Unraveling growth molecular mechanisms in *Pinus taeda* with GWAS, machine learning, and gene coexpression networks

Stephanie Karenina Bajay[a,b†], Alexandre Hild Aono[a,b†], Felipe Roberto Francisco[a,b†], Anete Pereira de Souza[a,b,*]

[a]Center for Molecular Biology and Genetic Engineering (CBMEG), University of Campinas (UNICAMP), Campinas, SP, Brazil

[b]Department of Plant Biology, Biology Institute, University of Campinas (UNICAMP), Campinas, SP, Brazil

[†]These authors have contributed equally to this work.

***Corresponding Author**

E-mail: anete@unicamp.br

Anete Pereira de Souza: 0000-0003-3831-9829

ORCID

Stephanie Karenina Bajay: 0000-0002-7018-6491

Alexandre Hild Aono: 0000-0003-4871-6625

Felipe Roberto Francisco: 0000-0001-6488-7016

Abstract

Pinus taeda (loblolly pine [LP]), a long-lived tree species, is one of the world's most economically important forest trees. Among growth traits, volume is the most widely considered in tree improvement programs. The challenge of revealing the genetic variants involved in growth trait variation is even greater in conifers such as LP because of the extremely large size and high complexity of their genomes. Here, we present a comprehensive genetic analysis of LP involving genetic markers associated with volume (via genome-wide association studies [GWASs] and machine learning [ML]) to reveal pathways involved in high-performance phenotypes. We used a population of LP in the 2nd

cycle of breeding composed of full-sib progenies established at seven sites. A total of 1,692 individuals were phenotyped and genotyped using sequence capture probes. A total of 31,589 SNPs were identified and used to perform a GWAS through a multilocus mixed model. A precision core marker set with 7,864 SNPs selected through ML, i.e., feature selection (FS), was used for genomic selection (GS), providing a predictive accuracy of approximately 0.79. For genomic annotation and the construction of gene coexpression networks, three transcriptomes were assembled. We selected genes associated with the target trait and assessed the cascade of related molecular mechanisms within gene coexpression networks. These results advance the understanding of the genetics of growth traits, reveal candidate genes for future functional studies in LP, and increase the understanding of quantitative genetics and the genomics of complex phenotypic variations in LP.

Keywords: forest breeding, genomic selection, genome-wide association studies [GWAS], loblolly pine, machine learning [ML], volume trait

Key Message

From the combination of GWAS and ML, we identified genetic markers associated with volume trait in loblolly pine, providing information into its genetic architecture and enabling more efficient breeding strategies for improved plant growth.

1. Introduction

Since the 1950s, forest product companies have intensively invested in tree breeding, plantation silviculture, and forest management to meet the global demand for wood and fiber (Carter et al. 2015). *Pinus taeda* (loblolly pine, LP) is one of the most economically important forest trees worldwide (Plomion et al. 2007; Lu et al. 2017), particularly in the timber, pulp, and paper industries in the southeastern United States (US). Conifers such as LP and other commercially grown tree species are still considered largely undomesticated despite more than six decades of breeding. There has been increasing pressure to improve this species to increase productivity, forest value, and ecosystem services (Isik and McKeand 2019; Telfer et al. 2019).

One breeding cycle of LP using traditional phenotypic selection requires 15 to 30 years, including at least 8 years of field testing (Harfouche et al. 2012). Although classical pine breeding strategies have generated significant gains over the past decades, efforts in biotechnology and omics-based approaches have been suggested as a tool to accelerate and expand such genetic gains (Poovaiah et al. 2020; Caballero et al. 2021; Goralogia et al. 2021; Shalizi et al. 2022). However, for LP and other conifers, the development of such strategies is hindered by the complex genome of the species.

Even with recent advances in sequencing technologies, the pine genome could only be assembled into a highly fragmented draft reference with more than one million scaffolds and up to an 82% content of various highly repetitive DNA elements (Perera et al. 2018). This fact is explained by its complex, poorly characterized and extremely large genome (1C=21.6 Gb) (O'Brien et al. 1996; Wegrzyn et al. 2014; Zimin et al. 2017). Such complexity has led LP studies to focus on restricted genomic analyses, mostly on nuclear gene sequences, which have been the main source of molecular data for the improvement of this conifer (Carrasco et al. 2017; Baker et al. 2018; Junkes et al. 2019; Mao et al. 2019). In this context, the inclusion of additional molecular data sources can benefit LP improvement efforts, especially for highly quantitative traits of commercial importance, such as stem volume (SV).

The targeting of genetic breeding programs to produce varieties with high SV is prioritized in most cultivated tree species, including LP (Lu et al. 2018; Shalizi and Isik 2019). However, the realization of volume increases based on molecular data is hampered by the polygenic nature of this complex trait and substantial variation between populations (Lauer et al. 2021). Statistical methods for associating genotypes with phenotypes have been of great interest for breeding, especially as a tool for genomic selection (GS) and a means of deciphering causative loci

using genome-wide association studies (GWASs). These models use multiple methods with varying degrees of complexity, computational efficiency and predictive accuracy (Jannink et al. 2010; Desta and Ortiz 2014; Wang et al. 2018). A number of proof-of-concept studies have demonstrated the success of GS implementation in forest trees, leading to more dynamic breeding programs and ensuring adaptation to the influence of abiotic and biotic stresses in less time (Resende et al. 2012b; Zapata-Valenzuela et al. 2012; Beaulieu et al. 2014; Isik et al. 2016; Grattapaglia et al. 2018; Lenz et al. 2020).

Genomic prediction (GP) models are often based on the estimation of additive effects through a large number of molecular markers capable of representing the phenotypic contributions of coding and noncoding regions of the genome (Meuwissen et al. 2001). Such a theory has not been sufficient to elucidate which evolutionary forces maintain the wide variation in the complex genome of coniferous trees. Even when combining large-scale phenotypic and genotypic datasets with robust statistical models, the genetic markers identified by a GWAS often reside outside annotated gene boundaries and can be relatively far from the actual causal polymorphisms (Schaefer et al. 2018). For deeper comprehension, additional methodologies are needed and can provide a greater picture of unknown molecular interactions upon integration with different layers of omics.

Nonparametric models for GP represent an alternative for modeling genotype-phenotype associations, and the role of machine learning (ML) is still not fully understood (Wang et al. 2018). Additionally, multiomics joint analyses have demonstrated a high potential to dissect and better understand the architecture of complex traits, mapping genetic variants with unknown biological roles to molecular mechanisms through integrative methodologies, such as gene coexpression networks (Schaefer et al. 2018; Francisco et al. 2021). Loci detected under association with a trait can be linked to genes within coexpressed modules, providing a deeper comprehension of the mechanisms underlying complex traits (Francisco et al. 2021).

In this study, we used a targeted genotyping approach and SV phenotypes of 1,999 individuals from a breeding population of LP to identify genetic markers in association with volume via GWAS and ML-based prioritization. The employed GWAS approach elucidated volume's genetic architecture in seven Florida populations. In addition, we showed that in addition to accurately predicting volume phenotypes, we could employ GP models for broader marker prioritization with ML approaches. Finally, all the selected markers were linked to genes, and a gene coexpression network was leveraged to assess how the volume influenced plant growth, ultimately assisting breeders in the direction of the selection.

2. Material and Methods

2.1. Breeding Population and Plant Material

Phenotypic measurements were collected from the 2nd cycle of breeding and selection performed by the Cooperative Forest Genetics Research Program (CFGRP) at the University of Florida. The population was derived from crosses of twenty-eight LP progenitors, which were used as females and males. A total of 25,080 full-sib progenies from 45 crosses (hereafter families) were field tested at seven different sites established in 2006: AT238, BT240, FT239, HT243, IT244, LT242 and WT245. The complete individual distribution is described in the Supplementary Material (Section 1). The experimental design was a randomized complete block design with single-tree plots arranged in an alpha lattice with 20 incomplete blocks in 12 replications with 300 trees. The analysis was focused on the trait SV, which was calculated based on the stem height (SH) and the diameter at breast height (DBH) at 1.4 meters. These measures were obtained at ages 3 and 4 years, converted into decimeters (dm) and used for estimating SV in dm^3 based on the inside-bark total stem equation (Sherrill et al. 2011):

$$SV = 0.20571 - 0.00237 \times (DBH^2 \times SH)$$

2.2. Phenotypic Data Analysis

Phenotypes were analyzed using ASReml v.2 (Gilmour et al. 2006) according to the following linear mixed model:

$$Y_{ijrmn} = \mu + R_r + B_{jr} + P_{rm} + Q_{rn} + G_i + F_i + e_{ijrmn}$$

where Y is the phenotypic measure of the ith genotype considering the jth block, the rth replication, and the mth and n^{th} positions (rows and columns). The overall mean is represented by μ, and the fixed effects of the rth replication are represented by R_r. The random effects were modeled to estimate the contributions of (1) the jth block in the rth replication (B_{jr}); (2) the mth row in the rth replication (P_{rm}); (3) the nth column in the rth replication (Q_{rn}); (4) the ith genotype (G_i); (5) the family of the ith genotype (F_i); and (6) the residual error e.

All the remaining analyses were performed using R statistical software (R Core Team 2013). Individual narrow-sense heritability h^2 was calculated as the ratio of the additive variance σ^2_{Gi} to the total phenotypic variance ($\sigma^2_R + \sigma^2_B + \sigma^2_P + \sigma^2_Q + \sigma^2_F + \sigma^2_e$). The dominance ratio ($d^2$) was estimated as $4 \times \sigma^2{}_R$ divided by the total

phenotypic variance. Term 4 is included because of the presence of half-siblings within the block and the replication.

Data distributions were plotted using the ggplot2 R package (Wickham 2016). As the final measure, we discarded phenotypic measurements of trees that died and considered for each genotype the adjusted predicted least square means across sites. Differences among sites were assessed using analysis of variance (ANOVA).

2.3. Genotypic Analyses

Genomic DNA from 1,692 LP seedling samples was extracted and sequenced using sequence capture by Rapid Genomics (Gainesville, FL). These genotypes were selected from four sites (AT23804, BT24004, HT24304 and IT24404) and 45 families. Polymorphisms were evaluated in genomic regions captured by 11,867 probes selected from a previously developed set containing 54,773 probes representing 14,729 *P. taeda* unigenes (Neves et al. 2013). Target-enriched libraries were sequenced using Illumina HiSeq 2000.

Raw reads from sequence capture data from genotypes were aligned against the LP reference genome v.2.0 (GenBank accession GCA_000404065.3) (Zimin et al. 2017) using Bowtie2 v.2.2.9 (Langmead and Salzberg 2012). Only the regions of the probes with a 500 bp extension for each side of the sequence were considered for read mapping. SNP variant discovery was performed with FreeBayes v1.3.1 (Garrison and Marth 2012) with default parameters, followed by the use of VCFtools (Danecek et al. 2011) and BCFtools v.1.3.1 (Li et al. 2011) for SNP filtering. The filtering criteria included the removal of SNPs with a quality less than 20, a minimum depth of 8, a minimum QUAL score of 30, minimum and maximum allele frequencies of 0.05 and 0.95, respectively, and a maximum of 10% of individuals with missing information per SNP. Missing data were imputed as the means.

Principal component analysis (PCA) was performed using the snpReady R package (Granato et al. 2018). Linkage disequilibrium (LD) between loci was estimated using squared Pearson correlation coefficients (R^2) for markers within the same scaffold. LD decay was assessed using the exponential model $y = a + be^{(-cx)}$ (Tenesa et al. 2004), where x and y are the physical distance (base pairs) and R^2, respectively $a + b$ is the mean level of disequilibrium for loci at the same location, and e is the exponential term.

2.4. RNA-Seq Analysis

Data available in the SRA database from three independent experiments were employed in this research (Carrasco et al. 2017; Junkes et al. 2019; Mao et al. 2019). The first experiment (SRP155070) involved RNA sequencing (RNA-Seq) of three Illumina libraries from secondary xylem tissues of three 15-year-old *P. taeda* (Pta) trees from the Yingde Research Institute of Forestry in Guangdong Province (China). The second experiment (PRJNA550136) involved RNA-Seq of sixteen Illumina libraries extracted from the vascular cambium of four sixteen-year-old trees of *P. elliottii* (Pel, slash pine). The third experiment (PRJNA295331) involved RNA-Seq of twenty-six Illumina nonnormalized cDNA libraries synthesized using the stem tissue of 17-month-old genotypes (resistant (G1) and susceptible (G10)) of *P. radiata* (Pra). We developed a common bioinformatics pipeline to be applied to these different sets of data, including (I) quality filtering using Trimmomatic v.0.39 (Bolger et al. 2014), (II) transcriptome assembly with Trinity v.2.11.0 (Grabherr et al. 2011), (III) transcript quantification with Salmon v.1.1.0 (Patro et al. 2015) (k-mer of 31), (IV) assembly evaluation with BUSCO v.5.1.2 (Seppey et al. 2019), and (V) transcriptome annotation with the Trinotate v.3.2.1 pipeline and SwissProt database (Bryant et al. 2017).

To detect correspondence of the transcripts to the genomic regions containing probes, all the assembled transcripts were aligned to the corresponding scaffolds with the BLASTn v.2.11.0 tool (Altschul et al. 1990), considering only alignments with a minimum e-value of 1e-30 and a transcript coverage of at least 75%. To assign Gene Ontology (GO) terms and Enzyme Commission (EC) numbers to the assembled contigs, we considered the annotations obtained with the Trinotate pipeline.

2.5. Genome-Wide Association

GWAS was the first analysis performed in this study to identify SNPs associated with phenotypic variation. Polymorphisms showing a significant association with SV were used for functional inferences based on the neighboring genes and putative corresponding proteins. The GWAS was conducted with the Fixed and Random Model Circulating Probability Unification (FarmCPU) method implemented in the rMVP R package v.1.0.6 (Yin et al. 2020). Additionally, to include genotype-environment correlation in the association mapping, we also incorporated the site information into FarmCPU. SNPs were considered significantly associated with the phenotypes if the p value was below 0.05/*N* (Bonferroni correction), with *N* representing the number of tests performed. Additionally, the set of markers identified as being associated with SV was expanded based on LD estimates (minimum Pearson correlation coefficient of 0.7).

The functional profiles of significant SNPs were assessed using the assembled transcriptome data. We assigned for each marker the correspondences from the alignments performed on RNA-Seq data, considering a window of 10,000 bp from the right and left sides of the SNP position (20,001 bp in total).

2.6. Genomic Prediction and Feature Selection

In addition to the GWAS performed, we also investigated the potential of the set of identified SNP markers for predicting the SV trait in a GP model. For such a task, we used a Bayesian ridge regression (BRR) approach with the R package BGLR (Perez and de los Campos 2014). Considering our genotype dataset as a matrix $Z_{n\times m}$, with *n* genotypes and *m* loci codified as 0 (reference homozygote), 1 (heterozygote) and 2 (alternative homozygote), the SNP effects (γ) were estimated considering the equation:

$$y = 1\mu + Z\gamma + e$$

where *y* represents the SV trait, μ is the overall population mean, and *e is* the model residuals ($e \sim MVN(0_n, I\sigma^2{}_e)$). The SNP effects were assumed to follow the same normal prior distribution ($\gamma|\sigma^2{}_\gamma \sim N(0, \sigma^2{}_\gamma)$) across all loci, with $\sigma^2{}_\gamma$ representing the genetic variance.

In contrast to the predictions obtained with the models estimated using the entire set of markers, we evaluated the feasibility of feature selection (FS) techniques for subsetting the SNP data through putative phenotype-genotype associations. We selected the intersection between at least two out of three methods established (Aono et al. 2020, 2022a): (i) the gradient tree boosting (GTB) regressor model, (ii) Pearson correlations (maximum p value of 0.05), and (iii) the support vector machine (SVM) regression system. With such a dataset, we evaluated the importance of each SNP for prediction by calculating their feature importance using two different tree-based ML algorithms: decision tree (DT) and random forest (RF). Outliers of these measures were calculated by considering 1.5 times the interquartile range.

All the models created for prediction were evaluated considering a k-fold (k=10) cross-validation scenario repeated 100 times and created with the scikit-learn Python v.3 module (Pedregosa et al. 2011). As evaluation metrics, we used the R Pearson correlation coefficient and the mean squared error, calculated based on the comparison of the predicted and real values.

2.7. Coexpression Networks

To evaluate the impact of phenotype-associated markers from a broader perspective, we created gene coexpression networks using Pel and Pra RNA-Seq data and a weighted gene coexpression network analysis (WGCNA) approach implemented in the R package WGCNA (Langfelder and Horvath 2008). After calculating a Pearson correlation matrix based on the expression of gene pairs, we estimated a soft power β for fitting the network into a scale-free topology. With the transformed correlations, we calculated a topological overlap matrix (TOM), which was converted into a dissimilarity matrix for the definition of groups of coexpressed genes using an average linkage hierarchical clustering approach together with adaptive branch pruning (Langfelder and Horvath 2008).

After identifying coexpressed groups containing the genes surrounding SV-associated markers, we performed an enrichment analysis in each gene set to retrieve enriched biological process GO terms. We used the topGO R package (Alexa et al. 2006) considering Fisher's exact test and a multiple-comparison Bonferroni correction (p value of 0.01).

Based on the proportion of genes putatively associated with SVs within each coexpressed group, we selected the specific modules for modeling another gene coexpression network; however, we used the highest reciprocal rank (HRR) approach (Mutwil et al. 2010) to investigate specific gene connections and the network topology through centrality measures: (i) degree (Barabasi and Oltvai 2004); (ii) stress (Brandes 2001); (iii) short path length value (Watts and Strogatz 1998); (iv) betweenness centrality (Brandes 2001); and (V) neighborhood connectivity (Maslov and Sneppen 2002). To build the network, we considered Pearson correlations with a minimum coefficient of 0.8 and the 30 strongest connections for each gene. To visualize the behavior of the groups modeled by WGCNA in the HRR networks, we used Cytoscape software v3.7.0 (Shannon et al. 2003).

Additionally, we used the REVIGO tool (Supek et al. 2011) to summarize GO categories.

3. Results

3.1. Phenotyping

Seedlings derived from 45 LP families were planted across seven sites and analyzed in this study (Supplementary Fig. 1 and Supplementary Table 1). Stem volume measurements obtained from 19,933 surviving individuals were used in the phenotypic analyses (see Methods, Supplementary Fig. 2), and trait heritabilities and dominance ratios were estimated across sites (Supplementary Table 2).

3.2. Genotyping

A set of 1,692 individuals from 45 families planted in four environments were selected for genotyping (Supplementary Table 3). Genotyping by sequence capture generated ~996,611 million reads, which were aligned against the sequences of 11,867 probes (± 500 bp) obtained from the *P. taeda* reference genome. Probe regions located less than 500 bp from each other were combined, resulting in 10,412 sequences to which the reads were aligned.

Among the sequenced reads, 84.5% were aligned to the probe sequences, resulting in an initial count of 133,199 polymorphisms, including only biallelic SNPs with a minimum quality of 30 and inside sites with a quality above 10 and mapping quality of 30 (minimum allele frequency of 0.01 and no more than 30% missing data). To produce the final dataset used in the GWAS, we used more stringent parameters (maximum of 10% missing data and minimum and maximum allele frequencies of 0.05 and 0.95, respectively), resulting in 31,589 SNPs distributed among almost all probes (Supplementary Fig. 3).

A PCA performed with the SNP data indicated the presence of genetic clusters for each family (Fig. 1-ii). LD decay was estimated with the squared Pearson correlation coefficient (R^2) between SNP pairs. We observed a rapid decay to the half maximum $R^2 = 0.05$ within less than 100 kb (Fig. 1-i).

3.3. RNA-Seq Analysis

We selected data from transcriptome experiments previously performed in Pta, Pra and Pel (Supplementary Table 4). All the data processing procedures were the same for the three experiments. The filtering procedures applied to the RNA-Seq data retained approximately 85% of the reads, which were assembled with Trinity. We generated a total of 75,897, 137,103, and 50,781 genes for Pta, Pra and Pel, respectively (isoforms described in Supplementary Table 5). To evaluate the general quality, precision, contiguity and integrity of the three de novo assembled transcriptomes, we used different metrics (Supplementary Table 5) together with Benchmarking Universal Single Copy Orthologs (BUSCO) evaluations. Over 90.4% (Pel), 78.1% (Pta), and 98.4% (Pra) of the assembled genes could be recovered by BUSCO analysis, but with an elevated level of duplications due to the high abundance of isoforms (Supplementary Fig. 4). Additionally, we observed that ~40% of these genes could be related to SwissProt proteins through the Trinotate pipeline (Supplementary Table 6). As our final set of assembled genes,

we considered only genes expressed in at least 75% of the samples in each dataset, generating final numbers of 38,472, 37,499, and 33,374 genes for Pta, Pra and Pel, respectively.

A total of 9,537 (92%), 8,132 (78%) and 9,521 (91%) scaffolds containing probes could be related to the Pta, Pel and Pra transcriptomes, respectively (Supplementary Fig. 5).

3.4. GWAS

Initially, we identified seven candidate loci for the SV trait without including covariates in the GWAS model (Supplementary Table 7 and Supplementary Figs. 6 and 7). However, to avoid environmental biases, we included the sites and family structure as covariates in the final GWAS model. Fig. 2-i shows the quantile–quantile (QQ) plot for the GWAS. This approach enabled the identification of seven SNPs associated with the SV phenotype (Supplementary Table 8 and Supplementary Fig. 8), with five of them also identified using the model without covariates.

This set of seven markers was extended by linkage disequilibrium (LD) tests, which resulted in the identification of seven additional markers (Fig. 2-ii, Supplementary Table 9 and Supplementary Fig. 9). Surrounding the markers identified through GWAS and LD tests, we could establish a wide set of genes in such genomic regions through comparative alignments against the transcriptomes assembled. From the Pra transcriptome, we found 236 and 191 genes associated with SNPs using GWAS and LD, respectively. From the Pel dataset, 65 genes were found through GWAS and 96 through the LD set. Finally, from the Pta transcriptome, we found 564 genes associated with GWAS results and 472 associated with the LD set.

Although different quantities of genes could be retrieved for every transcriptome aligned against these genomic regions, there was a clear consensus for the biological functions retrieved (Supplementary Tables 8 and 9). The genes identified in the neighborhood of the SNPs identified by GWAS were mostly transposon elements (TEs) of retrovirus-related Pol polyprotein from transposon 17.6 (9, TE-17.6), retrovirus-related Pol polyprotein from transposon RE1 (9, RE1)) and retrovirus-related Pol polyprotein from transposon TNT 1-94 (9, TNT 1-94), representing 16.17% of these annotations. When we analyzed the annotations of the expanded SNPs in LD, we noticed a similar pattern containing a large number of TEs, mainly of the TNT-1-94 type (13), representing 6% of the identified genes.

3.5. Genomic Prediction

Using the SNP dataset, we created GP models using the BRR methodology in a 10-fold cross-validation scenario repeated 100 times. With the total set of 31,589 SNPs, the mean predictive accuracy (R Pearson coefficient) was 0.79 (mean squared error of 0.00023) (Fig. 3-iii), showing the high predictive capability of the dataset employed for predicting SV.

In addition, different FS techniques were used to decrease the dimensionality of the SNP data: (i) GTB (498 SNPs retrieved); (ii) Pearson correlations (20,957 SNPs retrieved); and (iii) SVM (11,440 SNPs retrieved) (Fig. 3-i). From such techniques, we selected a final subset of SNPs for estimating the SV predictive model as the intersection of at least two out of the three FS methods, totaling 7,864 SNPs (~24.89% of the initial marker data). From such markers, we obtained similar accuracy results (Fig. 3-iii), showing that most of the phenotypic variance could be associated with this smaller dataset. The similar results of predictive capacity using the total of markers and those selected with FS (reducing the initial dataset of SNPs by approximately 75%) point to the potential of SNP selection in GS, supplying reduced markers for studying phenotype and genotype associations.

Using this dataset selected, we employed an ML-based approach for retrieving additional phenotype-genotype associations. Through the DT and RF models, we estimated the feature importance of each SNP for predicting SV, evaluated their distribution in separated boxplots (Fig. 3-ii), and retrieved the outliers (1,506 for DT and 635 for RF). From such outliers, we selected the intersection of these markers found by each model, forming a final set of 128 SNPs, which were investigated together with GWAS results. Interestingly, all GWAS-associated SNPs were included in this defined ML set.

3.6. Coexpression Networks

To investigate a possible role of GWAS-associated genes in the molecular mechanisms controlling SV in pine, we created coexpression networks using Pra and Pel data. We did not include Pta quantifications due to the low sample size. Using a WGCNA approach, we modeled a different network for each species, considering an estimated soft-power β of 14 for Pel (R^2 of ~0.90 and mean connectivity of ~53.10) and Pra (R^2 of ~0.81 and mean connectivity of ~143.20). A total of 251 (minimum and maximum sizes of 81 and 14616 with a mean of 471.79) and 83 groups (minimum and maximum sizes of 50 and 4858 with a mean of 132.96) could be established for Pra and Pel, respectively.

For each group of genes that are part of a network, we evaluated the presence of (i) GWAS markers (7 SNPs); (ii) markers associated by LD with (i) (7 SNPs); and (iii) ML established markers (128 SNPs). Considering a minimum quantity of two genes associated with (i) in each group, we performed enrichment analyses for biological process GO terms (Supplementary Table 10).

From the groups identified in the network analyses, we found two in the Pra network (G39 and G68) with a significant proportion of genes associated with the selected markers (~92.78% of 194 genes in G39 and ~57.69% from 78 genes in G68) when contrasted to the other modules (mean proportion of ~1%). For this reason, we considered such groups of genes as a promising set of data for investigating the molecular mechanisms of SV regulation and used it to create a novel network based on HRR methodology. We created a cohesive network with 270 genes (only two did not present significant connections) and 3,008 connections (Fig. 4-i). From a treemap created with the biological process GO terms observed from this set of genes (Fig. 4-ii), we identified biological processes involved with DNA integration and recombination, cellular aromatic compound metabolic process, cellular nitrogen compound metabolism, nitrogen compound metabolic process and transposition, viral latency, and viral genome integration into host DNA and cells (Supplementary Table 10).

In addition to the associated GO terms, the presence of important elements in such a network was also investigated through centrality measures (Supplementary Table 11). We identified a gene annotated for the peptidyl-prolyl cis-trans isomerase enzyme and a gene without known annotation as hubs of the HRR network. Both genes are located close to SNPs identified by the GWAS. Other metrics were also analyzed to identify genes with great importance in the network, such as a gene annotated for GTP cyclohydrolase 1 enzyme with the highest stress value (320134) and a gene annotated for the nonlysosomal glucosylceramidase enzyme, showing the highest value of neighborhood connectivity (48).

4. Discussion

4.1. Pine Populations

Two key questions for forest breeders are which phenotypes to measure and which trees to select (Matallana-Ramirez et al. 2021). Although the answer to this question depends on the location of the populations evaluated and the level of diversity involved, volume growth, a trait evaluated in this work, is one of the traits that

forest breeders most often focus on for the evaluation and selection of phenotypes (Zobel and Talbert 1984; Shalizi and Isik 2019; Shalizi et al. 2022). Since the 1960s and early 1970s, breeders have been able to produce 7-12% more volume per hectare at harvest than trees grown from wild seeds (Li et al. 1999). In our work, we performed an in-depth investigation of genomic associations with SV configurations, providing a rich characterization of not only such putative associations but also the genes and biological processes involved.

The selection of elite genotypes in LP populations can provide a gain in volume of these trees of up to 50% in relation to unimproved genotypes (McKeand et al. 2006). The response to selection is closely related to the heritability of the evaluated phenotype, and the greater this heritability is, the greater the response to selection (Schmidt et al. 2019; Li et al. 2021). However, breeding programs that use genetic markers for selection, even to improve traits with low heritability values, such as volume, can have their efficiency increased considerably. The narrow-sense heritability estimates for SV in LP vary considerably depending on the statistical strategies used and are substantially higher for more complex models (Schmidt et al. 2019; Shalizi and Isik 2019). We found heritability values for SV ranging from 0.17 to 0.32 among the seven sites analyzed, indicating that there are different trends of heritability at each site but similar to those found by other authors (Xiang et al. 2003; Shalizi et al. 2020). In general, SV heritability values increase over time and present correlations with plant age. In this sense, the recommended age for selection can be as early as 3 years for height and as early as 4 years for DBH or SV (Xiang et al. 2003), which was the age of the samples in the data from the present work. The distribution of SVs among the sites indicates that the distinct genetic origin obtained by crossing the families influences the heritability of each site; however, the specific characteristics of each site make it difficult to generalize the heritability by family, considering the different genetic architectures between the populations.

In addition to the SV distribution across populations, we observed in the PCA performed with the genetic data a structure that corroborates the division by group families (Fig. 1-ii) and not by population locations. This structure caused by families was included in the GWAS models to minimize their possible negative effects, especially those related to spurious associations (Wu et al. 2007; Liu et al. 2016; Norman et al. 2018; Sul et al. 2018).

4.2. Genomic Selection

In practice, because of the quantitative aspect of most economically important traits, the application of GWAS-associated markers as a marker-assisted selection tool in genetic improvement programs is not feasible (Lebedev et al. 2020), paving the way for GS. Instead of relying on specific QTL regions, GS approaches employ genome-wide variants to predict breeding values (Spindel and McCouch 2016; Xu et al. 2020). The implementation of GS has been very promising for increasing the gain per unit of time and improving the accuracy of selection (Grattapaglia and Resende 2011). Despite the low heritability of the volume trait, this parameter has a relatively small impact on the predicted accuracy of GS in pines, as described by Grattapaglia and Resende (2011); therefore, we also tested the efficiency of GP models for such an estimation.

Especially in terms of the time of tree selection, which can take approximately 20 to 30 years (Lebedev et al. 2020), GS was already associated with a reduction of 50% in the LP breeding cycle, in addition to achieving an efficiency 53-112% higher than that in traditional breeding (Resende et al. 2012a, b). These promising results were obtained despite the PA being in the range of 0.36 to 0.49 and using a group of 4,853 SNPs in the best scenarios (Resende et al. 2012a, b). Here, such BP values are lower than those obtained in this study (0.79), showing a potential gain in GS efficiency. GS has already been implemented in LP selection, both to predict polygenic growth traits, such as height and wood quality, and oligogenic traits (resistance to fusiform rust) (Resende et al. 2012a, b; Filho et al. 2016). For several generations, important species of coniferous trees, such as LP, *Pinus elliottii* Engelm. and *Larix gmelinii* (Rupr.), have been genetically improved to select the best parental material, and the prediction of breeding values generated individual volume gains of approximately 20% (Emhart et al. 2007; Lai et al. 2017; Zhang et al. 2020).

Previous experiences with GS for growth and wood quality traits in forest trees showed moderate to high accuracy of selection models with correlations of 0.6 to 0.8 for full-sib families (Lenz et al. 2017). The half-sibling family structure seems to require more markers, and the GS accuracy is approximately 0.3 to 0.5 (Lenz et al. 2017). Lower GS prediction values have been reported for unrelated individuals (Lenz et al. 2017). Here, we achieve high GS prediction values, similar to those previously obtained in full-sib families, with the advantage of using a reduced set of markers. Estimates of GP accuracy in *P. taeda* were already observed in previous studies as moderate to high, ranging from 0.34 to 0.76 (Resende et al. 2012a; Zapata-Valenzuela et al. 2012, 2013). More recent research has also estimated the difference between the predictive ability of estimated breeding values in SV considering scenarios with randomly assigned clonal varieties and a family cross-validation scenario. The results showed a greater

predictive ability of SV by random genotypes (0.43) than considering the canaries of families (0.36) (Shalizi et al. 2022).

The high quality of the set of genotypic markers used here for GS may also have been influenced by the sequence capture method. Reducing genome complexity in LP is a smart strategy to address the challenges of representing and characterizing the extremely large genome in LP, which makes it difficult to analyze GS and capture most QTL effects in large breeding populations (Chen et al. 2018). The adoption of this type of low-cost and time-saving genotyping strategy is an important way to continue making the application of GS in conifers attractive (Neves et al. 2013). In this sense, the choice of targeted hybridization methods has become popular with its use for the GS of different forest trees (Chen et al. 2019; Thistlethwaite et al. 2019). The choice of sequence capture genotyping using probes was based on the challenges of interrogating the LP megagenome. The sequencing approach considering the selected probes and the polymorphic nature of the populations combined with bioinformatics strategies provided high-quality variants for this work, which were critical to the success of the methodologies employed.

Although there are different GS models that are being used for the GP of forest species, as well as different methodologies to evaluate the effectiveness or accuracy of these models, these studies have shown similar results for the application of most of these models (Isik et al. 2016; Chen et al. 2018; Calleja-Rodriguez et al. 2020). Thus, in this work, we chose to use only the BRR model since no significant influence on the effectiveness of volume estimates is expected when the forecast model is changed. On the other hand, what can have an impact on the SG, such as the ability to predict breeding values, is the selection of subsets of SNPs. The selection of marker subsets can be done in several ways, and here, we apply FS to compare predictive performance using a selected subset of SNPs and a total set of markers for SG. The use of FS with ML tools emerged as a useful method to address a large dimensionality of data, which is often due to redundant, irrelevant or noise elements (Aono et al. 2022b). Using different evaluation criteria, FS aims to build a subset of elements that provide an optimized input for the construction of ML models (Kumar and Minz 2014).

Although GS has driven significant improvements in forest breeding programs, correctly estimating the effects of a wide set of markers on a quantitative trait through GP models is still challenging (Lebedev et al. 2020). The large dimensionality of markers associated with a small number of observations (small *p*, big *n* problem) makes it difficult to estimate such a large number of parameters (Jannink et al. 2010). Some previous experimental studies

on forest tree species have already discussed the predictive ability using relatively moderate genotyping densities (Grattapaglia 2017). For example, in Scotch pine, subsets of markers representing 40% of the full set were sufficient to achieve the same predictive abilities and accuracies (Calleja-Rodriguez et al. 2020); in *Picea abies*, accuracy reached a plateau using 4,000 to 8,000 SNPs (Chen et al. 2019). Here, we also note that the AP observed with the full set of SNPs was very similar to that observed from the group of SNPs selected using FS (Fig. 3-iii).

Satisfactory accuracy results for prediction using lower densities of markers may be efficient, as found in this work, but a higher density may be necessary where the training and test populations do not genetically originate from the same primary population (White et al. 2007). The good result of selecting a reduced set of markers with good predictive power depends on how this selection is made. Selection based on FS, unlike a random selection of marker subsets, is able to indicate a possible set of predictors that are stable and less sensitive to changes in the training set (Piles et al. 2021). In this way, it is possible to work with data that do not require excessive adjustments (necessary due to high dimensionality) and allow for a reduction in computational requirements (Chandrashekar and Sahin 2014).

4.3. GWAS for Pine Breeding

Despite studies employing GWAS in some pine populations (Brown et al. 2003; De La Torre et al. 2019; Santini et al. 2021; Ding et al. 2022), several challenges remain that are related to the many SNPs associated with complex quantitative traits and the limited understanding of the molecular mechanisms by which these genetic variants are associated (Khatiwada et al. 2022). The identification of a gene affected by a QTL represents an enormous task, even more so in a species with a genome as large as LP.

Among the genotypic markers found in association with SV in our study, it was possible to observe annotations of genes and proteins with an important role in the development of LP volume. One of these related proteins was cysteine protease RD19, which was already associated with responses to biotic stresses, including mechanisms impacting genes of resistance to *Ralstonia solanacearum* (RSS1-R) (Bernoux et al. 2008), a bacterium already described as responsible for extensive damage in forest trees (Alfenas et al. 2006). It was also possible to identify genes involved in the production of important plant compounds, such as the tryptophan synthase alpha chain protein, which is necessary for the biosynthesis of indole-3-glycerol phosphate, an essential compound for plant growth (Ouyang et al. 2000). Another important protein identified was ABC transporter C family member 5, which

is related to the regulation of processes mediated by abscisic acid (ABA), including stomatal opening and germination (Gaedeke et al. 2001; Klein et al. 2003; Lee et al. 2004). Additionally, we also found acyl-CoA-binding domain-containing protein 6, which acts as an intracellular transporter of acyl-CoA esters, conferring resistance to cold and interacting with phosphatidylcholine and its derivatives, impacting phospholipid metabolism (Engeseth et al. 1996; Chen et al. 2008).

In addition to the markers identified by GWAS, five new markers were also selected through LD associations, indicating a putative impact on SV, but with smaller effects due to this co-association (Yuan et al. 2012). Interestingly, the annotation of this set of markers revealed many TEs, which have already been described as impacting plant development, responses to external stimuli and gene expression (Kashkush et al. 2003; Matsunaga et al. 2015; Traylor-Knowles et al. 2017; Tran and Choi 2020; Francisco et al. 2021).

In addition to GWAS-selected markers, through DT and RF models, we recovered other putative phenotype-genotype associations, identifying a reduced number of 128 SNPs within the large arsenal of important markers for GS. Such an approach expanded the set of identified markers putatively associated with QTL regions for the SV trait, providing additional insights into the genetic architecture of this quantitative characteristic. Further strengthening the results of the importance of these 128 markers identified by ML, we observed that the GWAS markers were also contained in this subset of SNPs, minimizing the identification of false positives.

4.4. Integrative Approach

Although growth is a complex trait governed by several genes (Francisco et al. 2021; Santini et al. 2021), only seven markers were identified to be associated with *Pinus* by GWAS. Santini et al. (2021) argue that despite the use of high-throughput phenotyping, growth-associated SNPs do not capture most of the observed phenotypic variation. GWAS approaches present limitations intrinsic to the technique, mainly related to the low heritability and proportion of phenotypic variance explained by the identified markers (Manolio et al. 2009; Korte and Farlow 2013; Tam et al. 2019). This is the main reason why we incorporated ML-based models on the definition of genotype-phenotype associations, increasing the set of 7 markers associated with SV to 128 SNPs. Alternatively, Ding et al. (2022) tried to overcome such limitations of GWAS by integrating the results of these analyses with coexpression networks in *P. elliottii*, making several biological inferences about the interactions of genes involved in species

growth. In this sense, we also included the investigation of gene coexpression patterns associated with SV in our study.

Gene coexpression networks have been successfully used to characterize GWAS results (Lee and Lee 2018; Francisco et al. 2021). This combination of quantitative analyses with gene expression data minimizes the problem of false positives or negatives and improves the understanding of the molecular mechanisms underlying the configuration of complex traits (Schaefer et al. 2018). The global measure of functional relationships that the coexpression network presents allows the association of certain phenotypes with biological processes controlled by several coexpressed genes. Thus, if the genetic variation that drives the phenotype is encoded by coregulated genes detected by GWAS, we will be able to capture a broader set of functional relationships of the genotypes studied (Schaefer et al. 2018).

Such integrative strategies are increasingly being used in tree species to overcome the individual limitations of different omics layers in addition to showing great importance for a deep understanding of the molecular mechanisms involved in the definition of an agronomic trait (Francisco et al. 2021; Ding et al. 2022). The strategy of combining the transcriptomes of LP-related species and LP genomes to annotate the regions of the markers made it possible to identify variants within and around genes along with variants substantially further upstream or downstream that can capture important regulatory elements. Additionally, we could investigate all these genes through a more global perspective by coupling coexpression network methodologies in our study.

4.5. Inferences Regarding Pine Growth

Gene coexpression networks allow us to associate a function with genes coexpressed in the same cluster through the principle known as "guilt-by-association" (Oliver 2000). By organizing the set of genes in coexpressed groups, it is possible to correlate several transcripts from pine trees with the genetic markers that we identified as associated with the volume trait. Considering that the phenotypic effect of each SNP selected by GWAS, LD, and FS can be very small, grouping the important genes according to their expression tendencies allows the identification of key physiological and regulatory processes (Baranzini et al. 2009).

For both the Pta and Pel transcriptomes, we defined different gene coexpression networks with WGCNA methodology, each with modules corresponding to different biological processes and with different quantities of SV-associated markers (through GWAS, LD, and ML). It was also possible to identify the importance of biological

processes linked to the storage of carbohydrates in different network modules, e.g., group 1 of the Pra network was associated with 14 ML-selected SNPs and 1 GWAS-selected marker. This significant participation of such processes available during the green period of the plant may be linked to the recent results of phenology measurements documented by repeated digital photographs by Luo et al. (2020). Volume growth is positively related to longer periods of green or at least carbohydrate storage to meet carbon demand at the beginning of the season (Sampson et al. 2001). This same group 1 of the Pra gene coexpression network shows expressive participation of cold response processes in pine, which corroborates the findings of Gao et al. (2021) that volume growth is associated with winter values. Trees with more active foliage during the winter are likely to have more stored carbohydrates, which can meet growth demands when the current photosynthate is insufficient during the green period (Gao et al. 2021).

Group 39, composed of more than 92% of markers selected by ML, showed significant transcripts related to the nitrogen metabolic process. The low availability of nitrogen (N) is an important limiting factor for the growth and development of trees because its acquisition, assimilation, storage, recycling, and efficient metabolic use largely determine vascular development, forest tree productivity and biomass production (Canovas et al. 2018). Another interesting module was group 68 of the Pra network, the second group in terms of the representativeness of ML and GWAS markers. Such a module is directly linked to group 39 and showed significance for biological responses linked to the nitrogen compound metabolic process. However, curiously, the great representation of GO terms in this group is linked to viral responses, indicating the importance of such responses in the development of more productive individuals.

The fluoride response was identified by group 14 of the Pel network as a mechanism linked to ML-selected markers and markers in association with the volume trait (2 GWAS markers and 8 ML-selected markers). Fluoride (F-) is a phytotoxin that can cause extensive damage to plants by disturbing their metabolism and enzymatic activities (Ruan et al. 2004). Thus, LP may be sensitive to F^- and present a mechanism that aims to protect and tolerate the toxicity of this phytotoxin in relation to tolerance to F^-, as identified in other perennial species (Camellia sinensis) (Pan et al. 2021). Cluster 32 in the Pel network showed enrichment for genes linked to light stimulus and radiation. The involvement of radiation as a limiting factor for growth is already known and seems to be involved in pine responses. In LP, volume growth is related to intercepted radiation (Cannell 1989); light interception, in which volume growth per unit of absorbed light increases with increasing stand density (Albaugh et al. 2018); and increasing leaf area index (Vose and Allen 1988). Recent studies in LP indicate that radiation

intensity and intercepted light are captured by different densities of LP stands, which becomes a significant predictor variable of volume growth using a linear model and equally important in the random forest model (Gao et al. 2021).

The purine biosynthesis pathway appears to be linked to the markers selected in group 18 of the Pra network. Rescue enzymes in this pathway may play a special role in activating resting cells and responding to environmental changes (Stasolla et al. 2003). Purine and pyrimidine catabolism is one of the components of nucleotide metabolism, which is an essential process in plants, as it is a signaling component and precursor of cytokinin phytohormone biosynthesis (Buchanan et al. 2002), provides energy and blocks construction for nucleic acids. In Arabidopsis, altered levels of nucleobases or nucleosides (purines) can act as physiological signals that affect stress responses (Daumann et al. 2015).

To deepen the biological meaning of the main gene modules associated with SV found in the networks modeled with the WGCNA methodology, another gene coexpression network was constructed with this restricted set of genes using the HRR method. The analysis of the most significant modules in the gene coexpression network using HRR allowed a global understanding of the most important biological processes for volume configuration in pine according to the significant quantity of SV-associated genes.

The two hub genes of the HRR coexpression network match exactly with two of the genes found by GWAS, one of which was annotated for the enzyme peptidyl-prolyl cis-trans isomerase (PPIase). PPIases are associated with plant growth and development, and the related genes have previously been indicated as suitable candidates to increase the abiotic stress tolerance of plants for industrial processes and agricultural applications (Singh et al. 2021). Although the other hub gene has no known annotation, it was identified as maintaining the highest degree in the network, being of great importance to maintain its stability (Barabasi and Oltvai 2004) and therefore of great importance for the definition of the phenotype of interest. Another gene with great importance in the coexpression network presenting the highest stress value was annotated as the GTP cyclohydrolase 1 enzyme and was also associated with a marker present in the selection of ML, reinforcing the significance of this marker for the development of volume in LP. GTP is part of the biosynthesis pathway of plant folate, which is an essential micronutrient for plant health and development, playing essential roles in photorespiration and in the synthesis of chlorophyll, plastoquinone, and tocopherol (Simkin 2019). Additionally, we identified a nonlysosomal glucosylceramidase enzyme (also identified by ML) with the highest value of neighborhood connectivity, indicating its influence over other genes and the definition of SV. The nonlysosomal glycosylceramidase enzyme, related in

humans, was recently discovered as a plant GCD homolog in *Arabidopsis thaliana*, indicating its participation in the catabolism of sphingolipid glucosylceramide (GlcCer), a lipid with important structural and signaling functions (Dai et al. 2020).

The characterization of these volume-associated marks by ML and GWAS allows us to investigate the biological processes linked to loci of known QTLs of growth in pine trees, which are important for predicting the genetic merit of volume. The networks of genes and transcription modules identified in volume show that the genetic interactions involved in the developmental processes in pine are abundant, as reported in other omics studies (Ingvarsson and Street 2011). This complex structure of interaction of gene expression networks makes it a complex task to select genes in isolation for characterization studies, since the variation in the expression of most genes will exert at least some degree of influence on several other genes. In this way, the integrated results of quantitative analyses (GWAS and GS) with transcriptome data and gene coexpression networks allow the targeting of which genomic modules are important for the expression of good volume phenotypes.

4.6. Adoption of Different Genomic Strategies during Forest Improvement Cycles

Plant breeding is a process of continuous genetic improvement carried out by the selection and recombination of superior strains. Optimizing available resources to maximize the selection response is essential to the success of a breeding program (Rutkoski 2019). Therefore, breeders need to constantly re-evaluate breeding strategies, considering the arsenal of available technologies and the challenge of effectively implementing these new technologies (Santantonio et al. 2020).

The implementation of GS has great potential in a phased approach in breeding programs, especially when the initial step is the routine genotyping of all materials that are evaluated for yield (Santantonio et al. 2020). In the early stages of a breeding program, the adoption of ML for the selection of small subsets of molecular markers, with medium accuracy but with low cost, may be sufficient to select genotypes within a large arsenal of population genotypic diversity. In the final cycles, when the number of genotypes is reduced, a greater density of molecular markers can be used, focusing on a more accurate result on the accuracy of prediction.

The adoption of GWAS in the phases with greater refinement of the selection of genotypes to be improved can be coupled with gene coexpression networks to identify target genes that can be used in genetic editing strategies, as with the use of CRISPR, for example. Studies for gene editing aim to study the overexpression or

cancellation of target genes, observing the results in the phenotype. With our study addressing the study of biological pathways linked to genes in association, transcriptome and ML-selected marks, we can achieve significant advances in more assertive selection of target genes for genomics breeding, presenting great potential to face the technical obstacles of the selection of genes associated with CRISPR (Goralogia et al. 2021).

This work made it possible to identify a reduced and precise set of genetic markers (128 markers) with a significant influence on the volume trait in the pine populations analyzed, as validated by the combination of GWAS, ML and GS analyses. The application of ML models, DT and RF to recover additional phenotype-genotype associations using the dataset identified by FS for GP narrowed the large list of important SNPs for volume prediction in pine. The genetic markers identified in our results are important advances for the genetic improvement of LP. The unprecedented combination of different quantitative genetics and bioinformatics strategies for data analysis of LP full-sib populations was demonstrated to be an efficient method to select genotypes with better results for volume. The modules indicated by the gene coexpression networks, based on the analyzed transcriptomes, demonstrate that genes identified by GWAS and by ML involving volume traits allowed a greater understanding of which important biological pathways are associated with pine growth and development.

Acknowledgments

We acknowledge Matias Kirst from the Forest Genomics Laboratory at the University of Florida for providing us with the data analyzed in this research.

Funding

This work was supported by grants from the Fundação de Amparo à Pesquisa do Estado de São Paulo, the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES, Computational Biology Programme). SB received a PhD fellowship from FAPESP (2019/13452-3) and CAPES (PE-BOLSAS 7349056). AA received a PhD fellowship from FAPESP (2019/03232-6). FF received a PhD fellowship from FAPESP (2018/18985-7). APS received a Research fellowship from CNPq (312777/2018-3).

Competing Interests

The authors have no relevant financial or nonfinancial interests to disclose.

Author Contributions

SB, AA and FF performed all analyses and wrote the manuscript. APS conceived the project. All authors reviewed, read, and approved the manuscript.

Data Availability

The datasets analyzed for this study are included in this article (and its Supplementary Information files).

References

Albaugh TJ, Fox TR, Maier CA, Campoe OC, Rubilar RA, Cook RL, Raymond JE, Alvares CA, Stape JL (2018) A common garden experiment examining light use efficiency and heat sum to explain growth differences in native and exotic *Pinus taeda*. For Ecol Manag 425:35-44. https://doi.org/10.1016/j.foreco.2018.05.033

Alexa A, Rahnenfuhrer J, Lengauer T (2006) Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics 22:1600-1607. https://doi.org/10.1093/bioinformatics/btl140

Alfenas AC, Mafia RG, Sartório RC, Binoti DHB, Silva RR, Lau D, Vanetti CA (2006) *Ralstonia solanacearum* em viveiros clonais de eucalipto no Brasil. Fitopatol Bras 31:357-366. https://doi.org/10.1590/s0100-41582006000400005

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403-410. https://doi.org/10.1016/S0022-2836(05)80360-2

Aono AH, Costa EA, Rody HVS, Nagai JS, Pimenta RJG, Mancini MC, dos Santos FRC, Pinto LR, Landell MGA, de Souza AP, Kuroshu RM (2020) Machine learning approaches reveal genomic regions associated with sugarcane brown rust resistance. Sci Rep 10:20057. https://doi.org/10.1038/s41598-020-77063-5

Aono AH, Ferreira RCU, Moraes A et al (2022a) A joint learning approach for genomic prediction in polyploid grasses. Sci Rep 12:12499. https://doi.org/10.1038/s41598-022-16417-7

Aono AH, Francisco FR, Souza LM, Goncalves PS, Scaloppi Junior EJ, Le Guen V, Fritsche-Neto R, Gorjanc G, Quiles MG, de Souza AP (2022b) A divide-and-conquer approach for genomic prediction in rubber tree using machine learning. Sci Rep 12:18023. https://doi.org/10.1038/s41598-022-20416-z

Baker EAG, Wegrzyn JL, Sezen UU et al (2018) Comparative transcriptomics among four white pine species. G3 Genes Genomes Genet 8:1461-1474. https://doi.org/10.1534/g3.118.200257

Barabasi AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet 5:101-113. https://doi.org/10.1038/nrg1272

Baranzini SE, Galwey NW, Wang J et al (2009) Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. Hum Mol Genet 18:2078-2090. https://doi.org/10.1093/hmg/ddp120

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

Bernoux M, Timmers T, Jauneau A, Briere C, de Wit PJ, Marco Y, Deslandes L (2008) RD19, an Arabidopsis cysteine protease required for RRS1-R-mediated resistance, is relocalized to the nucleus by the *Ralstonia solanacearum* PopP2 effector. Plant Cell 20:2252-2264. https://doi.org/10.1105/tpc.108.058685

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for illumina sequence data. Bioinformatics 30:2114-2120. https://doi.org/10.1093/bioinformatics/btu170

Brandes U (2001) A faster algorithm for betweenness centrality. J Math Sociol 25:163-177. https://doi.org/10.1080/0022250x.2001.9990249

Brown GR, Bassoni DL, Gill GP, Fontana JR, Wheeler NC, Megraw RA, Davis MF, Sewell MM, Tuskan GA, Neale DB (2003) Identification of quantitative trait loci influencing wood property traits in loblolly pine (*Pinus taeda* L.). III. QTL verification and candidate gene mapping. Genetics 164:1537-1546. https://doi.org/10.1093/genetics/164.4.1537

Bryant DM, Johnson K, DiTommaso T et al (2017) A tissue-mapped axolotl de novo transcriptome enables identification of limb regeneration factors. Cell Rep 18:762-776. https://doi.org/10.1016/j.celrep.2016.12.063

Buchanan BB, Gruissem W, Jones RL (2002) Biochemistry and molecular biology of plants: American society of plant physiologists. Wiley & Sons, Hoboken, NJ

Caballero M, Lauer E, Bennett J et al (2021) Toward genomic selection in *Pinus taeda*: integrating resources to support array design in a complex conifer genome. Appl Plant Sci 9:e11439. https://doi.org/10.1002/aps3.11439

Calleja-Rodriguez A, Pan J, Funda T, Chen Z, Baison J, Isik F, Abrahamsson S, Wu HX (2020) Evaluation of the efficiency of genomic versus pedigree predictions for growth and wood quality traits in Scots pine. BMC Genom 21:796. https://doi.org/10.1186/s12864-020-07188-4

Cannell MGR (1989) Physiological basis of wood production: a review. Scand J For Res 4:459-490. https://doi.org/10.1080/02827588909382582

Canovas FM, Canas RA, de la Torre FN, Pascual MB, Castro-Rodriguez V, Avila C (2018) Nitrogen metabolism and biomass production in forest trees. Front Plant Sci 9:1449. https://doi.org/10.3389/fpls.2018.01449

Carrasco A, Wegrzyn JL, Durán R, Fernández M, Donoso A, Rodriguez V, Neale D, Valenzuela S (2017) Expression profiling in *Pinus radiata* infected with *Fusarium circinatum*. Tree Genet Genomes 13:46. https://doi.org/10.1007/s11295-017-1125-0

Carter MC, Kellison RC, Wallinger RS (2015) Forestry in the U.S. South: a history. LSU Press, Baton Rouge, LA

Chandrashekar G, Sahin F (2014) A survey on feature selection methods. Comput Electr Eng 40:16-28. https://doi.org/10.1016/j.compeleceng.2013.11.024

Chen QF, Xiao S, Chye ML (2008) Overexpression of the Arabidopsis 10-kilodalton acyl-coenzyme A-binding protein ACBP6 enhances freezing tolerance. Plant Physiol 148:304-315. https://doi.org/10.1104/pp.108.123331

Chen ZQ, Baison J, Pan J, Karlsson B, Andersson B, Westin J, Garcia-Gil MR, Wu HX (2018) Accuracy of genomic selection for growth and wood quality traits in two control-pollinated progeny trials using exome capture as the genotyping platform in Norway spruce. BMC Genom 19:946. https://doi.org/10.1186/s12864-018-5256-y

Chen ZQ, Baison J, Pan J, Westin J, Gil MRG, Wu HX (2019) Increased prediction ability in Norway spruce trials using a marker X environment interaction and non-additive genomic selection model. J Hered 110:830-843. https://doi.org/10.1093/jhered/esz061

Dai GY, Yin J, Li KE, Chen DK, Liu Z, Bi FC, Rong C, Yao N (2020) The *Arabidopsis AtGCD3* protein is a glucosylceramidase that preferentially hydrolyzes long-acyl-chain glucosylceramides. J Biol Chem 295:717-728. https://doi.org/10.1074/jbc.RA119.011274

Danecek P, Auton A, Abecasis G et al (2011) The variant call format and VCFtools. Bioinformatics 27:2156-2158. https://doi.org/10.1093/bioinformatics/btr330

Daumann M, Fischer M, Niopek-Witz S, Girke C, Mohlmann T (2015) Apoplastic nucleoside accumulation in Arabidopsis leads to reduced photosynthetic performance and increased susceptibility against *Botrytis cinerea*. Front Plant Sci 6:1158. https://doi.org/10.3389/fpls.2015.01158

De La Torre AR, Puiu D, Crepeau MW, Stevens K, Salzberg SL, Langley CH, Neale DB (2019) Genomic architecture of complex traits in loblolly pine. New Phytol 221:1789-1801. https://doi.org/10.1111/nph.15535

Desta ZA, Ortiz R (2014) Genomic selection: genome-wide prediction in plant improvement. Trends Plant Sci 19:592-601. https://doi.org/10.1016/j.tplants.2014.05.006

Ding X, Diao S, Luan Q, Wu HX, Zhang Y, Jiang J (2022) A transcriptome-based association study of growth, wood quality, and oleoresin traits in a slash pine breeding population. PLoS Genet 18:e1010017. https://doi.org/10.1371/journal.pgen.1010017

Emhart VI, Martin TA, White TL, Huber DA (2007) Clonal variation in crown structure, absorbed photosynthetically active radiation and growth of loblolly pine and slash pine. Tree Physiol 27:421-430. https://doi.org/10.1093/treephys/27.3.421

Engeseth NJ, Pacovsky RS, Newman T, Ohlrogge JB (1996) Characterization of an acyl-CoA-binding protein from *Arabidopsis thaliana*. Arch Biochem Biophys 331:55-62. https://doi.org/10.1006/abbi.1996.0282

Filho JEDA, Guimaraes JF, FF ES, de Resende MD, Munoz P, Kirst M, Resende MF, Jr (2016) The contribution of dominance to phenotype prediction in a pine breeding and simulated population. Heredity (Edinb) 117:33-41. https://doi.org/10.1038/hdy.2016.23

Francisco FR, Aono AH, da Silva CC, Goncalves PS, Scaloppi Junior EJ, Le Guen V, Fritsche-Neto R, Souza LM, de Souza AP (2021) Unravelling rubber tree growth by integrating GWAS and biological network-based approaches. Front Plant Sci 12:768589. https://doi.org/10.3389/fpls.2021.768589

Gaedeke N, Klein M, Kolukisaoglu U, Forestier C, Muller A, Ansorge M, Becker D, Mamnun Y, Kuchler K, Schulz B, Mueller-Roeber B, Martinoia E (2001) The *Arabidopsis thaliana* ABC transporter AtMRP5 controls root development and stomata movement. EMBO J 20:1875-1887. https://doi.org/10.1093/emboj/20.8.1875

Gao X, Gray J, Cohrs CW, Cook R, Albaugh TJ (2021) Longer greenup periods associated with greater wood volume growth in managed pine stands. Agric For Meteorol 297:108237. https://doi.org/10.1016/j.agrformet.2020.108237

Garrison E, Marth G (2012) Haplotype-based variant detection from short-read sequencing. arXiv preprint arXiv:12073907

Gilmour A, Cullis B, Harding S, Thompson R (2006) ASReml update: what's new in release 2.00. VSN International Ltd., Hemel Hempstead, UK

Goralogia GS, Redick TP, Strauss SH (2021) Gene editing in tree and clonal crops: progress and challenges. In Vitro Cell Dev Biol Plant 57:683-699. https://doi.org/10.1007/s11627-021-10197-x

Grabherr MG, Haas BJ, Yassour M et al (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29:644-652. https://doi.org/10.1038/nbt.1883

Granato ISC, Galli G, Couto EGDO, Souza MBE, Mendonça LF, Fritsche-Neto R (2018) snpReady: a tool to assist breeders in genomic analysis. Mol Breed 38:102. https://doi.org/10.1007/s11032-018-0844-8

Grattapaglia D (2017) Status and perspectives of genomic selection in forest tree breeding. In: Varshney R, Roorkiwal M, Sorrells M (eds) Genomic Selection for Crop Improvement. Springer International Publishing, Cham, pp 199-249

Grattapaglia D, Resende MDV (2011) Genomic selection in forest tree breeding. Tree Genet Genomes 7:241-255. https://doi.org/10.1007/s11295-010-0328-4

Grattapaglia D, Silva-Junior OB, Resende RT, Cappa EP, Müller BSF, Tan B, Isik F, Ratcliffe B, El-Kassaby YA (2018) Quantitative genetics and genomics converge to accelerate forest tree breeding. Front Plant Sci 9:1693-1693. https://doi.org/10.3389/fpls.2018.01693

Harfouche A, Meilan R, Kirst M, Morgante M, Boerjan W, Sabatti M, Mugnozza GS (2012) Accelerating the domestication of forest trees in a changing world. Trends Plant Sci 17:64-72. https://doi.org/10.1016/j.tplants.2011.11.005

Ingvarsson PK, Street NR (2011) Association genetics of complex traits in plants. New Phytol 189:909-922. https://doi.org/10.1111/j.1469-8137.2010.03593.x

Isik F, Bartholome J, Farjat A, Chancerel E, Raffin A, Sanchez L, Plomion C, Bouffier L (2016) Genomic selection in maritime pine. Plant Sci 242:108-119. https://doi.org/10.1016/j.plantsci.2015.08.006

Isik F, McKeand SE (2019) Fourth cycle breeding and testing strategy for *Pinus taeda* in the NC state university cooperative tree improvement program. Tree Genet Genomes 15:70. https://doi.org/10.1007/s11295-019-1377-y

Jannink JL, Lorenz AJ, Iwata H (2010) Genomic selection in plant breeding: from theory to practice. Brief Funct Genom 9:166-177. https://doi.org/10.1093/bfgp/elq001

Junkes CFDO, Júnior ATda, de Lima JC, de Costa F, Füller T, de Almeida MR, Neis FA, Rodrigues-Corrêa KCDS, Fett JP, Fett-Neto AG (2019) Resin tapping transcriptome in adult slash pine (*Pinus elliottii* var. *elliottii*). Ind Crops Prod 139:111545. https://doi.org/10.1016/j.indcrop.2019.111545

Kashkush K, Feldman M, Levy AA (2003) Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat. Nat Genet 33:102-106. https://doi.org/10.1038/ng1063

Khatiwada A, Wolf BJ, Yilmaz AS, Ramos PS, Pietrzak M, Lawson A, Hunt KJ, Kim HJ, Chung D (2022) GPA-Tree: statistical approach for functional-annotation-tree-guided prioritization of GWAS results. Bioinformatics 38:1067-1074. https://doi.org/10.1093/bioinformatics/btab802

Klein M, Perfus-Barbeoch L, Frelet A, Gaedeke N, Reinhardt D, Mueller-Roeber B, Martinoia E, Forestier C (2003) The plant multidrug resistance ABC transporter AtMRP5 is involved in guard cell hormonal signalling and water use. Plant J 33:119-129. https://doi.org/10.1046/j.1365-313x.2003.016012.x

Korte A, Farlow A (2013) The advantages and limitations of trait analysis with GWAS: a review. Plant Methods 9:29. https://doi.org/10.1186/1746-4811-9-29

Kumar V, Minz S (2014) Feature selection: a literature review. Smart Comput Rev 4:211-229. https://doi.org/10.6029/smartcr.2014.03.007

Lai M, Dong L, Yi M, Sun S, Zhang Y, Fu L, Xu Z, Lei L, Leng C, Zhang L (2017) Genetic variation, heritability and genotype × environment interactions of resin yield, growth traits and morphologic traits for *Pinus elliottii* at three progeny trials. Forests 8:409. https://doi.org/10.3390/f8110409

Langfelder P, Horvath S (2008) WGCNA: an R package for weighted correlation network analysis. BMC Bioinform 9:559. https://doi.org/10.1186/1471-2105-9-559

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357-359. https://doi.org/10.1038/nmeth.1923

Lauer E, Sims A, McKeand S, Isik F (2021) Genetic parameters and genotype-by-environment interactions in regional progeny tests of *Pinus taeda* L. in the Southern USA. For Sci 67:60-71. https://doi.org/10.1093/forsci/fxaa035

Lebedev VG, Lebedeva TN, Chernodubov AI, Shestibratov KA (2020) Genomic selection for forest tree improvement: methods, achievements and perspectives. Forests 11:1190. https://doi.org/10.3390/f11111190

Lee EK, Kwon M, Ko JH, Yi H, Hwang MG, Chang S, Cho MH (2004) Binding of sulfonylurea by AtMRP5, an Arabidopsis multidrug resistance-related protein that functions in salt tolerance. Plant Physiol 134:528-538. https://doi.org/10.1104/pp.103.027045

Lee T, Lee I (2018) araGWAB: network-based boosting of genome-wide association studies in *Arabidopsis thaliana*. Sci Rep 8:2925. https://doi.org/10.1038/s41598-018-21301-4

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

Lenz PRN, Nadeau S, Mottet MJ, Perron M, Isabel N, Beaulieu J, Bousquet J (2020) Multi-trait genomic selection for weevil resistance, growth, and wood quality in Norway spruce. Evol Appl 13:76-94. https://doi.org/10.1111/eva.12823

Li B, McKeand S, Weir R (1999) Tree improvement and sustainable forestry–impact of two cycles of loblolly pine breeding in the USA. For Genet 6:229-234

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics, 27(21), 2987-2993. https://doi.org/10.1093/bioinformatics/btr509

Li Y, Hu Y, Yang Y, Zheng W, Chen C, Li Z (2021) Selective breeding for juvenile survival in Chinese tongue sole (*Cynoglossus semilaevis*): heritability and selection response. Aquaculture 531:735901. https://doi.org/10.1016/j.aquaculture.2020.735901

Liu X, Huang M, Fan B, Buckler ES, Zhang Z (2016) Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. PLoS Genet 12:e1005767. https://doi.org/10.1371/journal.pgen.1005767

Lu M, Krutovsky KV, Nelson CD, West JB, Reilly NA, Loopstra CA (2017) Association genetics of growth and adaptive traits in loblolly pine (*Pinus taeda* L.) using whole-exome-discovered polymorphisms. Tree Genet Genomes 13:57. https://doi.org/10.1007/s11295-017-1140-1

Lu M, Seeve CM, Loopstra CA, Krutovsky KV (2018) Exploring the genetic basis of gene transcript abundance and metabolite levels in loblolly pine (*Pinus taeda* L.) using association mapping and network construction. BMC Genet 19:100. https://doi.org/10.1186/s12863-018-0687-7

Luo Y, El-Madany T, Ma X et al (2020) Nutrients and water availability constrain the seasonality of vegetation activity in a Mediterranean ecosystem. Glob Chang Biol 26:4379-4400. https://doi.org/10.1111/gcb.15138

Manolio TA, Collins FS, Cox NJ et al (2009) Finding the missing heritability of complex diseases. Nature 461:747-753. https://doi.org/10.1038/nature08494

Mao J, He Z, Hao J, Liu T, Chen J, Huang S (2019) Identification, expression, and phylogenetic analyses of terpenoid biosynthesis-related genes in secondary xylem of loblolly pine (*Pinus taeda* L.) based on transcriptome analyses. PeerJ 7:e6124. https://doi.org/10.7717/peerj.6124

Maslov S, Sneppen K (2002) Specificity and stability in topology of protein networks. Science 296:910-913. https://doi.org/10.1126/science.1065103

Matallana-Ramirez LP, Whetten RW, Sanchez GM, Payn KG (2021) Breeding for climate change resilience: a case study of loblolly pine (*Pinus taeda* L.) in North America. Front Plant Sci 12:606908. https://doi.org/10.3389/fpls.2021.606908

Matsunaga W, Ohama N, Tanabe N, Masuta Y, Masuda S, Mitani N, Yamaguchi-Shinozaki K, Ma JF, Kato A, Ito H (2015) A small RNA mediated regulation of a stress-activated retrotransposon and the tissue specific transposition during the reproductive period in Arabidopsis. Front Plant Sci 6:48. https://doi.org/10.3389/fpls.2015.00048

McKeand SE, Jokela EJ, Huber DA, Byram TD, Allen HL, Li B, Mullin TJ (2006) Performance of improved genotypes of loblolly pine across different soils, climates, and silvicultural inputs. For Ecol Manag 227:178-184. https://doi.org/10.1016/j.foreco.2006.02.016

Meuwissen TH, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819-1829. https://doi.org/10.1093/genetics/157.4.1819

Mutwil M, Usadel B, Schutte M, Loraine A, Ebenhoh O, Persson S (2010) Assembly of an interactive correlation network for the Arabidopsis genome using a novel heuristic clustering algorithm. Plant Physiol 152:29-43. https://doi.org/10.1104/pp.109.145318

Neves LG, Davis JM, Barbazuk WB, Kirst M (2013) Whole-exome targeted sequencing of the uncharacterized pine genome. Plant J 75:146-156. https://doi.org/10.1111/tpj.12193

Norman A, Taylor J, Edwards J, Kuchel H (2018) Optimising genomic selection in wheat: effect of marker density, population size and population structure on prediction accuracy. G3 Genes Genomes Genet 8:2889-2899. https://doi.org/10.1534/g3.118.200311

O'Brien IEW, Smith DR, Gardner RC, Murray BG (1996) Flow cytometric determination of genome size in Pinus. Plant Sci 115:91-99. https://doi.org/10.1016/0168-9452(96)04356-7

Oliver S (2000) Guilt-by-association goes global. Nature 403:601-603. https://doi.org/10.1038/35001165

Ouyang J, Shao X, Li J (2000) Indole-3-glycerol phosphate, a branchpoint of indole-3-acetic acid biosynthesis from the tryptophan biosynthetic pathway in *Arabidopsis thaliana*. Plant J 24:327-333. https://doi.org/10.1046/j.1365-313x.2000.00883.x

Pan W, Liang J, Sui J et al (2021) ABA and bud dormancy in perennials: current knowledge and future perspective. Genes 12:1635. https://doi.org/10.3390/genes12101635

Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C (2015) Salmon: accurate, versatile and ultrafast quantification from RNA-seq data using lightweight-alignment. bioRxiv. https://doi.org/10.1101/021592

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825-2830

Perera D, Magbanua ZV, Thummasuwan S, Mukherjee D, Arick M, 2nd, Chouvarine P, Nairn CJ, Schmutz J, Grimwood J, Dean JFD, Peterson DG (2018) Exploring the loblolly pine (*Pinus taeda* L.) genome by BAC sequencing and Cot analysis. Gene 663:165-177. https://doi.org/10.1016/j.gene.2018.04.024

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

Piles M, Bergsma R, Gianola D, Gilbert H, Tusell L (2021) Feature selection stability and accuracy of prediction models for genomic prediction of residual feed intake in pigs using machine learning. Front Genet 12:611506. https://doi.org/10.3389/fgene.2021.611506

Plomion C, Chagné D, Pot D et al (2007) Pines. In: Kole C (ed) Forest trees. Springer, Berlin, Heidelberg, pp 29-92

Poovaiah C, Phillips L, Geddes B, Reeves C, Sorieul M, Thorlby G (2020) Genome editing with CRISPR/Cas9 in *Pinus radiata* (D. Don). BMC Plant Biol 21:363. https://doi.org/10.1186/s12870-021-03143-x

R Core Team R (2013) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria

Resende MF, Jr., Munoz P, Resende MD, Garrick DJ, Fernando RL, Davis JM, Jokela EJ, Martin TA, Peter GF, Kirst M (2012a) Accuracy of genomic selection methods in a standard data set of loblolly pine (*Pinus taeda* L.). Genetics 190:1503-1510. https://doi.org/10.1534/genetics.111.137026

Resende MFR, Jr., Munoz P, Acosta JJ, Peter GF, Davis JM, Grattapaglia D, Resende MDV, Kirst M (2012b) Accelerating the domestication of trees using genomic selection: accuracy of prediction models across ages and environments. New Phytol 193:617-624. https://doi.org/10.1111/j.1469-8137.2011.03895.x

Ruan J, Ma L, Shi Y, Han W (2004) The impact of pH and calcium on the uptake of fluoride by tea plants (*Camellia sinensis* L.). Ann Bot 93:97-105. https://doi.org/10.1093/aob/mch010

Rutkoski JE (2019) A practical guide to genetic gain. In: Sparks DL (ed) Advances in agronomy. Academic Press, Cambridge, MA, pp 217-249

Sampson DA, Johnsen KH, Ludovici KH, Albaugh TJ, Maier CA (2001) Stand-scale correspondence in empirical and simulated labile carbohydrates in loblolly pine. For Sci 47:60-68. https://doi.org/10.1093/forestscience/47.1.60

Santantonio N, Atanda SA, Beyene Y et al (2020) Strategies for effective use of genomic information in crop breeding programs serving Africa and South Asia. Front Plant Sci 11:353. https://doi.org/10.3389/fpls.2020.00353

Santini F, Kefauver SC, Araus JL, de Dios VR, Garcia SM, Grivet D, Voltas J (2021) Bridging the genotype-phenotype gap for a Mediterranean pine by semi-automatic crown identification and multispectral imagery. New Phytol 229:245-258. https://doi.org/10.1111/nph.16862

Schaefer RJ, Michno JM, Jeffers J, Hoekenga O, Dilkes B, Baxter I, Myers CL (2018) Integrating coexpression networks with gwas to prioritize causal genes in maize. Plant Cell 30:2922-2942. https://doi.org/10.1105/tpc.18.00299

Schmidt P, Hartung J, Bennewitz J, Piepho HP (2019) Heritability in plant breeding on a genotype-difference basis. Genetics 212:991-1008. https://doi.org/10.1534/genetics.119.302134

Seppey M, Manni M, Zdobnov EM (2019) BUSCO: assessing genome assembly and annotation completeness. In: Kollmar M (ed) Gene prediction: methods and protocols. Springer, New York, NY, pp 227-245

Shalizi MN, Gezan SA, McKeand SE, Sherrill JR, Cumbie WP, Whetten RW, Isik F (2020) Correspondence between breeding values of the same *Pinus taeda* L. genotypes from clonal trials and half-sib seedling progeny trials. For Sci 66:600-611. https://doi.org/10.1093/forsci/fxaa016

Shalizi MN, Isik F (2019) Genetic parameter estimates and GxE interaction in a large cloned population of *Pinus taeda* L. Tree Genet Genomes 15:46. https://doi.org/10.1007/s11295-019-1352-7

Shalizi MN, Payn KG, Isik F (2022) Genetic linkage between the training and selection sets impacts the predictive ability of SNP markers in a cloned population of *Pinus taeda* L. Tree Genet Genomes 18:4. https://doi.org/10.1007/s11295-021-01532-9

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13:2498-2504. https://doi.org/10.1101/gr.1239303

Sherrill JR, Bullock BP, Mullin TJ, McKeand SE, Purnell RC (2011) Total and merchantable stem volume equations for midrotation loblolly pine (*Pinus taeda* L.). South J Appl For 35:105-108. https://doi.org/10.1093/sjaf/35.3.105

Simkin AJ (2019) Genetic engineering for global food security: photosynthesis and biofortification. Plants (Basel) 8:586. https://doi.org/10.3390/plants8120586

Singh M, Kaur K, Sharma A, Kaur R, Joshi D, Chatterjee M, Dandapath I, Kaur A, Singh H, Singh P (2021) Genome-wide characterization of peptidyl-prolyl cis-trans isomerases in *Penicillium* and their regulation by salt stress in a halotolerant *P. oxalicum*. Sci Rep 11:12292. https://doi.org/10.1038/s41598-021-91602-8

Spindel JE, McCouch SR (2016) When more is better: how data sharing would accelerate genomic selection of crop plants. New Phytol 212:814-826. https://doi.org/10.1111/nph.14174

Stasolla C, Katahira R, Thorpe TA, Ashihara H (2003) Purine and pyrimidine nucleotide metabolism in higher plants. J Plant Physiol 160:1271-1295. https://doi.org/10.1078/0176-1617-01169

Sul JH, Martin LS, Eskin E (2018) Population structure in genetic studies: confounding factors and mixed models. PLoS Genet 14:e1007309. https://doi.org/10.1371/journal.pgen.1007309

Supek F, Bosnjak M, Skunca N, Smuc T (2011) REVIGO summarizes and visualizes long lists of gene ontology terms. PLoS One 6:e21800. https://doi.org/10.1371/journal.pone.0021800

Tam V, Patel N, Turcotte M, Bosse Y, Pare G, Meyre D (2019) Benefits and limitations of genome-wide association studies. Nat Rev Genet 20:467-484. https://doi.org/10.1038/s41576-019-0127-1

Telfer E, Graham N, Macdonald L, Li Y, Klapste J, Resende M, Jr., Neves LG, Dungey H, Wilcox P (2019) A high-density exome capture genotype-by-sequencing panel for forestry breeding in *Pinus radiata*. PLoS One 14:e0222640. https://doi.org/10.1371/journal.pone.0222640

Tenesa A, Wright AF, Knott SA, Carothers AD, Hayward C, Angius A, Persico I, Maestrale G, Hastie ND, Pirastu M, Visscher PM (2004) Extent of linkage disequilibrium in a Sardinian sub-isolate: sampling and methodological considerations. Hum Mol Genet 13:25-33. https://doi.org/10.1093/hmg/ddh001

Thistlethwaite FR, Ratcliffe B, Klapste J, Porth I, Chen C, Stoehr MU, El-Kassaby YA (2019) Genomic selection of juvenile height across a single-generational gap in Douglas-fir. Heredity (Edinb) 122:848-863. https://doi.org/10.1038/s41437-018-0172-0

Tran KN, Choi JI (2020) Comparative transcriptome analysis of high-growth and wild-type strains of *Pyropia yezoensis*. Acta Bot Croat 79:148-156. https://doi.org/10.37427/botcro-2020-020

Traylor-Knowles N, Rose NH, Sheets EA, Palumbi SR (2017) Early transcriptional responses during heat stress in the coral *Acropora hyacinthus*. Biol Bull 232:91-100. https://doi.org/10.1086/692717

Vose JM, Allen HL (1988) Leaf area, stemwood growth, and nutrition relationships in loblolly Pine. For Sci 34:547-563. https://doi.org/10.1093/forestscience/34.3.547

Wang X, Xu Y, Hu Z, Xu C (2018) Genomic selection methods for crop improvement: current status and prospects. Crop J 6:330-340. https://doi.org/10.1016/j.cj.2018.03.001

Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393:440-442. https://doi.org/10.1038/30918

Wegrzyn JL, Liechty JD, Stevens KA et al (2014) Unique features of the loblolly pine (*Pinus taeda* L.) megagenome revealed through sequence annotation. Genetics 196:891-909. https://doi.org/10.1534/genetics.113.159996

White TL, Adams WT, Neale DB (2007) Forest genetics. CABI Publishing, Cambridge, UK

Wickham H (2016) ggplot2: elegant graphics for data analysis. Springer, New York, NY

Wu HX, Eldridge KG, Matheson AC, Powell MB, McRae TA, Butcher TB, Johnson IG (2007) Achievements in forest tree improvement in Australia and New Zealand 8. Successful introduction and breeding of radiata pine in Australia. Aust For 70:215-225. https://doi.org/10.1080/00049158.2007.10675023

Xiang B, Li B, Isik F (2003) Time trend of genetic parameters in growth traits of *Pinus taeda* L. Silvae Genet 52:114-120

Xu Y, Liu X, Fu J, Wang H, Wang J, Huang C, Prasanna BM, Olsen MS, Wang G, Zhang A (2020) Enhancing genetic gain through genomic selection: from livestock to plants. Plant Commun 1:100005. https://doi.org/10.1016/j.xplc.2019.100005

Yin L, Zhang H, Tang Z, Xu J, Yin D, Zhang Z, Yuan X, Zhu M, Zhao S, Li X, Liu X (2020) rMVP: a memory-efficient, visualization-enhanced, and parallel-accelerated tool for genome-wide association study. Genom Proteom Bioinform 19:619-628. https://doi.org/10.1016/j.gpb.2020.10.007

Yuan Z, Gao Q, He Y, Zhang X, Li F, Zhao J, Xue F (2012) Detection for gene-gene co-association via kernel canonical correlation analysis. BMC Genet 13:83. https://doi.org/10.1186/1471-2156-13-83

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

Zapata-Valenzuela J, Whetten RW, Neale D, McKeand S, Isik F (2013) Genomic estimated breeding values using genomic relationship matrices in a cloned population of loblolly pine. G3 Genes Genomes Genet 3:909-916. https://doi.org/10.1534/g3.113.005975

Zhang H, Zhang Y, Zhang D, Dong L, Liu K, Wang Y, Yang C, Chiang VL, Tigabu M, Zhao X (2020) Progeny performance and selection of superior trees within families in *Larix olgensis*. Euphytica 216:60. https://doi.org/10.1007/s10681-020-02596-9

Zimin AV, Stevens KA, Crepeau MW, Puiu D, Wegrzyn JL, Yorke JA, Langley CH, Neale DB, Salzberg SL (2017) An improved assembly of the loblolly pine mega-genome using long-read single-molecule sequencing. Gigascience 6:1-4. https://doi.org/10.1093/gigascience/giw016

Zobel B, Talbert J (1984) Applied forest tree improvement. The Blackburn Press, Caldwell, NJ

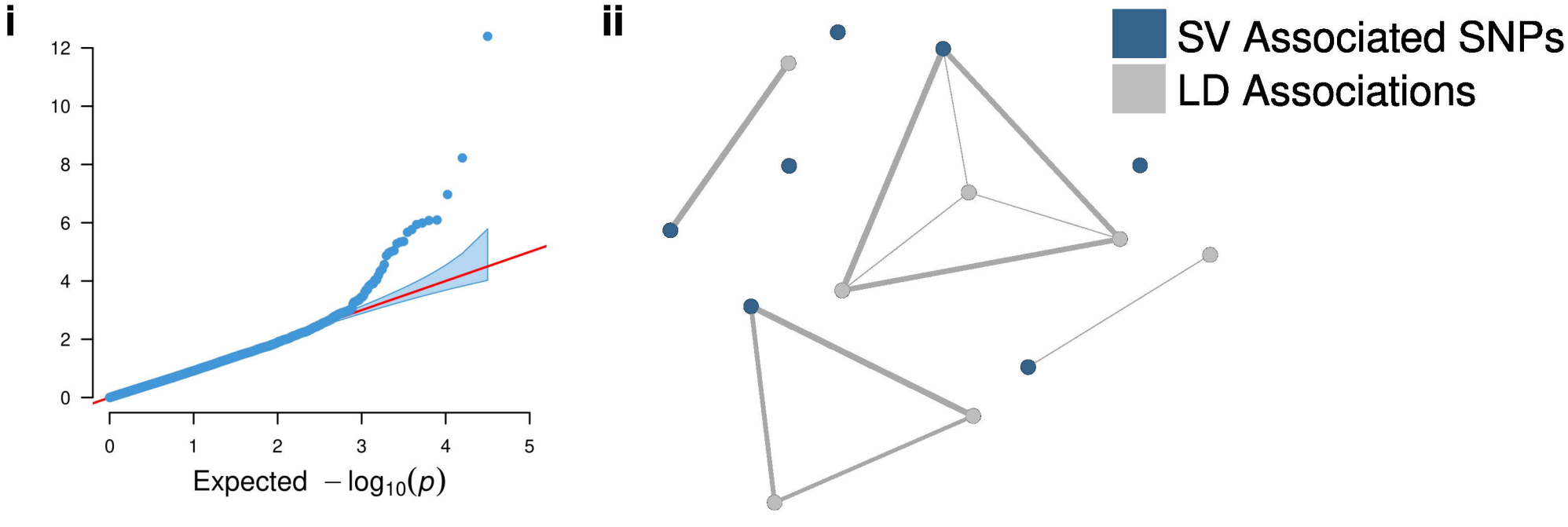
Figure Legends

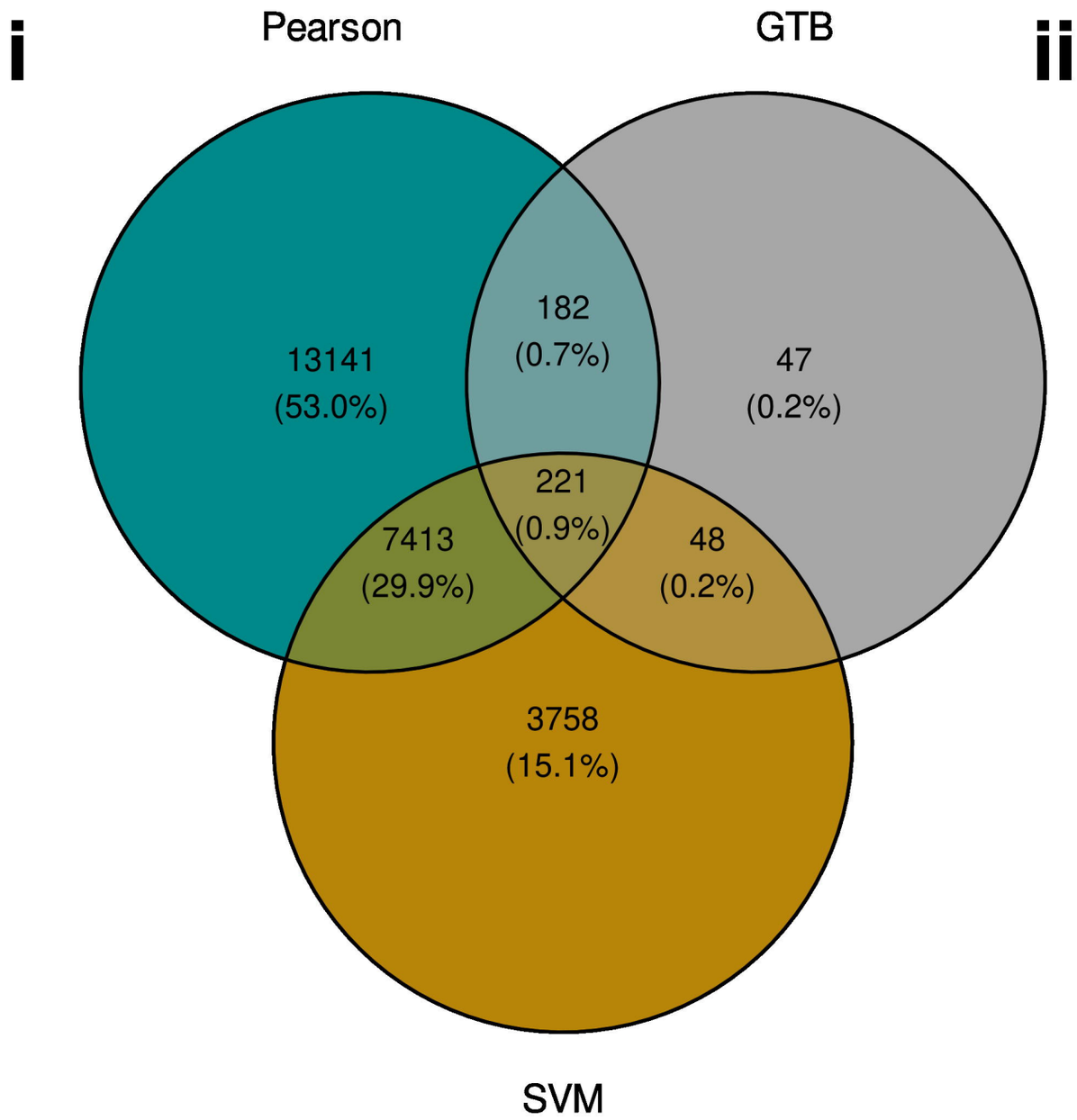
Fig. 1 Genotypic analyses performed on the SNP data separated into (i) linkage disequilibrium (LD) decay considering squared Pearson correlations (R^2) and distances calculated based on the respective probes (measured in kilobase pairs); (ii) principal component analysis (PCA) scatter plot of the *Pinus taeda* individuals colored according to family (the percentage of variance explained was approximately 5% and 4% for the first and second principal components, respectively)

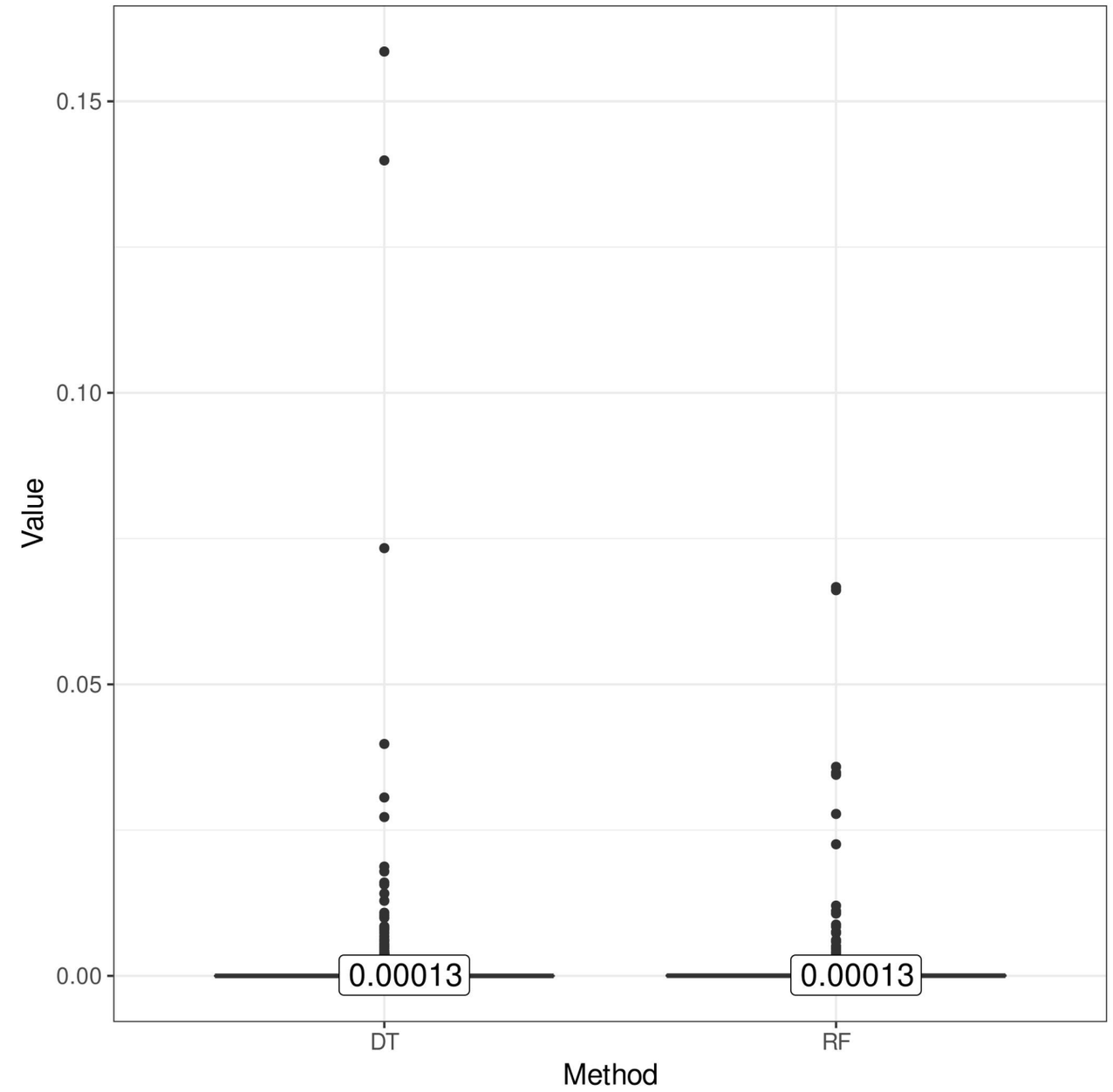
Fig. 2 (i) Quantile–quantile (QQ) plot of FarmCPU models in the genome-wide association study (GWAS) with the stem volume (SV) trait. The set of markers associated with SV together with the other SNPs in linkage disequilibrium ($R^2 \geq 0.5$) is displayed in (ii), which is a network representing these associations within the set

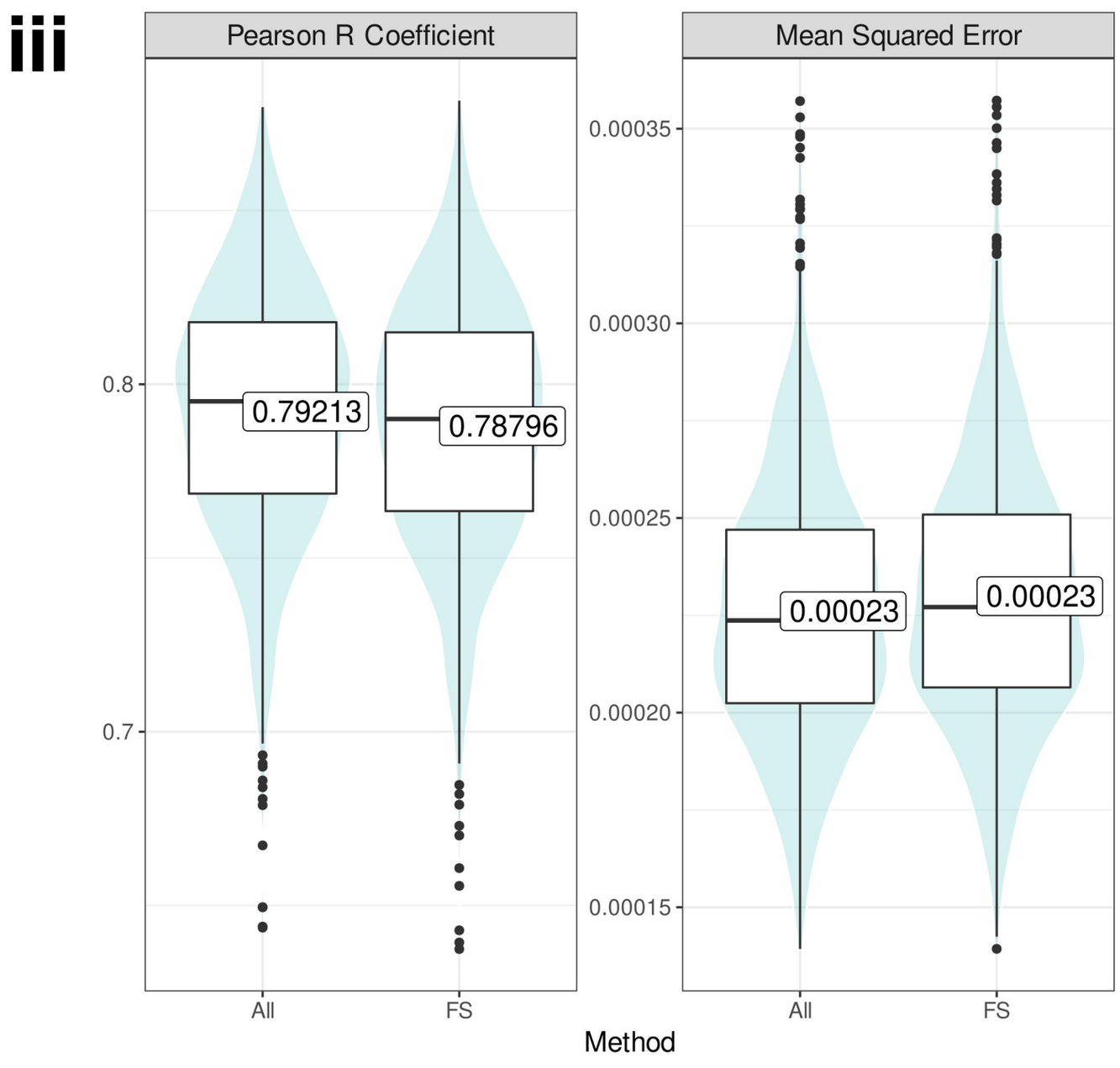
Fig. 3 Genomic prediction results for stem volume using a total of 31,589 SNPs (All) and 7,864 selected through feature selection (FS) based on (a) Pearson correlations (p value of 0.05), (b) gradient tree boosting (GTB), and (c) support vector machine (SVM). (i) Markers selected through each FS method. (ii) Distribution of the feature importance estimated for each of the 7,864 SNPs obtained through a decision tree (DT) and a random forest (RF) model. (iii) Predictive evaluations

Fig. 4 (i) Coexpression network created with the highest reciprocal rank (HRR) methodology considering groups of coexpressed genes selected in the weighted gene coexpression network analysis (WGCNA) modeled for *Pinus radiata* transcriptome data. Each node represents a gene, colored with (a) gray (nondirect association with stem volume (SV) phenotype), (b) light blue (genes surrounding markers associated with SV through machine learning methodologies), and dark blue (genes of (b) and surrounding GWAS SNPs). (ii) Gene Ontology (GO) terms from the network summarized in a tree map

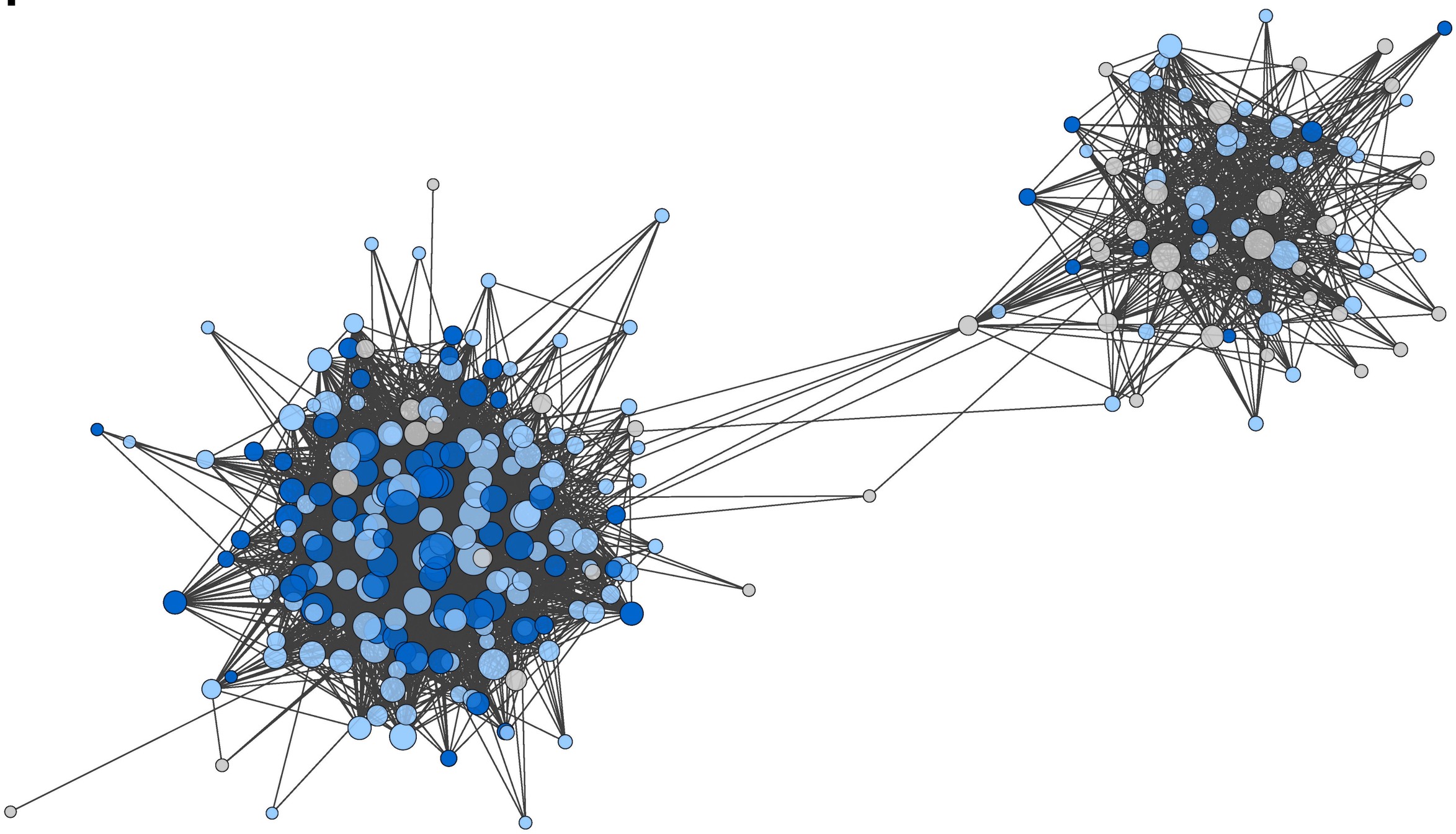

i
12
10
8
6
4
2
0
0
1
2
3
4
5
Expected $-\log_{10}(p)$
ii
SV Associated SNPs
LD Associations


i
Pearson
GTB
SVM
13141
(53.0%)
182
(0.7%)
47
(0.2%)
221
(0.9%)
7413
(29.9%)
48
(0.2%)
3758
(15.1%)

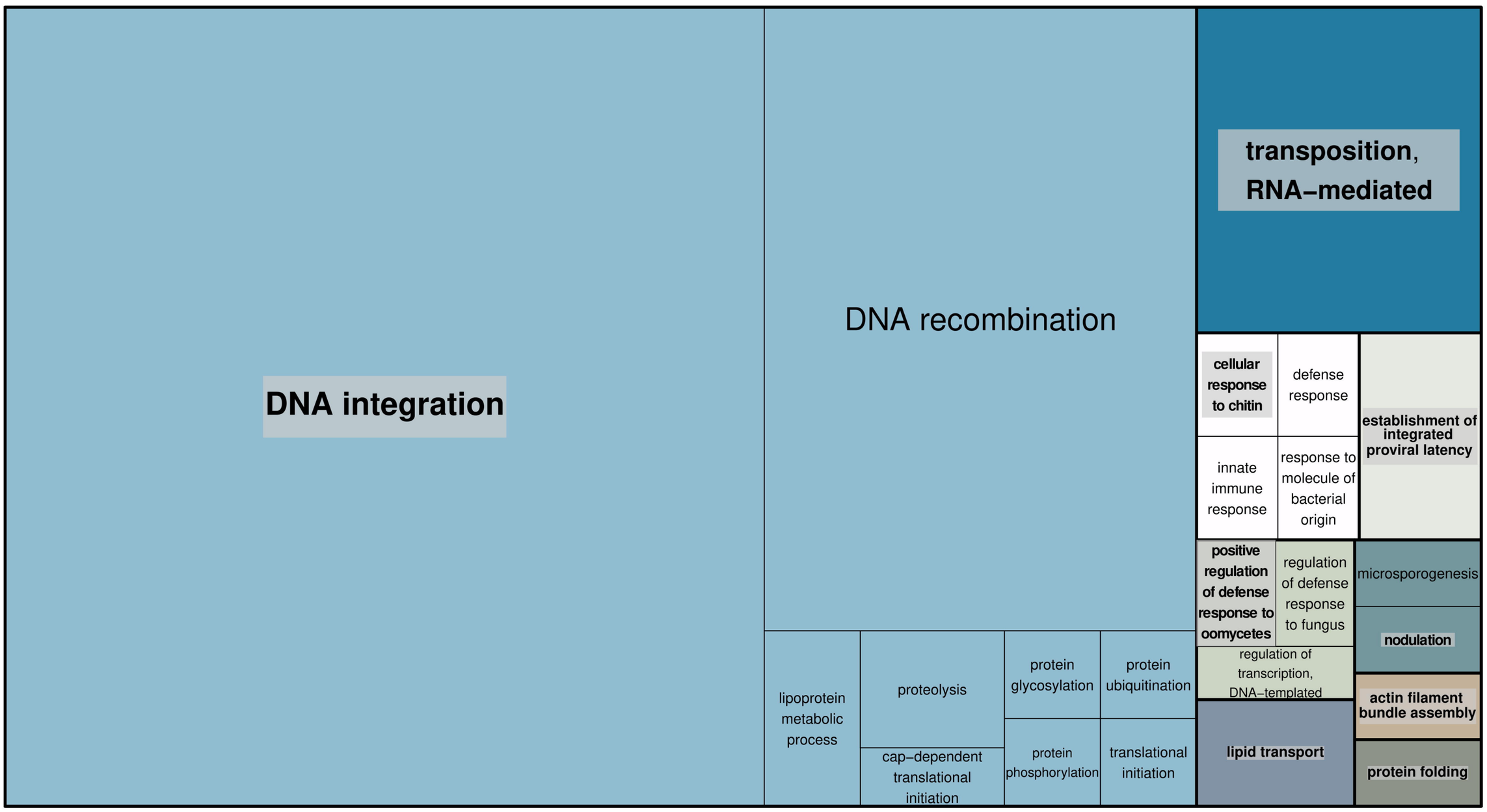

ii
Value
0.15
0.10
0.05
0.00
0.00013
0.00013
DT
RF
Method

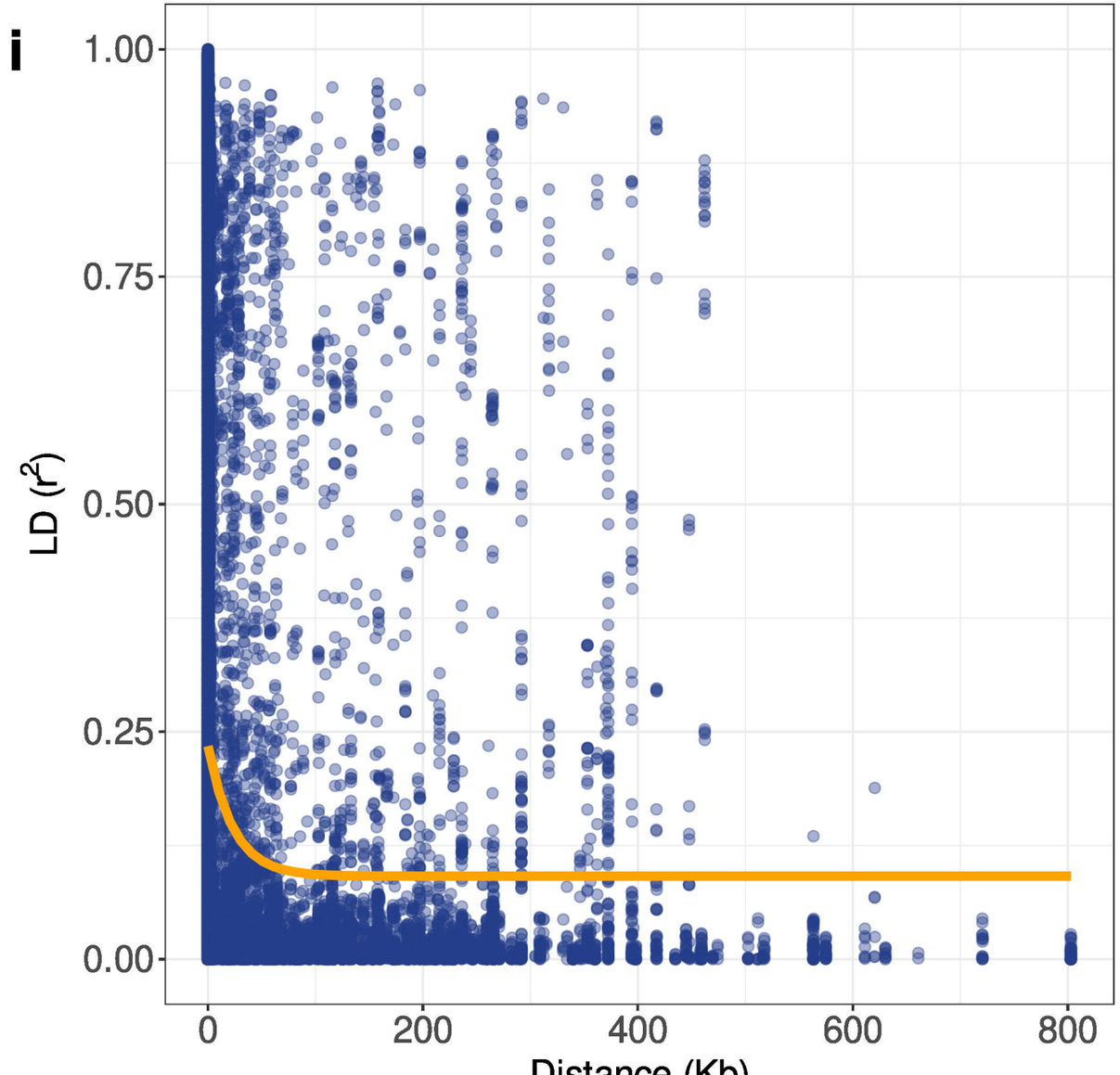

iii
Pearson R Coefficient
Mean Squared Error
0.8
0.7
0.79213
0.78796
0.00035
0.00030
0.00025
0.00020
0.00015
0.00023
0.00023
All
FS
All
FS
Method

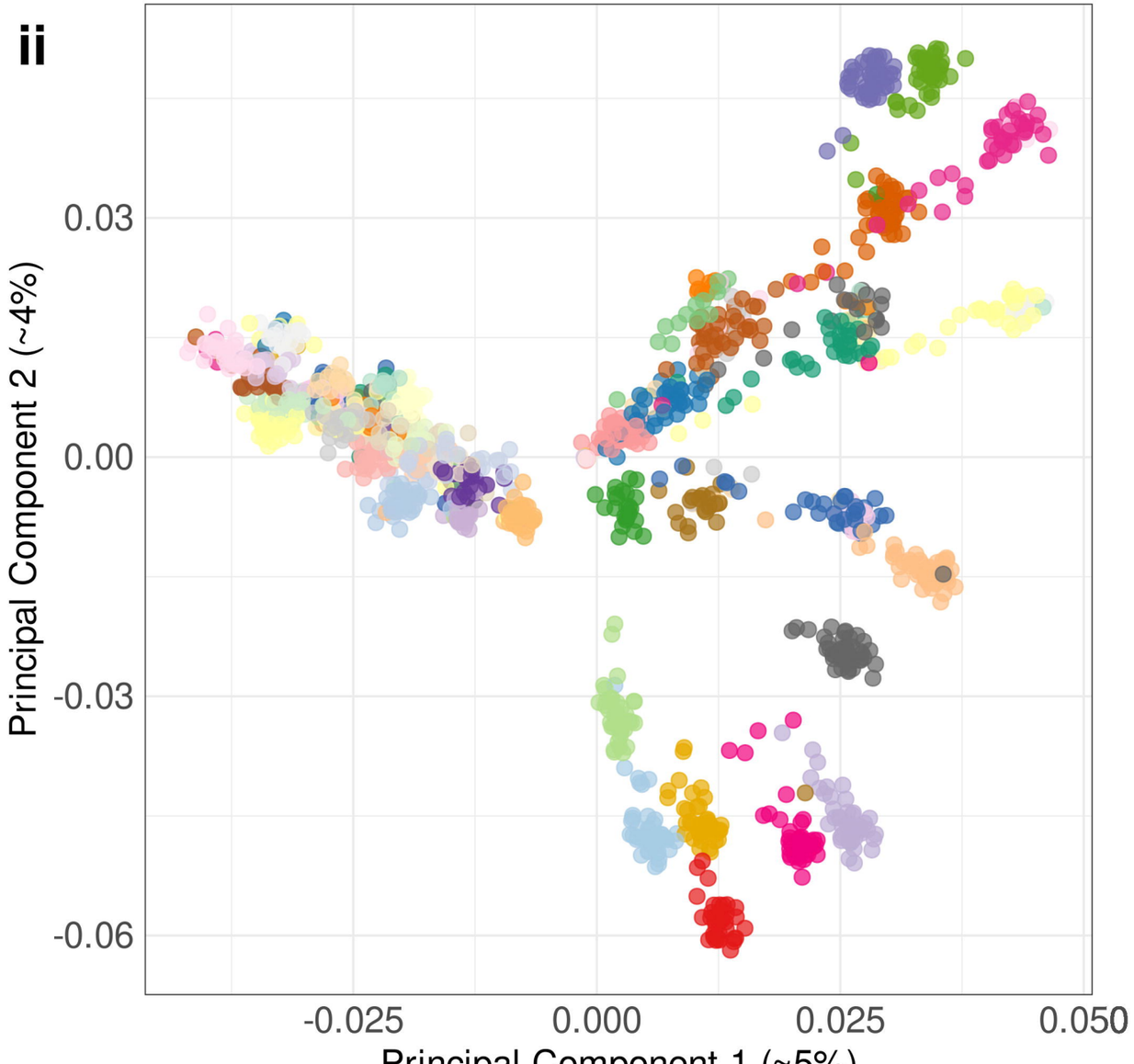
i



ii


DNA integration
DNA recombination
transposition, RNA−mediated
cellular response to chitin
defense response
establishment of integrated proviral latency
innate immune response
response to molecule of bacterial origin
positive regulation of defense response to oomycetes
regulation of defense response to fungus
microsporogenesis
nodulation
regulation of transcription, DNA−templated
actin filament bundle assembly
lipid transport
protein folding
lipoprotein metabolic process
proteolysis
protein glycosylation
protein ubiquitination
cap−dependent translational initiation
protein phosphorylation
translational initiation


i
1.00
0.75
0.50
0.25
0.00
LD (r^2)
0
200
400
600
800
Distance (Kb)


ii
0.03
0.00
-0.03
-0.06
Principal Component 2 (~4%)
-0.025
0.000
0.025
0.050
Principal Component 1 (~5%)