UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE BIOLOGIA

STEPHANIE KARENINA BAJAY

ABORDAGENS MULTI-ÔMICAS PARA O MELHORAMENTO FLORESTAL

MULTI-OMIC APPROACHES FOR FOREST BREEDING

Campinas
2022

STEPHANIE KARENINA BAJAY

ABORDAGENS MULTI-ÔMICAS PARA O MELHORAMENTO FLORESTAL

MULTI-OMIC APPROACHES FOR FOREST BREEDING

Tese apresentada ao Instituto de Biologia da Universidade Estadual de Campinas como parte dos requisites exigidos para a obtenção de título de Doutora em Genética e Biologia Molecular na área de Genética Vegetal e Melhoramento

Thesis presented to the Institute of Biology of the University of Campinas in partial fulfillment of the requirements for the degree of Doctor in Genetics and Molecular Biology in the area of Plant Geneticsand Genetic Breeding

Orientadora: Anete Pereira de Souza

ESTE ARQUIVO DIGITAL CORRESPONDE À VERSÃO FINAL DA TESE DEFENDIDA PELA ALUNA STEPHANIE KARENINA BAJAY, E ORIENTADA PELA PROFESSORA DOUTORA ANETE PEREIRA DE SOUZA.

Campinas

2022

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Biologia
Mara Janaina de Oliveira - CRB 8/6972

Bajay, Stephanie Karenina, 1989-
B167a Abordagens multi-ômicas para o melhoramento florestal / Stephanie Karenina Bajay. – Campinas, SP : [s.n.], 2022.

Orientador: Anete Pereira de Souza.
Tese (doutorado) – Universidade Estadual de Campinas, Instituto de Biologia.

1. Genômica. 2. Estudo de associação genômica ampla. 3. Aprendizado de máquina. 4. Plantas - Melhoramento genético. 5. Florestas. 6. Herança multifatorial. I. Souza, Anete Pereira de, 1962-. II. Universidade Estadual de Campinas. Instituto de Biologia. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: Multi-omic approaches for forest breeding
Palavras-chave em inglês:
Genomics
Genome-wide association study
Machine learning
Plant breeding
Forests
Multifactorial inheritance
Área de concentração: Genética Vegetal e Melhoramento
Titulação: Doutora em Genética e Biologia Molecular
Banca examinadora:
Anete Pereira de Souza [Orientador]
Maria Imaculada Zucchi
Prianda Rios Laborda
Lorena Guimarães Batista
Luís Felipe Ventorim Ferrão
Data de defesa: 14-06-2022
Programa de Pós-Graduação: Genética e Biologia Molecular

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0002-7018-6491
- Currículo Lattes do autor: http://lattes.cnpq.br/3120902594300785

COMISSÃO EXAMINADORA

Profa. Dra. Anete Pereira de Souza (orientadora)

Dra. Maria Imaculada Zucchi

Dra. Prianda Rios Laborda

Dr. Luís Felipe Ventorim Ferrão

Dra. Lorena Guimarães Batista

A Ata da defesa com as respectivas assinaturas dos membros encontra-se no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria do Programa da Unidade.

À minha família, minha fonte de amor e inspiração

Especialmente ao meu pai, Sergio Bajay, meu melhor amigo e grande mentor, e minha mãe, Terezinha Scaramelli, meu grande suporte emocional e companheira
Dedico

AGRADECIMENTOS

À Professora Dra. Anete Pereira de Souza, minha orientadora, por todos os ensinamentos, suporte profissional e emocional. Agradeço pela oportunidade de crescer junto à um grupo que foi fundamental para minha formação profissional e pessoal.
Agradeço pela confiança depositada em mim e por sempre me incentivar a ir atrás de novos desafios.

À Universidade Estadual de Campinas e ao programa de pós-graduação em Genética e Biologia Molecular.

À CAPES, pois o presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001 e conceção da bolsa de doutorado sanduíche (PE-BOLSAS 7349056). À Fundação de Amparo a Pesquisa do Estado de São Paulo (FAPESP), pela bolsa doutorado, **Processo 2019/13452-3,** que possibilitaram a realização deste trabalho.

Aos membros do Exame de Qualificação e Banca de Defesa de Tese.

Ao meu supervisor do doutorado sanduiche na Universidade da Florida, Prof. Matias Kirst, e todo seu grupo Forest Genomic Lab, especialmente ao coordenador de laboratório Christopher Dervinis e o Rodrigo Santos, pela paciência e tantos ensinamentos. Agradeço as queridas amizades que fiz durante esse período, no laboratório: Wendell, Nathaniel, Hank e Daniel; e as especiais amizades de Gainesville: Bia, Bianca e Govinda.

Ao Prof. Dr. Ian Mackay e Prof. Dr. Wayne Power por me receberem tão bem no Scotland Rural College, e a todos os colegas do SRUC que tornaram um país frio, um dos lugares mais calorosos por onde passei. Agradeço em especial, toda a credibilidade e confiança com que fui recebida. Certamente esta experiência contribuiu muito para que eu pudesse fortalecer a minha autoconfiança.

Ao querido Alexandre Aono pela paciência, disponibilidade e inestimável ajuda com a análise dos dados e escrita do artigo. Ao Felipe Francisco pela dedicação às análises e ajuda com a escrita do artigo.

Aos técnicos do laboratório LAGM: Aline, Danilo, Carlão e Pati, obrigada por sempre auxiliarem todos do laboratório.

Aos colegas do LAGM/CBMEG, obrigada pela convivência, aprendizados e pela troca de experiências. Agradeço pelas valiosas amizades que pude construir, Lívia, Melina, Fernanda e Carla. Em especial, a Rebecca, obrigada por ser essa amigona que esteve sempre presente em todos os momentos da minha vida, como uma grande ouvinte e concelheira!

Agradeço aos meus amigos de longa data, Gabriela, And Yara e em especial a Bruna, por todo apoio, carinho e compreensão em todas as fases dessa longa jornada. Os sucessos das minhas experiências no exterior, com certeza, se devem muito ao suporte das visitas da Bruna!

Ao Afonso, pelo carinho, paciência e companheirismo!

Aos meus irmãos Arpad e Miklos, e meus sobrinhos Vicente e Jaime, que são grandes amores da minha vida. Obrigada, Miklos, por ter me iniciado neste universo da genética, e continuar me mostrando como podemos acrescentar à sociedade com nossas pesquisas e engajamentos. Ao tango, por me mostrar o que é um amor incondicional e por trazer tanta alegria aos meus dias!

Ao meu pai, meu grande mentor e inspiração, pelo amor e apoio em todas minhas decisões, desde o âmbito acadêmico, ao pessoal.

À minha mãe, minha grande companheira de vida, meu porto seguro e inspiração. Obrigada pela paciência e amor.

RESUMO

A seleção de genótipos superiores, através da aplicação de seleção genômica (SG) e estudos de associação genômica ampla (GWAS), melhoraram drasticamente a velocidade e a escala da genética aplicada no melhoramento florestal. No entanto, a escolha da melhor metodologia a ser adotada varia de acordo com os objetivos a serem alcançados em cada etapa de desenvolvimento do programa de melhoramento. A SG é uma excelente estratégia a ser adotada nas fases iniciais e intermediária de programas de melhoramento, para a seleção de genótipos superiores. Nos últimos anos, os esforços para o melhoramento de espécies perenes se voltaram para a SG, pois ela deve permitir que a maioria das fontes de variação para características complexas sejam rastreadas. Neste contexto de SG, também vem se destacando a incorporação de métodos avançados de aprendizado de máquina, devido ao fato destes algoritmos permitirem treinamento usando representação de dados mais complexos e por não exigirem pressuposições quanto ao modelo. Dentre os vários métodos de aprendizado existentes, a seleção de atributos foi o escolhido para estudo na presente tese. A seleção de atributos permite a redução da densidade de marcadores e construção de modelos simples e abrangentes de predição, evitando a atribuição de efeitos não genéticos aos marcadores e aumentando o poder preditivo dos fenótipos de interesse. Complementarmente à SG, com o objetivo de ampliar o entendimento da arquitetura e base genética dos fenótipos estudados, também foi adotada a estratégia de estudo quantitativo da interação genótipo-fenótipo através do GWAS. Para a interpretação de como os genes descobertos por GWAS e SG influenciam nas características analisadas, foram adotadas estratégias de anotação de vias genéticas e ontologias identificadas com transcriptomas. Redes de co-expressão gênica foram construídas, com intuito de desenvolver uma compreensão global da expressão gênica e função biológica possivelmente correlacionada com os genes candidatos a modulação dos fenótipos de interesse. A combinação de diferentes estatísticas e análises genômicas, como SG, aprendizado de máquina, GWAS e redes de co-expressão gênica, se torna uma estratégia promissora para que seja possível lidar efetivamente com o melhoramento de caracteres complexos. Desta forma, o objetivo central da tese que se apresenta foi integrar múltiplas análises ômicas: SG, GWAS, ML e rede de co-expressão gênica para seleção de genótipos superiores em espécies arbóreas.

Palavras-chave: Seleção Genômica, Associação genômica ampla (GWAS), Aprendizado de Máquina, Melhoramento Florestal, Características Complexas

ABSTRACT

The selection of superior genotypes, through the application of genomic selection (GS) and genome-wide association studies (GWAS), has dramatically improved the speed and scale of genetics applied to forest breeding. However, the choice of the best methodology to be adopted varies according to the aims to be achieved in each stage of development of the breeding program. GS is an excellent strategy to be adopted in the initial and intermediate steps of breeding programs, for the selection of superior genotypes. In recent years, efforts to improve perennial species have turned to GS, as it should allow tracking of most sources of variation for complex traits. In this context of GS, the incorporation of advanced methods of machine learning has also been highlighted, due to the fact that these algorithms allow training using more complex data representation and because they do not require assumptions about the model. Among the various existing learning algorithms, we applied in the present study, the feature selection, which aims to reduce the density of markers and build simple and comprehensive prediction models, avoiding the attribution of non-genetic effects to the markers and increasing the predictive power of the phenotypes. In addition to GS, with the objective of expanding the understanding of the architecture and genetic basis of the phenotypes studied, the strategy with GWAS was also adopted. For the interpretation of how the genes discovered by GWAS and GS influence the analyzed traits, strategies of annotation of genetic pathways and ontologies identified with transcriptomes were adopted. Gene co-expression networks were constructed in order to develop a global understanding of gene expression and biological function possibly correlated with candidate genes for modulating the phenotypes of interest. The combination of different statistics and genomic analyses, such as SG, machine learning, GWAS and gene co-expression networks, becomes a promising strategy to effectively deal with the improvement of complex traits. Thus, the main objective of this thesis was to integrate multiple omic analyses: SG, GWAS, ML and gene co-expression networks for the selection of superior genotypes in forest trees.

Keywords: Genomic Selection, Genome Wide Association Studies (GWAS), Machine Learning, Forestry Breeding, Complex Traits

SUMÁRIO

2. PREFÁCIO

A presente tese está organizada no formato de capítulos. As diferentes etapas desenvolvidas neste trabalho são apresentadas no formato de dois artigos científicos e um capítulos de livro.

Inicialmente, serão apresentadas uma introdução geral e uma revisão bibliográfica, que compreende os seguintes tópicos: (1) Melhoramento de pinus e (2) Abordagem integrada de análises de Seleção Genômica, Aprendizado de Máquina, GWAS e redes de co-expressão gênica no melhoramento de espécies arbóreas.

Em seguida, é apresentado o capítulo I, em formato de artigo, dedicado a abordagem das estratégias multi-ômicas de melhoramento florestal em pinus, proposta nos principais objetivos da presente tese: Integrar múltiplas análises ômicas: SG, GWAS, ML e rede de co-expressão gênica para seleção de genótipos superiores em pinus.

No capítulo II é apresentado o artigo sobre resposta a frio em seringueira, onde foi aplicado a análise de redes de co-expressão gênica para descoberta das principais vias metabólicas envolvidas na expressão gênica dessas árvores sobre estresse de baixas temperaturas.

O capítulo III apresenta dois tópicos escritos pela aluna, que irão compor dois capítulos de um livro sobre transcriptome da editora Elsevier, sendo eles: (1) “Analysis of transcript expression and gene regulation (research involving fungi that act mainly on the degradation of biomass)” e (2) “Transcriptomics in the context of abiotic or biotic stress / agricultural sciences (mainly cultivated plants and models with wild forest trees)”.

Ambos os capítulos, são parte do livro: TRANSCRIPTOME PROFILING: PROGRESS AND PROSPECTS” by M.A. Ali and J. Lee, em processo de edição. Os dois tópicos escritos pela aluna e compõem uma revisão bibliográfica sobre estudos utilizando transcriptomas em plantas. Os tópicos escritos pela aluna são:

- Expression analysis/differential expression analysis: Programs, statistical analysis, validation of differential expression analysis
- Transcriptome and breeding in Orphan crops

Os resultados obtidos nos capítulos I e II são apresentados no Capítulo IV- “Principais resultados”, seguido das conclusões e perspectivas para o prosseguimento dos estudos genéticos e genômicos em espécies florestais cultivadas.

No Anexo, encontram-se a descrição e principais os resultados obtidos nos dois períodos de em estágio no exterior, (1) Universidade da Flórida – uma ano e (2) SRUC – 3 meses, realizados durante o período de doutorado.

3.1 INTRODUÇÃO GERAL

Loblolly pine, a espécie de pinus com um dos maiores genomas conhecidos, e seringueira, a árvore com maior produção de látex para borracha natural, são importantes espécies florestais, com relevância econômica a nível mundial. *L. pine,* além da sua grande importância para os EUA, também tem papel significativo para a economia do Brasil, pois é a espécie de pinus mais cultivada no país e oferece diversos produtos fundamentais para a sociedade, como madeira, celulose, borracha, algodão etc.

Assim como em seringueira, devido ao longo ciclo de vida de pinus, os programas de melhoramento dependem de novas tecnologias para aumentar a produtividade das florestas plantadas, selecionar características de interesse no setor e reduzir impactos ambientais. Portanto, as informações genômicas nestas árvores florestais são necessárias para preparar o caminho para o melhoramento das próximas gerações.

A aplicação da seleção genômica (SG), desde seu desenvolvimento há quase duas décadas (Meuwissen et al. 2001), melhorou drasticamente a velocidade e a escala da genética aplicada e da pesquisa de melhoramento (Daetwyler et al. 2013). Modelos de SG são treinados em uma população reprodutora com um subconjunto de dados genotípicos e fenotípicos, permitindo aos criadores prever valores genéticos para indivíduos genotipados com fenótipos desconhecidos (Meuwissen et al. 2001).

O desenvolvimento de tais modelos de previsão permite acelerar o melhoramento e compreender melhor o controle genético da resistência de pragas, bem como características de interesse econômico como crescimento e qualidade da madeira (Burdon and Wilcox, 2011).

Por causa da proliferação de tecnologias genômicas e um declínio constante nos custos de genotipagem (Heffner et al. 2010; Thomson 2014), há um forte impulso para os melhoristas aproveitarem ao máximo a disponibilidade de extensos conjuntos de marcadores moleculares para ganhos de seleção na agricultura e silvicultura (Grattapaglia e Resende, 2011; Iwata et al., 2011; Crossa et al., 2011; Hayes et al., 2013; De Los Campos et al., 2013; Souza et al., 2019).

Os esforços para o melhoramento de espécies perenes se voltaram para a SG, pois a disponibilidade do grande conjunto de SNPs, que essa estratégia requer, deve permitir que a maioria das fontes de variação para características complexas sejam rastreadas. Assim, estimar os

efeitos de milhares de marcadores permite a inclusão de informações de todos os marcadores genotipados, em vez de basear conclusões apenas em loci que alcançam significância em todo o genoma. Esta estratégia, portanto, permite capturar mais da variância genética aditiva do que na seleção tradicional assistida por marcadores com base em loci de características quantitativas (QTL) detectados em Genome-Wide Association Studies (GWAS) (Heffner et al. 2009, Beaulieu et al., 2011; Porth et al., 2013, Turner-Hissong et al., 2020).

Os melhoristas têm sido capazes de implementar o SG aproveitando ao máximo os avanços nas tecnologias de genotipagem de alto rendimento, acumulando milhares de alelos favoráveis para desenvolver linhas de cultivo resilientes, com alto potencial de rendimento. Um grande número de marcadores de DNA que cobrem todo o genoma deve ser considerado para que o desequilíbrio de ligação (LD) entre os marcadores e a maioria das fontes de variação para fenótipos complexos valiosos possa ser rastreado. Portanto, o sucesso da SG depende da extensão da LD entre os marcadores de DNA e os loci que afetam as características complexas para prever o mérito genético dos indivíduos.

Esta não dependência de loci de grande efeito dá uma vantagem estatística à SG (Meuwissen et al. 2001), funcionando bem para características complexas (Goddard et al. 2009; de Los Campos et al. 2013), como volume para espécies florestais, altura das plantas, qualidade e crescimento da madeira. A aplicação da SG para prever o crescimento e a qualidade da madeira mostraram resultados bem-sucedidos em diferentes espécies de árvores, por exemplo, seringueira (Souza et al., 2019; Cros et al., 2019), pinus (Resende, et al., 2012; Zapata – Valenzuela et al., 2013; Isik et al., 2016), eucalipto (Mphahlele et al., 2020)) e abeto (Beaulieu, et al., 2014; Beaulieu, Doerksen, et al., 2014; Ratcliffe et al., 2015; Lenz et al., 2017; Chen et al., 2018).

Complementarmente à associação genômica e à predição genômica, notou-se a necessidade de associar os genes e vias biossintéticas envolvidas com as características complexas. Prever o desempenho e a estabilidade do genótipo é ainda mais desafiador devido às condições ambientais futuras incertas (de Los Campos et al., 2020).

As plantas precisam desenvolver uma série de mecanismos de ajuste para lidar com muitos tipos diferentes de estresses abióticos (temperatura, seca, sal e metais pesados) e bióticos (interação das plantas com insetos e patógenos). Esses estresses podem impedir que a planta atinja o

crescimento ideal e gerar um impacto significativo na produção (Bai et al., 2018; Chen et al., 2020; Garbeva et al., 2020).

Para tolerar esses diferentes estresses, as plantas desenvolveram ampla regulação de vários genes para mediar processos fisiológicos e bioquímicos. A compreensão do papel dos genes envolvidos nesses mecanismos é necessária para o desenvolvimento de ferramentas biotecnológicas que melhorem características agronômicas desejáveis, como crescimento e produtividade das plantas (Chen et al., 2020).

A inclusão de informações de mapeamento genético, estudos GWAS, anotação de vias genéticas e ontologias identificadas em transcriptomas e experimentos "in silico", além de abertura de cromatina e características evolutivas (Ramstein et al. 2020, Lozano et al. 2017), possibilita explorar com mais detalhes a base genética de diferentes fenótipos (Li et al. 2012).

Essas informações adicionais têm grande potencial para aprimorar os estudos de SG e GWAS. Atualmente, há uma nova tendência de incorporar a partição genômica em métodos de predição. O particionamento genômico é capaz de melhorar a SG por ser um framework para testar hipóteses guiadas, por meio da relação entre a influência das características genômicas na variância explicada para características complexas. A incorporação de partições genômicas, como múltiplos efeitos aleatórios na avaliação de características complexas, permite avaliar o tamanho dos efeitos aleatórios entre diferentes categorias de variantes (Turner-Hissong et al., 2020).

Diferentes camadas de dados ômicos podem ser integradas por meio da genética de sistemas, ajudando a dissecar e entender melhor a arquitetura de características complexas. O crescente interesse pela genética de sistemas se deve à capacidade de identificar polimorfismos que modulam a expressão gênica e, a partir do mapeamento dessas variantes genéticas, podemos decifrar redes biológicas e caminhos envolvidos nessas características complexas (Balmant et al., 2020). Nesse contexto, podemos vincular os loci detectados por GWAS a informações de redes de regulação gênica para obter uma compreensão mais profunda dos mecanismos subjacentes às características complexas. Apesar do potencial promissor, essa integração entre GS, GWAS e dados genéticos, como expressão gênica e dados de anotação, ainda não foi feita para a maioria das espécies arbóreas.

Uma forma atual e promissora para integração de dados genômicos de seleção e associação genômica, com foco na capacidade preditiva e conexão lógica por trás dos padrões biológicos

identificados em marcadores moleculares e dados fenotípicos, é através do uso de modelos de aprendizado de máquina (Machine Learning - ML) (Azodi et al., 2020).

O uso de ML vem ganhando espaço ao se tratar do desafio de analisar grande quantidade de dados e em situações em que o número de parâmetros é muito maior que o número de observações. O aprendizado de máquina tem sido amplamente aplicado em dados de processamento de imagem, reconhecimento de áudio e mineração de texto, e os algoritmos de aprendizado são livres de especificação de modelo e podem capturar informações imprevistas de conjuntos de dados de alto rendimento (Namin et al., 2018). Isso é atraente em estudos de associação genómica (Tong and Nikoloski, 2021). Alguns algoritmos avançados de aprendizado de máquina, como métodos de conjunto e algoritmos de aprendizado profundo, podem ajudar na previsão habilitada para genoma (Aono et al., 2020).

3.2 REVISÃO BIBLIOGRÁFICA

3.2.1 Melhoramento de pinus

Em um contexto em que as mudanças climáticas estão ameaçando florestas e afetando o rendimento atual e futuro de espécies agrícolas, se faz urgente a aplicação de esforços para lidar com essas mudanças (NOAA National Centers for Environmental Information, 2020).

As espécies florestais são organismos estacionários que enfrentam o grande desafio de ter que ajustar continuamente sua fisiologia para responder aos estresses durante seu longo ciclo de vida (Gonzalez et al., 2018; Polle et al., 2019; Vieira et al., 2019). Esses estresses abióticos, como extremos de temperatura, seca, sal e metais pesados, e estresses bióticos, como, por exemplo, resistência a pragas e doenças, podem impedir que a planta atinja o crescimento ideal, impactando significativamente a produção (Chen et al., 2020).

Desde a década de 1950, grandes empresas de produtos florestais começaram a investir em silvicultura de plantações e melhoramento de árvores por meio de manejo florestal agressivo para atender à demanda global por madeira e fibra (Carter et al., 2015). A substituição da seleção fenotípica pela seleção genotípica trouxe maior eficiência e precisão ao melhoramento genético. Várias espécies importantes de árvores coníferas, como *Pinus taeda*, (loblolly pine, LP), *Larix gmelinii* e *Pinus elliottii*, foram geneticamente melhorados por várias gerações. O melhor material parental foi selecionado para cruzamento através da predição de valores genéticos, gerando ganhos de 20% em volume individual (Lai et al., 2017; Zhang et al., 2020; Emhart et al., 2007).

Considerada a principal árvore florestal para as indústrias de madeira, celulose e papel no sudeste dos Estados Unidos (EUA), LP é uma das árvores florestais de maior importância econômica do mundo (Plomion et al., 2007; Lu et al., 2017). LP é uma conífera diplóide e de vida longa, que pertence a família Pinaceae. Para alcançar melhorias em características economicamente importantes em LP, como altura da árvore, volume do caule, qualidade da madeira e resistência à ferrugem fusiforme, através do melhoramento clássico, requer cerca de 30 anos de ciclo. Além disso, esses traços complexos variam substancialmente entre as populações (Lauer et al., 2020).

O ciclo de melhoramento tradicional, em pinus, inclui: o período para reprodução, de pelo menos 10 anos, testes de campo, por pelo menos 8 anos, e propagação, por pelo menos mais 8 anos

(Harfouche et al., 2012). Nos EUA, por exemplo, foram necessários 55 anos para completar três ciclos de reprodução em um programa de melhoramento de LP (Zapata-Valenzuela, 2013).

Mesmo com mais de seis décadas de reprodução, esta espécie ainda é considerada em grande parte não domesticada, e tem havido uma pressão crescente para o seu melhoramento (Isik and McKeand, 2019, Telfer et al., 2019). Nesse contexto, os esforços de melhoramento de LP se voltaram para abordagens de todo o genoma devido ao seu potencial para acelerar o melhoramento genético (Ingvarsson and Street, 2011).

Avanços recentes no sequenciamento de próxima geração (NGS) e metodologias biocomputacionais, levaram ao rápido desenvolvimento de recursos genômicos para LP, como sequências genômicas (Zimin et al., 2017; Lu et al., 2016), marcadores moleculares de larga escala (Eckert et al., 2009; De La Torre et al., 2019; Eckert et al., 2010) e mapas genéticos (Neves et al., 2014). LP possui um genoma complexo, mal caracterizado e extremamente grande (aproximadamente 21,6 Gb) (Wegrzyn, 2014; O'Brien et al., 1996), que até o momento, pode ser montado em apenas uma sequência de referência altamente fragmentada, com mais de um milhão de scaffolds e até 82 % conteúdo de vários elementos de DNA altamente repetitivos (Perera et al., 2018). Devido a essa complexidade, a maioria dos estudos genômicos de LP tem se concentrado na análise de sequências de genes nucleares, que são importantes para a aplicação de dados genômicos para o aprimoramento desta conífera (Carrasco et al., 2017; Baker et al., 2018; de Oliveira Junkes et al., 2019; Mao et al., 2019), mesmo em casos de informações limitadas (Soltis and Soltis, 2013).

Inicialmente, os melhoristas florestais investiram na seleção assistida por marcadores (MAS), que é uma técnica utilizada para prever o ganho genético através do uso uma pequena coleção de variantes (tipicamente polimorfismos de nucleotídeo único [SNPs]), ligadas a loci de características quantitativas conhecidas (QTL) (Zapata-Valenzuela, 2012). Porém, considerando a alta complexidade do genoma de LP, são esperados erros na definição de um limite estatístico suficiente para declarar um efeito significativo de um determinado QTL. O poder estatístico para detecção de associações entre marcadores e os fenótipos é inversamente proporcional às correlações quadradas (r2) entre os alelos em um locus do marcador e a variante causal (Pritchard e Przeworski 2001). Portanto, a extensão do desequilíbrio de ligação é importante para se determinar as densidades de marcadores necessárias para representar segmentos de haplótipos não

recombináveis nas populações florestais (Yan et al. 2009). A estimativa do desequilíbrio de ligação entre loci distantes no mesmo cromossomo ou, em cromossomos diferentes, deve ser utilizada para investigar a possibilidade de associações falso-positivas ou falso-negativas (Platt et al. 2010).

LP, por ser considerada uma espécie florestal ainda não completamente domesticadas, é caracterizada por populações que apresentam rápido decaimento no desequilíbrio de ligação (Neale & Kremer, 2011). O cruzamento dentro de populações de coníferas, com grandes tamanhos populacionais efetivos, geralmente apresenta esse rápido decaimento (r2) para menos de 0,1 em 500 a 1500 bases (Pavy et al. 2012; Brown et al. 2004). Por outro lado, em cruzamentos multifamiliares usados para estudos de associação genética e seleção genômica, espera-se que r2 decaia em distâncias maiores, proporcionais aos níveis de parentesco (Flint-Garcia et al. 2003).

Além das grandes porcentagens de falso-positivo encontrados com o uso de MAS, esta técnica também apresenta outra relevante limitação, com relação a capacidade de prever características complexas controladas por muitos loci de efeito pequeno não identificados, como é frequentemente visto nas características de maior crescimento econômico, como crescimento, tolerância à seca e qualidade da madeira de árvores florestais.

Considerando tais limitações do MAS, os melhoristas passaram a voltar suas atenções a uma outra técnica de uso de marcadores moleculares correlacionados a fenótipos de interesse, a chamada seleção genômica. A SG utiliza variantes distribuídas ao longo do genoma para prever o valor genético de genótipos com fenótipos não observados (Spindel e McCouch, 2016; Xu et al., 2020). O uso de SG reduziu a fase de testes de campo para vários meses (Harfouche et al., 2012), tornando desnecessária a avaliação baseada no fenótipo e permitindo a genotipagem baseada em marcadores assim que as sementes estavam disponíveis. Tal estratégia reduziu quase pela metade o ciclo de reprodução de LP por possibilitar a avaliação de plantas em uma idade muito jovem (Resende et al., 2012).

SG em LP aumento a eficiência do melhoramento dos traços de interesse econômico entre 53-112% em comparação com o melhoramento tradicional (Resende et al., 2012), proporcionando um aumento no rendimento desta cultura florestal, uma vez que acelera o melhoramento de plantas. No entanto, as ameaças climáticas estão ocorrendo cada vez com maior intensidade, tornando necessário acelerar ainda mais a reprodução e melhoramento das plantas em direção a adaptação

e resistências a essas novas condições climáticas extremas. Neste sentido, incorporar novas abordagens de melhoramento, como a de aprendizado de máquina apoiará a melhoria das culturas para ajudar a combater os resultados adversos das mudanças climáticas na produção agrícola (Bayer et al., 2021).

3.2.2 Abordagem integrada de análises de Seleção Genômica, Aprendizado de Máquina, GWAS e redes de co-expressão gênica no melhoramento de espécies arbóreas

Seleção Genômica

A seleção genômica envolve a estimativa dos efeitos de um grande número de marcadores moleculares em uma população de melhoramento através do desenvolvendo um modelo misto para prever o valor genético dos fenótipos em um diferente conjunto populacional (Meuwissen, Hayes e Goddard, 2001).

A SG surgiu como uma variante da seleção assistida por marcadores se vem se consolidando como uma ferramenta poderosa no auxílio na seleção de genótipos superiores, possibilitando a identificação de indivíduos superiores a partir da predição de seus valores genéticos, obtidos pela análise de milhares de dados de polimorfismo de DNA e de um reduzido conjunto de dados fenotípicos (Meuwissen et al., 2001).

Nesta abordagem, marcadores distribuídos por todo o genoma são utilizados para predizer o mérito genético de indivíduos, visando a seleção de materiais com desempenho desejado de acordo com o caráter em estudo (Meuwissen et al. 2001). Em geral, os modelos estatísticos utilizados em SG são baseados na pressuposição de normalidade dos valores fenotípicos. Visando contornar essa limitação, Pérez e de los Campos (2014) propuseram a utilização de modelos lineares generalizados sob o enfoque bayesiano (BGLR) estendendo assim a seleção genômica a modelos contínuos e discretos. Desta forma, os métodos mais tradicionais para análises de SG são baseados em diferentes suposições sobre os dados de entrada. Dois principais pressupostos dividem tais métodos: 1) métodos Bayesianos, que permitem que a maioria dos SNPs tenha variância independente, como os BLUP de regressão de crista: BayesA, BayesB, propostos por Meuwissen et al., 2001; Bayes C, proposto por Habier et al., 2011, Bayesian Lasso de Park & Casella, 2008, reproduzindo o espaço de Hilbert do kernel (Gianola et al. al., 2006) e; 2) métodos

que assumem que todos os SNPs contribuem igualmente para o fenótipo, como o BLUP genômico (GBLUP) (VanRaden, 2008).

Aprendizado de Máquina

Apesar do consagrado uso de SG no melhoramento de plantas, existem fatores complicadores aos modelos convencionais de predição, tais como epistasia e dominância, que são efeitos que precisam ser estabelecidos a priori para a utilização do modelo. Uma alternativa para lidar com as restrições associadas aos dados de alta dimensão utilizados na SG, é o uso de abordagens de aprendizado de máquina e aprendizado profundo nas tarefas de predição genômica (Montesinos-López et al., 2021).

O aprendizado profundo é um subcampo do aprendizado de máquina, que por sua vez, é um subcampo da inteligência artificial. O aprendizado profundo compreende métodos que implementam modelos não paramétricos, sendo interessantes por sua capacidade de se adaptar a associação complicada entre os dados e seus padrões ocultos de estrutura desconhecida, que não poderiam ser incorporados de início em modelos paramétricos (Kononenko et al., 2007).

A crescente incorporação de métodos avançados de ML, como por exemplo as Redes Neurais Artificiais (RNA) e Florestas Aleatórias, no contexto de seleção genômica, se deve pelo fato dos algoritmos permitirem treinamento usando representação de dados mais complexos e por não exigirem pressuposições quanto ao modelo. Os resultados de ML não dependem da distribuição das variáveis e si, mas sim do processo de aprendizagem. Essa característica dos métodos de aprendizado estatístico permite que se possa modelar efeitos não aditivos menores, tal como a interação entre genótipos e entre fenótipos (Bayer et al., 2021).

Muitas variantes dos métodos estatísticos tradicionais de GS foram propostas utilizando ML (Montesinos-López et al., 2019), como o BLUP de múltiplos loci ajustados por parentesco (Yin et al., 2020), que utiliza métodos desenvolvidos em ML para otimizar parâmetros de um modelo linear misto enquanto se ajusta à estrutura populacional. Outra alternativa para otimização dos modelos de previsão, é o uso de ML para seleção de atributos (feature selection – FS).

A FS tem por objetivo, a redução do número de SNPs em um conjunto de dados de sequenciamento, identificando um subconjunto de marcadores com maior capacidade preditiva, removendo marcadores com menor significado ao fenótipo predito (Dash e Liu, 1997). A redução

da densidade de marcadores e construção de modelos simples e abrangentes de predição, evitando a atribuição de efeitos não genéticos aos marcadores (Hickey at al., 2014), torna a FS, um método de grande interesse para o aumento do poder preditivo de fenótipos de interesse (Li et al., 2018). Além do ganhos em precisão, a FS também oferece vantagem à SG em termos de redução de tempo, necessário para análises computacionais, e custos de genotipagem para criação de um conjunto de SNPs suficientes para predição fenotípica (Aono et al., 2020).

GWAS e redes de co-expressão gênica

Além da predição genômica, há outro método estatístico denominado associação genômica ampla (GWAS), do qual podemos fazer uso para identificação das complexas relações entre as variações genotípicas e fenotípicas dos porta- enxertos em seringueira. O estudo de GWAS é um método onde marcadores associados com os traços de interesse são determinados via desequilíbrio de ligação e frequência de alelos (Weir, 2008).

Atualmente, com o alto nível de rendimento tecnológico para sequenciamento e maior facilidade de acesso a informações fenotípicas, presentes na grande quantidade de população de melhoramento florestal de referência, o GWAS também se tornou uma interessante estratégia para seleção de genótipos superiores. A associação genômica possibilita o aprofundamento do conhecimento sobre o grau de associação entre alguns marcadores genéticos e regiões gênicas/cromossômicas com a expressão fenotípica de características de interesse (HUANG & HAN, 2014).

O GWAS tem potencial para identificar as variantes genéticas com maior efeito sobre a característica fenotípica estudada, especialmente características que sofrem efeito de um ou alguns genes, a fim de utilizá-las nas avaliações genéticas das plantas. Para fenótipos que são afetados por numerosos loci de características quantitativas menores, o GWAS tem se mostrado menos eficiente (Kainer et al., 2015), principalmente pelo desafio na identificação de genes causais. Além do desafio na identificação de uma lista de potenciais genes causais, há poucas fontes de informações funcionais que podem ajudar na identificação desses genes ligados a um fenótipo, dificultando a interpretação biológica de locus identificados.

Desta forma, para suprir este gargalo no desenvolvimento de uma nova compreensão global de como os genes influenciam as características moduladas por vários genes em plantas, surge a

necessidade de conciliar os estudos de GWAS com outras estratégias genômicas, como o uso de redes de co-expressão gênica. A partir de dados de transcriptomas, fonte valiosa de informações gênicas e anotações, é possível construir redes de co-expressão, que permitem o levantamento do perfil de expressão gênica em diferentes contextos (diferentes tecidos e fases de desenvolvimento), permitindo a contextualização das informações de expressão gênica, sua função biológica e possíveis associações com variações no fenótipo. Como a rede de co-expressão fornece uma medida global de relacionamentos funcionais, inclusive a de variações genéticas que conduzem o fenótipo, espera-se que as redes forneçam um poderoso recurso para interpretar os loci capturados pelo GWAS (Schaefer et al., 2018).

Adicionalmente a possibilidade de capturar relações funcionais por meio das relações de co-expressão de vários genes diferentes que podem estar associados a um fenótipo, as redes de co-expressao apresentam uma forma potencial de classificação dos genes identificados por GWAS. Essa classificação pode ser realizada quanto a centralidade dos genes candidatos do GWAS na modulação da característica fenotípica. Genes identificados como centrais na rede, podem ser classificados como moduladores diretos da característica associada, enquanto, os genes periféricos, afetam tal característica por meio de seus efeitos indiretos nas redes, através das suas correlações com os genes centrais (Liu et al., 2019).

A combinação de diferentes estatísticas e análises genômicas, como SG, aprendizado de máquina, GWAS e redes de co-expressão gênica, se torna uma estratégia promissoras para que seja possível lidar efetivamente com o melhoramento de caracteres complexos. A implementação destas técnicas de melhoramento em espécies florestais se apresenta como solução inovadora e promissora para lidar com as características biológicas e genéticas destes organismos complexos, considerando a importância dos caracteres quantitativos complexos e, muitas vezes estruturados nestas espécies. Além disso, a dificuldade e o elevado custo da fenotipagem em árvores, também torna atrativo o uso de genotipagem em larga escala para predição de fenótipos com características melhoradas.

4. OBJETIVOS

4.1 OBJETIVO GERAL

Integrar múltiplas análises ômicas: SG, GWAS, ML e rede de co-expressão gênica para seleção de genótipos superiores em espécies arbóreas.

4.2 OBJETIVOS ESPECÍFICOS

1) Aplicar modelos de predição genômica usando características de importância econômica para espécies florestais, como volume, para a seleção precoce de genótipos superiores usando um extenso conjunto de dados de marcadores SNP e dados fenotípicos de LP.

2) Avaliar a aplicação de técnicas de aprendizado de máquina, através da seleção de atributos (FS), para construção de subconjuntos dos dados SNP utilizados em tarefa preditiva, comparando o ML como ferramenta complementar as técnicas tradicionais de SG.

3) Aplicar Associação Genômica (GWAS) para identificar marcadores moleculares associados a característica fenotípica de volume em pinus

4) Usar redes de co-expressão gênica para investigar as vias metabólicas relacionadas à volume em pinus e, resposta à frio em seringueira.

5. *CAPÍTULO I*: ARTIGO CIENTÍFICO SOBRE SG, ML, GWAS E REDES DE CO-EXPRESSÃO GÊNICA EM PINUS

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

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 pine trees have focused on survival, early and rapid growth, resistance to diseases and pests, and stem shape. Among the growth traits, volume is one of the most important variables in forests, and monitoring it is one of the main tasks of regional management plans. Despite the great interest in increasing forest volume, there is a great challenge in unraveling the molecular mechanisms behind this quantitative trait since it is presumably influenced by the action of an unknown network of genes interacting through complex molecular mechanisms. This challenge becomes even greater considering the extremely large size and high complexity of the Pinus genome, making the characterization, sequencing and computational analysis difficult. In this study, we present the first comprehensive integrated analysis of LP involving genetic markers in association with volume (via GWAS) and ML-selected markers for genomic selection (GS) to understand which biological pathways are 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 machine learning, i.e., feature selection (FS), was used for GS, providing a predictive accuracy (R Pearson coefficient) of 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: Machine learning, RNA-Seq, genomic prediction, GWAS, gene coexpression network, forest breeding

1. Introduction

Since the 1950s, large forest product companies have started to invest in plantation silviculture and tree breeding through aggressive forest management to meet the global demand for wood and fiber (Carter et al., 2015). Considered the leading forest tree for the timber, pulp and paper industries in the southeastern United States (US), *Pinus taeda* (loblolly pine, LP) is one of the most economically important forest trees in the world (Plomion et al., 2007; Lu et al., 2017). Even with more than six decades of breeding, this species is still considered largely undomesticated, and there has been increasing pressure for its improvement with lasting effects on forest productivity, forest values and ecosystem services (Isik and McKeand, 2019; Telfer et al., 2019).

The LP breeding cycle with traditional phenotypic selection based on economically important traits requires 15 to 30 years, including at least 8 years of field testing (Harfouche et al., 2012). Even though classical pine breeding strategies have provided significant gains for the last 30 years, efforts in biotechnology and omics-based approaches have been suggested as a tool to expand such genetic gains (Poovaiah et al., 2020; Caballero et al., 2021; Goralogia et al., 2021; Shalizi et al., 2022); for LP, 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 (Mao et al., 2019; de Oliveira Junkes et al., 2019; Carrasco et al., 2017; Baker et al., 2018). In this context, the inclusion of additional molecular data sources can benefit LP efforts, especially for high quantitative traits, such as stem volume (SV).

SV is one of the most important physiological variables in forest trees, and its monitoring is one of the main tasks of regional management plans (Shalizi and Isik, 2019). The targeting of genetic breeding programs to produce varieties with high volume is increasing for many forest species, including LP (Shalizi and Isik, 2019; Lu et al., 2018), in which an enlarged volume is related to increased fuelwood production (Bouvet et al., 2020). However, the challenge of increasing volume in LP forestry production is precisely the greater clarification of the polygenic basis of this complex trait that can explain the phenotypic variations, which vary substantially between populations (Lauer et al., 2020).

Statistical models 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 relevant associations through 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; Grattapaglia 2014; Beaulieu et al. 2014; Isik et al. 2016; Lenz et al. 2020).

Even though theoretical genomic prediction models are 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. On the other hand, through GWASs, it is possible to assess how this genomic variation is maintained and to estimate the forces that shape the genetic architecture of complex traits (De La Torre et al., 2019). However, 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 a deeper comprehension, additional methodologies to GS and GWAS are needed, and the integration of different layers of omics can provide a greater picture of unknown molecular interactions.

Nonparametric models for genomic prediction can be an alternative for modeling genotype-phenotype associations, with the role of machine learning (ML) 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 genomic prediction 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 stem volume (SV) trait, 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

SV phenotypes were analyzed using ASReml v.2 (Gilmour, et al., 2006) according to the linear mixed model created:

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

where *Y* is the phenotypic measure of the *ith* genotype considering the *jth* block, the *rth* replication, and the *mth* and *nth* 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 designed based on the probe set used for sequence capture containing 54,773 probes representing 14,729 *P. taeda* unigenes (Neves et al., 2013). Target enriched libraries were sequenced by the company RapidGenomic using outputs from Illumina HiSeq 2000. The approaches of reduced representation of the LP genome using sequence capture, as well as other approaches such as exome capture (Neves et al., 2013) and Pita50K array (Caballero et al., 2021), are robust strategies for detecting variants from different fragments from the many genome copies within each sample. In this way, sequence capture has become a cost-effective alternative to whole-genome sequencing (Lebedev et al., 2020). The use of probes (single-stranded sequence of DNA selected from reference sequence data including genomes and/or transcriptomes used) were used to hybridize with its complementary sequence in the samples of genotypes, allowing target SNPs, genes and QTLs at the same time to produce sequence data in each region.

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 the Bowtie2 v.2.2.9 algorithm (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 software (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 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 (Mao et al., 2019; de Oliveira Junkes, et al., 2019; Carrasco, et al., 2017). 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, to determine the genes involved in the terpenoid biosynthesis pathway. 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) under field commercial resinosis to better understand the molecular bases of resin production, a major source of terpenes for the industry. 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, separated into (I) quality filtering using Trimmomatic v.0.39 (Bolger et al., 2014), (II) transcriptome assembly with Trinity v.2.11.0 software (Grabherr et al., 2011), (III) transcript quantification with Salmon v.1.1.0 software (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 (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 GO terms and EC numbers to the assembled contigs, we used Trinotate software together with the SwissProt database (Supplementary Table 6). All of these assembled contigs were aligned against LP scaffolds to obtain putative correspondence of probes with transcripts.

2.5. Genome-Wide Association

The first analysis performed in this study for investigation of genotipic markers with significant effects on phenotypic variation were GWAS. The markers with the main effects in SV were used for functional inferences based on the neighboring genes and the putative corresponding

proteins. A 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). To include genotype-environment correlation in the association mapping, we 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. The first tests of the GWAS were performed without covariates (Supplementary Table 7, Supplementary Figs. 6-7).

The functional profiles of significant SNPs were estimated 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). We retrieved Gene Ontology (GO) terms and Enzyme Commission (EC) numbers from the transcriptome annotations.

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 genomic prediction (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 $\sigma^2{}_\gamma$ 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: (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 an estimated 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).

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 inside 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

Several LP families representing seedlings planted across seven sites were used in this study (Supplementary Fig. 1, Supplementary Table 1). All the raw measures for SV phenotyped in these individuals were corrected based on the mixed model described (Supplementary Fig. 2). Individuals who did not survive in the field until the phenotyping period were excluded from further analysis. A total of 19,933 (~80%) surviving individuals were considered for phenotypic analyses. Heritabilities and dominance ratios were estimated across sites (Supplementary Table 2).

3.2. Genotyping

For genotyping, 1,692 individuals from four environments and 45 families were selected (Supplementary Table 3). The genotyping process was performed using a probe set used for sequence capture containing 54,773 probes representing 14,729 *P. taeda* unigenes (Neves et al., 2013), resulting in a total of 11,867 probes with adjacent nonoverlapping 120-nt-long regions extracted as putative probes from the inner part of the UniGene sequence (or predicted exon). These probes are located in 10,412 sequences from the *P. taeda* reference genome. The sequencing experiment generated 996,611 million reads, which were mapped against the probe region extracted from the *P. taeda* reference genome with an extension of 500 bp. This step led to the collapse of probe sequences with close proximity (reducing the 11,867 probes to 10,412 sequences).

Among the sequenced reads, 84.5% were mapped against the probe sequences, resulting in an initial count of 133,199 markers, 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 to be used in the GWAS, we used more stringent parameters (maximum of 10% missing genotypes at a SNP and minimum

and maximum allele frequencies of 0.05 and 0.95, respectively), resulting in 31,589 SNPs distributed among almost all probes with different quantities depending on the reference sequence length (Supplementary Fig. 3).

The PCA performed with the SNP data suggested the presence of genetic clusters for each family (Fig. 1-I), which was not observed when considering the different locations of the populations. LD decay was estimated with the squared Pearson correlation coefficient (R^2) between SNP pairs linked within respective probes. We observed a rapid decay to the half maximum $R2 = 0.05$ within less than 100 kb (Fig. 1-II).

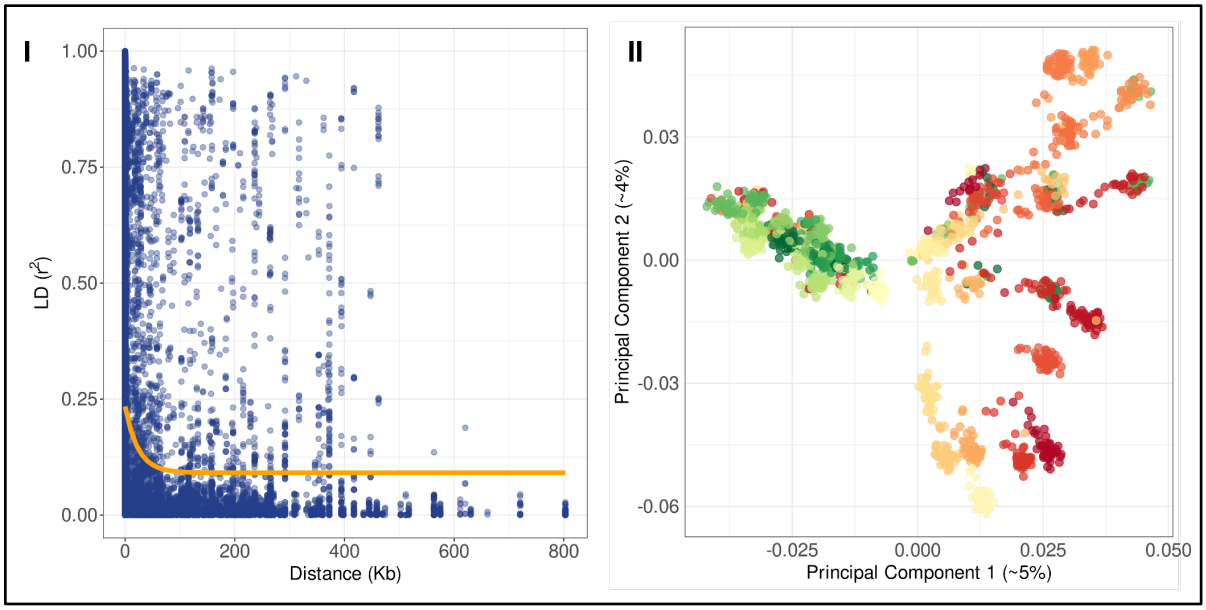


Fig. 1. Genotypic analyses performed on the SNP data separated into (I) linkage disequilibrium (LD) decay considering squared Pearson correlations (R^2) and distances calculated based on the respective probes (measured in kilobase pairs); (II) principal component analysis (PCA) scatter plot of the *Pinus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

The functional characterization genotypic data can only be better understood through functional genomics (Ding et al., 2022). In this sense, to refine the biological and annotation

information of the SNPs, three independent RNA-Seq experiments were used to generate a gene coexpression network and to supply evidence of biological inferences for genomic regions containing SV-associated markers. Gene expression profiles can be compared and quantified in different contexts (tissues, developmental stage, or stress conditions) to indicate joint gene regulation and shared function (Eisen et al., 1998).

We selected experiments from 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 software. We generated a total of 75,897, 137,103, and 50,781 genes for Pta, Pra and Pel, respectively. 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 BUSCO annotation (Supplementary Fig. 4). These evaluations revealed that the transcriptome sets are well represented and can serve as a de novo assembled reference; however, we still selected a subset of genes based on the expression results (Supplementary Table 5), 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

We identified seven candidate loci for the SV trait. 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.

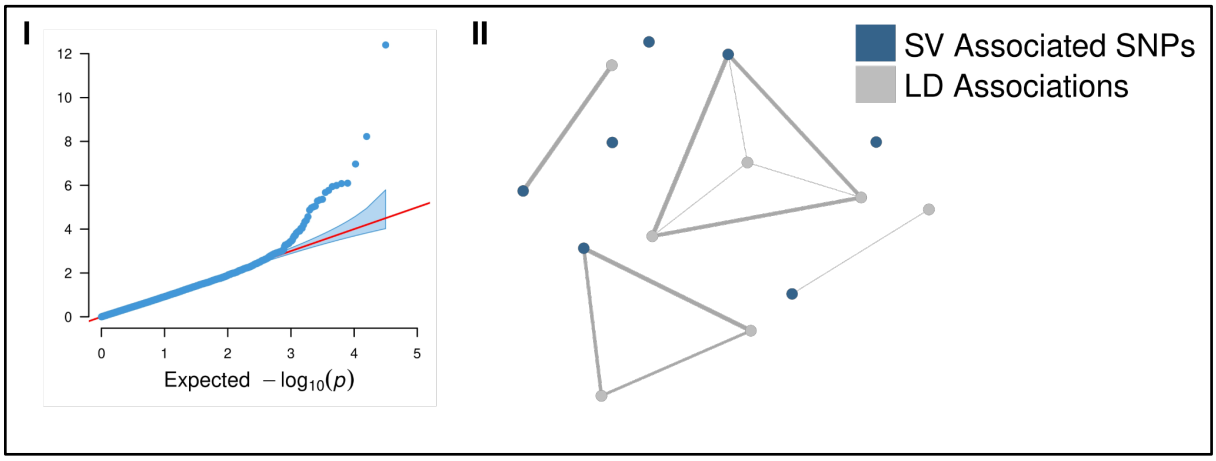


Fig. 2. (I) Quantile-quantile (QQ) plot of FarmCPU models in the genome-wide association study (GWAS) with the stem volume (SV) trait. The set of markers associated with SV together with the other SNPs in linkage disequilibrium ($R^2 \geq 0.5$) is displayed in (II), which is a network representing these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

This set of markers was extended by LD tests, as shown in Figure 2-II and listed in Supplementary Table 9. No correlations were identified between the SNPs found in association with SV through GWAS (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 genes for GWAS (830 isoforms) and 191 genes for LD (725 isoforms). From the Pel dataset, 65 genes were found through GWAS (133 isoforms) and 96 through the LD set (139 isoforms). Finally, from the Pta transcriptome, we found 564 genes (1218 isoforms) associated with GWAS results and 472 (1004 isoforms) 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). Interestingly, the genes identified in the neighborhood of the marks found 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

marks 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 genomic prediction 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.

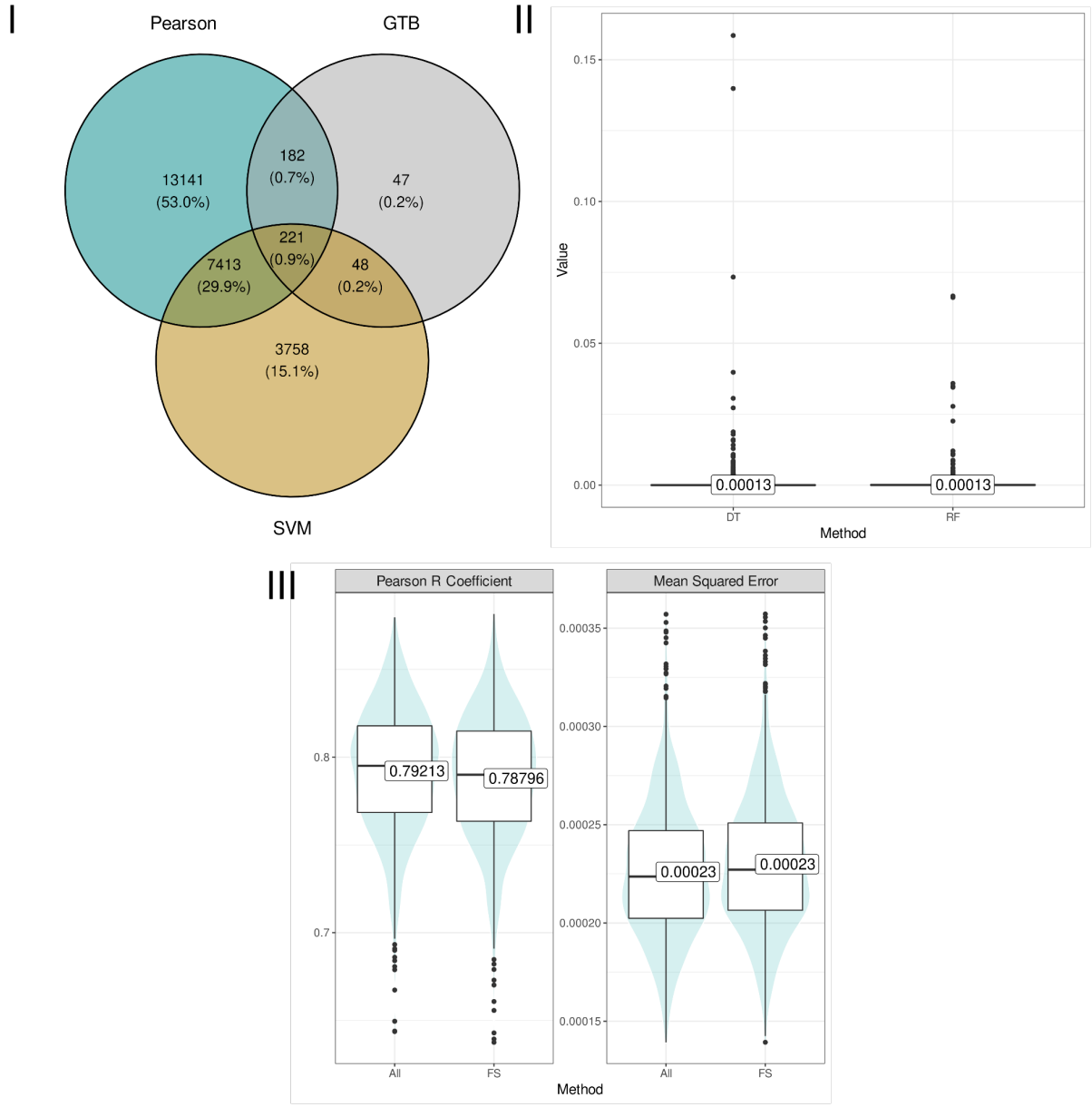


Fig. 3. Genomic prediction results for stem volume using a total of 31,589 SNPs (All) and 7,864 selected through feature selection (FS) based on (a) Pearson correlations (p value of 0.05), (b) gradient tree boosting (GTB), and (c) support vector machine (SVM). (I) Markers selected through each FS method. (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, representing a major advance in reducing the time required for predictive computational analyses and 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 the implications of GWAS-associated genes in the molecular mechanisms of pine from a broader perspective, we created different 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). 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 the Pra and Pel networks, respectively.

For each one of the groups defined in each network, we evaluated the presence of genes associated with (i) GWAS markers (7 SNPs); (ii) markers associated with 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). Two groups (G12 and G159) of the Pel expression network, with

the expression of 3 ML-selected markers, were enriched for catabolic lignin, relating the importance of these genes involved in chemical reactions and pathways that result in lignin breakdown to associate genotypes with the volume trait. The lignin catabolic process also appeared enriched in the Pra expression network in 3 groups (G13, G36 and G57) involving the coexpression of 8 markers selected by ML.

Another interesting genetic orthology found in groups (G13, G15 and G19, G65) with coexpression of 9 markers selected by ML was cellulose metabolic process and biosynthetic cellulose in Pra. G12, from the Pel network, also showed enrichment of hemicellulose metabolic processes.

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 SV molecular mechanisms 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 from such a set of genes (Fig. 4-II), we can identify 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 as hubs of the HRR network the genes TRINITY_DN34225_c0_g1, annotated for the peptidyl-prolyl cis-trans isomerase enzyme, and the gene TRINITY_DN37756_c0_g1 without known annotation, both identified by the GWAS, presenting the highest degree value (57). Other metrics were also analyzed to identify genes with great importance in the network, such as TRINITY_DN12054_c0_g3, annotated for GTP cyclohydrolase 1 enzyme, with the highest stress value (320134), TRINITY_DN3090_c0_g1 with the lowest short path length value (2) and the highest value of betweenness centrality (0.206), identified by the FS, and the gene

TRINITY_DN72617_c0_g1, annotated for the nonlysosomal glucosylceramidase enzyme, showing a higher value of neighborhood connectivity (48).

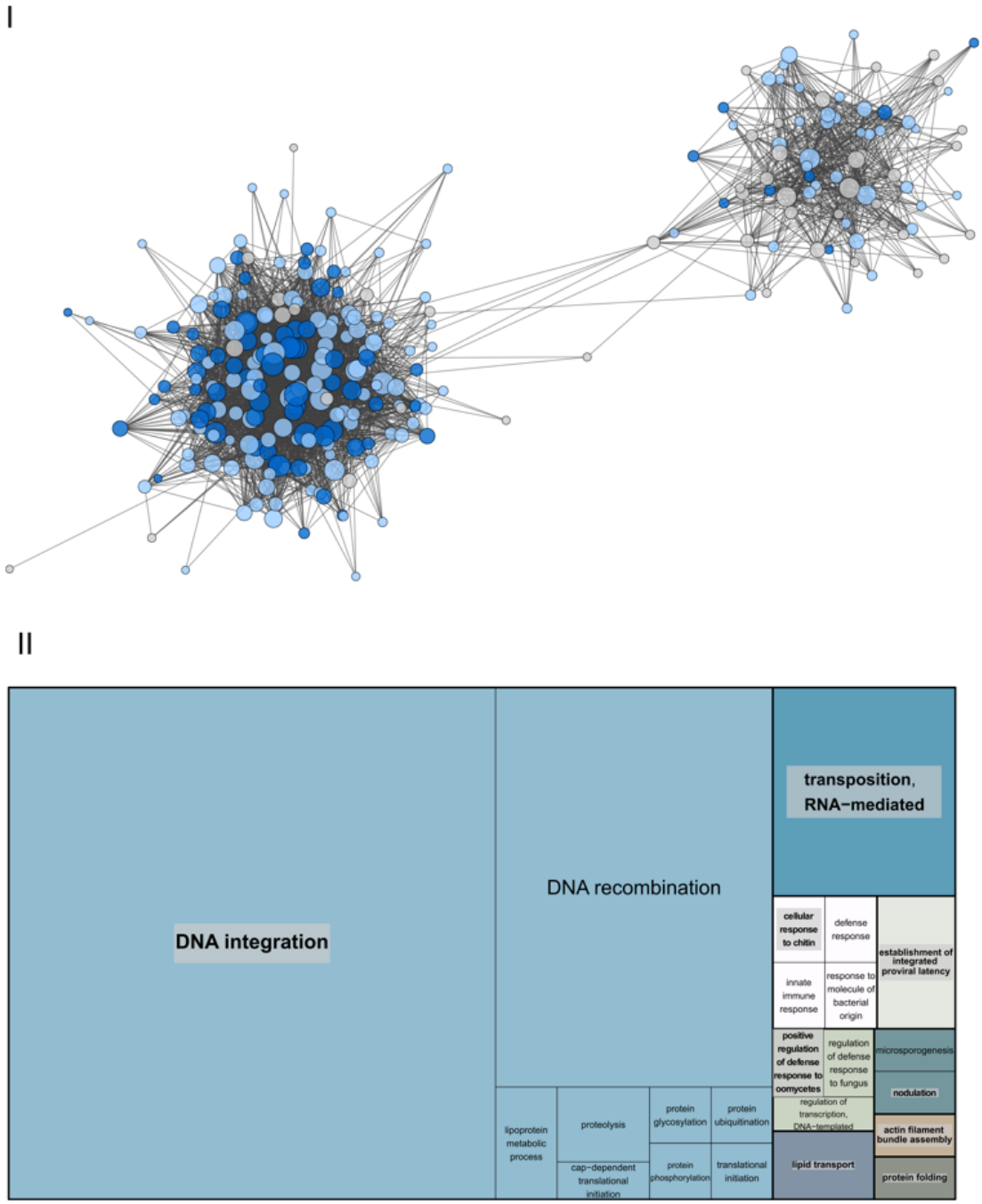


Fig. 4. (I) Coexpression network created with the highest reciprocal rank (HRR) methodology considering groups of coexpressed genes selected in the weighted gene coexpression network

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 (Shalizi and Isik, 2019, Schmidt et al., 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 in general, the heritability values are moderate considering the ages of the plants. Such heritability values were

similar to those of other authors (Shalizi et al., 2020; Xiang et al., 2003). We know that the trend in volume heritability in LP is increasing over time and that there are also age-age correlations with other growth traits. The recommended age for selection can be as early as 3 for height and as early as 4 for diameter at breast height (DAP) or volume (Xiang et al., 2003), which was the age measured 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 very 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 predictive 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).

The results of the SV phenotypic distribution in the breeding populations of LP half-sib trees used in this study are presented as a good example for the application of quantitative genetics methods to study 1) the association of genotypes and their phenotypes, as well as 2) the selection of superior genotypes through GS. GS is an important strategy for long breeding cycles, such as those required by pine trees, early selection of genotypes with good performance of complex traits (Bhat et al., 2016) and with good dispersion of the phenotypes, as found here. 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. The half-sibling family structure seems to require more markers and the GS accuracy is approximately 0.3 to 0.5. Lower GS prediction values have been reported for unrelated individuals (Lenz et al., 2017). Here, we achieve high GS values, as has been previously obtained in full-sib families, with the advantage of using a very small set of markers.

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, based on probes, was based on the challenges of interrogating the LP megagenome and enabled the detection of high confidence variants. 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.

The use of Capture-Seq in this work allowed efficient genotyping, with the identification of a total of 31,589 markers, even using high-quality filters. These SNPs showed LD (Figure 1) along the genome similar to those reported in other studies using probes designed based on loblolly pine exome capture: the NimbleGen SeqCap EZ method (Lu et al. al., 2016). That wide range of genomic information about the breeding material used, is critical to the success of the GS, GWAS and ML analyses performed.

4.2. GWAS for pine breeding

Despite studies employing GWAS in some pine populations (De La Torre et al., 2019; Ding et al. 2022), especially for growth traits (Santini et al., 2020), several challenges still remain related to the many single nucleotide polymorphisms (SNPs) associated with complex traits and the limited understanding of the functional mechanisms by which these genetic variants are associated (Khatiwada et al., 2022). Finding a gene underlying a QTL is an enormous task, even more so in a species with genomes as large as LP.

Brown et al., 2003 used GWAS to associate LP genotypes with wood property traits and found the relationship of C4H, C3H, 4CL and CCoAOMT of monolignol synthesis with the trait studied. However, limitations regarding the characteristics of the population studies (strong linkage disequilibrium) required validation in different populations to test the association of SNPs in candidate genes with wood property phenotypes. In this sense, the populations studied here include genomic evaluations in natural populations with a wide distribution on the east coast of the USA, which is an interesting resource for the study of volume in LP. De La Torre et al., 2018 also used LP populations to find associations between genotypes and 409 variables, including

morphological and molecular traits. All these studies, including our results, contribute to the current genomic resources available for LP.

Among the genotypic markers found in association with SV, 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, which is essential 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 showed 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).

Looking at this data thinking about deepening the genotypic and phenotypic associations, the nest step was contemplate additional association analysis using genetic markers selected by ML and SF, which help to reduce the risks of false positives and negatives. The genetic markers identified by GWAS, and ML form a successful proof of concept to combine volume phenotyping with capture sequence genotyping. In addition to helping to better elucidate the genetic architecture

of phenotypic variation in pine, our work also used association analysis to integrate with a gene coexpression network to clarify the hub and key genes that were strongly associated with volume.

Although growth is a complex trait governed by several genes (Santini et al. 2020; Francisco et al., 2021), several 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 all 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). Ding et al. 2022 tried to overcome such limitations of GWAS by integrating the results of these analyses with coexpression networks in P. elliottii. This approach identified important functional modules in which it was possible to make biological inferences about the interactions of genes involved in species growth (Ding et al. 2022). Such multiomica strategies are increasingly being used in tree species, aiming to overcome the individual limitations of these different omics in addition to showing great importance for a deep understanding of the molecular mechanisms involved in the definition of the agronomic trait (Francisco et al., 2021; Ding et al. 2022).

4.3. 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 genomic prediction 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). 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).

Despite the current interest and application of GS in forest species (Lebedev et al., 2020), there is still concern about optimizing genotypic data and bioinformatics analysis, which inspired our choice to use FS to direct a more accurate selection of genetic markers for genomic prediction.

Testing for growth traits in clonal varieties of P. taeda planted in different environments is expensive and time-consuming; however, estimates of genomic prediction accuracy already observed in previous studies are moderate to high, ranging from 0.34 to 0.76 (Zapata-Valenzuela et al. 2012, 2013; Resende et al. 2012a). More recent research has also estimated the difference between the predictive ability of estimated breeding values in stem volume considering scenarios with randomly assigned clonal varieties and a family cross-validation scenario. The results showed a greater predictive ability of stem volume by random genotypes (0.43) than considering the canaries of families (0.36) (Shalizi et al., 2022).

Although there are different GS models that are being used for genomic predictions 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 (Chen et al., 2018.; Isik et al., 2016; Calleja-Rodriguez et al., 2020). Calleja-Rodriguez et al., 2020 demonstrated that the application of genomic best linear unbiased prediction (GBLUP), BRR or Bayesian LASSO (BL) models provided similar prediction efficiencies for growth characteristics and wood quality in maritime pine. Despite this trend of similarity of SG results for most complex forest traits, in LP, Resende et al., 2012, found better results for disease resistance using Bayesian methods when compared with BLUP-based methods. 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. FS can offer a more efficient way of selecting genotypes, optimizing the use of markers considering the influence of SNPs on the phenotype to be selected (Aono et al., 2020).

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 genomic prediction 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).

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 feature selection, unlike a random selection of marker subsets, is able to indicate independent datasets and 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). The difficulty of selecting subsets of markers for genomic prediction lies in the existence of the basic requirement that the markers must be distributed along the genome with at least one marker in linkage disequilibrium (LD) with each QTL (Hayes and Goddard, 2001). However, when you have a number of predictors, which are the genetic markers, much greater than the number of genotypes, it is very difficult to fit the prediction models due to overfitting between the predictors (Neves et al., 2012).

Faced with these challenges, the use of FS with ML tools emerged as a useful tool to deal with a large dimensionality of data, often due to redundant, irrelevant or noise elements (Aono et al., 2022). 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).

Running DT and RF models to recover additional phenotype-genotype associations, it was possible to identify that within the large arsenal of important markers for GS, a reduced number of 217 markers are associated with the volume trait, which enhances the role of these markers in QTL regions for this trait that is difficult to monitor due to its quantitative characteristic. Further strengthening the results of the importance of these 217 markers identified by ML, we observed that the GWAS markers were also contained in this subset of SNPs, minimizing the identification of false positives.

4.4.1 Inferences into pine growth

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. Given the difficulty of finding functional and quality data to represent the genomic complexity of Pinus taeda, we chose to use RNA-seq experiments of correlated species, with a significant number of repetitions and good experimental design, to contribute functional information on the coexpression developed.

Furthermore, gene coexpression network has been successfully used to characterize GWAS results (Francisco et al., 2021; Lee and Lee, 2018), such as identifying functionally related genes (Wem et al., 2018). This combination of quantitative analyses with gene expression data minimizes the problem of false positives or negatives and improves the understanding of the genomic structure of complex traits. The analysis of transcripts and their correlations is a highly informative and easily measurable source of functional information, in addition to allowing the tracking of expression patterns related to phenotype variation (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, detected by GWAS, is

encoded by coregulated genes, we will be able to capture part of the functional relationships of the genotypes studied (Schaefer et al., 2018).

Gene coexpression networks allow us to associate a function with genes coexpressed in the same cluster through the logic known as the “guilt-by-association principle” (Oliver S. 2000). In this way, it is possible to correlate several transcripts from pine trees with the genetic markers that we found to be associated with the volume trait. Considering that the phenotypic effect of each SNP selected by GWAS, and FS can be very small, grouping the important genes according to their expression tendencies allows the identification of a key physiological process or regulatory network involving most of these genes (Baranzini et al., 2009). Genes with high correlation and influence on networks are called hub genes and may be related to mutations (Lamara et al., 2016) and QTLs of interest.

4.4.2 WGCNA results

In a recent study, a combination of methodologies was also used with transcriptome and metabolome to investigate gene expression and metabolic differences in different tissues of *P. taeda* to provide information on the genetic association of small molecule metabolites (Mao et al., 2021). The results of the present study can complement such information, considering that some groups of the gene coexpression network analyzed here showed enrichment for small-molecule metabolite processes, such as group 13 of Pra, which was associated with 3 ML markers. Processes of "flavonoid biosynthetic process", "flavonoid metabolic process", "flavone biosynthetic process", and "flavonol biosynthetic process" were linked to 3 ML markers, indicating that these metabolic processes are associated with SV in LP.

It was also possible to identify the importance of biological pathways linked to the "carbohydrate transmembrane transport", "carbohydrate metabolic process", "carbohydrate derivative catabolic processes. ", "carbohydrate transport" and "cellular carbohydrate metabolic process" in group 1 of the Pra network, linked to 14 Ml brands and 1 GWA brand linked to volume.

This significant participation of processes linked to the storage of carbohydrates, 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 pathways in pine, such as "response to cold", "cellular response to freezing", "response to temperature stimulus", "response to freezing", "response to external stimulus", and "response to abiotic stimulus", which corroborates the findings of Gao et al., 2021, that volume growth is associated with winter values, where higher winter values were associated with further growth in volume. 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.

Group 39, composed of more than 92% of markers selected by ML, showed significant transcripts related to the nitrogen metabolic process. The annotation of the genes coexpressed in group 39 was enriched with the "cellular nitrogen compound metabolic process" and "nitrogen compound metabolic process" pathways. The low availability of nitrogen (N) is an important limiting factor for the growth and development of trees. The acquisition, assimilation, storage, recycling, and efficient metabolic use of nitrogen largely determine vascular development, forest tree productivity and biomass production (Canovas et al., 2018). This important biological pathway of tree development indicates that this group of the Pra network was able to capture an important genetic region of LP.

In this same group 39, the participation of the "metabolic process of cellular aromatic compounds" among the ML marks is notorious. Zhang et al. 2018 reported an improvement in nitrogen use efficiency when, under low nitrogen, excess carbon is redirected to the biosynthesis of aromatic amino acids and lignin. This nitrogen use efficiency mechanism deserves special attention (Cánovas et al., 2018) and has practical value in terms of biomass production.

Group 68 of Pra, the second group in terms of representativeness of ML and GWAS markers, 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 in this group is linked to viral responses such as "viral latency", "establishment of viral latency", "viral genome integration into host DNA", "viral entry into host cell entry into host cell", "entry into cell of other organism involved", "viral life cycle" and "viral process". In this sense, it is

possible to link the importance of the viral response to the development of volume in more productive individuals.

Considering that glutamine synthetase is a key enzyme in nitrogen metabolism for popular growth and biomass production (Canovas et al., 2018; Fu et al., 2003), we also found interesting results in the Pel coexpressed network related to "glutamine family amino acid catabolic process" in group 39 (with 2 ML markers and 1 GWAS mark); 2 ML markers and 1 GWAS mark associated with "regulation of glutamine family amino acid" in group 58; and "glutamine family amino acid metabolic process" and "glutamine family amino acid catabolic process" linked to 3 ML markers in group 48.

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, such as "response to light stimulus", "response to radiation", "regulation of cellular response to X-ray", and "response to UV-C". In this group, enrichment directed at the metabolism of cell walls and macromolecules was also noted.

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), to light interception, in which volume growth per unit of absorbed light increases with increasing stand density (Albaugh et al., 2018) and with 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). The enrichment of GO terms associated with the processes of 'purine metabolism', "purine nucleotide metabolic process", "purine nucleoside triphosphate metabolic process", "purine ribonucleoside triphosphate metabolic process", "purine nucleoside monophosphate metabolic process", "purine ribonucleoside monophosphate metabolic process", and "purine ribonucleotide metabolic process" in this group suggests that physiological signals for volume development in LP, involving "photosynthesis", "cellular respiration" and "ATP metabolic process", are directly involved with nucleoside levels.

4.5 HRR results

A search for overlap between genes implicated by WGCNA allowed the correlation between genes from LP transcriptomes with the markers associated with volume phenotypes, thus allowing the functional understanding of genes involved in the same modules identified by coexpression networks. All GWAS genes overlap with the ML gene pool. This matching of the results of the GWAS and ML analyses allowed WGCNA to detect many biological pathways related to the phenotypic variation studied. To deepen the biological meaning of the main gene modules found in the WGCNA, another gene coexpression network was constructed using the HRR method, which made it possible to increase the correlation between the global coexpression network and the enrichment of the annotation of the Gene Ontology. The analysis of the most significant modules in the gene coexpression network using HRR, considering the 128 markers identified by ML and the candidate genes from GWAS, allowed a global understanding of which biological processes are important for volume in pine. The two hub, for the degree metric, genes of the HRR coexpression network match exactly with two of the genes found by GWAS, one of which, TRINITY_DN34225_0_g1, was annotated for the enzyme peptidyl-prolyl cis-trans isomerase (PPIases).

These enzymes are associated with plant growth and development, and PPIase genes have previously been indicated as suitable candidate genes to increase the abiotic stress tolerance of

plants for industrial processes and agricultural applications (Singh et al., 2021). Although the TRINITY_DN37756_c0_g1 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, TRINITY_DN12054_c0_g3, annotated for GTP cyclohydrolase 1 enzyme, with the highest stress value (320134), was also 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). We also identified a nonlysosomal glucosylceramidase enzyme (TRINITY_DN72617 pude_c0_g1 gene) identified by the FS with the highest value of neighborhood connectivity (48), which the other genes in the network are of great importance, thus playing a role in defining the phenotype role. 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 TRINITY_DN3090_c0_g1 gene, with no known annotation, was also fundamental to maintain the network topology, with the lowest short path length value, also indicating great importance for the dissemination of information through the network (Watts and Strogatz, 1998), and a higher value of betweenness places it as fundamental for the integration of network components (Brandes, 2001).

The characterization of these volume-associated marks by ML and GWAS allows us to investigate the biological processes linked to loci of known quantitative traits (QTL) 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 SG) with transcriptome data and gene coexpression networks allow the targeting of which genomic modules are important for the expression of good volume phenotypes.

4.6. Adoption of different genomic strategies during forest improvement cycles

Plant breeding is a process of continuous genetic improvement carried out by the selection and recombination of superior strains. The "breeder equation" expresses multiple factors that influence the selection response or rate of genetic gain (Lush, 1936). Optimizing available resources to maximize this 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 populations analyzed in pine, as

validated by the combination of GWAS, ML and SG analyses. The application of machine learning models, DT and RF, to recover additional phenotype-genotype associations using the dataset identified by FS for GS allowed the narrowing of 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 LB full-sib populations proved to be an efficient way 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 for SG 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), and FF received a PhD fellowship from FAPESP (2018/18985-7). AS received a Research fellowship (312777/2018-3) from CNPq.

Author Contribution Statement

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

Data Availability

All phenotypic and genotypic data used in this study will be available upon request to the Cooperative Forest Genetics Research Program (CFGRP): https://programs.ifas.ufl.edu/cfgrp/.

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.

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 controlpollinated 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., Schuؼtte, M., Loraine, A., Ebenhoؼh, 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.

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.

Supplementary Material

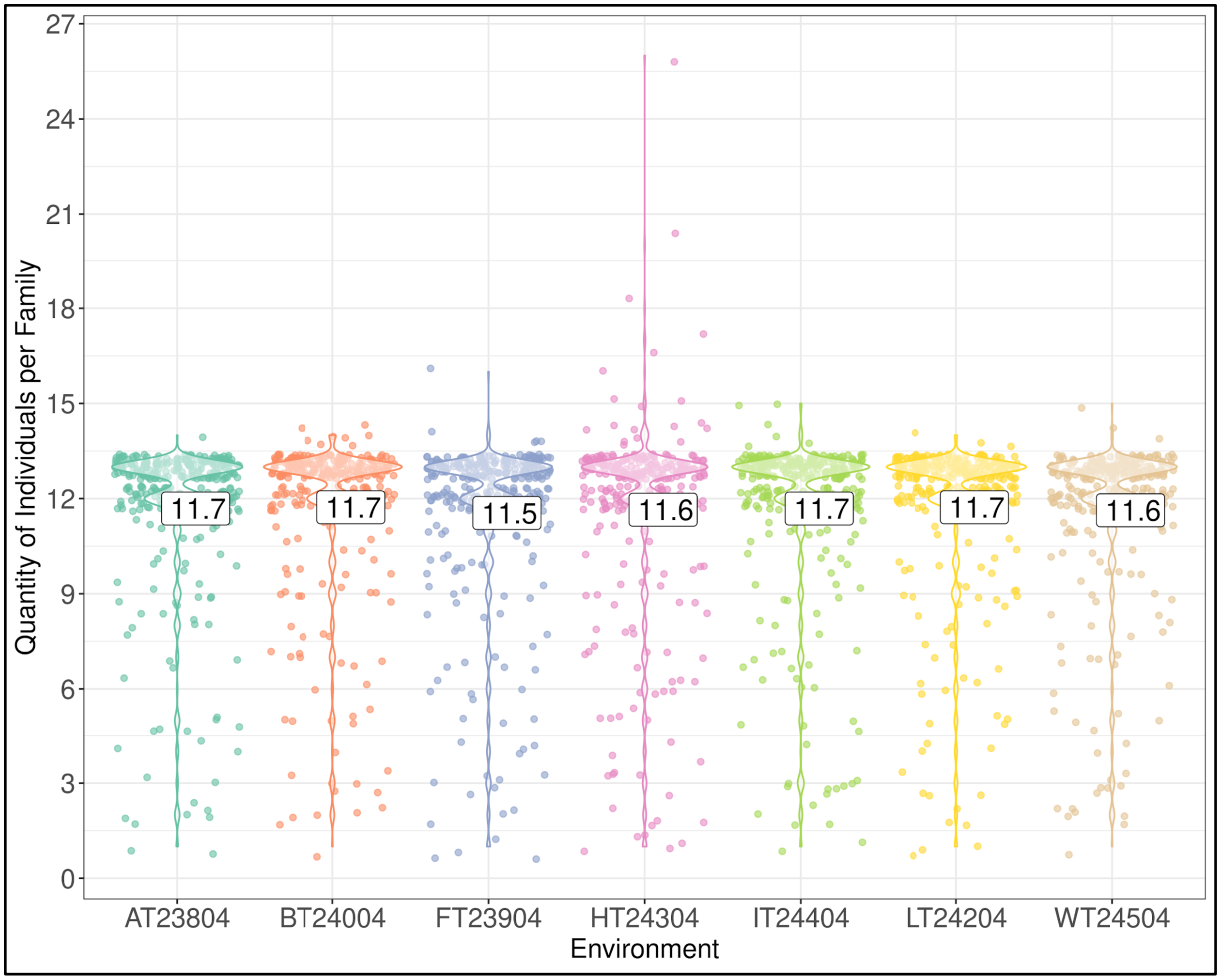
1. Plant Material

The experimental design included 12 replicates of 300 trees in the seven evaluated environments (AT23804, BT24004, FT23904, HT24304, IT24404, LT24204 and WT24504), resulting in 25,080 trees. Herein, single-tree plots represent the experimental units. A check lot (CCK) for comparisons among tests was developed for each region to represent the average genetic performance. One hundred twenty-eight loblolly pine progenitors were used in the mating scheme and separated into females and males:

Females (128): TB01, TB02, TB04, TB05, TB06, TB07, TB08, TB09, TB10, TB14, TB15, TB17, TC01, TC02, TC03, TC05, TC08, TC09, TC10, TC11, TC16, TC18, TC21, TE05, TE06, TE07, TE08, TE09, TE10, TE11, TE12, TE13, TE17, TE18, TE19, TE20, TF01, TF02, TF03, TF04, TF05, TF06, TF07, TF08, TF10, TF11, TF12, TF13, TF14, TF15, TF16, TF17, TF18, TF20, TF21, TF24, TH01, TH02, TH04, TH05, TH07, TH09, TH10, TH11, TH13, TH14, TH16, TH18, TH19, TH21, TH22, TH24, TI01, TI03, TI06, TI07, TI08, TI10, TI11, TI12, TI16, TI17, TI18, TI19, TL15, TM01, TM02, TM03, TM04, TM06, TM07, TM08, TM09, TM10, TM11, TM12, TM13, TM14, TM15, TM16, TM17, TM18, TM19, TM20, TS01, TS03, TS05, TS06, TS14, TW02, TW03, TW04, TW05, TW06, TW07, TW08, TW09, TW10, TW11, TW12, TW13, TW14, TW15, TW16, TW17, TW18, TW19, TW20.

Males (126): TB01, TB02, TB04, TB06, TB07, TB08, TB09, TB11, TB14, TB15, TB16, TB17, TB18, TC03, TC07, TC08, TC09, TC11, TC13, TC15, TC16, TC19, TC21, TE03, TE05, TE07, TE08, TE09, TE10, TE11, TE13, TE16, TE17, TE18, TF01, TF02, TF03, TF04, TF05, TF06, TF08, TF09, TF13, TF14, TF15, TF16, TF18, TF20, TF21, TF22, TF24, TF25, TF26, TH02, TH04, TH05, TH06, TH11, TH12, TH14, TH16, TH21, TH22, TH23, TH24, TI03, TI06, TI07, TI11, TI12, TI16, TI17, TI18, TI19, TI20, TI21, TI22, TI23, TL17, TM01, TM02, TM03, TM04, TM05, TM06, TM07, TM08, TM09, TM10, TM11, TM12, TM13, TM14, TM15, TM16, TM17, TM18, TM19, TM20, TM21, TS03, TS05, TS06, TS14, TS15, TS18, TW01, TW02, TW03, TW04, TW05, TW06, TW07, TW08, TW09, TW10, TW11, TW12, TW13, TW14, TW15, TW16, TW17, TW18, TW19, TW20.

These progenitors generated a total of 308 families with a mean of approximately 81 individuals per family. Among environments, the distribution of quantities was similar, as shown in Supplementary Fig. 1 and Supplementary Table 1.



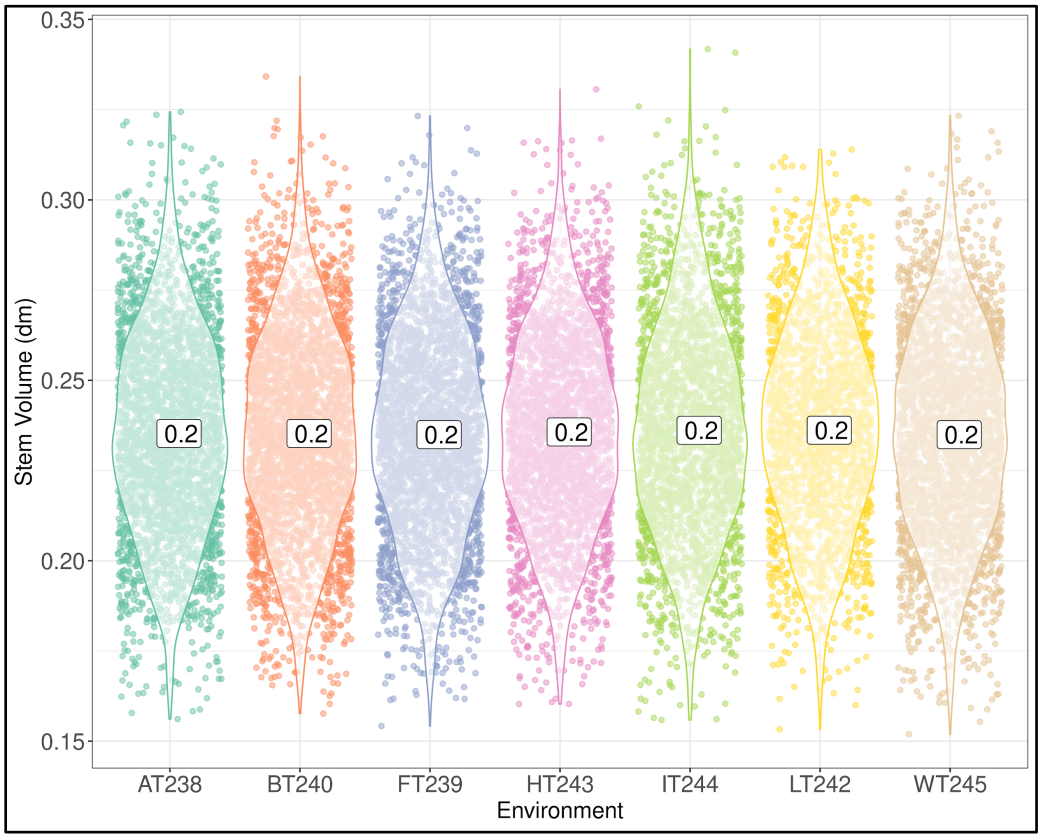
Supplementary Fig. 1. Distribution of quantities of individuals per family across environments. The mean is shown in the centre of the violin plots.

Supplementary Table 1. Quantities of families, individuals and progenitors per environment.

	Individuals	**Families**
AT23804	3,588	307
BT24004	3,600	307
FT23904	3,545	307
HT24304	3,588	308
IT24404	3,588	307
LT24204	3,600	307
WT24504	3,571	307
All	25,080	308
Filler/CCK	108	-

2. Phenotypic Analyses

Volume (V) measures presented significant differences among environments; however, significant differences were not observed after correction (Supplementary Fig. 2). Using a Tukey test of multiple comparisons, we separated the measures into groups: (a) BT240, FT239 and WT245; (ab) AT238 and IT244; (b) HT243; and (c) LT242.



Supplementary Fig. 2. Distribution of volume (V) measures after correction separated according to the seven environments used (ANOVA p-value = 0.469).

Supplementary Table 2. Posterior estimates for individual-tree heritability and dominance (± standard error) for volume.

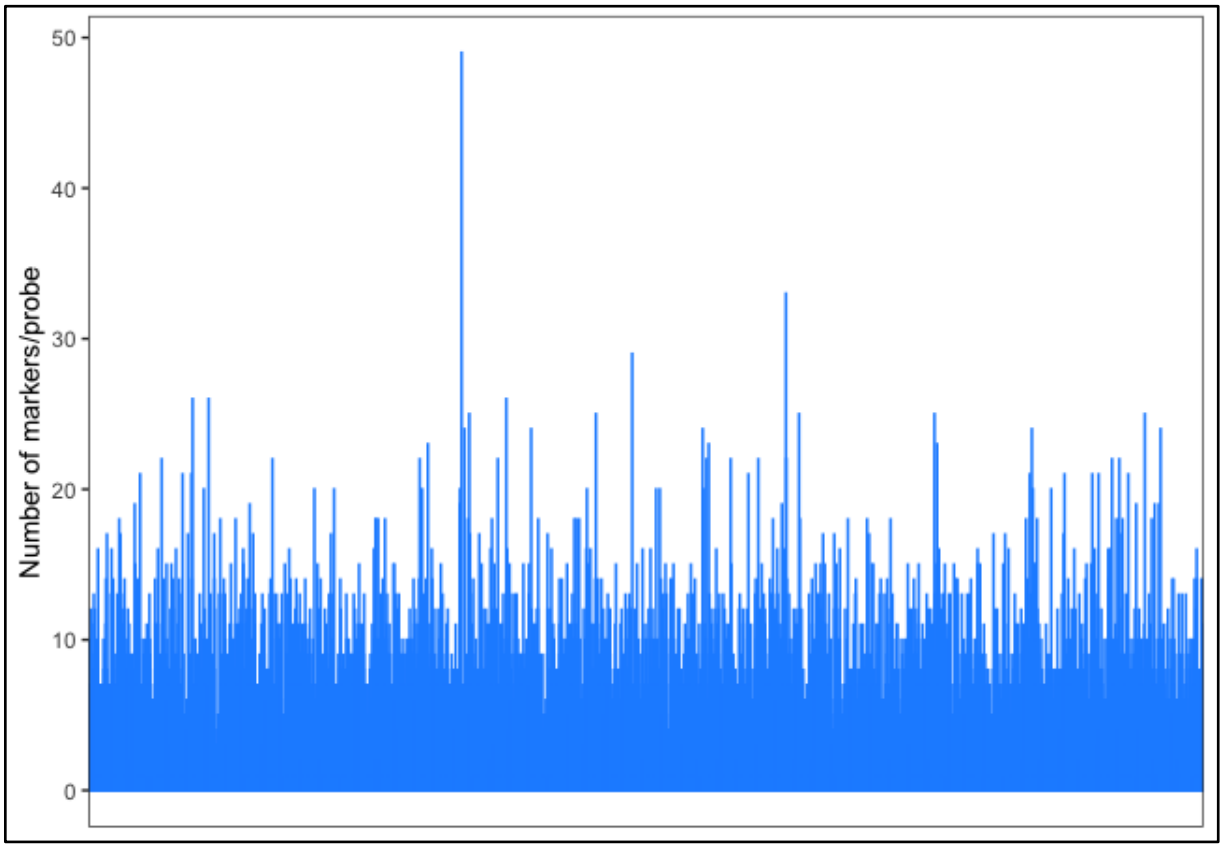
Environment	Individual-tree Heritability	Dominance Ratio
AT238	0.22 (± 0.04)	0.20 (± 0.06)
BT240	0.22 (± 0.04)	0.14 (± 0.05)
FT239	0.18 (± 0.04)	0.15 (± 0.05)
HT243	0.17 (± 0.04)	0.11 (± 0.05)
IT244	0.32 (± 0.05)	0.13 (± 0.06)
LT242	0.18 (± 0.04)	0.11 (± 0.07)
WT245	0.22 (± 0.04)	0.12 (± 0.05)

3. Genotypic Analyses

For genotyping, 1,692 individuals from four environments (AT238, BT240, HT243, and IT244) and 45 families (TC01 x TC11, TC01 x TC16, TC02 x TC07, TC02 x TC09, TC02 x TC13, TC02 x TC21, TC03 x TC11, TC08 x TC16, TC09 x TC07, TC09 x TC11, TC10 x TC08, TC10 x TC13, TC10 x TC15, TC11 x TC07, TC11 x TC13, TC18 x TC03, TC18 x TC07, TC18 x TC09, TC18 x TC11, TC21 x TC07, TC21 x TC13, TC21 x TC15, TE05 x TE16, TE06 x TE07, TE06 x TE08, TE06 x TE11, TE06 x TE18, TE08 x TE09, TE09 x TE11, TE09 x TE13, TE10 x TE05, TE10 x TE16, TE10 x TE17, TE11 x TE07, TE11 x TE08, TE12 x TE09, TE12 x TE17, TE13 x TE16, TE17 x TE13, TE18 x TE07, TE19 x TE03, TE19 x TE05, TE19 x TE07, TE20 x TE05, and TE20 x TE10) were selected (Supplementary Table 2).

Supplementary Table 3. Quantities of families, individuals and progenitors per environment considering only the genotyped trees.

	Individuals	**Families**	**Mean Individuals Per Family**
AT23804	477	45	10.6
BT24004	490	45	10.9
HT24304	447	45	9.9
IT24404	278	29	9.6
All	1,692	45	37.6



Supplementary Fig. 3. Distribution of SNP markers across probes (x axis) used for sequencing.

4. RNA-Seq Assembly and Annotation

Supplementary Table 4. RNA-Seq experiments were selected for the developed study from the Sequence Read Archive (SRA) of the National Center for Biotechnology Information (NCBI).

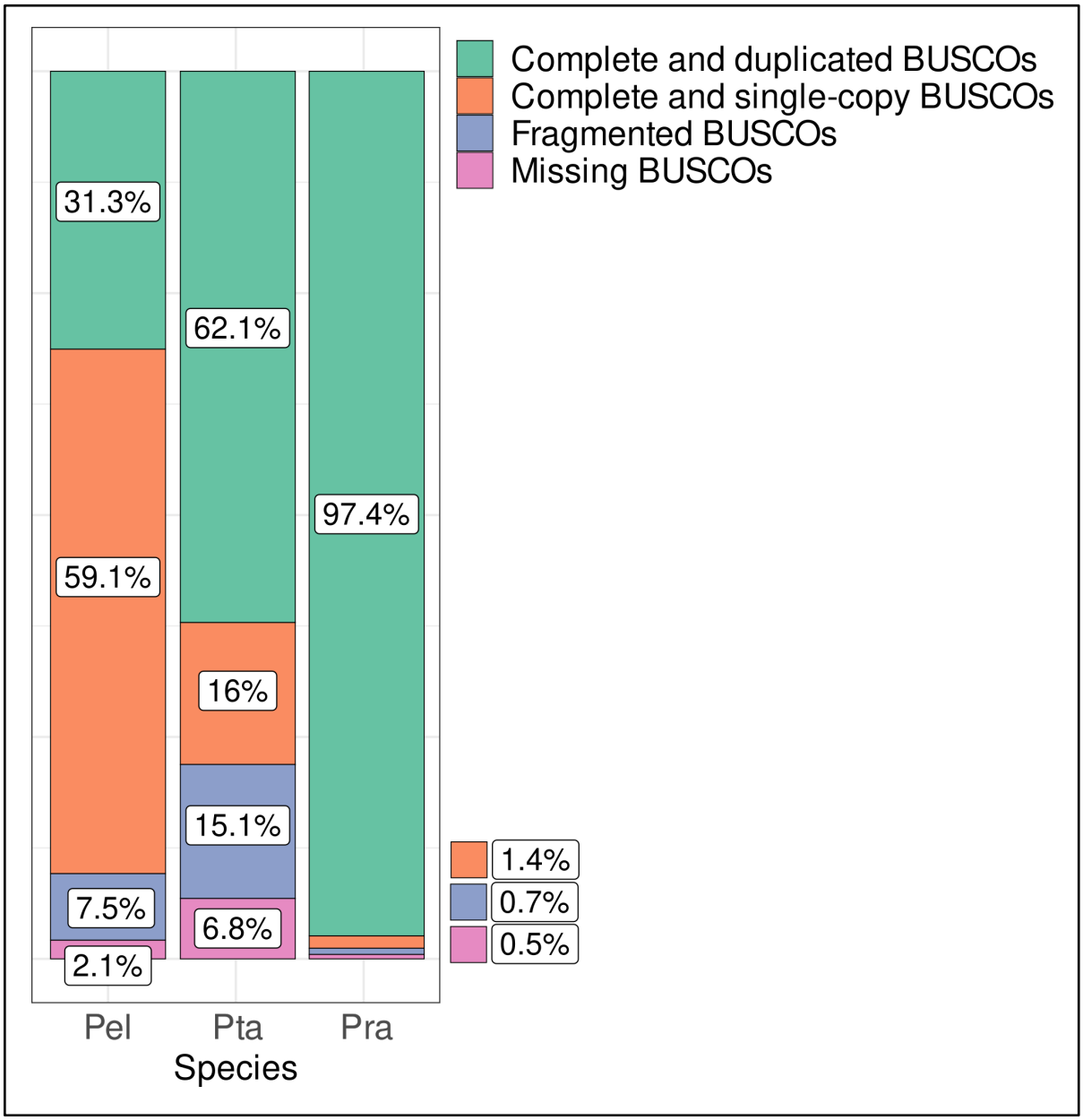
SRA	Age	Instrument	Organism	Tissue
SRR7589385	15 years	Illumina HiSeq 4000	*Pinus taeda*	Cambium
SRR7589386	15 years	Illumina HiSeq 4000	*Pinus taeda*	Cambium
SRR7589387	15 years	Illumina HiSeq 4000	*Pinus taeda*	Cambium
SRR9417402	16 years	Illumina HiSeq 2500	*Pinus elliottii*	Vascular cambium
SRR9417403	16 years	Illumina HiSeq 2500	*Pinus elliottii*	Vascular cambium
SRR9417404	16 years	Illumina HiSeq 2500	*Pinus elliottii*	Vascular cambium
SRR9417405	16 years	Illumina HiSeq 2500	*Pinus elliottii*	Vascular cambium
SRR9417406	16 years	Illumina HiSeq 2500	*Pinus elliottii*	Vascular cambium
SRR9417407	16 years	Illumina HiSeq 2500	*Pinus elliottii*	Vascular cambium
SRR9417408	16 years	Illumina HiSeq 2500	*Pinus elliottii*	Vascular cambium
SRR9417409	16 years	Illumina HiSeq 2500	*Pinus elliottii*	Vascular cambium
SRR9417386	16 years	Illumina HiSeq 2500	*Pinus elliottii*	Vascular cambium
SRR9417387	16 years	Illumina HiSeq 2500	*Pinus elliottii*	Vascular cambium

SRR9417388	16 years	Illumina HiSeq 2500	*Pinus elliottii*	Vascular cambium
SRR9417389	16 years	Illumina HiSeq 2500	*Pinus elliottii*	Vascular cambium
SRR9417390	16 years	Illumina HiSeq 2500	*Pinus elliottii*	Vascular cambium
SRR9417391	16 years	Illumina HiSeq 2500	*Pinus elliottii*	Vascular cambium
SRR9417392	16 years	Illumina HiSeq 2500	*Pinus elliottii*	Vascular cambium
SRR9417393	16 years	Illumina HiSeq 2500	*Pinus elliottii*	Vascular cambium
SRR3111638	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3113360	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3130544	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3111615	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3113316	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3130546	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3111624	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3113338	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3130548	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3123622	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3123623	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3123625	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3123626	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3123627	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3123642	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3122144	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3122249	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3122250	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3123366	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3123370	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3123374	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3123606	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3123607	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3123610	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3123611	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3123612	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3123613	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3111571	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3113314	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3130567	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3111627	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3113349	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem

SRR3130550	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3111576	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3113315	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3130551	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3111623	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3113327	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3130553	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3123614	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3123615	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3123616	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3123617	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3123619	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3123620	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3130556	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3130558	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3122248	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3123334	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3123336	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3123368	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3123376	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3123378	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3130559	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3130561	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3130562	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3123350	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3123352	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3123358	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3123379	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3123383	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3123386	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3130563	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3130564	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3130566	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3111565	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3113312	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem
SRR3130568	17 months	Illumina Genome Analyzer IIx	*Pinus radiata*	Stem

Supplementary Table 5. *De novo assembly* Trinity statistics for each species.

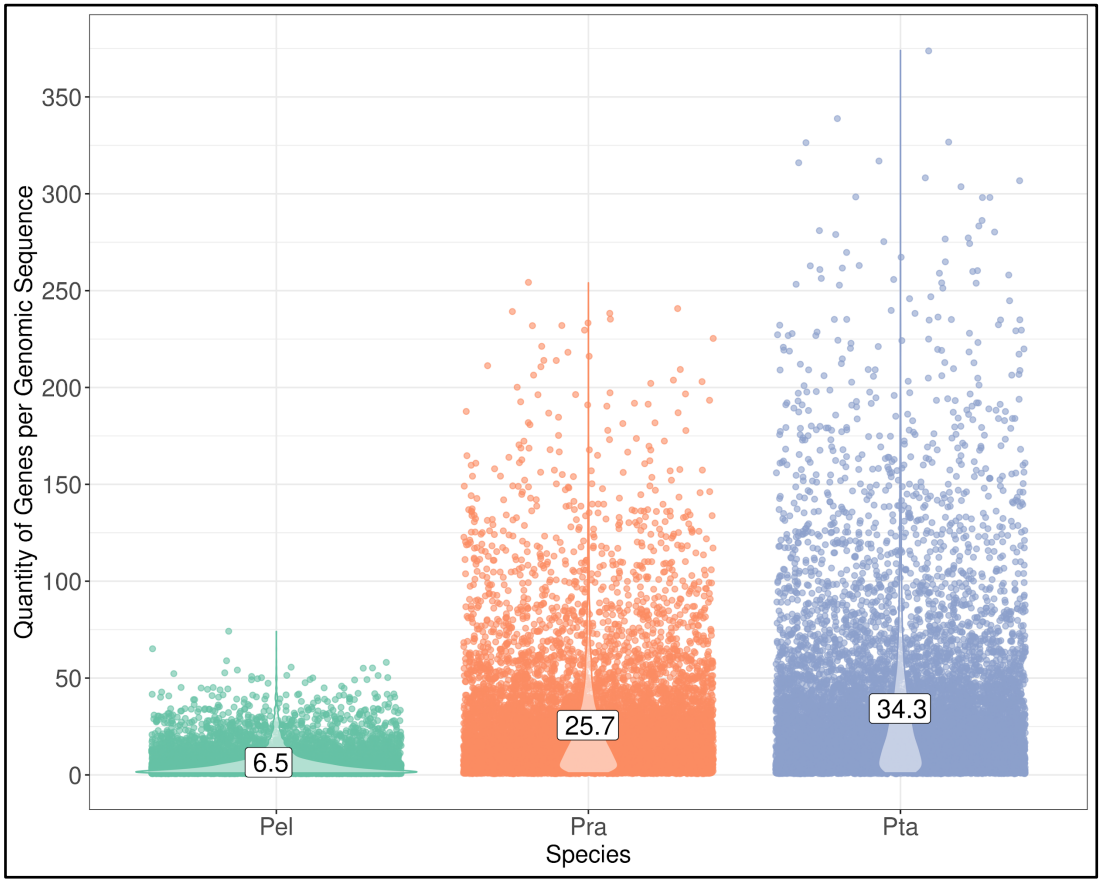
Species ID	***Pinus taeda***	***Pinus radiata***	***Pinus elliottii***
Illumina sequencing raw reads	60,899,426	482,672,571	278,299,413
Reads maintained after filtering	51,121,782	455,253,550	276,114,142
Total assembled bases	190,245,856	343,900,274	102,646,851
Total isoforms	146,652	271,721	96,424
Total genes	75,897	137,103	50,781
Median contig length	772 bp	849 bp	789 bp
Average contig length	1,297.26 bp	1,265.64 bp	1,064.54 bp
Contig N50	2071	1921	1499



Supplementary Fig. 4. Total groups of orthologues inferred from the BUSCO search of plant databases for *Pinus elliottii* (Pel), *Pinus taeda* (Pta) and *Pinus radiata* (Pra).

Supplementary Table 6. Retrieved annotations from transcriptomes of *Pinus elliottii*, *Pinus taeda* and *Pinus radiata* against the SwissProt database and the correspondences considering Gene Ontology (GO) terms and Enzyme Commission (EC) numbers.

Species	Isoforms (Gene Correspondences)	Isoforms with Correspondences in SwissProt (Genes)	Isoforms with GO Annotations (Genes)	Isoforms with EC Numbers (Genes)
P. elliottii	96,424 (50,781)	47,634 (23,973)	46,804 (23,565)	14,251 (7,115)
P. radiata	271,721 (137,103)	136,234 (57,090)	133,917 (56,176)	38,468 (15,661)
P. taeda	146,652 (75,897)	77,897 (33,221)	76,592 (32,567)	23,583 (10,016)

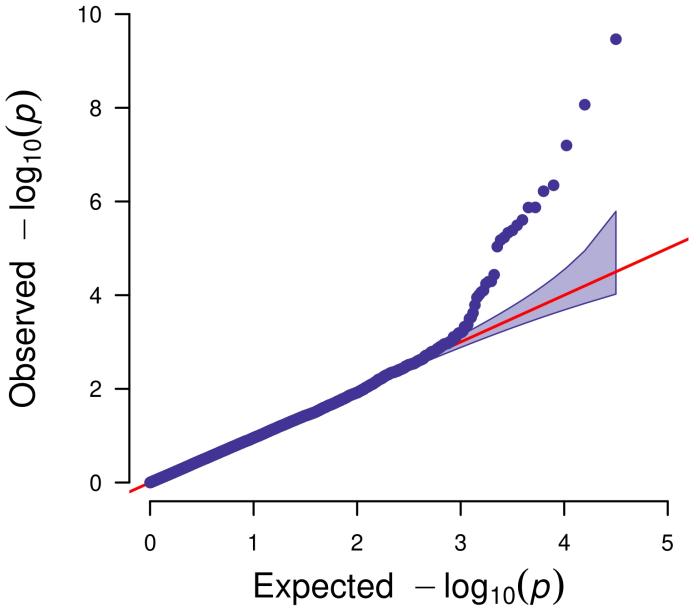


Supplementary Fig. 5. Distribution of genes (assembled from RNA-Seq experiments) per genomic sequence (scaffolds with probes) according to comparative alignments. Genes from the transcriptomes of *Pinus elliottii*, *Pinus radiata* and *Pinus taeda* were used.

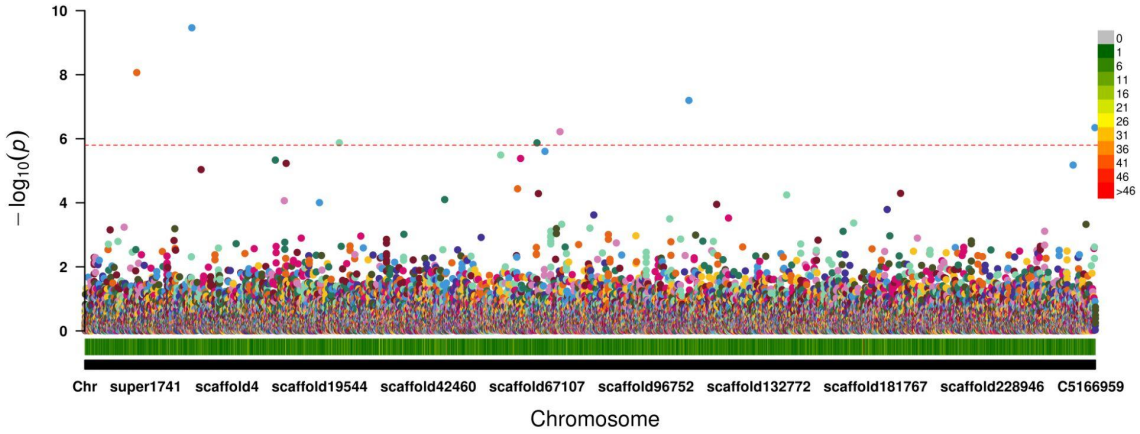
5. Genome-Wide Association Analyses

Supplementary Table 7. Significant SNP markers were found to be associated with volume phenotypes used: volume (V). The model used was FarmCPU, and no covariates were included in the genome-wide association analysis.

Phenotype	Sequence	Position	P-value
V	super1496	457445	8.60e-09
V	super3461	42253	3.43e-10
V	scaffold23393	29322	1.34e-06
V	scaffold67087	208847	1.35e-06
V	scaffold73604	618978	6.04e-07
V	scaffold109055	54249	6.36e-08
V	C5175585	188933	4.51e-07



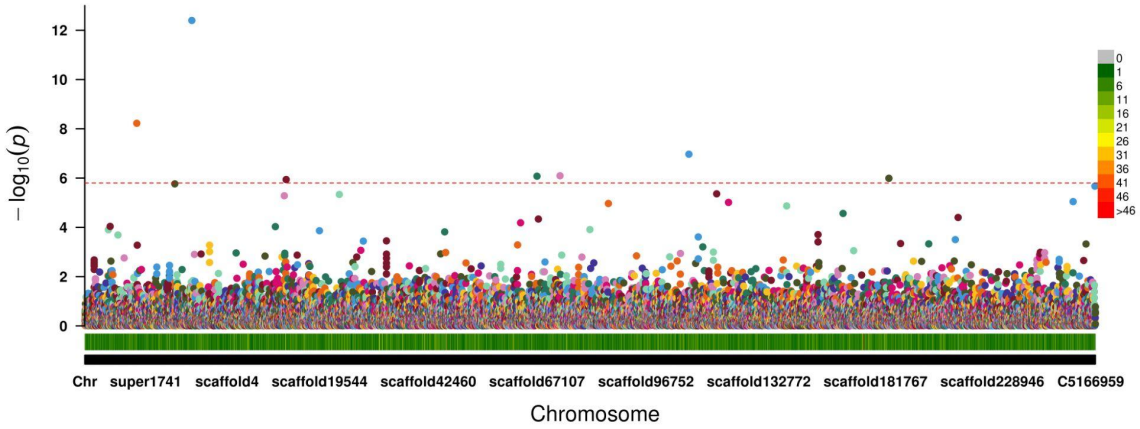
Supplementary Figure 6. Quantile-quantile plots for the FarmCPU models created for the identification of SNP associations with volume without covariates.



Supplementary Figure 7. Manhattan plots for the p-values obtained from the association tests (FarmCPU models without covariates) performed for the identification of SNP associations with volume. Each point in the scatter plot denotes a SNP coloured according to the scaffold on which it is located.

Supplementary Table 8. Significant SNP markers found to be associated with volume. The model used was FarmCPU, and the environmental information was included as a covariate in the genome-wide association analysis. Annotations of these genomic regions were performed with genes assembled from the transcriptomes of *Pinus elliottii* (Pel), *Pinus radiata* (Pra) and *Pinus taeda* (Pta).

PH	Sequence	Position	P-value	Annotation (Enzyme/Gene Ontology)
SV	super1496	457445	5.96e-09	(i) Chloroplast unusual positioning protein, putative, expressed; (ii) Hydroxyproline-rich glycoprotein family protein
SV	super3461	42253	4.00e-13	(i) Cystathionine beta-synthase (CBS) protein; (ii) CBS domain containing membrane protein, putative, expressed
SV	scaffold11937	131008	1.15e-06	alpha/beta-Hydrolases superfamily protein
SV	scaffold67087	208847	8.39e-07	
SV	scaffold73604	618978	8.05e-07	
SV	scaffold109055	54249	1.08e-07	
SV	scaffold187212	226376	1.02e-06	



Supplementary Figure 8. Manhattan plots for the p-values obtained from the association tests (FarmCPU models with environment as covariates) performed for the identification of SNP associations with volume. Each point in the scatter plot denotes a SNP coloured according to the scaffold on which it is located.

Supplementary Table 9. SNPs found in linkage disequilibrium ($r^2 \geq 0.5$) with the markers identified by the association tests against volume (V). Annotations of these genomic regions were performed with genes assembled from the transcriptomes of *Pinus elliottii* (Pel), *Pinus radiata* (Pra) and *Pinus taeda* (Pta).

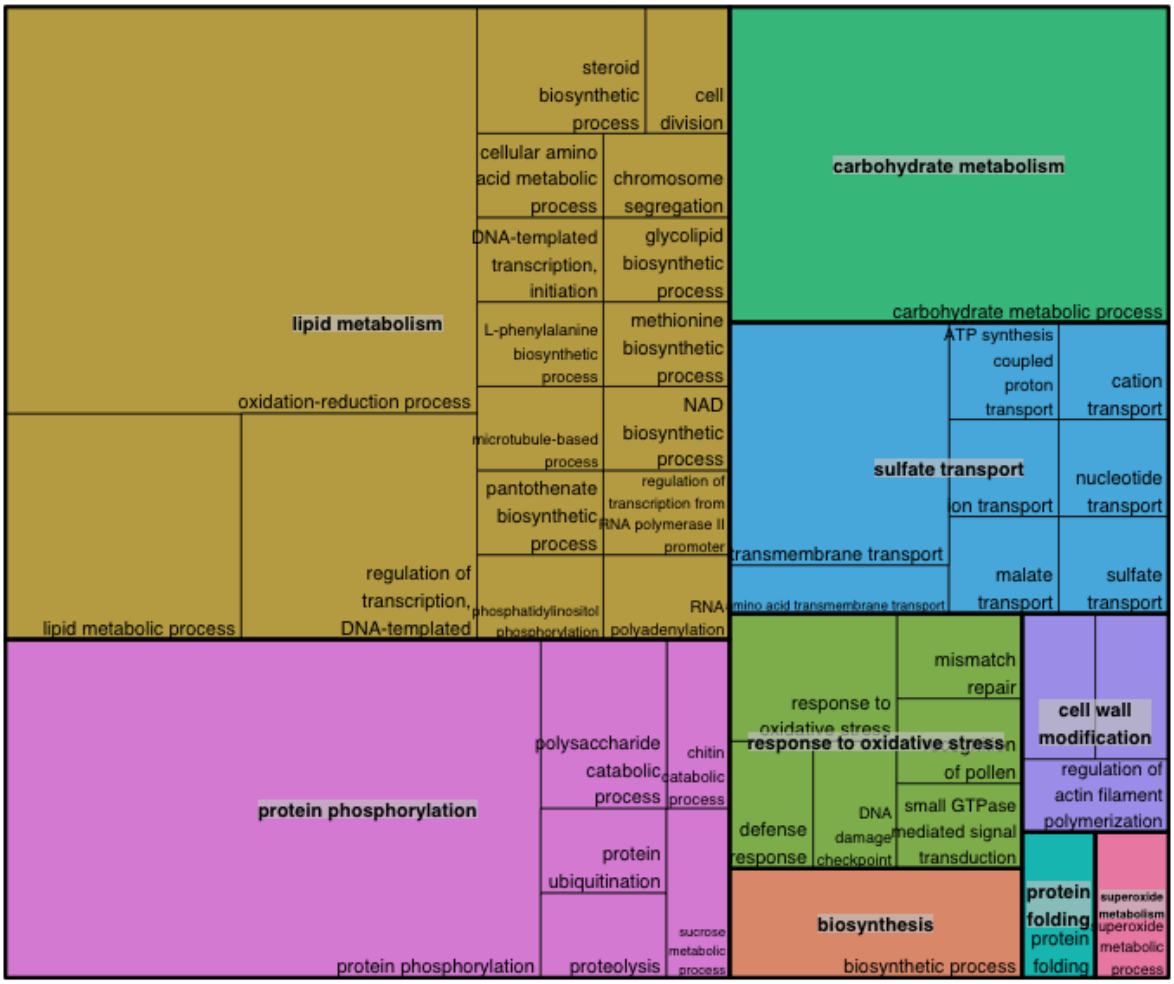
PH	Sequence	Position	Annotation (Enzyme/Gene Ontology)
SV	super1496	457298	Hydroxyproline-rich glycoprotein family protein
SV	super1496	457450	(i) Chloroplast unusual positioning protein, putative, expressed; (ii) Hydroxyproline-rich glycoprotein family protein
SV	scaffold67087	208715	
SV	scaffold67087	208932	Uncharacterized protein At4g04980
SV	scaffold73604	618877	

Supplementary Figure 9. Linkage disequilibrium network constructed for the SNPs found in linkage disequilibrium ($R^2 \geq 0.5$) with the markers. Each point labelled in the network represents a SNP.

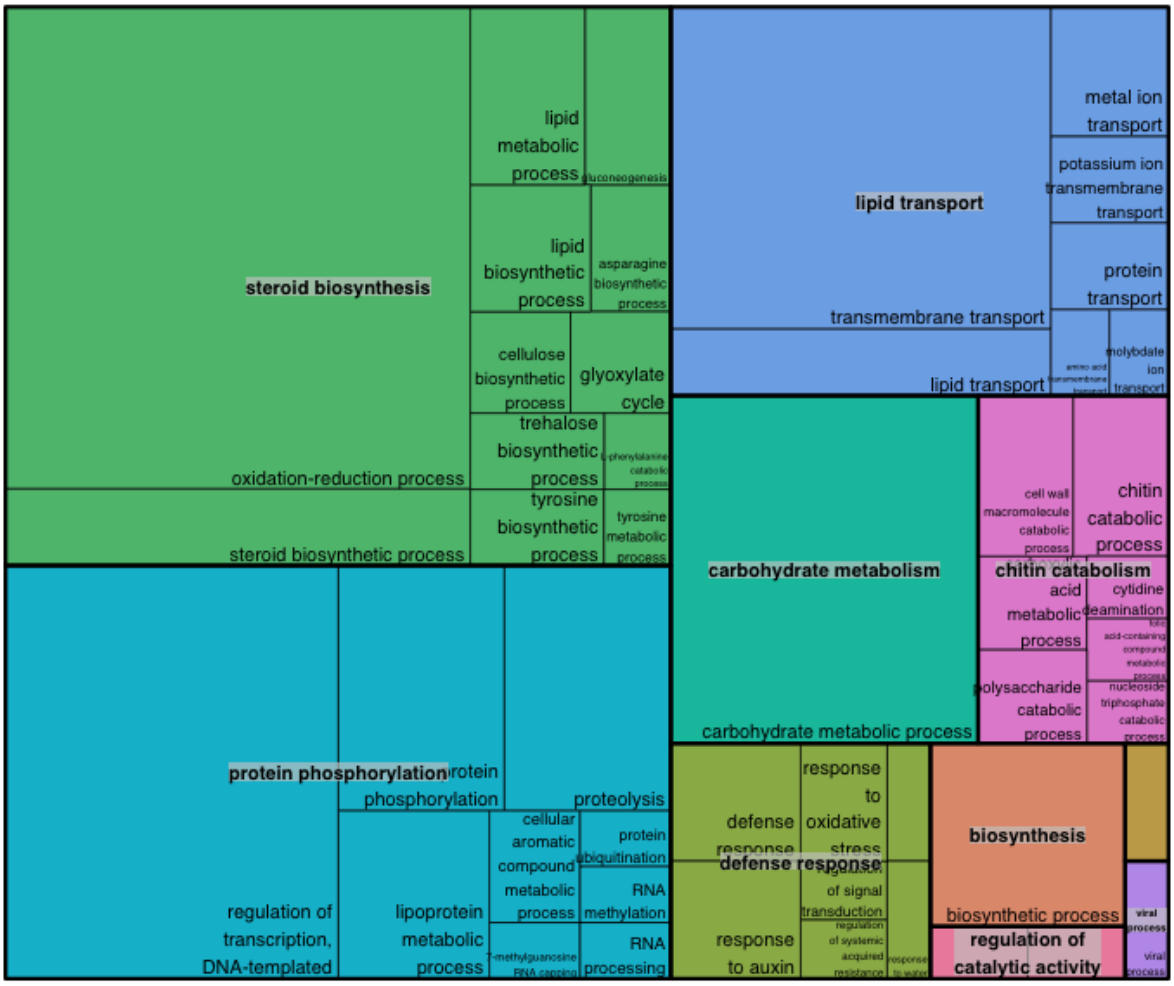
6. Co-expression Networks

Supplementary Table 10. Gene Ontology (GO) terms detected for genes associated with volume (V), identified through co-expressed genes from the assembled transcriptomes of *Pinus elliottii* (Pel) and *Pinus radiata* (Pra).

Phenotype	GO Terms
Pel_V(148 GO terms)	GO:0016887, GO:0030896, GO:0006457, GO:0004674, GO:0051539, GO:0004664, GO:0004568, GO:0022857, GO:0006811, GO:0008272, GO:0016702, GO:0030983, GO:0003864, GO:0015299, GO:0006979, GO:0004602, GO:0005515, GO:0003854, GO:0030833, GO:0015743, GO:0016779, GO:0030244, GO:0000077, GO:0005524, GO:0004553, GO:0003871, GO:0005094, GO:0030599, GO:0042626, GO:0016760, GO:0008134, GO:0007264, GO:0005669, GO:0004650, GO:0046854, GO:0006357, GO:0043631, GO:0006352, GO:0004652, GO:0016747, GO:0016161, GO:0016762, GO:0006032, GO:0003333, GO:0010333, GO:0004672, GO:0016157, GO:0004386, GO:0009094, GO:0016998, GO:0016616, GO:0006801, GO:0045735, GO:0005975, GO:0016829, GO:0034314, GO:0033926, GO:0005092, GO:0006812, GO:0006073, GO:0016853, GO:0004601, GO:0046983, GO:0048544, GO:0008270, GO:2001070, GO:0015629, GO:0006952, GO:0015940, GO:0006862, GO:0016788, GO:0008168, GO:0046872, GO:0006629, GO:0006694, GO:0003677, GO:0007017, GO:0007059, GO:0005471, GO:0005737, GO:0030286, GO:0046527, GO:0004842, GO:0009435, GO:0051301, GO:0005509, GO:0009247, GO:0016491, GO:0004857, GO:0016846, GO:0006520, GO:0009086, GO:0005516, GO:0042578, GO:0020037, GO:0005634, GO:0043565, GO:0004556, GO:0004806, GO:0008171, GO:0016592, GO:0015116, GO:0016787, GO:0003723, GO:0030170, GO:0016705, GO:0009058, GO:0043531, GO:0000272, GO:0006298, GO:0042545, GO:0005525, GO:0016020, GO:0006468, GO:0016021, GO:0006508, GO:0005985, GO:0003924, GO:0016773, GO:0000940, GO:0016209, GO:0016567, GO:0055085, GO:0005885, GO:0003824, GO:0016765, GO:0005506, GO:0005216, GO:0032977, GO:0008017, GO:0004523, GO:0006355, GO:0055114, GO:0048046, GO:0003676, GO:0015986, GO:0005876, GO:0009250, GO:0005992, GO:0000786, GO:0050662, GO:0006813, GO:0008987, GO:0003712, GO:0008289, GO:0005618, GO:0016758, GO:0004252
Pra SV (139 GO terms)	GO:0008234, GO:0005506, GO:0008017, GO:0006355, GO:0004665, GO:0055114, GO:0005992, GO:0043666, GO:0000786, GO:0042157, GO:0004866, GO:0004506, GO:0050662, GO:0008061, GO:0004126, GO:0008289, GO:0016758, GO:0004252, GO:0043531, GO:0045095, GO:0000272, GO:0016020, GO:0003700, GO:0008610, GO:0047429, GO:0006468, GO:0016021, GO:0016032, GO:0006508, GO:0016773, GO:0006855, GO:0016567, GO:0055085, GO:0003824, GO:0009143, GO:0006094, GO:0004474, GO:0009055, GO:0006725, GO:0020037, GO:0005634, GO:0004864, GO:0043565, GO:0030001, GO:0005680, GO:0004806, GO:0005576, GO:0008171, GO:0006760, GO:1902600, GO:0009341, GO:0019752, GO:0030170, GO:0016705, GO:0009058, GO:0006529, GO:0006694, GO:0050661, GO:0003677, GO:0004869, GO:0010112, GO:0050660, GO:0009972, GO:0005507, GO:0050790, GO:0004411, GO:0009452, GO:0006097, GO:0009678, GO:0004499, GO:0004842, GO:0003777, GO:0005509, GO:0009966, GO:0004451, GO:0016491, GO:0004970, GO:0015079, GO:0071805, GO:0016853, GO:0008198, GO:0046983, GO:0004601, GO:0008270, GO:0004066, GO:0006952, GO:0071949, GO:0016788, GO:0008168, GO:0030246, GO:0046872, GO:0006629, GO:0008977, GO:0006396, GO:0006559, GO:0004197, GO:0016747, GO:0016161, GO:0006032, GO:0003333, GO:0004672, GO:0007018, GO:0006570, GO:0004150, GO:0015098, GO:0016998, GO:0010309, GO:0045735, GO:0016616, GO:0008173, GO:0006571, GO:0005975, GO:0004612, GO:0004427, GO:0003950, GO:0030244, GO:0005524, GO:0004553, GO:0016757, GO:0042626, GO:0016760, GO:0015689, GO:0015031, GO:0009415, GO:0001510, GO:0008233, GO:0016887, GO:0015297, GO:0004568, GO:0022857, GO:0016614, GO:0016702, GO:0004565, GO:0009733, GO:0016714, GO:0006979, GO:0006869, GO:0003854, GO:0005515



Supplementary Figure 10. REVIGO summarization of Gene Ontology (GO) terms for genes associated with volume found in the *Pinus elliottii* transcriptome.



Supplementary Figure 11. REVIGO summarization of Gene Ontology (GO) terms for genes associated with volume found in the *Pinus radiata* transcriptome.

6. *CAPÍTULO II*: ARTIGO SOBRE RESPOSTA A FRIO EM SERINGUEIRA

Novel insights into cold resistance in *Hevea brasiliensis* through coexpression networks

Carla Cristina Da Silva[1], Stephanie Karenina Bajay[1], Alexandre Hild Aono[1], Felipe Roberto Francisco[1], Ramir Bavaresco Junior[1], Anete Pereira de Souza[1,2], Camila Campos Mantello[3], Renato Vicentini dos Santos[4]*

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

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

[3]Hazera Seeds B. V., Warmenhuizen, The Netherlands

[4]Department of Genetics, Evolution, Microbiology and Immunology, Biology Institute, University of Campinas (UNICAMP), Campinas, SP, Brazil

*** Correspondence:**
Renato Vicentini dos Santos
shinapes@unicamp.br

Keywords: Breeding Strategies, Cold Stress, Edaphoclimatic Adaptation, Heuristic Cluster Chiseling Algorithm, RNA-Seq, Rubber Tree, Stress Gene Modules

Abstract

Hevea brasiliensis, a tropical tree species from the Amazon rainforest, is the main world source of natural rubber. Due to the high pressure of fungal diseases in hot humid regions, rubber plantations have been moved to “escape areas”, which are dryer and present lower temperatures during the winter. We carried out a study with a primary (GT1) and a secondary (RRIM 600) young rubber tree clones, which present different cold tolerance strategies, to analyze gene expression regulation while under 24 hours of cold exposure (10°C). Together with traditional differential expression approaches, a RNA-seq gene co-expression network (GCN) was established with 27,220 genes that were grouped into 205 gene clusters. Through the GCN, we could relate the predominance of

rubber tree molecular responses on cold stress to 31 clusters, which were divided in three GCN modules: a down-regulated group with 16 clusters and two up-regulated groups with twelve and three clusters, respectively. The hub genes of the cold-responsive modules were identified and analyzed as well. Each genotype has a well defined group of gene clusters with little occurrence of both clones in the same cluster. The GT1 gene clusters were enriched mostly for cell wall biogenesis, furthermore there are also several DEGs present in the photosynthesis clusters. RRIM600 clone clusters were enriched for protein translation, with DEGs located in clusters related to stress response and respiration. As a result of the GCN strategy applied in this study, we could not only access single DEGs related to *Hevea* cold responses, but also provide insights into a deeper cascade of associated mechanisms involved in the response to cold stress in young rubber trees. Our results may represent the species genetic stress responses developed during its evolution since the varieties chosen for this work are genotypes selected during the early years of rubber tree domestication. The understanding of *H. brasiliensis* cold response mechanisms can greatly improve the breeding strategies for this crop that has such a narrow genetic base, is being impacted by climate change and is the only source for large scale rubber production.

1 Introduction

Hevea brasiliensis, mainly known as rubber tree, is an allogamous perennial tree species native from the Amazon Rainforest, belonging to the Euphorbiaceae family (Priyandarshan and Clément-Demange, 2004). Although rubber trees are the main world source of natural rubber (Pootakham et al., 2017; Gonçalves and Fontes, 2009; De Faÿ and Jacob, 1989), these tree species are a recent crop, being still in domestication (Priyadarshan and Clément-Demange, 2004). The dispersal and domestication of rubber trees around the world was based on only about 20 seedlings that were introduced in Southeast Asia in the late 19th century. These seedlings were the only survivors of a collection of thousands of seeds from the Amazon Basin, originating the selection of elite trees and controlled hybridization for reproduction in Hevea. Thus, until the present day, almost all commercial varieties of *H. brasiliensis* planted are derived from these seedlings, therefore, their genetic variability is quite narrow (Gonçalves and Fontes, 2009; Souza et al., 2015).

Although the Amazon Basin offers ideal conditions for rubber tree cultivation, the occurrence of the fungal South American Leaf Blight (SALB) disease in the region hinders the establishment of rubber tree plantations in the area. To overcome this situation, Brazilian rubber

tree plantations were moved to sub-optimum or escape areas such as the Brazilian Center and South East regions, which are dryer and present lower temperatures during the winter (Priyadarshan et al., 2009). The South East Asian countries, which have similar Amazonian climate conditions and successfully avoided the introduction of SALB, are the major rubber producers worldwide. Nevertheless, these Asian plantations have already suffered heavy losses because of other fungal disease outbreaks (Priyadarshan and Gonçalves, 2003; IRCo, 2019; Pornsuriya et al., 2020), being expanded to suboptimal areas as well, such as Southern China and Northeast India (Priyadarshan et al., 2009). Thus, rubber tree varieties adapted to areas with new edaphoclimatic conditions are of key interest.

Among the stressing factors characteristic of the escape areas, low temperatures heavily affect *H. brasiliensis* development, causing damage to leaves and latex production (Priyadarshan et al., 2005). For a rubber tree plantation, a prolonged low temperature period or a frost occurrence may cause the plants' death (Priyadarshan et al., 2009). Cold stress triggers a massive reprogramming of plants' gene expression to adjust their metabolic processes to cope with the cold environment (Theocharis et al., 2012). In order to promote breeding towards cold tolerance in *H. brasiliensis*, a deeper understanding of the species cold response mechanisms is necessary.

In the last years, due to the impact of low temperatures in rubber tree plantations and rubber production, there is an increase of studies about *H. brasiliensis* molecular cold stress responses, especially through comparative transcriptomics (Deng et al., 2018, Gong et al., 2018, Mantello et al., 2019). The modification of Hevea gene expression under cold treatments has already enabled the identification of thousands of differentially expressed genes (DEGs) (Mohamed Sathik et al, 2018; Deng et al., 2018; Cheng et al., 2018). However, despite these comparative transcriptome analysis between resistant and cold-sensitive clones show that cold is a key factor in latex production, more precise information on the main biological pathways and metabolic processes involved in the cold response has not yet been elucidated in rubber trees. Recently, Ding et al. (2020) modeled a gene coexpression network (GCN) and associated a network module with cold treatment. Although the cold response mechanisms were not assessed, the GCN created represented a rich source of data for investigating cold response.

Here we carried out a study based on a GCN modeled with transcriptomic data from two of the earliest *H. brasiliensis* clones to identify novel patterns between the genes involved in *H.*

brasiliensis response to cold stress, aiming to supply a deeper insight on how young rubber tree plants cope with low temperatures. The use of the transcriptome from a cold experiment with two genotypes, one resistant and one susceptible, allowed the identification of central genes in the created networks underlying cold stress in rubber trees. Thus, this study is one of the pioneers in the association of molecular mechanisms of Hevea through GCNs, being the first one to unveil rubber tree cold stress responses in important commercial varieties.

2 Material and Methods

2.1 RNA-Seq Experiment

In this work, two genotypes of *H. brasiliensis* recommended for escape areas (Mantello et al., 2019), GT1 and RRIM600, were evaluated. Under low temperatures, RRIM 600 presents an avoidance physiological strategy while GT1 balances between photosynthetic/growth activities and stress response (Mai et al., 2010; Mantello et al., 2019). The cold experiment consisted of exposing three six-months-old seedlings of each genotype to a temperature of 10ºC for 24 hours (12h light/12h dark). During this period, leaf material was collected at a time series of 0 hours (0h - control), 90 minutes (90m), 12 hours (12h), and 24 hours (24h). The RNA extracted from these samples was employed to construct cDNA libraries, and it was subsequently sequenced using an Illumina Genome Analyzer IIx with the TruSeq SBS 36-Cycle Kit (Illumina, San Diego, CA, USA) for 72 bp PE reads. The transcriptome assembled by Campos Mantello et al. (2019) was used for the analyzes of the present work.

2.2 Bioinformatics and Differential Gene Expression Analysis

The reads of each genotype evaluated with their respective times and replicates of treatment were mapped with the Bowtie2 aligner (Langmead & Salzberg, 2012) in the reference genome (Tang et al., 2016). The estimated abundance of reads for each sample was calculated by RSEM (Li & Dewey, 2011) and, subsequently, a matrix was constructed with expression values and abundance estimates between samples and replicates. Genes with a minimum of 10 counts per million (CPM) in at least three samples were maintained for further analyses. These samples were normalized by the quantile method and transformed into log2 counts per million (log2cpm) by the edgeR package (Robinson et al., 2010).

From the normalized data, a principal component analysis (PCA) was performed using the R software, to explore the data and verify the behavior of the samples. For DEG analysis, the datasets from GT1 and RRIM600 were combined in an attempt to create an unique *H. brasiliensis* expression dataset. An empirical Bayes smoothing method implemented in the edgeR package (Robinson et al., 2010) was used to identify DEGs between cold exposure sampling times, in which the following comparisons were performed: (i) 0h vs 90m; (ii) 0h vs 12h; (iii) 0h vs 24h; (iv) 90m vs 12h; (v) 90m vs 24h; and (vi) 12h vs 24h. The p-values were adjusted using a Bonferroni correction and genes with p-value ≤ 0.05 were considered DEGs.

Rubber tree DEGs were compared between the time-points evaluated with Venn plots constructed with the venn R package (Dusa, 2017). For all genes, we selected the respective annotation obtained with the Trinotate v2.0.1 pipeline (Bryant et al., 2017; https://trinotate.github.io/). Untranslated transcripts were searched against the SwissProt/UniProt database using BLASTX and filtered with an e-value of 1e-5 and placed into a tab-delimited file. Transcripts were also associated with Gene Ontology (GO) (Harris et al., 2004).

2.3 Gene Coexpression Network (GCN) Analysis

The GCN was inferred through R Pearson correlation coefficients with a cut-off of 0.8 (Van Noort et al., 2004). The R values of the GCN were normalized using the highest reciprocal rank (HRR) approach limited to the top 10 connections, and the heuristic cluster chiseling algorithm (HCCA) was used to organize the network into communities (Mutwil et al., 2010). The HRR networks and the HCCA results were visualized with Cytoscape software v.3.8.0 (Shannon et al., 2003).

2.4 Cold Stress Association

The number of DEGs in the clusters was used to pinpoint the most probable groups involved in *Hevea* cold stress response. Hub genes in the selected clusters were analysed using the node degree distribution (DD), betweenness centrality (BC) and stress centrality distribution (SC) parameters of the networks that were obtained using the Cytoscape software v.3.8.0 (Shannon et al., 2003). Genes with high DD that presented high values of BC and SC were considered highly influential in the selected clusters' co-expression dynamics. For each network cluster, at least one

hub gene was identified, and they were evaluated by their functional annotation and/or the genes they interacted with.

2.5 Functional Enrichment

Overrepresented GO terms in the clusters were determined with BiNGO plugin v.3.0.4 (Maere et al., 2005) using a customized reference set created from the transcriptome annotation data. A hypergeometric distribution and false discovery rate (FDR) < 0.05 were used in the analyses. GO enrichment analyses were performed with EnrichmentMap plugin v.3.3.1 (Merico et al., 2010) with FDR Q-values < 0.05 and the BiNGO output files. Both plugins were used in the Cytoscape software v.3.8.0 (Shannon et al., 2003).

3 Results

3.1 Bioinformatics and Differential Expression Analysis

Approximately 530 million paired-end (PE) reads were obtained for the RRIM600 genotype and 633 million for the GT1 genotype. After quality filtering, around 933 million PE reads were retained and used to assemble the transcriptome in the Trinity software. A total of 104,738 isoforms (51,697 genes) were identified with sizes between 500 bp and 22,333 bp, with an average of 1,874 bp and N50 of 2,369. From these genes, we could identify a total of 74,398 unique proteins related to 9,417 different GOs (Supplementary Table 1 - Access in https://docs.google.com/spreadsheets/d/1AkQICDVW60e6Ftrdfr8QjHVx814Ve22qGNvUKkjkBqI/edit?usp=sharing). For DEG analyses, we employed a filtering criteria based on genes that presented at least 10 counts per million (CPM) in at least 3 samples, which resulted in a final amount of 30,407 genes. From these genes, the PCA employed revealed a distinct profile of samples belonging to each different genotype (Figure 1A).

The number of DEGs per time period comparison showed different responses, regarding not only the number of DEGs (Table 1), but also the associated genes (Supplementary Table 2 – Access in https://docs.google.com/spreadsheets/d/1oGunmT7okMwkCnXkxvV1XkaTmVSoxjRc4z04hi4OLvI/edit?usp=sharing). Although there were evident intersections among the tests established (Figure 1B), it is clearly observed a joint mechanism of several genes for cold response. Regarding

the GO enrichment analysis, the full DEG set was enriched for response to stress and response to chitin terms. The down-regulated group of genes did not present any overrepresented category. In contrast, the up-regulated genes' group was enriched for the following terms: response to chitin, defense response, response to wounding, cell death, photoprotection, calcium ion transmembrane transporter activity and integral to plasma membrane.

Table 1. Number of *Hevea brasiliensis* differentially expressed genes (DEGs) that were modulated by cold stress.

Time period comparison	**DEGs**	**Up-regulated DEGs**	**Down-regulated DEGs**
0h vs 90m	32	32	0
0h vs 12h	673	558	115
0h vs 24h	1,208	852	356
90m vs 12h	526	455	71
90m vs 24h	917	700	217
12h vs 24h	18	15	3
(0h + 90min) vs (12h + 24h)	2,436	1,575	861

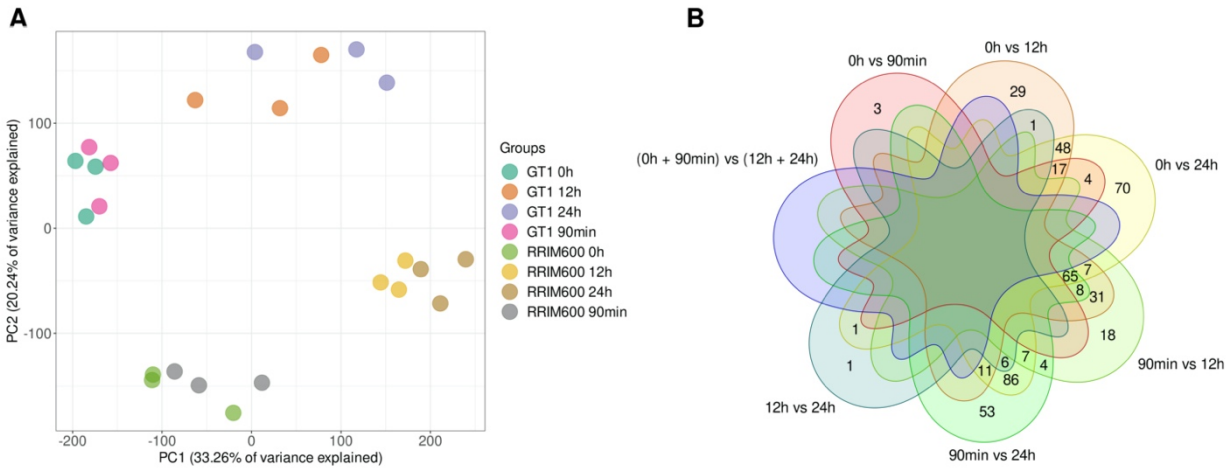


Figure 1. (A) Principal component analysis (PCA) scatter plot performed on the first two principal components (PCs) of rubber tree transcriptome data (GT1 and RRIM600 genotypes). Each point represents a different sample, colored according to the respective genotype. (B) Venn diagrams for differentially expressed genes (DEGs) found on the conditions established.

After 90 minutes of cold treatment, only thirty-two transcripts were characterized as DEGs, all of them were up-regulated during the plantlets' early response to cold stress. Of these, 26 genes were successfully annotated (Supplementary Table 1 - Access in https://docs.google.com/spreadsheets/d/1AkQICDVW60e6Ftrdfr8QjHVx814Ve22qGNvUKkjkBqI/edit?usp=sharing). The annotated DEGs are known to be induced by cold as well as other abiotic stressors, such as drought and salt stresses. They are involved with cell wall tightening processes, growth inhibition, oxidative-stress protection, calcium (Ca2+) signaling and cold-induced RNA processing. The late response, considered as the interval between twelve hours and 24 hours of cold treatment, contained 18 DEGs, including genes that stimulate an increase in the membrane diffusion barriers and cell wall modifications.

One hundred and thirty eight DEGs were annotated with the GO term DNA-binding transcription factor activity, being 103 up-regulated and 35 down-regulated. All of them, but one, were grouped in the cold stress response gene clusters (see below). An analysis of overrepresented GO terms in the 138 DEGs showed that they are involved with the regulation of cellular metabolic processes: negative and positive regulations are modulated by the up- and down-regulated groups

respectively. While the up-regulated genes are responsive to stress and other stimuli, the down-regulated genes promote plant development and growth.

Non-specific serine/threonine Protein kinases (EC: 2.7.11.1) were the most represented enzymes among the annotated DEGs: 165 were up-regulated while 57 were down-regulated. The up-regulated kinases have an overrepresentation of GO terms related to reproduction processes, response to ABA stimulus, calmodulin binding and signal transmission, while the down-regulated kinases were enriched for post-translational protein modification. The second most differentially expressed enzyme-types were RING-type E3 Ubiquitin Transferases (EC: 2.3.2.27) with 42 and 11 genes up- and down-regulated respectively.

3.2 Rubber Tree GCN Analysis

From the transcriptome data of both genotypes, we modeled an HRR network in order to represent the cold response of the species (Figure 2A). This network was composed of 27,220 nodes across 50,650 connections (edges). Only the top 10 Pearson R correlation coefficients considering a 0.8 cut-off were retained for each gene, being disconsidered the genes disconnected. The major connected component within the network is composed of 25,608 nodes (94.1%). There were also other 626 components for the remaining 1,592 genes, with a gene quantity inside each component ranging from 2 to 11 nodes. In order to identify putative gene associations within network structure, we analysed the network with the HCCA, identifying 832 clusters (Figure 2B). From these sets of genes, 626 corresponded to the disconnected components (mean cluster size of ~3), and 206 to the core GCN structure (mean cluster size of ~124), comprising clusters with sizes ranging from 45 to 280 genes (Supplementary Table 3 - Access in https://docs.google.com/spreadsheets/d/1lg0DBV6vXIO7zTrgqNdqpBoJJYSDlpurZ_3mUc5Kmr4/edit?usp=sharing).

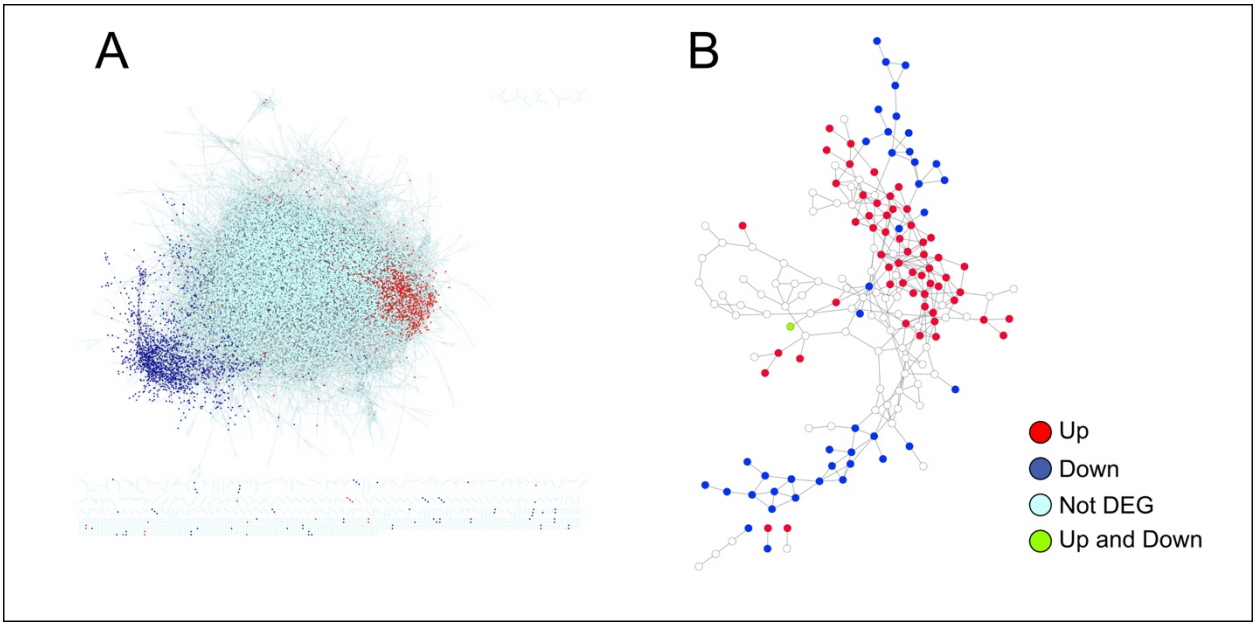


Figure 2. (A) Gene coexpression network (CGN) modeled; each node represents a gene colored according to the differential expression analysis, and each connection a minimum R Pearson correlation coefficient of 0.8 (corrected with the highest reciprocal rank approach). (B) GCN network contracted according to the groups identified by the heuristic cluster chiseling algorithm (HCCA); each node represents a cluster colored according to the presence of differentially expressed genes.

A functional coherence is expected across co-expressed genes belonging to the same cluster. As such, an enrichment analysis of GO terms was performed in each cluster of genes identified by the HCCA. From the 206 main clusters, 39 (14%) had GO terms that were significantly enriched, as highlighted in Figure 3 for biological process GO terms. Response to stress was the most overrepresented term among clusters, encompassing 1,039 genes across 9 groups, which was expected considering the cold experimental design. Two of these clusters (c14 and c28) were enriched for developmental processes as well, indicating the tight correlation between environmental stimulus and growth. Interestingly, we also observed seven clusters enriched for virus replication and transposition.

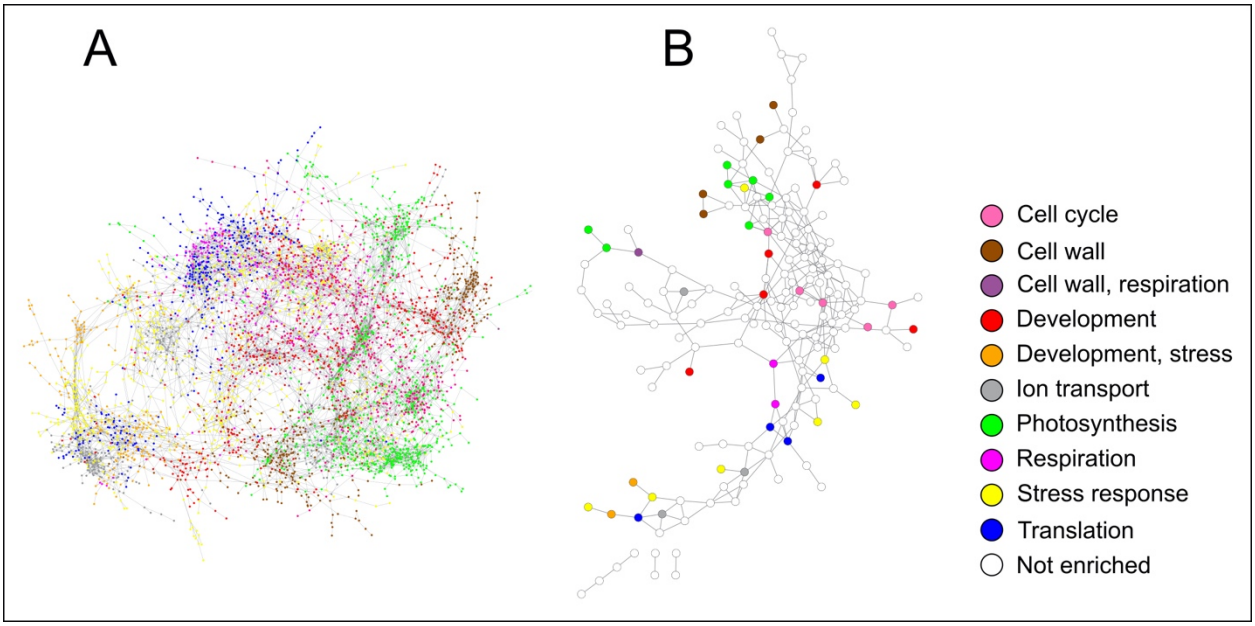


Figure 3. (A) Gene coexpression network (CGN) modeled; each node represents a gene colored according to the enriched gene ontology (GO) term found in the related cluster, and each connection a minimum R Pearson correlation coefficient of 0.8 (corrected with the highest reciprocal rank approach). (B) GCN network contracted according to the groups identified by the heuristic cluster chiseling algorithm (HCCA); each node represents a cluster colored according to the GO term enriched.

In addition to the clustering structure evaluated on the GCN topology, we also assessed the gene centrality measures (Supplementary Table 3 - Access in https://docs.google.com/spreadsheets/d/1lg0DBV6vXIO7zTrgqNdqpBoJJYSDlpurZ_3mUc5Kmr4/edit?usp=sharing). From the total of 27,220 genes inserted into the network, there were only 507 genes with the maximum permitted connections (ten). One of them, annotated as "probable cytochrome c biosynthesis protein" (CCBS), also presents the highest values of SC and BC centralities, indicating that this gene exerts great influence in the network. The gene belongs to cluster c8, which was enriched for photosynthesis, but also for cellular respiration and transposition. By evaluating the inter cluster connections in the HCCA-based contracted network (Figure 3B and Supplementary Table 4 - Access in https://docs.google.com/spreadsheets/d/1-peHUKs684MmvI74qUyinc7m9MF9QiyIqX_xHgmIK7U/edit?usp=sharing), we found that the cluster with the highest number of connections is c117 (nine edges), being enriched for developmental processes.

3.3 Gene Clusters Mostly Related to Cold Stress Response

Even though the DEGs identified were dispersed among the GCN (Figure 2A), specific patterns could be observed in several clusters. For identifying molecular associations with cold stress response, we selected the clusters that, together, account for 70.2% (1,751) of the DEGs identified. We considered as associated clusters those that contain at least 50% of their genes as DEGs (Supplementary Table 5 - Access in https://docs.google.com/spreadsheets/d/1LU2geL64NmrEPg88SwoTZ5JYPhevc72y32pD5UvVzCA/edit?usp=sharing), expanding the module across these clusters' neighbors considering a minimum required frequency of 20% of DEGs inside a cluster (Figure 4). Interestingly, we could establish three GCN modules for the 31 clusters selected: (i) a down-regulated group with 12 clusters, (ii) an up-regulated group with 11 clusters; and (iii) an up-regulated group with 3 clusters. Even though there were additional clusters containing DEGs, only the most pronounced ones were used for novel inferences. Among the DEGs in the selected clusters, 480 genes were annotated with enzyme code (EC) numbers, being 339 up-regulated and 141 down-regulated transcripts, and 180 DEGs were assigned to KEGG pathway maps.

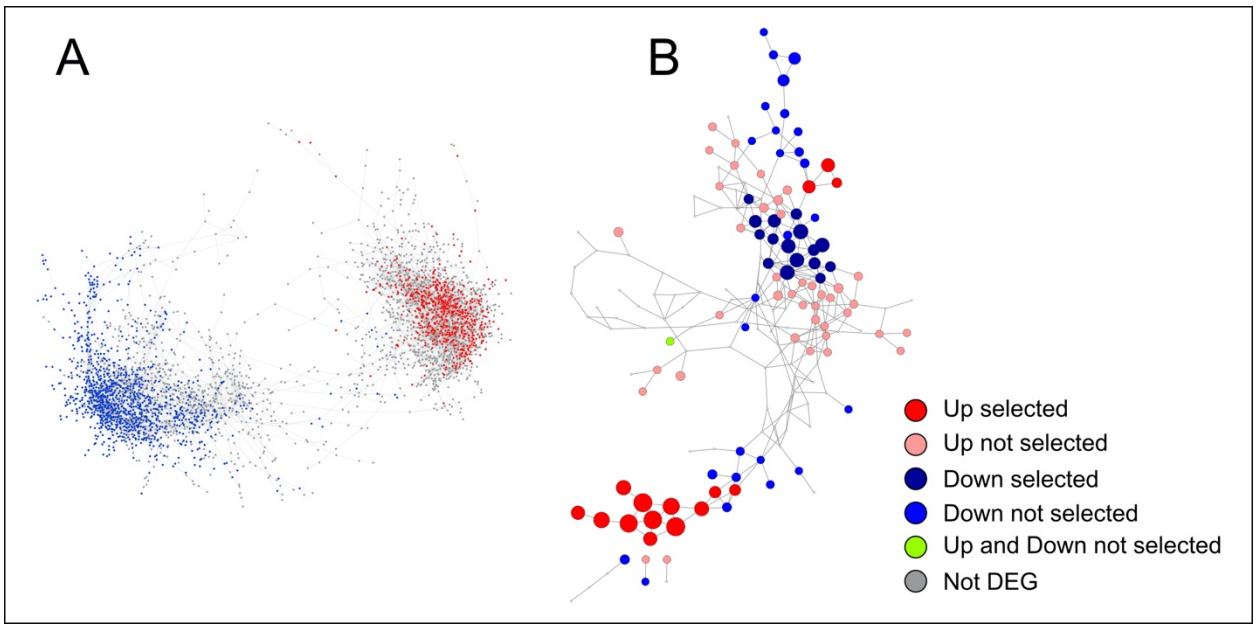


Figure 4. (A) Gene coexpression network (GCN) colored according to the modules considered as associated with cold response; each node represents a gene, and each connection a minimum R Pearson correlation coefficient of 0.8 (corrected with the highest reciprocal rank approach). (B) GCN network contracted according to the groups identified by the heuristic cluster chiseling algorithm (HCCA); each node represents a cluster colored according to modules selected and sized according to the proportion of DEGs (Supplementary Table 5 - Access in https://docs.google.com/spreadsheets/d/1LU2geL64NmrEPg88SwoTZ5JYPhevc72y32pD5UvVzCA/edit?usp=sharing).

The down-regulated module (i), as a whole, did not present any overrepresented GO category; however the cluster c68, by itself, was enriched for cell cycle. The GO enrichment of the module (ii) showed that signaling processes were amplified since MAP kinase kinase kinase and calcium ion (Ca^{2+}) transmembrane transporter activities were overrepresented. The group was also enriched for response to stress, photoprotection and plant-type hypersensitive response. Additionally, we also observed in (ii) clusters enriched for specific GO terms. Twenty-five early response genes were grouped into cluster c14, which was the cluster with more overrepresented GO categories among the groups of the modules selected, being enriched for response to high light intensity, response to reactive oxygen species and transcription regulation activity. Other clusters from this module were enriched for stress response (c28, c126 and c159), ion transport (c34), development (c28) and translation (c143). Intriguingly, the second transcript with the highest BC and SC values in the whole network is in cluster c28 and does not have a functional annotation. It has seven connections, four of which with genes from different clusters. Its sequence is conserved among plant species and it is described as an uncharacterized protein.

The genes clustered in the second up-regulated module (iii), according to the enrichment analysis, are mostly involved with carbohydrate metabolic processes, lipid phosphorylation and phosphatidylinositol phosphate biosynthetic processes. By itself, cluster c106 was enriched for developmental growth and water homeostasis.

The GO terms identified in the modules selected were summarized into several biological processes, separated into up (Figure 5A) and down-regulated (Figure 5B) associated categories. Although up and down-regulated DEGs from these modules present evident similarities into the GO visualization, it is clearly observed that the same biological processes are affected in specific ways.

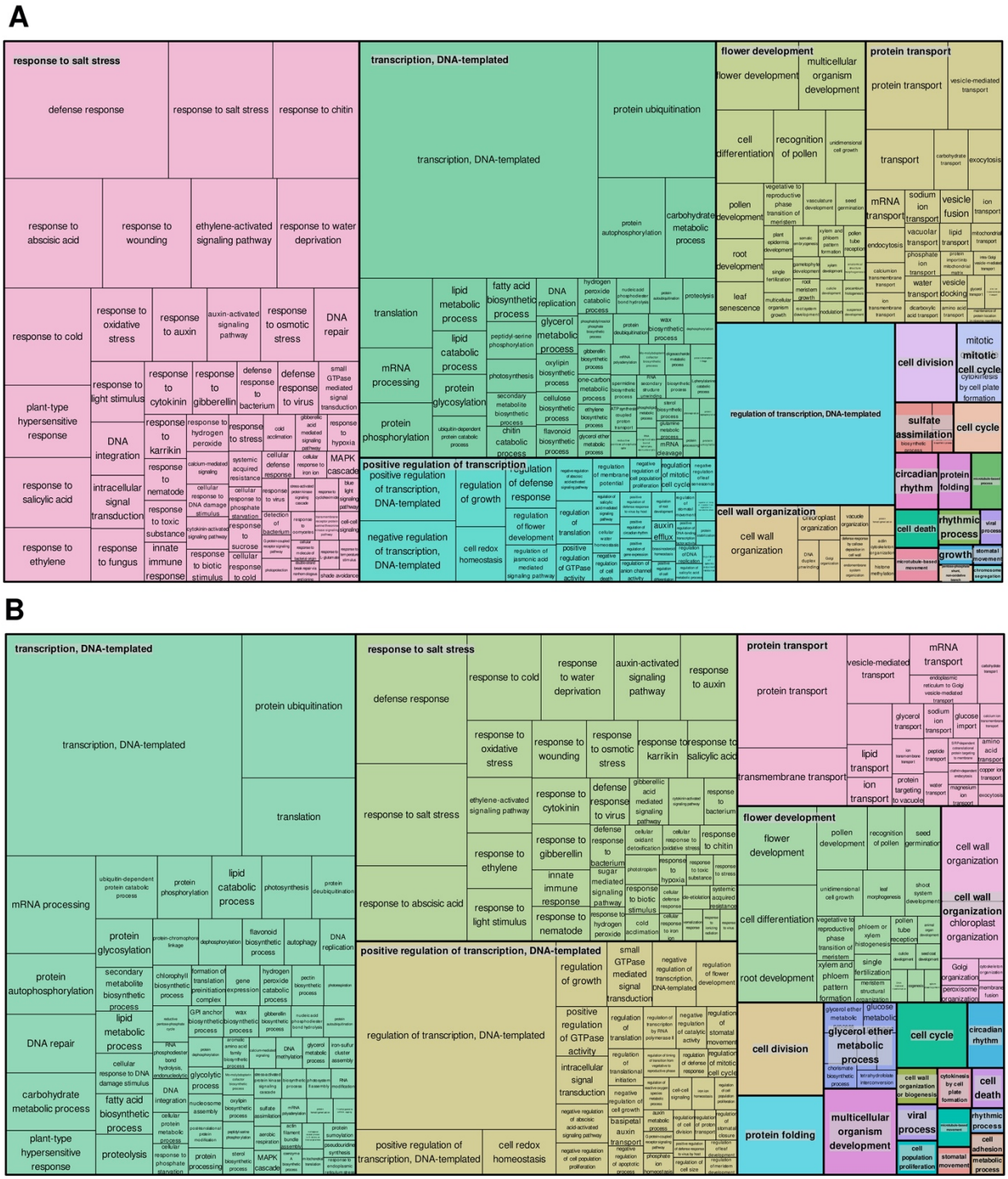


Figure 5. Gene ontology (GO) terms for the cold response modules selected: (A) Up modules; (B) Down module.

In order to evaluate the importance of genes inside these modules, we evaluated the presence of central genes inside each cluster, the hubs. For defining a gene as a hub, we selected those ones with the highest degree value (Supplementary Table 6 - Access in https://docs.google.com/spreadsheets/d/1UynT3MQvcdZVKOEBUKuAwpMaYVZ17Ds3v_BdCU9hAWE/edit?usp=sharing). Although the network structure allows the insertion of ten edges for each node respecting the minimum Pearson correlation of 0.8, we did not observe this maximum quantity in all the clusters. Interestingly, one two hubs were not among the set of DEGs.

4 Discussion

The optimal conditions for *H. brasiliensis* development are temperatures between 22-30 ºC, relative humidity of 70% or higher, and annual rainfall between 1,500 and 3,000 mm. In rubber trees, suboptimal conditions directly affect plant development, impacting rubber production (Rao & Kole, 2016). Currently, rubber tree plantations are mostly located in Southeast Asia, which has the ideal edaphoclimatic conditions for the crop development. Interestingly, such varieties are genetically closer to genotypes derived from the southern part of the Amazon basin (Souza et al., 2015), also presenting an adaptation capacity to the Amazon Rainforest dynamics. In this context, Brazilian rubber tree breeding programs have been focused on the development of genotypes with both high productivity and tolerance to stress imposed by escape areas, especially cold, which inhibits biochemical reactions, reduces photosynthetic capacity and alters membrane permeability (Sage & Kubien, 2007; Sevillano et al., 2009; Mai et al., 2010).

Acclimatization is essential for the development of rubber tree crops not adapted to the environmental conditions of the place to be developed, many plants increase their tolerance to freezing when exposed to low temperatures (Tomashow, 1999). The expression of specific genes is altered at the transcriptional level during the cold acclimation process, forming several cooling-responsive pathways from a signaling network (Tomashow, 1999; Carvallo et al., 2011; Jeon & Kim, 2013). Thus, the work carried out in this paper, which starts from transcripts of genotypes that represent a good part of the genomic characterization of rubber trees, opened paths for understanding the response to cold in the species and which biological pathways to focus on the next steps of rubber tree breeding.

GT1 is a primary clone obtained through open pollination of the wild seedlings introduced in South East Asia and their unselected progeny. RRIM600 was selected in the first rubber tree breeding cycle, when breeders first used controlled pollinations. Both genotypes are prime progenitors of modern rubber tree clones (Priyadarshan et al., 2009). Considering rubber tree long life cycle and the few generations that occurred these clones are substantially close to its wild ancestors, therefore they may be considered as genotypes bearing the wild *Hevea* natural genetic variation. For such a reason, we performed inferences on *H. brasiliensis* cold resistance using these genotypes.

We started our analyses by identifying genes modulated in different periods of cold response through differential expression tests. We observed contrasted quantities of DEGs for the conditions established, being the highest quantity of DEGs identified when comparing 0h and 24h and the smallest amount for 0h and 90min. These results may indicate that the seedlings are able to withstand a short period of cold exposure without massive activation or repression of genes. After twelve hours of exposure to the cold, we can observe that both genotypes underwent a reprogramming in their expression profiles (DEGs identified in 0h vs 12h), with the new changes being practically maintained over the next twelve hours (DEGs identified in 12h vs 24h). About 93% of the down-regulated genes and 96% of the up-regulated genes could be annotated, being the ubiquitin protein the most prominent annotation among the DEGS. The ubiquitin proteasome system (UPS) plays an important role in the plant response to diverse environmental stimuli, and RING E3 ubiquitin proteins are essential for UPS because of their functions in different developmental processes and stress responses since these proteins recruit the substrate to be degraded (Cho et al., 2017).

In addition to DEG analyses, we expanded the inferences on cold response using GCNs. Analyses based on such a structure is a promising approach to better elucidate the participation of genes linked to cooling stress. Considering that such a tolerance is the result of a network of complex biochemical pathways, an appropriate way to understand such a regulation is not evaluating a single gene, but a set of connected genes corresponding to a broad molecular mechanism. The correlation of how the coordinated expression of genes involved in multiple metabolic pathways act in response regulation can be elucidated through the definition of clusters within the GCN (Umer et al, 2020; Sun et al., 2021). Recently, Ding et al. (2020) modeled a GCN

using several public rubber tree transcriptomes, providing insights into rubber biosynthesis. Even with a high potential, such approaches are incipient in rubber tree research and this study is the first initiative on evaluating *Hevea* cold response in a more holistic perspective.

4.1 Selection of cold response associated clusters

From the GCN clusters selected as associated with cold response, we could supply a range of molecular mechanisms triggered by such a stress. It was possible to observe an over representation of the transmembrane transporter activities of MAP kinase kinase and calcium ion (Ca2+) transmembrane. This mechanism of action to cold stress has been reported in other studies where several metabolites act as low temperature sensors, after which cells release calcium through phosphatidylinositol signaling and inositol phosphate metabolism through the MAPK signaling pathway . This MAPK signaling pathway triggers signal transduction via TFs such as ERF, bHLH and MYB TFs (Sun, 2021).

By performing an analysis of enzymes and correspondent metabolic pathways within such groups, we could find diverse associations into cold response. The cold exposure triggered the up-regulation of enzymes involved in the synthesis of protective photosynthetic pigments, such as zeaxanthin and lutein (c106) (Dall'Osto et al., 2006), and lipoic acid (lipoamide) (c194), which protects against oxidative stress (Navari-Izzo et al., 2002). Several enzymes from the nucleotide sugars' metabolic pathway were up-regulated (c101, c106, c121, c130, c143, c16, c161, c34, c53). Nucleotide sugars are essential for cell wall biosynthesis and can function as signaling molecules, being sugar donors for the targeted-glycosylation of different compounds (Figueroa et al., 2021). Interestingly, enzymes from the methionine metabolism pathway that produce ethylene as their final product were also up-regulated (c121, c28, c34), as well as, α-linolenic acid metabolism enzymes that synthetize oxylipins, which are jasmonic acid (JA) precursor molecules that also have signaling functions in cold stress (c126, c159, c34, c53) (Du et al., 2013; Maynard et al., 2018; Wang et al., 2020).

The starting point of rubber tree cold stress response mainly resides in cluster c14, which harbors several sequences annotated as *ethylene-responsive transcription factors* (*ERFs*). *ERFs* are activated by the ethylene signaling pathway and regulate the transcription of ethylene-responsive genes, which negatively impact plant growth (Müller and Munné-Bosch, 2015; Van

den Broeck et al., 2017). The *ERF* family in rubber tree is comprised of 115 members (Duan et al., 2013) and the overexpression of an arabidopsis *ERF1* orthologue in rubber tree plants enhanced their tolerance to abiotic stresses and augmented laticifer cell differentiation (Lestari et al., 2018). In our analysis, all degs annotated as ERFs were up-regulated, suggesting that *H. brasiliensis* plantlets maintained the ethylene-mediated stress response active during the whole experiment. The up-regulation of ethylene biosynthesis enzymes also indicates that ethylene plays a constitutive role in rubber tree cold stress response. Surprisingly, only six enzymes were early responsive to the cold treatment and all were placed in this cluster as well and four of them are RING-type E3 ubiquitin transferases. Two were annotated as RING-H2 finger proteins ATL2 (PASA_cluster_59724 and PASA_cluster_65327). RING-H2 finger proteins, in special the *ATL* genes, are involved in the regulation of plant response to different type of stresses (Bopopi et al., 2010; Serrano et al., 2006; Song et al., 2016; Kim et al., 2020), being *ATL2* an early responsive gene in Arabidopsis (Salinas-Mondragon et al., 1999). The other two were annotated as E3 ubiquitin-protein ligase RING1 (PASA_cluster_150375 and PASA_cluster_47183), which are involved in the regulation of programmed cell death during plant hypersensitive response (Lin et al., 2008; Lee et al., 2011).

Considering the subnetwork formed with the first up-regulated module, c14 is solely connected to cluster c159, indicating a direct mechanism associated. In addition to being enriched for response to stress and protein ubiquitination, the cluster c159 has transcripts annotated as ERFs, calmodulin binding proteins, serine/threonine protein kinases and disease-responsive genes. Three early-responsive transcripts are present in this cluster, but only two were annotated: Nematode resistance protein-like HSPRO2, a hub gene of this cluster (see below), and BON1-associated protein 2 (BAP2), a gene that is connected to two c14 genes. HSPRO-like proteins are associated with disease resistance, being up-regulated after pathogen infection (Murray et al., 2007; Nemchinov et al., 2017). BAP proteins are negative regulators of cell death in plants, which is essential for plant development and hypersensitive response, and their expression is modulated by temperature (Yang et al., 2007; Cao et al., 2019; Hou et al., 2018). c159 is connected to two other clusters: c53, which did not present any overrepresented GO category but has transcripts annotated as proteins involved with protein phosphorylation and serine/threonine kinase activity, and cluster c143.

Other up-regulated clusters presented important associations into cold response. Cluster c143 was enriched for the translation process, more specifically for negative regulation of translation. This group has three members of the C Carbonatabolite-Repressor 4 (CCR4)–associated factor 1 (CAF1) gene family that were up-regulated in response to the cold treatment. These deadenylase proteins are part of the CCR4- negative on Tata (CCR4-NOT) complex that promotes the deadenylation of mRNAs, which leads to the suppression of translation and mRNA degradation (Collart 2016, Fang et al., 2020). In plants, CAF1 proteins were shown to be necessary for proper development and stress response processes (Sarowar et al., 2007; Liang et al., 2009; Walley et al., 2010; Shimo et al., 2019; Fang et al., 2020, 2021; Wang et al., 2021). In addition, stress-responsive TFs, nucleotide sugar biosynthesis' enzymes and cell wall biogenesis' proteins were assigned to c143 as well. In the HCCA network, c143 is connected to clusters c34 e c28.

In addition to being enriched for stress responses and developmental processes, the GO terms "positive regulation of protein kinase activity" and "DNA-binding transcription factor activity" were overrepresented in cluster c28. Among the 108 transcripts that have a functional annotation, 25 (23%) are TFs involved in the regulation of plant defense response. The most represented TF family was WRKY with eight members. WRKYs are one of the largest plant-specific TF families and they participate in the regulation of plant growth and development as well as promoters of plant stress response mechanisms by activation and inhibition of target-genes transcription and modulation of signaling cascades (Jiang et al., 2017). This cluster also harbors transcripts similar to enzymes from the ethylene biosynthesis pathway, calmodulin-binding proteins and small auxin up-regulated RNA (SAUR) proteins. Plants have several SAUR genes and they are essential growth factors for regular plant development and adaptive growth under stress conditions. They are responsive to different environmental factors and to other phytohormones besides auxin (Stortenbeker and Bemer, 2019; He et al., 2021). In addition, this cluster has a strong candidate for further analysis into rubber tree cold stress response: PASA_cluster_65283, the uncharacterised protein with the second highest BC and SC values in the whole GCN. Five of its connected transcripts were annotated and the proteins are involved with plant meiotic recombination (Homologous-pairing protein 2 homolog, c143) (Uanschou et al., 2013; Shi et al., 2019), zeaxanthin synthesis (Beta-carotene 3-hydroxylase 1, chloroplastic, c106) (Fiore et al., 2006), JA-dependent defense response (Suppressor of NPR1-1 Constitutive 4,

c28) (Bi et al., 2010), stress-induced osmotic solutes accumulation (Probable polyol transporter 4, c28) (Noiraud et al., 2001) and Ca^{2+}-signaling (Probable calcium-binding protein CML36, c28) (Astegno et al., 2017).

Another cluster that also showed enrichment for stress response was cluster c126, which only showed connection with cluster c28 in the HCCA network. There was an overrepresentation of GO terms for phenylalanine biosynthesis and sphingolipid biosynthesis processes. Phenylalanine, apart from its role in protein biosynthesis, is the precursor of several plant phenolic compounds, such as lignin, flavonoids and anthocyanins (Maeda and Dudareva, 2012), and salicylic acid (SA) (Lefevere et al., 2020). Sphingolipids constitute a significant part of plant plasma membrane lipids and are also involved in the control of PCD and mediating SA-signaling in plant stress response (Pata et al., 2010; Alden et al., 2011; Rivas-San Vicente et al., 2013).

Stress responses and plant growth can also be associated with the c34 cluster, enriched for ionic transport, more specifically for transmembrane transport of Ca2+. Ca2+ is widely known as a second messenger in plants, being (Zhu, 2016; Yuan et al., 2018; Michailidis et al., 2020). Cluster c34 is the most connected cluster in the first up-regulated group, being connected to clusters c143, c53, c161 and c121 in the HCCA network. Low temperature stimulus is well established as a main trigger of cell Ca2+ influx by the activation of Ca2+ channels, which leads to the Ca2+-signaling cascade of plant cold stress response (Mori et al., 2018; Liu et al., 2021; Mao et al., 2021). Considering the overrepresentation of proteins with Ca2+ transporter activity in cluster c34, it could be seen as the cluster controlling Ca2+-influx, therefore regulating Ca2+-signaling, which agrees with the cluster central position in the up-regulated module (ii).

The remaining five clusters in module (ii) did not present any overrepresentation of GO terms in the enrichment analysis, nevertheless the genes present in these clusters are involved with signaling processes, response to stress and transcription regulation. Cluster c53 houses members of *ERF* and *TIFY*/*JAZ* transcription factor families. *TIFY*/*JAZ* TF family is specific to plants and its members are negative regulators of JA-induced responses. Nevertheless, their expression is induced by JA, which creates a regulatory feedback loop for fine-tuning of JA-signaling (Chini et al., 2007). Different types of stresses induce the expression of these genes and they are essential

for plant growth-defense balance (Zhu et al., 2014; Major et al., 2017; Guo et al., 2018; He et al., 2020).

Cold acclimatization and response to low temperatures was also associated with cluster 101, where there is positive expression of cognate 70 kilodalton heat shock proteins (Neven et al., 1992) in the period between 90 min and 24 hours of stress by cold in rubber tree. In addition, cluster c101 has the action of BAM1, a receptor protein kinase involved in plant development (DeYoung et al., 2006). This protein also has a role in drought stress response, being required for long distance signaling for stomatal closure (Takahashi et al., 2018). c101 also houses a transcript similar to DELLA protein RGL1, which is a negative regulator of growth promoting phytohormone gibberellin (GA) signaling pathway (Wen and Chan, 2002). Poplar plants overexpressing this gene showed severe dwarfed phenotypes (Busov et al., 2006). This cluster has a regulator of PCD as well. Myb108 is a negative regulator of ABA biosynthesis and ABA-induced cell death (Cui et al., 2013) and is highly induced in rose plants under chilling (4°C) and freezing (-20°C) treatments. Arabidopsis plants overexpressing the rose ortholog showed improved cold tolerance and better performance under other stresses as well as a shorter growth cycle than the wild type (Dong et al., 2021).

The cluster with the most enzymes mapped to the KEGG pathways was c121, with forty-four transcripts assigned to EC numbers. Enzymes that synthesize scopolin and lignins in the phenylpropanoid biosynthesis pathway were placed in this cluster. The coumarin scopolin and its precursor, scopoletin, accumulate in plant tissues in response to stress conditions (Kai et al., 2006; Zandalinas et al., 2017) and are believed to be involved in ROS scavenging (Chong et al., 2002; Lee et al., 2013). Members of the plant ACT domain-containing protein family (ACR4 and ACR8) are present in cluster c161, and their expression is induced by cold treatment in Arabidopsis (Hsieh and Goodman, 2002). Regarding PCD, c161 bears both proteins that positively (E3 ubiquitin-protein ligase RING1) or negatively (Accelerated cell-death ACD11) regulate cell death. Expression of RING1 is induced by biotic stress and the protein is necessary for the activation of the PCD process (Lin et al., 2008; Lee et al., 2011). ACD11 is a sphingolipid transfer protein that controls cell death through the regulation of sphingolipid levels (Brodersen et al., 2002; Simanchu et al., 2014). It is down-regulated by the UPS machinery after pathogen infection (Liu et al., 2017),

however ACD11 transiently accumulates under low concentrations of ABA and under salt and drought stress and its overexpression confers abiotic stress tolerance (Li et al., 2020).

Cluster c194 has proteins involved with negative regulation of SA-mediated PCD (MACPF domain-containing proteins CAD1 and NSL1) (Tsutsui et al., 2006, Noutoshi et al., 2006), regulation of photoperiod gene expression and circadian clock (WD repeat-containing protein LWD1) (Wu et al., 2008; Wang et al., 2011), sugar transporters involved with cell wall biogenesis (UDP-galactose transporter 2) (Norambuena et al., 2005), and plant growth and stress response (Sugar transport protein 13) (Schofield et al., 2009; Lee and Seo 2021), besides having lignin synthesizing proteins as well.

4.2 Identification of key elements within cold response associated modules

In addition to evaluating the functional profile of all the associated genes within the modules selected, we performed inferences into the importance of these genes for the cold response mechanisms. Genes with high centrality within a GCN may be described as core regulators (Amrine et al., 2015), being known as hubs. Through the criteria defined for selecting hubs, we could identify for each cluster of the cold stress gene modules the hub genes, being considered key elements into the cold resistance definition (Carlson et al., 2006; Koido et al., 2018).

From the clusters selected as associated with cold down-regulation, we could identify nineteen hub genes. Almost all hubs of these clusters were identified as DEGs in the comparison performed both at 0h and 90min with 12h and 24h. The functional annotation of two of these hubs indicated a down-regulation of photosynthetic pathway genes (clusters c42 and c166). The c42 hub gene was annotated as lycopene beta cyclase, chloroplastic/chromoplastic (LCYB), an enzyme of the carotenoid biosynthesis process. Carotenoids form pigment-protein complexes in the photosystems protecting the machinery from reactive oxygen species (ROS) created by abiotic stresses (Dall'Osto et al., 2014; Shi et al., 2015; Kang et al., 2018). The hub from cluster c166 was annotated as OTP51, a pentatricopeptide repeat-containing protein, that is required for the proper assembly of photosystems I and II in Arabidopsis and rice (de Longevialle et al., 2008; Ye et al., 2012). Combined with low temperature, light promotes an enhancement in transcription of cold-responsive and photosynthesis-related genes (Soitamo et al., 2008). The down-regulation of both

genes in rubber trees might be linked to the reduction of photosynthetic activity due to cold stress; nevertheless this modulation was detected in the dark period of the experiment.

In addition to the regulation of photosynthetic activity from the cold down-regulation genes, it was also possible to identify two hubs representing proteins involved in abscisic acid (ABA) signaling, which is essential for development and responses to abiotic stress. A transcript similar to a protein phosphatase 2C (PP2C) is one of the hubs in cluster c171. Chao et al. (2020), in a genome-wide identification and expression analysis of the phosphatase 2A family in rubber trees, identified cis-acting elements related to the low temperature responsiveness (LTR) category. Under non-stress conditions, PP2C proteins negatively regulate the ABA-mediated signaling pathway by inactivating ABA-responsive genes. Once under stress, ABA-receptors bind PP2Cs and inhibit their activity, consequently activating ABA signals (Cai et al., 2017). The other hub gene belongs to cluster c134 and was annotated as zinc finger CONSTANS-like 4 protein (COL4), a transcription factor. COL4 is a flowering inhibitor by repressing the flowering locus T (FT) gene expression (Steinbach, 2019) and, strikingly, it is also involved in abiotic stress responses through the ABA-dependent signaling pathway in Arabidopsis (Min et al., 2015). Considering that ABA concentration might have increased in the rubber trees during the cold exposure, which is enforced by the down-regulation of a transcript similar to a PP2C protein, one could expect that this gene would be up-regulated instead. ABA is known for delaying flowering, nevertheless, under severe drought conditions, ABA up-regulates florigen genes, which is part of the drought escape (DE) strategy. DE prompts plants to accelerate their vegetative growth and reproduction stages during the high water availability period. In the drought season, the plants enter a dormant stage (Verslues and Juenger 2011; Yildirim and Kaya 2017).

The hub from the cluster c156 was annotated as histone acetyltransferase of the MYST family (HAM) 1/2 proteins, which are involved in UV-B DNA damage repair (Campi et al., 2012). In Arabidopsis, HAM1 and HAM2 proteins are expressed mainly in growing tissues and double mutants are lethal (Latrasse et al., 2008). Interestingly, these proteins positively regulate the expression of FLOWERING LOCUS C (FLC) gene, therefore being a flowering repressor as well (Xiao et al., 2013). The down-regulation of this transcript in *H. brasiliensis* plantlets might be due to a decrease in damaged DNA by UV-B because of the absence of light. In addition, the down-

regulation of c196 hub annotated with Down Homeobox protein BEL1 homolog from different metabolic processes that are flowering repressors (Bhatt et al., 2003) suggests that rubber trees may have a DE-type strategy to deal with abiotic stresses.

The down-regulation of three other strongly co-expressed hub genes, from the c184, c171 and c153 clusters, detected in the last phase of the cold experiment, also suggests an increase in ABA synthesis by the *H. brasiliensis* plantlets connected to an inhibition of growth and the photoperiod. The hub identified for cluster c184 was annotated as homeobox-leucine zipper (HD-Zip) protein HAT5 (AtHB1) and it was detected as a down-regulated gene after 12 hours of rubber tree cold exposure. HD-Zip transcription factors seem to be a plant-specific TF family and its members take part in plant development and stress response (Ariel et al., 2007), each member having its own expression pattern under different environmental conditions (Li et al., 2019; Li et al., 2020). In Arabidopsis, HAT5 expression is down-regulated by salt stress and low temperature, but up-regulated by darkness (Henriksson et al., 2005) and is a positive regulator of growth by modulating the expression of genes involved in cell elongation and cell wall composition (Capella et al., 2015). This hub was connected to not only c171 hub, but also to c153 hub, which was annotated as leaf rust receptor-like kinase 10-like (LRK10L) protein. LRK10 is a disease resistance gene first described in wheat (Feuillet et al., 1997) and wheat members of its family were shown to be induced by stresses and light. Interestingly, in the dark period these genes were down-regulated even though the plants were under biotic stress (Zhou et al., 2007). The Arabidopsis' homolog gene codes for two different proteins: one is involved in ABA-signaling and is down-regulated by drought stress and ABA treatment, while the other is up-regulated in these conditions, which indicates that these transcripts might regulate each other (Lim et al., 2015). These three hub genes were strongly co-expressed together in rubber trees, directly connecting three different clusters. Their down-regulation was detected in the latter phase of the cold experiment, which is suggestive of an increase in ABA synthesis by the *H. brasiliensis* plantlets connected to an inhibition of growth and the photoperiod.

Considering the differences in cold stress response from two *H. brasiliensis* clones obtained in the early years of rubber tree domestication, we attempted to have a glimpse of *Hevea* general genetic responses to cold stress. As a tropical tree species, rubber tree most probably did

not have to deal with prolonged periods of low temperatures during its evolution; nevertheless the few wild genotypes that were successfully introduced in Southeast Asia carried so much genetic diversity that breeding for varieties that are cold tolerant is still possible. As a result of the GCN strategy applied in this study, we could visualize *Hevea*'s primary reprogramming of gene expression and the relationship among the genes involved in the cold stress response. In the short period of cold exposure used in this work, the plantlets activated the ethylene-mediated signaling pathway since the beginning of the stress treatment and kept ethylene signaling active. PCD plays a major role in the rubber tree cold response process, being tightly regulated by the signaling cascades. Growth inhibition and cell wall thickening are implemented by the plantlets, which can be correlated to a possible drought escape strategy triggered by the cold stress. In view of the genotypes analyzed, our results may represent the species' genetic stress responses developed during its evolution. The understanding of how *H. brasiliensis* copes with low temperature stress can improve greatly the breeding strategies for this crop as well as emphasize the importance of rubber tree genetic diversity preservation, since it has such a narrow genetic base, is being impacted by climate change and is the only source for large scale rubber production.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

AS, CS, CM and RV conceived the project; CS, SB, AA, FF, RB and RV performed the analyses and wrote the manuscript. All authors reviewed, read and approved the manuscript.

Funding

The authors gratefully acknowledge the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) for PD fellowship to CS (2015/24346-9), and Ph.D. fellowships to SB (2019/13452-3), AA(2019/03232-6); FF (18/18985-7); master's fellowship to RB (2018/23831-9); the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) for financial support (Computational Biology Program - 88882.160095/2013–01) and internship fellowship to SB (PE-BOLSAS7349056), and the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for research fellowship to AS (312777/2018-3).

References

Alden K P,Dhondt-Cordelier S, McDonald K L, Reape T J, Ng C K Y, McCabe P F, Leaver C J. (2011). Sphingolipid long chain base phosphates can regulate apoptotic-like programmed cell death in plants. Biochem Biophys Res Commun. 410(3):574-80. doi: 10.1016/j.bbrc.2011.06.028.

Amrine, K. C., Blanco-Ulate, B., & Cantu, D. (2015). Discovery of core biotic stress responsive genes in Arabidopsis by weighted gene co-expression network analysis. *PloS one*, *10*(3), e0118731.

Ariel F D , Manavella P A, Dezar C A, Chan R L. (2007). The true story of the HD-Zip family. Trends Plant Sci.12(9):419-26. doi: 10.1016/j.tplants.2007.08.003.

Ashraf M A, Rahman A. (2019). Cold stress response in Arabidopsis thaliana is mediated by GNOM ARF-GEF. Plant J. 97(3):500-516. doi: 10.1111/tpj.14137. Epub 2018 Dec 17.

Bhatt AM, Etchells JP, Canales C, Lagodienko A, Dickinson H. (2004). VAAMANA--a BEL1-like homeodomain protein, interacts with KNOX proteins BP and STM and regulates inflorescence stem growth in Arabidopsis. Gene. 17;328:103-11. doi: 10.1016/j.gene.2003.12.033. PMID: 15019989.

Bolaños-Villegas P, Guo C L & Jauh G Y. (2015). Arabidopsis Qc-SNARE genes BET11 and BET12 are required for fertility and pollen tube elongation. Botanical Studies volume 56, Article number: 21

Bryant DM, Johnson K, DiTommaso T, Tickle T, Couger MB, Payzin-Dogru D, Lee TJ, Leigh ND, Kuo TH, Davis FG, Bateman J, Bryant S, Guzikowski AR, Tsai SL, Coyne S, Ye WW, Freeman RM Jr, Peshkin L, Tabin CJ, Regev A, Haas BJ, Whited JL. (2017) . A Tissue-Mapped Axolotl De Novo Transcriptome Enables Identification of Limb Regeneration Factors. Cell Rep. 17;18(3):762-776. doi: 10.1016/j.celrep.2016.12.063.

Campos Mantello, C., Boatwright, L., da Silva, C.C. et al., (2019). Deep expression analysis reveals distinct cold-response strategies in rubber tree (Hevea brasiliensis). BMC Genomics 20, 455 . https://doi.org/10.1186/s12864-019-5852-5

Cao Shifeng, Xiao Zeyu, Jiranek Vladimir, Tyerman Stephen D. (2019) The VvBAP1 gene is identified as a potential inhibitor of cell death in grape berries. Functional Plant Biology 46, 428-442.

Cattaneo, P, Hardtke, C S. (2017). BIG BROTHER uncouples cell proliferation from elongation in the Arabidopsis primary root. Plant and Cell Physiology, v. 58, n. 9, p. 1519-1527.

Carlson, M. R., Zhang, B., Fang, Z., Mischel, P. S., Horvath, S., & Nelson, S. F. (2006). Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks. BMC genomics, 7(1), 1-15.

Carvallo MA, Pino MT, Jeknic Z, Zou C, Doherty CJ, Shiu SH, Chen TH, Thomashow MF. (2011). A comparison of the low temperature transcriptomes and CBF regulons of three plant species that differ in freezing tolerance: Solanum commersonii, Solanum tuberosum, and Arabidopsis thaliana. Journal of experimental botany, v. 62, n. 11, p. 3807-3819.

Chao J, Huang Z, Yang S, Deng X, Tian W (2020) Genome-wide identification and expression analysis of the phosphatase 2A family in rubber tree (Hevea brasiliensis). PLOS ONE 15(2): e0228219. https://doi.org/10.1371/journal.pone.0228219

Cheng H, Chen X, Fang J, An Z, Hu Y, Huang H. (2018). Comparative transcriptome analysis reveals an early gene expression profile that contributes to cold resistance in Hevea brasiliensis (the Para rubber tree), Tree Physiology, Volume 38, Issue 9 Pages 1409–1423, https://doi.org/10.1093/treephys/tpy014

Chini A, Fonseca S, Fernández G, Adie B, Chico J M, Lorenzo O, García-Casado G, López-Vidriero I, Lozano F M, Ponce M R, Micol J L & Solano R. (2007). The JAZ family of repressors is the missing link in jasmonate signaling. Nature volume 448, pages 666–671.

Cho S K, Ryu M Y , Kim J H , Hong J S , Oh T R , Kim W T , Yang S W. (2017). RING E3 ligases: key regulatory elements are involved in abiotic stress responses in plants. BMB Rep. 50(8):393-400. doi: 10.5483/bmbrep.2017.50.8.128.

Dall'Osto L, Lico C, Alric J, Giuliano G, Havaux M, Bassi R. (2006). Lutein is needed for efficient chlorophyll triplet quenching in the major LHCII antenna complex of higher plants and effective photoprotection in vivo under strong light. BMC Plant Biol. 6:32. doi: 10.1186/1471-2229-6-32.

Du, H., Liu, H., & Xiong, L. (2013). Endogenous auxin and jasmonic acid levels are differentially modulated by abiotic stresses in rice. Frontiers in plant science, 4, 397.

de Faÿ, E., & Jacob, J. L. (1989). Anatomical organization of the laticiferous system in the bark. In Physiology of rubber tree latex (pp. 3-14). CRC Press.

de Longevialle AF, Hendrickson L, Taylor NL, Delannoy E, Lurin C, Badger M, Millar AH, Small I. (2008). The pentatricopeptide repeat gene OTP51 with two LAGLIDADG motifs is required for the cis-splicing of plastid ycf3 intron 2 in Arabidopsis thaliana. Plant J. 56(1):157-68. doi: 10.1111/j.1365-313X.2008.03581.x.

den Broeck L V , Dubois M , Vermeersch M, Storme V , Matsui M, Inzé D. (2017). From network to phenotype: the dynamic wiring of an Arabidopsis transcriptional network induced by osmotic stress. Mol Syst Biol. 21;13(12):961. doi: 10.15252/msb.20177840.

Deng, X., Guo, D., Yang, S., Shi, M., Chao, J., Li, H., et al., (2018). Jasmonate signalling in regulation of rubber biosynthesis in laticifer cells of rubber tree (Hevea brasiliensis muell. arg.). J. Exp. Bot. 69, 3559–3571. doi: 10.1093/jxb/ery169

Ding, Z, Fu, L, Tan, D, Sun, X, Zhang, J. (2020). An integrative transcriptomic and genomic analysis reveals novel insights into the hub genes and regulatory networks associated with rubber synthesis in H. brasiliensis, Industrial Crops and Products, Volume 153, 112562

Driedonks, N., Xu, J., Peters, J. L., Park, S., & Rieu, I. (2015). Multi-level interactions between heat shock factors, heat shock proteins, and the redox system regulate acclimation to heat. Frontiers in plant science, 6, 999.

Dong, Hui et al., (2017). Ubiquitylation activates a peptidase that promotes cleavage and destabilization of its activating E3 ligases and diverse growth regulatory proteins to limit cell proliferation in Arabidopsis. Genes & development, v. 31, n. 2, p. 197-208.

Dusa, A. (2017). *venn: Draw Venn Diagrams*. R package version 1.2. Available online at: https://cran.microsoft.com/snapshot/2018-03-30/web/packages/venn/venn.pdf

Ewald R., Hoffmann C., Florian A., Neuhaus E., Fernie A.R., Bauwe H. (2004). Lipoate-protein ligase and octanoyltransferase are essential for protein lipoylation in mitochondria of Arabidopsis. Plant Physiol. 165:978-990

Fisch, G.; Marengo, J. A.; Nobre, C.A. (1998). Uma Revisão Geral Sobre o Clima da Amazônia. Acta Amazonica 28(2): 101-126.

Gonçalves PS & Fontes JRA (2009) Domesticação e Melhoramento da Seringueira. In: BORÉM A, LOPES MTG, CLEMENT CR & NODA H (Eds). Domesticação e Melhoramento: Espécies Amazônicas. Viçosa: UFV, p.395-423.

Gong, XX., Yan, BY., Hu, J. et al., (2018). Transcriptome profiling of rubber tree (Hevea brasiliensis) discovers candidate regulators of the cold stress response. Genes Genom 40, 1181–1197 https://doi.org/10.1007/s13258-018-0681-5

Guo M, Liu J H, Ma X, Luo DX, Gong Z H, Lu M H. (2016). The plant heat stress transcription factors (HSFs): structure, regulation, and function in response to abiotic stresses. Frontiers Plant Sci. 7:114.

Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, et al., The gene ontology (GO) database and informatics resource. Nucleic Acids Res. 2004;32:D258–61.

Huang, X., Shi, H., Hu, Z., Liu, A., Amombo, E., Chen, L., & Fu, J. (2017). ABA is involved in regulation of cold stress response in bermudagrass. Frontiers in Plant Science, 8, 1613.

Huang YC, Niu CY, Yang CR, Jinn TL. (2016). The Heat Stress Factor HSFA6b Connects ABA Signaling and ABA-Mediated Heat Responses. Plant Physiol. 172(2):1182-1199. doi: 10.1104/pp.16.00860.

International Rubber Consortium Limited (IRCo) https://ircorubber.com/2019/pestalotiopsis-endanger-national-rubber-production/

Jeon, J., Kim, J. Cold stress signaling networks in Arabidopsis. J. Plant Biol. 56, 69–76 (2013). https://doi.org/10.1007/s12374-013-0903-y

Jiang J, Ma S, Ye N, Jiang M, Cao J, Zhang J. (2017). J Integr Plant Biol. WRKY transcription factors in plant responses to stresses. 59(2):86-101. doi: 10.1111/jipb.12513.

Lai, R., Feng, X., Chen, J., Zhang, Y., Wei, X., Chen, Y., ... & Wu, R. (2021). De novo transcriptome assembly and comparative transcriptomic analysis provide molecular insights into low temperature stress response of Canarium album. Scientific Reports, 11(1), 1-17.

Latrasse D, Benhamed M, Henry Y, Domenichini S, Kim W, Zhou D X, Delarue M. (2008). The MYST histone acetyltransferases are essential for gametophyte development in Arabidopsis. BMC Plant Biol. 28;8:121. doi: 10.1186/1471-2229-8-121.

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

Latrasse D, Benhamed M, Henry Y, Domenichini S, Kim W, Zhou D X, Delarue M. (2008). The MYST histone acetyltransferases are essential for gametophyte development in Arabidopsis. BMC Plant Biol. 28;8:121. doi: 10.1186/1471-2229-8-121.

Lisa Van den Broeck 1 2 , Marieke Dubois 1 2 , Mattias Vermeersch 1 2 , Veronique Storme 1 2 , Minami Matsui 3 , Dirk Inzé From network to phenotype: the dynamic wiring of an Arabidopsis transcriptional network induced by osmotic stress. Mol Syst Biol. 2017 Dec 21;13(12):961. doi: 10.15252/msb.20177840.

Kim MH, Cho JS, Park EJ, Lee H, Choi YI, Bae EK, Han KH, Ko JH. (2020). Overexpression of a Poplar RING-H2 Zinc Finger, Ptxerico, Confers Enhanced Drought Tolerance via Reduced Water Loss and Ion Leakage in Populus. Int J Mol Sci. 11;21(24):9454. doi: 10.3390/ijms21249454.

Koido, M., Tani, Y., Tsukahara, S., Okamoto, Y., & Tomida, A. (2018). InDePTH: detection of hub genes for developing gene expression networks under anticancer drug treatment. Oncotarget, 9(49), 29097.

Maren Müller, Sergi Munné-Bosch. Ethylene Response Factors: A Key Regulatory Hub in Hormone and Stress Signaling. Plant Physiol. 2015 Sep;169(1):32-41. doi: 10.1104/pp.15.00677

Min J H, Chung J S, Lee K H, Kim C S. (2014). The CONSTANS-like 4 transcription factor, AtCOL4, positively regulates abiotic stress tolerance through an abscisic acid-dependent manner in Arabidopsis. J Integr Plant Biol. 57(3):313-24. doi.org/10.1111/jipb.12246

Merico D, Isserlin R, Stueker O, Emili A, Bader GD PLoS One. (2010). Enrichment Map: A Network-Based Method for Gene-Set Enrichment Visualization and Interpretation Nov 15;5(11):e13984.

Mohamed Sathik, M.B., Luke, L.P., Rajamani, A. et al., (2018). De novo transcriptome analysis of abiotic stress-responsive transcripts of Hevea brasiliensis. Mol Breeding 38, 32 https://doi.org/10.1007/s11032-018-0782-5

Müller M, Munné-Bosch S. (2015). Ethylene Response Factors: A Key Regulatory Hub in Hormone and Stress Signaling. Plant Physiol. 169(1):32-41. doi: 10.1104/pp.15.00677

Nakagami, H., Pitzschke, A., & Hirt, H. (2005). Emerging MAP kinase pathways in plant stress signalling. Trends in plant science, 10(7), 339-346.

Navari-Izzo F, Quartacci MF, Sgherri C. (2002). Lipoic acid: a unique antioxidant in the detox-ification of activated oxygen species. Plant Physiol Biochem. 40:463–70.

Neven L G, Haskell D W, Guy C L, Denslow N, Klein P A, Green L G, Silverman A.(1992). Association of 70-Kilodalton Heat-Shock Cognate Proteins with Acclimation to Cold , *Plant Physiology*, Volume 99, Issue 4: 1362–1369, https://doi.org/10.1104/pp.99.4.1362

Oliveira, P J de, Rocha E J P., Fisch G. (2004). Efeitos de um evento de friagem nas condições meteorológicas na Amazônia: um estudo de caso. Acta Amazonica, v. 34, n. 4

Pata M O , Hannun Y A, Ng C K Y. (2010). Plant sphingolipids: decoding the enigma of the Sphinx. New Phytol. 185(3):611-30. doi: 10.1111/j.1469-8137.2009.03123.x.

Pootakham, W., Sonthirod, C., Naktang, C., Ruang-Areerate, P., Yoocha, T., Sangsrakru, D., ... & Tangphatsornruang, S. (2017). De novo hybrid assembly of the rubber tree genome reveals evidence of paleotetraploidy in Hevea species. Scientific reports, 7(1), 1-15.

Priyadarshan PM, Goncalves PS. *Hevea* gene pool for breeding. Genet Resour Crop Evol. 2003;50: 101–114.

Priyadarshan P.M. Hoa T.T.T. Huasun H. Gonc ̧alves P. de S. (2005) Yielding potential of rubber(Hevea brasiliensis) in sub-optimal environments. J. Crop Improv. 14, 221–247.

Priyadarshan. P. M. et al., Breeding Hevea Rubber. In: Jain, S.M.; Priyadarshan, P. M. (Ed.). Breeding Plantation Crops. Tropical Species, New York: Springer Science, 2009. p. 469-524.

Priyadarshan, P. M., & Clément-Demange, A. (2004). 3 Breeding Hevea Rubber: Formal and Molecular Genetics. *Advances in Genetics*, *52*, 51-116.

Rivas-San Vicente, M., Larios-Zarate, G. & Plasencia, J. (2013). Disruption of sphingolipid biosynthesis in Nicotiana benthamiana activates salicylic acid-dependent responses and compromises resistance to Alternaria alternata f. sp. lycopersici . Planta 237, 121–136. https://doi.org/10.1007/s00425-012-1758-z

Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010 Jan 1;26(1):139-40.

Saez A., Apostolova N., Gonzalez-Guzman M., Gonzalez-Garcia M.P., Nicolas C., Lorenzo O., Rodriguez P.L. (2004). Gain-of-function and loss-of-function phenotypes of the protein phosphatase 2C HAB1 reveal its role as a negative regulator of abscisic acid signalling. Plant J. 37:354-369

Saez A., Robert N., Maktabi M.H., Schroeder J.I., Serrano R., Rodriguez P.L. (2006). Enhancement of abscisic acid sensitivity and reduction of water consumption in Arabidopsis by combined inactivation of the protein phosphatases type 2C ABI1 and HAB1. Plant Physiol. 141:1389-1399.

Serrano M, Parra S, Alcaraz LD, Guzman P. (2006). The ATL gene family from Arabidopsis thaliana and Oryza sativa comprises a large number of putative ubiquitin ligases of the RING-H2 type. J Mol Evol.62:434–445. doi: 10.1007/s00239-005-0038-y.

de Souza LM, Le Guen V, Cerqueira-Silva CBM, Silva CC, Mantello CC, et al., Genetic Diversity Strategy for the Management and Use of Rubber Genetic Resources: More than 1,000 Wild and Cultivated Accessions in a 100-Genotype Core Collection. PLOS ONE, v. 10, n. 8, p. 1-20, 2015.

Soitamo A J, Piippo M, Allahverdiyeva Y, Battchikova N, Aro E M. (2008). Light has a specific role in modulating Arabidopsis gene expression at low temperature. BMC Plant Biol. 29;8:13. doi: 10.1186/1471-2229-8-13.

Song J, Xing Y, Munir S, Yu C, Song L, Li H, Wang T, Ye Z. (2016). An ATL78-Like RING-H2 Finger Protein Confers Abiotic Stress Tolerance through Interacting with RAV2 and CSN5B in Tomato. Front Plant Sci. 29;7:1305. doi: 10.3389/fpls.2016.01305.

Steinbach Y. (2019). The Arabidopsis thaliana CONSTANS- LIKE 4 (COL4) - A Modulator of Flowering Time. Front Plant Sci.28;10:651. doi: 10.3389/fpls.2019.00651.

Sun, S., Lin, M., Qi, X. et al., (2021). Full-length transcriptome profiling reveals insight into the cold response of two kiwifruit genotypes (A. arguta) with contrasting freezing tolerances. BMC Plant Biol 21, 365 https://doi.org/10.1186/s12870-021-03152-w

Theocharis, A., Clément, C., and Barka, E. A. (2012). Physiological and molecular changes in plants grown at low temperatures. *Planta* 235, 1091–1105. doi: 10.1007/s00425-012-1641-y

Thomashow, Michael F. (1999). Plant cold acclimation: freezing tolerance genes and regulatory mechanisms. Annual review of plant biology, v. 50, n. 1, p. 571-599

Umer MJ, Bin Safdar L, Gebremeskel H, Zhao S, Yuan P, Zhu H, et al., (2020). Identification of key gene networks controlling organic acid and sugar metabolism during watermelon fruit development by integrating metabolic phenotypes and gene expression profiles. Hortic Res.7(1):193. https://doi.org/10.1038/s41438-020-00416-8.

Wang Y, Yang L, Tang Y, Tang R, Jing Y, Zhang C, Zhang B, Li X, Cui Y, Zhang C, Shi J, Zhao F, Lan W, Luan S. (2017). Arabidopsis choline transporter-like 1 (CTL1) regulates secretory trafficking of auxin transporters to control seedling growth. PLoS Biol. 28;15(12):e2004310. doi: 10.1371/journal.pbio.2004310.

Willige B C, Ahlers S, Zourelidou M, Barbosa I C, Demarsy E, Trevisan M, Davis P A, Roelfsema MR, Hangarter R, Fankhauser C, Schwechheimer C. (2013). Plant Cell 25:1674-1688.

Wolters H , Anders N, Geldner N, Gavidia R, Jürgens G. (2011). Coordination of apical and basal embryo development revealed by tissue-specific GNOM functions. Development. 138(1):117-26. doi: 10.1242/dev.059147.

Xiao J, Zhang H, Xing L, Xu S, Liu H, Chong K, Xu Y. (2013). Requirement of histone acetyltransferases HAM1 and HAM2 for epigenetic modification of FLC in regulating flowering in Arabidopsis. J Plant Physiol.1;170(4):444-51. doi: 10.1016/j.jplph.2012.11.007.

Yang, H., Yang, S., Li, Y., & Hua, J. (2007). The Arabidopsis BAP1 and BAP2 genes are general inhibitors of programmed cell death. Plant physiology, 145(1), 135–146.

Ye J W, Gong Z Y, Chen CG, Mi HL, Chen GY. (2012). A mutation of OSOTP 51 leads to impairment of photosystem I complex assembly and serious photo-damage in rice. J Integr Plant Biol.54(2):87-98. doi: 10.1111/j.1744-7909.

Yu X Y, Yao Y, Hong Y H, Hou P Y, Li C X, Xia Z Q, Geng M T, Chen Y H. (2009). Differential expression of the Hsf family in cassava under biotic and abiotic stresses. Genome. 62(8):563-569. doi: 10.1139/gen-2018-0163.

Zourelidou M, Muller I, Willige B C, Nill C, Jikumaru Y, Li H, Schwechheimer C. (2009). The polarly localized D6 PROTEIN KINASE is required for efficient auxin transport in Arabidopsis thaliana. Development 136:627-636

7. *CAPÍTULO III*: TÓPICOS PARA LIVRO " TRANSCRIPTOME PROFILING: PROGRESS AND PROSPECTS" by M.A. Ali and J. Lee – ELSEVIER

1)Transcriptome and breeding in Orphan crops

Seventy percent of the calories that are part of human food correspond to the consumption of only 15 cultivated species. More than 50% of these calories consumed come from three major cereal crops in the world: rice, corn and wheat (Dawson et al., 2019). Despite this consumption concentrated in a few varieties of food sources, there are more than 2,500 species in the process of domestication, of which 300 have been fully domesticated (Fernie and Yan, 2019).

The concentration of food production makes up a food system far from a sustainable food supply for humanity (Khoury et al., 2014; Dawson et al., 2019). The lack of diversification in agricultural production reduces global food security against the current scenario of accelerated population growth and increased demand for food (Chang et al., 2019). In this context, the cultivation of orphan species play an important role in global food and nutritional security, particularly in the developing world (Jamnadass et al. 2020), in addition to presenting a complex of valuable gene pools for the improvement of future crops (Mabhaudhi et al., 2019). Thus, the nutritional and energy situation of the population, the conservation of biodiversity, sustainable use of natural resources and the consequences of global climate change (Kameny et al., 2021) are the main challenges for the genetic improvement of plants, especially for orphan crops.

One of the solutions with good prospects to face these challenges is the use of orphan crops, increasing agricultural production, food sources, sources of biocompounds and contributing to expand the genetic resources of resilient plants. In the approaching genomic Era, there are great prospects for the genetic improvement of orphan and main crops, including "de novo" domestication of orphan crops (Ye and Fan, 2021). New techniques for plant breeding, such as the transcriptome and other "omics", have made it possible to accelerate basic research and facilitate the exploration of orphan crops.

Many orphan cultures have been identified as intelligent in their ability to respond to the abiotic stresses they are exposed to. This developed adaptation to climatic conditions presents opportunities to promote these crops to a higher level of production and increase the degree of

molecular knowledge and improvement of more resilient plants. Increased drought and heat resulting from climate change have made current agricultural production particularly challenging (Mabhaudhi et al., 2019). Drought is one of the main abiotic constraints limiting agricultural production worldwide, and several orphan species exhibit high levels of tolerance to water stress, such as millet (Krishnamurthy et al. 2016), millet (Setaria italica) (Puranik et al. . 2011), pea (Lathyrus sativa L.) (Hanbury et al. 2000) and quinoa (Chenopodium quinoa Willd.) (Hinojosa) (Hinojosa et al. 2018).

Although the drought response stability mechanisms are not yet fully elucidated, among the known drought response mechanisms reported in these orphan plantations are the efficient antioxidant potential (Jiang et al. 2013), association with arbuscular mycorrhiza (Tyagi et al. 2017).), osmotic adjustment (Jiang et al. 2013) et al. 2013; Tyagi et al. 2017), reduction in green leaf area and stomatal conductance (Cullis and Kunert 2017).

In addition, orphan crops have been a reference in the development of genetic and molecular mechanisms to survive to the others abiotic stress factors such as soil salinity tolerance (Bray et al. 2000; Shailaja and Thirumeni 2007; Rahman et al. 2014). One of the most salt-tolerant species in the world is an orphaned annual dicotyledon with diverse uses (Patel 2016), the common Glasswort (Salicornia europaea L.), (Loconsole et al. 2019).

Other adaptations to specific agrosystems make orphan crops examples of responses to poor soils, flooding and consequent hypoxia, temperature extremes (Levizou et al. 2004), soils poor in fertility and deficiencies (Thilakarathna and Raizada 2015; Cullis et al. 2018; Mabhaudhi et al. 2019), nutrients and contaminated with toxic metals (Chandra et al. 2016; Sharma et al. 2010; Shinozaki et al. 2015; Mkumbo et al. 2012; Raskin et al. 1997).

Orphan crops, in addition to their importance as a source of food diversity, production of biofuels, soil phytoremediation, also have a large role in the production of medicines and pharmaceuticals and cosmetics (Pootakham and Nawae et al., 2021). For centuries native people have applied empirical knowledge to the medicinal use of plants, and today this knowledge has been extended to sciences such as herbal medicine, herbology and other ways of exploring the medicinal resources of plants. Caper (Capparis spinosa L.), a rich source of bioactive compounds with important nutritional and medicinal values, had its transcriptome sequenced and its transcription dataset provided an important resource for discovering genes and markers of a plant

with promising potential for agrosystems threatened by global warming (Mercati et al., 2019). Another plant of high nutritional value that had its transcriptome sequenced with the aim of improving the placement of the pulp and peel was Pitaya. These characteristics are essential for the cultivation and improvement of the species, and the transcriptome allowed to elucidate the candidate genes and the main metabolites involved in these phenotypes (Zhou et al., 2020).

Orphan crops have also been presented as a solution to the problem of biotic stress and plant health, since many of them have been shown to be tolerant to some of the pests and diseases when grown in diversified agricultural systems (Hendre et al. 2019).

The growing demand for biofuels has led to the identification of yet another interest for orphan crops, their use as sources of "second generation" biofuels. The use of translational genomics allows the identification and modification of several of the genes, mainly in plant clades that share high levels of genomic synteny between members and include bioenergy cultures (Pancaldi and Trindade, 2020). There are even platforms specifically designed for orphan cultures without a sequenced genome, but for which transcriptomes can be developed, such as, for example, Orphan Crops Browser; (Kamei et al., 2016).

Sequences of genomic data, and especially transcribed data from orphan crops and their wild relatives, provide great potential benefits to plant breeding through: 1) providing valuable genetic resources related to environmental adaptation for the improvement of major crops; 2) construction of new models of biological responses, such as a model for photosynthesis and characteristics of C4 plants to increase cereal productivity, facilitating genome editing; 3) revealing mechanisms that support the ability to compete with major crops and helped to create cultivars with competitive advantages, for example, with elucidation of mechanisms of allelopathic interactions (Guo et al., 2018). Transcriptome allows for the discovery and search of genes involved in the main metabolic pathways and important biotic and abiotic responses to agricultural production.

Transcriptome sequencing has been widely used to detect expression profiles relevant to abiotic stress resilience in plants. Furthermore, this technique has helped most complete genome assembly projects, guiding the complete annotation of the generated genome. This scenario favored the characterization of most model species, but still does not include most orphan species.

Although the ideal for genome annotation is to use transcriptome from the same species that was used for sequencing (Yang et al. 2016; Pati et al. 2019, Song et al. 2019, VanBuren et al. 2020), reality stops in most cases of orphan species, it is the use of the existing transcriptome from a close relative or a well-annotated transcriptome from a model culture. Thus, despite resource limitations, there has been a growing interest in expanding the characterization of transcriptomes from orphan species, in order to broaden the knowledge of specific answers to varied biological questions. Before NGS was cheaper and expanded beyond model species, the microarray method was widely used for transcriptome analysis in orphan cultures, such as white lupine (Zhu et al. 2010), tef (Degu 2019), African belladonna (Solanum nigrum) (Schmidt and Baldwin 2009), wild mustard (Srivastava et al. 2015) and buckwheat (Fagopyrum esculentum) (Golisz et al. 2008). The currently most used method of choice has been RNA sequencing (RNA-seq) (Ozsolak and Milos 2011). For example, Ranasinghe et al. (2019) identified 2,416 differentially expressed genes during the salt stress response profile in quinoa (Chenopodium quinoa). In Juta-malva, a transcription analysis was done to identify genes related to water stress (Yang et al. 2017).

A new trend in this area of transcript analysis is the ISO-seq sequencing by PacBio Sequel II platform. This platform makes it possible to sequence entire transcribed isoforms, from the 5' end to the 3' polyA tail in high quality. The generation of complete transcribed sequences can greatly simplify the analysis, eliminating the need to reconstruct the transcripts by genomic assembly and thus impacting the quality of isoform inference, which is very prone to errors due to the use of short reads. Furthermore, these transcripts can help annotate the construction of new genomes (Pootakham and Nawae et al., 2021).

References

Bray EA, Bailey-Serres J, Weretilnyk E (2000) Responses to abiotic stresses. In: Gruissem W, Buchannan B, Jones R (eds) Biochemistry and molecular biology of plants. American Society of Plant Biologists, Rockville, MD, pp 158–1249

Chandra V, Bajpai O, Singh N (2016) Energy crops in sustainable phytoremediation. Renew Sustain Energy Rev 54:58–73. https://doi.org/10.1016/j.rser.2015.09.078

Chang, Y.,Liu H, Liu M., Liao XZ, Sahu SK, Fu Y, Song B, Cheng SF, Kariba R, Muthemba S, *et al.* (2019). **The draft genomes of five agriculturally important African orphan crops**. Gigascience, 8, p. giy152

Cullis C, Chimwamurombe P, Barker N, Kunert K, Vorster J (2018) Orphan legumes growing in dry environments: marama bean as a case study. Front Plant Sci 9:1–7. https://doi.org/10.3389/fpls.2018.01199

Dawson I.K. , Powell W. , Hendre P. , Bancic J., Hickey J.M. , Kindt R. , Hoad S., Hale I., Jamnadass R. (2019).**The role of genetics in mainstreaming the production of new and orphan crops to diversify food systems and support human nutrition**. New Phytol., 224, pp. 37-54

Fernie A.R. , Yan J.B. (2019). **De novo domestication: an alternative route toward new crops for the future**. Mol. Plant, 12, pp. 615-631.

Hendre PS, Muthemba S, Kariba R, Muchugi A, Fu Y, Chang Y, Song B, Liu H, Liu M, Liao X, Sahu SK, Wang S, Li L, Lu H, Peng S, Cheng S, Xu X, Yang H, Wang J, Liu X, Simons A, Shapiro HY, Mumm RH, Van Deynze A, Jamnadass R (2019) African Orphan crops consortium (AOCC): status of developing genomic resources for African orphan crops. Planta 250:989–1003. https://doi.org/10.1007/s00425-019-03156-9

Hinojosa L, González JA, Barrios-Masias FH, Fuentes F, Murphy KM (2018) Quinoa abiotic stress responses: a review. Plants 7:1–32. https://doi.org/10.3390/plants7040106

Hanbury CD, White CL, Mullan BP, Siddique KHM (2000) A review of the potential of *Lathyrus sativus* L. and *L. cicera* L. grain for use as animal feed. Anim Feed Sci Technol 87:1–27. https://doi.org/10.1016/S0377-8401(00)00186-3

Jamnadass R, Mumm RH, Hale I, Hendre P, Muchugi A, Dawson IK, Powell W, Graudal L, Yana-Shapiro H, Simons AJ, Van Deynze A (2020) Enhancing African orphan crops with genomics. Nat Genet 52:356–360. https://doi.org/10.1038/s41588-020-0601-x

Jiang J, Su M, Chen Y, Gao N, Jiao C, Sun Z, Li F, Wang C (2013) Correlation of drought resistance in grass pea (*Lathyrus sativus*) with reactive oxygen species scavenging and osmotic adjustment. Biol 68:231–240. https://doi.org/10.2478/s11756-013-0003-y

Kamenya, S.N., Mikwa, E.O., Song, B. *et al.* Genetics and breeding for climate change in Orphan crops. *Theor Appl Genet* **134,** 1787–1815 (2021). https://doi.org/10.1007/s00122-020-03755-1

Krishnamurthy L, Upadhyaya HD, Kashiwagi J, Purushothaman R, Dwivedi SL, Vadez V (2016) Variation in drought-tolerance components and their interrelationships in the minicore collection of finger millet germplasm. Crop Sci 56:1914–1926. https://doi.org/10.2135/cropsci2016.03.0191

Khoury CK, Bjorkman AD, Dempewolf H, Ramirez-Villegas J, Guarino L, Jarvis A, Rieseberg L H, Struik P C. (2014). **Increasing homogeneity in global food supplies and the implications for food security**. Proc. Natl. Acad. Sci. U S A, 111, pp. 4001-4006

Mabhaudhi T, Chimonyo VGP, Hlahla S, Massawe F, Mayes S, Nhamo L, Modi AT (2019) Prospects of orphan crops in climate change. Planta 250:695–708. https://doi.org/10.1007/s00425-019-03129-y

Mercati, F., Fontana, I., Gristina, A.S. *et al.* Transcriptome analysis and codominant markers development in caper, a drought tolerant orphan crop with medicinal value. *Sci Rep* **9,** 10411 (2019). https://doi.org/10.1038/s41598-019-46613-x

Pancaldi F, Trindade L M. (2020). Marginal Lands to Grow Novel Bio-Based Crops: A Plant Breeding Perspective. Frontiers in Plant Science: 11, p 227. DOI=10.3389/fpls.2020.00227

Pootakham, W, Nawae, W, Naktang, C, et al. A chromosome-scale assembly of the black gram (*Vigna mungo*) genome. *Mol Ecol Resour*. 2021; 21: 238– 250. https://doi.org/10.1111/1755-0998.13243

Puranik S, Bahadur RP, Srivastava PS, Prasad M (2011) Molecular cloning and characterization of a membrane associated NAC family gene, SiNAC from foxtail millet [*Setaria italica* (L.) P. Beauv.]. Mol Biotechnol 49:138–150. https://doi.org/10.1007/s12033-011-9385-7

Rahman H, Jagadeeshselvam N, Valarmathi R, Sachin B, Sasikala R, Senthil N, Sudhakar D, Robin S, Raveendran M (2014) Transcriptome analysis of salinity responsiveness in contrasting genotypes of finger millet (*Eleusine coracana* L.) through RNA-sequencing. Plant Mol Biol 85:485–503. https://doi.org/10.1007/s11103-014-0199-4Return to ref 2014 in article

Raskin I, Smith RD, Salt DE (1997) Phytoremediation of metals: using plants to remove pollutants from the environment. Curr Opin Biotechnol 8:221–226

Shailaja HB, Thirumeni S (2007) Evaluation of salt-tolerance in finger millet (*Eleusine coracana*) genotypes at seedling stage. Indian J Agric Sci 77:672–674

Shinozaki K, Uemura M, Bailey-Serres J, Bray EA, Weretilnyk E (2015) Responses to abiotic stress. In: Buchanan BB, Gruissem W, Jones RL (eds) Biochemistry and molecular biology. Wiley, Chichester, pp 1051–1100

Thilakarathna MS, Raizada MN (2015) A review of nutrient management studies involving finger millet in the semi-arid tropics of Asia and Africa. Agronomy 5:262–290. https://doi.org/10.3390/agronomy5030262

Tyagi J, Varma A, Pudake RN (2017) Evaluation of comparative effects of arbuscular mycorrhiza (*Rhizophagus intraradices*) and endophyte (*Piriformospora indica*) association with finger millet (*Eleusine coracana*) under drought stress. Eur J Soil Biol 81:1–10. https://doi.org/10.1016/j.ejsobi.2017.05.007

Ye CY, Fan L (2021). Orphan Crops and their Wild Relatives in the Genomic Era. Molecular Plant. Volume 14:1, 27-39. https://doi.org/10.1016/j.molp.2020.12.013.

Zhou, Z., Gao, H., Ming, J. *et al.* Combined Transcriptome and Metabolome analysis of Pitaya fruit unveiled the mechanisms underlying Peel and pulp color formation. *BMC Genomics* **21,** 734 (2020). https://doi.org/10.1186/s12864-020-07133-5

2) Expression analysis/differential expression analysis: Programs, statistical analysis, validation of differential expression analysis

One of the most definitive steps of a transcriptome analysis is the inference of relative, or absolute, gene and/ or isoform expression, by means of quantification and statistical modeling of genetic sequences with a pipeline of bioinformatics tools. An example of a classic approach adopted for molecular expression measurements is the Tuxedo Suite, a combination of the execution of TopHat and Cufflinks programs, in sequence (Trapnell et. al, 2013). Since gene expression is a very dynamic biological mechanism (Roundtree et. al., 2017), the characterization

of its profiles is heavily dependent on time based measurements, the experimental design and control samples (Robles et. al, 2012). Therefore, if the experimental design is not planned cautiously and the data not manipulated accordingly, it can lead to very discrepant outcomes and conclusions.

With the increasing popularity of RNA-Sequencing (RNA-Seq) technology, new and refined approaches to manipulate transcripts data are frequently published and get obsolete very quickly. Considering this scenario, it is very easy for a molecular data analyst to lose track of recent advances and most sophisticated software available. Here, it is described the concepts of the essential steps for a differential expression analysis with RNA-Seq data, in addition to the most accepted and up-to-date software to perform them.

Counting methods

The essential measure that initiates the analyses of a gene expression dataset is the raw count of transcribed reads that were accurately mapped against a reference genome or transcriptome. This preliminary mensuration gives us a more immediate overview of the transcriptome assembly, and check for issues detected in the dataset used or with the parameters applied in previous analysis steps, like wrong sample labeling, detection of outliers, mapping constraints that were too broad or too tight, and even problems in an extraction batch can be identified.

The abundance of transcripts in a RNAseq library can be estimated based on gene level or transcript level. The first can be done by initially removing reads that map to multiple reference locations (multi-mapped) and then count the remaining reads overlapping the same targets (genes, exons) (Soneson et. al., 2015). Tools like HTSeq-count (Anders et. al., 2015) and featureCounts (Liao et. al., 2014) are among the most used for gene level quantification.

The transcript level quantification aims to quantity individual transcripts, and usually can be achieved by two distinct approaches: 1) By performing a transcriptome alignment-based quantification, like employed by RSEM (Li and Dewey, 2011), which aligns reads against an assembled transcriptome using mapping softwares like bowtie2 (Langmead & Salzberg, 2012) or CuffLinks (Trapnell et al., 2010), and then estimates transcripts counts and abundance based on accurate aligned sequences; 2) The second method is based on free-alignment quantification,

which became very popular due to is high accuracy, even in the absence of a traditional mapping and robust genome reference. By eschewing the mapping step also makes them considered 'wicked fast' when compared to other counting methods (Patro et. al., 2014). One way to perform a free-alignment quantification is by constructing a *de Bruijn* graph based on k-mers built from sequence reads, perform a hash lookup (key-values associations) of these k-mers with an indexed transcriptome and then estimate its abundance based on a probabilistic likelihood expectation-maximization algorithm (EM). This procedure was named *pseudoalingment,* and was proposed by the kallisto tool (Bray et. al., 2016), as a way to circumvent, efficiently, the mapping step. Salmon (Patro et. al., 2017) is another, widely used, free-alignment quantification tool, it's computing strategy is similar with the one applied by kallisto, but it executes the hash look up step with a suffix array instead of a *de Brujing* graph*,* this procedure, called *quasi-mapping*, was introduced by the RapMap program (Srivastava et. al., 2016), and it allows the extraction of more information than the *pseudoaligment*. The high popularity of Salmon is also due to its great efficiency on correcting sample-specific bias, like positional and fragment GC content bias and normalization for library size and transcript length, which are overlooked or not properly implemented in other alignment tools, and Salmon manages to achieve it without loosing its speed performance.

Data Normalization and Statistical Procedures

On top of the raw count of reads, a data normalization statistic is applied to decrease noise from sequencing, sampling or preference lower gene counts. The data from gene expression quantification is initially skewed by biases that are primarily originated from library preparation protocols done prior to sequencing and, if overlooked, can greatly impact the results of a different expression analysis (Bullard et. al., 2010). A few normalization methods have already been proposed to adjust the biased data for downstream analysis, and the most relevant are discussed in this topic (Dillies et al. 2013; Li et al. 2015).

The library sequencing does not produce identical number of sequenced reads (sequencing depth) between different samples, and it is one of the most impactful bias present in a RNAseq data. If a library contains more sequenced reads than another, it is most likely to have more counts for a specific read present in both samples, even if they are built from repetitions of a same experimental condition. Ergo, the read counts should be expressed proportionally to its library size,

and this is done by counting up the total reads sequenced present in the library sample and dividing it by 1,000,000, resulting in a scale factor (CPM) that the individual reads count will then be divided by, producing then a reads per million (RPM) value, that will be the sequence depth adjusted read count value (Mortavazi et. al., 2008).

During the RNAseq library preparation, the transcripts molecules are broken down into smaller fragments in order to increase coverage of the entire transcript's lengths, which leads to a higher portion of the sequenced reads to be mapped to longer transcripts than to shorter ones, even if they have a similar expression (Oshlack & Wakefield, 2009), which makes the counting step to be biased by the transcript length, hence it should also be submitted to a normalization correction prior to downstream analysis. This correction is done by dividing the reads count value of a transcript by its length in kilobases, resulting in a read per kilobase value (RPK). These two normalization techniques (RPM and RPK) are commonly done in sequence, resulting in a reads per kilobase per million value, also known as RPKM, or, for paired-end RNAseq, fragments per kilobase per million (FPKM), in which the fact that two reads could be originated from the same transcript will be accounted for, and then won't be counted twice. Li and Dewey (2011) proposed, in the RSEM publication paper, a reversion of the FPKM operations order resulting in unit-less values that would be easier to manipulate and to be used in comparisons between samples, the TPM (transcript per million) metric, as it is called, is done by correcting the length bias first (RPK), then creating a scaling factor by dividing the sum of these RPK values in the samples by 1,000,00 (RPM), and finally dividing these RPM values by the scaling factor, resulting in a TPM count value. Although this combination of normalization procedures accounts for both sequencing depth and gene length, and therefore, they are recommended for gene count comparisons within a sample or between samples of the same group, they, alone, are not suited for different expression analysis.

To be fully adequate for a DE analysis, there is another feature of the RNAseq protocol that should be accounted for: the library composition. If there is a greatly different expressed gene in a sample, it would substantially increase the total number of counts in the sample, resulting in a greater scale factor (CPM) that the read counts of the transcripts of its sample would be divided by, resulting in an apparent lower expression of these reads when compared to a different sample that doesn't have this gene with a very high expression. To make up for this feature, the edgeR packet for R programming language (Mark et al., 2010) proposed a weighted trimmed mean of

log expression ratios between samples, that were named trimmed mean of M-values (TMM). In a whole different fashion, edgeR doesn't use RPKM or TPM at all, because they don't adjust for library composition. EdgeR's normalization approach is based on the presumption that none of the genes are differentially expressed in the sample, so it starts by removing all genes with 0 read counts in all samples, then it picks one of the samples, independently of experimental condition or treatment, that is the most balanced in terms of noise and average read count, to be the reference sample, this sample will then be used to normalize all other samples against. This normalization is done by comparing, with different statistical procedures, sample by sample against the reference sample, filtering out genes that are biased by the log fold differences between them, these biased genes are so called when they are much more expressed in a sample than in the other, or when they are too much or too little expressed in both samples all together, leaving out, then, genes that don't have much bias and have moderate reads mapped to them in both samples. After some weighting, scaling and centering procedures it outputs the final TMM values, that, edgeR then will use for different expression analysis, alongside with calculated p-values and false discovery rates (FDR). To the present moment, the TMM normalization method have been widely accepted and applied in the academic community, and have been classified, in different benchmark papers(), as the most appropriated normalization technique for different expression analysis to the date.

Comparison of different expression profiles

The choice of different methods for differential expression analysis depends on the sample design. Since some of these tools can only perform pairwise comparisons, others such as edgeR, limma-voom (Law et. al., 2014), DESeq (Love et. al., 2014) and maSigPro (Conesa & Nueda, 2006) can perform multiple comparisons. Differential expression analysis packages differ, mainly, in the statistical methods adopted. There are software based on negative binomial distributions (NB), such as edgeR and DESeq; and there are software that use Bayesian approaches based on a negative binomial model, such as baySeq (Hardcastle & Kelly, 2010) and EBSeq (Leng et. al., 2013).

After the comparisons and distribution estimation, a common approach is to test them for the null hypothesis, or the presumption that no genes are expressed differentially. Initially, the

probabilistic significance of the difference between the count of a gene in one sample compared to others (p-value) is calculated, then, this value is adjusted for a False Discovery Rate (FDR), becoming the q-value. If a p-value threshold of 0.05 is set, it means that 5% of all tests are false positives, on the other hand, a q-value of 0.05 would imply that 5% of the significant tests are false positives, resulting in a less stringent and more specific way to filter the abundance data for the most relevant values. The threshold or cutoff value of the q-value can be selected to filter out less significant expression differences, this cutoff can be set depending on the type of the analysis, eg. a cutoff for a co-expression analysis and another for GO enrichment analysis.

The estimation of the log fold change between different expression values is another common procedure in DE analysis, it provides a ratio value related to the changes in genes expression across two samples or conditions. It's value can be negative, meaning that an expression decreased and, therefore, the gene expression is down-regulated, positive for increased expressions (up-regulated genes) and values closer to the baseline (0) for genes not expressed significantly between the samples samples compared.

Visualization

Different expression data are usually organized into tables and/or matrices with thousands of rows representing the genes and columns expressing different types of numeric metrics, like raw counts, normalized counts, p-values, FDR and so on. Several types of charts can be plotted to monitor every step of the pipeline to ensure that the tools and the data are being properly manipulated.

A bar plot is very useful to compare samples abundance before and after a library size normalization procedure.

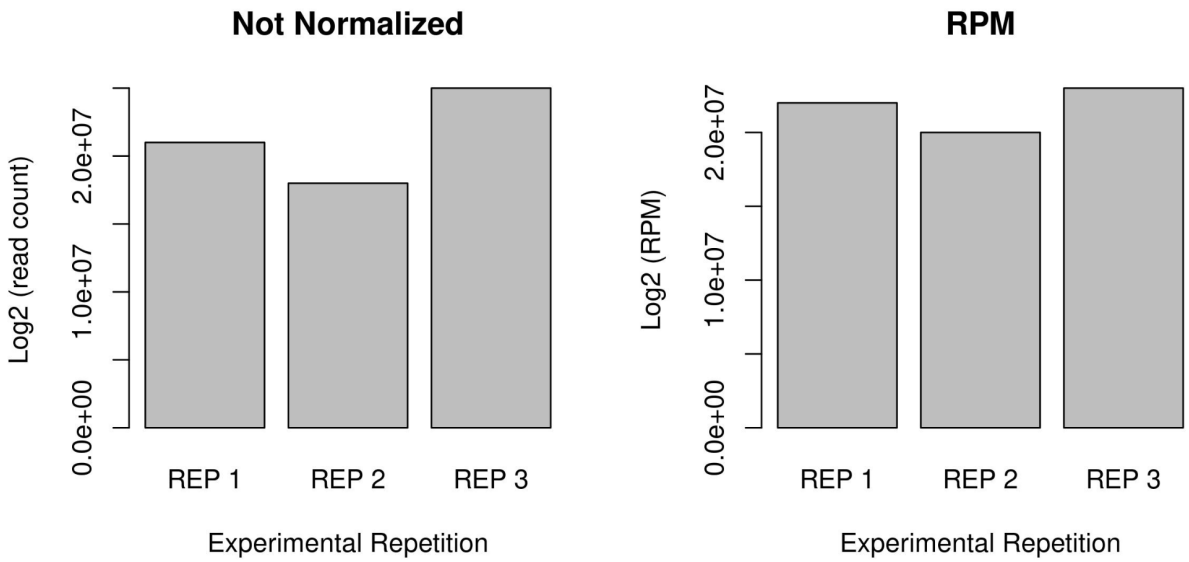


Figure. Bar plots contrasting genes abundance in different samples after a library size normalization

Multi-Dimensional Scaling plots (MDS) are widely used to check if the expression profiles of the samples are behaving accordantly with the experimental conditions, it converts the distance between the gene counts values across samples to a 2D graph, grouping samples based on their similarities, allowing for validation of experimental replication and identification of outliers that will have to be accounted for, or excluded from subsequent analysis.

The correlation plot is another alternative to visualize similarities between gene expression profiles across different samples, but instead of Cartesian data points as produced in the MDS plot, the correlation plots present the similarities based on a spectrum of colors, the more similar the samples are, the closer their colors will look on the chart.

For an overview of the relationship between expression changes across two experimental conditions (log fold change, M) and the log of the mean of normalized gene counts(A), an MA plot (2D scatter plot) can be used. Genes with no relevant expression changes between the experimental conditions will cluster around M=0, those that are plotted away from M=0 are colored in opposition to those that are not, resulting in a colored contrast. As stated above, this plot is only useful for a general comparison between experimental conditions, since it doesn't consider

p-values or adjusted p-values it can not tell if the different expression of the genes are actually statistical significant.

Validation

The validation process is an important part of gene expression studies, as it allows verifying the reliability and accuracy of the transcriptome data and validating the differential expression results. In this scenario, reverse transcription quantitative PCR (RT-qPCR) has become a powerful technology to detect and quantify gene expression in biological samples because of its sensitivity, rapid execution and accuracy, even at low expression levels (Kubista et al., 2006; Derveaux et al., 2010).

In summary, RT-qPCR technique involves seven main steps: (i) availability of experimental samples; (ii) extraction of total RNA from experimental samples; (iii) evaluation of RNA concentration and quality; (iv) synthesis of cDNA from extracted total RNA through reverse transcription; (v) optimizing conditions for the qPCR assay; (vi) running the qPCR reaction under optimized conditions; and (vii) data analysis using appropriate normalization methods (Kuang et al., 2018; Taylor et al., 2019). Despite the accuracy and sensitivity of the qPCR technique to quantify gene expression, the quality of the results obtained may be affected by variations in the way the steps described above are conducted in different laboratories. Also, the precision of this technique is influenced by several factors including the quality of RNA, PCR efficiency, and very important, for accurate normalization strategies and selection of suitable reference genes, which must be stable for analyzed experimental conditions (Phillips et al., 2009). These factors can influence the result of the technique and make it difficult to compare results between transcriptomic studies (Radonic et al., 2014; Kuang et al., 2018).

The main motivation in the validation step using the RT-qPCR technique is to confirm whether the genes are really expressed differently in the analyzed samples and if there is biological reproducibility between them. For this, it is important to access a minimal expression variation in the tissue, treatment or condition to be analyzed (Pombo et al., 2019). Also, it is recommended to validate a set that includes genes that exhibited down-regulation, up-regulation or no significant change once the selected genes are validated by the technique and confirm the reliability of RNA-Seq results.

Reference

Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics (Oxford, England)*, *31*(2), 166–169. https://doi.org/10.1093/bioinformatics/btu638

Berardini, T. Z., Mundodi, S., Reiser, L., Huala, E., Garcia-Hernandez, M., Zhang, P., Mueller, L. A., Yoon, J., Doyle, A., Lander, G., Moseyko, N., Yoo, D., Xu, I., Zoeckler, B., Montoya, M., Miller, N., Weems, D., & Rhee, S. Y. (2004). Functional annotation of the Arabidopsis genome using controlled vocabularies. *Plant physiology*, *135*(2), 745–755. https://doi.org/10.1104/pp.104.040071

Bray, N., Pimentel, H., Melsted, P. *et al.* Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* 34, 525–527 (2016). https://doi.org/10.1038/nbt.3519

Bullard, J.H., Purdom, E., Hansen, K.D. *et al.* Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* 11, 94 (2010). https://doi.org/10.1186/1471-2105-11-94

Conesa, A., Nueda, M. J., Ferrer, A., & Talón, M. (2006). maSigPro: a method to identify significantly differential expression profiles in time-course microarray experiments. *Bioinformatics*, *22*(9), 1096-1102.

Derveaux, S., Vandesompele, J. & Hellemans, J. How to do successful gene expression analysis using real-time PCR. *Methods* 50, 227–230 (2010).

Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D, Estelle J, Guernec G, Jagla B, Jouneau L, Laloë D, Le Gall C, Schaëffer B, Le Crom S, Guedj M, Jaffrézic F; French StatOmique Consortium. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Brief Bioinform. 2013 Nov;14(6):671-83. doi: 10.1093/bib/bbs046. Epub 2012 Sep 17. PMID: 22988256.

Francesca Finotello, Barbara Di Camillo, Measuring differential gene expression with RNA-seq: challenges and strategies for data analysis, *Briefings in Functional Genomics*, Volume 14, Issue 2, March 2015, Pages 130–142, https://doi.org/10.1093/bfgp/elu035

Hardcastle, T.J., Kelly, K.A. baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics* 11, 422 (2010). https://doi.org/10.1186/1471-2105-11-422

Kuang, J., Yan, X., Genders, A. J., Granata, C., & Bishop, D. J. (2018). An overview of technical considerations when using quantitative real-time PCR analysis of gene expression in human exercise research. *PloS one*, *13*(5), e0196438. https://doi.org/10.1371/journal.pone.0196438

Kubista M, Andrade JM, Bengtsson M, Forootan A, Jonak J, Lind K, Sindelka R, Sjoback R, Sjogreen B, Strombom L, et al. The real-time polymerase chain reaction. Mol Aspects Med. 2006;27(2–3):95–125.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Law, CW, Chen, Y, Shi, W, and Smyth, GK (2014). Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biology 15, R29

Leng, N., Dawson, J. A., Thomson, J. A., Ruotti, V., Rissman, A. I., Smits, B. M., Haag, J. D., Gould, M. N., Stewart, R. M., & Kendziorski, C. (2013). EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics (Oxford, England)*, *29*(8), 1035–1043. https://doi.org/10.1093/bioinformatics/btt087

Li, B., Dewey, C.N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12, 323 (2011). https://doi.org/10.1186/1471-2105-12-323

Li, P., Piao, Y., Shon, H.S. *et al.* Comparing the normalization methods for the differential analysis of Illumina high-throughput RNA-Seq data. *BMC Bioinformatics* 16, 347 (2015). https://doi.org/10.1186/s12859-015-0778-7
Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014 Apr 1;30(7):923-30. doi: 10.1093/bioinformatics/btt656. Epub 2013 Nov 13. PMID: 24227677.

Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15, 550 (2014). https://doi.org/10.1186/s13059-014-0550-8

Mark D. Robinson, Davis J. McCarthy, Gordon K. Smyth, edgeR: a Bioconductor package for differential expression analysis of digital gene expression data, Bioinformatics, Volume 26, Issue 1, 1 January 2010, Pages 139–140, https://doi.org/10.1093/bioinformatics/btp616

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods, 5(7), 621–628. doi:10.1038/nmeth.1226

Oshlack, A., Wakefield, M.J. Transcript length bias in RNA-seq data confounds systems biology. *Biol Direct* 4, 14 (2009). https://doi.org/10.1186/1745-6150-4-14

Patro, R., Mount, S. M., & Kingsford, C. (2014). Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nature biotechnology*, *32*(5), 462–464. https://doi.org/10.1038/nbt.2862

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature methods*, *14*(4), 417–419. https://doi.org/10.1038/nmeth.4197

Phillips, M. A., D'Auria, J. C., Luck, K., & Gershenzon, J. (2009). Evaluation of Candidate Reference Genes for Real-Time Quantitative PCR of Plant Samples Using Purified cDNA as Template. *Plant molecular biology reporter*, *27*(3), 407–416. https://doi.org/10.1007/s11105-008-0072-1

Pombo, M.A., Ramos, R.N., Zheng, Y. et al. Transcriptome-based identification and validation of reference genes for plant-bacteria interaction studies using Nicotiana benthamiana. Sci Rep 9, 1632 (2019). https://doi.org/10.1038/s41598-018-38247-2

Radonic A, Thulke S, Mackay IM, Landt O, Siegert W, Nitsche A. Guideline to reference gene selection for quantitative real-time PCR. Biochem Biophys Res Commun. 2004;313(4):856–62.

Robles, J.A., Qureshi, S.E., Stephen, S.J. et al. Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing. BMC Genomics 13, 484 (2012). https://doi.org/10.1186/1471-2164-13-484

Roundtree IA, Evans ME, Pan T, He C. Dynamic RNA Modifications in Gene Expression Regulation. Cell. 2017 Jun 15;169(7):1187-1200. doi: 10.1016/j.cell.2017.05.045. PMID: 28622506; PMCID: PMC5657247.

Srivastava A, Sarkar H, Gupta N, Patro R. RapMap: a rapid, sensitive and accurate tool for mapping RNA-seq reads to transcriptomes. Bioinformatics. 2016 Jun 15;32(12):i192-i200. doi: 10.1093/bioinformatics/btw277. PMID: 27307617; PMCID: PMC4908361.

Taylor, S.C., Nadeau, K., Abbasi, M., Lachance, C., Nguyen, M., & Fenrich, J. (2019). The Ultimate qPCR Experiment: Producing Publication Quality, Reproducible Data the First Time. *Trends in biotechnology, 37* *7*, 761-774.

Trapnell, C., Roberts, A., Goff, L. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* 7, 562–578 (2012). https://doi.org/10.1038/nprot.2012.016

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L., & Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols*, *7*(3), 562–578. https://doi.org/10.1038/nprot.2012.016

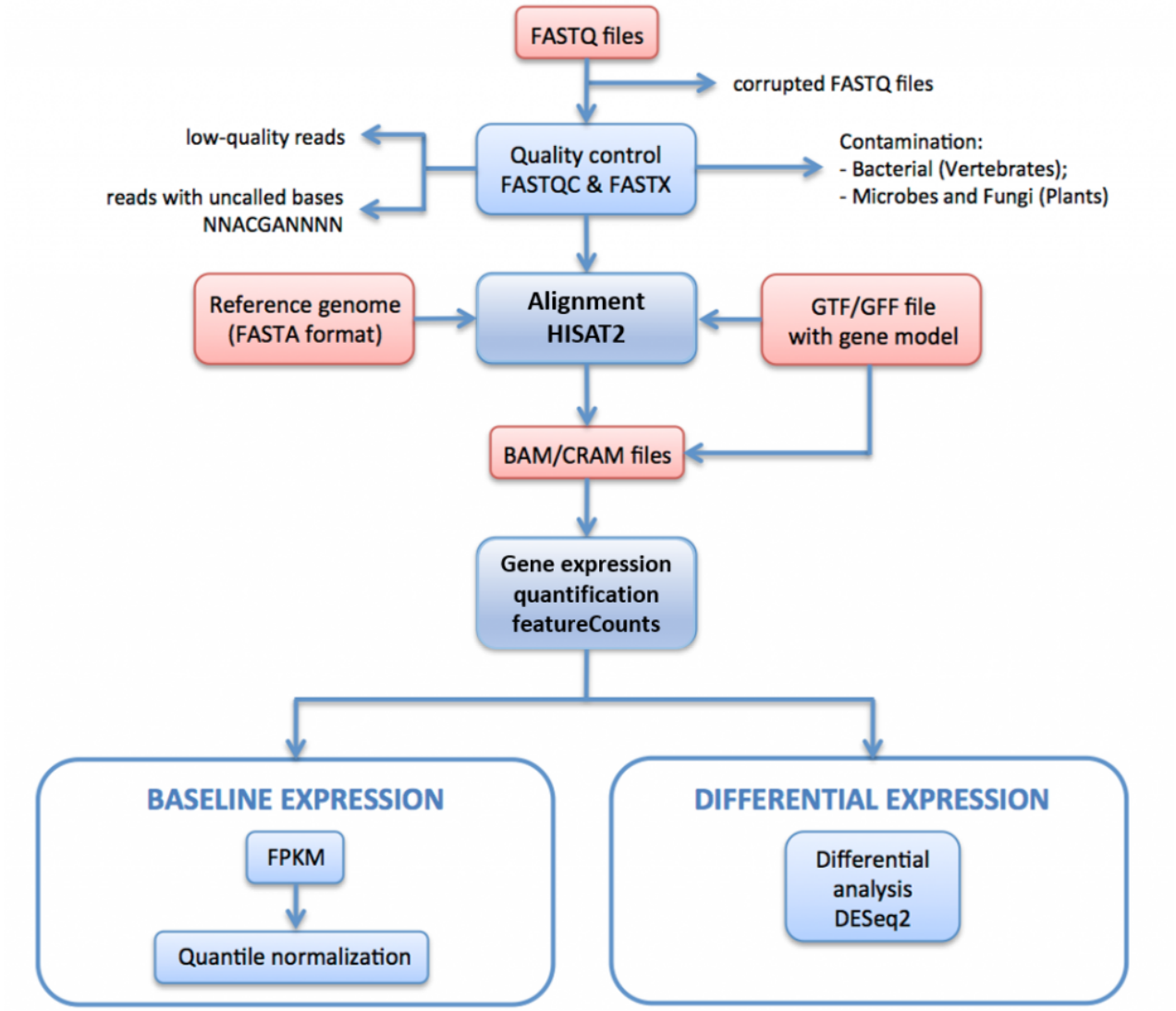


Figure: RNA-seq processing pipeline used to generate gene expression data in Expression Atlas.

8. *CAPÍTULO IV*: PRINCIPAIS RESULTADOS

A seguir estão apresentados os principais resultados referentes ao capítulo I da tese, cujo objetivo foi a aplicação de SG, ML, GWAS e redes de co-expressão gênica em populações de melhoramento de *P. taeda*:

- Um total de 19.933 (~80%) indivíduos sobreviventes foram avaliados nas análises fenotípicas para característica de volume, apresentando herdabilidade moderada variando entre 0.17 e 0.32 nos sete locais de coleta.
- 1.692 indivíduos foram genotipados em quatro ambientes e 45 famílias. O processo de genotipagem foi realizado com um protocolo de captura de sequências por sondas previamente desenvolvidas.
- Para produzir o conjunto de dados final usado no GWAS, usamos parâmetros rigorosos (máximo de 10% de genótipos ausentes em um SNP e frequências alélicas mínimas e máximas de 0,05 e 0,95, respectivamente), resultando em 31.589 SNPs distribuídos entre quase todas as sondas, com quantidades diferentes dependendo do comprimento da sequência de referência.
- A PCA realizada com os dados do SNP sugeriu a presença de clusters genéticos para cada família, o que não foi observado quando consideradas as diferentes localizações das populações.
- O decaimento de LD foi estimado com o quadrado do coeficiente de correlação de Pearson (R^2) entre pares de SNP ligados dentro das respectivas sondas. Observamos decaimento rápido para a metade do máximo R2 = 0,05 em menos de 100 Kb.
- Foram selecionados experimentos de transcriptomas, em bancos de dados públicos, de três espécies de pinus, Pta, Pra e Pel, para complementar a anotação das sequências genotípicas analisadas e para a construção das redes de co-expressão gênica.
- O conjunto de transcritos de cada experimento foi utilizado para montagem *de novo* do transcriptoma de cada espécie, gerando uma quantidade final de 38.472, 37.499 e 33.374 genes para Pta, Pra e Pel, respectivamente.

- A partir das análises de GWAS, foram identificados sete *loci* candidatos para o volume em *P. taeda*. Para evitar vieses ambientais, incluímos os locais e a estrutura familiar como covariável no modelo GWAS final.
- Usando o conjunto de dados do SNP, criamos modelos de previsão genômica usando a metodologia BRR em um cenário de validação cruzada de 10 vezes repetido 100 vezes. Com o conjunto total de 31.589 SNPs, a acurácia preditiva média (coeficiente R Pearson) foi de 0,79 (erro quadrático médio de 0,00023), mostrando a alta capacidade preditiva do conjunto de dados empregado para predizer VS.
- Diferentes técnicas de seleção de atributos foram usadas para reduzir o tamanho dos dados de SNP para análise de SG (i) GTB (498 SNPs recuperados); (ii) Correlações de Pearson (20.957 SNPs recuperados); e (iii) SVM (11.440 SNPs recuperados). A partir de tais técnicas, selecionamos um subconjunto final de SNPs para estimar o modelo preditivo de volume como a interseção de pelo menos dois dos três métodos de FS, totalizando 7.864 SNPs (~24,89% dos dados iniciais do marcador). A partir desses marcadores, conseguimos obter resultados de precisão semelhantes, (0,78), mostrando que a maior parte da variância fenotípica pode estar associada a esse conjunto de dados menor.
- Usando esse conjunto de dados selecionado, empregamos uma abordagem baseada em ML para recuperar associações adicionais de fenótipo-genótipo. Usando modelos DT e RF, estimamos a importância das características de cada SNP para prever volume, avaliamos sua distribuição em boxplots separados e recuperamos outliers (1.506 para DT e 635 para RF). A partir desses outliers, selecionamos a interseção desses marcadores encontrados em cada modelo, formando um conjunto final de 128 SNPs, que foram investigados juntamente com os resultados do GWAS. Curiosamente, todos os SNPs associados ao GWAS foram incluídos neste conjunto de ML definido.
- Usando uma abordagem WGCNA, modelamos uma rede diferente para cada espécie Pra e Pel, considerando um soft-power β estimado de 14 para Pel (R^2 de ~0,90 e conectividade média de ~53,10) e Pra (R^2 de ~0,81 e conectividade média de ~143,20). 251 (tamanhos mínimo e máximo de 81 e 14616 com média de 471,79) e 83 grupos (tamanhos mínimo e máximo de 50 e 4858 com média de 132,96) puderam ser estabelecidos para as redes Pra e Pel, respectivamente.

- Para cada um dos grupos definidos em cada rede, avaliamos a presença de genes associados a: (i) marcadores GWAS (7 SNPs); (ii) marcadores associados por LD com (i) (7 SNPs); e (iii) marcadores ML estabelecidos (128 SNPs).
- Considerando uma quantidade mínima de dois genes associados a (i) em cada grupo, realizamos análises de enriquecimento para termos GO do processo biológico. Dois grupos (G12 e G159) da rede de expressão Pel, com expressão de 3 marcadores selecionados ML, foram enriquecidos para catabolismo de lignina, relacionando a importância desses genes envolvidos em reações químicas e vias que resultam na quebra de lignina para associar genótipos com a característica de volume. O processo catabólico da lignina também apareceu enriquecido na rede de expressão de Pra, em 3 grupos (G13, G36 e G57) envolvendo a co-expressão de 8 marcadores selecionados por ML.
- Outra ortologia genética interessante encontrada em grupos (G13, G15 e G19, G65) com co-expressão de 9 marcadores selecionados por ML foram o processo metabólico da celulose e a celulose biossintética em Pra. G12, da rede Pel, também apresentou enriquecimento do processo metabólico da hemicelulose.

Como principais resultados para a análise da expressão diferencial e construção de redes de co-expressão gênica em experimento de frio em seringueira, apresentada no capítulo II da tese, temos:

- O experimento de frio consistiu na exposição de três mudas com seis meses de idade de cada genótipo (RRIM600 e GT1) a uma temperatura de 10ºC por 24 horas (12h claro/12h escuro). Durante esse período, o material foliar foi coletado em séries temporais de 0 horas (0h - controle), 90 minutos (90m), 12 horas (12h) e 24 horas (24h). O RNA extraído dessas amostras foi empregado para construir bibliotecas de cDNA, e posteriormente foi sequenciado usando um Analisador de Genoma Illumina IIx.
- A partir do sequenciamento foi realizado uma montagem *de novo*, no qual foram identificadas 104.738 isoformas (51.697 genes) com tamanhos entre 500 bp e 22.333 bp, com média de 1.874 bp e N50 de 2.369. A partir desses genes, pudemos identificar um total de 74.398 proteínas únicas relacionadas a 9.417 GOs diferentes. Para as análises de DEG, empregamos um critério de filtragem baseado em genes que apresentaram pelo

menos 10 contagens por milhão (CPM) em pelo menos 3 amostras, o que resultou em uma quantidade final de 30.407 genes.

- PCA empregado revelou um perfil distinto de amostras pertencentes a cada genótipo diferente.
- A comparação do número de DEGs por período de tempo apresentou respostas diferentes, não apenas quanto ao número de DEGs mas também quanto aos genes associados. Pode se observar claramente um mecanismo conjunto de vários genes para resposta ao frio. Em relação à análise de enriquecimento GO, o conjunto DEG completo foi enriquecido para resposta ao estresse e resposta aos termos de quitina. O grupo de genes com regulação negativa não apresentou nenhuma categoria super-representada. Em contraste, o grupo de genes regulados positivamente foi enriquecido para os seguintes termos: resposta à quitina, resposta de defesa, resposta ao ferimento, morte celular, fotoproteção, atividade do transportador transmembranar de íons de cálcio e integral à membrana plasmática.
- Após 90 minutos de tratamento a frio, apenas trinta e dois transcritos foram caracterizados como DEGs, todos eles foram regulados durante a resposta inicial das plântulas ao estresse por frio. Destes, 26 genes foram anotados com sucesso. Os DEGs anotados são conhecidos por serem induzidos pelo frio, bem como por outros estressores abióticos, como a seca e o sal. Eles estão envolvidos com processos de endurecimento da parede celular, inibição do crescimento, proteção contra o estresse oxidativo, sinalização de cálcio (Ca2+) e processamento de RNA induzido pelo frio. A resposta tardia, considerada como o intervalo entre doze horas e 24 horas de tratamento a frio, continha 18 GDEs, incluindo genes que estimulam o aumento das barreiras de difusão da membrana e modificações da parede celular.
- A partir dos dados do transcriptoma de ambos os genótipos, modelamos uma rede HRR para representar a resposta ao frio da espécie. Essa rede é composta por 27.220 nós em 50.650 conexões (bordas). Apenas os 10 melhores coeficientes de correlação de Pearson R considerando um ponto de corte de 0,8 foram retidos para cada gene, sendo desconsiderados os genes desconectados. O principal componente conectado dentro da rede é composto por 25.608 nós (94,1%).

- Para avaliar a importância dos genes dentro desses módulos, avaliamos a presença de genes centrais dentro de cada cluster, os hubs. Para definir um gene como hub, selecionamos aqueles com o maior valor de degree. Embora a estrutura da rede permita a inserção de dez arestas para cada nó, respeitando a correlação mínima de Pearson de 0,8, não observamos essa quantidade máxima em todos os clusters.
- Segundo a rede de co-expressão gênica, construída neste trabalho, o ponto de partida da resposta ao estresse por frio da seringueira reside principalmente no cluster c14, que abriga várias sequências anotadas como fatores de transcrição responsivos ao etileno (ERFs). Os ERFs são ativados pela via de sinalização do etileno e regulam a transcrição de genes responsivos ao etileno, que impactam negativamente o crescimento das plantas. A família ERF em seringueira é composta por 115 membros e a superexpressão de um ortólogo ERF1 de arabidopsis em seringueiras aumentou sua tolerância a estresses abióticos e diferenciação celular de laticíferos aumentada.
- Em nossa análise, todos os graus anotados como ERFs foram regulados positivamente, sugerindo que as plântulas de H. *brasiliensis* mantiveram a resposta ao estresse mediada por etileno ativa durante todo o experimento. A regulação positiva das enzimas da biossíntese de etileno também indica que o etileno desempenha um papel constitutivo na resposta ao estresse por frio da seringueira. Surpreendentemente, apenas seis enzimas responderam precocemente ao tratamento a frio e todas foram colocadas neste cluster também e quatro delas são ubiquitina transferases E3 do tipo RING.

9. DISCUSSÃO

A aplicação inovadora de metodologias de aprendizado de máquina e biologia de sistemas no desenvolvimento do estudo de associação gênica e seleção genômica em uma espécie florestal, de genoma altamente complexo, possibilitou bons resultados de acurácia preditiva e identificação de genes candidatos para serem implementados no melhoramento de pinus, como apresentado no capítulo I da presente tese.

Neste sentido, os resultados apresentados no artigo do capítulo I possibilitaram a identificação de um reduzido e preciso conjunto de marcadores genéticos (128 marcadores) com influência significativa na característica de volume, nas populações analisadas em pinus, validado pela combinação das análises GWAS, ML e SG. A aplicação dos modelos de DT e RF, para recuperar associações adicionais de fenótipo-genótipo, utilizando o dataset identificado por FS para GS, possibilitou o estreitamento da grande lista de SNPs importantes para predição genômica, que influencia para volume em pinus.

Esses marcadores genéticos identificados em nossos resultados são avanços importantes para o melhoramento genético de pinus. A combinação inédita de diferentes estratégias de genética quantitativa e bioinformática para análise de dados de populações de irmãos completos pinus provou ser uma forma eficiente de selecionar genótipos com melhores resultados para volume. O uso das redes de coexpressão genica em combinação com as marcas de GWAS e ML, possibilitaram a identificação de módulos funcionais chaves para o entendimento dos processos biológicos envolvidos com expressão dos genes associados a volume.

Os resultados do capítulo II desta tese mostram a eficiência da adoção das análises utilizando redes de co-expressão gênica em associação com estudos de expressão diferencial, para identificação de genes chaves na resposta ao frio em seringueira. O artigo apresentado complementou um estudo anterior do grupo, realizado apenas com análises de genes diferencialmente expressos, possibilitando aprofundar a investigação e análise dos dados de transcriptomas em seringueira. A construção de redes de co-expressão permitiu a visualização da reprogramação primária da expressão gênica de *H. brasiliensis* e a relação entre os genes envolvidos na resposta ao estresse pelo frio. Foi possível identificar que, no curto período de exposição ao frio, as seringueiras ativam a via de sinalização mediada pelo etileno e mantem a

sinalização do etileno ativa. Também foi possível identificar que o PCD (*programmed cell death*) desempenha um papel importante no processo de resposta ao frio da seringueira, sendo fortemente regulado pelas cascatas de sinalização. A inibição do crescimento e o espessamento da parede celular são implementados pelas plântulas, o que pode ser correlacionado a uma possível estratégia de fuga à seca desencadeada pelo estresse por frio. A elucidação de como o *H. brasiliensis* lida com o estresse por baixas temperaturas pode melhorar muito as estratégias de melhoramento para esta cultura, bem como enfatizar a importância da preservação da diversidade genética na espécie.

10. CONCLUSÃO

A aplicação inovadora de diferentes ômicas (genômica, transcritpômica, proteômica, etc) integrada às redes de co-expressão gênica, associada às metodologias de aprendizado de máquina e biologia de sistemas no desenvolvimento do estudo de associação gênica, seleção genômica e identificação de genes e marcadores de interesse, lavaram ao aumento da acurácia preditiva e à identificação de genes de grande importância para abordagens visando o melhoramento molecular de espécies florestais.

11. PERSPECTIVAS

Os trabalhos desta tese apresentaram múltiplas estratégias para se trabalhar com um amplo conjunto de dados genômicos e fenotípicos em espécies florestais. Ao longo das análises, nós identificamos a necessidade de usar novas abordagens estatísticas, incluindo o uso de diferentes algoritmos de aprendizado de máquina para predição genômica, em pinus, e uso de integração de dados genômicos e transcriptômicos para modelar interações de SNPs, genes e fenótipos, em pinus e seringueira.

Desta forma, os esforços do presente trabalho em desenvolver uma nova metodologia de análise integrada de dados genômicos e fenotípicos, puderam não apenas validar a metodologia em Pinus (com dados do Lab. do Prof Matias Kirst, da UF), mas como também tem grande perspectiva para implementação em uma ampla gama de espécies florestais, incluindo a seringueira, espécie alvo de estudo do segundo capítulo da tese. A aplicação destes novos modelos preditivos em pinus, tal como sua combinação com análises de associação genômica e redes de co-expressão gênica, também servirá de modelo para a aplicação destas múltiplas análises ômicas em uma população de melhoramento de seringueira, cujo enfoque é o estudo da influência de porta-enxerto para seleção de clones superiores.

Este é um dos trabalhos pioneiros do nosso grupo voltado à aplicação conjunta da biologia de sistemas em espécies florestais, unindo as análises de Seleção Genômica, GWAS e redes de coexpressão gênica. Os resultados gerados irão contribuir com novas perspectivas, tanto para trabalhos em andamento quanto para futuras pesquisas. Além disso, o contínuo avanço nos métodos estatísticos e de bioinformática deste trabalho, abre um grande leque de possibilidades para que outros estudos sejam realizados utilizando o mesmo conjunto de dados.

12. BIBLIOGRAFIA

Amadeu, R. R., C. Cellon, J. W. Olmstead, A. A. F. Garcia, M. F. R. Resende, and P. R. Muñoz. 2016 AGHmatrix: R Package to Construct Relationship Matrices for Autotetraploid and Diploid Species: A Blueberry Example. The Plant Genome 9(30).

Anacleto, R., Badoni, S., Parween, S., Butardo Jr, V.M., Misra, G., et al. 2018 Integrating a genome- wide association study with a large-scale transcriptome analysis to predict genetic regions influencing the glycaemic index and texture in rice. Plant Biotechnol. J. 17: 1261-1275.

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.

Azodi CB, Tang J, Shiu SH. 2020 Opening the Black Box: Interpretable Machine Learning for Geneticists. Trends Genet. 36(6):442-455.

Bai Y, Kissoudis C, Yan Z, Visser RGF, van der Linden G. 2018 Plant behaviour under combined stress: tomato responses to combined salinity and pathogen stress. Plant J. 93(4):781–93.

Balmant K M, Noble J D, Alves F C, Dervinis C, Conde D, Schmidt H W, Vazquez A I, Barbazuk W B, de los Campos G, Resende M F R and Kirst M. 2020 Xylem systems genetics analysis reveals a key regulator of lignin biosynthesis in Populus deltoides. Genome Res. 30: 1131-1143

Baker, E. A. et al. Comparative transcriptomics among four white pine species. G3: Genes, Genomes, Genet. 8, 1461–1474 (2018).
Brown G R, Gill G P, Kuntz R J, Langley C H, Neale D B, 2004 Nucleotide diversity and linkage disequilibrium in loblolly pine. *Proc. Natl. Acad. Sci. USA* 101: 15255–15260

Bayer, P. E., Petereit, J., Danilevicz, M. F., Anderson, R., Batley, J., & Edwards, D. (2021). The application of pangenomics and machine learning in genomic selection in plants. *The Plant Genome*, *14*(3), e20112.

Breiman, L. 2001 Random forests. Mach. Learn. 45 (1), 5–32.

Burdon RD, Wilcox PL . (2011). Integration of molecular markers in breeding. In: Plomion C, Bousquet J, Kole C (eds) *Genetics, Genomics and Breeding of Conifers*. CRC Press and Edenbridge Science Publishers: New York. pp 276–322.

Calleja-Rodriguez, A., Pan, J., Funda, T., Chen, Z., Baison, J., Isik, F., ... & Wu, H. X. (2020). Evaluation of the efficiency of genomic versus pedigree predictions for growth and wood quality traits in Scots pine. *BMC genomics*, *21*(1), 1-17.

Chen, Z.-Q.; Baison, J.; Pan, J.; Westin, J.; Gil, M.R.G.; Wu, H.X. Increased prediction ability in Norway spruce trials using a marker × environment interaction and non-additive genomic selection model. *J. Hered.* **2019**, *110*, 830–843.

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

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

Chen, C., Chen, X., Han, J., Lu, W. & Ren, Z. Genome-wide analysis of the wrky gene family in the cucumber genome and transcriptome-wide identification of wrky transcription factors that respond to biotic and abiotic stresses. BMC plant biology 20, 1–19 (2020).

Conson, André R. O. ; Taniguti, Cristiane H. ; Amadeu, Rodrigo R. ; Andreotti, Isabela A. A. ; De Souza,

Cristianini, N., J. Shawe-Taylor, et al. 2000 An introduction to support vector machines and other kernel- based learning methods. Cambridge university press.

Cros, D., Mbo-Nkouloud, L., Bell, J. M., Oum, J., Masson, A., Soumahoro, M., et al. 2019 Within-family genomic selection in rubber tree (Hevea brasiliensis) increases genetic gain for rubber production. Ind. Crops Prod. 138, 111464.

Dash, M. & Liu, H. Feature selection for classification. *Intell. Data Anal.* **1**, 131–156 (1997).

de Los Campos, G., J. M. Hickey, R. Pong-Wong, H. D. Daetwyler, and M. P. L. Calus, 2013 Whole- genome regression and prediction methods applied to plant and animal breeding. Genetics 193: 327– 345.

de los Campos, G., Pérez-Rodríguez, P., Bogard, M. et al. 2020 A data-driven simulation platform to predict ”cultivars’ performances under uncertain weather conditions. Nat Commun 11, 4876

De La Torre, A. R. et al. Genomic architecture of complex traits in loblolly pine. New Phytol. 221, 1789–1801 (2019).

de Oliveira Junkes, C. F. et al. Resin tapping transcriptome in adult slash pine (pinus elliottii var. elliottii). Ind. Crop. Prod.139, 111545 (2019).

Eckert, A. J. et al. High-throughput genotyping and mapping of single nucleotide polymorphisms in loblolly pine (pinus taeda l.). Tree Genet. & Genomes 5, 225–234 (2009).

Eckert, A. J. et al. Patterns of population structure and environmental associations to aridity across the range of loblolly pine (pinus taeda l., pinaceae). Genetics 185, 969–982 (2010).

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.

Esposito S, Carputo D, Cardi T, Tripodi P. 2019 Applications and Trends of Machine Learning in Genomics and Phenomics for Next-Generation Breeding.Plants (Basel). 9(1):34.

Flint-Garcia S A, Thornsberry J M, Buckler E S, 2003 Structure of linkage disequilibrium in plants. *Annu. Rev. Plant Biol.* 54: 357–374.

Ferreira, M.A., Gamazon, E.R., Al-Ejeh, F. Aittomäki, K., Andrulis, AL, et al. 2019 Genome-wide association and transcriptome studies identify target genes and risk loci for breast cancer. Nat. Commun. 10: 1741.

Francisco, F. R., Hild Aono, A., Da Silva, C. C., Gonçalves, P. S., Scaloppi Junior, E. J., Le Guen, V., ... & Pereira de Souza, A. (2021). Unravelling rubber tree growth by integrating GWAS and biological network-based approaches.

Freund, Y. and R. E. Schapire (1997). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of computer and system sciences 55 (1), 119–139.

Garbeva P, Weisskopf L. 2020 Airborne medicine: bacterial volatiles and their influence on plant health. New Phytol;226(1):32–43.

Garrison, EMarthG. 2012 Haplotype-based variant detection from short-read sequencing. preprint arXiv:12073907

Gilmour A. R., Gogel B. J., Cullis B. R., Thompson R. 2006 ASReml User Guide Release 2.0. VSN International, Hemel Hempstead, UK.

Gonzalez de Andres, E. et al. Tree-to-tree competition in mixed european beech–scots pine forests has different impacts on growth and water-use efficiency depending on site conditions. J. Ecol. 106, 59–75 (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

Heffner, E. L., M. E. Sorrells, and J.-L. Jannink, 2009 Genomic Selection for Crop Improvement. Crop Sci. 49: 1–12.

Hickey, J. M. *et al.* Evaluation of genomic selection training population designs and genotyping strategies in plant breeding programs using simulation. *Crop Sci.* **54**, 1476–1488 (2014).

Ingvarsson, P. K. & Street, N. R. Association genetics of complex traits in plants. New Phytol. 189, 909–922 (2011).
Isik, F. & McKeand, S. E. Fourth cycle breeding and testing strategy for pinus taeda in the nc state university cooperative tree improvement program. Tree Genet. & Genomes 15, 1–12 (2019).

Jombart T., Solymos P. 2008 adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics 24:1403-1405.

Kainer, D.; Lanfear, R.; Foley, W.J.; Külheim, C. Genomic approaches to selection in outcrossing perennials: Focus on essential oil crops. *Theor. Appl. Genet.* **2015**, *128*, 2351–2365. Schaefer, R. J., Michno, J. M., Jeffers,

Kononenko I, Kukar M. Machine Learning and Data Mining: Introduction to Principles and Algorithms. London: Horwood Publishing; 2007.

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.
Li, J. *et al.* Feature selection: A data perspective. *ACM Comput. Surv. CSUR* **50**, 94 (2018).

Livia M. ; Dos Santos, Luciano H. B. ; Rosa, João R. B. F. ; Mantello, Camila C. ; Da Silva, Carla C. ; José Scaloppi Junior, Erivaldo ; Ribeiro, Rafael V. ; Le Guen, Vincent ; Garcia, Antonio A. F. ; Gonçalves, Paulo de Souza ; De Souza, Anete P 2018 High-Resolution Genetic Map and QTL Analysis of Growth-Related Traits of Hevea brasiliensis Cultivated Under Suboptimal Temperature and Humidity Conditions. Frontiers in Plant Science , v. 9, p. 1-16

Liu, X., Li, Y. I., & Pritchard, J. K. (2019). Trans effects on gene expression can drive omnigenic inheritance. *Cell*, *177*(4), 1022-1034.

Lu, M. et al. Exome genotyping, linkage disequilibrium and population structure in loblolly pine (pinus taeda l.). BMC genomics 17, 1–11 (2016).

Lu, M. et al. Association genetics of growth and adaptive traits in loblolly pine (pinus taeda l.) using whole-exome-discovered polymorphisms. Tree Genet. & Genomes 13, 57 (2017).

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).

Mantello, C; Boatwright, L; Silva, C C; Scaloppi Jr, E J; Goncalves, P S; Barbazuk W B and De Souza, A P. 2019 Deep expression analysis reveals distinct cold-response strategies in rubber tree (Hevea brasiliensis). BMC Genomics. 20:455. https://doi.org/10.1186/s12864-019-5852-5

Mantello, C. C., Cardoso-Silva, C. B., Da Silva, C. C., De Souza, L. M., Scaloppi, E. J., de Gonçalves, P. S., et al. 2014 De Novo assembly and transcriptome analysis of the rubber tree (Hevea brasiliensis) and SNP markers development for rubber biosynthesis pathways. PLoS One 9:e102665. doi: 10.1371/journal.pone.0102665

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

Montesinos-López, O. A., Martín-Vallejo, J., Crossa, J., Gianola, D., Hernández-Suárez, C. M., Montesinos-López, A., Juliana P., & Singh, R. (2019b). New deep learning genomic-based prediction model for multiple traits with binary, ordinal, and continuous phenotypes. *G3: Genes, Genomes, Genetics*, **9**, 1545–1556

Mphahlele, M.M., Isik, F., Mostert-O'Neill, M.M. et al. 2020. Expected benefits of genomic selection for growth and wood quality traits in Eucalyptus grandis. Tree Genetics & Genomes 16, 49.

Namin, S. T., Esmaeilzadeh, M., Najafi, M., Brown, T. B., Borevitz, J. O. (2018). Deep phenotyping: deep learning for temporal phenotype/genotype classification. *Plant Methods* 14, 66. doi: 10.1186/s13007-018-0333-4

NOAA National Centers for Environmental Information. (2020). *State of the climate: Global climate report for July 2020*. https://www.ncdc.noaa.gov/sotc/global/202007

Neves, L. G., Davis, J. M., Barbazuk, W. B. & Kirst, M. A high-density gene map of loblolly pine (pinus taeda l.) based on exome sequence capture genotyping. G3: Genes, Genomes, Genet. 4, 29–37 (2014).

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).

Pavy N, Namroud M-C, Gagnon F, Isabel N, Bosquet J, 2012a The heterogeneous levels of linkage disequilibrium in white spruce genes and comparative analysis with other conifers. *Heredity* 108: 273–284.

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

Pedregosa, Fabian, et al. 2011 ""Scikit-learn: Machine learning in Python"." Journal of machine learning research 12: 2825-2830.

Platt A, Vilhjálmsson B J, Nordborg M, 2010 Conditions under which genome–wide association studies will be positively misleading. *Genetics* 186: 1045–1052

Pritchard J K, Przeworski M, 2001 Linkage disequilibrium in humans: Models and data. *Am. J. Hum. Genet.* 69: 1–14.

Plomion, C., Chagné, D., Pot, D., Kumar, S., Wilcox, P. L., Burdon, R. D., ... & Islam-Faridi, M. N. (2007). Pines. In *Forest Trees* (pp. 29-92). Springer, Berlin, Heidelberg.

Polle, A., Chen, S. L., Eckert, C. & Harfouche, A. Engineering drought resistance in forest trees. Front. plant science 9,1875 (2019).

Popescu, M.-C., V. E. Balas, L. Perescu-Popescu, and N. Mastorakis 200. Multilayer perceptron and neural networks. WSEAS Trans. Cir. and Sys. 8 (7), 579–588.

Rasmussen, C. E. 2004 Gaussian Processes in Machine Learning. Advanced Lectures on Machine Learning. Col: Lecture Notes in Computer Science. 3176. pp. 63–71

Schaefer, R. J., Michno, J. M., Jeffers, J., Hoekenga, O., Dilkes, B., Baxter, I., & Myers, C. L. (2018). Integrating coexpression networks with GWAS to prioritize causal genes in maize. *The Plant Cell*, *30*(12), 2922-2942.

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

Souza LM, Francisco FR, Gonçalves PS, et al. 2019.Genomic Selection in Rubber Tree Breeding: A Comparison of Models and Methods for Managing G×E Interactions.Front Plant Sci. 2019;10:1353.

Spindel, J. E. , and McCouch S. R.. 2016. When more is better: How data sharing would accelerate genomic selection of crop plants. *New Phytologist* 212(4): 814–826. 10.1111/nph.14174.

Souza, L. M; Santos L. H; Rosa, J. R. J; Silva, C. C; Mantello, C. C; Conson, A. R; Scaloppi, JR. E. J. Fialho, J. D; Moraes, M. L. T; Gonçalves, P; Margarido, A. A; Le Guen, V; Souza, A. P. 2018. Linkage Disequilibrium and

Population Structure in Wild and Cultivated Populations of the Rubber Tree (Hevea brasiliensis); Frontiers in Plant Science, 2018.

Telfer, E. et al. A high-density exome capture genotype-by-sequencing panel for forestry breeding in pinus radiata. PloSone 14, e0222640 (2019).

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

Turner-Hissong S. D, Bird K. A, Lipka A. E, King E. G, Beissinger T. M and Angelovici R, 2020 Genomic Prediction Informed by Biological Processes Expands Our Understanding of the Genetic Architecture Underlying Free Amino Acid Traits in Dry Arabidopsis Seeds. G3: Genes, Genomes, Genetics: 120.401240

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

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

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

Yan J, Shah T, Warburton M L, Buckler E S, McMullen M D et al. , 2009 Genetic characterization and linkage disequilibrium estimation of a global maize collection using SNP markers. *PLoS ONE* 4: e8451.

Xu, Y. , Liu X., Fu J., Wang H., Wang J., Huang C., Prasanna B. M., et al. 2020. Enhancing genetic gain through genomic selection: From livestock to plants. *Plant Communications* 1(1): 100005. 10.1016/j.xplc.2019.100005.

Zapata-Valenzuela, J.; Isik, F.; Maltecca, C.; Wegrzyn, J.; Neale, D.; McKeand, S.; Whetten, R. SNP markers trace familial linkages in a cloned population of *Pinus taeda*—Prospects for genomic selection. *Tree Genet. Genomes* **2012**, *8*, 1307–1318.

Zapata-Valenzuela, J.; Whetten, R.W.; Neale, D.; McKeand, S.; Isik, F. Genomic estimated breeding values using genomic relationship matrices in a cloned population of loblolly pine. *G3 Genes Genomes Genet.* **2013**, *3*, 909–916.

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.

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

13. ANEXOS

Durante o doutorado, eu realizei dois períodos de estágio no exterior: um ano no laboratório do Prof. Matias Kirst, na Universidade da Florida (abril de 2018 a abril de 2019) e; três meses de estágio com o grupo de pesquisa dos Professores Ian Mackay e Wayne Powell, no Scotland's Rural College (SRUC) (novembro de 2019 a fevereiro de 2020). Durante este período de doutorado sanduíche, foram desenvolvidos trabalhos estreitamente relacionados à linha de estudos quantitativos e genômicos que desenvolvemos em meu projeto de pesquisa de doutorado.

No começo no doutorado, com o interesse em aplicar genômica ao melhoramento de árvores, eu iniciei um contato com o Prof. Matias Kirst, líder do laboratório de genômica florestal na Universidade da Flórida, Estados Unidos, visando um estágio em um dos laboratórios pioneiros na aplicação de seleção genômica em espécies arbóreas. O Dr. Kirst, além de professor de genética quantitativa, é coordenador do programa de pós graduação "Plant Molecular and Cellular Biology Program (PMCB)" e possui amplo conhecimento em Bioinformática e estudos quantitativos concentrados na arquitetura genômica de populações de espécies florestais.

Contemplada por uma bolsa sanduiche vinculada ao projeto "Biologia Computacional na Análise da variação Genética" (aprovado na Chamada CAPES Biologia Computacional Edital 051/2013), coordenado pela Profa. Anete Pereira de Souza, eu puder realizar um ano de estágio no laboratório do Dr. Kirst, trabalhando em um projeto para desenvolver análises de seleção genômica aplicada a duas espécies de pinus do sul dos Estados Unidos.

Considerando as limitações de orçamento destinados a tecnologia por parte de produtores florestais e, o fato da genotipagem requerida para as análises de seleção genômica demandar bastante investimento, eu trabalhei em meu projeto de sanduiche, com análises voltadas a seleção genômica por família florestais, ao invés de genotipagem por indivíduo. A seleção de famílias pode ser uma alternativa à SG tradicional, reduzindo custos de genotipagem ao focar o desenvolvimento e seleção do modelo para as famílias e, reduzindo o número de amostras necessárias para análise.

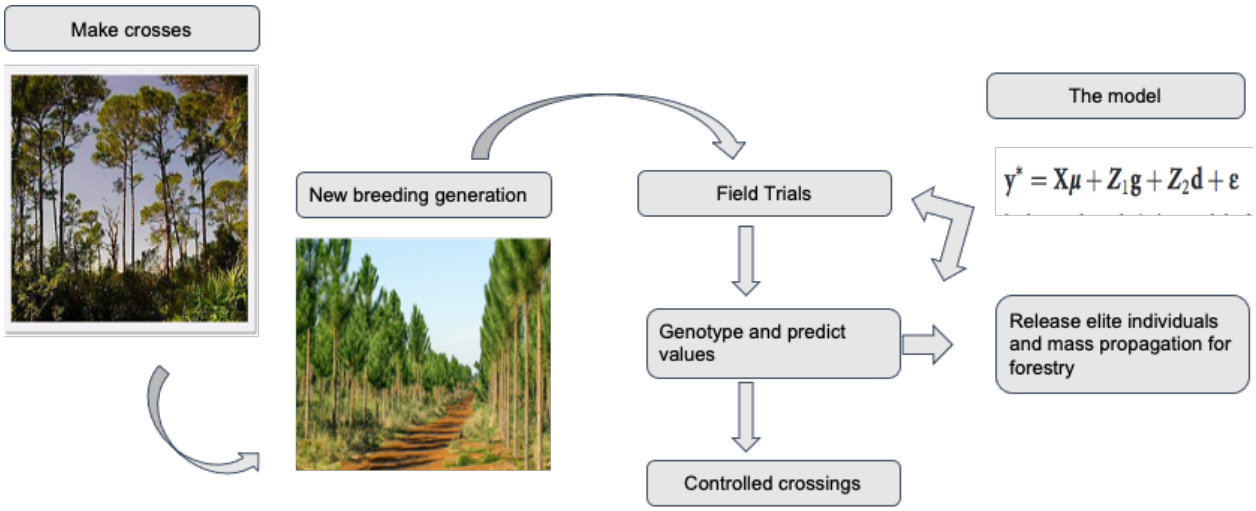


Figura 1: Seleção Genômica no programa de melhoramento de pinus

Para avaliar essa abordagem, foi desenvolvido um pipeline para obter e avaliar dados por família de uma população de melhoramento de pinus, para futura implementação da seleção de famílias usando SG.

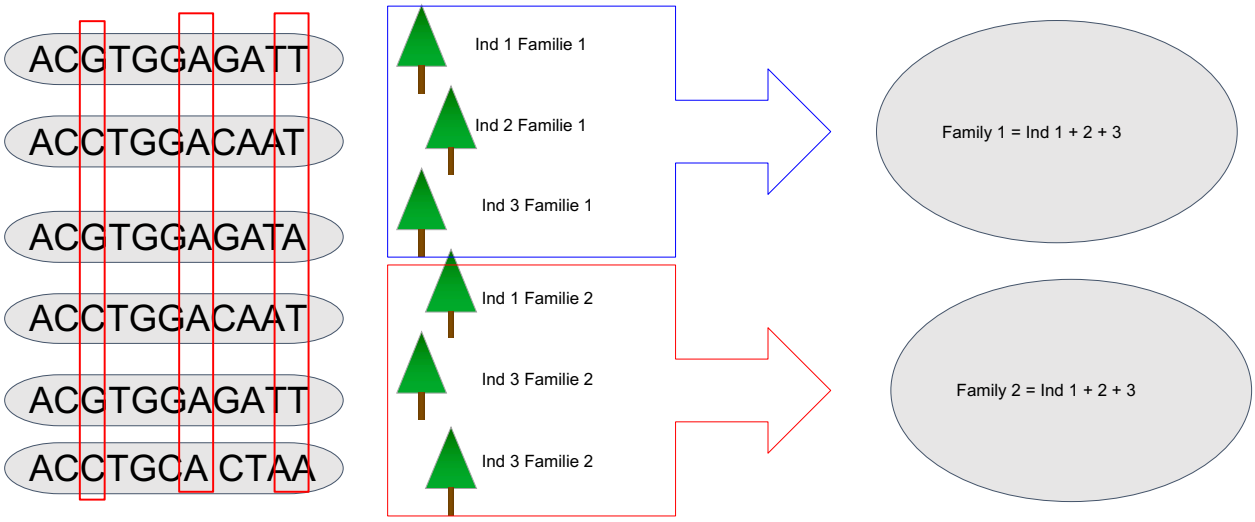


Figura 2: Metodologia para genotipagem por família.

Para esta análise, foram obtidos dados genômicos de 60 famílias compostas por 999 indivíduos de uma população reprodutora. Os dados genômicos foram obtidos usando captura de sequência e a chamada de SNPs foi realizada usando o software FreeBayes, seguidos de etapas de filtragem usando vcflib. Após a filtragem, um índice de informações da família foi adicionado no arquivo vcf, e foram geradas estimativas de profundidade de leitura total (DP), profundidade de leitura de alelos alternativos e de referência (AD e RD, respectivamente). A informação da família de profundidade de leitura foi definida como a soma da profundidade de leitura de todos os indivíduos dentro de uma família. Os genótipos foram obtidos pela razão entre AD e DP. Esses dados variaram continuamente entre 0 e 1. Um total de 60.333 SNPs foi usado na análise. Para avaliar esses dados uma matriz de relacionamento foi calculada usando o pacote AGHmatrix e os parâmetros de genética de populações foram estimados usando o pacote R adegenet. Após a análise com dados faltantes, foram selecionados 56896 (3437 dados excluídos) marcadores. Após remover valores faltantes <0,10 tivemos 2% de dados faltantes remanescentes que foram substituídos pela média das frequências contínuas. Qualquer família tinha menos de 50% de data ausente. O DAPC indicou a presença de dois grupos entre as famílias. A análise de heterozigosidade mostrou uma frequência populacional de AA (p2) = 151443846; Aa (2pq) = 30937102; Aa (q2) = 13562113 e frequência relativa AA = 0,77289721; Aa = 0,15788822 e aa = 0,06921456. Os valores de acurácia de seleção genômica por família foram correspondentes aos valores indicados para a seleção por indivíduos, em torno de 0.64.

Este trabalho foi fundamental para a consolidação do aprendizado desta metodologia de análises de genética quantitativa em meu projeto, que de sequência as análises de seleção genômica em pinus por indivíduos, devido ao número de famílias analisadas.

Em meu segundo período de doutorado sanduíche, no SRUC, fui selecionada para trabalhar no projeto “Desenvolvimento de uma plataforma para a genética e melhoramento da alga vermelha *Palmaria palmata* para compostos bioativos de alto valor".

Palmaria palmata é uma alga vermelha quem vem sendo utilizada para consumo humano durante séculos. A sua distribuição ocorre ao redor das zonas de maré baixa e subtidal das costas americanas e europeias do Atlântico Norte. Atualmente *P. palmata* não é cultivada comercialmente, embora isso seja necessário, pois a crescente demanda por algas, para várias utilidades, resulta em maiores pressões sobre os estoques selvagens. A produção experimental no mar em cordas e em terra em tanques está sob investigação. O desenvolvimento deste projeto e interesse em seleção genômica de *P. palmata* surgiu diante de um mercado em ascensão focado na utilização desta alga para seguintes finalidades:

- Produtos baseados em seu valor nutricional
- Produtos da extração de compostos bioativos de alto valor para uso em produtos farmacêuticos e cosméticos,
- Produção de biocombustíveis
- Aditivo alimentar para ruminantes, considerando o efeito antimetanogênico das algas vermelhas, que tem sido atribuído à presença de bioativos como bromofórmio e diclorometano. Estudos recentes mostraram o potencial das algas vermelhas para reduzir as emissões de metano em até 99%, o que poderia promover maiores taxas de crescimento e eficiência de conversão alimentar em ruminantes.

Desta forma, eu trabalhei para o desenvolvimento de estudos genéticos com ênfase nos efeitos antimetanogênicos e de características bioquímicas de alto valor de *P. palmata*. O objetivo desta pesquisa foi implementar melhorias na identificação de locais para produção off-shore com germoplasma coletado localmente e para melhoramento genético de estoque para cultivo on-shore. Além disso, o projeto focou no desenvolvimento de marcadores moleculares para *P.palmata*, pesquisando a variação genética ao redor das costas do Reino Unido, para o desenvolvimento de populações de mapeamento e detectando QTL para características bioquímicas nessas populações.

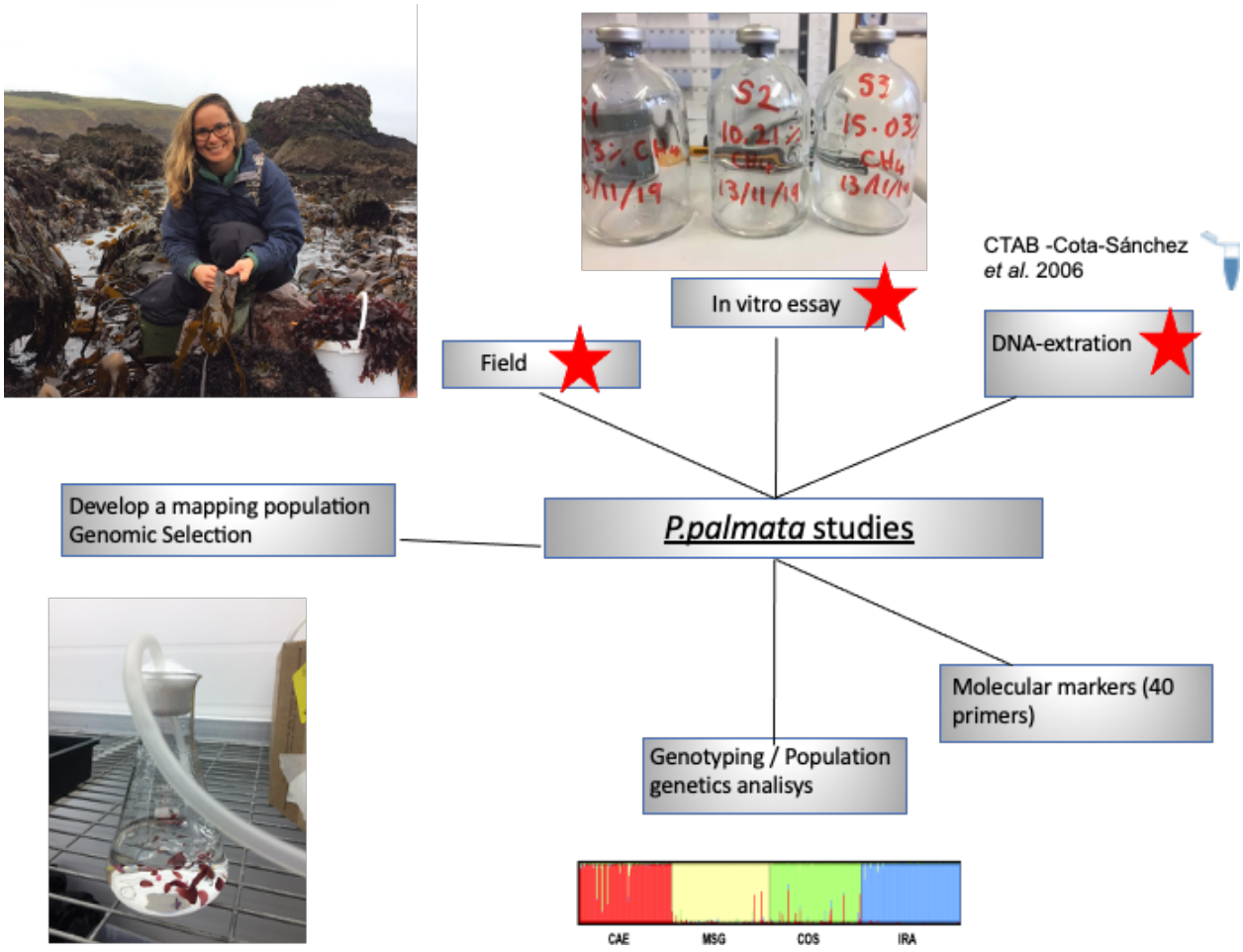


Figura 3: Plano de trabalho de *P. palmata* no SRUC.

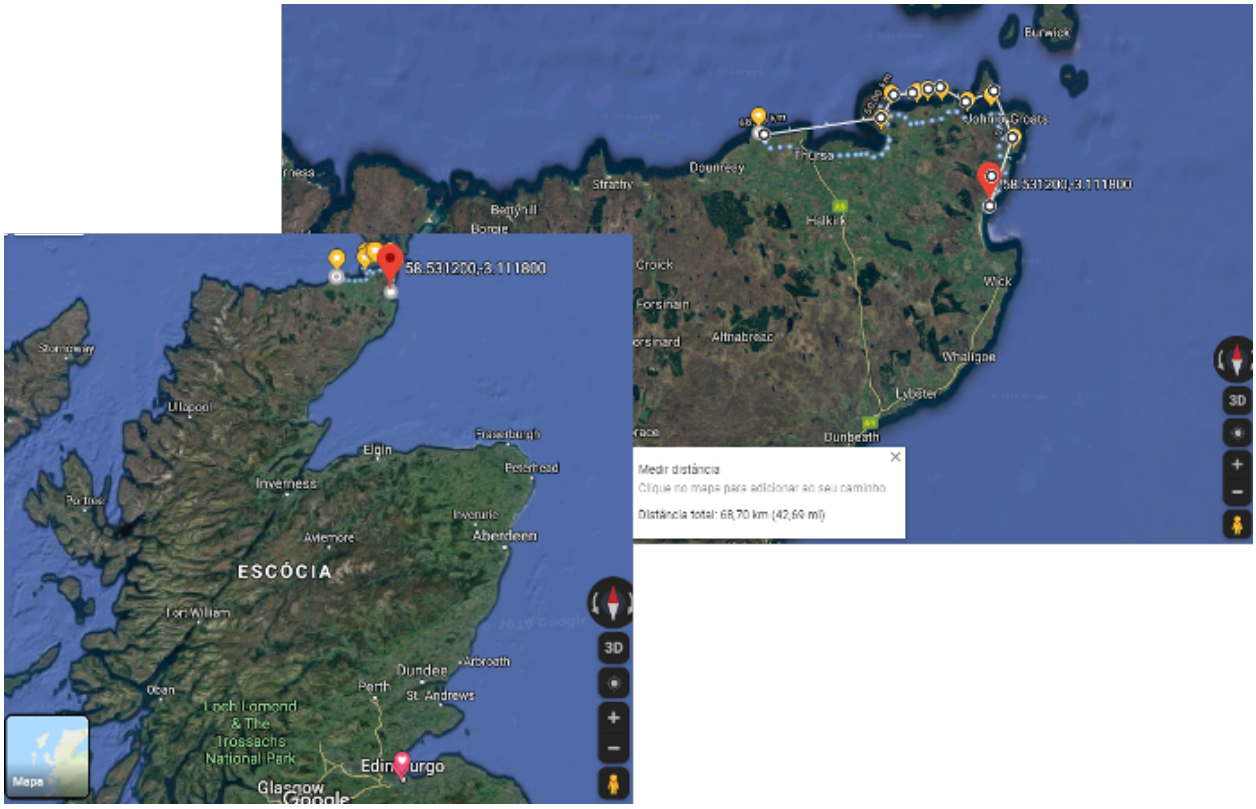


Figura 4: Locais de coleta de *P. palmata* para análise do experimento in vitro.

Para o estudo da associação entre características de interesse econômico nas algas vermelhas, medidas *in silico* e dados genéticos das mesmas amostras (base para seleção genômica), foi primeiro realizada uma etapa de desenvolvimento de primers para estudo da diversidade genética nas amostras coletas.

Tabela 1: *Primers* desenvolvidos para análise da diversidade de P.palmata nas amostras coletadas.

Primers	Based on	Size	Application
1) rrl ribosomal RNA	mitochondrion, complete genome	INCLUDED REGION SIZE: 3638 PRODUCT SIZE: 692	Phylogenetic and to explore deeper relationships providing a summary of PCR primers and profiles to that time
2) Cytochrome **apocytochrome b (COB)**	mitochondrion, complete genome	INCLUDED REGION SIZE: 1143 PRODUCT SIZE: 876	Phylogenetic
3) cytochrome oxidase subunit 1 (COI) gene	BARCODE (Bringloe,T. and Saunders,G.W) - mitochondrial	INCLUDED REGION SIZE: 642 PRODUCT SIZE: 605	Phylogenetic
4) rps3 ribosomal protein S3	mitochondrion, complete genome	INCLUDED REGION SIZE: 684 PRODUCT SIZE: 613	Phylogenetic
5) rbcL ribulose-1,5-bisphosphate	plastid, complete genome	INCLUDED REGION SIZE: 1467 PRODUCT SIZE: 849	Phylogenetic (barcode)
6) ITS (**internal transcribed spacer 1)** B	genomic sequence (Saunders,G.W. et l, 2017)	REGION SIZE: 923 PRODUCT SIZE: 799	Population level analyses, and for resolving relationships
7) **cox2 cytochrome c oxidase subunit 2**	mitochondrion, complete genome	INCLUDED REGION SIZE: 774 PRODUCT SIZE: 705	Population level
8) psaA photosystem I P700 chlorophyll a apoprotein A1	plastid, complete genome	INCLUDED REGION SIZE: 2259 PRODUCT SIZE: 1620 Literature (psaA, ~1,600 bp)	Phylogenetic
9) psaB photosystem I P700 chlorophyll a apoprotein A2	plastid, complete genome	INCLUDED REGION SIZE: 2205 PRODUCT SIZE: 1270 Literature (psaB, ~1,250 bp)	Phylogenetic
10) psbA photosystem II protein D1	plastid, complete genome	INCLUDED REGION SIZE: 1083 PRODUCT SIZE: 909 Literature (psbA, ~950 bp)	Phylogenetic

Para gerar os dados fenotípicos a serem avaliados nas análises de genética quantitativa, eu realizei um experimento *in vitro*, para avaliação da implementação de *P. palmata* na suplementação bovina e consequente redução na emissão de metano durante o processo de digestão. Este experimento consistiu na avaliação do efeito *in vitro* de amostras de *Palmaria palmata* (diferentes locais, estações e anos) na fermentação ruminal e produção de metano.

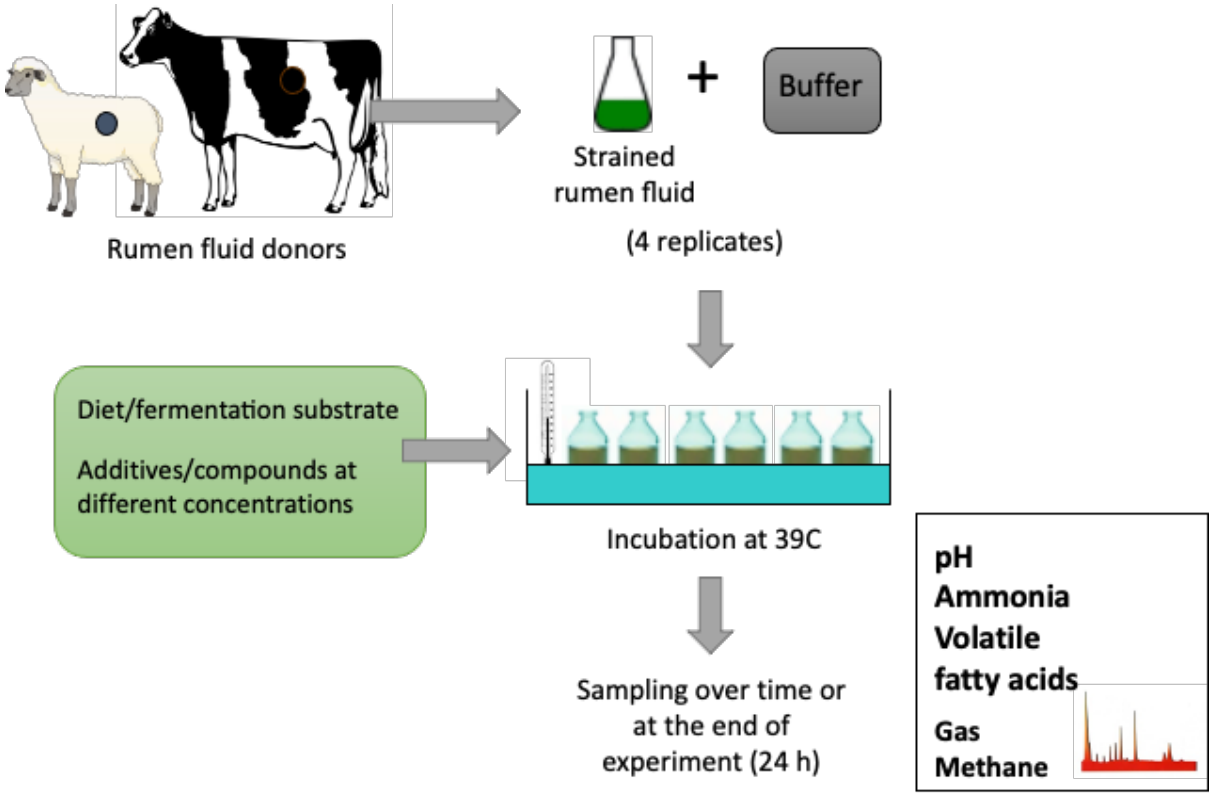


Figura 5: Culturas em *in vitro* - Parâmetros de fermentação

A realização deste experimento consistiu inicialmente na seleção de 50 amostras secas de *Palmaria palmata* de 122 amostras (diferentes locais, estações e anos) fornecidas pela New Wave Food para teste. O delineamento experimental consistiu de um controle (sem aditivo) e 50 amostras de *P. palmata* adicionadas a 2 g/L da incubação -> 60 mg por frasco). O experimento foi conduzido em quadruplicata, utilizando fluido ruminal de quatro vacas Jersey canuladas em fenação, em duas semanas consecutivas (2 vacas por semana).

O conteúdo ruminal foi filtrado através de musselina e diluído 1:2 em solução tampão de saliva (Menke e Steingass (1978). Alíquotas (30 mL) do líquido ruminal coado diluído foram adicionados anaerobicamente a frascos de 120 mL de trigo contendo 0,3 g de uma mistura ração (50:50 feno:cevada), previamente moída para passar por peneira de malha de 1 mm2. Os frascos foram lacrados e incubados a 39◦C, recebendo uma mistura suave antes da amostragem às 24 h.

O padrão de fermentação, em termos de pH, amônia, ácidos graxos voláteis (AGV) e emissões de metano, foi determinado após 24 h da incubação. Após o registro da produção de gás, uma amostra foi retirada do headspace para a determinação da concentração de metano por cromatografia gasosa. Em seguida, as garrafas foram abertas e o pH medido. Uma subamostra (1,6

mL) foi diluída com 0,4 mL de solução desproteinizante (200 mL/L de ácido ortofosfórico contendo 20 mmol L/ de ácido 2-etilbutírico como padrão interno) para a determinação de AGV por cromatografia gasosa, conforme descrito por Stewart e Duncan (1985). Outra subamostra (0,8 mL) foi diluída com 0,2 mL de ácido tricloroacético a 25% (p/vol) para análise de amônia por método colorimétrico (Weatherburn 1967).

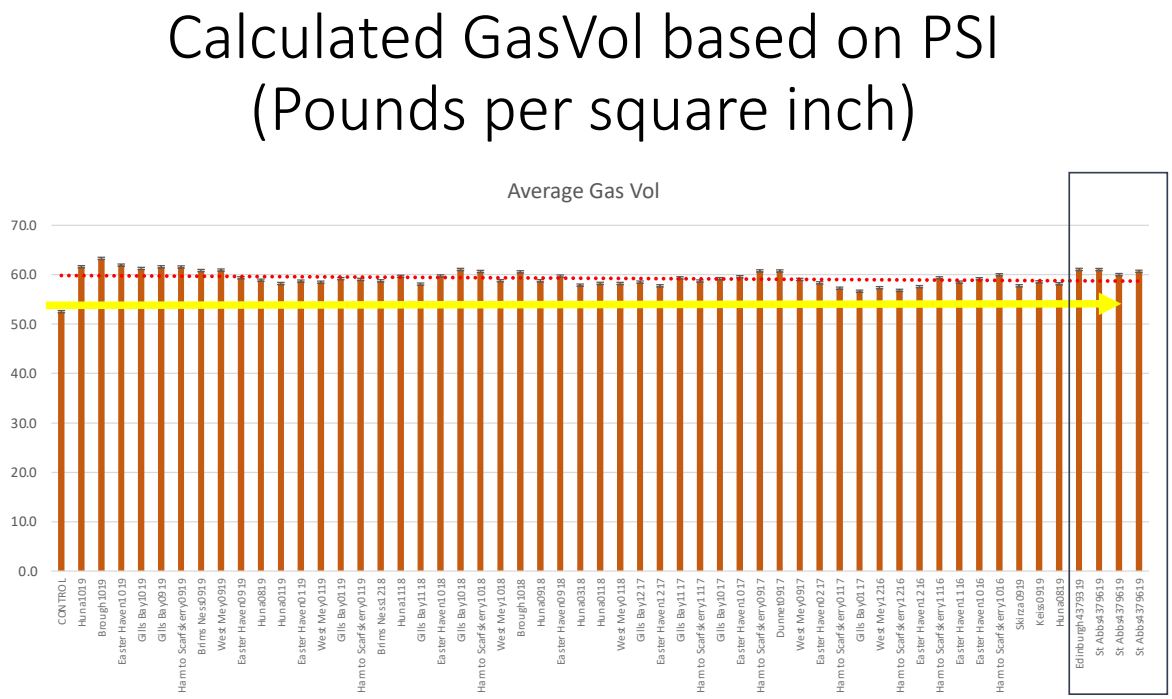


Figura 6: Resultado da cromatografia gasosa da produção de gases voláteis por amostra.

Methane (mL)

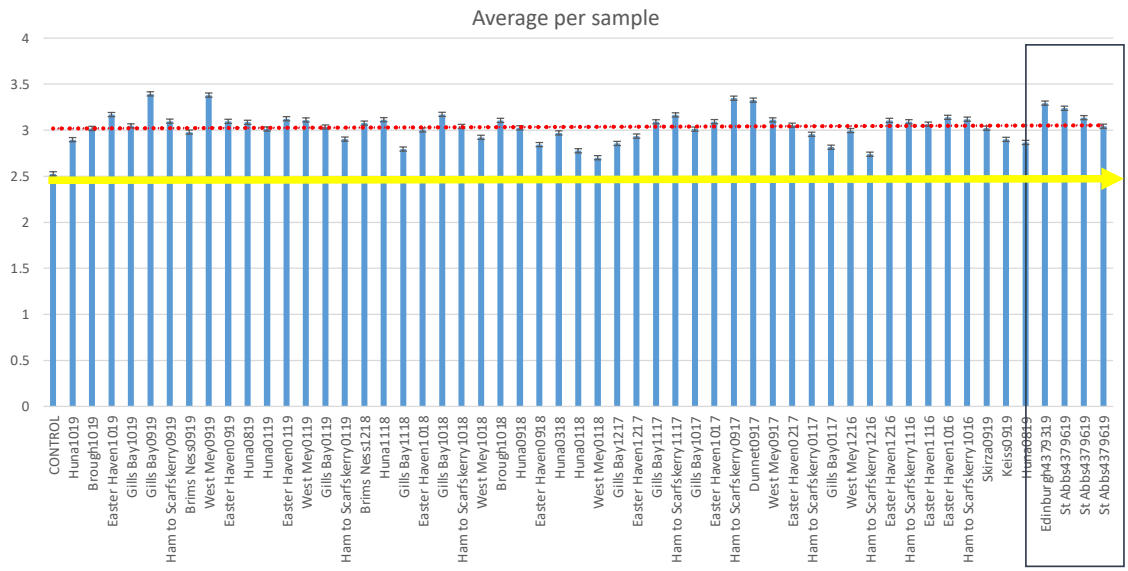


Figura 7: Resultado da cromatografia gasosa da produção de metano por amostra.

Alguns resultados preliminares foram obtidos e analisados, ainda durante o meu período de estágio, mas requerem maiores investigações para as seguintes análises de combinação das informações genéticas e fenotípicas.

Referencias

Menke KH, Steingass H. 1988. Estimation of the energetic feed value obtained from chemical analysis and gas production using rumen fluid. Animal Research and Development. 28: 7–55.

Stewart CS, Duncan SH. 1985. The effect of avoparcin on cellulolytic bacteria of the ovine rumen. *J Gen Microbiol.* **131**:427–35.

Weatherburn MW. 1967. Phenol-hypochlorite reaction for determination of ammonia. *Anal Chem* **39**:971–4.

COORDENADORIA DE PÓS-GRADUAÇÃO
INSTITUTO DE BIOLOGIA
Universidade Estadual de Campinas
Caixa Postal 6109. 13083-970, Campinas, SP, Brasil
Fone (19) 3521-6378. email: cpgib@unicamp.br

DECLARAÇÃO

Em observância ao **§5º do Artigo 1º da Informação CCPG-UNICAMP/001/15**, referente a Bioética e Biossegurança, declaro que o conteúdo de minha Tese de Doutorado, intitulada "***Abordagens multi-ômicas para o melhoramento florestal*** ", desenvolvida no Programa de Pós-Graduação em Genética e Biologia Molecular do Instituto de Biologia da Unicamp, não versa sobre pesquisa envolvendo seres humanos, animais ou temas afetos a Biossegurança.

Assinatura: _____
Nome do(a) aluno(a): Stephanie Karenina Bajay

Assinatura: _____
Nome do(a) orientador(a): Anete Pereira de Souza

Data: 24/08/2022

Declaração

As cópias de artigos de minha autoria ou de minha co-autoria, já publicados ou submetidos para publicação em revistas científicas ou anais de congressos sujeitos a arbitragem, que constam da minha Dissertação/Tese de Mestrado/Doutorado, intitulada **Abordagens multi-ômicas para o melhoramento florestal**, não infringem os dispositivos da Lei n.° 9.610/98, nem o direito autoral de qualquer editora.

Campinas, 24 de agosto de 2022

Assinatura: _____
Nome do(a) autor(a): **Stephanie Karenina Bajay**
RG n.° 35.838.266-x

Assinatura: _____
Nome do(a) orientador(a): **Anete Pereira de Souza**
RG n.° 8680325