RAFAEL TASSINARI RESENDE

REGIONAL HERITABILITY MAPPING AND GWAS FOR MOLECULAR BREEDING IN EUCALYPTUS HYBRIDS

Thesis presented to the Universidade Federal de Viçosa, in partial fulfillment of the requirements of Genetic and Breeding Program for degree of Doctor Scientiae.

VIÇOSA
MINAS GERAIS - BRAZIL
2017

Ficha catalográfica preparada pela Biblioteca Central da Universidade Federal de Viçosa - Câmpus Viçosa

T
R433r
2017

Resende, Rafael Tassinari, 1986-
Regional heritability mapping and GWAS for molecular breeding in eucalyptus hybrids / Rafael Tassinari Resende. – Viçosa, MG, 2017.
x, 47f. : il. ; 29 cm.

Inclui apêndice.
Orientador: Marcos Deon Vilela de Resende.
Tese (doutorado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. Eucalipto - Melhoramento genético. 2. Genômica. 3. Madeira - Pesquisa. I. Universidade Federal de Viçosa. Departamento de Biologia Geral. Programa de Pós-graduação em Genética e Melhoramento. II. Título.

CDD 22 ed. 634.922

RAFAEL TASSINARI RESENDE

REGIONAL HERITABILITY MAPPING AND GWAS FOR MOLECULAR BREEDING IN EUCALYPTUS HYBRIDS

Thesis presented to the Universidade Federal de Viçosa, in partial fulfillment of the requirements of Genetic and Breeding Program for degree of Doctor Scientiae.

APROVED: February 20, 2017.

Fabyano Fonseca e Silva

Camila Ferreira Azevedo

Elizabete Keiko Takahashi

Felipe Lopes da Silva

Luiz Alexandre Peternelli
(President of the examining commission)

A Deus,

OFEREÇO

Aos meus pais Cezar e Auxiliadora,
E minha tia Adélia
Por todo amor envolvido.

DEDICO

ACKNOWLEDGEMENTS

To God.
My advisor Marcos Deon V. Resende;
To Dario Grattapaglia;
CENIBRA S.A. and the researcher Elizabete K. Takahashi;
Professors Camila Azevedo and Fabyano F. Silva;
Professors Luiz A. Peternelli and Felipe L. da Silva;
My parents Cezar and Auxiliadora, and my Aunt Adélia;
My girlfriend Ursula;
My friends and colleagues of the Genetics and Breeding Program;
To the professors of the disciplines that I took during my doctorate;
Marco Tulio and Odilon from the Secretary of Genetic and Breeding Program;
My friends and colleagues from the Department of Forest Engineering;
To my friends and colleagues of the República;
To the friends Roberta Corsino and Gustavo Marcilio;
To the professors Leonardo L. Bhering and Cosme Damião Cruz for the support of Biometrics and Bioinformatics laboratories;
To Luciano Braatz for help with Structure software;
To Ricardo Cabral for the scientific discussions.

O homem só envelhece
quando o seu lamento
substitui seus sonhos

(Chiquinho da Floresta, Viçosa-MG, 1994)

SUMMARY

LIST OF TABLES

LIST OF FIGURES

LIST OF SUPPLEMENTARY MATERIAL

ABSTRACT

RESENDE, Rafael Tassinari, D.Sc., Universidade Federal de Viçosa, February, 2017. **Regional Heritability Mapping and GWAS for molecular breeding in eucalyptus hybrids**. Adviser: Marcos Deon Vilela de Resende.

Although genome-wide association studies (GWAS) have provided valuable insights into the decoding of the relationships between sequence variation and complex phenotypes, they have explained little heritability. Regional heritability mapping (RHM) provides heritability estimates for genomic segments containing both common and rare allelic effects that individually contribute too little variance to be detected by GWAS. We carried out GWAS and RHM for seven growths, wood and disease resistance traits in a breeding population of 768 Eucalyptus hybrid trees using EuCHIP60K. Total genomic heritabilities accounted for large proportions (64–89%) of pedigree-based trait heritabilities, providing additional evidence that complex traits in eucalypts are controlled by many sequence variants across the frequency spectrum, each with small contributions to the phenotypic variance. RHM detected 26 quantitative trait loci (QTLs) encompassing 2,191 single nucleotide polymorphisms (SNPs), whereas GWAS detected 13 single SNP–trait associations. RHM and GWAS QTLs individually explained 5–15% and 4–6% of the genomic heritability, respectively. RHM was superior to GWAS in capturing larger proportions of genomic heritability. Equated to previously mapped QTLs, our results highlighted genomic regions for further examination towards gene discovery. RHM-QTLs bearing a combination of common and rare variants could be useful enhancements to incorporate prior knowledge of the underlying genetic architecture in genomic prediction models.

RESUMO

RESENDE, Rafael Tassinari, D.Sc., Universidade Federal de Viçosa, fevereiro de 2017. **Regional Heritability Mapping e GWAS no melhoramento molecular de híbridos de eucalipto**. Orientador: Marcos Deon Vilela de Resende.

Embora os estudos de associação genômica (GWAS) forneçam informações valiosas na descodificação das relações entre a variação gênica e os fenótipos complexos, esta técnica explica uma pequena fração da herdabilidade. O mapeamento de herdabilidades regionais (RHM) fornece estimativas de herdabilidade para segmentos genômicos que contêm efeitos alélicos raros e que contribuem individualmente com baixa variação ao ponto de serem detectados pela GWAS. Neste estudo foi realizado a GWAS e o RHM para sete características de crescimento, madeira e resistência à doenças em uma população 768 árvores híbridas de *Eucalyptus* usando o moderno Chip Illumina EuCHIP60K. As herdabilidades genômicas totais representaram grandes proporções (64-89%) de herdabilidades baseadas em pedigree, fornecendo evidências adicionais de que características complexas em eucaliptos são controlados por muitas variantes ao longo do genoma, cada uma com pequenas contribuições para a variância fenotípica. O RHM detectou 26 QTLs (Q*uantitative Trait Loci*) abrangendo 2.191 SNPs (S*ingle Nucleotide Polymorphism*), enquanto que a GWAS detectou 13 associações. Os QTLs detectados via RHM e GWAS explicaram individualmente 5 a 15% e 4 a 6% da herdabilidade genômica, respectivamente. O RHM foi superior à GWAS na captura de maiores proporções de herdabilidade genômica. Semelhantemente a QTLs previamente mapeados, os resultados destacaram as regiões genômicas que podem ser utilizadas em estudos mais aprofundados para descoberta de genes. Os RHM-QTLs contendo uma combinação de variantes comuns e raras representam um avanço para incorporar conhecimento prévio da arquitetura genética subjacente em modelos de predição genômica.

1. Regional heritability mapping and genome-wide association identify loci for complex growth, wood and disease resistance traits in Eucalyptus

1.1 Background

1.1.1 The Eucalyptus Crop and Breeding

The eucalyptus breeding aims to increase its productivity, enable the achievement of greater performance in traits of interest and reduce environmental impacts (GOLLE et al., 2009). In this sense, different strategies are established combining crossing design, selection methods and population structure in order to increase the frequency of alleles of interest in the population (RESENDE, 2000), which results in genetic breeding to their offspring. In this same perspective, the biotechnology emerged as a feasible tool capable to provide a better understanding of individuals' genotype and allow the indirect obtaining of predicted genotypic data with high reliability in each cycle, reducing its extent (GOLLE et al., 2009).

Evidence of the effectiveness of eucalyptus breeding programs may be obtained by genetic values prediction of individuals using BLUP procedure (best linear unbiased prediction). Prediction via BLUP allows the maximization of selective accuracy and prediction of genotypic values under unbalanced experiments, allows the use of information from different locations besides being superior to the method of least squares (RESENDE, 2000). The effectiveness of the methodology can also be measured by the estimation of genetic gains over a period of time so that it is possible to carry out the selection of the traits of interest in the species. For this, knowledge of the variability between and within populations and the genetic control related to them are fundamental (VENCOVSKY, 1986).

Some phenotypic variables are associated with forest productivity. The main evaluated variable in eucalyptus experiments still is the volume of wood, which in turn depends on the diameter and height of the tree. In pulp production sector, the yield of pulp as well as the paper qualities are important economic indicators (GOMES et al., 2015). Eucalyptus has also problems with pathogen attack, such as rust (caused by the *Puccina pisidii* fungus) (JUNGHANS et al., 2003), so the study of resistance pathways is also desirable. Except for the volume of wood, all other phenotypic traits

related to wood quality, pulp production, and resistance to diseases are difficult to measure (CASTRO et al., 2016). The use of molecular genomic techniques has shown to be beneficial in the accurate prediction of these variables.

1.1.2 QTL Detections

Chromosomal regions presenting genes or loci controlling the variation in polygenic traits are called QTLs (Quantitative Trait Loci) (FALCONER; MACKAY, 1996). The main importance of QTLs mapping is to estimate the relations between genotypes and phenotypes of a quantitative trait along the genome. At first it seems a simple procedure, especially with all available knowledge of current genomics and computational tools (DOERGE, 2002). However, it is a complex procedure due the large number of QTLs associated with a given trait and possible epistatic interactions. Moreover, many other additional sources of variation may influence the results (MACKAY, 2001).

The full mapping should provide information about the position, number, and effects of the markers. Moreover, they may also rely on other information such as allelic interactions between (epistasis) and within loci (dominance) of QTLs (GARCIA et al., 2008). QTLs identification also allows a better understanding of genetic bases and correlation between traits governed by linked or pleiotropic QTLs (FALCONER; MACKAY, 1996); and the identification of interactions between QTLs and environment (BOER et al., 2007). Thus, firstly, the availability of a map of locus markers is required, which provides support for the allocation of QTLs throughout the genome. Molecular markers are fast and effective tools for genomic studies, in terms that they detect the polymorphism directly at DNA level and do not have environmental influence (SOUZA, 2001). Numerous types of molecular markers are currently used for genomes studies of species of agronomic interest, or of natural diversity. However, single nucleotide polymorphism (SNP) markers are gaining special prominence given the large volume of information, reliability, codominance and reproducibility of application.

There are several genetic-statistical methodologies available for association analysis between phenotypic and molecular data. Among them, we can highlight those with specific control of the type of family (such as full-sibs, half-sibs, Recombinant Inbred Lines, Biparental Outcrossing, and others), which includes: the analysis for

each individual marker, interval mapping, composite interval mapping, and multiple interval mapping (for more details see ANONI, 2011). However, given the high availability of genomic and phenotypic data (in a new concept of big data), genome wide association (GWAS) analyzes without segregation control have shown to be more efficient in QTLs detection (BAZAKOS et al., 2017).

The general idea of GWAS is to evaluate the effects of individual markers or groups of markers on the phenotypic expression, considering that the marker will affect the trait only if it is in linkage disequilibrium with the supposed QTL. The main GWAS procedure separately estimates the effect of each marker on the phenotype by fixed regression on single marks or by mixed model equations. In this approach, the existence of associations between markers and quantitative traits of interest (significance for the markers effects) are evaluated through traditional statistical tests such as t-test, analysis of variance and maximum likelihood ratio test (DOERGE; ZENG; WEIR, 1997). In these tests, the null hypothesis is defined as that "the marker does not show any effect on the trait". If the null hypothesis is rejected, it is inferred that a possible QTL is linked to that marker (DOERGE, 2002). Analysis of individual markers is an easy-to-carry method, requiring no specialized software neither complex statistical tools; furthermore, it does not require the prior building of the linkage map. However, its results are biased, in terms that there is a misunderstanding between the QTL effect and its distance from the marker, resulting in low QTL detection power. Considering intervals between pairs of adjacent markers, the existence of a possible QTL within the windows is tested. The results are expressed in LOD-Score or $-\log_{10}$ of the p-value, and a greater QTLs detection power is observed compared to individual markers analyses. However, by not controlling the interference of QTLs that are located outside the analyzed intervals, there are chances of false positives occurrence (false inference of association).

Population structure as well as familial linkage disequilibrium may bias the genomic association in order to detect specific QTLs, not observable in other populations (SEHGAL; SINGH; RAJPAL, 2016). The capture of general effects becomes more attractive from genetic breeding point of view; in terms that it is desirable to identify the markers that lead to more adapted genotypes. Association mapping is generally divided into two categories: mapping associations based on candidate genes, which relates polymorphisms in selected candidate genes; and

genome-wide association mapping (GWAS), which examines genetic variation throughout the genome to identify evidence of association for complex traits. The most important aspect in deciding between a candidate gene approach and a complete genome study is the extent of Linkage Disequilibrium (LD) in the population of interest. The extent of LD determines not only the resolution of the mapping but also the number of markers needed for proper genome coverage in a genomic study (WHITT; BUCKLER, 2003).

This study used two approaches of Genomic Association, the classic GWAS, which considers marker-to-marker in the genome and the Regional Heritability Mapping (RHM) method proposed by NAGAMINE et al. (2012). In both procedures, a correction was made for families' effect and population structure, in order to control possible false positives and obtain more accurate results. The RHM method considers genomic windows instead of marker-to-marker and is effective in harnessing information from rare alleles, in terms that it uses this information in association with other neighboring markers with a higher degree of polymorphism. Other advantages of RHM are described by RIGGIO; PONG-WONG (2014) and detailed in the following topic.

1.2 Introduction

The deciphering of the complex relationships between sequence variation and complex phenotypes has been a key driver of modern genetics and genomics with important consequences for fundamental biology and applied breeding practice. Efforts in this direction have been undertaken in forest tree genomics to understand the adaptive diversity of trees and to provide tools to accelerate the long time necessary to complete a breeding cycle. For species of Eucalyptus, notwithstanding their fast growth, breeding cycles generally take 12–16 years to deliver elite genotypes (REZENDE; DE RESENDE; DE ASSIS, 2014). Although growth traits are measured in all trees of a progeny trial, the assessment of wood properties is typically carried out in a considerably smaller number of trees in the late stages of the breeding cycle, such that the full range of genetic variation in wood properties is not exploited (GRATTAPAGLIA, 2014).

The development of DNA markers fueled great expectations of the acceleration of selection for complex traits in forest trees in general and Eucalyptus in particular. Several quantitative trait locus (QTL) mapping studies were carried out in bi-parental populations and some attempts to use this information for breeding were made (reviewed in GRATTAPAGLIA et al., 2012). When meant for use in selection, QTL mapping data suffer from substantial drawbacks that have been discussed extensively (BERNARDO, 2008; GRATTAPAGLIA; RESENDE, 2011). Only a small proportion of QTLs underlying the target trait are detected given the low mapping power, and the variance explained by the QTLs is largely overestimated (BEAVIS, 1998). Association genetics, proposed as a solution to the quandaries of QTL mapping, was expected to result in major advances in the dissection of multifactorial traits (NEALE; SAVOLAINEN, 2004), warranting a number of association studies based on candidate genes (NEALE; KREMER, 2011). In the eucalypts, a few associations between polymorphisms in candidate genes and wood traits have been reported (DENIS et al., 2013; DILLON et al., 2012; MANDROU et al., 2012; THAVAMANIKUMAR et al., 2014; THUMMA et al., 2005, 2009). Candidate gene studies, however, suffer from severe bias introduced by the a priori choice of genes and do not account for more than a few percent of trait variation, with estimates inflated because of the 'winner's curse' effect (GODDARD et al., 2009). The first genome-wide association studies (GWAS) in forest trees have been reported for wood (PORTH et al., 2013), biomass, ecophysiological and phenological traits in poplar (MCKOWN et al., 2014). Despite the higher genotyping density in these studies (c. 29,000 single nucleotide polymorphisms, SNPs), only c. 3,500 of the 45,557 (7.7%) annotated genes were targeted, and GWAS results captured very modest proportions of the phenotypic variance. The only GWAS in Eucalyptus to date, carried out with relatively low power, reported 16 marker–trait associations for growth and two for lignin traits (CAPPA et al., 2013). Despite the efforts to discover polymorphisms associated with economically relevant traits, much of the genetic contribution to complex traits in forest trees remains unexplained.

The recent development of a high-density 60,000 SNP chip for Eucalyptus has opened up opportunities for GWAS and genomic prediction experiments (SILVA-JUNIOR; FARIA; GRATTAPAGLIA, 2015). The multispecies SNP discovery strategy adopted for chip design has been shown to sample variants across most of the site frequency spectrum, mitigating ascertainment bias. Nevertheless, because of the

limited linkage disequilibrium (LD) between rare segregating alleles and genotyped SNPs on the chip, GWAS have been shown to lack detection power for rare genetic variants (BODMER; TOMLINSON, 2010; ROBINSON; WRAY; VISSCHER, 2014). Instead of evaluating each variant individually, tests that assess the cumulative effects of multiple genetic variants in a gene or a genomic region increase power when multiple variants in the group are associated with a given trait, therefore increasing the likelihood of capturing the complete effect of a QTL (LEE et al., 2014).

A relatively novel method, called regional genomic relationship mapping or regional heritability mapping (RHM), has been proposed, showing superiority when compared with other methods in uncovering variance not accounted for by GWAS (NAGAMINE et al., 2012). RHM uses a genomic relationship matrix (GRM) between individuals based on common and rare SNP variants found in short segments of the genome to estimate the trait variance explained by such regions. This method has been shown to have greater power for the detection of true QTLs, with considerably lower rates of false positives and larger fractions of explained variance when compared with a number of other local test methods, either SNP-by-SNP or segment testing (USAI et al., 2014). The power advantage of RHM has also been shown when compared with gene-based association methods, such as sequence kernel association test (SKAT) (WU et al., 2011) and canonical correlation analysis (TANG; FERREIRA, 2012), under a range of scenarios for QTLs controlled by both common and rare alleles (RIGGIO et al., 2013; UEMOTO et al., 2013). The use of RHM, assessed by simulation on full sequence data, detected a larger number of QTLs than did GWAS, although QTLs individually explained a slightly smaller amount of genetic variance (CABALLERO; TENESA; KEIGHTLEY, 2015).

Although the RHM method has received increasing attention in humans (SHIRALI et al., 2015) and domestic animals (MATIKA et al., 2016; RIGGIO et al., 2013; USAI et al., 2014), to the best of our knowledge it has not yet been applied to the study of complex traits in plant species. In this work, we were interested in evaluating the performance of RHM and GWAS in an operational breeding population that has undergone selection to pinpoint regions that would capture larger fractions of the additive genetic variance. We mapped regional QTLs and carried out GWAS for seven productivity and disease resistance traits in a breeding population of Eucalyptus. Genomic heritabilities accounted for large fractions of narrow-sense heritabilities and

RHM captured considerably more of the genomic heritability than GWAS. In addition, RHM and GWAS results were compared with previous bi-parental QTL mapping data, revealing genomic regions that could merit further examination towards gene discovery.

1.3 Materials and Methods

1.3.1 Population and phenotypes

The study was carried out in a *Eucalyptus grandis* × *Eucalyptus urophylla* hybrid breeding population belonging to Celulose Nipo-Brasileira (CENIBRA S.A.). The population involved 768 trees distributed in 37 outbred F2 full-sib families derived from mating 10 unrelated elite interspecific F1 hybrids. Trees were deployed in a field trial in a randomized incomplete block design with single-tree plots and 24–36 reps per family. Height growth (HEI) by means of a Suunto PM-5 clinometer and diameter at breast height (DBH) were measured at 3 years. *Puccinia psidii* rust (PPR) disease resistance was assessed by artificial inoculation in a glasshouse on five vegetatively propagated ramets of each one of 559 trees, as described previously (JUNGHANS et al., 2003). Four wood properties that impact pulp yield (GOMES et al., 2015) were measured: basic wood density (BWD) by the water displacement method using a 3–5-cm-thick wood disk sampled at breast height; effective alkali demand (EA) for bleached pulp; screened pulp yield (SPY) by batch kraft digestion of 150 g of wood chips at 15–18% effective alkali; and pulp bleaching content, also called kappa number (KN).

1.3.2 SNP genotyping and quality control

SNP genotypes were obtained using the Illumina Infinium (GUNDERSON et al., 2005) EuCHIP60K (SILVA-JUNIOR; FARIA; GRATTAPAGLIA, 2015), which includes 47,069 SNPs located inside or at < 10 kb distance of 30,444 of the 36,376 (84%) annotated gene models. SNP genotypes were called from intensity files obtained through GENESEEK (Lincoln, NE, USA) using GENOMESTUDIO 2011.1 (Illumina Inc., San Diego, CA, USA) following standard genotyping and quality control

procedures with no manual editing of clusters (SILVA-JUNIOR; FARIA; GRATTAPAGLIA, 2015). The average SNP call frequency across samples was > 90% and the sample call rate across SNPs was > 95%. SNP data were then filtered by keeping SNPs with minimum allele frequency (MAF) > 0.01.

1.3.3 Linkage disequilibrium and structure analysis

Pairwise estimates of LD, measured by the squared correlation of allele frequencies r^2, were obtained using all 24,806 filtered SNPs with MAF > 0.01 to inform the appropriate window length for RHM. LD was estimated using LDcorSV (MANGIN et al., 2012), correcting for population structure (r^2_s). Population structure analysis was performed with STRUCTURE v.2.3.1 (PRITCHARD; STEPHENS; DONNELLY, 2000) and the most probable value of K (K = 2) defined by DK (EVANNO; REGNAUT; GOUDET, 2005). Decay curves were fitted using a standard exponential function.
Genome-wide association study The following model was used for the GWAS:

$$y = X\beta + Fs + Sb + Zg + m_i + \varepsilon, \quad [1]$$

where the vector y represents the phenotypic values, β is the vector of fixed effects (i.e. overall mean and environmental effects), s is the vector of the fixed effect of population structure, b is the vector of the random effect of blocks within sites, g is the vector of the random genomic polygenic additive genetic effect, m_i is a scalar referring to the fixed effect of the ith marker, ε is the vector of residual effects, X and F are the incidence matrices of fixed effects, and S and Z are the incidence matrices of random effects. Analyses were run nm times, where nm is the total number of SNP markers. The random polygenic additive genetic effect (in g) was fitted to provide an estimate of the overall additive effect whilst accounting for both family and population structure. The optimal threshold for declaring a significant association was estimated for each trait using a permutation test. A Bonferroni correction for multiple tests (HOCHBERG, 1988) with a global a $\alpha = 0.05$ was then applied to the threshold following the method originally used for RHM (NAGAMINE et al., 2012). In the traditional GWAS context with the analysis of one marker of fixed effect at a time, the marker coefficient of determination (analogous to a heritability) is estimated by Equation 2:

$$h^2_{m_i} = \frac{2p_i(1-p_i)m_i^2}{\sigma^2_y}, \quad [2]$$

where p_i is the allele frequency of the ith SNP marker, m_i is the marker effect from Equation 1 and σ^2_y is the phenotypic variance. Regional heritability mapping QTL mapping using the RHM method was performed as described previously (NAGAMINE et al., 2012; RIGGIO et al., 2013). The mixed model was fitted using the R package {regress} (CLIFFORD; MCCULLAGH; CLIFFORD, 2012) as shown in Equation 3:

$$y = X\beta + Fs + Sb + Z_1 g + Z_2 r + \varepsilon, \quad [3]$$

where r is the vector of random regional genomic additive effects, Z_1 and Z_2 are the incidence matrices of random effects, and the other terms are as described for Equation 1. The distributions and variance structures of the elements of both models 1 and 3 are described as follows:

$$y|\beta, V \sim N(X\beta, V);$$
$$b|\sigma^2_b \sim N(0, \mathbf{I}\sigma^2_b);$$
$$g|\sigma^2_g \sim N(0, \mathbf{G}\sigma^2_g);$$
$$r|\sigma^2_r \sim N(0, \mathbf{G}_{reg}\sigma^2_r);$$
$$\varepsilon|\sigma^2_\varepsilon \sim N(0, \mathbf{I}\sigma^2_\varepsilon);$$

Whole-genome and regional heritabilities are $h^2_g = \sigma^2_g/\sigma^2_y$ (in a model fitted without $\mathbf{r}$) and $h^2_r = \sigma^2_r/\sigma^2_y$, respectively, and, in the full model 3, the phenotypic variance is given by $\sigma^2_y = \sigma^2_g + \sigma^2_r + \sigma^2_b + \sigma^2_\varepsilon$. Variance components were estimated using restricted maximum likelihood (REML) (PATTERSON; THOMPSON, 1971). The $\mathbf{G}$ matrix is the full genomic relationships matrix associated with the random effect accounting for relatedness and family structure, and is represented by Equation 4 as follows:

$$\mathbf{G} = \frac{WW^T}{\sum_1^{m_i} 2p_i(1-p_i)}, \quad [4]$$

where W is the SNP marker incidence matrix assuming $\mathbf{W} \subset \{-1,0,1\}$ and $\mathrm{p_i}$ is the allele frequency of the ith SNP marker present in the $\mathbf{W}$ matrix. The $\mathbf{G_{REG}}$ matrix follows Equation 3, but using a subset from $\mathbf{W}$. To circumvent numerical linear algebra problems, the diagonal of the $\mathbf{G_{REG}}$ matrix is constrained to have an average value of unity. The subsets were determined by genomic 'regions' of 2Mb in length overlapping by 1Mb (e.g. the first three regions were therefore 0–2, 1–3 and 2–4Mb), calibrated to the extent of usable LD estimated for this population ($r^2 \leq 0.2$; see the Results section). According to the density of polymorphic SNPs across the genome, a minimum of nine and a maximum of 158 SNPs were present in a genomic region spanning 2Mb. The mean, median and mode of the number of SNPs within a region were 81, 82 and 62, respectively. To test for the presence of regional variance σ_r^2 against the null hypothesis of no regional variance, the maximum likelihood ratio test (LRT) was adopted as follows:

$$\mathrm{LRT} = -2\log_e\left(\frac{\mathrm{L0}}{\mathrm{L1}}\right), \qquad [5]$$

where $\mathrm{L0}$ and $\mathrm{L1}$ represent the likelihood values for the hypothesis of the absence (H0) or presence (Ha) of regional variance, respectively (i.e. complete model of Equation 3 vs the reduced model without the term r). To explore the statistical test distribution of the null hypothesis and to find the optimal threshold for each trait, the same permutation test with Bonferroni correction for multiple tests with a global $\alpha = 0.05$ used for GWAS was also used for RHM. The genomic segments displaying significant r^2 were declared as regional QTLs. The same whole-genome relationship G matrix was used to analyze all genomic regions, and the SNPs in the particular region under analysis were not removed. Thus, any consequential correlation generated between whole-genome and regional relationships would be very small and would nevertheless reduce the likelihood of detecting a regional effect (NAGAMINE et al., 2012). Under such conditions, the likelihood of a detected regional effect was therefore very high.

1.3.4 Definition of threshold heritabilities

We estimated the threshold heritability values for declaring significant GWAS or regional mapping QTLs. These express the significance levels taking into account multiple tests applied simultaneously. This relationship can be derived theoretically from the biometrical expressions for sample sizes for a model with SNPs treated as fixed effects (N_f) and for a model with SNPs treated as random effects (N_r) required for the detection of significance at a specific a level for a stated power b, given by:

$$N_f \approx \frac{(Z_{(1-\alpha/2)}+Z_{(1-\beta)})^2}{h^2_{GWAS}}, \quad [6]$$

for associations treated as fixed effects in GWAS, and

$$N_r \approx \frac{(Z_{(1-\alpha/2)}+Z_{(1-\beta)})^2(1-h^2)}{h^2_{REG}}, \quad [7]$$

for regions treated as random effects in RHM. The Z values are ordinates of the normal curve. The derivations of these expressions are given in Supporting Information Methods S1. From these equations and for the sample size for each trait analyzed, we estimated the heritability thresholds for significance for the GWAS and regional heritability QTLs using the expressions:

$$h^2_{GWAS} \approx \frac{(Z_{(1-\alpha/2)}+Z_{(1-\beta)})^2}{N_f}, \quad [8]$$

$$h^2_{REG} \approx \frac{(Z_{(1-\alpha/2)}+Z_{(1-\beta)})^2(1-h^2)}{N_r}, \quad [9]$$

for $\alpha = 10^{-5}$ and a power of detection $1 - \beta = 0.90$.

1.4 Results

1.4.1 SNP data and LD

Of the 49,042 genotyped SNPs at a call rate > 95%, 24,806 were polymorphic at MAF > 0.01 and were kept for GWAS and RHM analyses (Table 1). As expected, the r^2 estimates at 1Mb were slightly larger than those at 2Mb, although the genome-wide averages were similar. Considerable differences were seen in the average LD across chromosomes (Table 1), as shown by the LD decay plot (Fig. 1), consistent with the variable recombination rates reported previously in Eucalyptus (SILVA-JUNIOR; GRATTAPAGLIA, 2015). For example, chromosome 6 showed a much slower rate of LD decay when compared with chromosome 3, suggesting considerable differences in their rate of recombination. Using $r^2 < 0.2$ as a canonical threshold for usable LD, the LD decays in this population at < 1Mb, consistent with its small effective population size (Ne = 10) and hybrid origin.

Table 1. Summary statistics of the numbers of single nucleotide polymorphisms (SNPs) genotyped and effectively used in the analyses and the individual chromosome estimates of linkage disequilibrium r^2 in the *Eucalyptus grandis* × *Eucalyptus urophylla* hybrid breeding population

Chr.	Chr. Size (Mb)	Total SNPs	SNPs (MAF>0.01)	r^2 average	r^2 (1 Mb)	r^2 (2 Mb)
1	40.275	3,595	1,874	0.072	0.158	0.145
2	64.221	5,388	2,748	0.055	0.142	0.133
3	79.945	4,860	2,460	0.054	0.134	0.128
4	41.928	3,469	1,769	0.075	0.173	0.159
5	74.729	4,555	2,217	0.053	0.116	0.111
6	53.886	5,425	2,625	0.068	0.212	0.190
7	52.405	3,977	1,945	0.065	0.174	0.159
8	74.308	5,937	3,113	0.050	0.135	0.127
9	39.001	3,557	1,860	0.078	0.179	0.162
10	39.352	3,939	1,977	0.081	0.196	0.177
11	45.396	4,340	2,218	0.064	0.175	0.158
Wide	605.446	49,042	24,806	0.062	0.164	0.153

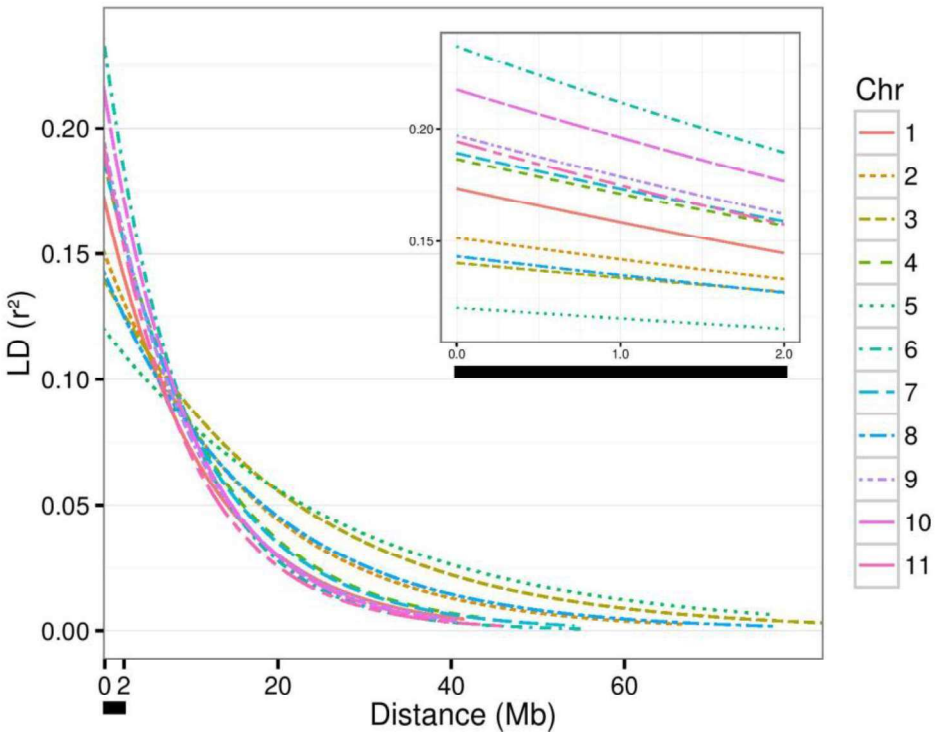

Figure 1. Decay of linkage disequilibrium (LD) estimated by r^2 (y-axis) along physical distance in Mb (x-axis) with correction for family and population structure for each of the 11 *Eucalyptus* chromosomes.

1.4.2 GWAS

Summary statistics and genetic correlations among the seven traits provide a general overview of the range of variation observed in the population (Tables S1, S2). All traits displayed a considerable amount of phenotypic variation and heritabilities were moderate to high (0.36–0.57), with the highest values observed for BWD and DBH. A total of 173,642 association tests was performed (24,806 SNPs vs seven traits). Following multiple testing correction, 13 SNPs with genome-wide significant associations were found. Between one and four associated SNPs were found for each trait, some positioned very close on the same chromosome (e.g. for KN) (Table 2). Fractions of genomic heritability explained by single associations were small, capturing between 3.7% for EA and 6.6% for HEI of the additive genetic variation. The most significant association based on the difference between the $-\log_{10}$ scaled P-value (5.56) and the trait specific threshold (4.13) was found for PPR on chromosome 3, followed by the association for BWD (5.22 to 4.04) on chromosome 2. All SNP variants associated were common with $2p\,(1 - p) > 0.4$, with the exception of an SNP for DBH on chromosome 7.

Table 2. Detected single nucleotide polymorphism (SNP)–trait associations by genome-wide association (GWAS) for the seven traits in the *Eucalyptus grandis* × *Eucalyptus urophylla* hybrid breeding population

Trait	SNP	Significance threshold	Chr.	Position (Mb)	h^2_{GWAS}	–log10	2p(1–p)
DBH	EuBR05s21063566	4.05	5	21.06	0.044	4.11	0.47
DBH	EuBR07s52103787	4.05	7	52.10	0.050	4.39	0.18
HEI	EuBR02s21671254	4.83	2	21.67	0.066	5.29	0.42
BWD	EuBR01s30635188	4.04	1	30.64	0.052	4.21	0.42
BWD	EuBR01s30899860	4.04	1	30.90	0.050	4.09	0.41
BWD	EuBR02s14161950	4.04	2	14.16	0.060	5.22	0.44
BWD	EuBR08s64106447	4.04	8	64.11	0.057	4.78	0.48
EA	EuBR03s65446288	5.02	3	65.45	0.037	5.09	0.41
SPY	EuBR02s42876352	4.61	2	42.88	0.043	5.56	0.42
KN	EuBR02s42875938	4.1	2	42.88	0.038	4.18	0.42
KN	EuBR02s42888917	4.1	2	42.89	0.038	4.19	0.42
KN	EuBR02s42997872	4.1	2	43.00	0.038	4.20	0.42
PPR	EuBR03s56400715	4.13	3	56.40	0.041	5.56	0.07

Also listed is the heritability explained by each association (h^2_{GWAS}), the P-value expressed as $-\log_{10}$ and the expected heterozygosity of the associated SNP ($2p(1-p)$). [1]Permutation threshold is the $-\log_{10}$ scaled P value. 2p is the minor allele frequency of the associated SNP. DBH, diameter at breast height; HEI, height growth; BWD, basic wood density; EA, effective alkali; SPY, screened pulp yield; KN, pulp bleaching kappa number; PPR, Puccinia psidii rust disease resistance.

1.4.3 Regional Heritability Mapping

A total of 603 genomic windows was subjected to RHM, each covering a variable number of polymorphic SNPs, providing an average genome-wide scanning density of one SNP every 24.3 kb. Twenty-six QTLs were mapped by RHM, each encompassing between 41 and 150 SNPs across six chromosomes (Table 3). Regional QTLs for DBH, HEI, SPY, KN and PPR were located on single chromosomes, whereas QTLs for BWD and EA were mapped on several different chromosomes. Nearby QTLs on chromosome 2 were detected for HEI, SPY and KN. Although SPY and KN are correlated traits (Table S2), the overlapping result for HEI was unexpected. Overlapping QTLs for PPR and EA on chromosome 3 were observed, consistent with putatively shared genetic control of secondary metabolites of the lignin pathway involved in disease resistance and pulp yield (see the Discussion section). Heritabilities captured by individual regional QTLs were, on average, higher (0.062) than the estimates for the GWAS QTLs (0.047), and a number of regional QTLs explained considerably larger proportions of the additive variation (e.g. on chromosome 5 for BWD, $h^2_{REG} = 0.154$, and for DBH, $h^2_{REG} = 0.106$) (Table 3). A heat map of the genome-wide distribution of RHM QTLs, with corresponding heritabilities explained, summarizes the results (Fig. 2). In addition to the regional QTLs that surpassed the significance threshold, indicated with arrows, the heat map also displays suggestive additional associations across the genome, such as those on chromosomes 2 and 7 for DBH, for example.

Table 3. Results of quantitative trait loci (QTL) detection via regional heritability mapping (RHM) using 2-Mb genomic segments with 1-Mb sliding window for the seven traits in the *Eucalyptus grandis* × *Eucalyptus urophylla* hybrid breeding population

Trait	Chr.	Number of Snps in the region	Region start position (Mb)	SNP at starting position	Region end position (Mb)	SNP at ending position	h^2_{REG}	-log10	Threshold[1]
DBH	5	45	19.53	EuBR05s19534354	21.28	EuBR05s21282371	0.080	3.45	3.13
DBH	5	53	34.5	EuBR05s34502605	36.43	EuBR05s36434115	0.106	4.03	3.13
HEI	2	105	23.5	EuBR02s23504270	25.50	EuBR02s25496072	0.039	3.24	3.10
HEI	2	101	31.54	EuBR02s31535213	33.46	EuBR02s33462558	0.046	3.90	3.10
HEI	2	65	36.51	EuBR02s36508813	38.43	EuBR02s38425707	0.042	3.62	3.10
HEI	2	43	37.55	EuBR02s37551409	39.49	EuBR02s39485452	0.042	3.54	3.10
HEI	2	121	46.58	EuBR02s46578051	48.49	EuBR02s43498180	0.044	3.18	3.10
BWD	1	121	30.51	EuBR01s30510351	32.47	EuBR01s32472039	0.099	2.86	2.55
BWD	5	150	2.5	EuBR05s19534354	4.49	EuBR05s4493894	0.154	3.95	2.55
BWD	5	130	3.54	EuBR05s3537597	5.50	EuBR05s5496276	0.124	3.63	2.55
BWD	7	114	49.52	EuBR07s49523369	51.49	EuBR07s51489213	0.087	2.71	2.55
EA	1	51	4.88	EuBR01s4878866	6.49	EuBR01s6490451	0.041	2.88	2.65
EA	3	65	64.1	EuBR03s64100540	65.45	EuBR03s65446465	0.052	2.66	2.65
EA	3	64	64.52	EuBR03s64517190	66.38	EuBR03s65446465	0.052	2.68	2.65
EA	8	41	22.66	EuBR08s22660709	24.41	EuBR08s24409285	0.072	3.44	2.65
SPY	2	58	20.52	EuBR02s20516940	22.50	EuBR02s22497719	0.045	2.91	2.99
SPY	2	101	31.54	EuBR02s31535213	33.46	EuBR02s33462558	0.041	3.00	2.99
SPY	2	65	36.51	EuBR02s36508813	38.43	EuBR02s38425707	0.041	3.13	2.99
SPY	2	104	41.56	EuBR02s41557558	43.50	EuBR02s43498180	0.039	3.42	2.99
SPY	2	136	42.51	EuBR02s42511757	44.49	EuBR02s44494452	0.055	4.03	2.99
KN	2	104	41.56	EuBR02s41557558	43.50	EuBR02s43498180	0.040	3.22	2.94
KN	2	136	42.51	EuBR02s42511757	44.49	EuBR02s44494452	0.059	4.03	2.94
PPR	3	52	54.52	EuBR03s54520115	56.46	EuBR03s56463896	0.050	3.48	2.71
PPR	3	44	56.32	EuBR03s56321245	57.50	EuBR03s57499880	0.058	4.03	2.71
PPR	3	65	58.51	EuBR03s58507251	60.49	EuBR03s60492240	0.045	2.89	2.71
PPR	3	57	59.69	EuBR03s59692264	61.44	EuBR03s61444996	0.049	3.10	2.71

Also listed is the heritability explained by each regional QTL (h^2_{REG}) and the P-value of the detected QTL expressed as $-\log_{10}$. [1]Threshold used to declare significance calculated by permutation (also in the $-\log_{10}$ scaled P-value). SNP, single nucleotide polymorphism; DBH, diameter at breast height; HEI, height growth; BWD, basic wood density; EA, effective alkali; SPY, screened pulp yield; KN, pulp bleaching kappa number; PPR, Puccinia psidii rust disease resistance.

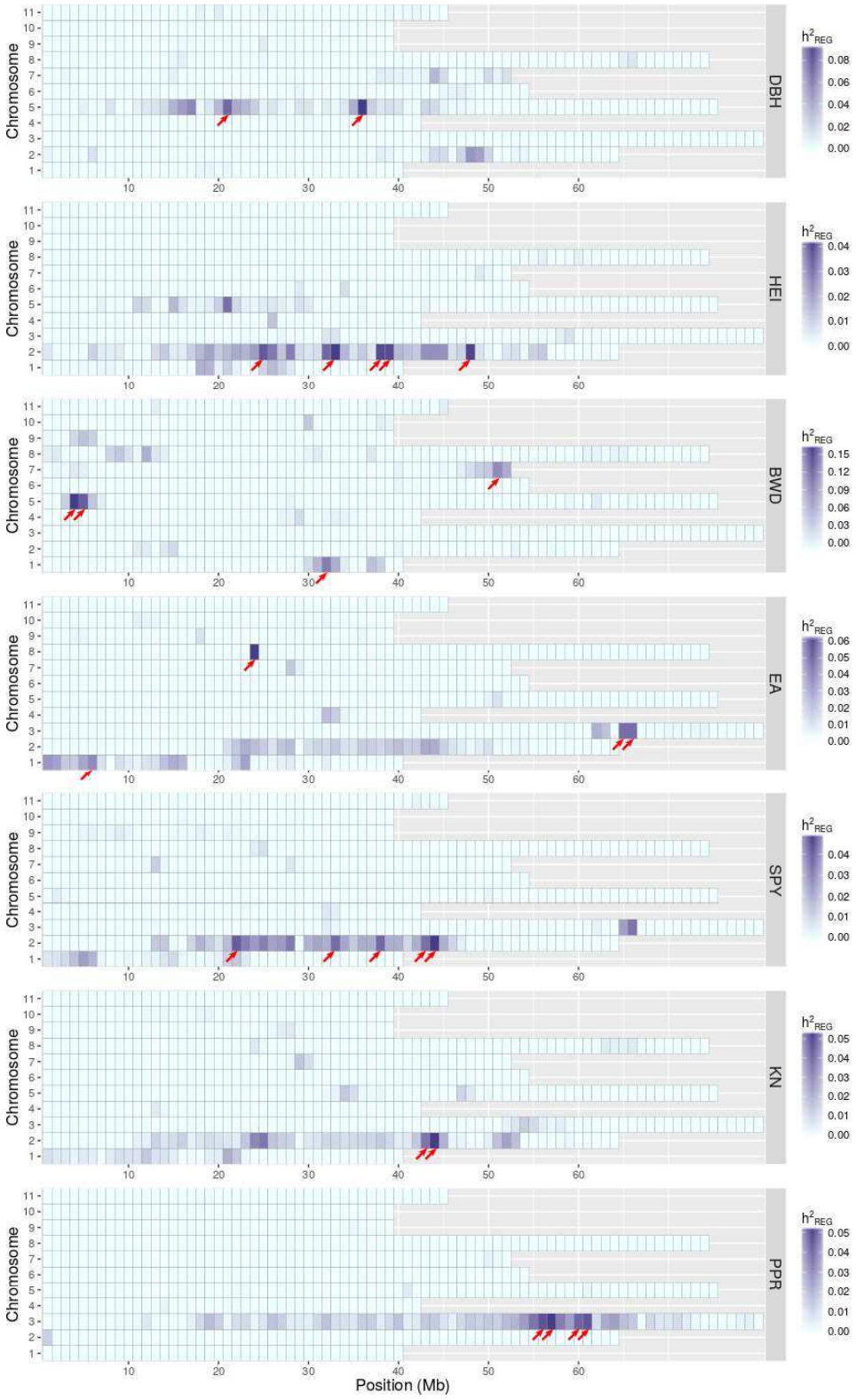

Figure 2. Genome-wide distribution of regional heritability quantitative trait loci (QTLs) mapped by regional heritability mapping (RHM) long the 11 *Eucalyptus* chromosomes (y-axis), subdivided into 1-Mb windows, for the seven traits (right). The heat map bar legend on the right corresponds to the regional heritability estimate. The genomic segments indicated with red arrows were declared as significant RHM QTLs. DBH, diameter at breast height; HEI, height growth; BWD, basic wood density; EA, effective alkali demand; SPY, screened pulp yield; KN, kappa number pulp bleaching content; PPR, *Puccinia psidii* rust disease severity.

1.4.4 Comparative analysis of GWAS and RHM results

Total genomic heritability (h^2_{SNP}) captured between 64 and 89% of the heritability estimated from pedigree data (h^2_{PED}) (Table 4). Larger fractions of trait heritabilities were captured by the SNP data for wood properties when compared with growth traits. When comparing the aggregate fractions of the genomic heritability explained by the two mapping methods, clearly, RHM was superior to GWAS for all traits, capturing, in general, twice or three times the amount of genomic heritability, as in the case of BWD (69% vs 33%), HEI (78% vs 26%), EA (51% vs 14%) and PPR (63% vs 19%) (Table 4). Of the 13 associations detected by GWAS, nine were also detected by RHM overlapping in the same intervals. A visual summary of the comparative results and genomic positions of the QTLs found by GWAS and RHM is provided in a Manhattan plot (Fig. 3). Overall, the two approaches tend to pinpoint the same genomic regions, although the RHM approach more frequently reaches the threshold. Interestingly, the Manhattan plot also shows a contrast of the global profile of P-values between the growth traits (DBH, HEI and BWD) when compared with the wood chemical traits and PPR. A much larger number of SNPs or regions show a signal for growth traits, despite not reaching significance, suggesting a more complex genetic architecture, as opposed to a less complex one for the wood chemical and disease resistance traits studied.

Table 4. Trait heritability (pedigree data) (h^2_{PED}), genomic heritability (single nucleotide polymorphism (SNP) data) (h^2_{SNP}) and fraction of genomic heritability captured by the associations (n_{GWAS}) detected by genome-wide association (GWAS) (h^2_{GWAS}) and by the (n_{RHM}) quantitative trait loci (QTLs) mapped by regional heritability mapping (RHM) (h^2_{REG}) in the *Eucalyptus grandis* × *Eucalyptus urophylla* hybrid breeding population

Trait	N	n_{RHM}	n_{GWAS}	h^2_{PED}	h^2_{SNP}	h^2_{GWAS}	h^2_{REG}
DBH	768	2	2	0.53	0.35 (66%)1	0.09 (26%)2	0.19 (54%)3
HEI	768	5	1	0.42	0.27 (64%)	0.07 (26%)	0.21 (78%)
BWD	764	4	4	0.69	0.67 (97%)	0.22 (33%)	0.46 (69%)
EA	761	4	1	0.49	0.36 (73%)	0.05 (14%)	0.22 (61%)
SPY	761	5	1	0.47	0.42 (89%)	0.05 (12%)	0.22 (52%)
KN	761	2	3	0.34	0.22 (65%)	0.11 (50%)	0.10 (45%)
PPR	559	4	1	0.36	0.32 (89%)	0.06 (19%)	0.20 (63%)

[1] h^2_{SNP}/h^2_{PED}; [2] h^2_{GWAS}/h^2_{SNP}; [3] h^2_{REG}/h^2_{SNP} (%). DBH, diameter at breast height; HEI, height growth; BWD, basic wood density; EA, effective alkali; SPY, screened pulp yield; KN, pulp bleaching kappa number; PPR, Puccinia psidii rust disease resistance.

1.4.5 Threshold heritabilities for GWAS and RHM

As expected from theory, threshold values for GWAS were considerably higher than those for RHM (Table S3). The lowest heritability estimates obtained for the GWAS hits and for the regional QTLs, respectively, were at least the same (for EA in GWAS) or higher than the theoretically expected thresholds at a power of 90%. Therefore, all the declared associations and QTLs were associated at a power level of at least 90%, strengthening the reliability of the GWAS and RHM results. This analysis also shows that the RHM approach requires smaller expected threshold values to reach the same power of the GWAS approach. This is most likely a result of the fact that SNPs are treated as random effects in RHM modeling, whereas they are fitted as fixed effects in GWAS.

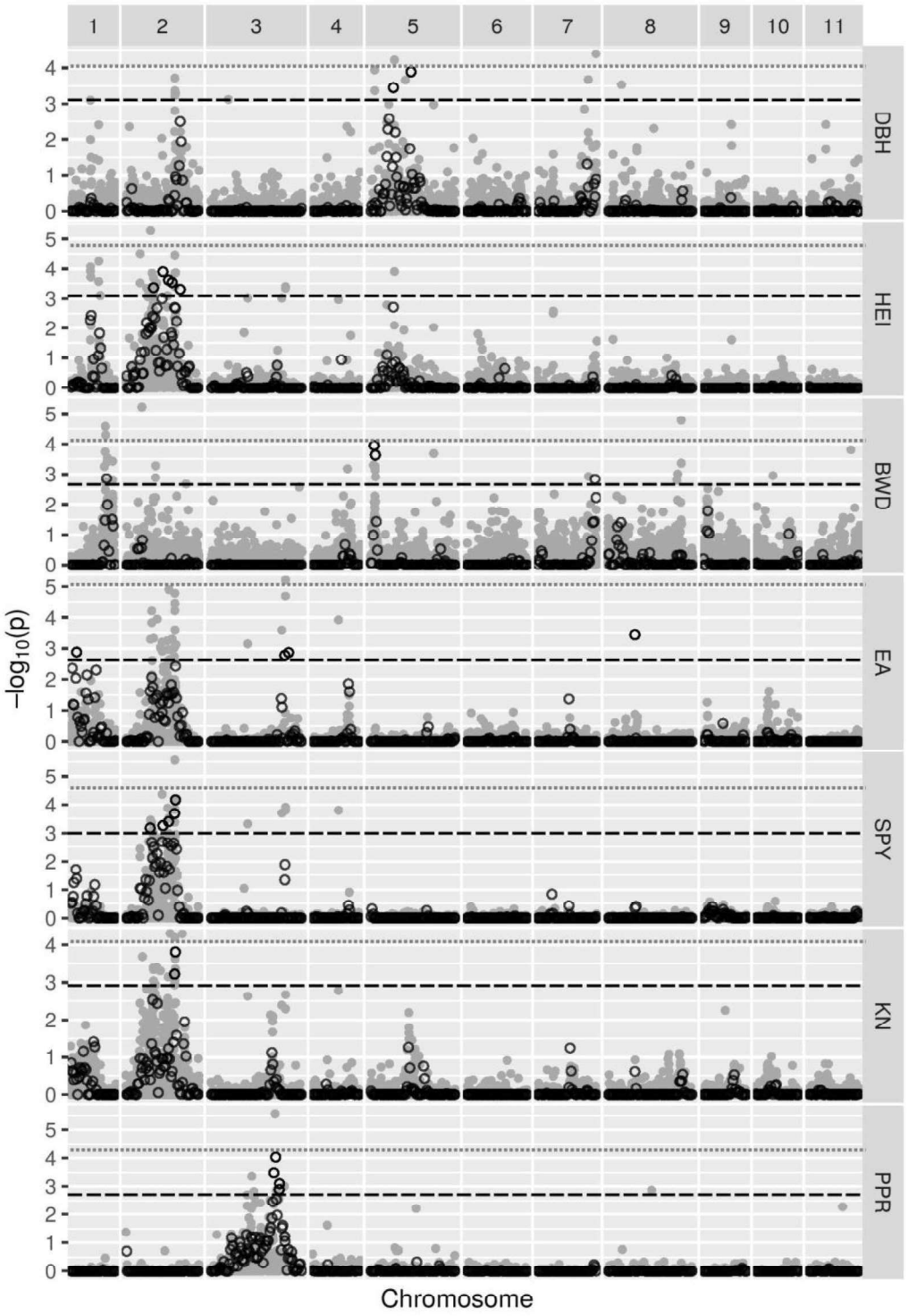

Figure 3. Manhattan plots of the genome-wide association study (GWAS, gray solid circles) and regional heritability mapping (RHM, black open circles) results for the seven traits (right y-axis): DBH, diameter at breast height; HEI, height growth; BWD, basic wood density; EA, effective alkali demand for bleached pulp; SPY, screened pulp yield; KN, kappa number pulp bleaching content; PPR, Puccinia psidii rust disease severity; based on the $-\log_{10}$ P-value (left y-axis) along the 11 *Eucalyptus* chromosomes (x-axis). The gray horizontal dotted line indicates the significant threshold level after permutation test for a genome-wide significant level of 5% by Bonferroni correction for GWAS, and the black horizontal dashed line indicates the significant threshold level of genome-wide significance after permutation test at 5% Bonferroni correction for RHM.

1.5 Discussion

This study represents a further step towards the identification of the genomic regions underlying complex growth, wood properties and *Puccinia rust* resistance traits in *Eucalyptus*. Although both GWAS and RHM successfully identified associations, RHM confirmed its expected superior performance when compared with GWAS. By combining the power of linkage mapping with the ability of association analysis to capture variance across the whole population, RHM accounted for larger fractions of the additive genetic variance, probably as a result of the many segregating alleles at the tagged locus and the combined effect of several closely linked loci in the mapped region. To the best of our knowledge, this is the first study to apply RHM in plants.

Association studies in forest trees have mostly targeted candidate genes and, only recently, the first genome-wide analyses have been reported in *Populus* (EVANS et al., 2014; MCKOWN et al., 2014; PORTH et al., 2013) and *Eucalyptus* (CAPPA et al., 2013). These experiments were carried out using collections of trees derived from natural populations with no selection. The objective of these studies was to maximize the probability of detecting associations at the level of genes that would potentially support tree breeding efforts based on tracking their desirable allelic variants. However, despite the well-intentioned rhetoric, it remains to be seen how such SNP–trait associations found in GWAS in undomesticated natural populations, far removed from selected breeding material, will be translated into useful information to breeding. In our study, we had a different perspective. We were interested in evaluating the performance of alternative genome wide mapping approaches in an operational breeding population that had undergone selection to pinpoint regions that would capture larger fractions of the additive genetic variance. Although less genetic variation is available in a closed elite breeding population, associations found in such selected material should be considerably more useful to inform practical breeding decisions, including whole-genome prediction, a concept successfully explored in a recent rice GWAS (BEGUM et al., 2015). Low-frequency alleles in natural populations become considerably more common when sampled in closed breeding populations and, once revealed by genome-wide mapping approaches, can be easily tracked in generations of breeding. Moreover, as single SNP associations have been consistently shown to explain small proportions of the heritability, and given the extensive LD in our

population, we aimed not to discover genes, but rather to estimate the proportion of heritability explained by the different ways in which GWAS and RHM exploit the SNP data.

1.5.1 Genome coverage and LD

This is the first experimental study to use dense whole-genome SNP data to investigate phenotype–genotype associations in species of *Eucalyptus*. It constitutes the first truly GWAS in the genus by interrogating the genome at several thousand SNP loci. These results provide a valuable demonstration of the usefulness of the EuCHIP60k for future GWAS or RHM efforts in eucalypts. Of the 56,932 successfully genotyped SNPs, c. 44% (24,806) were polymorphic in this particular population, despite its small effective population size

A larger number of informative SNPs would be observed in populations closer to the wild (SILVA-JUNIOR; FARIA; GRATTAPAGLIA, 2015), although the multi-species nature of the chip was deliberately planned to accommodate several *Eucalyptus* taxa, such that not all 60,000 SNPs on the chip are expected to be informative in any single species. The silver lining of this chip format, however, is that ascertainment bias towards more common SNPs was minimized (SILVA-JUNIOR; FARIA; GRATTAPAGLIA, 2015), such that SNPs were probably sampled across most of the frequency range, a particularly useful advantage for RHM, which aims to capture the joint variation contributed by all SNPs with variable frequency in the mapped interval. Consistent with the reduced effective population size, a considerably slower LD decay at c. 1 Mb was seen in this breeding population when compared with the decay at 4–6 kb reported in natural populations of E. grandis (SILVA-JUNIOR; FARIA; GRATTAPAGLIA, 2015). The regional mapping window of 2 Mb, sliding by 1 Mb, was calibrated to the extent of usable LD, such that the resolution of the RHM was consistent with the proposed approach. However, the long-range LD, although appropriate for the detection of associations, was certainly not ideal to provide the resolution to pinpoint genes, an objective we evidently precluded at the start of this study.

1.5.2 Accounting for population structure in the breeding population

As expected from the germplasm origin of the breeding population (see the Materials and Methods section), both population and family structure were present. STRUCTURE analysis showed the presence of k=2 populations (Fig. S1) corresponding to the two species involved, such that the F2 individuals contain variable proportions of either the E. grandis or E. urophylla genomes. The inclusion of the GRM in the GWAS and RHM models was therefore essential to provide unbiased results. The approach used to estimate genomic heritability did not explicitly specify the existing family and population structures, following an earlier approach (ZAITLEN et al., 2013). In addition, this model using the genomic relationship simultaneously accounts for the fact that many SNPs are in LD. Such a model fits all the SNPs jointly in a random effect model, so that each SNP effect is fitted conditioned on the joint effects of all the other SNPs, therefore accounting for the LD between the SNPs.

1.5.3 Genomic heritabilities capturing trait heritability

Heritabilities estimated for growth traits (Table S1) were in the same range as reported previously for equivalent *Eucalyptus* hybrids (BOUVET; VIGNERON, 1995). Estimated genomic heritabilities accounted for relatively large proportions (64–89%) of trait heritabilities (Table 4). Our results are comparable with those in white spruce (*Picea glauca*), in which a 6,385 SNP model captured 64% and 62% of the additive genetic variance for growth and wood density, respectively (BEAULIEU et al., 2014). Similar results have been reported for interior spruce, in which heritabilities from genomic best linear unbiased prediction (GBLUP) were generally 30–60% of the heritability from pedigree (EL-DIEN et al., 2015). Replacing the average relationship matrix derived from pedigree with the realized relationship matrix has been shown to increase the accuracy of breeding values (HAYES; VISSCHER; GODDARD, 2009), although this will depend on the number of effective loci involved. In forest tree breeding, in which half-sib families are used or full-sib pedigrees may contain errors, GBLUP estimates of heritability have been considered to be advantageous as they allow a more accurate ascertainment of the genealogical relationships among individuals, and consequently provide more realistic gain estimates, as a result of the

adjustment for the Mendelian sampling term (EL-DIEN et al., 2015). Genomic heritability and trait heritability parameters have been shown to be equal only when all causal variants are typed (DE LOS CAMPOS; SORENSEN; GIANOLA, 2015). Furthermore, genomic heritability estimates were also shown to be unbiased when close relatives sharing long chromosome segments are analyzed, such that the patterns of allele sharing at markers and at QTLs will also be similar. Given the log range LD in our small effective size breeding population and the genome-wide SNP coverage adopted, our estimates of genomic heritability should therefore be largely unbiased. Genomic heritability is seen as the amount of variation that would be explained by GWAS when the sample size is so large that all associated variants would be statistically significant (VINKHUYZEN et al., 2013). As proposed in human studies (YANG et al., 2010), our results therefore indicate that the still missing heritability is most probably a result of rare variants having very low MAF. The alternative possibility of imperfect LD with the set of SNPs genotyped is less likely in our high-LD breeding population. Evoking more complex arguments, such as epistasis, for the still unexplained heritable variation does not seem to be warranted (ROBINSON; WRAY; VISSCHER, 2014), as experimental evidence in model systems (MACKAY, 2014) and a recent study in hybrid *Eucalyptus* (BOUVET et al., 2016) converge on the fact that most genetic variation for quantitative traits is additive.

1.5.4 Associations for growth and wood property traits

The proportion of genomic heritability explained by RHM for DBH, HEI and BWD was considerably higher than that by GWAS (Table 4). These results support the expectation that RHM can detect trait-associated regions for complex traits which GWAS does not identify as significant, in line with results described for different traits in humans (NAGAMINE et al., 2012; SHIRALI et al., 2015; UEMOTO et al., 2013) and domestic animals (MATIKA et al., 2016; RIGGIO et al., 2013). BWD showed the highest trait heritability h^2=0.69 and the highest proportion of trait heritability explained by genomic heritability (97%). Both RHM and GWAS detected four associations, but RHM captured twice as much heritability than GWAS (Table 4). It was for BWD that two of the four non-overlapping associations between GWAS and RHM were observed, specifically the GWAS hits on chromosomes 2 and 8. One would expect that

not all loci detected by RHM would have been detected by GWAS, but the opposite seems counterintuitive, unless they correspond to GWAS false positives. This apparent discrepancy can be explained by the fact that the combined effect of a set of SNPs in RHM changes the likelihood of detection comparatively to detecting the effect of a single marker in GWAS. Joint fitting of genomic regions and the GRM in a model and comparison with single marker fitting and GRM in GWAS can therefore lead to changes in the inference. In other words, the contribution of many small effect markers can reach significance in an RHM segment, whereas a large single marker effect can be non-significant in GWAS. However, the opposite is also possible when a region with a relatively large single marker effect detected by GWAS does not reach significance by RHM because of additional small SNP effects fitted in the RHM model that contribute to the variance in opposite directions, cancelling the overall effect of the segment and thus precluding detection or avoiding a false positive. Such observations have been reported in comparative studies between GWAS and RHM for nematode resistance and body weight in sheep (RIGGIO et al., 2013), and blood lipid traits in humans (SHIRALI et al., 2015), where additional evidence from meta-analyses indicated that RHM avoided false positives identified by GWAS.

BWD has been consistently reported as a high heritability trait in *Eucalyptus* (REZENDE; DE RESENDE; DE ASSIS, 2014), suggesting that it might involve some loci of large effect. High heritability, however, does not, by itself, imply that there is a relationship between heritability and the number or effect size of detected genomic regions, nor that any such regions will explain a large proportion of the genetic variance (VISSCHER; HILL; WRAY, 2008). This seems to be the case for BWD. Despite its high trait and genomic heritabilities, the detection of a small number of QTLs suggests that BWD is, in fact, polygenic and that the inability to detect further associations is most probably a result of the existence of many small effects underlying this trait. The same two chromosomes, 2 and 5, in which most GWAS and RHM associations for DBH, HEI and BWD were detected, have been consistently shown to contain QTLs for these same traits in biparental mapping, across different *Eucalyptus* species, since the early studies (GRATTAPAGLIA et al., 1996; VERHAEGEN et al., 1997), up to the more recent ones (FREEMAN et al., 2013; GION et al., 2011; THUMMA et al., 2010a). As expected, correlations amongst wood chemical traits were either positive (EA vs KN) or negative (EA vs SPY and SPY vs KN) (Table S2). All these traits are actual

downstream industrial level traits which are strongly impacted by lignin content and, especially, by the relative proportion of syringyl to guayacil lignin (S:G ratio). A higher S:G ratio reduces the demand for alkali (EA) during wood digestion, increases SPY and lowers KN necessary for wood deconstruction without excessive alkali charge (GOMES et al., 2015). Genomewide association studies and RHM QTLs for wood chemical traits were mapped mainly on chromosomes 2 and 3, for which bi-parental QTL mapping studies have also identified loci for the same or correlated traits (FREEMAN et al., 2013; GION et al., 2011; THUMMA et al., 2010b). Evidently, comparisons of the associations reported here with linkage mapped QTLs in previous studies should be seen as tentative at best, given the large and coarse intervals to which bi-parental QTLs are mapped, their generally overestimated effect and the lack of their exact physical position. As an indirect validation, however, the concentration of growth QTLs on specific chromosomes, especially for growth and BWD on chromosomes 2 and 5, provides additional credibility to the associations reported in our work.

1.5.5 *Eucalyptus* chromosome 3 and fungal resistance loci

All GWAS and RHM associations for PPR were detected on chromosome 3, essentially in the same position between c. 54 and 61 Mb (Tables 2, 3). The first major effect QTL identified for Puccinia rust resistance, Ppr1 (JUNGHANS et al., 2003), was later validated in unrelated pedigrees and positioned on chromosome 3 (MAMANI et al., 2010). Ppr1 on chromosome 3 was again validated in further pedigrees and additional epistatic QTLs were described (ALVES et al., 2012). The genomic heritability captured 89% of the trait heritability, and the RHM associations 63% of the genomic heritability (Table 4). By accounting for relatively large proportions of the additive genetic variance, these results corroborate our early view that, although PPR involves one or more major effect QTLs, it is largely complex and multifactorial in nature (JUNGHANS et al., 2003). In a recent bi-parental mapping study, this same view was corroborated and four additional major effect QTLs were mapped, one again on chromosome 3 at position 57 Mb (BUTLER et al., 2016). Major effect QTLs for resistance to other fungal disease in *Eucalyptus* have also been reported on chromosome 3 for *Mycosphaerella cryptica* leaf disease (FREEMAN; POTTS;

VAILLANCOURT, 2008) and *Ceratocystis fimbriata* wilt (ROSADO et al., 2016). The fact that QTLs for fungal disease resistance have been repeatedly reported on chromosome 3 in independent studies clearly points to a major involvement of this chromosome in the pathogen resistance response. Interestingly, at the gene level, 104 syntenic blocks of genes are shared between *Eucalyptus* chromosome 3 and *Populus* chromosome XVIII, containing 778 genes in 522 gene families, with the most common family in this conserved gene space represented by 33 disease resistance genes (MYBURG et al., 2014). Recently, the highest densities of clusters and superclusters of NBS-LRR (nucleotide binding site–leucine-rich repeat) resistance genes were reported on *Eucalyptus* chromosomes 3, 5, 6, 8 and 10, and a clear overlap of Ppr1 with a supercluster on chromosome 3 at position c. 54Mb was observed in *Eucalyptus* (CHRISTIE et al., 2015), overlapping the same genomic interval in which we mapped our GWAS and RHM hits for PPR. Furthermore, the Manhattan plot for PPR reveals some SNPs almost reaching significance on chromosomes 5, 8 and 11 (Fig. 3), suggesting the presence of several gene effects clustered on these chromosomes, although not reaching significance in our experiment.

1.5.6 Concluding remarks

We showed that RHM successfully captured larger fractions of trait heritability when compared with GWAS. Clearly, however, our experimental population, despite its considerably larger size and extensive LD, only provided power to detect a small fraction of the loci controlling trait heritability. The proportion of the heritability explained by the GWAS hits varied according to the trait, and was probably inflated because of the selective reporting of the significant associations, therefore potentially subject to the ‘winner’s curse’ effect (GARNER, 2007). By precluding the exclusive dependence on single SNP associations, however, the RHM approach enabled us to incorporate effects over multiple causative variants, thus providing a joint estimate of the combined effects of common and rare variants in the genomic regions detected (NAGAMINE et al., 2012). One would therefore expect that regions known to contain effects large enough to be detected by GWAS would always be captured by RHM, whereas the opposite would not necessarily be true because of the additional small effect variants accounted for by RHM. Our results, however, showed that this was not always the

case, and coincidence in genomic position between GWAS and RHM QTLs varied depending on the trait, with wood chemical traits and PPR showing better coincidence than growth traits, highlighting the higher complexity of the latter.

Despite the availability of a reference genome for *Eucalyptus*, an attempt to relate or co-locate our findings with previous QTL mapping or gene models is very speculative. We have, however, pointed out a few plausible, although coarse, genomic regions that would merit further examination towards specific gene discovery if that becomes an objective. These include the regions highlighted on chromosomes 2 and 5 for growth traits, and chromosome 3 for fungal disease resistance. Nevertheless, these discrete associations represent only a fraction of those that control the traits investigated, and even if genes were found and validated, a considerable proportion of the heritability would still be left missing. This is in line with results from whole-genome prediction for growth and wood traits in *Eucalyptus* (RESENDE et al., 2012), showing that large proportions of trait heritability were captured only when all genome-wide markers were considered simultaneously without the application of any rigorous statistical test.

This study also substantiated the fact that genome-wide SNP data were able to account for large proportions (53–92%) of trait heritability for all traits. It showed, however, that the total genomic heritability was larger than the fraction explained by the statistically significant associations detected. Notwithstanding the difference between the structure of our breeding population and the natural populations used in previous reports (CAPPA et al., 2013; EVANS et al., 2014; MCKOWN et al., 2014; PORTH et al., 2013), all of these GWAS results converge to an increasingly undisputable view that traits of ecological and economic interest in forest trees are controlled by many sequence variants across the frequency spectrum, each with only a small average contribution to the phenotypic variance. Moreover, these common undetected variants probably account for the large difference between the heritability obtained by GWAS or RHM hits and the genomic heritability estimated from the use of all SNPs, such that, rather than ‘missing’, we should refer to it as ‘hidden’ heritability (VINKHUYZEN et al., 2013). The challenge to detect more of such variants will require considerably larger sample sizes, possibly several tens of thousands trees, better phenotyping and the integration of multiple sources of phenotypic and genetic information (ROBINSON; WRAY; VISSCHER, 2014).

In the meantime, whole-genome regression methods, although unable and agnostic to the identification of genes, but proven to provide a powerful approach to the incorporation of genomic data into selection decisions (MEUWISSEN; HAYES; GODDARD, 2001), are delivering the long expected impact of genomics into tree breeding that association genetics alone has not yet been able (GRATTAPAGLIA, 2014; GRATTAPAGLIA; RESENDE, 2011; RESENDE et al., 2012). Nevertheless, robust local mapping data from RHM QTLs, bearing a combination of common and rare variants contributing large fractions of the heritability, will be useful to enhance the predictive ability of whole-genome regression. RHM provides information on the genetic architectures of traits which can be used by assigning locus- or trait-specific priors to genomic prediction models (DAETWYLER et al., 2010). By assigning different marker weights in building a trait-specific numerator relationship matrix, improvements in prediction accuracies have been achieved in recent experimental studies (GAO et al., 2015; ZHANG et al., 2014). The incorporation of the RHM data reported in this work into whole-genome prediction models should prove a promising avenue for upcoming research.

1.6 References

ALVES, A. A. et al. Genetic mapping provides evidence for the role of additive and non-additive QTLs in the response of inter-specific hybrids of Eucalyptus to Puccinia psidii rust infection. **Euphytica**, v. 183, n. 1, p. 27–38, 2012.

ANONI, C. DE O. **Mapeamento de QTLs em progênie de irmãos-completos de cana-de-açúcar utilizando imputação múltipla**Escola Superior de Agricultura “Luiz de Queiroz, , 2011.

BAZAKOS, C. et al. New Strategies and Tools in Quantitative Genetics: How to Go from the Phenotype to the Genotype. **Annual Review of Plant Biology**, n. 0, 2017.

BEAULIEU, J. et al. Accuracy of genomic selection models in a large population of open-pollinated families in white spruce. **Heredity**, v. 113, n. 4, p. 343–52, 2014.

BEAVIS, W. D. QTL Analyses: Power, Precision, and Accuracy. In: **Molecular Dissection of Complex Traits**. [s.l: s.n.]. v. 3p. 145–162.

BEGUM, H. et al. Genome-wide association mapping for yield and other agronomic traits in an elite breeding population of tropical rice (Oryza sativa). **PloS one**, v. 10, n. 3, p. e0119873, 2015.

BERNARDO, R. Molecular markers and selection for complex traits in plants: learning from the last 20 years. **Crop science**, v. 48, n. 5, p. 1649–1664, 2008.

BODMER, W.; TOMLINSON, I. Rare genetic variants and the risk of cancer. **Current opinion in genetics & development**, v. 20, n. 3, p. 262–267, 2010.

BOER, M. P. et al. A mixed-model quantitative trait loci (QTL) analysis for multiple-environment trial data using environmental covariables for QTL-by-environment interactions, with an example in maize. **Genetics**, v. 177, n. 3, p. 1801–1813, 2007.

BOUVET, J.-M. et al. Modeling additive and non-additive effects in a hybrid population using genome-wide genotyping: prediction accuracy implications. **Heredity**, v. 116, n. 2, p. 146–157, 2016.

BOUVET, J.-M.; VIGNERON, P. Age trends in variances and heritabilities in Eucalyptus factorial mating designs. **Silvae genetica**, v. 44, n. 4, p. 206–216, 1995.

BUTLER, J. B. et al. Evidence for different QTL underlying the immune and hypersensitive responses of Eucalyptus globulus to the rust pathogen Puccinia psidii. **Tree Genetics & Genomes**, v. 12, n. 3, p. 1–13, 2016.

CABALLERO, A.; TENESA, A.; KEIGHTLEY, P. D. The Nature of Genetic Variation for Complex Traits Revealed by GWAS and Regional Heritability Mapping Analyses. **Genetics**, v. 201, n. 4, p. 1601–1613, 2015.

CAPPA, E. P. et al. 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**, v. 8, n. 11, p. e81267, 2013.

CASTRO, C. A. DE O. et al. Brief history of Eucalyptus breeding in Brazil under

perspective of biometric advances. **Ciência Rural**, v. 46, n. 9, p. 1585–1593, 2016.

CHRISTIE, N. et al. The Eucalyptus grandis NBS-LRR gene family: physical clustering and expression hotspots. **Frontiers in plant science**, v. 6, 2015.

CLIFFORD, D.; MCCULLAGH, P.; CLIFFORD, M. D. The regress package. **R package version**, p. 1–3, 2012.

DAETWYLER, H. D. et al. The impact of genetic architecture on genome-wide evaluation methods. **Genetics**, v. 185, n. 3, p. 1021–1031, 2010.

DE LOS CAMPOS, G.; SORENSEN, D.; GIANOLA, D. Genomic heritability: what is it? **PLoS Genet**, v. 11, n. 5, p. e1005048, 2015.

DENIS, M. et al. Genetic variation of wood chemical traits and association with underlying genes in Eucalyptus urophylla. **Tree genetics & genomes**, v. 9, n. 4, p. 927–942, 2013.

DILLON, S. K. et al. Association genetics in Corymbia citriodora subsp. variegata identifies single nucleotide polymorphisms affecting wood growth and cellulosic pulp yield. **New Phytologist**, v. 195, n. 3, p. 596–608, 2012.

DOERGE, R. W. Mapping and analysis of quantitative trait loci in experimental populations. **Nature Reviews Genetics**, v. 3, n. 1, p. 43–52, 2002.

DOERGE, R. W.; ZENG, Z. B.; WEIR, B. S. Statistical issues in the search for genes affecting quantitative traits in experimental populations. **Statistical Science**, p. 195–219, 1997.

EL-DIEN, O. G. et al. Prediction accuracies for growth and wood attributes of interior spruce in space using genotyping-by-sequencing. **BMC genomics**, v. 16, n. 1, p. 370, 2015.

EVANNO, G.; REGNAUT, S.; GOUDET, J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. **Molecular ecology**, v. 14, n. 8, p. 2611–2620, 2005.

EVANS, L. M. et al. Population genomics of Populus trichocarpa identifies signatures of selection and adaptive trait associations. **Nature genetics**, v. 46, n. 10, p. 1089–1096, 2014.

FALCONER, D. S.; MACKAY, T. F. C. 1996. **Introduction to quantitative genetics**, v. 4, 1996.

FREEMAN, J. S. et al. Stability of quantitative trait loci for growth and wood properties across multiple pedigrees and environments in Eucalyptus globulus. **New Phytologist**, v. 198, n. 4, p. 1121–1134, 2013.

FREEMAN, J. S.; POTTS, B. M.; VAILLANCOURT, R. E. Few Mendelian genes underlie the quantitative response of a forest tree, Eucalyptus globulus, to a natural fungal epidemic. **Genetics**, v. 178, n. 1, p. 563–571, 2008.

GAO, N. et al. Improving accuracy of genomic prediction by genetic architecture based priors in a Bayesian model. **BMC genetics**, v. 16, n. 1, p. 120, 2015.

GARCIA, A. A. F. et al. Quantitative trait loci mapping and the genetic basis of heterosis in maize and rice. **Genetics**, v. 180, n. 3, p. 1707–1724, 2008.

GARNER, C. Upward bias in odds ratio estimates from genome-wide association studies. **Genetic epidemiology**, v. 31, n. 4, p. 288–295, 2007.

GION, J.-M. et al. Comprehensive genetic dissection of wood properties in a widely-grown tropical tree: Eucalyptus. **Bmc Genomics**, v. 12, n. 1, p. 301, 2011.

GODDARD, M. E. et al. Estimating effects and making predictions from genome-wide marker data. **Statistical Science**, v. 24, n. 4, p. 517–529, 2009.

GOLLE, D. P. et al. Forestry improvement: emphasis on biotechnology application. **Ciência Rural**, v. 39, n. 5, p. 1607–1614, 2009.

GOMES, F. J. B. et al. Evaluation of alkaline deconstruction processes for Brazilian new generation of eucalypt clones. **Industrial Crops and Products**, v. 65, p. 477–487, 2015.

GRATTAPAGLIA, D. et al. Genetic mapping of quantitative trait loci controlling growth and wood quality traits in Eucalyptus grandis using a maternal half-sib family and RAPD markers. **Genetics**, v. 144, n. 3, p. 1205–1214, 1996.

GRATTAPAGLIA, D. et al. Progress in Myrtaceae genetics and genomics: Eucalyptus as the pivotal genus. **Tree Genetics & Genomes**, v. 8, n. 3, p. 463–508, 2012.

GRATTAPAGLIA, D. 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.

GRATTAPAGLIA, D.; RESENDE, M. D. V. Genomic selection in forest tree breeding. **Tree Genetics & Genomes**, v. 7, n. 2, p. 241–255, 2011.

GUNDERSON, K. L. et al. A genome-wide scalable SNP genotyping assay using microarray technology. **Nature genetics**, v. 37, n. 5, p. 549–554, 2005.

HAYES, B. J.; VISSCHER, P. M.; GODDARD, M. E. Increased accuracy of artificial selection by using the realized relationship matrix. **Genetics research**, v. 91, n. 1, p. 47–60, 2009.

HOCHBERG, Y. A sharper Bonferroni procedure for multiple tests of significance. **Biometrika**, v. 75, n. 4, p. 800–802, 1988.

JUNGHANS, D. T. et al. Resistance to rust (Puccinia psidii Winter) in Eucalyptus: mode of inheritance and mapping of a major gene with RAPD markers. **Theoretical and Applied Genetics**, v. 108, n. 1, p. 175–180, 2003.

LEE, S. et al. Rare-variant association analysis: study designs and statistical tests. **The American Journal of Human Genetics**, v. 95, n. 1, p. 5–23, 2014.

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

MACKAY, T. F. C. Epistasis and quantitative traits: using model organisms to study gene-gene interactions. **Nature Reviews Genetics**, v. 15, n. 1, p. 22–33, 2014.

MAMANI, E. M. C. et al. Positioning of the major locus for Puccinia psidii rust resistance (Ppr1) on the Eucalyptus reference map and its validation across unrelated pedigrees. **Tree Genetics & Genomes**, v. 6, n. 6, p. 953–962, 2010.

MANDROU, E. et al. A candidate gene for lignin composition in Eucalyptus: cinnamoyl-CoA reductase (CCR). **Tree Genetics & Genomes**, v. 8, n. 2, p. 353–364, 2012.

MANGIN, B. et al. Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. **Heredity**, v. 108, n. 3, p. 285–291, 2012.

MATIKA, O. et al. Genome-wide association reveals QTL for growth, bone and in vivo carcass traits as assessed by computed tomography in Scottish Blackface lambs. **Genetics Selection Evolution**, v. 48, n. 1, p. 1, 2016.

MCKOWN, A. D. et al. Genome-wide association implicates numerous genes underlying ecological trait variation in natural populations of Populus trichocarpa. **New Phytologist**, v. 203, n. 2, p. 535–553, 2014.

MEUWISSEN, T. H. E.; 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.

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

NAGAMINE, Y. et al. Localising loci underlying complex trait variation using Regional Genomic Relationship Mapping. **PloS one**, v. 7, n. 10, p. e46501, 2012.

NEALE, D. B.; KREMER, A. Forest tree genomics: growing resources and applications. **Nature Reviews Genetics**, v. 12, n. 2, p. 111–122, 2011.

NEALE, D. B.; SAVOLAINEN, O. Association genetics of complex traits in conifers. **Trends in plant science**, v. 9, n. 7, p. 325–330, 2004.

PATTERSON, H. D.; THOMPSON, R. Recovery of inter-block information when block sizes are unequal. **Biometrika**, p. 545–554, 1971.

PORTH, I. et al. Genome-wide association mapping for wood characteristics in Populus identifies an array of candidate single nucleotide polymorphisms. **New Phytologist**, v. 200, n. 3, p. 710–726, 2013.

PRITCHARD, J. K.; STEPHENS, M.; DONNELLY, P. Inference of population structure using multilocus genotype data. **Genetics**, v. 155, n. 2, p. 945–959, 2000.

RESENDE, M. D. V. **Análise estatística de modelos mistos via REML/BLUP na experimentação em melhoramento de plantas perenes**. [s.l.] Embrapa Florestas, 2000.

RESENDE, M. D. V et al. Genomic selection for growth and wood quality in Eucalyptus: capturing the missing heritability and accelerating breeding for complex

traits in forest trees. **New Phytologist**, v. 194, n. 1, p. 116–128, 2012.

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**. [s.l.] Springer, 2014. p. 393–424.

RIGGIO, V. et al. Genome-wide association and regional heritability mapping to identify loci underlying variation in nematode resistance and body weight in Scottish Blackface lambs. **Heredity**, v. 110, n. 5, p. 420–429, 2013.

RIGGIO, V.; PONG-WONG, R. **Regional Heritability Mapping to identify loci underlying genetic variation of complex traits**. BMC proceedings. **Anais**...BioMed Central, 2014

ROBINSON, M. R.; WRAY, N. R.; VISSCHER, P. M. Explaining additional genetic variation in complex traits. **Trends in Genetics**, v. 30, n. 4, p. 124–132, 2014.

ROSADO, C. C. G. et al. QTL mapping for resistance to Ceratocystis wilt in Eucalyptus. **Tree Genetics & Genomes**, v. 12, n. 4, p. 1–10, 2016.

SEHGAL, D.; SINGH, R.; RAJPAL, V. R. Quantitative Trait Loci Mapping in Plants: Concepts and Approaches. In: **Molecular Breeding for Sustainable Crop Improvement**. [s.l.] Springer, 2016. p. 31–59.

SHIRALI, M. et al. Regional heritability mapping method helps explain missing heritability of blood lipid traits in isolated populations. **Heredity**, 2015.

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.

SILVA-JUNIOR, O. B.; GRATTAPAGLIA, D. Genome-wide patterns of recombination, linkage disequilibrium and nucleotide diversity from pooled resequencing and single nucleotide polymorphism genotyping unlock the evolutionary history of Eucalyptus grandis. **New Phytologist**, v. 208, n. 3, p. 830–845, 2015.

SOUZA, A. P. DE. Biologia molecular aplicada ao melhoramento. **Recursos genéticos e melhoramento–plantas. Rondonópolis: Fundação MT**, p. 939–966, 2001.

TANG, C. S.; FERREIRA, M. A. R. A gene-based test of association using canonical correlation analysis. **Bioinformatics**, v. 28, n. 6, p. 845–850, 2012.

THAVAMANIKUMAR, S. et al. Association mapping for wood quality and growth traits in Eucalyptus globulus ssp. globulus Labill identifies nine stable marker-trait associations for seven traits. **Tree genetics & genomes**, v. 10, n. 6, p. 1661–1678, 2014.

THUMMA, B. R. et al. Polymorphisms in cinnamoyl CoA reductase (CCR) are associated with variation in microfibril angle in Eucalyptus spp. **Genetics**, v. 171, n. 3, p. 1257–1265, 2005.

THUMMA, B. R. et al. Identification of a cis-acting regulatory polymorphism in a eucalypt COBRA-like gene affecting cellulose content. **Genetics**, v. 183, n. 3, p. 1153–1164, 2009.

THUMMA, B. R. et al. Quantitative trait locus (QTL) analysis of wood quality traits in Eucalyptus nitens. **Tree Genetics & Genomes**, v. 6, n. 2, p. 305–317, 2010a.

THUMMA, B. R. et al. Quantitative trait locus (QTL) analysis of growth and vegetative propagation traits in Eucalyptus nitens full-sib families. **Tree Genetics & Genomes**, v. 6, n. 6, p. 877–889, 2010b.

UEMOTO, Y. et al. The power of regional heritability analysis for rare and common variant detection: simulations and application to eye biometrical traits. **Frontiers in genetics**, v. 4, p. 232, 2013.

USAI, M. G. et al. **XVI th QTLMAS: simulated dataset and comparative analysis of submitted results for QTL mapping and genomic evaluation**. BMC proceedings. **Anais**...BioMed Central, 2014

VENCOVSKY, R. Amostragem genética em populações naturais. **Silvicultura, São Paulo**, v. 41, p. 95–96, 1986.

VERHAEGEN, D. et al. Quantitative trait dissection analysis in Eucalyptus using RAPD markers: 1. Detection of QTL in interspecific hybrid progeny, stability of QTL expression across different ages. **Theoretical and Applied Genetics**, v. 95, n. 4, p. 597–608, 1997.

VINKHUYZEN, A. A. E. et al. Estimation and partition of heritability in human populations using whole-genome analysis methods. **Annual review of genetics**, v. 47, p. 75–95, 2013.

VISSCHER, P. M.; HILL, W. G.; WRAY, N. R. Heritability in the genomics era—concepts and misconceptions. **Nature Reviews Genetics**, v. 9, n. 4, p. 255–266, 2008.

WHITT, S. R.; BUCKLER, E. S. Plant functional genomics: Methods and protocols. 2003.

WU, M. C. et al. Rare-variant association testing for sequencing data with the sequence kernel association test. **The American Journal of Human Genetics**, v. 89, n. 1, p. 82–93, 2011.

YANG, J. et al. Common SNPs explain a large proportion of the heritability for human height. **Nature genetics**, v. 42, n. 7, p. 565–569, 2010.

ZAITLEN, N. et al. Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. **PLoS Genet**, v. 9, n. 5, p. e1003520, 2013.

ZHANG, Z. et al. Improving the accuracy of whole genome prediction for complex traits using the results of genome wide association studies. **PloS one**, v. 9, n. 3, p. e93017, 2014.

1.7 R scripts

1.7.2 GWAS procedure

```
library(rrBLUP)
library(lme4)
library(regress)
source("gbuild.R");

##Carregando adaptando os dados
struc <- read.table("Dados/popstruc.txt",header=TRUE); struc$geno <-
as.character(struc$geno-1)
dat <- read.table("Dados/data.txt",header=TRUE) #dados fenotípicos
dat$BWD[dat$BWD<200] <- NA;
M <- read.table("Dados/M.txt",header=TRUE) #Matriz de marcas
map <- read.table("Dados/map.txt",header=TRUE)[,1:3]
colnames(map) <- c("mkr","Chr","Pos")
map$Pos <- map$Pos/1e+06

codes <-
as.numeric(sapply(as.character(rownames(M)),function(x){strsplit(x,"CNB")
[[1]][2]}))+10
rownames(M) <- codes

dat$id <- as.character(dat$id)

map$Dist <- round(map$Pos)+1 #Generating position classes
map$Chr <- as.numeric(map$Chr)

##Removendo efeito de ambientes controlaveis
for(tr in colnames(dat)[9:15]){
model <- paste(tr,"~-1+tph+(1|block)",sep="")
dat[!is.na(dat[,tr]),tr] <- residuals(lmer(model,dat))
}

phe <- droplevels(dat[,9:15])
rownames(phe) <- dat$id
phe <- phe[rownames(M),]
rownames(struc) <- struc$geno; phe <- data.frame("stru"=struc[,3],phe)

compor <- data.frame(rownames(phe)); rownames(compor) <- rownames(phe);
colnames(compor) <- "blup"

#blup <- NULL
#for(i in 1:7){    compor$blup <- NA
tr = colnames(phe)[-1][i] # loop

    fixed <- formula(paste(tr,"~stru",sep=""))

    set <- phe[,c("stru",tr),drop=FALSE];
    set <- set[!is.na(set[,2]),,drop=FALSE]; dim(set)
    M1 <- M[rownames(set),]
```

```
G0 <- gbuild(M1)
    system.time(L0 <- summary(m <-
regress(fixed,~G0,data=set,pos=rep(1,2))))
    h2G <- round(L0$sigma[1]/sum(L0$sigma,na.rm=TRUE),4);
print(c(tr,h2G))
#fit <- m$predicted
#compor[rownames(compor)%in%rownames(fit),] <- m$predicted[,1]
#blup <- cbind(blup,compor$blup)
#colnames(blup)[i] <- tr
#}

gwas <- NULL
for(i in 1:ncol(M)){
    fixed <- formula(paste(tr,"~stru+M1[,",i,"]",sep=""))
    p <- pbuild(M1[,i])
    loci <- colnames(M1[,i,drop=FALSE])
    G1 <- gbuild(M1[,!colnames(M)%in%loci]);
    err <- try(L1 <- summary(regress(fixed,~G1,data=set,pos=rep(1,2))))
    if(class(err)=="regress"){
        LRT <- 2*(L1$llik-L0$llik)
        mi =L1$beta[3]
        vg=L1$sigma[1]
        ve=L1$sigma[2]
    } else LRT <- mi <- vg <- ve <- NA
    gwas <- rbind(gwas,
                  data.frame("trait"=tr,h2G,loci,mi,vg,ve,p,LRT))

print(c(tr,nrow(gwas),paste(round(i/ncol(M),5)*100,"%"),round(LRT,3),roun
d(mi,3)))
}
save(gwas,file=paste("RData/GWAS_",tr,".RData",sep=""))
```

1.7.2 RHM procedure

```
library(rrBLUP); library(ggplot2)
library(lme4)
library(regress)
source("gbuild.R");

##Carregando adaptando os dados
struc <- read.table("popstruc.txt",header=TRUE); struc$geno <-
as.character(struc$geno-1)
dat <- read.table("Dados/data.txt",header=TRUE)
dat$BWD[dat$BWD<200] <- NA;
M <- read.table("Dados/M.txt",header=TRUE)
map <- read.table("Dados/map.txt",header=TRUE)[,1:3]
colnames(map) <- c("mkr","Chr","Pos")
map$Pos <- map$Pos/1e+06
```

```
codes <-
as.numeric(sapply(as.character(rownames(M)),function(x){strsplit(x,"CNB")
[[1]][2]}))+10
rownames(M) <- codes

dat$id <- as.character(dat$id)

map$Dist <- round(map$Pos)+1 #Generating position classes
map$Chr <- as.numeric(map$Chr)

##Removendo efeito de ambientes controlaveis
for(tr in colnames(dat)[9:15]){
model <- paste(tr,"~-1+tph+(1|block)",sep="")
dat[!is.na(dat[,tr]),tr] <- residuals(lmer(model,dat))
}

phe <- droplevels(dat[,9:15])
rownames(phe) <- dat$id
phe <- phe[rownames(M),]
rownames(struc) <- struc$geno; phe <- data.frame("stru"=struc[,3],phe)

for(tr in colnames(phe)[-1]){
    fixed <- formula(paste(tr,"~stru",sep=""))

    set <- phe[,c("stru",tr),drop=FALSE];
    set <- set[!is.na(set[,2]),,drop=FALSE]
    M1 <- M[rownames(set),]

    G0 <- gbuild(M1)
    L0 <- summary(regress(fixed,~G0,data=set,pos=rep(1,2)))
    h2G <- round(L0$sigma[1]/sum(L0$sigma,na.rm=TRUE),4); h2G

rhm <- NULL
for(j in 1:11){
        temp <- droplevels(map[map$Chr==j,])
        ##
        for(i in (min(temp$Dist)):(max(temp$Dist)-1)){
            pos <- mean(temp[temp$Dist%in%c(i,i+1),]$Pos)
            min_pos <- min(temp[temp$Dist%in%c(i,i+1),]$Pos)
            max_pos <- max(temp[temp$Dist%in%c(i,i+1),]$Pos)
            mkrs <- as.character(temp[temp$Dist%in%c(i,i+1),]$mkr)
            if(length(mkrs)>1){
                G1 <- gbuild(M1[,!colnames(M)%in%mkrs]);
                Reg <- gbuild(M1[,mkrs]);
                err <- try(L1 <-
summary(regress(fixed,~Reg+G1,data=set,pos=rep(1,3))))
                if(class(err)=="regress"){
                    LRT <- 2*(L1$llik-L0$llik)
                    h2reg <- L1$sigma[1]/sum(L1$sigma,na.rm=TRUE)
                    h2Greg <- L1$sigma[2]/sum(L1$sigma,na.rm=TRUE)
                } else LRT <- h2reg <- h2Greg <- NA
            } else LRT <- h2reg <- h2Greg <- NA
```

```
            rhm <- rbind(rhm,
                         data.frame("Chr"=j,h2Greg, min_pos,max_pos,
                                    "Region"= paste(i,i+1,sep="_"),
                                    "Pos"=pos, "Num"=length(mkrs),LRT,
h2reg))

print(c(tr,j,paste(i,i+1,sep="_"),round(LRT,3),round(h2Greg,3),round(h2re
g,3)))
            rm(G1,Reg,h2reg,LRT)
         }
     save(rhm,file=paste("RData/RHM_",tr,".RData",sep=""))
}
}
```

1.8 Appendix

The following Supporting Information is available for this thesis:

Table S1 Summary statistics of the phenotypic traits studied

Table S2 Genetic correlations among the seven traits estimated using SNP-based genomic relationships (below diagonal) and significance levels (above diagonal)

Table S3 Expected and observed heritability thresholds above which the individual RHM regions or GWAS hits were declared significant given the sample size (n) of the population with a significance level $\alpha = 10^{-5}$ and a power of detection $\beta = 0.90$

Fig. S1 STRUCTURE analysis of the 768 individual trees across the 37 full-sib families showing the presence of k = 2 populations.

Methods S1 Required sample sizes for the detection of fixed and random QTL effects and heritability threshold for genome-wide association study.

Table S1. Summary statistics of the phenotypic traits studied

Trait	Max.	Min.	Mean	σ_f	σ_g	h^2_{PED}
DBH (cm)	16.6	5	10.56	2.075	1.105	0.533
HEI (m)	22.63	7.8	14.7	1.878	0.783	0.417
PPR (score)	3	0	2.037	0.056	0.02	0.364
BWD (kg m^{-3})	575.2	188.8	468	731.386	428.738	0.586
EA (%)	21	13	16.04	0.761	0.373	0.49
SPY (%)	59.88	40.2	50.73	2.74	1.292	0.472
KN	18	14	16.13	0.397	0.133	0.335

DBH, diameter at breast height; HEI, height growth; PPR, *Puccinia psidii* rust disease severity; BWD, basic wood density; EA, effective alkali demand; SPY, screened pulp yield; KN, pulp bleaching content also called Kappa number. σ_p, phenotypic variance; σ_p, genotypic variance; h^2_{PED}, trait pedigree-based heritability;

Table S2. Genetic correlations among the seven traits estimated using SNP-based genomic relationships (below diagonal) and significance levels (above diagonal)

	DBH	HEI	BWD	EA	SPY	KN	PPR
DBH							
HEI	0.88		ns				
BWD	-0.16	-0.09					
EA	-0.31	-0.41	0.20				ns
SPY	0.45	0.53	-0.12	-0.84			ns
KN	-0.16	-0.38	0.15	0.69	-0.66		ns
PPR	-0.37	-0.35	0.23	-0.08	-0.09	-0.11	

DBH, diameter at breast height; HEI, height growth; PPR, *Puccinia psidii* rust disease severity; BWD, basic wood density; EA, effective alkali demand; SPY, screened pulp yield; KN, pulp bleaching content also called Kappa number. Significance levels were based on multiple testing Bonferroni correction at an overall $\alpha \leq 0.01$.

Table S3. Expected and observed heritability thresholds above which the individual RHM regions or GWAS hits were declared significant given the sample size (N) of the population with a significance level $\alpha = 10^{-5}$ and a power of detection 1-β = 0.90

				Expected		Observed[+]	
	$\mathbf{Z}$	**N**	h^2_{PED}	h^2_{REG}	h^2_{GWAS}	h^2_{REG}	h^2_{GWAS}
DBH	27.77	768	0.533	0.017	0.036	0.080	0.044
HEI	27.77	768	0.417	0.021	0.036	0.039	0.066
BWD	27.77	764	0.686	0.015	0.036	0.087	0.050
EA	27.77	761	0.490	0.019	0.036	0.041	0.036
SPY	27.77	761	0.472	0.019	0.036	0.039	0.042
KN	27.77	761	0.335	0.024	0.036	0.040	0.038
PPR	27.77	559	0.364	0.032	0.050	0.045	0.041

[+]Lowest h^2 values observed for individual regions or associations in Tables 2 and 3. $\mathbf{Z}$ is defined as $\mathbf{Z} = (\mathbf{Z}_{(1-\alpha/2)} + \mathbf{Z}_{(1-\beta)})^2$, where α (significance level) and 1-β (power of detection).

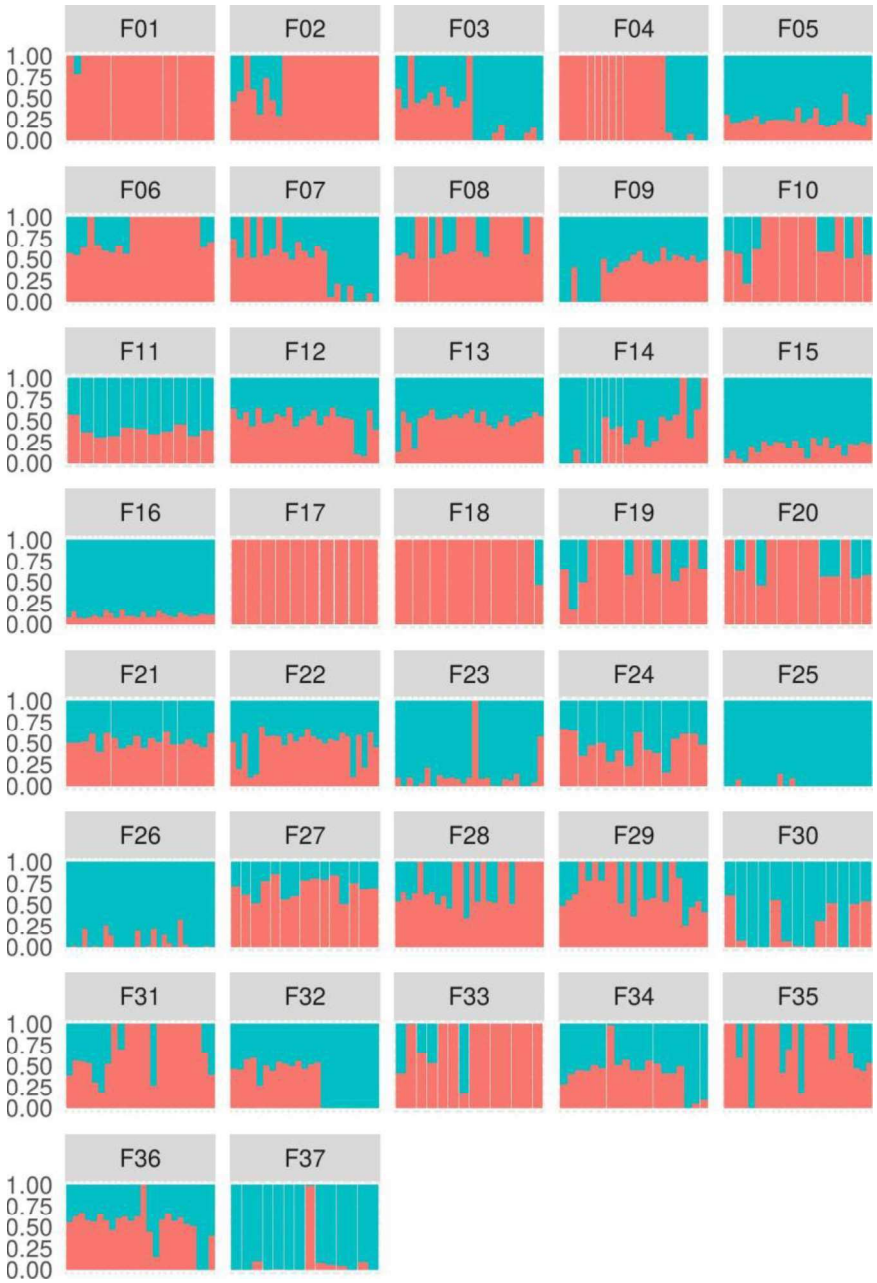

Figure S1. STRUCTURE analysis of the 768 individual trees across the 37 full-sib families showing the presence of k = 2 populations corresponding to the two ancestral Eucalyptus species involved, *E. grandis* and *E. urophylla*. The *Y*-axis corresponds to the genomic proportion of each one of the two species for each individual (vertical bars).

Methods S1. Required sample sizes for the detection of fixed and random QTL effects and heritability threshold for genome-wide association study.

QTLs treat as fixed effects. Statistics books (Steel & Torrie, 1960) provide the general expression to calculate the required sample size (N) which is $N = \frac{(\mathbf{Z}_{\alpha}+\mathbf{Z}_{\beta})^2\sigma_D^2}{\delta^2}$, where Z_{α} and Z_{β} are the values of accumulated distribution functions of type I (α) and type (β) erros, under unilateral hypothesis tests or $N = \frac{(\mathbf{Z}_{\alpha/2}+\mathbf{Z}_{\beta})^2\sigma_D^2}{\delta^2}$ for bilateral tests, where σ_D^2 is the variance of the difference between two treatments means; δ^2 is the size of the true difference between the two means which is intended to be declared as significant. The quantity $(1-\beta)$ is the probability that the experiment shows a significant difference between treatments means. A probability of 90% is common and adequate in practice. The variance of σ_D^2 is a function of the residual variance and δ^2 can be taken as the squared contrast between one effect and the zero point of mass.

In traditional GWAS context with the analysis of one marker i (fixed effect) under the model $\mathrm{y} = \mathrm{Xb} + \mathrm{Ts} + \mathrm{Wm_i} + \mathrm{e}$ for phenotypes we have the following definitions $\sigma_D^2 = \sigma_e^2/Var(W_i)$, $Var(W_i) = 2p_iq_i$ and $\delta^2 = m_i^2$. Then we have:

$N = \frac{(\mathbf{Z}_{(1-\alpha/2)}+\mathbf{Z}_{(1-\beta)})^2\sigma_D^2}{\delta^2} = \frac{(\mathbf{Z}_{(1-\alpha/2)}+\mathbf{Z}_{(1-\beta)})^2\sigma_e^2}{2p_iq_im_i^2}$, where $\sigma_e^2 = (1-"h_{m_i}^2")\sigma_y^2$ and as a consequence, $N = (\mathbf{Z}_{(1-\alpha/2)} + \mathbf{Z}_{(1-\beta)})^2\frac{(1-"h_{m_i}^2")}{"h_{m_i}^2"}$, where $"\mathrm{h}_{\mathrm{m_i}}^2" = \frac{2p_iq_im_i^2}{\sigma_y^2}$ is the determination coefficient of the i^{th} SNP effect, which is similar to a marker "heritability". Assuming, $\sigma_y^2 - 2p_iq_im_i^2 \approx \sigma_y^2$, we have $1 - \mathrm{h}_{\mathrm{m_i}}^2 = 1$ and $N_f \approx \frac{(\mathbf{Z}_{(1-\alpha/2)}+\mathbf{Z}_{(1-\beta)})^2}{"h_{m_i}^2"}$. From this last equation we arrive at the threshold heritability value required to declare a marker effect as significant given the sample size used in the experiment, which is $h_{GWAS}^2 \approx \frac{(\mathbf{Z}_{(1-\alpha/2)}+\mathbf{Z}_{(1-\beta)})^2}{N_f}$.

QTLs treat as random effects. In the random effects context with the simultaneous analysis of a vector m, with $m \sim N(0, I\sigma_m^2)$ of all markers (random effects) and under the model $y = Xb + Ts + Wm + e$ we have the following definitions: $Var(W_im_i) = 2p_iq_i\sigma_{m_i}^2 = 2p_iq_i\tilde{m}_i^2$. Then we have: $N_r = \frac{(\mathbf{Z}_{(1-\alpha/2)}+\mathbf{Z}_{(1-\beta)})^2\sigma_D^2}{\delta^2} = \frac{(\mathbf{Z}_{(1-\alpha/2)}+\mathbf{Z}_{(1-\beta)})^2\sigma_e^2}{2p_iq_i\tilde{m}_i^2}$, where $\sigma_e^2 = (\sigma_y^2 - \sum_{i=1}^{n} 2p_iq_i\tilde{m}_i^2 = (1-h^2)\sigma_y^2$ and then $N_r = \frac{(\mathbf{Z}_{(1-\alpha/2)}+\mathbf{Z}_{(1-\beta)})^2(1-h^2)}{h_{m_i}^2}$ with $\mathrm{h}_{\mathrm{m_i}}^2 =$

$\frac{2p_iq_im_i^2}{\sigma_y^2}$ and $h^2_{m_i} = \frac{\sum_{i=1}^{n} 2p_iq_i\tilde{m}_i^2}{\sigma_y^2}$, which are the marker i and trait heritabilities, respectively.

Table A. Sample sizes (N) and power to detect a QTL with $\alpha = 10^{-5}$ according to the magnitude of the QTL effect (h^2) treated as random (N_r) or fixed (N_f) for a trait displaying a heritability $h^2 = 0.30$

$\mathbf{Z}$ for $\beta = 0.90$	$\mathbf{Z}$ for $\alpha = 10^{-5}$	$(\mathbf{Z}_{(1-\alpha/2)} + \mathbf{Z}_{(1-\beta)})^2$	$h^2_{m_i}$	h^2	N_r	N_f
1.28	3.99	27.7729	0.001	0.30	19441	27773
1.28	3.99	27.7729	0.005	0.30	3888	5555
1.28	3.99	27.7729	0.01	0.30	1944	2777
1.28	3.99	27.7729	0.05	0.30	389	555
1.28	3.99	27.7729	0.1	0.30	194	278
1.28	3.99	27.7729	0.2	0.30	97	139
1.28	3.99	27.7729	0.3	0.30	65	93

$N_r = \frac{(\mathbf{Z}_{(1-\alpha/2)}+\mathbf{Z}_{(1-\beta)})^2(1-h^2)}{h^2_{m_i}}$; $N_f \approx \frac{(\mathbf{Z}_{(1-\alpha/2)}+\mathbf{Z}_{(1-\beta)})^2}{"h^2_{m_i}"}$