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

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

[a]*Molecular Biology and Genetic Engineering Center (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

Abstract

Pinus taeda (loblolly pine [LP]), a long-lived tree species, is one of the world's most economically important forest trees. Genetic improvement programs for loblolly and other *Pinus* species have focused on survival, biomass growth, resistance to diseases and pests, and stem shape. Among the growth traits, volume is the most widely considered in tree improvement programs. Despite the great interest in increasing volume growth, there are significant challenges to unraveling the molecular mechanisms behind this quantitative trait since it is presumably influenced by the action of large numbers of genes that may interact epistatically through unknown molecular mechanisms. The challenge of uncovering the genetic variants involved in variation for growth traits and their interaction is even greater in conifers such as LP because of the extremely large size and high complexity of *Pinus* genomes. Here we present a comprehensive genetic analysis of LP involving genetic markers in association with volume (via genome-wide association studies [GWAS] and machine learning [ML]) to uncover pathways involved with good phenotypes for volume. The objective of this data integration was to provide a functional characterization of the genetic marker regions selected by GWAS and ML. We used a population of LP in the 2nd cycle of breeding and testing composed of full-sib progenies established at seven sites by the Cooperative Forest Genetics Research Program at the University of Florida. A total of 1,692 individuals were phenotyped and genotyped using sequence capture

probes targeting putative genes. Probes were developed based on an elite germplasm transcriptome from LP. 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 (R Pearson coefficient) of approximately 0.79. For genome annotation and the construction of gene coexpression networks, three transcriptomes were assembled based on data from different pine species (*Pinus taeda*, *Pinus elliottii* and *Pinus radiata*). We selected genes associated with the target trait and assessed the cascade of related molecular mechanisms within coexpression networks. These results advance our understanding of the genetics influencing wood traits and reveal candidate genes for future functional studies as well as increase our understanding of quantitative genetics and the genomics of complex phenotypic variations in LP.

Keywords: Forest breeding, genomic prediction, genomic selection, loblolly pine, RNA-Seq, stem volume

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 for the timber, and 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. In the meantime, there is increasing pressure for its improvement to increase productivity, forest values, and ecosystem services (Isik and McKeand, 2019; Telfer et al., 2019).

A 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). Even though 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; Mao et al., 2019; de Oliveira Junkes 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 for most cultivated tree species, including LP (Lu et al., 2018; Shalizi and Isik, 2019). However, increasing volume based on molecular data is hampered by the polygenic nature of this complex trait and substantial variation between populations (Lauer et al., 2020). 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; Grattapaglia 2014; Isik et al. 2016; 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 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, upon integration with different layers of omics, can provide a greater picture of unknown molecular interactions.

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 the 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 the genetic architecture of volume in seven populations in Florida. 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 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 (Team et al., 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, H, 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 (Danecek et al., 2015) 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; Mao et al., 2019; de Oliveira Junkes, 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 with 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 associated with SV was expanded through LD estimations (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 (Pérez & 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 the overall population mean, and *e* 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 σ^2_γ representing the genetic variance.

Contrasted with 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; Aono et al., 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 from 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.

2.8. 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 & 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 & 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 in 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 (Barabási 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, 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. Probes regions located less than 500 bp from each other were combined, resulting in 10,412 sequences to which the reads were aligned to.

Among the sequenced reads, 84.5% 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).

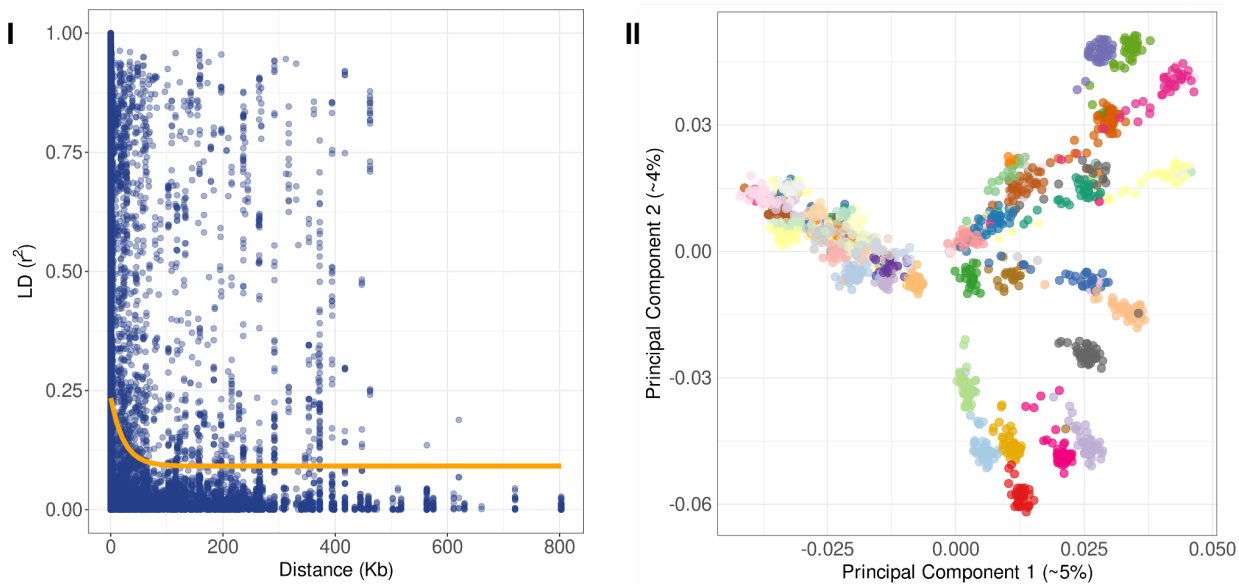

Fig. 1. Genotypic analyses performed on the SNP data separated into (I) linkage disequilibrium (LD) decay considering squared Pearson correlations (R^2) and distances calculated based on the respective probes (measured in kilobase pairs); (II) principal component analysis (PCA) scatter plot of the *Pinus taeda* individuals colored according to family (the percentage of variance

explained was approximately 5% and 4% for the first and second principal components, respectively).

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 assembled genes could be recovered by BUSCO, however 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 Trinotate pipeline (Supplementary Table 6). As our final set of genes assembled, we only considered genes expressed in at least 75% of the samples of 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, Supplementary Figs. 6-7). However, to avoid environmental biases, we included the sites and family structure as covariates in the final GWAS model. Figure 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, Supplementary Fig. 8), with five of them also identified using the model without covariates.

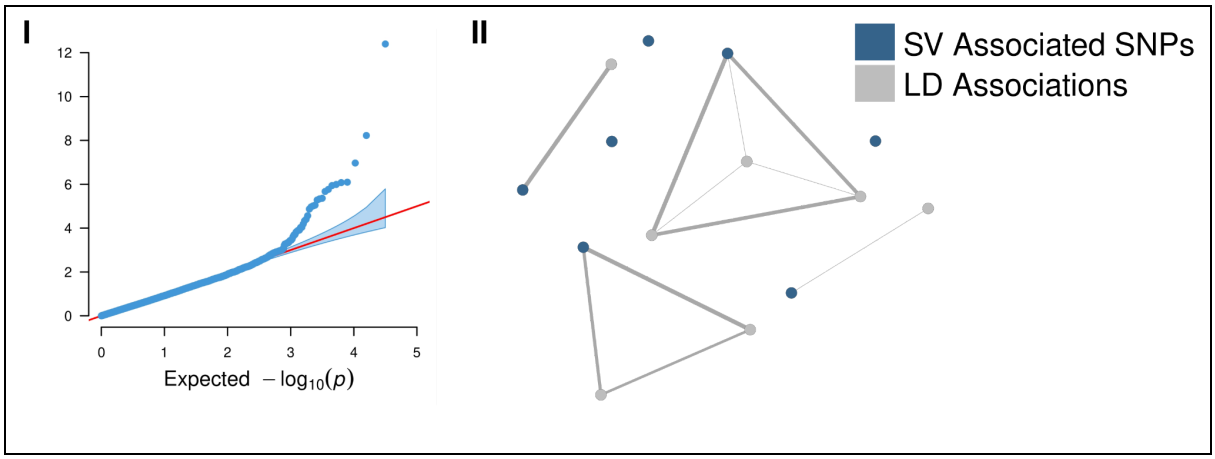

Fig. 2. (I) Quantile-quantile (QQ) plot of FarmCPU models in the genome-wide association study (GWAS) with the stem volume (SV) trait. The set of markers associated with SV together with the other SNPs in linkage disequilibrium ($R^2 \geq 0.5$) is displayed in (II), which is a network representing these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

This set of seven markers was extended by linkage disequilibrium (LD) tests, which resulted in the identification of seven additional markers (Figure 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.

Even though 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 8-9). 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.

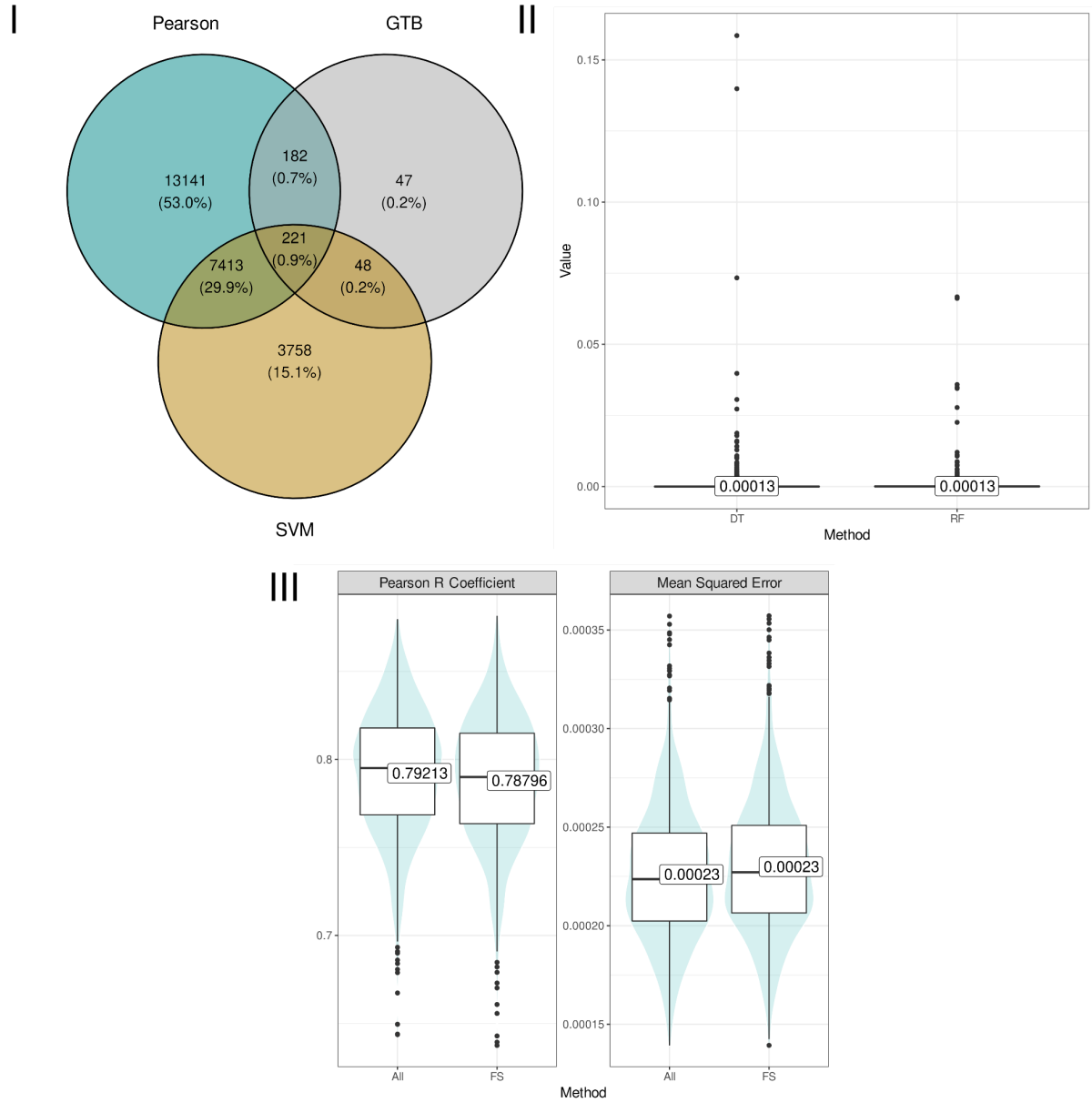

Fig. 3. Genomic prediction results for stem volume using a total of 31,589 SNPs (All) and 7,864 selected through feature selection (FS) based on (a) Pearson correlations (p value of 0.05), (b)

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.

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 molecular mechanisms of SV regulation, and used it to create a novel network based on HRR methodology. We could create 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 metabolic, nitrogen compound metabolic process and transposition, viral latency, and viral genome integration into host DNA and cell (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 closely located 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).

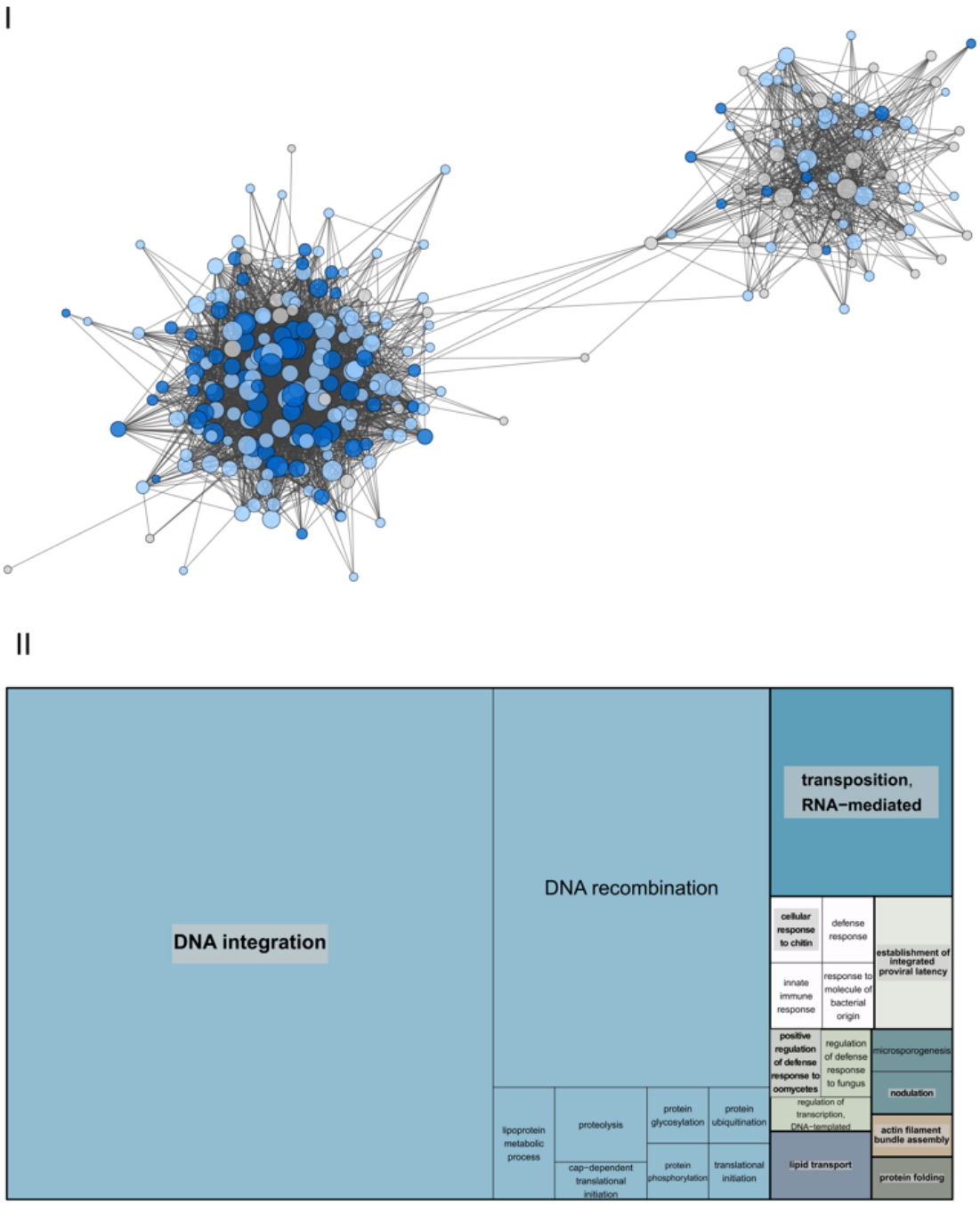

Fig. 4. (I) Coexpression network created with the highest reciprocal rank (HRR) methodology considering groups of coexpressed genes selected in the weighted gene coexpression network analysis (WGCNA) modeled for *Pinus radiata* transcriptome data. Each node represents a gene, colored with (a) gray (nondirect 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.

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).

4.1. 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, 2012). 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 (2012); 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., 2012). These promising results were obtained despite the PA being in the range 0.36 to 0.49 and using a group of 4,853 SNPs in the best scenarios (Resende et al., 2012). 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., 2012; de Almeida 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 deal with 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. 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; Li 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 GP of forest species, as well as different methodologies to evaluate the effectiveness or accuracy of these models, what these studies have shown are 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 deal with 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).

Even though GS has driven significant improvements in forest breeding programs, correctly estimating the effects of a wide set of markers into 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 (Neves et al., 2021). 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., 2021); 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). et al., 2018).

4.2. 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 still 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; Martinoia et al., 2001; Martinoia et al., 2002; 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 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 coassociation (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 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.

4.3. Integrative Approach

Although growth is a complex trait governed by several genes (Santini et al. 2021; Francisco et al., 2021), only seven markers were identified to be associated with *Pinus* by GWAS. Santini et al. (2020) argues 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 to 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.

4.4 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 S., 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 Pta and Pel transcriptomes, we defined different gene coexpression networks with WGCNA methodology, each one 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 with 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 carbohydrates stored, 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 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 the plant by disturbing its metabolism and enzymatic activities (Ruan et al. 2003). 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 (Barabási 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 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 et al., 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 studies of omics (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.

4.5. 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).

Conclusion

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.

Acknowledgements

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

Author Contribution Statement

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

References

Aono, A. H., Costa, E. A., Rody, H. V. S., Nagai, J. S., Pimenta, R. J. G., Mancini, M. C., ... & Kuroshu, R. M. (2020). Machine learning approaches reveal genomic regions associated with sugarcane brown rust resistance. *Scientific reports*, *10*(1), 1-16.

Aono, A. H., Costa, E. A., Rody, H. V. S., Nagai, J. S., Pimenta, R. J. G., Mancini, M. C., ... & Kuroshu, R. M. (2022a). A joint learning approach for genomic prediction in polyploid grasses. *Scientific reports*, *12*, 12499.

Aono, A. H., Francisco, F. R., Souza, L. M., de Souza Gonçalves, P., Scaloppi, E. J., Le Guen, V., ... & de Souza, A. P. (2022b). A divide-and-conquer approach for genomic prediction in rubber tree using machine learning. *Scientific reports*, *12, 18023*.

Albaugh, T.J. ; Fox, T.R. ; Maier, C.A. ; Campoe, O.C. ; Rubilar, R.A. ; Cook, R.L.; Raymond, J.E.; Alvares, C.A. ; Stape, J.L. (2018). A common garden experiment examining light use efficiency and heat sum to explain growth differences in native and exotic Pinus taeda. Forest Ecology and Management, 425 , pp. 35-44, 10.1016/j.foreco.2018.05.033

Alfenas, A. C., Mafia, R. G., Sartório, R. C., Binoti, D. H., Silva, R. R., Lau, D., & Vanetti, C. A. (2006). Ralstonia solanacearum em viveiros clonais de eucalipto no Brasil. Fitopatologia Brasileira, 31(4), 357-366.

Alexa, A., Rahnenführer, J., & Lengauer, T. (2006). Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics, 22(13), 1600-1607.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. molecular biology* 215, 403–410 (1990).

Atwood, Ryan A., Timothy L. White, and Dudley A. Huber. "Genetic parameters and gains for growth and wood properties in Florida source loblolly pine in the southeastern United States." Canadian Journal of Forest Research 32.6 (2002): 1025-1038.

Baker, E. A. *et al.* Comparative transcriptomics among four white pine species. *G3: Genes, Genomes, Genet.* 8, 1461–1474 (2018)

Barabási, A. L., and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* 5, 101–113. doi: 10.1038/nrg1272
Brandes, U. (2001). A faster algorithm for betweenness centrality. *J. Math. Sociol.* 25, 163–177. doi: 10.1080/0022250X.2001.9990249

Baranzini SE, Galwey NW, Wang J, Khankhanian P, Lindberg R, Pelletier D, et al. Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. Human molecular genetics. 2009; 18(11):2078–90. pmid:19286671

Bhat, J.A.; Ali, S.; Salgotra, R.K.; Mir, Z.A.; Dutta, S.; Jadon, V.; Tyagi, A.; Mushtaq, M.; Jain, N.; Singh, P.K.; et al. Genomic selection in the era of next generation sequencing for complex traits in plant breeding. *Front. Genet.* **2016**, *7*, 221.

Bernoux, M., Timmers, T., Jauneau, A., Briere, C., de Wit, P. J., 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. The Plant Cell, 20(8), 2252-2264.

Blazier, M.A., Tyree, M.C., Sword Sayer, M.A., KC, D., Hood, W.G., Osbon, B.S. 2018. Gas exchange and productivity in temperate and droughty years of four eastern, elite loblolly pine genotypes grown in the Western Gulf region. Int. J. Agronomy.

Brandes, U. (2001). A faster algorithm for betweenness centrality. *J. Math. Sociol.* 25, 163–177. doi: 10.1080/0022250X.2001.9990249

Bryant, D. M. *et al.* A tissue-mapped axolotl de novo transcriptome enables identification of limb regeneration factors. *Cell reports* 18, 762–776 (2017).

Brown, G. R., Bassoni, D. L., Gill, G. P., Fontana, J. R., Wheeler, N. C., Megraw, R. A., ... & Neale, D. B. (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(4), 1537-1546.

Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics* 30, 2114–2120 (2014).

Bouvet, J. M., Ekomono, C. G. M., Brendel, O., Laclau, J. P., Bouillet, J. P., & Epron, D. (2020). Selecting for water use efficiency, wood chemical traits and biomass with genomic selection in a Eucalyptus breeding program. *Forest Ecology and Management*, *465*, 118092.

Bryant, D. M. *et al.* A tissue-mapped axolotl de novo transcriptome enables identification of limb regeneration factors. *Cell reports* 18, 762–776 (2017).

Buchanan, B. B., Gruissem, W., and Jones, R. L. (2002). *Biochemistry and Molecular Biology of Plants: American Society of Plant Physiologists*. Hoboken: Wiley & Sons.

Caballero, M., Lauer, E., Bennett, J., Zaman, S., McEvoy, S., Acosta, J., ... & Isik, F. (2021). Toward genomic selection in Pinus taeda: Integrating resources to support array design in a complex conifer genome. *Applications in plant sciences*, *9*(6), e11439.

Cánovas, F. M., Cañas, R. A., de la Torre, F. N., Pascual, M. B., Castro-Rodríguez, V., & Avila, C. (2018). Nitrogen metabolism and biomass production in forest trees. *Frontiers in Plant Science*, *9*, 1449.

Cannell, M.G.R. (1989). Physiological basis of wood production: A review. Scandinavian Journal of Forest Research, 4 (1–4), pp. 459-490

Carrasco, A. *et al.* Expression profiling in pinus radiata infected with fusarium circinatum. *Tree Genet. & Genomes* 13, 46(2017).

Calleja-Rodriguez, A., Pan, J., Funda, T. et al. Evaluation of the efficiency of genomic versus pedigree predictions for growth and wood quality traits in Scots pine. BMC Genomics 21, 796 (2020).

Carter, M. C., Kellison, R. C., & Wallinger, R. S. (2015). Forestry in the US South: a history. LSU Press.

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

Chen, Q. F., Xiao, S., & Chye, M. L. (2008). Overexpression of the Arabidopsis 10-kilodalton acyl-coenzyme A-binding protein ACBP6 enhances freezing tolerance. Plant Physiology, 148(1), 304-315.

Chen, Z.-Q.; Baison, J.; Pan, J.; Karlsson, B.; Andersson, B.; Westin, J.; García-Gil1, M.R.; Wu, H.X. 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.* **2018**, *19*, 946.

Chen, Z. Q., Baison, J., Pan, J., Westin, J., Gil, M. R. G., & Wu, H. X. (2019). Increased prediction ability in Norway spruce trials using a marker x environment interaction and non-additive genomic selection model. Journal of Heredity, 110(7), 830-843.

Childs, K. L., Davidson, R. M., and Buell, C. R. (2011). Gene coexpression network analysis as a source of functional annotation for rice genes. PLoS One 6:e22196. doi: 10.1371/journal.pone.0022196

Dai, G. Y., Yin, J., Li, K. E., Chen, D. K., Liu, Z., Bi, F. C., ... & Yao, N. (2020). The Arabidopsis AtGCD3 protein is a glucosylceramidase that preferentially hydrolyzes long-acyl-chain glucosylceramides. Journal of Biological Chemistry, 295(3), 717-728.

Danecek, P. *et al.* The variant call format and vcftools. *Bioinformatics* 27, 2156–2158 (2011).

Danecek, P., McCarthy, S., Li, H. *et al.* bcftools—utilities for variant calling and manipulating vcfs and bcfs (2015).

Daumann, Manuel, et al. "Apoplastic nucleoside accumulation in Arabidopsis leads to reduced photosynthetic performance and increased susceptibility against Botrytis cinerea." *Frontiers in plant science* 6 (2015): 1158.
de Oliveira Junkes, C. F. *et al.* Resin tapping transcriptome in adult slash pine (pinus elliottii var. elliottii). *Ind. Crop. Prod.*139, 111545 (2019).

de Almeida Filho, J.E.; Guimarães, J.F.R.; de Silva, F.F.; de Resende, M.D.V.; Muñoz, P.; Kirst, M.; Resende, M.F.R., Jr. The contribution of dominance to phenotype prediction in a pine breeding and simulated population. Heredity 2016, 117, 33–41

De La Torre, A. R., Puiu, D., Crepeau, M. W., Stevens, K., Salzberg, S. L., Langley, C. H., & Neale, D. B. (2019). Genomic architecture of complex traits in loblolly pine. *New Phytologist*, *221*(4), 1789-1801.

Ding, X., Diao, S., Luan, Q., Wu, H. X., Zhang, Y., & Jiang, J. (2022). A transcriptome-based association study of growth, wood quality, and oleoresin traits in a slash pine breeding population. PLoS genetics, 18(2), e1010017.

Eisen M.B., Spellman P.T., Brown P.O., Botstein D. (1998). Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA **95**: 14863–14868.

Emhart, V. I., Martin, T. A., White, T. L., & Huber, D. A. (2007). Clonal variation in crown structure, absorbed photosynthetically active radiation and growth of loblolly pine and slash pine. *Tree Physiology*, *27*(3), 421-430.

Engeseth, N. J., Pacovsky, R. S., Newman, T., & Ohlrogge, J. B. (1996). Characterization of an Acyl-CoA-Binding Protein fromArabidopsis thaliana. Archives of Biochemistry and Biophysics, 331(1), 55-62.

Francisco FR, Aono AH, da Silva CC, Gonçalves PS, Scaloppi Junior EJ, Le Guen V, Fritsche-Neto R, Souza LM, de Souza AP. Unravelling Rubber Tree Growth by Integrating GWAS and Biological Network-Based Approaches. Front Plant Sci. 2021 Dec 21;12:768589.

Fu J., Sampalo R., Gallardo F., Cánovas F. M., Kirby E. G. (2003). Assembly of a cytosolic pine glutamine synthetase holoenzyme in the leaf of transgenic poplar leads to enhanced vegetative growth in young plants. *Plant Cell Environ.* 26 411–418. 10.1046/j.1365-3040.2003.00972.x

Gaedeke, N., Klein, M., Kolukisaoglu, U., Forestier, C., Müller, A., Ansorge, M., ... & Martinoia, E. (2001). The Arabidopsis thaliana ABC transporter AtMRP5 controls root development and stomata movement. The EMBO journal, 20(8), 1875-1887.

Gao, X., Gray, J., Cohrs, C. W., Cook, R., & Albaugh, T. J. (2021). Longer greenup periods associated with greater wood volume growth in managed pine stands. *Agricultural and Forest Meteorology*, *297*, 108237.
Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907* (2012).

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

González-Camacho, J. M., Ornella, L., Pérez-Rodríguez, P., Gianola, D., Dreisigacker, S., & Crossa, J. (2018). Applications of machine learning methods to genomic selection in breeding wheat for rust resistance. Plant Genome 11: 1–15.

Gonzalez de Andres, E., Camarero, J. J., Blanco, J. A., Imbert, J. B., Lo, Y. H., Sangüesa‐Barreda, G., & Castillo, F. J. (2018). Tree‐to‐tree competition in mixed European beech–Scots pine forests has different impacts on growth and water‐use efficiency depending on site conditions. Journal of Ecology, 106(1), 59-75.

Goralogia, G. S., Redick, T. P., & Strauss, S. H. (2021). Gene editing in tree and clonal crops: Progress and challenges. *In Vitro Cellular & Developmental Biology-Plant*, *57*(4), 683-699

Grabherr, M. G. *et al.* Trinity: reconstructing a full-length transcriptome without a genome from rna-seq data. *Nat. biotechnology* 29, 644 (2011).

Granato, I. S. *et al.* snpready: a tool to assist breeders in genomic analysis. *Mol. Breed.* 38, 1–7 (2018).

Harfouche, A.; Meilan, R.; Kirst, M.; Morgante, M.; Boerjan, W.; Sabatti, M.; Mugnozza, G.S. Accelerating the domestication of forest trees in a changing world. *Trends Plant Sci.* **2012**, *17*, 64–72.

Hayes B, Goddard ME. The distribution of the effects of genes affecting quantitative traits in livestock. Genet Sel Evol. 2001;33(3):1–21.

Heslot, N., Jannink, J. L., & Sorrells, M. E. (2015). Perspectives for genomic selection applications and research in plants. *Crop Science*, *55*(1), 1-12.

Ingvarsson, P. K., & Street, N. R. (2011). Association genetics of complex traits in plants. *New Phytologist*, *189*(4), 909-922.

Isik F, Bartholomé J, Farjat A, Chancerel E, Raffin A, Sanchez L, Plomion C, Bouffier L. Genomic selection in maritime pine. Plant Sci. 2016; 242:108–19.

Isik, F., & McKeand, S. E. (2019). Fourth cycle breeding and testing strategy for Pinus taeda in the NC State University Cooperative Tree Improvement Program. Tree Genetics & Genomes, 15(5), 1-12.

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

Khatiwada, Aastha, et al. "GPA-Tree: statistical approach for functional-annotation-tree-guided prioritization of GWAS results." Bioinformatics 38.4 (2022): 1067-1074.

Klein, M., Perfus‐Barbeoch, L., Frelet, A., Gaedeke, N., Reinhardt, D., Mueller‐Roeber, B., ... & Forestier, C. (2003). The plant multidrug resistance ABC transporter AtMRP5 is involved in guard cell hormonal signalling and water use. The Plant Journal, 33(1), 119-129.

Knapp, S. J. (1998). Marker‐assisted selection as a strategy for increasing the probability of selecting superior genotypes. Crop Science, 38(5), 1164-1174.

Korte, A., and Farlow, A. (2013). The advantages and limitations of trait analysis with GWAS: a review. Plant Methods 9, 1–9. doi: 10.1186/1746-4811-9-29

Kumar, V., & Minz, S. (2014). Feature selection: a literature review. *SmartCR*, *4*(3), 211-229.

Lai, M., Dong, L., Yi, M., Sun, S., Zhang, Y., Fu, L., ... & 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(11), 409.

Lamara M, Raherison E, Lenz P, Beaulieu J, Bousquet J, MacKay J. Genetic architecture of wood properties based on association analysis and co-expression networks in white spruce. New Phytologist. 2016; 210(1):240–55. pmid:26619072

Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. BMC bioinformatics, 9(1), 1-13.

Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with bowtie 2. *Nat. methods* 9, 357 (2012).

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. *Forest Science*, *67*(1), 60-71.

Lee T., Lee I. (2018). araGWAB: Network-based boosting of genome-wide association studies in *Arabidopsis thaliana*. Sci. Rep. **8**: 2925.

Lebedev, V. G., Lebedeva, T. N., Chernodubov, A. I., & Shestibratov, K. A. (2020). Genomic selection for forest tree improvement: Methods, achievements and perspectives. Forests, 11(11), 1190.

Lee, E. K., Kwon, M., Ko, J. H., Yi, H., Hwang, M. G., Chang, S., & Cho, M. H. (2004). Binding of sulfonylurea by AtMRP5, an Arabidopsis multidrug resistance-related protein that functions in salt tolerance. Plant Physiology, 134(1), 528-538.

Lenz, P.R.N.; Beaulieu, J.; Mansfield, S.D.; Clément, S.; Desponts, M.; Bousquet, J. 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.* **2017**, *18*, 335.

Lenz, P.R.N.; Nadeau, S.; Mottet, M.-J.; Perron, M.; Isabel, N.; Beaulieu, J.; Bousquet, J. Multi-trait genomic selection for weevil resistance, growth, and wood quality in Norway spruce. *Evol. Appl.* **2020**, *13*, 76–94.

Li, B., McKeand, S. E., and 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, 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.

Liu, X., Huang, M., Fan, B., Buckler, E. S., & Zhang, Z. (2016). Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. PLoS genetics, 12(2), e1005767.

Lu, M., Krutovsky, K. V., Nelson, C. D., Koralewski, T. E., Byram, T. D., & Loopstra, C. A. (2016). Exome genotyping, linkage disequilibrium and population structure in loblolly pine (Pinus taeda L.). BMC genomics, 17(1), 1-11.

Lu, M., Krutovsky, K. V., Nelson, C. D., West, J. B., Reilly, N. A., & Loopstra, C. A. (2017). Association genetics of growth and adaptive traits in loblolly pine (Pinus taeda L.) using whole-exome-discovered polymorphisms. Tree Genetics & Genomes, 13(3), 57.

Lu, M., Seeve, C. M., Loopstra, C. A., & Krutovsky, K. V. (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 genetics*, *19*(1), 1-13.

Luo, Y. ; El‐Madany, T. ; Ma, X.; Nair, R. ; Jung, M. ; Weber, U. ; Filippa, G. ; Bucher, S.F. ; Moreno, G. ; Cremonese, E. ; Carrara, A.; Gonzalez‐Cascon, R. ; Escudero, Y.C. ; Galvagno, M. ; Pacheco‐Labrador, J. ; Martín, M.P.; Perez‐Priego, O.; Reichstein, M. ; Richardson, A.D.; Migliavacca M.. (2020). Nutrients and water availability

constrain the seasonality of vegetation activity in a Mediterranean ecosystem. Global Change Biology (2020), 10.1111/gcb.15138

Lush, J. L. (1936). Genetics and animal breeding∗. *J. Hered.* 27, 201–203. doi: 10.1093/oxfordjournals.jhered.a104206

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., et al. (2009). Finding the missing heritability of complex diseases. Nature 461, 747–753. doi: 10.1038/nature08494

Mao, J. *et al.* 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 (2019).

Martinoia, E., Klein, M., Geisler, M., Bovet, L., Forestier, C., Kolukisaoglu, Ü., ... & Schulz, B. (2002). Multifunctionality of plant ABC transporters–more than just detoxifiers. Planta, 214(3), 345-355.

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

Matallana-Ramirez, L. P., Whetten, R. W., Sanchez, G. M., & Payn, K. G. (2021). Breeding for climate change resilience: a case study of loblolly pine (Pinus taeda L.) in North America. Frontiers in plant science, 12, 790.

Matsunaga, W., Ohama, N., Tanabe, N., Masuta, Y., Masuda, S., Mitani, N., et al. (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. doi: 10.3389/fpls.2015.00048

McKeand, S. E., Jokela, E. J., Huber, D. A., Byram, T. D., Allen, H. L., Li, B., & Mullin, T. J. (2006). Performance of improved genotypes of loblolly pine across different soils, climates, and silvicultural inputs. Forest Ecology and Management, 227(1-2), 178-184.

Meuwissen T H, Hayes B J, Goddard M E, 2001 Prediction of total genetic value using genome-wide dense marker maps. Genetics 157: 1819–1829.

Moreau, L., Lacoudre, F., Charcosset, A., & Gallais, A. (1997). More on the efficiency of marker-assisted selection. Theoretical and applied genetics, 95(8), 1181-1189.

Mutwil, M., Usadel, B., Schu tte, M., Loraine, A., Ebenho h, O., & Persson, S. (2010). Assembly of an interactive correlation network for the Arabidopsis genome using a novel heuristic clustering algorithm. Plant physiology, 152(1), 29-43.

Neale DB, Kremer A. Forest tree genomics: growing resources and applications. Nature Reviews Genetics 12: 111–121 (2011).

Neves, H. H., Carvalheiro, R., & Queiroz, S. A. (2012). A comparison of statistical methods for genomic selection in a mice population. *BMC genetics*, *13*(1), 1-17.

Neves, L. G., Davis, J. M., Barbazuk, W. B. & Kirst, M. Whole-exome targeted sequencing of the uncharacterized pine genome. *The Plant J.* 75, 146–156 (2013).

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, Genetics, 8(9), 2889-2899.

O'Brien, I. E., Smith, D. R., Gardner, R. C. & Murray, B. G. Flow cytometric determination of genome size in pinus. *Plant Sci.* 115, 91–99 (1996).

Oliver S. Proteomics: Guilt-by-association goes global. Nature. 2000;403(6770):601–3. pmid:10688178

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. *The Plant Journal*, *24*(3), 327-334.

Patro, R., Duggal, G. & Kingsford, C. Salmon: accurate, versatile and ultrafast quantification from rna-seq data using lightweight-alignment. *Biorxiv* 021592 (2015).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.

Pérez, P., & de Los Campos, G. (2014). Genome-wide regression and prediction with the BGLR statistical package. Genetics, 198(2), 483-495.

Perera, D. *et al.* Exploring the loblolly pine (pinus taeda l.) genome by bac sequencing and cot analysis. *Gene* 663, 165–177 (2018)

Peterson, B. G. *et al.* Package 'performanceanalytics'. *R Team Coop.* (2018).

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. *Frontiers in genetics*, *12*, 137.

Plomion, C. et al. Pines. In Forest Trees, 29–92 (Springer, 2007).

Polle, A., Chen, S. L., Eckert, C., & Harfouche, A. (2019). Engineering drought resistance in forest trees. Frontiers in Plant Science, 9, 1875.

Poovaiah C, Phillips L, Geddes B, Reeves C, Sorieul M, Thorlby G (2020) Genome editing with CRISPR/Cas9 in Pinus Radiata (D. Don).

Resende, R. T., Resende, M. D. V., Silva, F. F., Azevedo, C. F., Takahashi, E. K., Silva-Junior, O. B., & Grattapaglia, D. (2017). Assessing the expected response to genomic selection of individuals and families in Eucalyptus breeding with an additive-dominant model. *Heredity*, *119*(4), 245-255.

Resende MFR, Muñoz PR, Resende MDV, 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, Muñoz 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, J. E. (2019). "A practical guide to genetic gain," in *Advances in Agronomy*, ed. D. L. Sparks (Cambridge, MA: Academic Press Inc.), 217–249. doi: 10.1016/bs.agron.2019.05.001

Sampson, D.A. ; Johnsen, K.H.; Ludovici, K.H. ; Albaugh, T.J.; Maier C.A. (2001). Stand-scale correspondence in empirical and simulated labile carbohydrates in loblolly pine. Forest Science, 47 (1) (2001), pp. 60-68

Santantonio, N., Atanda, S. A., Beyene, Y., Varshney, R. K., Olsen, M., Jones, E., ... & Robbins, K. R. (2020). Strategies for effective use of genomic information in crop breeding programs serving Africa and South Asia. *Frontiers in Plant Science*, *11*, 353.

Santini, F., Kefauver, S. C., Araus, J. L., Resco de Dios, V., Martín García, S., Grivet, D., & Voltas, J. (2021). Bridging the genotype–phenotype gap for a Mediterranean pine by semi‐automatic crown identification and multispectral imagery. New Phytologist, 229(1), 245-258.

Schmidt, P., Hartung, J., Bennewitz, J., & Piepho, H. P. (2019). Heritability in plant breeding on a genotype-difference basis. Genetics, 212(4), 991-1008.

Seppey, M., Manni, M. & Zdobnov, E. M. Busco: assessing genome assembly and annotation completeness. In *Gene Prediction*, 227–245 (Springer, 2019).

Shalizi, M. N., & Isik, F. (2019). Genetic parameter estimates and GxE interaction in a large cloned population of Pinus taeda L. Tree Genetics & Genomes, 15(3), 1-13.

Shalizi, M. N., Payn, K. G., & 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 Genetics & Genomes, 18(1), 1-17.

Shalizi, M. N., Gezan, S. A., McKeand, S. E., Sherrill, J. R., Cumbie, W. P., Whetten, R. W., & Isik, F. (2020). Correspondence between breeding values of the same Pinus taeda L. Genotypes from clonal trials and half-sib seedling progeny trials. *Forest Science*, *66*(5), 600-611

Sherrill, J. R., Bullock, B. P., Mullin, T. J., McKeand, S. E., & Purnell, R. C. (2011). Total and merchantable stem volume equations for midrotation loblolly pine (Pinus taeda L.). Southern Journal of Applied Forestry, 35(3), 105-108.

Schaefer, R. J., Michno, J. M., Jeffers, J., Hoekenga, O., Dilkes, B., Baxter, I., et al. (2018). Integrating coexpression networks with GWAS to prioritize causal genes in maize. Plant Cell 30, 2922–2942. doi: 10.1105/tpc.18.00299

Singh, M., Kaur, K., Sharma, A. et al. 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 (2021).

Simkin, A. J. (2019). Genetic engineering for global food security: photosynthesis and biofortification. Plants, 8(12), 586.

Soltis, P. S. & Soltis, D. E. A conifer genome spruces up plant phylogenomics. *Genome biology* 14, 1–5 (2013).

Stasolla C, Katahira R, Thorpe TA, Ashihara H. Purine and pyrimidine nucleotide metabolism in higher plants. *J Plant Physiol.* 2003;**160**(11):1271–1295. doi: 10.1078/0176-1617-01169

Supek, F., Bošnjak, M., Škunca, N. & Šmuc, T. Revigo summarizes and visualizes long lists of gene ontology terms. *PloS one* 6, e21800 (2011).

Sul, J. H., Martin, L. S., & Eskin, E. (2018). Population structure in genetic studies: Confounding factors and mixed models. PLoS genetics, 14(12), e1007309.

Sun, C., Wu, X.L., Weigel, K.A., Rosa, G.J.M., Bauck, S., Woodward, B.W., Schnabel, R.D., Taylor, J.F., and Gianola, D.. 2012. An ensemble-based approach to imputation of moderate density genotypes for genomic selection with application to Angus cattle. *Genet. Res.* **94**:133–150.

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

Team, R. C. *et al. R: A language and environment for statistical computing* (2013).

Telfer, E., Graham, N., Macdonald, L., Li, Y., Klápště, J., Resende Jr, M., ... & Wilcox, P. (2019). A high-density exome capture genotype-by-sequencing panel for forestry breeding in Pinus radiata. PloS one, 14(9), e0222640.

Tenesa, A., Wright, A. F., Knott, S. A., Carothers, A. D., Hayward, C., Angius, A., ... & Visscher, P. M. (2004). Extent of linkage disequilibrium in a Sardinian sub-isolate: sampling and methodological considerations. Human Molecular Genetics, 13(1), 25-33.

Thistlethwaite, F. R., Ratcliffe, B., Klápště, J., Porth, I., Chen, C., Stoehr, M. U., & El-Kassaby, Y. A. (2019). Genomic selection of juvenile height across a single-generational gap in Douglas-fir. *Heredity*, *122*(6), 848-863.

Tong, H., & Nikoloski, Z. (2021). Machine learning approaches for crop improvement: Leveraging phenotypic and genotypic big data. *Journal of Plant Physiology*, *257*, 153354.

Tran, K. N., and Choi, J. I. (2020). Comparative transcriptome analysis of high-growth and wild-type strains of Pyropia yezoensis. Acta Bot. Croat. 79, 148–156. doi: 10.37427/botcro-2020-020

Traylor-Knowles, N., Rose, N. H., Sheets, E. A., and Palumbi, S. R. (2017). Early transcriptional responses during heat stress in the coral Acropora hyacinthus. Biol. Bull. 232, 91–100. doi: 10.1086/692717

Vieira, J., Moura, M., Nabais, C., Freitas, H., & Campelo, F. (2019). Seasonal adjustment of primary and secondary growth in maritime pine under simulated climatic changes. Annals of Forest Science, 76(3), 1-12.

Vose, J.M. and Allen, H.L. (1988). **Leaf area, stemwood growth, and nutrition relationships in loblolly pine.** Forest Science, 34 (3), pp. 547-563

Wang, X., Xu, Y., Hu, Z., & Xu, C. (2018). Genomic selection methods for crop improvement: Current status and prospects. *The Crop Journal*, *6*(4), 330-340.

Watts, D. J., and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature* 393, 440–442. doi: 10.1038/30918

Wegrzyn, J. L. *et al.* Unique features of the loblolly pine (pinus taeda l.) megagenome revealed through sequence annotation. *Genetics* 196, 891–909 (2014).

Wen Z., Tan R., Zhang S., Collins P.J., Yuan J., Du W., Gu C., Ou S., Song Q., An Y.-Q.C., Boyse J.F., Chilvers M.I., et al. (2018). Integrating GWAS and gene expression data for functional characterization of resistance to white mold in soya bean. Plant Biotechnol. J. **16**: 1825–1835.

Wickham, H. *ggplot2: elegant graphics for data analysis* (springer, 2016).

White, T.L.; Adams, W.T.; Neale, D.B. Forest Genetics; CABI Publishing CAB International: Cambridge, UK, 2007.

Wu, H. X., Eldridge, K. G., Matheson, A. C., Powell, M. B., McRae, T. A., Butcher, T. B., & Johnson, I. G. (2007). Achievements in forest tree improvement in Australia and New Zealand 8. Successful introduction and breeding of radiata pine in Australia. Australian forestry, 70(4), 215-225.

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

Yin, L. *et al.* rmvp: A memory-efficient, visualization-enhanced, and parallel-accelerated tool for genome-wide association study. *bioRxiv* (2020).

Xiang, B., Li, B., & Isik, F. (2003). Time trend of genetic parameters in growth traits of Pinus taeda L. *Silvae genetica*, *52*(3-4), 114-120.

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

Zhang, H., Zhang, Y., Zhang, D., Dong, L., Liu, K., Wang, Y., ... & Zhao, X. (2020). Progeny performance and selection of superior trees within families in Larix olgensis. Euphytica, 216(4), 1-10.

Zhang, J. *et al.* Allele-defined genome of the autopolyploid sugarcane *Saccharum spontaneum* L. *Nat. Genet.* **50**, 1565 (2018).

Zhang X., Misra A., Nargund S., Coleman G. D., Sriram G. (2018). Concurrent isotope-assisted metabolic flux analysis and transcriptome profiling reveal responses of poplar cells to altered nitrogen and carbon supply. *Plant J.* 3 472–488. 10.1111/tpj.13792

Zimin, A. V. *et al.* An improved assembly of the loblolly pine mega-genome using long-read single-molecule sequencing.*Gigascience* 6, giw016 (2017).

Zobel, B., and Talbert, J. (1984). Applied Forest Tree Improvement. Caldwell NJ: The Blackburn Press.