, the total solid content is also important since its presence or absence directly impacts the number of chemicals used in the cellulose pulp production process (DUTT; TYAGI, 2011), which has an impact on production costs. Thus, the development of multi-trait analysis using phenotypic information for growth traits, such as DBH3, can be effective in optimizing the predictive ability of GS models, potentially reducing the costs associated with breeding programs.

Finally, multi-trait analysis using the optimization of the training set (MT-OTS) also showed an increase in predictive ability estimates. Similar to the results for the random multi-trait model, PCY and TSC demonstrated the highest predictive accuracies (0.672 and 0.644, respectively). Thus, the high heritability of those traits and their genetic correlation might have the same effect on the MT-OTS models. Compared with the additive model, these traits showed the highest increase for predictive ability (7.2% and 9.7% for PCY and TSC, respectively). To the best of our knowledge, this is the first analysis involving both multi-trait and OTS models. It is also important to note that predictive ability for those traits reached a plateau with a TS size of 60%. These results indicate that MT-OTS can increase predictive ability and improve resource allocation in breeding programs since phenotyping can include a smaller number of individuals.

Most studies involving genomic selection have analyzed the success of selection considering only the correlation between the predicted and observed genotypes (HE et al., 2016). However, when only the predictive ability is analyzed, the result can produce several different rankings which do not necessarily identify the best

individuals. As with traditional selection strategies based only on phenotype, GS should focus on accurately ranking the best genotypes (BLONDEL et al., 2015) since high selection coincidence can provide a more accurate selection. Several studies have shown that predictive ability is not necessarily strictly correlated with CS (MENDONÇA; FRITSCHÉ-NETO, 2020; SABADIN et al., 2021). Blondel et al. (2015) showed that the same predictive ability could be obtained from different rankings. Thus, looking only at the CS, our study shows that a high selection intensity can lead to the inclusion of genotypes with low performance, indicating that the success of GS in *Eucalyptus* is improved with a high selection percentage. Although a high selection intensity may lead to the selection of low performance genetic materials, it also guarantees that the best genotypes are not discarded (MENDONÇA et al., 2020).

Although we found a relatively low selection coincidence, the selection gains for all methods and training set sizes were positive, demonstrating that even though the best progenies were not selected with low selection intensity, genomic selection showed positive gains. According to Heffner et al. (2010), moderate selection intensity can drastically increase genetic gains since there is a shorter breeding cycle when compared to traditional selection. Mendonça et al. (2020) offered two options for breeding using genomic selection: first, the impact of high selection intensity on genomic selection could be explored using a small number of plots, which could lead to the same selection gains; or second, the same number of plots is maintained, thus exploiting all variation from the population, which could increase selection gains. We suggest that different fields of research, such as microbiomics, metabolomics, and enviromics, should be combined with genomic evaluation to better predict genotype performance. Over time, new optimization methodologies will offer an improved understanding of genetic and environmental effects on the performance of individuals and will contribute to the development of more accurate prediction models.

2.5 CONCLUSION

Herein, we provide insights on strategies to analytically optimize the accuracy and resource allocation in breeding programs using different genomic selection methods. Generally, the GBLUP model was effective in predicting individual breeding values using different training set sizes and methods. First, we showed that dominance variance appears to have more of an effect on growth than on wood quality traits. Then,

we found moderate to high predictive ability with the GS model able to obtain selection gains using the suggested optimization methods. Further, the selection coincidence was more efficient when selecting around 3 to 5% of the best individuals (top 100 and top 150). Concerning the optimization of the training set, both methods were effective in predicting individual breeding values, but with relatively greater accuracy using the PEVmean statistic. In general, the multi-trait analysis showed high predictive ability and selection gains, indicating that growth traits can be used to increase the model's capacity to predict wood quality traits. As expected, MT-OTS achieved better results for both predictive accuracy and selection gains. Also, size of the training set can be optimized according to the method applied. Thus, they should be used together in genomic selection models to better optimize and allocate resources in *Eucalyptus* breeding programs.

2.6 ACKNOWLEDGMENTS

We thank SUZANO S.A. for providing phenotypic and genotypic data and AgroPartners consulting for relevant comments and indispensable collaboration. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. We also acknowledge the German Academic Exchange Service (DAAD) for the co-financed short-term research grant (ref. no.: 91781916). Evandro V. Tambarussi is supported by a research productivity fellowship granted by CNPq (grant number 304899/2019-4).

REFERENCES

- AKDEMIR, D. STPGA: An R-package for selection of training populations with a genetic algorithm. **arXiv preprint**, 2017.
- AKDEMIR, D. et al. Multi-objective optimized genomic breeding strategies for sustainable food improvement. **Heredity**, v. 122, n. 5, p. 672–683, 2019.
- AKDEMIR, D.; SANCHEZ, J. I.; JANNINK, J.-L. Optimization of genomic selection training populations with a genetic algorithm. **Genetics Selection Evolution**, v. 47, n. 1, p. 1-10, 2015.
- AROJJU, S. K. et al. Multi-Trait genomic prediction improves predictive ability for dry matter yield and water-soluble carbohydrates in perennial ryegrass. **Frontiers in plant science**, v. 11, p. 1197, 2020.
- ASORO, F. G. et al. Accuracy and training population design for genomic selection on quantitative traits in elite North American oats. **The Plant Genome**, v. 4, n. 2, p. 132, 2011.
- BASTIAANSEN, J. W. M. et al. Long-term response to genomic selection: effects of estimation method and reference population structure for different genetic architectures. **Genetics Selection Evolution**, v. 44, n. 1, p. 1-13, 2012.
- BERRO, I. et al. Training population optimization for genomic selection. **The Plant Genome**, v. 12, n. 3, p. 1-14, 2019.
- BLONDEL, M. et al. A ranking approach to genomic selection. **PloS one**, v. 10, n. 6, 2015.
- BUDHLAKOTI, N. et al. A comparative study of single-trait and multi-trait genomic selection. **Journal of Computational Biology**, v. 26, n. 10, p. 1100-1112, 2019.
- BUTLER, D. G. et al. **ASReml-R reference manual version 4**. VSN International Ltd, Hemel Hempstead, HP1 1ES, UK, 2017. Disponível em: <<https://asreml.kb.vsnl.co.uk/wp-content/uploads/sites/3/ASReml-R-Reference-Manual-4.pdf>> . Acesso em 18 de fevereiro de 2022.
- CALUS, M. P. L. et al. Accuracy of genomic selection using different methods to define haplotypes. **Genetics**, v. 178, n. 1, p. 553-561, 2008.
- CROS, D. et al. Within-family genomic selection in rubber tree (*Hevea brasiliensis*) increases genetic gain for rubber production. **Industrial Crops and Products**, v. 138, p. 111464, 2019.
- CROSSA, J. et al. Genomic selection in plant breeding: methods, models, and perspectives. **Trends in plant science**, v. 22, n. 11, p. 961-975, 2017.
- CROSSA, J. et al. The modern plant breeding triangle: optimizing the use of genomics, phenomics, and enviromics data. **Frontiers in Plant Science**, v. 12, 2021.

DA SILVA, A. S. et al. Constraints and advances in high-solids enzymatic hydrolysis of lignocellulosic biomass: a critical review. **Biotechnology for biofuels**, v. 13, n. 1, p. 1-28, 2020.

DE LOS CAMPOS, G. et al. Whole-genome regression and prediction methods applied to plant and animal breeding. **Genetics**, v. 193, n. 2, p. 327-345, 2013.

DENIS, M.; BOUVET, J.-M. Efficiency of genomic selection with models including dominance effect in the context of *Eucalyptus* breeding. **Tree Genetics & Genomes**, v. 9, n. 1, p. 37-51, 2013.

DO, C. et al. NeEstimator v2: re-implementation of software for the estimation of contemporary effective population size (N_e) from genetic data. **Molecular ecology resources**, v. 14, n. 1, p. 209-214, 2014.

DUTT, D.; TYAGI, C. H. Comparison of various *Eucalyptus* species for their morphological, chemical, pulp and paper making characteristics. 2011.

FERNANDES, S. B. et al. How well can multivariate and univariate GWAS distinguish between true and spurious pleiotropy? **Frontiers in genetics**, v. 11, p. 1747, 2021.

FLORES, F.; MORENO, M. T.; CUBERO, J. I. A comparison of univariate and multivariate methods to analyze G×E interaction. **Field crops research**, v. 56, n. 3, p. 271-286, 1998.

FRISTICHE-NETO, R.; AKDEMIR, D.; JANNINK, J.-L. Accuracy of genomic selection to predict maize single-crosses obtained through different mating designs. **Theoretical and Applied Genetics**, v. 131, n. 5, p. 1153-1162, 2018.

GALLO, R. et al. Growth and wood quality traits in the genetic selection of potential *Eucalyptus dunnii* Maiden clones for pulp production. **Industrial Crops and Products**, v. 123, p. 434-441, 2018.

GARRICK, D. J.; TAYLOR, J. F.; FERNANDO, R. L. Deregressing estimated breeding values and weighting information for genomic regression analyses. **Genetics Selection Evolution**, v. 41, n. 1, p. 1-8, 2009.

GILL, H. S. et al. Multi-trait multi-environment genomic prediction of agronomic traits in advanced breeding lines of winter wheat. **Frontiers in Plant Science**, p. 1619, 2021.

GRANATO, I. S. et al. snpReady: a tool to assist breeders in genomic analysis. **Molecular Breeding**, v. 38, n. 8, p. 1-7. 2018.

GRATTAPAGLIA, D. et al. Quantitative genetics and genomics converge to accelerate forest tree breeding. **Frontiers in Plant Science**, v. 1693. 2018.

HE, S. et al. Genomic selection in a commercial winter wheat population. **Theoretical and applied genetics**, v. 129, n. 3, p. 641-651, 2016.

HE, S. et al. Genomic prediction using composite training sets is an effective method for exploiting germplasm conserved in rice gene banks. **The Crop Journal**, 2022.

HEFFNER, E. L. et al. Plant breeding with genomic selection: gain per unit time and cost. **Crop science**, v. 50, n. 5, p. 1681-1690, 2010.

HENDERSON, C. R. Best linear unbiased estimation and prediction under a selection model. **Biometrics**, p. 423-447, 1975.

HESLOT, N. et al. Genomic selection in plant breeding: a comparison of models. **Crop science**, v. 52, n. 1, p. 146-160, 2012.

HESLOT, N.; JANNINK, J.; SORRELLS, M. E. Perspectives for genomic selection applications and research in plants. **Crop Science**, v. 55, n. 1, p. 1-12, 2015.

ISIDRO, J. et al. Training set optimization under population structure in genomic selection. **Theoretical and applied genetics**, v. 128, n. 1, p. 145-158, 2015.

JIA, Y.; JANNINK, J.-L. Multiple-trait genomic selection methods increase genetic value prediction accuracy. **Genetics**, v. 192, n. 4, p. 1513-1522, 2012.

JIA, Z. Controlling the overfitting of heritability in genomic selection through cross validation. **Scientific reports**, v. 7, n. 1, p. 1-9, 2017.

KADAM, D. C.; RODRIGUEZ, O. R.; LORENZ, A. J. Optimization of training sets for genomic prediction of early-stage single crosses in maize. **Theoretical and Applied Genetics**, v. 134, n. 2, p. 687-699, 2021.

KIEN, N. D. et al. Cellulose content as a selection trait in breeding for kraft pulp yield in *Eucalyptus urophylla*. **Annals of Forest Science**, v. 66, n. 7, p. 1-8, 2009.

LALOË, D. Precision and information in linear models of genetic evaluation. **Genetics Selection Evolution**, v. 25, n. 6, p. 557-576, 1993.

LALOË, D.; PHOCAS, F.; MENISSIER, F. Considerations on measures of precision and connectedness in mixed linear models of genetic evaluation. **Genetics selection evolution**, v. 28, n. 4, p. 359-378, 1996.

LENZ, P. R. N. et al. Multi-trait genomic selection for weevil resistance, growth, and wood quality in Norway spruce. **Evolutionary applications**, v. 13, n. 1, p. 76-94, 2020.

LOZADA, D. N.; CARTER, A. H. Accuracy of single and multi-trait genomic prediction models for grain yield in US Pacific Northwest winter wheat. **Crop Breeding, Genetics and Genomics**, v. 1, n. 1, 2019.

MAKOUANZI, G. et al. Assessing the additive and dominance genetic effects of vegetative propagation ability in *Eucalyptus* influence of modeling on genetic gain. **Tree genetics & genomes**, v. 10, n. 5, p. 1243-1256, 2014.

- MENDONÇA, L. DE F. et al. Genomic prediction enables early but low-intensity selection in soybean segregating progenies. **Crop Science**, v. 60, n. 3, p. 1346-1361, 2020.
- MENDONÇA, L. DE F.; FRITSCHÉ-NETO, R. The accuracy of different strategies for building training sets for genomic predictions in segregating soybean populations. **Crop Science**, v. 60, n. 6, p. 3115-3126, 2020.
- MEUWISSEN, T.; HAYES, B. J.; GODDARD, M. E. Prediction of total genetic value using genome-wide dense marker maps. **Genetics**, v. 157, n. 4, p. 1819-1829, 2001.
- MONTESINOS-LÓPEZ, O. A. et al. Multi-trait, multi-environment deep learning modeling for genomic-enabled prediction of plant traits. **G3: Genes, genomes, genetics**, v. 8, n. 12, p. 3829-3840, 2018.
- MPHAHLELE, M. M. et al. Expected benefits of genomic selection for growth and wood quality traits in *Eucalyptus grandis*. **Tree Genetics & Genomes**, v. 16, n. 4, p. 1-12, 2020.
- MUÑOZ F.; SANCHEZ L. **breedR: statistical methods for forest genetic resources analysts. R package version 0.7–16** 2014. Disponível em: <<https://github.com/famuvie/breedR>>. Acesso em: 15 de março de 2022.
- NORMAN, A. et al. Optimising genomic selection in wheat: effect of marker density, population size and population structure on prediction accuracy. **G3: Genes, Genomes, Genetics**, v. 8, n. 9, p. 2889-2899, 2018.
- OLIVOTO, T.; LÚCIO, A. D. metan: An R package for multi-environment trial analysis. **Methods in Ecology and Evolution**, v. 11, n. 6, p. 783-789, 2020.
- OSORIO, L.; WHITE, T.; HUBER, D. Age-age and trait-trait correlations for *Eucalyptus grandis* Hill ex Maiden and their implications for optimal selection age and design of clonal trials. **Theoretical and Applied Genetics**, v. 106, n. 4, p. 735-743, 2003.
- OU, J.-H.; LIAO, C.-T. Training set determination for genomic selection. **Theoretical and Applied Genetics**, v. 132, n. 10, p. 2781-2792, 2019.
- PALUDETO, J. G. Z. et al. Genomic relationship-based genetic parameters and prospects of genomic selection for growth and wood quality traits in *Eucalyptus benthamii*. **Tree Genetics & Genomes**, v. 17, n. 4, p. 1-20, 2021.
- RAMBOLARIMANANA, T. et al. Performance of multi-trait genomic selection for *Eucalyptus robusta* breeding program. **Tree Genetics & Genomes**, v. 14, n. 5, p. 1-13, 2018.
- RAYMOND, C. A. Genetics of *Eucalyptus* wood properties. **Annals of Forest Science**, v. 59, n. 5-6, p. 525-531, 2002.

RESENDE, J. F. R. et al. Accuracy of genomic selection methods in a standard data set of loblolly pine (*Pinus taeda* L.). **Genetics**, 2012.

RESENDE, R. T. et al. Assessing the expected response to genomic selection of individuals and families in *Eucalyptus* breeding with an additive-dominant model. **Heredity**, v. 119, n. 4, p. 245-255, 2017.

REZENDE, G. D. S. P. et al. Breeding forest trees by genomic selection: current progress and the way forward. In: **Genomics of plant genetic resources**. [s.l.] Springer, 2014. p. 651-682.

RIEDELSHEIMER, C.; MELCHINGER, A. E. Optimizing the allocation of resources for genomic selection in one breeding cycle. **Theoretical and applied genetics**, v. 126, n. 11, p. 2835-2848, 2013.

RINCENT, R. et al. Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: comparison of methods in two diverse groups of maize inbreds (*Zea mays* L.). **Genetics**, v. 192, n. 2, p. 715-728, 2012.

RINCENT, R.; CHARCOSSET, A.; MOREAU, L. Predicting genomic selection efficiency to optimize calibration set and to assess prediction accuracy in highly structured populations. **Theoretical and Applied Genetics**, v. 130, n. 11, p. 2231-2247, 2017.

SABADIN, F. et al. Population-tailored mock genome enables genomic studies in species without a reference genome. **Molecular Genetics and Genomics**, p. 1-14, 2021.

SARINELLI, J. M. et al. Training population selection and use of fixed effects to optimize genomic predictions in a historical USA winter wheat panel. **Theoretical and Applied Genetics**, v. 132, n. 4, p. 1247-1261, 2019.

SCHUMACHER, F. X. Logarithmic expression of timber-tree volume. **Journal of Agricultural Research**, v. 47, n.9, p. 719-734, 1933.

SCHWANNINGER, M.; RODRIGUES, J. C.; FACKLER, K. A review of band assignments in near infrared spectra of wood and wood components. **Journal of Near Infrared Spectroscopy**, v. 19, n. 5, p. 287-308, 2011.

SCUTARI, M.; MACKAY, I.; BALDING, D. Improving the efficiency of genomic selection. **Statistical applications in genetics and molecular biology**, v. 12, n. 4, p. 517-527, 2013.

SILVA-JUNIOR, O. B.; FARIA, D. A.; GRATTAPAGLIA, D. A flexible multi-species genome-wide 60K SNP chip developed from pooled resequencing of 240 *Eucalyptus* tree genomes across 12 species. **New Phytologist**, v. 206, n. 4, p. 1527-1540, 2015.

SUN, C. et al. Improvement of prediction ability for genomic selection of dairy cattle by including dominance effects. **PloS one**, v. 9, n. 8, 2014.

TAN, B. et al. Evaluating the accuracy of genomic prediction of growth and wood traits in two *Eucalyptus* species and their F1 hybrids. **BMC plant biology**, v. 17, n. 1, p. 110, 2017.

TAN, B. et al. Genomic relationships reveal significant dominance effects for growth in hybrid *Eucalyptus*. **Plant science**, v. 267, p. 84-93, 2018.

THUMMA, B. R.; JOYCE, K. R.; JACOBS, A. Genomic studies with preselected markers reveal dominance effects influencing growth traits in *Eucalyptus nitens*. **G3: Genes, Genomes, Genetics**, v. 12, n. 1, 2022.

VANRADEN, P. M. Efficient methods to compute genomic predictions. **Journal of dairy science**, v. 91, n. 11, p. 4414-4423, 2008.

VARONA, L. et al. Non-additive effects in genomic selection. **Frontiers in genetics**, v. 9, p. 78, 2018.

VITEZICA, Z. G.; VARONA, L.; LEGARRA, A. On the additive and dominant variance and covariance of individuals within the genomic selection scope. **Genetics**, v. 195, n. 4, p. 1223-1230, 2013.

WANG, X. et al. Genomic selection methods for crop improvement: Current status and prospects. **The Crop Journal**, v. 6, n. 4, p. 330-340, 2018.

WAPLES, R. S.; DO, C. H. I. LDNE: a program for estimating effective population size from data on linkage disequilibrium. **Molecular ecology resources**, v. 8, n. 4, p. 753-756, 2008.

WARD, B. P. et al. Multienvironment and multitrait genomic selection models in unbalanced early-generation wheat yield trials. **Crop Science**, v. 59, p. 491-507, 2019.

WEI, X.; BORRALHO, N. M. G. Genetic control of growth traits of *Eucalyptus urophylla* ST Blake in South East China. **Silvae genetica**, v. 47, n. 2-3, p. 158-165, 1998.

WERNER, C. R. et al. How population structure impacts genomic selection accuracy in cross-validation: implications for practical breeding. **Frontiers in plant science**, 2020.

WU, S. et al. Genotypic variation in wood properties and growth traits of *Eucalyptus* hybrid clones in southern China. **New Forests**, v. 42, n. 1, p. 35-50, 2011.

ZHANG, W. et al. Evaluation of genomic prediction for *Fusarium* head blight resistance with a multi-parental population. **Biology**, v. 10, n. 8, p. 756, 2021.

ZHAO, Y. et al. Accuracy of genomic selection in European maize elite breeding populations. **Theoretical and Applied Genetics**, v. 124, n. 4, p. 769-776, 2012.

ZHENG, X. et al. A high-performance computing toolset for relatedness and principal

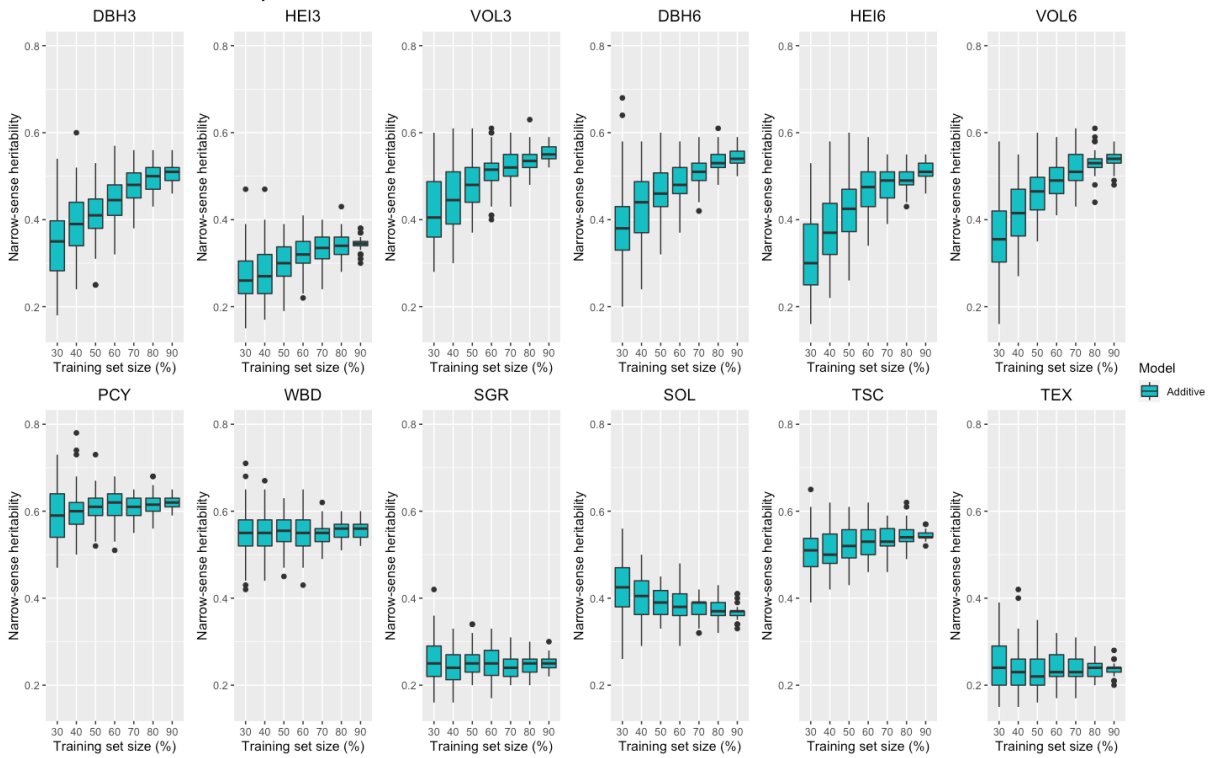
component analysis of SNP data. **Bioinformatics**, v. 28, n. 24, p. 3326-3328, 2012.

ZHOU, X.; STEPHENS, M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. **Nature methods**, v. 11, n. 4, p. 407-409, 2014.

ZHU, X. et al. Training set design in genomic prediction with multiple biparental families. **The Plant Genome**, v. 14, n. 3, 2021.

APPENDIX A - SUPPLEMENTARY MATERIAL FOR CHAPTER 2

Figure S1. Narrow-sense heritability for growth and wood-quality traits in a *Eucalyptus grandis* population located in São Miguel Arcando, São Paulo, Brazil.



FINAL CONSIDERATIONS

Using genomic information associated with phenotypic data is an effective option to improve *Eucalyptus* breeding programs. This study shows that both genome-wide association as well as genomic selection models can be key strategies to improve *Eucalyptus* breeding. Both GWAS and GS methods can be effective in selecting superior genotypes in two principal ways: *i*) using marker-assisted selection to identify genotypes with genes significantly linked to the expression of economically-relevant traits; and *ii*) reducing the time spent on tree breeding through early selection. Furthermore, this study highlights the importance of using accurate genomic information to improve tree breeding programs.

In terms of GWAS, the SNP repositioning using the second version of the *Eucalyptus* genome enabled us to accurately perform GWAS and identify the genes associated with trait expression. Additionally, both methods (single-trait - farmCPU and multi-trait - MTMM) were effective in finding significant candidate genes associated with the expression of phenotypic traits. Spurious associations were not evident since we found no significant deviation from the expected and observed p -values, as shown by the QQ-plots of significant markers. As expected, the phenotypic variance explained by the markers (PEV) were minimal for most significant SNPs, indicating that many genes are related to phenotypic expression in *E. grandis*. Furthermore, the pleiotropic effect of markers was evident for some traits using farmCPU since the same markers were identified as significant for more than one trait. Similarly, the MTMM model identified several significant markers related to the expression between and within growth and wood-quality traits. Through gene ontology analysis, we were able to identify several markers with different functions related to trait expression.

Considering the optimization of the selection models, the different methodologies applied proved to be effective in improving the GS prediction ability. In general, all applied methodologies (A, AD, MT, OTS, and MT-OTS) effectively improved the prediction ability of GS models. For most traits, the additive-dominant GBLUP model, using a training set of 80% of all genotypes, achieved the best prediction ability. Additionally, both OTS statistics (PEVmean and CDmean) were able to select the best genotypes to be used as the training set. However, the PEVmean statistic provided better results when considering the prediction ability. In terms of the MT analysis, the MT-CV2 scheme generally presented higher PA when compared with

the MT-CV1. The MT analysis also demonstrated the effectiveness of using growth trait information to improve the prediction ability for wood-quality traits, as well as using OTS information to increase the PA (MT-OTS; PEVmean). Considering the low coincidence of selection (CS) for top the genotypes (e.g., top20 and top30), we observed that the GS models tended to not select the best genotypes, but it proved to be effective for tree breeding since the worse performing genotypes were not selected due to the high CS for the other categories (top50, top100, and top150). Furthermore, the superiority of GS was achieved by the relatively high selection gains for all analyzed models. The results found here show that it is possible to use genomic selection models for early selection with a relatively high accuracy in *Eucalyptus* breeding programs.

REFERENCES

- ACOSTA, M. S.; MASTRANDREA, C.; LIMA, J.T. Wood technologies and uses of Eucalyptus wood from fast grown plantations for solid products. In: **Proceedings of the 51st international convention of society of wood science and technology**, Concepción, Chile, p. 10-12. 2008.
- AGARWAL, M.; SHRIVASTAVA, N.; PADH, H. Advances in molecular marker techniques and their applications in plant sciences. **Plant cell reports**, v. 27, n. 4, p. 617-631, 2008.
- ARCURI, M. L. C. et al. Genome-wide identification of multifunctional laccase gene family in *Eucalyptus grandis*: potential targets for lignin engineering and stress tolerance. **Trees**, p. 1-14, 2020.
- BALLESTA, P. et al. Genomic Predictions Using Low-Density SNP Markers, Pedigree and GWAS Information: A Case Study with the Non-Model Species *Eucalyptus cladocalyx*. **Plants**, v. 9, n. 1, p. 99, 2020.
- BASTIAANSEN, J. W. M. et al. Long-term response to genomic selection: effects of estimation method and reference population structure for different genetic architectures. **Genetics Selection Evolution**, v. 44, n. 1, p. 1-13, 2012.
- BERNARDO, R. Predictive breeding in maize during the last 90 years. **Crop Science**, 2021.
- BERRO, I. et al. Training population optimization for genomic selection. **The Plant Genome**, v. 12, n. 3, p. 1-14, 2019.
- BOLORMAA, S. et al. A genome-wide association study of meat and carcass traits in Australian cattle. **Journal of animal science**, v. 89, n. 8, p. 2297-2309, 2011.
- BUDHLAKOTI, N. et al. Comparative study of different non-parametric genomic selection methods under diverse genetic architecture. 2020.
- BURDON, R. D. Early selection in tree breeding: principles for applying index selection and inferring input parameters. **Canadian Journal of Forest Research**, v. 19, n. 4, p. 499-504, 1989.
- BUSH, W. S.; MOORE, J. H. Genome-wide association studies. **PLoS Comput Biol**, v. 8, n. 12, 2012.
- CAMPOE, O. C. et al. Stem production, light absorption and light use efficiency between dominant and non-dominant trees of *Eucalyptus grandis* across a productivity gradient in Brazil. **Forest Ecology and Management**, v. 288, p. 14-20, 2013.
- CAPPA, E. P. et al. Genomic-based multiple-trait evaluation in *Eucalyptus grandis* using dominant DArT markers. **Plant Science**, v. 271, p. 27-33, 2018.

CERICOLA, F. et al. Optimizing training population size and genotyping strategy for genomic prediction using association study results and pedigree information. A case of study in advanced wheat breeding lines. **PloS one**, v. 12, n. 1, 2017.

CHANG, L.-Y. et al. High density marker panels, SNPs prioritizing and accuracy of genomic selection. **BMC genetics**, v. 19, n. 1, p. 1-10, 2018.

CROSSA, J. et al. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. **Genetics**, v. 186, n. 2, p. 713-724, 2010.

CROSSA, J. et al. Genomic selection in plant breeding: methods, models, and perspectives. **Trends in plant science**, v. 22, n. 11, p. 961-975, 2017.

DE MORAES GONCALVES, J. L. et al. Integrating genetic and silvicultural strategies to minimize abiotic and biotic constraints in Brazilian eucalypt plantations. **Forest ecology and management**, v. 301, p. 6-27, 2013.

DENIS, M.; BOUVET, J.-M. Efficiency of genomic selection with models including dominance effect in the context of *Eucalyptus* breeding. **Tree Genetics & Genomes**, v. 9, n. 1, p. 37–51, 2013.

DUIJVESTIJN, N. et al. A genome-wide association study on androstenone levels in pigs reveals a cluster of candidate genes on chromosome 6. **BMC genetics**, v. 11, n. 1, p. 42, 2010.

GODDARD, M. E.; HAYES, B. J. Genomic selection. **Journal of Animal breeding and Genetics**, v. 124, n. 6, p. 323-330, 2007.

GRATTAPAGLIA, D. Breeding forest trees by genomic selection: current progress and the way forward. **Genomics of plant genetic resources**, 651-682, 2014.

GRATTAPAGLIA, D. Status and perspectives of genomic selection in forest tree breeding. In: **Genomic selection for crop improvement**. [s.l.] Springer, 2017. p. 199–249.

GRATTAPAGLIA, D. et al. Quantitative genetics and genomics converge to accelerate forest tree breeding. **Frontiers in Plant Science**, 2018.

GUIMARÃES, E. P. **Marker-assisted selection: current status and future perspectives in crops, livestock, forestry and fish**. [s.l.] Food & Agriculture Org., 2007.

HABIER, D. et al. Extension of the Bayesian alphabet for genomic selection. **BMC bioinformatics**, v. 12, n. 1, p. 1-12, 2011.

HAYES, B. J. et al. **Accuracy of genomic selection: comparing theory and results**. Proc Assoc Advmt Anim Breed Genet. **Anais...**2009

HENDERSON, C. R. Best linear unbiased estimation and prediction under a selection model. **Biometrics**, p. 423-447, 1975.

HIRSCHHORN, J. N.; DALY, M. J. Genome-wide association studies for common diseases and complex traits. **Nature reviews genetics**, v. 6, n. 2, p. 95-108, 2005.

HOWARD, R.; CARRIQUIRY, A. L.; BEAVIS, W. D. Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. **G3: Genes, Genomes, Genetics**, v. 4, n. 6, p. 1027-1046, 2014.

IBÁ - Indústria Brasileira de Árvores. **Relatório 2019**. 80 p. Disponível em: <<https://iba.org/datafiles/publicacoes/relatorios/iba-relatorioanual2019.pdf>>. Acessado em: 13 de setembro de 2021.

JABBARI, M. et al. GWAS analysis in spring barley (*Hordeum vulgare* L.) for morphological traits exposed to drought. **PloS one**, v. 13, n. 9, 2018.

JIA, Y.; JANNINK, J.-L. Multiple-trait genomic selection methods increase genetic value prediction accuracy. **Genetics**, v. 192, n. 4, p. 1513-1522, 2012.

KAINER, D. et al. High marker density GWAS provides novel insights into the genomic architecture of terpene oil yield in *Eucalyptus*. **New Phytologist**, v. 223, n. 3, p. 1489–1504, 2019.

LEBEDEV, V. G. et al. Genomic selection for forest tree improvement: Methods, achievements and perspectives. **Forests**, v. 11, n. 11, p. 1190, 2020.

LEGARRA, A. et al. Improved Lasso for genomic selection. **Genetics research**, v. 93, n. 1, p. 77-87, 2011.

LI, Z.; SILLANPÄÄ, M. J. Overview of LASSO-related penalized regression methods for quantitative trait mapping and genomic selection. **Theoretical and applied genetics**, v. 125, n. 3, p. 419-435, 2012.

MACKAY, T. F. C. The genetic architecture of quantitative traits. **Annual review of genetics**, v. 35, n. 1, p. 303-339, 2001.

MAMMADOV, J. et al. SNP markers and their impact on plant breeding. **International journal of plant genomics**, v. 2012, 2012.

MEUWISSEN, T.; HAYES, B. J.; GODDARD, M. E. Prediction of total genetic value using genome-wide dense marker maps. **Genetics**, v. 157, n. 4, p. 1819-1829, 2001.

MISZTAL, I.; STEIN, Y.; LOURENCO, D. A. L. Genomic evaluation with multibreed and crossbred data. **JDS Communications**, 2022.

MOSTERT-O'NEILL, M. M. et al. Genomic evidence of introgression and adaptation in a model subtropical tree species, *Eucalyptus grandis*. **Molecular Ecology**, v. 30, n. 3, p. 625-638, 2021.

MPHAHLELE, M. M. et al. Expected benefits of genomic selection for growth and wood quality traits in *Eucalyptus grandis*. **Tree Genetics & Genomes**, v. 16, n. 4, p.

1-12, 2020.

MÜLLER, B. S. F. et al. 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**, v. 18, n. 1, p. 524, 2017.

MYBURG, A. A. et al. The genome of *Eucalyptus grandis*. **Nature**, v. 510, n. 7505, p. 356-362, 2014.

NAMKOONG, G.; BARNES, R. D.; BURLEY, J. Screening for yield in forest tree breeding. **The Commonwealth Forestry Review**, p. 61-68, 1980.

NOVAES, E. et al. High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. **BMC genomics**, v. 9, n. 1, p. 312, 2008.

O'MALLEY, D. M.; MCKEAND, S. E. **Marker assisted selection for breeding value in forest trees**. [s.l.] Citeseer, 1994.

PALUDETTO, J. G. Z. et al. Genomic relationship-based genetic parameters and prospects of genomic selection for growth and wood quality traits in *Eucalyptus benthamii*. **Tree Genetics & Genomes**, v. 17, n. 4, p. 1-20, 2021.

POTTS, B. M.; DUNGEY, H. S. Interspecific hybridization of *Eucalyptus*: key issues for breeders and geneticists. **New Forests**, v. 27, n. 2, p. 115-138, 2004.

RAFALSKI, A. Applications of single nucleotide polymorphisms in crop genetics. **Current opinion in plant biology**, v. 5, n. 2, p. 94-100, 2002.

REZENDE, G. D. S. P.; DE RESENDE, M. D. V.; DE ASSIS, T. F. *Eucalyptus* Breeding for Clonal Forestry. In: **Challenges and Opportunities for the World's Forests in the 21st Century**. Springer, 2014, p. 393-424.

SILVA-JUNIOR, O. B.; FARIA, D. A.; GRATTAPAGLIA, D. A flexible multi-species genome-wide 60K SNP chip developed from pooled resequencing of 240 *Eucalyptus* tree genomes across 12 species. **New Phytologist**, v. 206, n. 4, p. 1527–1540, 2015.

TAO, Y. et al. Large-scale GWAS in sorghum reveals common genetic control of grain size among cereals. **Plant Biotechnology Journal**, v. 18, n. 4, p. 1093–1105, 2020.

TAYEH, N. et al. Genomic prediction in pea: effect of marker density and training population size and composition on prediction accuracy. **Frontiers in plant science**, v. 6, p. 941, 2015.

THAVAMANIKUMAR, S. et al. Dissection of complex traits in forest trees—opportunities for marker-assisted selection. **Tree Genetics & Genomes**, v. 9, n. 3, p. 627–639, 2013.

USAI, M. G.; GODDARD, M. E.; HAYES, B. J. LASSO with cross-validation for

genomic selection. **Genetics research**, v. 91, n. 6, p. 427-436, 2009.

VARSHNEY, R. K.; ROORKIWAL, M.; SORRELLS, M. E. Genomic Selection for Crop Improvement: An Introduction. In: **Genomic Selection for Crop Improvement**. [s.l.] Springer, 2017. p. 1-6.

WALDRON, L. et al. Optimized application of penalized regression methods to diverse genomic data. **Bioinformatics**, v. 27, n. 24, p. 3399-3406, 2011.

WANG, M. et al. Genome-wide association study (GWAS) of resistance to head smut in maize. **Plant science**, v. 196, p. 125-131, 2012.

WANG, W. Y. S. et al. Genome-wide association studies: theoretical and practical concerns. **Nature Reviews Genetics**, v. 6, n. 2, p. 109, 2005.

WANG, X.; YANG, Z.; XU, C. A comparison of genomic selection methods for breeding value prediction. **Science Bulletin**, v. 60, n. 10, p. 925-935, 2015.

WHITTAKER, J. C.; THOMPSON, R.; DENHAM, M. C. Marker-assisted selection using ridge regression. **Genetics Research**, v. 75, n. 2, p. 249-252, 2000.

WU, H. X. et al. Study of early selection in tree breeding. **Silvae Genetica, Frankfurt**, v. 47, n. 2-3, p. 146-155, 1998.

WU, X. et al. Genome wide association studies for body conformation traits in the Chinese Holstein cattle population. **BMC genomics**, v. 14, n. 1, p. 897, 2013.

XU, S. An empirical Bayes method for estimating epistatic effects of quantitative trait loci. **Biometrics**, v. 63, n. 2, p. 513-521, 2007.

XU, Y.; XU, C.; XU, S. Prediction and association mapping of agronomic traits in maize using multiple omic data. **Heredity**, v. 119, n. 3, p. 174-184, 2017.