UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE BIOLOGIA

FELIPE ROBERTO FRANCISCO

FERRAMENTAS GENÔMICAS E MOLECULARES PARA CARACTERIZAÇÃO GENÉTICA E PREDIÇÃO GENÔMICA EM SERINGUEIRA

GENOMIC AND MOLECULAR TOOLS FOR GENETIC CHARACTERIZATION AND GENOMIC PREDICTION IN RUBBER TREE

CAMPINAS
2023

FELIPE ROBERTO FRANCISCO

FERRAMENTAS GENÔMICAS E MOLECULARES PARA CARACTERIZAÇÃO GENÉTICA E PREDIÇÃO GENÔMICA EM SERINGUEIRA

GENOMIC AND MOLECULAR TOOLS FOR GENETIC CHARACTERIZATION AND GENOMIC PREDICTION IN RUBBER TREE

Tese apresentada ao Instituto de Biologia da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Doutor 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 Genetic Breeding

ESTE ARQUIVO DIGITAL CORRESPONDE À VERSÃO FINAL DA TESE DEFENDIDA PELO ALUNO FELIPE ROBERTO FRANCISCO, E ORIENTADA PELA PROF(A). DR(A). ANETE PEREIRA DE SOUZA

Orientadora/Supervisor: Prof.ª Dr.ª Anete Pereira de Souza

Coorientadora/Co-supervisor: Dr.ª Lívia Moura de Souza

Coorientadora/Co-supervisor: Prof. Dr. Roberto Fritsche-Neto

Campinas

2023

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

F847f

Francisco, Felipe Roberto, 1992-
Ferramentas genômicas e moleculares para caracterização genética e predição genômica em seringueira / Felipe Roberto Francisco. – Campinas, SP : [s.n.], 2023.

Orientador: Anete Pereira de Souza.
Coorientadores: Livia Moura de Souza e Roberto Fritsche-Neto.
Tese (doutorado) – Universidade Estadual de Campinas, Instituto de Biologia.

1. Seringueira. 2. Redes complexas. 3. Estudo de associação genômica ampla. 4. Melhoramento genético. 5. Genômica. I. Souza, Anete Pereira de, 1962-. II. Souza, Livia Moura de, 1980-. III. Fritsche-Neto, Roberto,1983-. IV. Universidade Estadual de Campinas. Instituto de Biologia. V. Título.

Informações Complementares

Título em outro idioma: Genomic and molecular tools for genetic characterization and genomic prediction in rubber tree
Palavras-chave em inglês:
Rubber plants
Complex networks
Genome-wide association study
Breeding
Genomics
Área de concentração: Genética Vegetal e Melhoramento
Titulação: Doutor em Genética e Biologia Molecular
Banca examinadora:
Anete Pereira de Souza [Orientador]
Marcelo Falsarella Carazzolle
Renato Vicentini dos Santos
Américo José Carvalho Viana
Marcelo Mollinari
Data de defesa: 14-02-2023
Programa de Pós-Graduação: Genética e Biologia Molecular

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0001-6488-7016
- Currículo Lattes do autor: http://lattes.cnpq.br/8013647627905495

Campinas, 14 de fevereiro de 2023

Comissão Examinadora

Profª. Drª. Anete Pereira de Souza
Dr. Marcelo Falsarella Carazzolle
Prof. Dr. Renato Vicentini dos Santos
Dr. Américo José Carvalho Viana
Dr. Marcelo Mollinari

Os membros da comissão examinadora acima assinaram a Ata da Defesa, que se encontra no processo de vida acadêmica do aluno na diretoria acadêmica.

Dedico essa tese as mulheres que permitiram que esse trabalho fosse realizado, a minha tia Tania, a minha avó Clara e a minha mãe Maria Angela

Agradecimentos

Agradeço primeiramente a prof. Anete por ter me dado a oportunidade e a confiança de trabalhar sob a sua orientação para a realização dessa tese, junto com o seu grupo de pesquisa que é referência no Brasil em análise genéticas e genômica em espécies não modelos, o Laboratório de Análises Genética e Molecular (LAGM), em uma das melhores universidades do país (Universidade Estadual de Campinas, UNICAMP). Muito obrigado por a todo momento acreditar em mim, e sempre me proporcionar tudo que foi necessário para a realização desse trabalho, me apoiando com seu otimismo e conselhos, o que me fez crescer na vida profissional e também pessoal.

Agradeço também ao Prof. Dr. Roberto Fritsche-Neto, pela paciência e ensinamentos sobre um assunto tão complexo que é a genética quantitativa. O seu entusiasmo, assim como a sua paixão por esse universo que me inspiraram a buscar novos conhecimentos e novas formas de analisar dados tão complicados. Também agradeço imensamente a Prof. Dr. Lívia Moura de Souza, que sempre esteve preocupada e próxima aos problemas da Heveicultura no Brasil, buscando sempre formas de ajudar no melhoramento genético dessa cultura tão importante. Agradeço a Dr. Lívia por todos os elogios, mas também as críticas e conselhos que fizeram os trabalhos aqui apresentados atingirem os resultados e o nível que atingiram.

Agradeço aos membros da banca Dr. Marcelo Falsarella Carazzolle, prof. Renato Vicentini dos Santos, Dr. Américo José Carvalho Viana e Dr. Marcelo Mollinari, que prontamente aceitaram participar da defesa de doutorado e contribuíram com as discussões dos trabalhos, assim como as devidas correções.

Agradeço aos meus companheiros de laboratório que não nomearei todos, por conta da grande quantidade e pelo medo de no meio desses tantos acabar esquecendo algum. Agradeço a essas amizades que foram firmadas ao longo de tantos cafés e bolos nos corredores do LAGM. Dessas amizades preciso agradecer especialmente a duas, Michele e Alexandre que não consigo expressar em palavras o quão importantes foram para a conclusão desta tese, tanto com ajuda nas análises, escritas e tantos outros âmbitos obrigatórios a uma tese, quanto ao ombro amigo que sempre pude contar, que nem mesmo a distância de intercâmbios ou as mazelas causadas por uma pandemia puderam enfraquecer.

Agradeço também ao corpo técnico do laboratório, em especial ao Danilo Sforça e a Aline Moraes que sempre nos ajudaram tanto dispondo de todos os equipamentos e reagentes

necessários à realização dos experimentos, quanto dos seus conhecimentos técnicos e teóricos em biologia molecular.

Agradeço aos meus amigos que a vida e a Unicamp me deram em especial, Angelo, Juninho, Jean, Ita, Gabriel, Rebeca, Gabi, Lara, Kauê, Anderson e Martinha, que sempre estiveram comigo ao longo desses 11 anos de Unicamp, me fortalecendo nos momentos mais difíceis. Não tenho como expressar o quão importante essas pessoas foram para amenizar as dificuldades dessa trajetória tão sinuosa.

Agradeço imensamente a minha família, ao meu irmão Danilo, aos meus primos Aline e Guilherme, aos meus primos de consideração Jéssica e Bruno, ao meu Tio Paulo e principalmente as mulheres que apesar de todas as dificuldades sempre me apoiaram e a quem dedico essa tese, Clara Manfrin, Tânia Regina e Maria Angela.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001, Programa Biologia Computacional (CAPES-88887.176241/2018-00). E também contamos com o apoio da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP - 2018/18985-7) que financiou os 4 últimos anos dessa pesquisa.

Resumo

A *Hevea brasiliensis,* mais conhecida como seringueira, é uma espécie emblemática com centro de origem na bacia amazônica. É considerada a espécie mais importante do gênero *Hevea* por ser a única espécie capaz de produzir borracha natural em quantidade e qualidade para suprir a demanda mundial dessa matéria prima, componente essencial para mais de 40.000 produtos. Apesar de conter o centro de origem da seringueira e grandes áreas com clima adequado para o cultivo da espécie, o Brasil é hoje um importador dessa matéria prima, pois nas áreas com condições climáticas ótimas para o cultivo da espécie a heveicultura é impossibilitada por conta da presença do fungo *Pseudocercospora ulei* causador da doença mal-das-folhas (SALB). Uma alternativa para o cultivo no país foi o plantio em regiões conhecidas como área de escape onde o clima seco e frio impossibilita a proliferação da SALB, da mesma forma que é um clima inadequado para o cultivo dos genótipos mais produtivos. Neste contexto existe uma grande necessidade do melhoramento genético principalmente visando o cultivo nessas regiões. A biologia molecular pode contribuir com o melhoramento genético da espécie possibilitando fazer uso da seleção assistida por marcadores (MAS) utilizando ferramentas como seleção genômica (GS) e associação genômica ampla (GWAS). Esta tese apresenta pela primeira vez o uso de seleção genômica em seringueira fazendo uso de marcadores do tipo *single nucleotide polymorphism* (SNPs) incluindo a interação genótipo x ambiente no modelo preditivo. Utilizando tais modelos foi possível obter um ganho genético esperado cinco vezes superior ao que se espera no melhoramento genético convencional. Além disso, também abordamos pela primeira vez a integração de associação genômica ampla com redes biológicas complexas para identificação e caracterização dos principais mecanismos moleculares envolvidos no crescimento da espécie. Mostrando principalmente a importância de genes envolvidos com resistência a estresse abiótico nesse processo, como SBT 4.6, GK 1, IQM 2 e elementos de transposons.

Palavras chaves: *Hevea brasiliensis;* seleção genômica; associação genômica ampla (GWAS); redes biológicas complexas; multiomicas; melhoramento genético.

Abstract

Hevea brasiliensis, better known as rubber tree, is an emblematic species with a center of origin in the Amazon basin. It is considered the most important species of the genus *Hevea* as it is the only species capable of producing natural rubber in quantity and quality to supply the world demand for this raw material, an essential component for more than 40,000 products. Despite containing the center of origin of the rubber tree and large areas with a suitable climate for the cultivation of the species, Brazil is today an importer of this raw material, because in areas with optimal climatic conditions for the cultivation of the species, rubber cultivation is impossible due to the presence of the fungus *Pseudocercospora ulei* that causes leaf blight (SALB). An alternative for cultivation in the country was planting in regions known as escape areas where the dry and cold climate makes it impossible for SALB to proliferate, just as it is an unsuitable climate for the cultivation of the most productive genotypes. In this context, there is a great need for genetic improvement, mainly aiming at cultivation in these regions. Molecular biology can contribute to the genetic improvement of the species, making it possible to use marker-assisted selection (MAS) using tools such as genomic selection (GS) and genomic wide association (GWAS). This thesis presents for the first time the use of genomic selection in rubber trees using single nucleotide polymorphism markers (SNPs) including the genotype x environment interaction in the predictive model. Using such models, it was possible to obtain an expected genetic gain five times higher than that expected in conventional genetic improvement. In addition, we also approach for the first time the integration of genomic wide association with complex biological networks for identification and characterization of the main molecular mechanisms involved in species growth. Mainly showing the importance of genes involved with resistance to abiotic stress in this process, such as SBT 4.6, GK 1, IQM 2 and transposon elements.

Keywords: *Hevea brasiliensis*; genomic selection; genome-wide association (GWAS); complex biological networks; multiomics; genetic improvement.

Sumário

Organização da Tese **13**
Introdução **14**
Referencial Bibliográfico **17**
2.1 Seringueira 17
2.2 Estudos Genômicos em Seringueira 18
2.3 Seleção Genômica 20
2.4 Associação Genômica Ampla e Redes de Co-expressão 21
Objetivos **24**
3.1 Objetivo Geral 24
3.2 Objetivo Específico 24
Capítulo I **25**
Genomic selection in rubber tree breeding: a comparison of models and methods for managing G× E interactions **25**
Abstract 26
Introduction 27
Materials and Methods 30
Populations and Phenotypes 30
Populations 30
Phenotypic Analysis 31
Genotypic Data and Single-Nucleotide Polymorphism Calling 32
Genomic Selection Analysis 33
SM 34
MM 35
MDs 36
MDe 36
Assessing Prediction Accuracy by Random Cross-Validation 38
Expected Genetic Gain 38
Expected Genetic Gain Obtained by a Classic Breeding Cycle With Only Phenotypic Information Used 38
Expected Selection Gain via Molecular Marker Information 39
Results 39
Single-Nucleotide Polymorphisms Calling 39
Estimates of Genetic Parameters by Single-Nucleotide Polymorphism Genotyping 40
Descriptive Statistics 40
Estimates of the Variance Components 41
Assessment of Prediction Accuracy 43
Single-Environment Model (SM) 44
Multienvironment Models (MM, MDe, and MDs) 44
Expected Genetic Gain 44
Discussion 46
Implementing Genomic Selection in Rubber Tree Breeding Programs 49
Data Availability Statement 51
Author Contributions 52

Funding 52
Conflict of Interest 52
Acknowledgments 52
Supplementary Material 53
References 53
Capítulo II **61**
Unravelling rubber tree growth by integrating GWAS and biological network-based approaches. **61**
Abstract 62
Introduction 63
Materials and Methods 65
Plant Material 66
Phenotypic Analyses 67
Genotypic Analyses 67
Genome-Wide Association Studies 68
Transcriptome 69
Gene-Associated Markers 70
Coexpression Networks 70
Metabolic Network Modeling 71
Results 71
Phenotypic and Genotypic Analyses 71
RNA-Seq Analyses 72
Genome Wide Association Study 72
Gene Coexpression Network 75
Metabolic Networks 78
Discussion 79
Genome-Wide Association Studies 81
Multiomics 84
Data Availability Statement 89
Funding 89
Conflict of Interest 89
Publisher's Note 89
Supplementary Material 90
Footnotes 90
References 90
Resumo dos Resultados **106**
7.1 Capítulo I 106
7.2 Capítulo II 106
Conclusão **108**
Perspectivas **109**
Referências **110**
ANEXOS **117**
8.1 Anexo 118
Machine learning for crop science: applications and perspectives in maize breeding 118

8.2 Anexo 120
A divide-and-conquer approach for genomic prediction in rubber tree using machine learning. 120
8.3 Anexo 122
The rubber tree kinome: genome-wide characterization and insights into coexpression patterns associated with abiotic stress responses 122
8.4 Anexo 124
Novel insights into the cold resistance of Hevea brasiliensis through coexpression networks 124
8.5 Anexo 126
Unraveling growth molecular mechanisms in Pinus taeda with GWAS, machine learning and gene coexpression networks 126
8.6 Anexo: Declaração Bioética e Biossegurança 128
8.7 Anexo: Declaração de Autoria 130

Organização da Tese

Esta tese está organizada em 8 partes. Que se iniciam a partir de uma **introdução geral**, onde são abordados os principais aspectos da heveicultura no Brasil, indicando a sua importância e seus principais problemas, principalmente aqueles relacionados com a obtenção de novas variedades para o melhorista. Em seguida apresentamos uma breve **revisão bibliográfica** acerca dos principais assuntos abordados nos capítulos I e II. Traremos posteriormente os **objetivos gerais** e **específicos** que foram investigados e alcançados nos capítulos I e II, um **resumo dos resultados encontrados** nesses capítulos, uma breve **conclusão**, **perspectivas** para novas pesquisas e as **referências bibliográficas**. Os capítulos I e II já estão publicados em revistas internacionais de alto impacto e foram escolhidos para compor a tese por terem sido os temas do projeto de doutorado inicial. Além desses capítulos o aluno se envolveu em outros trabalhos apresentados no final da tese como anexo.

No **capítulo I** fizemos uso pela primeira vez de estratégias de seleção genômica em seringueira (GS), utilizando duas matrizes de parentesco genético (VanRaden e Gaussian Kernel). Incluímos também nos modelos de predição diferentes efeitos da variação causada por consequência da interação genótipo x ambiente. As acurácias preditivas foram avaliadas utilizando duas formas de validações cruzadas, no qual os genótipos são avaliados em um único ambiente (CV1) ou em múltiplos ambientes (CV2). Os resultados mostraram acurácia preditiva em CV1 que foram de 0.19 a 0.27, enquanto que em CV2 as acurácias preditivas foram superiores a 0.80. Comparando tais resultados com os ganhos genéticos esperados pelo melhoramento tradicional, podemos notar ganhos de até cinco vezes superiores utilizando a estratégia de GS.

No **capítulo II** fizemos a integração dos resultados obtidos a partir da associação genômica ampla (GWAS) com redes biológicas complexas. A partir do GWAS, foram identificados 4 SNPs significativamente associados ao crescimento do caule da seringueira (snpsGWAS). Selecionamos outras 181 marcas significativamente associadas aos snpsGWAS (snpsLD). Todo esse conjunto de SNPs foi utilizado para selecionar 5 módulos altamente associados em uma rede de co-expressão global. Esses módulos se mostraram como *hubs* justamente os snpsLD, indicando a importância dessas marcas para a definição da característica estudada. Por fim construímos uma rede enzimática a partir dos genes presentes nesses módulos selecionados, identificando novos elementos que contribuem para a definição do fenótipo. Essa abordagem foi essencial para a identificação de diversos genes putativos com possível relação com o crescimento da seringueira em área de escape.

Introdução

Hevea brasiliensis, mais conhecida como seringueira, é uma espécie de grande importância no cenário econômico mundial, pois é a única capaz de produzir borracha natural (látex) em quantidade e qualidade suficiente para suprir o mercado mundial dessa matéria prima insubstituível para mais de 40.000 produtos (De Faÿ and Jacob, 1989; Hayashi, 2019; Pootakham et al., 2017; Mantello et al., 2019). Apesar de grande parte do centro de origem da seringueira estar localizado no Brasil, hoje o país é um importador dessa matéria prima, importando pouco mais da metade do consumo interno (Carvalho; 2020). Isso é devido a ocorrência da doença conhecida como *South American leaf blight* (SALB), causada pelo fungo *Pseudocercospora ulei* (Hora Júnior et al., 2014), que ocorre principalmente no centro de origem de diversidade do gênero *Hevea*, que são as áreas com clima adequado para o cultivo dos clones mais produtivos de seringueira (Jaimes et al., 2016; Mantello et al., 2019). Como forma de viabilizar a produção de borracha natural no país, a alternativa encontrada foi fazer o cultivo da *H. brasiliensis* em locais conhecidos como área de escape, onde o clima seco e frio não permite a proliferação e disseminação da SALB (Jaimes et al., 2016; Mantello et al., 2019).

Neste contexto se faz necessário o melhoramento da espécie visando o plantio nas áreas de escape (Gonçalves et al., 2021). Apesar da sua importância tal esforço ainda é incipiente, de modo que a espécie é considerada em domesticação. Isso é devido ao longo ciclo de melhoramento (~30 anos), a necessidade de grandes áreas para o plantio e do recente início das iniciativas voltadas ao cultivo de *H. brasiliensis* (Gonçalves, 1988; Dean, 1989; Clément-Demange et al., 2001; Jain & Priyadarshan, 2009; Priyadarshan, 2011, 2007). Esse longo tempo necessário ao melhoramento genético da seringueira é requerido principalmente por conta das etapas de avaliação que só podem ser iniciadas após o sexto ano de plantio para os características mais importantes agronomicamente como crescimento, vigor e produção de borracha natural (Priyadarshan, 2017, Conson et al., 2018, An et al., 2019 Wu et al., 2022). Deste modo a biologia molecular pode contribuir com os programas de melhoramento genético de *H. brasiliensis*, possibilitando fazer uma seleção precoce dos genótipos mais promissores a partir da seleção assistida por marcadores (MAS; figura 1; Collard & Mackill, 2008; Pootakham et al., 2017; Priyadarshan, 2017). Para tal tarefa é necessária a associação de marcadores moleculares com as características de interesse para o melhorista (Collard & Mackill, 2008).

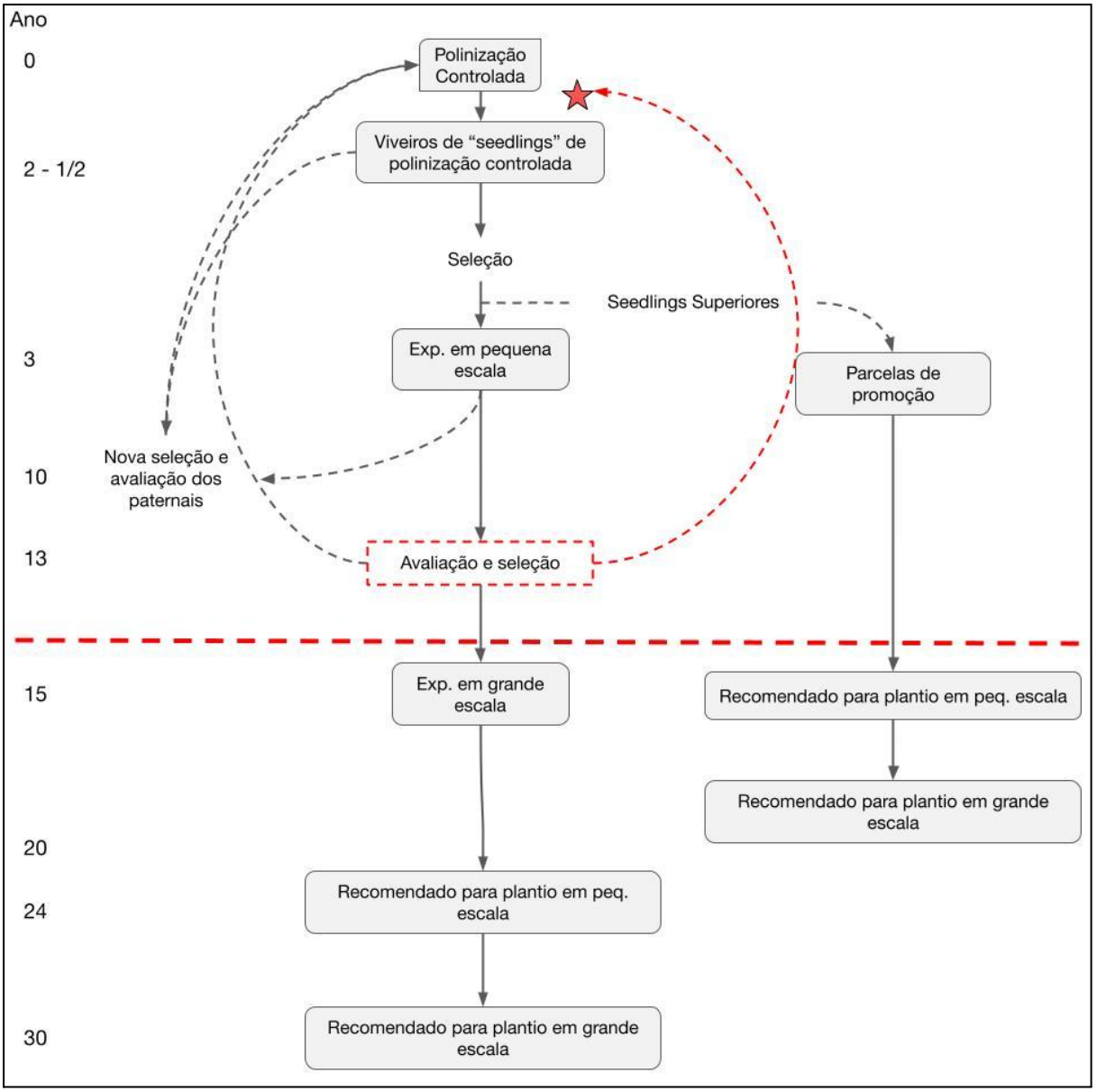

Figura 1: Esquema do melhoramento genético em *H. brasiliensis* adaptado de Gonçalves et al. (1986). A estrela vermelha destaca onde pode ser realizada a seleção utilizando a MAS reduzindo de forma drástica o tempo de melhoramento genético da espécie.

Uma das grandes dificuldades da aplicação da MAS nos programas de melhoramento genético da seringueira é justamente quanto a característica dos fenótipos de interesse que são geralmente controlados por uma grande quantidade de locus conhecido como QTLs (*quantitative trait loci*), o que torna a identificação dessas regiões uma tarefa difícil (Pootakham et al., 2020). Em seringueira diversas iniciativas foram tomadas para identificar QTLs de importância econômica, inicialmente fazendo uso de mapas de ligação contendo algumas centenas de marcadores moleculares (Lespinasse et al., 2000; Le Guen et al., 2003;

Le Guen et al., 2007; Souza et al., 2013). Com o advento das tecnologias de sequenciamento de nova geração, que revolucionou a biologia molecular (Sarig, & Sprecher, 2017), foi possível a identificação de um grande número de marcadores, principalmente do tipo *single nucleotide polimorphism* (SNPs), que cobrem grande parte do genoma das plantas (Pootakham et al., 2015). Esses marcadores foram então incorporados aos mapas de ligação da espécie possibilitando a identificação de QTLs importantes para características como vigor e produção (Xia et al., 2018; Rosa et al., 2018; Conson et al., 2018; An et al., 2019; Wu et al., 2022). Além de possibilitar a utilização novas abordagens como o *genome wide association study* (GWAS) e *genomic selection* (GS) (Chanroj et al., 2017; Souza et al., 2019; Francisco et al., 2021).

Apesar do grande número de QTLs identificado em seringueira, para diversos caracteres, esses ainda são pouco usuais, principalmente por conta da baixa proporção fenotípica explicada por esses marcadores, justificada pelo grande número de genes envolvidos no controle das características fenotípica quantitativas, assim como da influência ambiental sobre tais caracteres (Manolio et al., 2009; Nguyen et al., 2019). Por esse motivo, a integração de diferentes ferramentas ômicas é um avanço essencial para uma compreensão profunda da variação fenotípica observada, além de possibilitar a superação de dificuldades encontradas nesses estudos, tornando a MAS usual para os programas de melhoramento genético (Schaefer et al., 2018; Tam et al., 2019). Ainda assim, a integração de técnicas comuns de análises genômicas e transcriptoma ainda é pouco usada em espécies vegetais, principalmente em espécies arbóreas como no caso da *H. brasiliensis.*

Esta tese é pioneira na utilização de seleção genômica em seringueira fazendo uso de marcadores obtidos a partir de tecnologia de *next generation sequencing* (NGS), incluindo a interação genótipo x ambiente (Souza et al., 2019). Também utilizamos pela primeira vez a integração de metodologia genômica (GWAS) com transcriptoma (rede de co-expressão) na espécie para a melhor caracterizar os mecanismos moleculares envolvidos com o seu crescimento (Francisco et al., 2021).

Este trabalho contou com a colaboração entre pesquisadores do Laboratório de Análise Genética e Molecular (LAGM) - CBMEG/UNICAMP, laboratório de Melhoramento de Plantas Alógamas - Escola Superior de Agronomia “Luiz de Queiroz” (ESALQ), Instituto de Agronomia de Campinas (IAC), Universidade São Francisco (USF) e Coopération Internationale en Recherche Agronomique pour le Développement (CIRAD/França).

Referencial Bibliográfico

2.1 Seringueira

Hevea é um gênero natural bem definido, que se caracteriza por uma fácil identificação (Pires et al., 2002). Classificado como dicotiledônea, monóica, flores unissexuadas, apetaladas, amarelas e disposta em racimo, com folhas longamente pecioladas e repartidas em três folíolos com frutos grandes que geralmente apresentam três sementes (Paiva, 1992). O seu centro de origem está na região amazônica, com ocorrência natural por toda sua bacia, é encontrada no Brasil, Bolívia, Peru, Colômbia, Equador, Guianas, Suriname e Venezuela (Gonçalves e Fontes, 2009). Esse gênero pode ser considerado um complexo de espécies, por conta da alta ocorrência de híbridos naturais (Souza et al., 2018), tal característica facilita a introgressão gênica entre as espécies do grupo (de Souza, 2015). Dentre todas as 11 espécies do gênero *Hevea*, apenas *H.spruceana* e *H. microphylla* não possuem deiscência e látex em todas as partes da planta (Gonçalves & Marques, 2008). *H. brasiliensis* é considerada a maior fonte de borracha natural do mundo, produto que apresenta características únicas como resistência, elasticidade e dissipação de calor (Pootakham et al., 2017), por esse motivo é a espécie de maior importância econômica do gênero e a única capaz de suprir as necessidades do mercado deste produto em quantidade e qualidade (De Faÿ and Jacob, 1989; Gonçalves & Fontes, 2009; Hayashi, 2019; Pootakham et al., 2017; Mantello et al., 2019).

H. brasiliensis é uma espécie arbórea, perene e alógama da família das Euphorbiaceae (Priyandarshan e Clément-Demange, 2004) e contém 36 cromossomos (2n=2x=36), com uma origem possivelmente paleopoliplóide (Pootakham et al., 2017). A polinização é realizada de forma anemófila e entomófila (família Ceratopogonidae (Heleidae) e tripés) (IPEF, 2007). E existe a possibilidade das sementes serem dispersa pelas águas durante alagamentos sazonais, já que não é possível que as mesmas sejam carregadas pelo vento e a dispersão por animais, como pássaros e mamíferos, não ultrapassar algumas centenas de metros (Le Guen et al., 2009). Quando propagada por enxertia, produz látex com cerca de seis anos podendo perdurar até os 35 anos, existindo a possibilidade do aproveitamento da madeira ao fim deste período (Alika, 1980).

Apesar de a seringueira ser nativa do Brasil, e o país conter condições climáticas em grande parte do seu território adequadas para o cultivo da espécie, este é hoje um grande

importador da borracha natural, importando a maior parte do seu consumo (Carvalho; 2020). O principal problema para o cultivo de *H. brasiliensis* no país, está relacionado com a SALB causada pelo fungo *P. ulei* que leva a queda prematura das folhas jovens, resultando em prejuízos na produção e muitas vezes na morte das plantas (Jaimes et al., 2016). Iniciativas como a da montadora FORD de produzir borracha natural em grande escala no meio da floresta Amazônica no distrito de Fordlândia - Aveiro (PA-Amazônia), é um bom exemplo da inviabilidade do cultivo da seringueira nesta região (Duarte, 2015). Possíveis soluções para o problema causado pela enfermidade seriam: o plantio em áreas de escape, o uso de fungicidas, ou, o plantio de clones resistentes à doença. As duas últimas soluções são inviáveis por conta da elevação dos custos de produção e a facilidade de quebra de resistência pelo fungo. Em plantações fora do país o maior problema é causada por conta da baixa variabilidade genética das plantas, resultado da baixa amostragem dos indivíduos coletados por Henry Wickham em 1876 em um único local, que originaram todos os cultivares do sudeste asiático (Polhamus, 1962; Baulkwill, 1989).

No Brasil, o plantio em áreas de escape, como no estado de São Paulo, hoje maior produtor de borracha natural do país (Mazzaro, 2014), foi a melhor alternativa encontrada para contornar os problemas causados pelo SALB. Contudo, apesar desses locais apresentarem condições impróprias para o desenvolvimento do fungo, tais como frio e clima seco, os clones de alta produção de látex também têm um desenvolvimento prejudicado. Neste contexto existe uma grande necessidade de se realizar o melhoramento dessas plantas visando o plantio nas áreas de escape sem a perda da produção do látex e vigor.

2.2 Estudos Genômicos em Seringueira

Apesar da grande importância da seringueira para a economia mundial, o melhoramento genético da espécie ainda é bastante incipiente devido principalmente às longas etapas de seleção e a sua recente domesticação (Priyadarshan and Clément-Demange,2004). Neste sentido, a biologia molecular pode contribuir com os programas de melhoramento genético da *H. brasilieneis* possibilitando a utilização de marcadores moleculares, que podem ser utilizados para se fazer uma seleção precoce dos genótipos superiores (Jannink, 2010). Em espécies arbóreas é estimado que tal abordagem pode reduzir os ciclos de melhoramento, além de possibilitar um aumento considerável da

acurácia de seleção e consequentemente nos ganhos genéticos esperados (Vivek et al., 2017; Montesinos-López et al., 2021).

Contudo a utilização de tais ferramentas nos programas de melhoramento, requer antes de tudo, estudos genômicos que possibilitem a caracterização molecular, assim como as suas contribuições para a variação fenotípica. Diversos estudos genômicos vêm sendo realizados em seringueira possibilitando a identificação de QTLs importantes agronomicamente como resistência a SALB, crescimento e produção de borracha natural (Lespinasse et al., 2000; Le Guen et al., 2003; Le Guen et al., 2007; Souza et al., 2013; Chanroj et al., 2017; Rosa et al., 2018; Xia et al., 2018; Conson et al., 2018; An et al., 2019; Wu et al., 2022). Apesar dessas iniciativas mostrarem resultados importantes, os estudos genômicos em seringueira requerem avanços, principalmente no que diz respeito a genomas disponíveis para a espécie. O primeiro genoma só foi disponível para seringueira em 2013, sequenciado a partir do clone RRIM 600, contendo 1.150.326 *scaffolds* com um tamanho ~1,1 Gb do genoma haplóide, que foi estimado em 2,15 Gb, sendo desse total ~78% regiões repetitiva, com N50 de 2,972 bp e 68.995 genes preditos dos quais 12,7% são exclusivos de *Hevea* (Rahman et al., 2013). Em 2016 foram publicados outros dois genomas para a espécie (Lau et al., 2016; Tang et al., 2016). Lau et al., 2016, também utilizando RRIM 600, disponibilizou um genoma contendo 189.320 *scaffolds* com tamanho ~1,55 Gb, sendo 72,5% composto por sequências repetidas, identificando um total de 84.440 genes e com um N50 de 67,2 Kb (Lau et al., 2016). Tang et al., 2016 no mesmo ano, também disponibilizou um genoma de alta qualidade para a espécie, utilizando o clone Reyan 7-33-97, contendo 7.453 *scaffolds,* com tamanho de ~1,37 Gb, N50 de 1,28 Mb, 71% de regiões de repetição e 43.792 genes preditos (Tang et al., 2016). Pootakham et al., 2017, disponibilizaram um genoma para o cultivar BPM24, este genoma 1,26 Gb (N50 = 96,8 kb), ancorado em 592.579 *scafoldes,* contendo 69,2% de sequências repetidas (Pootakham et al., 2017). Essa grande quantidade de regiões repetidas encontrada no genoma da espécie é um grande complicador para a montagem do genoma, por conta disso o primeiro genoma a nível de pseudo cromossomo foi disponibilizado apenas em 2021 (Liu et al., 2020). Este genoma foi construído com base no genótipo GT1 e ancorou 1,47-Gb em 18 pseudo cromossomo, com evidência de importante expansão causada por retrotransposons LTR, que compreende ~65,88% (~970 Mbp) de todo genoma (Liu et al., 2020). Em comparação com outras Euphorbiaceae também é notado no genoma da seringueira uma expansão massiva de genes envolvidos com biossíntese de borracha natural (Liu et al., 2020). Apesar dos poucos estudos que visam investigar a relação dos transposons no genoma de seringueira é notória a sua importância para evolução e

adaptação da espécie tanto nos níveis moleculares quanto para definição de fenótipos relevantes para o melhorista, como aqueles envolvidos com produção e resistência a condições climáticas adversas (Wu et al., 2020; Francisco et al., 2021).

Mesmo com as limitadas informações genômicas disponíveis para seringueira, esforços foram realizados para a caracterização molecular da espécie e identificação de QTLs úteis principalmente para o melhoramento genético foram realizados. Inicialmente fazendo uso de mapas genéticos construídos com marcadores do tipo RFLPs, AFLPs, microssatélites, isoenzimas (Lespinasse et al., 2000; Le Guen et al., 2003; Le Guen et al., 2007; Souza et al., 2013). Esses mapas possibilitaram a identificação de QTLs importantes para a cultura como resistência ao fungo *P. uley* (Lespinasse et al., 2000, Le Guen et al., 2003; Le Guen et al., 2007), resistência a condições climáticas adversas e crescimento (Souza et al., 2013). Com o advento das tecnologias de NGS, novas abordagens puderam ser realizadas, revolucionando não só os estudos genômicos em seringueira, mas em toda a biologia. Nessa nova era de tecnologias de sequenciamento, muitos transcriptomas foram desenvolvidos para *H. brasiliensis* em diferentes tratamentos e tecidos como estresse ao frio (Mantello et al., 2019; Cheng et al., 2018), produção de látex (Lau et al., 2016; Tang et al. 2016), casca (Chow et al., 2014) e látex (Montoro et al., 2018), identificando diversos genes importante para a espécie.

2.3 Seleção Genômica

Proposta por Bernardo (1994) e Meuwissen et al. (2001), a seleção genômica (GS) é muito bem estabelecida como ferramenta que auxilia o melhoramento genético de espécies (Zhao et al., 2011; McKown et al., 2014; Weiss et al., 2020; Chen et al., 2021; Ding et al., 2022; Cerioli, et al., 2022). O conceito é baseado na utilização de um grande número de marcadores moleculares polimórficos, juntamente com os valores genéticos (BLUP) observada em uma população de treinamento, no qual os modelos preditivos são construídos, para predizer os valores genéticos (GEBVs) de uma população não fenotipada (população de seleção) com base apenas nos marcadores moleculares dessa população (Crossa et al., 2017; Montesinos-López et al., 2021). Neste caso, não se faz necessário o conhecimento prévio da localização, nem mesmo do efeito de cada marca ou QTL sobre o fenótipo, sendo assim são utilizadas todas os marcadores para estimar os GEBVs da população futura (Crossa et al., 2017; Lebedev et al., 2020). Com o barateamento da descoberta desses marcadores

moleculares, proporcionado principalmente pelas tecnologias de NGS, essa ferramenta se tornou comum em muitos programas de melhoramento.

Em contraste com as culturas agrícolas com histórico milenar de cultivo, o melhoramento de espécies florestais é relativamente recente (Isik et al., 2015). Além disso, o processo de melhoramento das espécies arbóreas é bem mais árduo e demorado, podendo levar entre 20 e 30 anos cada ciclo de seleção (Lebedev et al., 2020). Neste contexto, a incorporação de novas tecnologias como a SG pode contribuir de forma significativa para a obtenção de cultivares melhorados (Lebedev et al., 2020). Em comparação com o melhoramento genético tradicional baseado no fenótipo de cada indivíduo, a SG pode encurtar esse processo, já que a avaliação é feita com base nos marcadores moleculares, que podem ser avaliados assim que o DNA possa ser isolado sem prejudicar a planta, eliminando a etapa mais demorada e custosa do ciclo de seleção (Lebedev et al., 2020). Essa abordagem é especialmente importante quando se trata de características complexas que se manifestam tardiamente, que em última análise reduzirá o tempo de seleção aumentando assim os ganhos genéticos esperados (Bhat et al., 2016; Crossa et al., 2017, Lebedev et al., 2020).

A acurácia da SG, assim como da predição genômica (GP), pode ser afetada por diversos fatores como: i) tamanho e diversidade genética da população de treinamento (TRN) e sua relação genética com a população de teste (TST) (Pszczola et al., 2012); ii) herdabilidade da característica avaliada (Crossa et al., 2017); iii) características com grande número de marcadores em desequilíbrio de ligação (LD) com o fenótipo avaliado, tem a acurácia de predição aumentada com o aumento da TRN (Daetwyler et al., 2010).

2.4 Associação Genômica Ampla e Redes de Co-expressão

A contribuição dos mapas genéticos para identificação de QTLs em diversas espécies como *Psidium guajava* L. (Sohi et al., 2022), *Mylopharyngodon piceus* (Guo et al., 2022) e *Rhopilema esculentum* (Chen et al., 2022), é inestimável. Em seringueira, foram construídos diversos mapas de ligação altamente saturados identificando importantes QTLs (Xia et al., 2018; Rosa et al., 2018; Conson et al., 2018; An et al., 2019; Wu et al., 2022). Embora tal metodologia apresenta limitações como: i) necessidade da produção de uma população biparental; ii) uso limitado a população no qual o mapa foi desenvolvido; iii) baixa diversidade genética e iv) número de marcadores relativamente pequeno ancorados em grupos de ligação (Myles et al., 2009, Crossa et al., 2017). Essas restrições são ainda mais

evidentes quando se trata de espécies florestais, como é o caso da seringueira, essencialmente por conta das dificuldades para a construção de populações apropriadas.

Com a possibilidade da descoberta de um grande número de marcadores moleculares, principalmente do tipo SNPs, se tornou possível a utilização de ferramentas como GWAS para fazer tais descobertas de QTLs (Zargar et al., 2015). Em comparação com mapas de ligação o GWAS apresenta algumas vantagens como: i) utilização de uma população diversificada, não sendo necessário a construção de uma população biparental; ii) utiliza um grande número de marcadores, sem a necessidade do conhecimento prévio da sua localização em grupos de ligação; iii) faz uso do LD histórico entre os marcadores moleculares e os QTLs; iv) não está limitado o seu uso a população em que esse foi desenvolvido e v) utiliza populações com alta diversidade genética (Myles et al. 2009, Bernardo, 2016; Crossa et al., 2017).

A utilização do GWAS em diversas espécies vegetais mostra resultados bastante promissores para a identificação de QTLs envolvidos com características complexas em diferentes espécies vegetais como *Arabidopsis thaliana* (Atwell et al., 2010), *Oryza sativa* (Zhao et al., 2011, Cerioli, et al., 2022), *Populus trichocarpa* (McKown et al., 2014) e coníferas (Weiss et al., 2020; Chen et al., 2021; Ding et al., 2022). Em seringueira, essa abordagem já foi descrita com sucesso para a descobertas de QTLs associados com o crescimento em condições de estresse abiótico, assim como para a produção de borracha natural (Chanroj et al. 2017). Neste estudo Chanroj et al. (2017), identificou apenas dois SNPs associados à produção de borracha natural em duas estações do ano distintas e outro par de SNPs também associado ao crescimento, também para estação do ano chuvosa e seca. Apesar da sua imensa importância, o GWAS ainda necessita de adequações para o seu uso efetivo nos programas de melhoramento genético, principalmente para fenótipos complexos, que apresentam baixa herdabilidade (Manolioet al., 2009; Korte, & Farlow 2013; Tam et al., 2019). Essas limitações ocorrem principalmente por conta do grande número de falsos positivos, causados por um limiar de significância restritivo e o pequeno efeito de alguns QTLs importantes que compõem essas características (Tam et al., 2019).

Em espécies como no caso da *H. brasiliensis* com limitados dados genômicos disponíveis, abordagens alternativas devem ser tomadas para uma adequada caracterização genética dos mecanismos moleculares envolvidos com a característica de interesse. Schaefer et al (2018) argumenta que o uso de dados de transcriptoma, integrados ao GWAS, seria uma alternativa satisfatória para superar tais limitações.

A literatura apresenta métodos bastante robustos para a montagem de transcriptomas sem a necessidade de genomas de referências, principalmente para espécies não modelos (Grabherr et al., 2011). Também existe uma vasta literatura que mostra a importância das redes de co-expressão para identificação de genes em diferentes espécies, assim como a compreensão de como os elementos que a compõem e interagem (Koenig et al., 2013; Ma et al., 2021; Fajardo & Quecini, 2021).

As redes de co-expressão gênica são ferramentas que podem exibir e contextualizar, conjuntos de dados de co-expressão para se obter um nível de informação e compreensão molecular satisfatório dos processos biológicos chaves, além de nos possibilitar inferir relações biológicas de determinado gene ainda não conhecido (Serin et al., 2016). A construção de redes de co-expressão tem o poder de agrupar genes em módulos com base na semelhança da expressão gênica, resultando em agrupamento de genes semelhantes (Langfelder e Horvath, 2008). Em resumo, uma pontuação é atribuída para representar o nível de similaridade do padrão de expressão de um dado gene, pares de genes com uma correlação de expressão acima de um dado limiar são definidos como co-expressos (Rao & Dixon, 2019). Essa abordagem nos possibilita responder a questões biológicas importantes, que podem ser agrupadas em 3 categorias principais: (i) identificar relações reguladoras entre reguladores e seus alvos, (ii) predizer genes estruturais em vias metabólicas e (iii) transferência de anotação de genes por meio da análise de co-expressão comparativas em espécies de plantas distintas (Rao & Dixon, 2019).

Objetivos

3.1 Objetivo Geral

Desenvolver ferramentas genômicas e moleculares que possibilitem a aplicação de seleção genômica em seringueira e que possam auxiliar na compreensão dos mecanismos moleculares e genômicos envolvidos com o crescimento e vigor das plantas.

3.2 Objetivo Específico

- Aplicar modelos de predição genômica em seringueira
- Avaliar a capacidade preditiva para as matrizes de parentesco VanRaden e Gaussian Kernel
- Avaliar diferentes formas de validação cruzada
- Avaliar os efeitos da interação genótipo x ambiente na acurácia preditiva dos modelos de seleção genômica
- Detecção de QTLs importantes para os programas de melhoramento genético da espécie através do GWAS
- Anotação das regiões dos marcadores identificados pelo GWAS
- Construção de rede de co-expressão e integração dos resultados do GWAS.
- Construção de uma rede metabólica com base nos resultados da integração do GWAS com a rede de co-expressão

Capítulo I

Genomic selection in rubber tree breeding: a comparison of models and methods for managing G× E interactions

Livia M. Souza †, **Felipe R. Francisco** †, Paulo S. Gonçalves, Erivaldo J. Scaloppi Junior, Vincent Le Guen, Roberto Fritsche-Neto and Anete P. Souza*

Front. Plant Sci., **25 October 2019**
Sec. Plant Breeding
https://doi.org/10.3389/fpls.2019.01353

Keywords: *Hevea brasiliensi*s, breeding, multienvironment, single nucleotide, genotyping

Abstract

Several genomic prediction models combining genotype × environment (G×E) interactions have recently been developed and used for genomic selection (GS) in plant breeding programs. G×E interactions reduce selection accuracy and limit genetic gains in plant breeding. Two data sets were used to compare the prediction abilities of multienvironment G×E genomic models and two kernel methods. Specifically, a linear kernel, or GB (genomic best linear unbiased predictor [GBLUP]), and a nonlinear kernel, or Gaussian kernel (GK), were used to compare the prediction accuracies (PAs) of four genomic prediction models: 1) a single-environment, main genotypic effect model (SM); 2) a multienvironment, main genotypic effect model (MM); 3) a multienvironment, single-variance G×E deviation model (MDs); and 4) a multienvironment, environment-specific variance G×E deviation model (MDe). We evaluated the utility of genomic selection (GS) for 435 individual rubber trees at two sites and genotyped the individuals via genotyping-by-sequencing (GBS) of single-nucleotide polymorphisms (SNPs). Prediction models were used to estimate stem circumference (SC) during the first 4 years of tree development in conjunction with a broad-sense heritability (H^2) of 0.60. Applying the model (SM, MM, MDs, and MDe) and kernel method (GB and GK) combinations to the rubber tree data revealed that the multienvironment models were superior to the single-environment genomic models, regardless of the kernel (GB or GK) used, suggesting that introducing interactions between markers and environmental conditions increases the proportion of variance explained by the model and, more importantly, the PA. Compared with the classic breeding method (CBM), methods in which GS is incorporated resulted in a 5-fold increase in response to selection for SC with multienvironment GS (MM, MDe, or MDs). Furthermore, GS resulted in a more balanced selection response for SC and contributed to a reduction in selection time when used in conjunction with traditional genetic breeding programs. Given the rapid advances in genotyping methods and their declining costs and given the overall costs of large-scale progeny testing and shortened breeding cycles, we expect GS to be implemented in rubber tree breeding programs.

Introduction

Rubber tree (*Hevea brasiliensis*) breeding programs are generally characterized by breeding cycles of 25–30 years and include the crosses, evaluation, and selection of field progeny as well as the propagation of selected superior materials (Gonçalves et al., 2006). Compared with animal and annual crop species breeding, forest tree breeding is still in its infancy; the most advanced programs are in their third or fourth cycle of breeding, with very little differentiation occurring between the bred populations and natural populations (Isik, 2014). Rubber tree breeding programs are complex and costly because the large size of these trees necessitates experiments over large tracts of land, and progeny tests are expensive to establish, manage over many years, and evaluate via measurements.

The main objective of rubber tree breeding is the development of early selection methods that support the accurate prediction of mature phenotypes at a young stage; these methods are therefore important for shortening breeding cycles and thus improving the cost efficiency of such breeding programs. The time taken to derive a *Hevea* through breeding must be substantially reduced. Priyadarshan (2017) proposed two strategies: 1) truncating the breeding steps that follow conventional means and 2) incorporating genomics into breeding programs specifically to identify high-yielding genotypes in half-sib, full-sib, and polycross seedlings during the juvenile stage to minimize both space and time.

Classic plant breeding programs depend principally on phenotypic evaluation in several environments; selection and recombination are based on the resulting data and genotype information when available. Genomic selection (GS), a new approach in which whole-genome molecular markers are used, has the potential to quickly improve complex traits with low heritability, significantly reduce the cost of line and hybrid development and increase yields in reduced amounts of time, allowing improvements to quantitative traits within large plant breeding populations (Meuwissen et al., 2001).

Genomic prediction combines phenotypic and pedigree data with marker data in efforts to increase the prediction accuracy (PA) for breeding and genotypic values. This method depends on dense genome-wide marker coverage to produce genomic estimated breeding values (GEBVs) from a comprehensive analysis of all available markers.

According to Lorenz et al. (2011), the accuracy of GS, which is measured as the correlation between GEBVs and true breeding values, is affected by the relationship between the training (TRN) and testing (TST) sets, the number of individuals in the TRN set, linkage

disequilibrium (LD) between the markers and quantitative trait loci (QTLs), the distribution of the underlying QTL effects, the statistical method used to estimate the GEBVs, and the trait heritability.

According to Meuwissen et al. (2001), GS has received increasing interest from forest tree breeders. In reports of initial experiments involving Pinus and Eucalyptus (Resende et al., 2012a; Resende et al., 2012b), this new method showed encouraging prospects, thus confirming the potential of GS in studies of conifers, pines, and eucalypts (Zapata-Valenzuela et al., 2013; Lima, 2014; El-Dien et al., 2015; Ratcliffe et al., 2015; Bartholome et al., 2016; Isik et al., 2016), which further supports the potential for GS to accelerate the breeding of forest trees.

In rubber tree breeding programs, pedigree-based analysis has been widely used to evaluate field experiments, estimate genetic parameters, and predict breeding values (Furlani et al., 2005). However, due to the decreasing costs of genotyping thousands or millions of markers and the increasing costs of phenotyping (Krchov and Bernardo, 2015), GS is emerging as an alternative genome-wide marker-based method to predict future genetic responses.

Genomic prediction models were originally developed for use in a single environment. However, to implement GS strategies in plant breeding, genotype × environment (G×E) interactions must be predicted. Habier et al. (2007) used genetic marker information to identify associations between individuals via the genomic relationship matrix K. Two very frequently used matrix-based methods include the genomic best linear unbiased predictor (GBLUP) (GB) (VanRaden, 2007, VanRaden, 2008) and the nonlinear Gaussian kernel (GK) methods (Gonzalez-Camacho et al., 2012). Burgueño et al. (2012) extended this general methodology to incorporate G×E effects. A separate GB extension introduces interaction effects between markers and environmental factors, and studies have shown that modeling G×E can result in substantial gains in PA (Heslot et al., 2014; Jarquin et al., 2014; Crossa et al., 2016; Cuevas et al., 2016).

A GBLUP model was proposed by Lopez-Cruz et al. (2015) to explicitly model the partitioning of genomic values and marker effects into components that are stable among environments and others that are environment specific. Therefore, according Cuevas et al. (2016), the marker × environment interaction model is suitable for application in groups of positively correlated environments. However, in practice, this approach can be very restrictive in cases where several environments have correlations close to zero, as it can lead to a large G×E variance component compared with the genetic variance component

(Burgueño et al., 2011). VanRaden (2008) first suggested models in which the GBLUP was a linear model that included parameters associated with additive quantitative genetics theory.

A nonparametric and semiparametric method was proposed by Gianola et al. (2006) and accounted for small, complex epistatic interactions without explicitly modeling them. According to Heslot et al. (2012), the semiparametric reproducing kernel Hilbert space (RKHS method) uses a kernel function to convert the marker matrix into a set of distances between pairs of individuals. RKHS regression is thought to increase PA by capturing nonadditive variation, and several studies have confirmed this advantage (de los Campos et al., 2010; Perez-Rodriguez et al., 2013; Morota and Gianola, 2014).

Cuevas et al. (2016) applied GS with the marker × environment interaction model of Lopez-Cruz et al. (2015) and modeled the GB (linear kernel) and GK (nonlinear kernel) in a manner similar to that of de los Campos et al. (2010) in the RKHS with kernel averaging, and by estimating the bandwidth via an empirical Bayesian method (Pérez-Elizalde et al., 2015), and using wheat and maize data sets, they performed single-environment analyses and expanded them to account for G×E interactions. Compared with the other approaches, the GK combined with the G×E model provided greater flexibility and accounted for smaller, more complex marker main effects and marker-specific interaction effects (Cuevas et al., 2016). However, as in the study by Lopez-Cruz et al. (2015), this model assumes sets of environments that are positively correlated. To solve this problem, Cuevas et al. (2016) proposed two multienvironment genomic models to overcome some of the restrictions of previous genomic models.

Accurate predictions are obtained when the appropriate method is used even for untested genotypes, allowing considerable progress in breeding programs by reducing the number of field-tested genotypes and, consequently, the costs of phenotyping (Krchov and Bernardo, 2015). The benefits of GS are more evident when traits are difficult, time consuming, and expensive to measure and when several environments need to be evaluated.

The objective of this paper was to evaluate the predictive capability of GS implementation in rubber trees when linear and nonlinear kernel methods are used and to examine the performance of the predictions, including G×E interactions, of each of the four models described by Bandeira et al. (2017). Thus, for all data sets, we fitted models with a linear kernel via GB or GK with a bandwidth parameter estimated according to the methods of Pérez-Elizalde et al. (2015). We also compared the PA of the two kernel regression methods for the four models, which included the following: a single-environment, main genotypic effect model (SM); a multienvironment, main genotypic effect model (MM)

(Jarquin et al., 2014); a multienvironment, single-variance G×E deviation model (MDs) (Jarquin et al., 2014); and a multienvironment, environment-specific variance G×E deviation model (MDe) (Lopez-Cruz et al., 2015).

To the best of our knowledge, this is the first attempt to apply GS with a multienvironment technique to a rubber tree breeding program. The development of a robust methodology enables the implementation of GS in routine evaluations to accelerate genetic progress.

Materials and Methods

Populations and Phenotypes

The data set included 435 samples, which comprised 252 F1 hybrids derived from a PR255 × PB217 cross (Souza et al., 2013; Rosa et al., 2018), 146 F1 hybrids derived from a GT1 × RRIM701 cross (Conson et al., 2018), 37 genotypes from a GT1 × PB235 cross, and 4 testers (GT1, PB235, RRIM701, and RRIM600), which are described further below.

Populations

The PR255 × PB217 population is a full-sib segregating population with a total of 252 individuals (progeny). Seedlings acquired via controlled pollination were clonally propagated by budding onto rootstocks. PR255 is a rapidly growing clone with vigorous and high yield, good growth, and stable latex production. In contrast, clone PB217 is the opposite, presenting slow growth and delayed latex production in its early years of development, although its latex production increases rapidly during the early years; however, this clone has potential for superior yield performance in the long term (Souza et al., 2013; Rosa et al., 2018). The field trial was performed in Itiquira, Mato Grosso state, Brazil (17°24′03″S and 54°44′53″W), from March 2006 until March 2007. The climate of this region is characterized by very dry and relatively cold winters and hot and humid summers, which represent conditions typical of southeastern Brazil, the most productive region for rubber. The experimental design was a randomized block design with four replications, with four grafted trees of the same individual in each plot (Rosa et al., 2018).

The GT1 × RRIM701 population comprised 146 individuals, and the GT1 × PB235 population comprised a total of 37 individuals. These two groups of progeny were derived from open pollination, and their effective pollination was checked via microsatellite markers.

The hybrids were selected on the basis of polymorphisms between the parents. GT1 is a male-sterile clone that is classified as a primary clone (Shearman et al., 2014) and is tolerant to wind and cold. RRIM701 grows vigorously and presents an increased circumference after initial tapping (Romain and Thierry, 2011). PB235 is reported to be a high-yielding clone with an active metabolism and is known to be particularly susceptible to tapping panel dryness (Sivakumaran et al., 1988). These two groups of progeny were planted at the Center of Rubber Tree and Agroforestry Systems/Agronomic Institute (IAC) in the northwestern region of São Paulo state (20°25′00″S and 49°59′00″W at an altitude of 450 m), Brazil, in 2012 (Alvares et al., 2013). A modified block design was used (Federer and Raghavarao, 1975), and the trial was repeated four times. Each trial consisted of four blocks with two trees (clones) per plot spaced 4 m by 4 m. The experiment comprised a total of 656 (41 plots × 4 blocks × 4 replicates) plots and 1,312 trees (Conson et al., 2018).

Phenotypic Analysis

Stem circumference (SC, in cm) at 50 cm above ground level was measured to evaluate the growth of individual trees, where the average per plot was calculated. Growth traits were frequently measured only during the first 6 years, as height and SC are the main selection traits for rubber tree breeding (Rao and Kole, 2016). Measurements were taken at four different ages and are listed in Supplementary Table 1. Two sets of measurements were taken each year: one set applied to trees under low-water conditions (LW), and the other applied to trees under well-watered conditions (WW). These conditions were established according to the water distribution of each region in which the experiments were installed (Supplementary Figure 1, Supplementary Table 2).

Analyses of the SC traits were carried out via the breedR package (Munõz and Sanchez, 2017) in conjunction with the remlf90 function and method = “ai,” and the best linear unbiased predictors (BLUPs) of each genotype used with the following mixed linear model were taken:

$$y = Xb + Zg + e$$

where y is the adjusted mean phenotypic value (best linear unbiased estimated [BLUES]), × and Z are known incidence matrices, b is the vector of fixed effects (environmental effects), and g is the vector of random effects (genetic effects). In the general model (H^2), when the entire data set from both environments (LW and WW) is used, the fixed effects included locale (place where the experiment was performed), block, water, and year. The G × E

interaction and genotype were included as random effects in the model. When we considered each environment (LW or WW) separately (H^2_{env}), the fixed effects were the G × E interaction, locale, year, repetition, and block. Genotype was included as a random effect.

The broad-sense heritability (H^2) (clonal mean heritability) was estimated for SC for each water management system (LW and WW) and for every data set:

$$H^2 = \sigma^2_g \Big/ \left[\sigma^2_g + \frac{\sigma^2_{gxe}}{s} + \frac{e}{sa}\right]$$

where σ^2_g is the genetic variance, σ^2_{gxe} is the variance caused by the interaction between genotype and the environment, *e* is the residual variance, *s* is the number of environments, and *a* is the number of blocks.

For each environment, we estimated heritability (H^2_{env}) separately as follows:

$$H^2_{env} = \sigma^2_g \Big/ \left(\sigma^2_g + \frac{\sigma^2_e}{r}\right)$$

where σ^2_g is the genetic variance, σ^2_e is the residual variance, and r is the number of trees per replicate.

Genotypic Data and Single-Nucleotide Polymorphism Calling

Genomic DNA was extracted according to the methods of Souza et al. (2013) and Conson et al. (2018). Genotyping-by-sequencing (GBS) library preparation and sequencing were performed as described by Elshire et al. (2011). Genome complexity was reduced by digesting individual genomic DNA samples with EcoT22I, a methylation-sensitive restriction enzyme, and 96 samples were included in each sequencing lane. The resulting fragments from each sample were directly ligated to a pair of enzyme-specific adapters and combined into pools. PCR amplification was carried out to generate the GBS libraries. Library sequencing of GT1 × RRIM701 and GT1 × PB235 was performed on an Illumina GAIIx platform (Illumina Inc., San Diego, CA, United States), and sequencing of PR255 × PB217 was performed on the Illumina HiSeq platform.

The raw data were processed, and single-nucleotide polymorphism (SNP) calling was performed via TASSEL 5.0 (Glaubitz et al., 2014). Initially, the FASTQ files were demultiplexed according to their assigned barcodes. The reads from each sample were trimmed, and the tags were identified by the following parameters: Kmer length of 64 bp, minimum quality score within the barcode and read length of 20, minimum Kmer length of

20 and minimum count of reads for a tag of 6. The retained tags with a minimum count of six reads were aligned to the *H. brasiliensis* reference genome sequence (Tang et al., 2016) via Bowtie 2 version 2.1 (Langmead and Salzberg, 2012), with the very sensitive option enabled. SNP calling was performed via the TASSEL 5 GBSv2 pipeline (Glaubitz et al., 2014) and filtered with snpReady software (Granato and Fritsche-Neto, 2018). The following criteria were used: 20% missing data, minor allele frequency (MAF) greater than or equal to 5% (MAF of 0.05), and removal of individuals with more than 50% (sweep.sample = 0.5) missing data for the called SNPs. Only biallelic SNPs were maintained, which was performed via VCFtools (Danecek et al., 2011). After the data were filtered, the missing data were imputed by the knni method with snpReady software (Granato and Fritsche-Neto, 2018).

The genotypic data are available under NCBI accession PRJNA540286 (ID: 5440286) (GT1 × PB235 and GT1 × RRIM701) and accession PRJNA541308 (ID: 541308) (PR255 × PB217).

Genomic Selection Analysis

Phenotypic analysis was carried out jointly for all years of evaluation via the mixed model approach.

Prediction based on genomic relationships and predictive ability assessment was performed via a relationship matrix-based approach for genomic prediction (Habier et al., 2007); the matrix K was the central object denoting the genomic relationship matrix. Two kernel methods were used: the linear kernel method (GB) used by Jarquin et al. (2014) and Lopez-Cruz et al. (2015) and the nonlinear kernel method (GK) proposed by Cuevas et al. (2016). The matrix for the GB (VanRaden, 2008) and GK (Gonzalez-Camacho et al., 2012) methods was obtained via the function G.matrix in snpReady software (Granato and Fritsche-Neto, 2018). Statistical models for genomic predictions taking G×E interactions into account (Jarquin et al., 2014; Lopez-Cruz et al., 2015) combine genetic information from molecular markers or from pedigrees (Pérez-Rodríguez et al., 2015) with environmental covariates, while the López-Cruz model breaks down the marker effect across all environments and the interaction for each specific environment.

The PA was obtained from the correlation between the predicted BLUPs and the observed BLUPs. Four statistical prediction models were fitted to all the data sets to study their PA via random cross-validation (CV) schemes. The main objective was to compare the prediction ability of kinship matrices (GB and GK) and the proposed single-environment and multienvironment (G×E) genomic models.

The PA values were also compared for the single-environment and multienvironment models SM, MM (Jarquin et al., 2014), MDs (Jarquin et al., 2014), and MDe (Lopez-Cruz et al., 2015) and fitted with the GB and GK methods; this was applied to all the data sets for all the studied traits. These analyses were performed to derive estimates of variance components resulting from the main genetic effect, and genetic environment-specific effects and residual effects of the four models described above for SC in the data sets from the two different conditions (LW and WW) were computed. For all GS models, BLUPs were estimated via the mixed model with breedR software (Munõz and Sanchez, 2017), and all models were fitted with G×E interactions via BGGE software (Granato et al., 2018), in which 20,000 iterations were performed (ite = 20,000), 5,000 samples were discarded (burn-in = 5,000), and every fifth iteration was used to estimate the posterior mean (thin = 5).

SM

Using the main effect of the genotype, the single-environment model fits data from each environment (LW and WW) separately. Equation (1) shows the matrix representation of this model.

$$y = \mu 1 + Z_g g + e \quad (1)$$

where $y = (y_1, ..., y_n)$' is the response vector (BLUP), y_i is the observation of t he *i*th line ($i = 1, ..., n$) in each environment, μ is the general mean, Z_g is the incidence matrix that combines the random genetic effects and phenotypes, and *g* and *e* are the random genetic effect and the residual random effect, respectively, for each environment (LW and WW). In SM (1), *g* is considered to present a multivariate normal distribution with a mean of zero and a covariance matrix $\sigma^2_{gj}K$; that is, $g \sim \left(0, \sigma^2_{gj}K\right)$, where σ^2_{gj} is the genetic variance of *g* in the jth environment, and where *K* is a positive semidefinite symmetric matrix that shows the variance–covariance of the genetic values calculated from the molecular markers. Furthermore, the residual error e in each environment is considered to be separate from the homogeneous variance (σ^2_e) and is distributed as $e \sim N\left(0, \sigma^2_{ej}I\right)$, where *I* is the identity matrix and σ^2_{ej} is the residual variance in the jth environment. Thus, g is an estimation of the true unknown genetic values, and e includes the residual genetic effects that are not elucidated by g more other nongenetic effects that approximate the errors, as described by Bandeira et al. (2017). For SM (1), matrix K can be constructed using the linear kernel $K = \left(\frac{XX'}{p}\right)$ (de los

Campos et al., 2012) proposed by VanRaden (2007, 2008) for estimating the GBLUP, where x is the standardized matrix of molecular markers for the individuals, of order *n x p* and *p* is the number of markers. The entries of the GK are computed as $K\left(X_i X_i\right) = exp\left(- hd_{ii'}^2\right)$, where $d_{ii'}$ is the Euclidean distance between the ith and $i^{'th}$ individuals $\left(i = 1, \ldots, n_j\right)$ given by the markers and $h > 0$ is the bandwidth parameter that controls the rate of decay of K values (Pérez-Rodríguez et al., 2013; Cuevas et al., 2016). In this study, GK takes the form $K\left(x_i x_{i'}\right) = exp\left(- hd_{ii'}^2 / median\left(d_{ii'}^2\right)\right)$, where $h = 1$ and where the median of the distances is used as a scaling factor (Crossa et al., 2010). de los Campos et al. (2010) described the theory of the GK in the context of the RKHS KA (KA is well known as the GK, which is based on the Euclidian distance, aiming to capture additive and nonadditive effects).

MM

The MM takes into account the main fixed effects of environments, even in the presence of random genetic effects across environments. Equation (2) indicates the matrix representation of this model.

$$y = \mu 1 + X_E \beta_E + Z_g g + e \quad (2)$$

where $y = \left(y_1, \ldots, y_j, \ldots, y_s\right)'$ is the response vector and yj is the vector of line observations $\left(i = 1, \ldots, n_j\right)$ in the jth environment $(j = 1, \ldots, s)$. The fixed environmental effects in the data are models in the XE incidence matrix, where the intercept for each environment (β_E) is the parameter to be estimated. The incidence matrix Z_g is the other fixed effect that can be incorporated into the model, the matrix Z_g combines genotype with phenotype for each environment, and g is the variance in the main genetic effects across environments. The random vector of genetic effects g across environments is considered to follow a multivariate normal distribution with a mean of zero and a covariance matrix of $\sigma_{g0}^2 K$; that is, $g \sim N\left(0, \sigma_{g0}^2 K\right)$, where σ_{g0}^2 is the variance of the main genetic effects across environments and , $e \sim N\left(0, \sigma_e^2\right)$, as described by Bandeira et al. (2017). We used g with GB or GK.

MDs

The MDs extends the MM to implement the random interaction effect of the environments to incorporate more genetic information of the lines (*ge*).

$$y = \mu 1 + X_E\beta_E + Z_g g + ge + e \quad (3)$$

The vector of random effects of the interaction ge is considered to follow a multivariate normal distribution, $ge \sim N\left(0,\ \left[Z_g K Z_g'\right] o \left[X_E X_E'\right]\sigma^2_{ge}\right)$. Here, (◦) is the Hadamard product operator and indicates, according to Jarquin et al. (2014), the product (element to element) between two matrices in the same order, and σ^2_{ge} is the variance component of the interaction. The *K* matrix is defined the same as that above, and the vector of the main genetic effects is *g*, which presents a multivariate normal distribution with a mean of zero and a covariance matrix $\sigma^2_{g0}K$; that is, $g \sim N\left(0,\ \sigma^2_{g0}K\right)$, with variance of the main genetic effects $\left(\sigma^2_{g0}\right)$ and $e \sim \left(\sigma^2_e I\right)$, as described by Bandeira et al. (2017):

$$[\boldsymbol{Z_g K Z_g'}] \circ [\boldsymbol{X_E X_E'}] = \begin{bmatrix} \mathbf{K_1} & \cdots & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \boldsymbol{K_j} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \cdots & \boldsymbol{K_m} \end{bmatrix}$$

where K_j represents the kernel constructed from the molecular markers of the lines in the *j*th environment. As in the MM, the matrix *K* is used in the variance–covariance for *g* of the MDs and is also a component of the variance–covariance of *ge*. The kernel matrix *K* can be constructed with GB and GK

MDe

In the MDe (Lopez-Cruz et al., 2015), the genetic effects of markers are partitioned into main marker effects across all environments and specific marker effects within each environment. Equation (4) indicates the matrix representation of this model:

$$y = \mu 1 + X_E\beta_E + Z_g g_0 + gE + e \,(4)$$

where g_0 denotes the main effect of the markers with a variance–covariance $g_0 \sim N\left(0,\ \sigma^2_{g0} K\right)$ across all environments, σ^2_{g0} is common to all s environments, and the borrowing of information among environments is generated through the kernel matrix K. Otherwise, the specific effect of the markers $\left(g_E\right)$ in environments or even the effects of the interaction with a variance–covariance structure differ from those of model (4); in other words, $g_E \sim N\left(0,\ K_E\right)$, where K_E is as follows:

$$K_E = \begin{bmatrix} \sigma^2_{g_{E_1}} \boldsymbol{K} & \cdots & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \sigma^2_{g_{E_j}} \boldsymbol{K}_j & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \cdots & \sigma^2_{g_{E_j}} \boldsymbol{K}_m \end{bmatrix}$$

$$= \begin{bmatrix} \sigma^2_{g_{E_1}} \boldsymbol{K} & \cdots & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots & \cdots & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \cdots & \mathbf{0} \end{bmatrix} + \cdots$$

$$+ \begin{bmatrix} \mathbf{0} & \cdots & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \sigma^2_{g_{E_j}} \boldsymbol{K}_j & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \cdots & \mathbf{0} \end{bmatrix} + \cdots$$

$$+ \begin{bmatrix} \mathbf{0} & \cdots & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \cdots & \sigma^2_{g_{E_s}} \boldsymbol{K}_s \end{bmatrix}$$

The matrix K_E can be expressed as the sum of s matrices, and the effects given by gEj are specific for the *j*th environment, which has a variance–covariance matrix of $\sigma^2_{gEj} K_j$. These two terms (*g* and *gE*) of the MDe are given by the components of the estimated variance for the data. The kernel matrix *K* is used in the components of *g*, while kernel matrix K_E is used in the component of g_E; both *K* and K_E can be used with GB or GK, as described by Bandeira et al. (2017).

Assessing Prediction Accuracy by Random Cross-Validation

The PA of SM-method combinations was evaluated with the TRN set (which comprised 80% of the hybrids). The TST set comprised 20% of the individuals, and none of the lines to be predicted in the TST set were also in the TRN set, in which 5 random partitions were arranged 5-fold, with 100 random partitions each. This procedure was performed separately for each environment, namely, LW and WW, and the SMs were fitted separately for each environment.

The PA values of the multienvironment (LW and WW) model-method combinations were generated using two different cross-validation (CV) designs according to the methods of Burgueño et al. (2012). The random CV 1 design (CV1) assumes that new genotypes have not been tested or evaluated in either environment, where 20% of genotypes were not phenotyped in any environment and had to be predicted. The random CV 2 design (CV2) is a simulation of genotypes that has been evaluated in some environments but not in others. The CV2 design can be used only for multienvironment modeling methods (MM, MDs, and MDe) and not for single-environment (SM) modeling methods where the random CV is CV1.

The parameters of the models, which include the main genetic effects, variance components resulting from residual effects, G×E interaction effects, and environment-specific effects, were reestimated from the TRN data in each TRN-TST partition (50 random), and the models were fitted to the TRN data set. PA was assessed by computing Pearson's product-moment correlations between predictions and phenotypes in the TST data set within environments.

Expected Genetic Gain

Expected genetic gain (EGG) was estimated in two ways: the classic method used in rubber tree breeding via the breeder's equation and phenotypic data and with information from the SNPs obtained via GS. The EGG was calculated according to the methods of Matias et al. (2019) and Grattapaglia (2017).

Expected Genetic Gain Obtained by a Classic Breeding Cycle With Only Phenotypic Information Used

The EGGs obtained by a classic breeding cycle (EGGcs) were estimated under the assumption that the time for first selection is 10 years. In rubber tree breeding, 3 years are needed from pollination to planting in the field, and as rubber trees usually require 6 years or more to reach tapping girth, there is a wait time of 7–9 years until tapping is started and a

long period of 10 to 15 years before production and adaptation can be evaluated in the field (Gonçalves et al., 2005) according to the following equation:

$$EGGc = \frac{rc.i.\delta g}{T}$$

where *rc* is the accuracy of selection, in which the breeding improvement is equivalent to the square root of the H^2, *i* is the intensity selection, δg is the additive genetic standard deviation, and *T* is the selection cycle time.

Expected Selection Gain via Molecular Marker Information

The simulation of breeding cycles in which GS was used was based on the EGG via molecular marker information (EGGgs) equation, assuming a time of 3 years for each selection cycle and representing the time required for crossing, seed selection, and selection of the best individuals via molecular markers. The equation is as follows:

$$EGGgs = \frac{rgs.i.\delta g}{T}$$

where rgs is the selection accuracy with GS $\left(\frac{PA}{\sqrt{H^2}}\right)$, *i* is the intensity selection, δg is the additive genetic standard deviation, and *T* is the selection cycle time.

Results

Single-Nucleotide Polymorphisms Calling

We started with 435 genotypes, but three genotypes were replicates and thus were merged. We removed 27 individuals that had more than 50% missing SNPs, leaving 411 genotypes. After the data were analyzed, a total of 259.224 million reads of sequence data were obtained, of which 69.8% were high-quality barcoded reads. The overall alignment rate of these reads to the rubber tree reference genome (Tang et al., 2016) was 83.7%, and 23.1% were aligned exactly one time.

A total of 107,294 SNPs were identified. After markers 1) with more than 20% missing data, 2) with an MAF ≤ 0.05, 3) with more than two alleles were excluded, tags with a minimum depth of six reads were aligned to the *H. brasiliensis* reference genome sequence (Tang et al., 2016). This method was based on that of previous studies of other species, in which the authors argued that, compared with high-depth sequencing, low-depth (approximately 2–4X) sequencing enables more individuals to be genotyped for the same

cost, which, according to Li et al. (2011), is a good strategy for genome-wide association studies (GWAS). Gorjanc et al. (2015) obtained similar results for GS studies and reported that optimal PA was obtained via low-depth sequencing (approximately 1–2X) of many genotypes.

After the data were filtered, 6.7% were missing. The mean depth ranged from 444 to 521 for GT1 × RRIM701 and GT1 × PB235, respectively, and was 202 for PR255 × PB217 (Supplementary Figure 2). Although large variation was observed between populations, only SNPs with at least six reads were selected, and the entire data set was reduced to 30.546 SNPs.

Estimates of Genetic Parameters by Single-Nucleotide Polymorphism Genotyping

Using the genotyped SNPs, we assessed the population structure via principal component analysis (PCA), and the results indicated that the 411 genotypes fell into two major clusters (Supplementary Figure 3), which mainly contained hybrids derived from the PR255 × PB217 cross and hybrids derived from the GT1 × RRIM701 and GT1 × PB235 crosses. The first two PCs explained 19.5 and 2.2% of the total variance, respectively, clearly splitting the groups along the x- and y-axes.

Descriptive Statistics

Box plots of SC in each environment are depicted in Supplementary Figure 4. The distribution of this trait in the environments was symmetrical (data not shown). The LW environment exhibited relatively high increases in SC, while the WW environment exhibited relatively low increases (Supplementary Figure 4). The trees grew better under increased water availability (WW) (data not shown); however, because the phenotypic measurements were taken twice per year for each tree, the phenotypes of the trees under WW were always measured at the beginning of the year, whereas the phenotypes of the trees under the LW were taken half a year later. This method inevitably generates a small difference between the two phenotypes because the trees under LW are older than those under WW when the same measurements are taken. Because rubber populations require extensive field trial planting, it is not feasible for a breeding program to maintain two planting areas and to examine two hydric conditions with trees of the same age.

To assess how much of the phenotypic variation is genetically controlled and thus appropriate for GS, we first estimated the H^2 of SC. The H^2_{env} ranged from 0.51 to 0.50 for

LW and WW, respectively (Table 1). The populations PR255 × PB217, GT1 × RRIM701, and GT1 × PB235 were evaluated in different environments that presented different site indexes (soil and climate conditions). The phenotypic variance differed between the sites, but we had common check genotypes in both environments; these checks served the connection between populations and other factors. Furthermore, on the basis of the results of heritability and residual distribution, it is evident that this approach allowed a reliable estimate of the error, factor effects, and their interactions.

TABLE 1 Phenotypic variation: heritability (H^2), variance genotype × environment (G×E) interaction (σ^2_{gxe}), residual variance (σ^2_e), genetic variance main effect (σ^2_g), and coefficients of experimental variation (CVe%s) in environments with low-water conditions (LW) and with well-watered conditions (WW) considered together and alone, with p < .01 indicated by **.

	General	LW	WW
σ^2_g	3.61**	4.33	3.69
σ^2_{gxe}	0.81	–	–
σ^2_e	16.15	16.75	14.75
H^2	0.60	0.51	0.50
CVe%	20.00	20.40	19.10

On the basis of the phenotypic data, the estimates of genotypic variance (σ2g) and G×E interaction variance (σ^2_{gxe}) were relatively high (3.61) and relatively low (0.81), respectively, and both were significant. Under LW, σ^2_g (4.33) was greater than that under WW (3.69). Similarly, the residual variance (σ^2_e) estimate was greater under LW (16.75) than under WW (14.75). The coefficients of experimental variation (CVe%s) (Table 1) presented an overall value of 20%, ranging from 20.4% (LW) to 19.1% (WW), and were considered moderate.

Estimates of the Variance Components

The estimates of the variance components for each of the GS models derived from the full data analysis are presented in Table 2.

TABLE 2 Estimates of different variance components for the following genomic selection (GS) models: the single-environment, main genotypic effect model (SM); the multienvironment, main genotypic effect model (MM); the multienvironment, single-variance genotype × environment (G×E) deviation model (MDs); and the multienvironment, environment-specific variance G×E deviation model, with the genomic best linear unbiased predictor (GBLUP, GB) and Gaussian kernel (GK) for stem circumference (SC).

	SM				MM		MDs		MDe			
	GK		GB		GK	GB	GK	GB	GK		GB	
	WW	LW	WW	LW	–	–	–	–	LW	WW	LW	WW
σ^2_g	0.47	0.50	0.44	0.46	0.96	0.99	0.93	0.97	0.80		0.90	
σ^2_e	0.53	0.50	0.56	0.54	0.04	0.01	0.02	0.01	0.02		0.01	
σ^2_{gxe}	–	–	–	–	–	–	0.05	0.02	–	–	–	–
σ^2_{env}	–	–	–	–	–	–	–	–	0.11	0.07	0.06	0.03

The genetic variance (σ^2_g), *residual variance* (σ^2_e), *mean environmental genetic variance* (σ^2_{gxe}), *and environment-specific genetic variance* (σ^2_{env}) *are shown.*

The variance components of genetic effects were greater when the GB method was used rather than when the GK method was used in all environments for the SM. Both the genetic variance and the environmental variance were greater when analyzed in the LW (Table 2). The residual variance for SM-GB was lower than that for SM-GK in all environments.

Compared with exclusion of the interaction term (G×E), inclusion of the term when the MM, MDe, and MDs were used led to a more significant reduction in the estimated residual variance, and for all the environments, the residuals from the GK were smaller than the residuals from the GB for the MM, MDs, and MDe. However, the multienvironment model (LW versus WW) assumed that there was no marker-environment interaction between families tested at the different sites and that there could be an effect between marker effect estimations from families tested at different sites and environments (LW and WW). This should be taken into consideration and should be carefully analyzed in the GK approach when additive vs. additive epistasis is targeted.

The residual variance components of MM-GK (corresponding to 4% of the total variance) and MM-GB (corresponding to 1% of the total variance) were similar; the genetic variance corresponded to 99% for MM-GB and 96% for MM-GK. The percentage of total variance corresponding to variance components of the genetic main effects of MDs-GK (93%) and MDe-GK (80%) was consistently smaller than that of the genetic main effects of MDs-GB (97%) and MDe-GB (90%) (Table 2). These results indicate that the G×E model (MM, MDe, or MDs) fits the data better than do single-environment models.

Assessment of Prediction Accuracy

The estimated correlations between the phenotypes and predictions obtained from the CV test are shown in Figure 1 for the single-environment model (SM) and the multienvironment models (MM, MDs, and MDe).

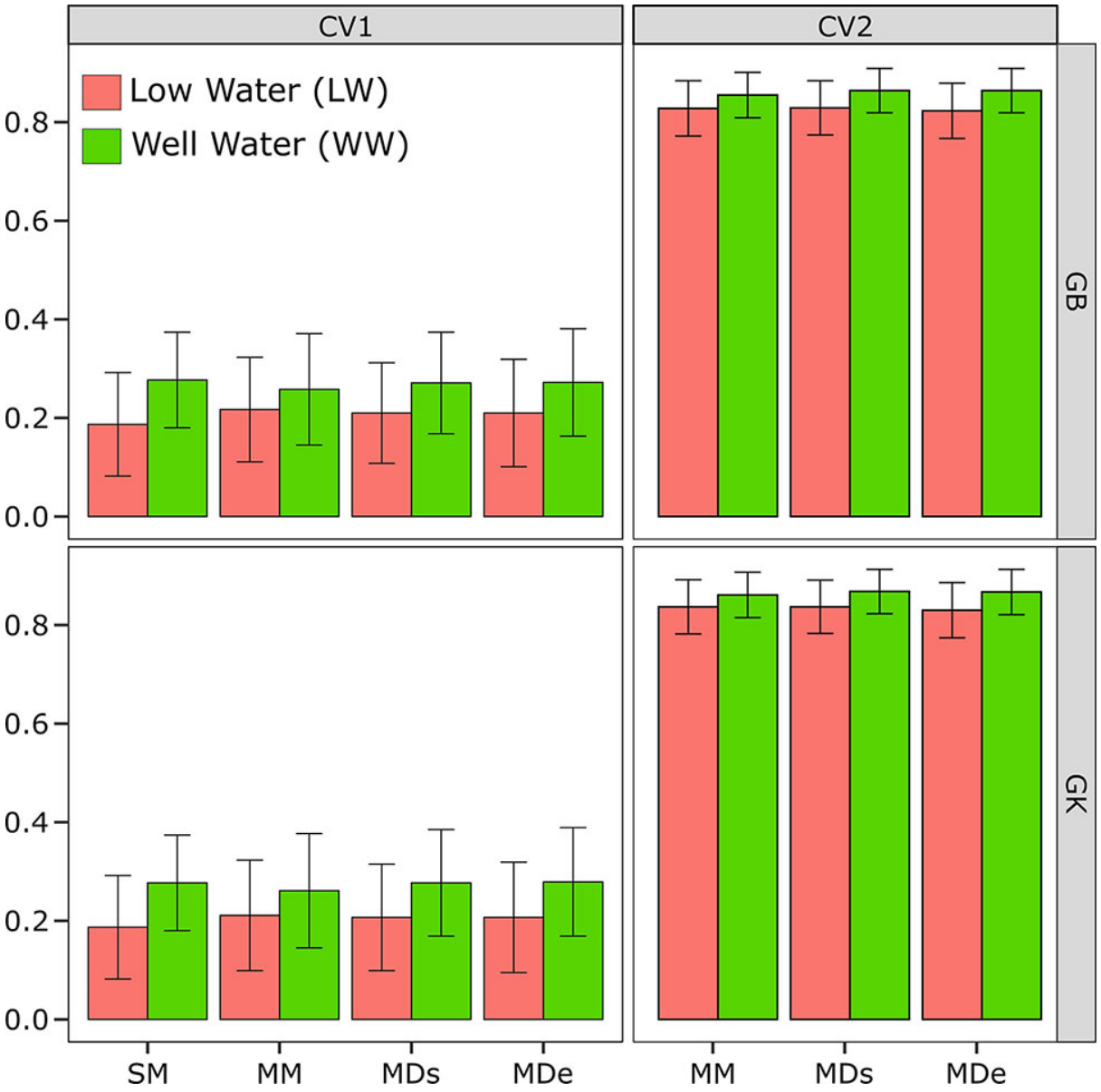

FIGURE 1 Correlations between phenotypes and prediction values for the single-environment, main genotypic effect model (SM) with the genomic best linear unbiased predictor (GBLUP) kernel method (SM-GB) and with the Gaussian kernel (GK) method (SM-GK); multienvironment, genotypic effect model with the GBLUP kernel (MM-GB) and with the GK (MM-GK); multienvironment, single-variance G×E model with the GBLUP kernel (MDs-GB) and with the GK (MDs-GK); and multienvironment, environment-specific variance G×E model with the GBLUP kernel (MDe-GB) and with the GK (MDe-GK) for

stem circumference (SC). The environments included one with low-water conditions (LW) and one with well-watered conditions (WW).

Single-Environment Model (SM)

The CV2 design can be used only for multienvironment modeling methods (MM, MDs, and MDe) and not single-environment modeling methods (SM). Therefore, a single environment (SM) is analyzed, the random CV is CV1, but it is applied to only one individual environment (LW or WW) (Figure 1).

The results showed that the PA of the SM-GK combination was greater than that of the SM-GB combination under both LW and WW. The SC results were 0.19 under LW for SM-GB and 0.19 for SM-GK, and under WW, the results were 0.27 for SM-GB and 0.28 for SM-GK.

Multienvironment Models (MM, MDe, and MDs)

In terms of evaluating the PA of a model based on the correlation between the observed and the predicted values, when the PAs obtained by implementing different models (MM, MDs, and MDe) were compared, all the models were most accurate when CV2 was applied. The PA varied considerably between the CV1 and CV2 conditions (Figure 1). When only a random CV2 was considered, the PA results were very similar and varied little between environments.

The PA varied very little between the WW and LW (Figure 1). The results obtained with the model-method combinations were very similar. Generally, under LW, the best model was the GK, which did not differ between the methods (0.84), and MM-GB exhibited similar results (0.82). Relatively low PA values were obtained using the GB; the PA was 0.82 for the MDs and 0.83 for the MDe (Figure 1). Under WW, the model-method combinations presented the same values; the PA ranged from 0.86 for the MM to 0.87 for the MDe and MDs with both GK and GB (Figure 1).

Expected Genetic Gain

The investigated alternative rubber tree breeding strategies differed considerably in the number of years required to finish one breeding cycle. For the classic improvement strategy, we considered a minimum duration of 10 years for the beginning of the selection of the best genotypes because 3 years are required from pollination to planting in the field and because rubber trees generally require several additional years (often 6 or more) to reach

tapping girth. In the case of GS, we considered 3 years for initial selection (from pollination to field planting).

The EGG calculations were performed for all the methods and models and were compared with classic improvements in both environments (Figure 2 and Supplementary Table 3). When the CBM, which takes into account only phenotypic data, was used, the selection gain without considering the environment was 0.08, and if the data were separated by environment, the EGGc was the same (0.07) for LW and WW (Figure 2). When we incorporated genotypic information in a single environment (SM), the genetic gain increased to 0.13 for the WW when GB was used and 0.09 when GK was used, while for LW, there was no difference between GK and GB.

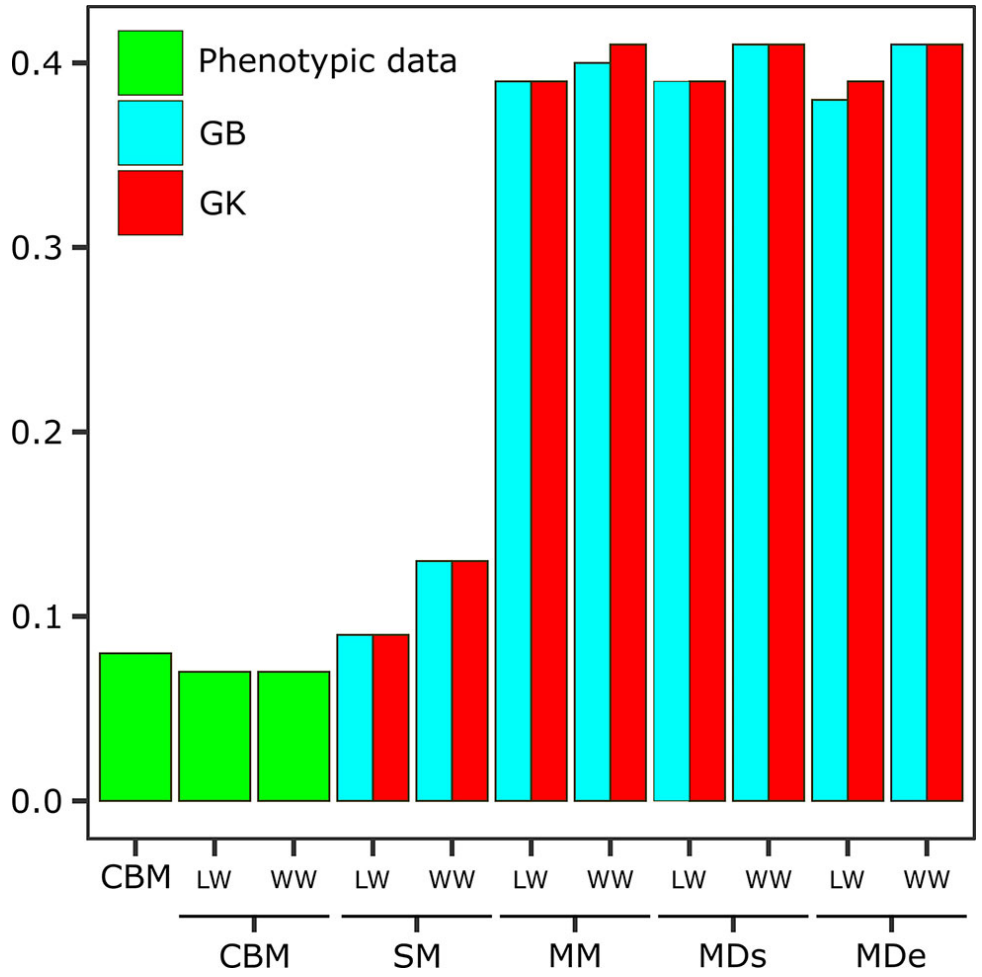

FIGURE 2 Expected genetic gain (EGG) obtained via the classic breeding method (CBM) with phenotypic data sets and analyzed in separate environments [one with low-water conditions (LW) and one with well-watered conditions (WW)] and EGG obtained via the following genomic selection (GS) models: the single-environment, main genotypic effect model (SM); multienvironment, genotypic effect model (MM); multienvironment, single-variance G×E model (MDs); and multienvironment, environment-specific variance

G×E model (MDe), with GB and GK shown in the two evaluated environments (LW and WW).

The genetic gains obtained when the information from the G×E interaction was incorporated were much greater gains than those obtained from a single environment. However, the results varied little between methods, with similar values resulting from most of the analyses. Considering the overall LW, the EGG was 0.39 for all models and methods except MDe-GB (0.38). For WW, the EGG was slightly greater than that under LW, and most models and methods had estimated gains of 0.41, with the exception of MM-GB (0.40) (Figure 2).

Discussion

Incorporating and improving the genomic PA of rubber trees are a challenge for the successful application of GS in breeding programs. In this research, genomic PA was studied in rubber tree data sets via the GB and GK methods in conjunction with multienvironment models that evaluated trees under contrasting hydric conditions in different seasons of the year (LW and WW).

Many factors such as genetic architecture, heritability, population structure, and marker density can influence GS (Crossa et al., 2017). According to Meuwissen et al. (2001), GS is expected to increase the accuracy of selection, particularly for traits that have a low heritability and that cannot be measured directly from breeding candidates.

The accuracy of GS also depends on the genetic architecture of traits, such as heritability which are positively related to PA. Complex traits that present low heritability and small marker effects are suitable for GS. Our analyses revealed moderate heritability estimates for SC ranging from 0.50 to 0.51, with the lowest value for WW and the highest for LW (Table 1). Nevertheless, the heritabilities estimated in this study were within the range of those estimated in other studies for SC in *Hevea*, which were H^2= 0.32 (Moreti et al., 1994) and H^2 = 0.47 (Gonçalves et al., 1999).

The CVe% for SC (Table 1) ranged from 20.4% (LW) to 19.1% (WW), which is considered moderate according to the classification proposed by Costa et al. (2010), who described the coefficient of variation as a useful tool to efficiently and accurately specify the experimental results: the lower the CVe% is, the more homogeneous the data, and the less

environmental interference. The environmental variation, genotypes, and interaction between these two factors were highly significant, indicating that the environments used were contrasting, that there was genetic variability among the genotypes, and that the genotypes performed differently depending on the environment.

Using the CBM, Moreti et al. (1994) estimated genetic parameters and expected gains via the selection of juvenile characters in rubber tree progeny, and some parameters (rubber production, bark thickness, and SC) positively stood out. Gonçalves et al. (1996) observed the same phenomenon in the results reported by Moreti et al. (1994), showing a correlation and its applicability to the selection process. Strong phenotypic and genetic correlations were observed between yield and SC, indicating the possibility of obtaining young clones of good productive capacity and high vigor (Gonçalves et al., 1984). This correlation in conjunction with moderate heritability could be used to perform early selection of more productive clones without the need to wait for the trees to enter production, which requires an extended evaluation period.

Trees with rapid SC development may be more productive, and this feature may be a useful way to predict more productive hybrids via GS. Given this and latex production having greater heritability than circumference growth because the influence of the rootstock is relatively low in production, this feature will be very important in future studies of this population.

In GS, G×E interactions can be modeled by a marker × environment interaction and by a linear kernel or a nonlinear GK (Cuevas et al., 2016). Multienvironment genomic prediction was successfully implemented using a GBLUP model; however, depending on the genetic architecture of the trait and germplasm, nonlinear semiparametric approaches such as GK could produce more accurate results than could linear approaches (Cuevas et al., 2016).

Here, the GK methods presented a small increase in the prediction ability of all single-environment and multienvironment models with CV2, confirming the results of Lopez-Cruz et al. (2015) and Zhang et al. (2015) and demonstrating that predicting new genotypes is more complicated than predicting genotypes that have been evaluated in correlated trials. The GB method was superior when analyzed only via CV1 under LW.

The multienvironment models and the GK method resulted in the best PA. Similar decreases in PA were reported by Lopez-Cruz et al. (2015) when wheat lines were used and by Bandeira et al. (2017) when a maize data set was analyzed in attempts to predict lines in untested environments under a CV1 random partitioning scheme.

Considering only random CV2, the PA was slightly greater under WW, ranging from 0.87 (MDe-GK-WW) to 0.82 (MDe-GB-LW), which is consistent with previously published results for forest tree species. Bartholome et al. (2016) reported medium to high PAs for all traits studied (0.52 to 0.91) in maritime pine. Similar accuracies were reported for the height of loblolly pine trees, with values ranging from 0.64 to 0.74 (Resende et al., 2012b), and eucalyptus hybrids (0.66 to 0.79) (Resende et al., 2012a), regardless of differences in GS models, species, and population structure between studies.

If information concerning WW and LW was combined with multienvironment models, the results were superior to those of single-environment genomic models with GB and GK. This finding suggests that introducing interactions between markers and environmental conditions can increase the proportion of variance accounted for by the model and, more importantly, can increase the PA. Optimized crosses via selection of the best stable parents can then be performed to improve hybrid stability and the EGG (Toro and Varona, 2010).

G×E interactions are essential in many aspects of a breeding program, and the increase in PA with the inclusion of environmental information represents a favorable result with important implications for both breeding and agronomic recommendations. In rubber tree breeding, progeny testing is commonly used to evaluate the performance of new genotypes. Thus, in this case, new hybrids identified as high-performance hybrids with stable development throughout the year can be selected for use in new biparental crosses or new population selections. Interactions in field trials affect both early selection and mature selection; therefore, when the effectiveness of early selections is evaluated, it is important to determine whether the G×E interactions among environments significantly affect the genetic correlation of early maturity.

Application of the combinations of four models (SM, MM, MDs, and MDe) and two kernel methods (GB and GK) to rubber tree data sets revealed that the PAs of the models with the nonlinear GK were similar to those of the models with the linear GB kernel. According to Gianola et al. (2014), the GK has a better predictive ability and a more flexible structure than does the GB, and the GK can capture nonadditive effects between markers.

Akdemir and Jannink (2015) presented different choices for estimating kernel functions: linear kernel matrices incorporate only the additive effects of the markers, polynomial kernels incorporate different degrees of marker interactions, and the GK function uses complex epistatic marker interactions. GK would be more appropriate for GS of rubber

trees because of the possibility of exploiting the local epistatic effects captured in the GK and their interactions with environments.

Many GS studies of plants have focused on breeding programs that generally evaluate crops in multiple environments, such as in different seasons/years or in geographic locations, to determine performance stability across environments (Crossa et al., 2016) and to identify markers whose effects are environment specific or whose effects are stable across environments (Crossa et al., 2016; Oakey et al., 2016). Previous studies in wheat (Lopez-Cruz et al., 2015) expanded the single-trait GB model to a multienvironment context and revealed substantial gains in PA with the multienvironment model compared with the single-environment model.

Advantages of GS applied to the improvement of forest species have been demonstrated. For example, Wong and Bernardo (2008) and Iwata et al. (2011) demonstrated the potential uses of GS and concluded that it could dramatically increase tree breeding efficiency. The advantage of marker-based relationship matrices is that gaps in pairwise relatedness in forest tree pedigrees are filled, which leads to increased accuracy of breeding candidate selection (Muller et al., 2017; Tan et al., 2017).

Using both genetic markers and environmental covariates, Cuevas et al. (2016) modeled G×E interactions, and Granato et al. (2018) introduced the Bayesian Genomic Genotype × Environment (BGGE) R package, which fits genomic linear mixed models to single environments and multiple environments with G×E models. These studies showed that modeling multienvironment interactions can lead to substantial gains in the PA of GS for rubber tree breeding programs.

GS is expected to increase the accuracy of selection, especially for traits that cannot be measured directly from breeding candidates and for traits with a low heritability (Meuwissen et al., 2001), and this effect was confirmed in the present study. The selection gain with GS for SC was on average 0.40, while the genetic gain with the CBM was 0.08. When the CBM for rubber trees was compared with the GS method while the multienvironment strategy was applied (MM, MDe and MDs), GS resulted in a five-fold greater genetic gain for SC.

Implementing Genomic Selection in Rubber Tree Breeding Programs

In the last decade, many statistical models have been proposed for applying GS in plant and animal breeding programs and have received increasing interest from forest tree breeders. Resende et al. (2012a, 2012b) demonstrated encouraging prospects of this new

method, and the potential for GS in conifers, pines, and eucalypts has since been confirmed (Zapata-Valenzuela et al., 2013; Lima, 2014; El-Dien et al., 2015; Ratcliffe et al., 2015; Bartholome et al., 2016; Isik et al., 2016), supporting further the potential for GS to accelerate the breeding of forest trees. In the case of rubber trees, a recent study explored GS in a breeding program (Cros et al., 2019).

In this study, we used three full-sib populations, taking advantage of breeding populations that had already been genotyped and phenotyped (Rosa et al., 2018; Conson et al., 2018). This type of population is favorable for GS because of the high LD between marker alleles and genetic alleles. Similar results were obtained in a recent study in which a biparental rubber tree population with 189 and 143 clones of the cross PB260 × RRIM600 was used; the population was genotyped with a limited number of markers (332 simple sequence repeat markers) (Cros et al., 2019), which resulted in a GS accuracy of 0.53. Other plant species have also been evaluated, with GS accuracies reaching moderate to high values (0.59 and 0.91) in a family of 180 Citrus clones (Gois et al., 2016).

For rubber trees, the time required to complete a breeding cycle and recommend a clone for commercial production can span multiple decades and is divided mainly into three selection stages. First, the aim is to obtain progeny by controlled or open pollination and to establish nurseries. At two and-a-half years, on the basis of early evaluations of yield, vigor, and tolerance to disease, breeding trees are selected and cloned for testing at a small scale. During this second stage of the selection cycle, after the first 2 years of tapping, promising clones are multiplied and subsequently evaluated in large-scale or regional trials. This last stage usually takes 12 to 15 years, until it is possible to recommend a clone for large-scale cropping. Therefore, it takes approximately 30 years to complete the breeding cycle, from controlled pollination to final cultivar recommendation (Gonçalves and Fontes, 2012).

In essence, implementing techniques that reduce the long breeding cycle of trees is urgently needed, and for this purpose, the use of a biparental population was a means of managing the difficulty of obtaining complex families, which can take many years to generate because of the low fecundity of trees and the long duration of the phenotypic evaluation needed. According to Cros et al. (2019), a GS approach in which a complex population involving several families is used could lead to variation in GS among selection candidates depending on their relationships with the TRN individuals, leading to GS accuracies lower than those from family-specific TRN populations.

In addition, large areas are required for the development of hybrids, which not only increases the costs associated with maintaining plants in the field but also limits the number

of genotypes that can be evaluated. GS can minimize these difficulties because selection can be performed on juvenile plants, which reduces the interval between generations and increases the intensity of selection, thus reducing the gain per unit time (Resende et al., 2008; Wong and Bernardo, 2008; Heffner et al., 2009).

The use of GS could dramatically reduce the time required for completion of a genetic improvement cycle by eliminating phenotypic progeny testing aimed at selecting the best individuals (replaced by GS), significantly increasing the genetic gain relative to that obtained by CBMs. Another advantage of GS compared with phenotypic selection is that more candidate genotypes are generated; therefore, the population size for selection is improved. All of the candidates are genotyped, and those with the best-predicted test cross values are evaluated in the field; this process can be considered a form of indirect selection.

According to Heffner et al. (2009), even when only moderate accuracy is obtained with GS, it is possible to obtain a genetic gain greater than that obtained by phenotypic selection, as GS reduces the duration of the selection cycle. According to Wong and Bernardo (2008), the selection cycle was shortened from 19 to 6 years when GS was implemented in oil palm. Similar results were observed in the present study, in that the length of the selection cycle was also reduced.

With declining costs and rapid advances in genotyping methods, even with the costs of maintaining large progeny trials and the potential for increased gains per unit time, we very cautiously expect GS to have excellent potential for implementation in rubber tree breeding programs. However, additional studies examining populations with different structures (which were not assessed in this initial work) are necessary before recommending GS for operational implementation in tree breeding programs.

This is the first study to incorporate models for G×E interaction when phenotypic and/or genotypic information was used simultaneously for genetic prediction in the context of GS in a rubber tree breeding program. The results presented here suggest that GS can be useful for rubber tree breeding because this technique can be used to accurately predict the phenotypes and reduce the length of the selection cycle. Thus, GS is a promising tool for improving rubber tree cultivation, and we look forward to exploring the historical phenotypic data collected during 15 years as part of national breeding programs.

Data Availability Statement

Publicly available datasets were analyzed in this study. This data can be found here: https://www.ncbi.nlm.nih.gov/bioproject/PRJNA540286 and https://www.ncbi.nlm.nih.gov/bioproject/PRJNA541308

Author Contributions

LS, FF, PG, RF-N and AS designed the study and performed the experiments; VG and EJ performed the experiments in the field; LS, FF, and RF-N analyzed the data; LS, FF and AS wrote the manuscript. and VG and EJ performed the experiments in the field.

Funding

The authors gratefully acknowledge the Fundação de Amparo a Pesquisa do Estado de São Paulo (FAPESP) for a Ph.D. fellowship to FF (18/18985-7); the Coordenação de Aperfeiçoamento do Pessoal de Nível Superior (CAPES) for financial support (Computational Biology Program and CAPES-Agropolis Program) and postdoctoral fellowships to LS (88887.334728/2019-00); and the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for financial support, a postdoctoral fellowship to LS (168028/2017-4), and research fellowships to AS and PG.

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.

Acknowledgments

We are grateful to the Michelin group in Brazil for allowing access to data from their breeding experiments in Itiquira (Mato Grosso, Brazil) as part of the data analyzed in this

article. This manuscript was previously posted to bioRxiv https://www.biorxiv.org/content/10.1101/603662v1.

Supplementary Material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2019.01353/full#supplementary-material

References

Akdemir, D., Jannink, J. L. (2015). Locally epistatic genomic relationship matrices for genomic association and prediction. Genetics 199, 857–871. doi: 10.1534/genetics.114.173658

Alvares, C. A., Stape, J. L., Sentelhas, P. C., De Moraes Gonçalves, J. L., Sparovek, G. (2013). Köppen's climate classification map for Brazil. Meteorol. Zeitschrift 22, 711–728. doi: 10.1127/0941-2948/2013/0507

Bandeira, E. S. M., Cuevas, J., De Oliveira Couto, E. G., Perez-Rodriguez, P., Jarquin, D., Fritsche-Neto, R., et al. (2017). Genomic-enabled prediction in maize using kernel models with genotype x environment interaction. G3 (Bethesda) 7, 1995–2014. doi: 10.1534/g3.117.042341

Bartholome, J., Van Heerwaarden, J., Isik, F., Boury, C., Vidal, M., Plomion, C., et al. (2016). Performance of genomic prediction within and across generations in maritime pine. BMC Genomics 17, 604. doi: 10.1186/s12864-016-2879-8

Burgueño, J., Crossa, J., Miguel Cotes, J., San Vicente, F., Das, B. (2011). Prediction assessment of linear mixed models for multienvironment trials. Crop Sci. 51, 944–954. doi: 10.2135/cropsci2010.07.0403

Burgueño, J., De Los Campos, G., Weigel, K., Crossa, J. (2012). Genomic prediction of breeding values when modeling genotype × environment interaction using pedigree and dense molecular markers. Crop Sci. 52, 707–719. doi: 10.2135/cropsci2011.06.0299

Conson, A. R. O., Taniguti, C. H., Amadeu, R. R., Andreotti, I. A. A., de Souza, L. M., Dos Santos, L. H. B., et al. (2018). High-resolution genetic map and QTL analysis of growth-related traits of Hevea brasiliensis cultivated under suboptimal temperature and humidity conditions. Front. Plant Sci. 9, 1255. doi: 10.3389/fpls.2018.00513

Costa, R. B., Resende, M. D. V., Gonçalves, P. S., Roa, R. A. R., Feitosa, K. C. O. F. (2010). Genetic parameters and values prediction for growth and latex production traits in rubber tree progenies. Bragantia 69, 49–56. doi: 10.1590/S0006-87052010000100007

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. doi: 10.1016/j.indcrop.2019.111464

Crossa, J., de los Campos, G., Perez, P., Gianola, D., Burgueño, J., Araus, J. L., et al. (2010). Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186, 713–724.

Crossa, J., De Los Campos, G., Maccaferri, M., Tuberosa, R., Burgueño, J., Pérez-Rodríguez, P. (2016). Extending the marker × environment interaction model for genomic-enabled prediction and genome-wide association analysis in durum wheat. Crop Sci. 56, 2193–2209. doi: 10.2135/cropsci2015.04.0260

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., et al. (2017). Genomic selection in plant breeding: methods, models, and perspectives. Trends Plant Sci. xx, 1–15. doi: 10.1016/j.tplants.2017.08.011

Cuevas, J., Crossa, J., Soberanis, V., Pérez-Elizalde, S., Pérez-Rodríguez, P., de los Campos, G., et al. (2016). Genomic prediction of genotype × environment interaction kernel regression models. Plant Genome 9, 3. doi: 10.3835/plantgenome2016.03.0024

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., Depristo, M. A., et al. (2011). The variant call format and VCFtools. Bioinformatics 27, 2156–2158. doi: 10.1093/bioinformatics/btr330

de los Campos, G., Gianola, D., Allison, D. B. (2010). Predicting genetic predisposition in humans: the promise of whole-genome markers. Nature Rev. Genet. 11, 880–886. doi: 10.1038/nrg2898

de los Campos, G., Hickey, J. M., Pong-Wong, R., Daetwyler, H. D., Calus, M. P. L. (2012). Whole genome regression and prediction methods applied to plant and animal breeding. Genetics 193, 1255–1268. doi: 10.1534/genetics.112.143313

El-Dien, O. G., Ratcliffe, B., Klápště, J., Chen, C., Porth, I., El-Kassaby, Y. A. (2015). Prediction accuracies for growth and wood attributes of interior spruce in space using genotyping-by-sequencing. BMC Genomics 16, 370. doi: 10.1186/s12864-015-1597-y

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., et al. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6, e19379. doi: 10.1371/journal.pone.0019379

Federer, A. W. T., Raghavarao, D. (1975). On augmented designs. Biometrics 31, 29–35.

Furlani, R. C. M., Moraes, M. L. T. D., Resende, M. D. V. D., Furlani Junior, E., Gonçalves, P. D. S., Valério, W. V. F., et al. (2005). Estimation of variance components and prediction of breeding values in rubber tree breeding using the REML/BLUP procedure. Genet. Mol. Biol. 28, 271–276. doi: 10.1590/S1415-47572005000200017

Gianola, D., Fernando, R. L., Stella, A. (2006). Genomic-assisted prediction of genetic value with semiparametric procedures. Genetics 173, 1761–1776. doi: 10.1534/genetics.105.049510

Gianola, D., Weigel, K. A., Krämer, N., Stella, A., Schön, C. C. (2014). Enhancing genome-enabled prediction by bagging genomic BLUP. PLoS One 9, e91693. doi: 10.1371/journal.pone.0091693

Glaubitz, J. C., Casstevens, T. M., Lu, F., Harriman, J., Elshire, R. J., Sun, Q., et al. (2014). TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. PLoS One 9, e90346. doi: 10.1371/journal.pone.0090346

Gois, I. B., Borém, A., Cristofani-Yaly, M., de Resende, M. D. V., Azevedo, C. F., Bastianel, C. F., et al. (2016). Genome wide selection in citrus breeding. Genet. Mol. Res. 15, gmr15048863. doi: 10.4238/gmr15048863

Gonçalves, P. S., Furtado, E. L., Bataglia, O. C., Ortolani, A. A., May, A., Belletti, G. O. (1999). Genetics of anthracnose panel canker disease resistance and its relationship with yield and growth character in half-sib progenies of rubber tree (Hevea brasiliensis). Genet. Mol. Biol. 22, 583–589.

Gonçalves, P. S., Bortoletto, N., Cardinal, Á. B. B., Gouvea, L. R. L., Costa, R. B., de Moraes, M. L. T. (2005). Age-age correlation for early selection of rubber tree genotypes in São Paulo State, Brazil. Genet. Mol. Biol. 28, 758–764. doi: 10.1590/S1415-47572005000500018

Gonçalves, P. S., Fontes, J. R. A. (2012). “Domestication and breeding of rubber tree,” in Domestication and breeding – amazonian species, vol. 393–41. Eds. Borém, A., Lopes, M. T. G., Clement, C. R. C., Noda, H. (Viçosa: Suprema Editora Ltda).

Gonçalves, P. S., Martins, A. L. M., Bortoletto, N., Tanzini, M. R. (1996). Estimates of genetic parameters and correlations of juvenile characters basead on open pollinated progenies of Hevea. Rev. Bras. Genet. 19, 105–111.

Gonçalves, P. S., Rossetti, A. G., Valois, A. C. C., Viegas, I. J. (1984). Genetic and phenotypic correlations between some quantitative traits in juvenile clonal rubber trees (Hevea spp.). Rev. Bras. Genet. II, 95–107.

Gonçalves, P. S., Silva, M. A., Gouvêa, L. R. L., Scaloppi-Junior, E. J. (2006). Genetic variability for girth growth and rubber yield traits in Hevea brasiliensis. Sci. Agric. 63, 246–254. doi: 10.1590/S0103-90162006000300006

Gorjanc, G., Cleveland, M. A., Houston, R. D., Hickey, J. M. (2015). Potential of genotyping-by-sequencing for genomic selection in livestock populations. Genet. Sel. Evol. 47, 12. doi: 10.1186/s12711-015-0102-z

González-Camacho, J. M., de los Campos, G., Pérez, P., GianolaJ, D., Cairns, E., Mahuku, G., et al. (2012). Genome-enabled prediction of genetic values using radial basis function neural networks. Theor. Appl. Genet. 125, 759–771. doi: 10.1007/s00122-012-1868-9

Granato, I., Cuevas, J., Luna-Vazquez, F., Crossa, J., Montesinos-Lopez, O., Burgueno, J., et al. (2018). BGGE: a new package for genomic-enabled prediction incorporating genotype x environment interaction models. G3 (Bethesda) 8, 3039–3047. doi: 10.1534/g3.118.200435

Granato, I., Fritsche-Neto, R. (2018). snpReady: Preparing genotypic datasets in order to run genomic; analysis. R package version 0.9.6. Available: https://CRAN.R-project.org/package=snpReady.

Grattapaglia, D. (2017). Status and Perspectives of Genomic Selection in Forest Tree Breeding. Springer International. doi: 10.1007/978-3-319-63170-7_9

Habier, D., Fernando, R. L., Dekkers, J. C. (2007). The impact of genetic relationship information on genome-assisted breeding values. Genetics 177, 2389–2397. doi: 10.1534/genetics.107.081190

Heffner, E. L., Sorrells, M. E., Jannink, J. L. (2009). Genomic selection for crop improvement. Crop Sci. 49, 1–12. doi: 10.2135/cropsci2008.08.0512

Heslot, N., Akdemir, D., Sorrells, M. E., Jannink, J. L. (2014). Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. Theor. Appl. Genet. 127, 463–480. doi: 10.1007/s00122-013-2231-5

Heslot, N., Yang, H.-P., Sorrells, M. E., Jannink, J.-L. (2012). Genomic Selection in plant breeding: a comparison of models. Crop Science 52, 146. doi: 10.2135/cropsci2011.06.0297

Isik, F. (2014). Genomic selection in forest tree breeding: the concept and an outlook to the future. New For. 45, 379–401. doi: 10.1007/s11056-014-9422-z

Isik, F., Bartholome, J., Farjat, A., Chancerel, E., Raffin, A., Sanchez, L., et al. (2016). Genomic selection in maritime pine. Plant Sci. 242, 108–119. doi: 10.1016/j.plantsci.2015.08.006

Iwata, H., Hayashi, T., Tsumura, Y. (2011). Prospects for genomic selection in conifer breeding: a simulation study of Cryptomeria japonica. Tree Genet. Genomes 7, 747–758. doi: 10.1007/s11295-011-0371-9

Jarquin, D., Crossa, J., Lacaze, X., Du Cheyron, P., Daucourt, J., Lorgeou, J., et al. (2014). A reaction norm model for genomic selection using high-dimensional genomic and environmental data. Theor. Appl. Genet. 127, 595–607. doi: 10.1007/s00122-013-2243-1

Krchov, L.-M., Bernardo, R. (2015). Relative efficiency of genomewide selection for testcross performance of doubled haploid lines in a maize breeding program. Crop Sci. 55, 2091–2099. doi: 10.2135/cropsci2015.01.0064

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

Li, Y., Sidore, C., Kang, H. M., Boehnke, M., Abecasis, G. R. (2011). Low-coverage sequencing: Implications for design of complex trait association studies. Genome Res. 21, 940–951. doi: 10.1101/gr.117259.110

Lima, B. M. (2014). Bridging genomics and quantitative genetics of Eucalyptus: genome-wide prediction and genetic parameter estimation for growth and wood properties using high-density SNP data.[thesis]. Piracicaba, Brazil: University of São Paulo.

Lopez-Cruz, M., Crossa, J., Bonnett, D., Dreisigacker, S., Poland, J., Jannink, J. L., et al. (2015). Increased prediction accuracy in wheat breeding trials using a marker x environment interaction genomic selection model. G3 (Bethesda) 5, 569–582. doi: 10.1534/g3.114.016097

Lorenz, A. J., Chao, S., Asoro, F. G., Heffner, E. L., Hayashi, T., Iwata, H., et al (2011). “Chapter Two - Genomic selection in plant breeding: knowledge and prospects,” in Advances in Agronomy. Ed. Sparks, D. L. (San Diego: Academic Press), 77–123. doi: 10.1016/B978-0-12-385531-2.00002-5

Matias, F. I., Alves, F. C., Meireles, K. G. X., Barrios, S. C. L., Vale, C. B., Endelman, J. B., et al. (2019). “On the The accuracy of genomic prediction models considering multi-trait and allele dosage in Urochloa spp,” in interspecific tetraploid hybrids (Netherlands: Springer). doi: 10.1007/s11032-019-1002-7

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.

Moreti, D., Gonçalves, P. S., Gorgulho, E. P., Martins, A. L. M., Bortoletto, N. (1994). Estimativas de parâmetros genéticos e ganhos esperados com a seleção de caracteres juvenis em progênies de seringueira. Pesquisa Agropecuária Brasileira, Brasília, DF v.7, n. 29, 1099–1109.

Morota, G., Gianola, D. (2014). Kernel-based whole-genome prediction of complex traits: a review. Front. Genet. 5, 363. doi: 10.3389/fgene.2014.00363

Muller, B. S. F., Neves, L. G., De Almeida Filho, J. E., Resende, M. F. R., Jr., Munoz, P. R., Dos Santos, P. E. T., et al. (2017). Genomic prediction in contrast to a genome-wide association study in explaining heritable variation of complex growth traits in breeding populations of Eucalyptus. BMC Genomics 18, 524. doi: 10.1186/s12864-017-3920-2

Munõz, F., Sanchez, L. (2017). breedR: statistical methods for forest genetic; resources analysts. R package version 0.12-2. Available: https://github.com/famuvie/breedR.

Oakey, H., Cullis, B., Thompson, R., Comadran, J., Halpin, C., Waugh, R. (2016). Genomic selection in multienvironment crop trials. G3 (Bethesda) 6, 1313–1326. doi: 10.1534/g3.116.027524

Pérez-Elizalde, S., Cuevas, J., Pérez-Rodríguez, P., Crossa, J. (2015). Selection of the bandwidth parameter in a bayesian kernel regression model for genomic-enabled prediction. J. Agric. Biol. Environ. Stat. 20 (4), 512–532. doi: 10.1007/s13253-015-0229-y

Perez-Rodriguez, P., Gianola, D., Gonzalez-Camacho, J. M., Manés, J. C., Dreisigacker, S. (2013). Comparison between linear and non-parametric regression models for genome-enabled prediction in wheat. G3 Genes Genomes Genet. 2, 1595–1605. doi: 10.1534/g3.112.003665

Pérez-Rodríguez, P., Crossa, J., Bondalapati, K., De Meyer, G., Pita, F., Campos, G. D. L. (2015). A pedigree-based reaction norm model for prediction of cotton yield in multienvironment trials. Crop Sci. 55, 1143–1151. doi: 10.2135/cropsci2014.08.0577

Priyadarshan, P. M. (2017). Refinements to Hevea rubber breeding. Tree Genet. Genomes 13, 20. doi: 10.1007/s11295-017-1101-8

Ratcliffe, B., El-Dien, O. G., Klápště, J., Porth, I., Chen, C., Jaquish, B., et al. (2015). A comparison of genomic selection models across time in interior spruce (Picea engelmannii × glauca) using unordered SNP imputation methods. Heredity (Edinb.) 115, 547–555. doi: Rao, G. P., Kole, P. C. (2016). Evaluation of Brazilian wild Hevea germplasm for cold tolerance: genetic variability in the early mature growth. J. For. Res. 27, 755–765. doi: 10.1007/s11676-015-0188-8

Resende, M. D. V., de Assis, T. F. (2008). Seleção recorrente recíproca entre populações sintéticas multi-espécies (SRR-PSME) de eucalipto. Pesquisa Florestal Brasileira 57, 57–60

Resende, M. D., Resende, M. F., Jr., Sansaloni, C. P., Petroli, C. D., Missiaggia, A. A., Aguiar, A. M., et al. (2012a). Genomic selection for growth and wood quality in Eucalyptus: capturing the missing heritability and accelerating breeding for complex traits in forest trees. New. Phytol. 194, 116–128. doi: 10.1111/j.1469-8137.2011.04038.x

Resende, M. F., Jr., Munoz, P., Acosta, J. J., Peter, G. F., Davis, J. M., Grattapaglia, D., et al. (2012b). Accelerating the domestication of trees using genomic selection: accuracy of prediction models across ages and environments. New. Phytol. 193, 617–624. doi: 10.1111/j.1469-8137.2011.03895.x

Romain, B., Thierry, C. (2011). RUBBERCLONES (Hevea Clonal Descriptions). Available at: http://rubberclones.cirad.fr.

Rosa, J. R. B. F., Mantello, C. C., Garcia, D., Souza, L. M., Silva, C. C., Gazaffi, R., et al. (2018). QTL detection for growth and latex production in a full-sib rubber tree population cultivated under suboptimal climate conditions. BMC Plant Biol. 18, 223. doi: 10.1186/s12870-018-1450-y

Shearman, J. R., Sangsrakru, D., Ruang-Areerate, P., Sonthirod, C., Uthaipaisanwong, P., Yoocha, T., et al. (2014). Assembly and analysis of a male sterile rubber tree mitochondrial genome reveals DNA rearrangement events and a novel transcript. BMC Plant Biol. 14, 45. doi: 10.1186/1471-2229-14-45

Sivakumaran, S., Haridas, G., Abraham, P. D. (1988). Problem of tree dryness with high yielding precocious clones and methods to exploit such clones. Proc. Coll. Hevea 88, IRRDB, Paris 1988, 253–267.

Souza, L. M., Gazaffi, R., Mantello, C. C., Silva, C. C., Garcia, D., Le Guen, V., et al. (2013). QTL mapping of growth-related traits in a full-sib family of rubber treetrees (Hevea brasiliensis) evaluated in a sub-tropical climate. PLoS One 8, e61238. doi: 10.1371/journal.pone.0061238

Tan, B., Grattapaglia, D., Martins, G. S., Ferreira, K. Z., Sundberg, B., Ingvarsson, P. K. (2017). Evaluating the accuracy of genomic prediction of growth and wood traits in two Eucalyptus species and their F1 hybrids. BMC Plant Biol. 17, 110. doi: 10.1186/s12870-017-1059-6

Tang, C., Yang, M., Fang, Y., Luo, Y., Gao, S., Xiao, X., et al. (2016). The rubber tree genome reveals new insights into rubber production and species adaptation. Nat. Plants 2, 16073. doi: 10.1038/nplants.2016.73

Toro, M. A., Varona, L. (2010). A note on mate allocation for dominance handling in genomic selection. Genet. Sel. Evol. 42, 1–9. doi: 10.1186/1297-9686-42-33

VanRaden, P. M. (2007). Efficient estimation of breeding values from dense genomic data. J. Dairy Sci. 90 (suppl 1), 374–375.

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. J. Dairy Sci. 91, 4414–4423. doi: 10.3168/jds.2007-0980

Wong, C. K., Bernardo, R. (2008). Genomewide selection in oil palm: increasing selection gain per unit time and cost with small populations. Theor. Appl. Genet. 116, 815–824. doi: 10.1007/s00122-008-0715-5

Zhang, Z., Erbe, M., He, J., Ober, U., Gao, N., Zhang, Z., et al. (2015). Accuracy of whole genome prediction using a genetic architecture enhanced variance–covariance matrix. G3 Genes Genomes Genet. 5, 615–627. doi: 10.1534/g3.114.016261

Zapata-Valenzuela, J., Whetten, R. W., Neale, D., Mckeand, S., Isik, F. (2013). Genomic estimated breeding values using genomic relationship matrices in a cloned population of loblolly pine. G3 (Bethesda) 3, 909–916. doi: 10.1534/g3.113.005975

Capítulo II

Unravelling rubber tree growth by integrating GWAS and biological network-based approaches.

Felipe Roberto Francisco,† Alexandre Hild Aono,† Carla Cristina da Silva, Paulo S. Gonçalves, Erivaldo J. Scaloppi Junior, Vincent Le Guen, Roberto Fritsche-Neto, Livia Moura Souza, and Anete Pereira de Souza*

Front. Plant Sci., **21 December 2021**
Sec. Plant Breeding
https://doi.org/10.3389/fpls.2021.768589

Keywords: GBS, GWAS, *Hevea brasiliensis*, linkage disequilibrium, metabolic networks, QTL, RNA-Seq, WGCNA

Abstract

Hevea brasiliensis (rubber tree) is a large tree species of the Euphorbiaceae family with inestimable economic importance. Rubber tree breeding programs currently aim to improve growth and production, and the use of early genotype selection technologies can accelerate such processes, mainly with the incorporation of genomic tools, such as marker-assisted selection (MAS). However, few quantitative trait loci (QTLs) have been used successfully in MAS for complex characteristics. Recent research shows the efficiency of genome-wide association studies (GWAS) for locating QTL regions in different populations. In this way, the integration of GWAS, RNA-sequencing (RNA-Seq) methodologies, coexpression networks and enzyme networks can provide a better understanding of the molecular relationships involved in the definition of the phenotypes of interest, supplying research support for the development of appropriate genomic based strategies for breeding. In this context, this work presents the potential of using combined multiomics to decipher the mechanisms of genotype and phenotype associations involved in the growth of rubber trees. Using GWAS from a genotyping-by-sequencing (GBS) *Hevea* population, we were able to identify molecular markers in QTL regions with a main effect on rubber tree plant growth under constant water stress. The underlying genes were evaluated and incorporated into a gene coexpression network modelled with an assembled RNA-Seq-based transcriptome of the species, where novel gene relationships were estimated and evaluated through in silico methodologies, including an estimated enzymatic network. From all these analyses, we were able to estimate not only the main genes involved in defining the phenotype but also the interactions between a core of genes related to rubber tree growth at the transcriptional and translational levels. This work was the first to integrate multiomics analysis into the in-depth investigation of rubber tree plant growth, producing useful data for future genetic studies in the species and enhancing the efficiency of the species improvement programs.

Introduction

Hevea brasiliensis (rubber tree) is an outbreeding forest species belonging to the Euphorbiaceae family with an inestimable importance in the world economy because it is the only crop capable of producing natural rubber with quantity and quality levels able to meet global demand (De Faÿ and Jacob, 1989). Possessing unique characteristics such as resistance, elasticity and heat dissipation, *Hevea* rubber is used as a feedstock for more than 40,000 products (Pootakham et al., 2017; Mantello et al., 2019). Although it is very important, *H. brasiliensis* is still in an early domestication stage due to its long breeding cycle (25–30 years), the large areas required for planting and its recent cultivation (Priyadarshan and Clément-Demange, 2004; Gonçalves et al., 2006). In this context, *Hevea* breeding programs aim to improve important agronomic traits for rubber fabrication, mainly those related to latex growth and production (Priyadarshan, 2003). The use of early genotype selection technologies has been proposed as a breeding alternative for accelerating this process, e.g., incorporating genomic tools for marker-assisted selection (MAS; Pootakham et al., 2017; Priyadarshan, 2017). Although the discovery of quantitative trait loci (QTLs) can benefit *Hevea* breeding programs (Souza et al., 2019), this characterization is hindered by the large number of genes and molecular interactions controlling such characteristics (Pootakham et al., 2020). To date, few QTLs have been successfully used for rubber tree MAS for complex quantitative traits due to the insufficient quantity of linked markers in the QTLs, small QTL effects on the phenotype, or strong environmental influences (Nguyen et al., 2019).

Several studies have been carried out in the last decade to identify QTLs in *H. brasiliensis* through genetic linkage maps (Souza et al., 2013; Pootakham et al., 2015; Conson et al., 2018; Rosa et al., 2018; Xia et al., 2018) and association mapping (Chanroj et al., 2017). Genome-wide association studies (GWAS) are important tools for the identification of candidate genetic variants underlying QTLs, with great potential to be incorporated into MAS. Compared to linkage maps, the use of GWAS methodologies has advantages such as using genetically diverse populations with different rates of recombination and linkage disequilibrium (LD; Myles et al., 2009). Despite the observed GWAS efficiency in several crops (Warraich et al., 2020; Zhang et al., 2020; Verzegnazzi et al., 2021), this methodology still presents limitations related to the low proportion of phenotypic variance explained by the identified genomic regions (Manolio et al., 2009). As an alternative, the combination of GWAS results with other molecular methodologies, such as

transcriptomics and proteomics analyses, can contribute to better knowledge of the genetic mechanisms involved in the definition of a trait (Tam et al., 2019), overcoming the statistical limitations on the characterization of a broad set of causal genomic regions.

Although the identification of genes with a great phenotypic effect is consolidated with GWAS methodologies (Nebel et al., 2011), there are no established methods for investigating the complete set of genes controlling complex traits through multiomics approaches, and such characterization is an open scientific challenge, especially in crops with complex genomes such as rubber trees (Schaefer et al., 2018). Different initiatives have associated GWAS results with RNA-Seq data (Yan et al., 2020), linking causal genes relevant to the observed phenotypic variation with cell transcription activity profiles (Schaefer et al., 2018; Nguyen et al., 2019). In *Hevea*, however, RNA-Seq-based studies have been mainly performed to investigate differentially expressed genes (DEGs) under different environmental or stress conditions and profiling rubber tree samples (Hurtado Páez et al., 2015; Sathik et al., 2018; Mantello et al., 2019; Ding et al., 2020). Although the integration of GWAS with RNA-Seq methodologies has proven to provide a deeper comprehension of the genetic relationships involved in trait definition, there is no study, to date, aggregating such data in *Hevea*.

We are currently undergoing a major revolution in omics sciences (genomics, transcriptomics, proteomics, and phenomics) with different methods for data integration enabling important advances in all phases of genetic improvement, ranging from the discovery of new variants to the understanding of important metabolic pathways (Scossa et al., 2021). The integration of data derived from multiomics can be combined to reveal, in a profound way, the relationships that represent the true biological meaning of the studied elements (Jamil et al., 2020; Wu et al., 2020b). This approach has become increasingly common in humans (Wu et al., 2018), animals (Fonseca et al., 2018), microorganisms (Wang et al., 2019), and combinations of species (Pinu et al., 2019). However, for plants, such integrated methodologies are still a great challenge, especially for nonmodel species with elevated genetic diversity and complex genomes (Jamil et al., 2020), which is the case for *H. brasiliensis* (Tang et al., 2016; Liu et al., 2020c, Wu et al., 2020a). Despite its economic importance, no study incorporating multiomics has been carried out on *H. brasiliensis*. With the wide availability of omics data, coexpression networks have become a tool with great potential for inferring gene interactions, mainly based on regulatory and structural relationships, allowing for a broader understanding of unknown molecular mechanisms (Rao and Dixon, 2019). The identification of these genes also allows us to indirectly assess,

through their enzymes, the global metabolic relationships involved in defining the evaluated characteristic (Pérez-Bercoff et al., 2011). In this way, we can make use of GWAS to select genes of great importance for the phenotype of interest. Such genes can be used as a guide to select modules of coexpressed genes and their enzymes, which may have minor effects on the phenotype but may be important to maintaining heritability.

In this context, this work presents for the first time a combination of omics data to determine the molecular mechanisms involved in rubber tree growth. For this task, we used a breeding population to infer QTLs using a GWAS approach. These results were incorporated into network analyzes based on RNA-Seq and enzymatic networks. By using this multiomics framework, our study supplies important cues on the interconnection of the metabolic mechanisms of rubber tree growth, providing novel growth-associated genes for future research on increasing *Hevea* production.

Materials and Methods

According to the analysis workflow performed, different molecular layers were investigated in this work (Figure 1). The study started with the identification of the SNPs with the greatest effect on stem diameter (SD) through a GWAS. After selecting these markers, the markers that presented a significant correlation were selected. This entire set of markers was annotated using a transcriptome assembled on the basis of two commercial genotypes that have been widely used in the genetic improvement of the species. Additionally, a weighted gene coexpression network was constructed, from which it is possible to select the functional modules containing the genes identified by the GWAS. An enzymatic network was also built based on the annotation of genes present in the functional modules selected, which supplied insights into the interaction of these enzymes with the studied phenotype.

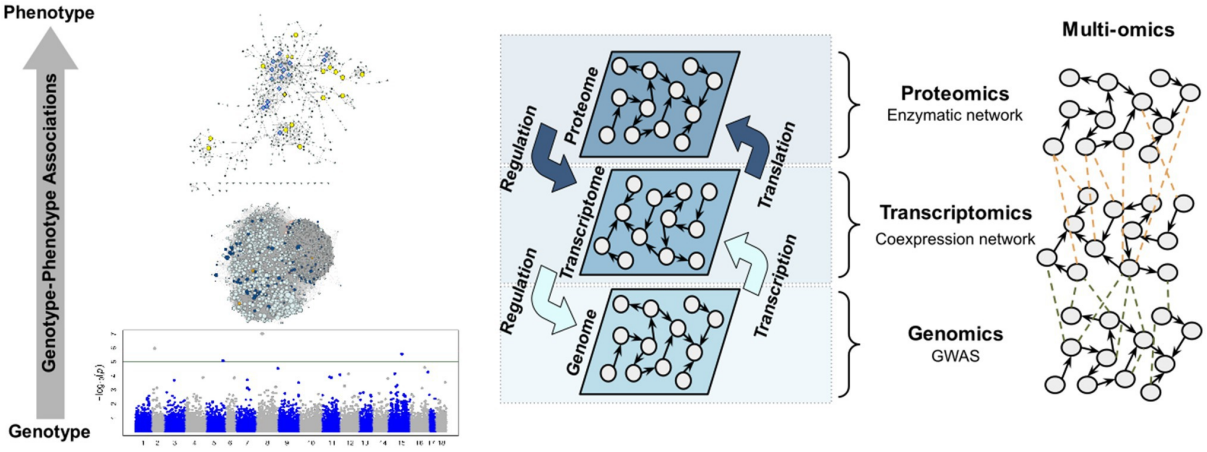

FIGURE 1. Workflow summarizing the main analyses performed.

Plant Material

For this work, we employed a population composed of four test clones (GT1, PB235, RRIM701, and RRIM600) and individuals from crosses between PR255 x PB217 (251 samples), GT1 x RRIM701 (143 samples) and GT1 x PB235 (40 samples; Souza et al., 2013, 2019; Conson et al., 2018; Rosa et al., 2018). The PR255 genotype was selected because of its early growth and high yield, as well as for being vigorous with stable latex production throughout life (Souza et al., 2013). In contrast, the PB217 genotype presents slow growth but has a rapid increase in latex production in its early years and great potential for long-term performance and yield (Souza et al., 2013; Rosa et al., 2018). These genotypes were planted in random blocks, with four replications of the same genotype grafted on the same plot. This plantation is located in Itiquira, Mato Grosso (MT), Brazil (17° 24′03″ S and 54° 44′53″ W). The GT1 genotype was selected because it is a sterile male and is classified as a primary clone that is tolerant to wind and cold (Shearman et al., 2014). The RRIM701 clone shows vigorous growth and a SD increase after the initial cut (Romain and Thierry, 2011). PB235 has been shown to be a high-yield genotype but is susceptible to panel dryness (Sivakumaran et al., 1988). These two populations (GT1 x RRIM 701 and GT1 x PB 235) were planted in an augmented block design that was repeated in four blocks containing two plants of the same genotype per plot with 4 meters of spacing between them. These populations were planted at the Center for Rubber and Agroforestry Systems/Instituto Agronômico (IAC; 20° 25′00″ S and 49° 59′00″ W) in the northwest region of the state of São Paulo (SP), Brazil. All of these genotypes are widely employed in commercial production and used in Brazilian breeding programs, representing the main rubber tree genetic sources in Latin America.

Crossing was carried out via open pollination, and paternity was confirmed using microsatellite markers (SSRs; Souza et al., 2013, Conson et al., 2018).

Phenotypic Analyses

As the main characteristic evaluated in rubber tree genetic breeding (Rao and Kole, 2016), SD was measured in the selected population during the first 4 years of genotype development. Each plant was individually phenotyped (in centimeters) at a height of 50 cm from the soil in two seasons with contrasting average rainfall (low precipitation and high precipitation), which are considered in *Hevea* studies as contrasting environments (Chanroj et al., 2017; Souza et al., 2019). The variance caused by the genotypic effects was estimated using the best linear unbiased predictor (BLUP) with the breedR package in R (Munõz and Sanchez, 2017). The linear mixed model was as follows:

$$y = \mu + X_{Bb} + X_{Rr} + X_{Ww} + X_{Mm} + Z_{Gg} + Z_{gw} + \sigma^2_{\varepsilon},$$

where y is the vector of the phenotypic measures; μ is the trait mean; and X_B, X_R, X_W, and X_M are the incidence matrices for the fixed effects of blocks (*b*), replicates (*r*), water levels (*w*) and month of the measurement (*m*), respectively. Z_G and Z are the incidence matrices of random effects for genotypic effects (*g*) and genotype x environment interactions (*gw*), respectively, and σ^2_{ε} is the residual variance. The significance of random effects was estimated by a likelihood ratio test (LRT) with a significance level of 0.05. We estimated the broad heritability (H^2) for genotypic means using the following equation:

$$H^2 = \frac{\sigma^2_g}{\sigma^2_g + \frac{\sigma^2_{gw}}{s} + \frac{\sigma^2_{\varepsilon}}{a}}$$

where σ^2_g is the genotypic variance, σ^2_{gw} is the variance caused by the environment x genotype interaction, *s* is the number of environments analyzed, σ^2_{ε} is the residual variance and a is the number of blocks.

Genotypic Analyses

The extraction of genomic DNA was performed according to Souza et al. (2013) and Conson et al. (2018). Genotyping-by-sequencing (GBS) libraries were prepared from genomic DNA using the method proposed by Elshire et al. (2011). Initially, the genomic DNA of each sample was digested using the methylation-sensitive enzyme EcoT22I to reduce the genomic

complexity. The resulting fragments of each sample were linked to specific barcodes and combined in pools. These fragments were amplified by PCR and sequenced. Sequencing of the PR255 x RRIM217 population was performed using the Illumina HiSeq platform, and sequencing of the GT1 x RRIM701 and GT1 x PB235 populations was performed with the GAIIx platform (Illumina Inc., San Diego, CA, United States). Processing of the GBS data from both experiments was carried out at the same time. SNPs were identified with TASSEL GBS 5 software (Glaubitz et al., 2014) using the following parameters: (i) k-mer size of 64 bp; (ii) minimum read quality (Q) score of 20; and (iii) minimum locus depth of six reads. Reads were aligned with the rubber tree reference genome proposed by Liu et al. (2020a) using Bowtie2 version 2.1 software (Langmead and Salzberg, 2012) with the very sensitive option. We only kept the biallelic markers selected with the VCFtools program (Danecek et al., 2011). Using snpReady software (Granato and Fritsche-Neto, 2018), SNPs with more than 20% missing data and minimum allele frequency (MAF) < 0.05 were filtered out. Imputation was performed using the k-nearest neighbor imputation (kNNI) algorithm (Hastie et al., 2001). LD estimations were calculated with the ldsep R package (Gerard, 2020) based on the squared Pearson correlation (R^2). For linkage decay investigation, we created a scatter plot of R^2 against the chromosomal distances, considering an exponential decay (Tenesa et al., 2004) created with a nonlinear least squares regression model using R software, calculated according to the following equation.

$$y = a + be^{(-cx)}$$

where y is the linkage disequilibrium, x is the physical distance in bp, $a+b$ is the mean level of disequilibrium for loci at the same location, and e is the exponential term.

Genome-Wide Association Studies

GWAS were performed using the Fixed and random model Circulating Probability Unification (FarmCPU) method implemented in the FarmCPU R package (Liu et al., 2016). This method tests the association of markers as fixed and random effects in a mixed linear model in separate steps (Liu et al., 2016). The kinship matrix and the first two principal components (PC1 and PC2) from a principal component analysis (PCA) were used as covariables in the mixed linear model to control the effects caused by the population structure (Challa and Neelapu, 2018). The significance threshold used for the association mapping was calculated based on 30 SD permutations and a 95% quantile value. Additionally, we expanded the set of putatively associated markers through LD. Considering a minimum R^2 of

0.7, we created a set of GWAS LD-associated markers (snpsLD), which was used for modeling an LD network with the igraph R package (Csardi and Nepusz, 2006).

Transcriptome

To estimate rubber tree gene expression, RNA-Seq data from RRIM600 and GT1 clones (Mantello et al., 2019) were used. From 6 months of age, these plants were transferred to a growth chamber at a temperature of 28°C with a 12-h photoperiod and were irrigated every 2 days for a period of 10 days. After this period, the plants were subjected to cold stress by changing the chamber temperature to 10°C for 24 h, with the leaf tissues being sampled at 0 h (control), 90 min, 12 and 24 h after exposure to the stress. RNA was extracted from the leaves of three biological replicates using the lithium chloride protocol (Dusotoit-Coucaud et al., 2009). From the total RNA, a cDNA library was built using the TruSeq RNA Sample Preparation Kit (Illumina Inc., San Diego, CA, United States). The 24 samples (three replicates per sample at each time) were randomly pooled (four samples per pool) and grouped using the TruSeq Paired-End Reads Cluster Kit on the cBot platform (Illumina Inc., San Diego, CA, United States). The cDNA libraries were posteriorly sequenced on the Illumina Genome platform Analyzer IIx with a TruSeq kit with 36 cycles (Illumina, San Diego, CA, United States) for 72 bp paired-end reads.

RNA-Seq barcodes were removed from FastQ files using Fastx-Tookit[1], and raw reads were filtered using the program NGS QC Toolkit 2.3 (Trivedi et al., 2014), keeping only sequences with a minimum Q-score of 20 across at least 70% of the sequence length. The filtered sequences were combined with bark reads (Mantello et al., 2014) and mapped to the reference genome of *H. brasiliensis* (Tang et al., 2016) using the HISAT2 aligner (Kim et al., 2015). The alignment was ordered and assembled using SAMtools (Li et al., 2009) and Trinity (Grabherr et al., 2011) software, respectively. *H. brasiliensis* scaffolds (Tang et al., 2016) were submitted for ab initio annotation using the Maker-P (Campbell et al., 2014) tool. The Trinity assembled transcripts and the Maker-P annotations were combined with nonredundant *H. brasiliensis* ESTs in the NCBI database (August 2016) and used as a database for aligning assemblies against the *H. brasiliensis* genome (Tang et al., 2016) with the PASA v2.0 pipeline (Haas et al., 2003) after removing redundant alternate splicing data. The obtained transcripts were filtered with a minimum size of 500 bp and evidence of transcription; we excluded sequences that were only predicted by ab initio genome annotation and with high identity for nonplant transcripts. To estimate the physical position of these sequences across *Hevea* chromosomes, we performed comparative alignments of these

transcripts against the *H. brasiliensis* genome proposed by Liu et al. (2020a) using BLASTn (Johnson et al., 2008). The annotation of these transcripts was performed using the Trinotate v3.2.1 program (Haas, 2015) and SwissProt database (downloaded in February 2021; Boeckmann et al., 2003).

Gene-Associated Markers

The analysis of candidate genes in QTL regions was performed based on transcript annotations. Candidate genes for the phenotypic variation of GWAS-discovered SNPs were considered by using the first transcripts positioned in the upstream and downstream regions of these markers. In addition to the SNPs significantly associated with the phenotype discovered by the GWAS, which we will call snpsGWAS here, we also searched for candidate genes in the neighboring snpsLD. The GO terms associated with these annotations (snpsGWAS and snpsLD) were investigated using REVIGO (Supek et al., 2011). The genomic regions of the phenotypically associated SNPs discovered in this work were compared with the QTLs discovered by Conson et al. (2018) from the mapping population GT1 x RRIM701. For this analysis, the sequences underlying the QTLs (Conson et al., 2018) were aligned to the reference genome of Liu et al. (2020a) using BLASTn. Alignments with identity above 90% and with the largest coverage area were selected (minimum e-value of e-10). Based on the position of this alignment in relation to the reference genome of Liu et al. (2020a), a representation of the 18 chromosomes of *H. brasiliensis* was made using snpsGWAS, snpsLD and QTLs (Conson et al., 2018) using the MapChart program v.2.2 (Voorrips, 2002).

Coexpression Networks

For modeling coexpression networks, we used RNA-Seq count data grouped into transcript clusters through PASA v2.0 software (Haas et al., 2003). Only transcripts with at least 10 counts per million (CPM) were retained and normalized with a quantile-based approach implemented in the edgeR package in R (Robinson et al., 2010). Weighted gene correlation analysis (WGCNA) was performed using the WGCNA R package (Langfelder and Horvath, 2008) together with Pearson correlation coefficients. A soft thresholding power β-value was estimated for fitting the network into a scale-free topology, and a topological overlap measure (TOM) for each gene pair was used for building a dissimilarity matrix and for performing unweighted pair group method with arithmetic mean (UPGMA) hierarchical clustering. The best clustering scheme was defined using a variable height pruning technique

implemented in the Dynamic Tree Cut R package (Langfelder et al., 2008). The groups containing genes associated with snpsGWAS were used to model a specific coexpression network using the igraph R package (Csardi and Nepusz, 2006) with Pearson correlation coefficients (minimum R value of 0.5), where we calculated the hub scores for each gene considering Kleinberg's hub centrality scores (Kleinberg, 1999).

Metabolic Network Modeling

From the annotations performed for genes surrounding the snpsGWAS and the snpsLD, we retrieved the enzyme commission (EC) numbers and investigated the related metabolic pathways using the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Kanehisa and Goto, 2000). All the *H. brasiliensis* metabolic pathways with enzymes related to snpsGWAS and snpsLD were retrieved and used to model a metabolic network using BioPython v.1.78 (Cock et al., 2009). From the created network, we evaluate the following topological properties: (i) degree (Barabási and Oltvai, 2004), (ii) betweenness centrality (Brandes, 2001), (iii) stress (Brandes, 2001), (iv) short path length value (Watts and Strogatz, 1998), and (v) neighborhood connectivity (Maslov and Sneppen, 2002), using Cytoscape v3.8.2 (Shannon et al., 2003). The network was also categorized regarding its community structure, with the enzymes organized into modules using the HiDeF algorithm (Zheng et al., 2021).

Results

Phenotypic and Genotypic Analyses

The SD values were adjusted according to the mixed model from which the BLUPs were extracted for further analysis (Supplementary Table 1). All fixed and random effects showed significant effects under the LRT test ($p < 0.01$). The estimated variances were 4.56, 0.0001 and 26.69 for the genotype (σ^2_g), genotype x environment interaction (σ^2_{gw}) and residual (σ^2_ε) effects, respectively. The experimental design was confirmed to show normality of the residual variance based on the quantile-quantile graph (Q-Q plot; Supplementary Figure 1). The estimated heritability (H^2) in the entire population was 0.55, which is close to those values found in previous studies on the species (Gonçalves et al., 1999; Chanroj et al., 2017).

The identification of SNPs was carried out using all 437 individuals. By employing the TASSEL pipeline, we produced 363,641 tags, which were aligned with the *Hevea* reference genome, producing an alignment rate of ~84.78%. We identified a total of 107,466 SNPs, which were filtered, resulting in a total of 30,266 high-quality markers (~28.16%), with an imputation rate of ~6.74%. This filtered SNP dataset was used for PCA, with 18.33 and 2.61% of the variance explained by the first two main components, respectively (Supplementary Figure 2). Although high LD decay was observed (Supplementary Figure 3A), we also assessed the LD decay rate only in the regions containing transposable elements (TEs; Supplementary Figure 3B), which was higher.

RNA-Seq Analyses

A total of ~530 million and ~633 million paired-end (PE) reads were obtained for the RRIM600 and GT1 genotypes, respectively. After quality filtering, we obtained ~933 million PE reads for assembling the transcripts through Trinity software. We identified 104,738 transcripts ranging from 500 to 22,333 bp (average transcript size of 1,874 bp and N50 of 2,369 bp) that were related to 49,304 genes. In total, 82,629 transcripts (78.89%) could be annotated using the Swiss-Prot database. We were able to associate Gene Ontology (GO) categories with 81,095 transcripts (77.42%) and metabolic pathways from the KEGG database with 74,668 transcripts (71.29%). A total of 11,150 different proteins could be associated with the estimated set of genes for rubber trees, with a high incidence of TEs; the retrovirus-related Pol polyprotein from transposon RE1 (RE1) (4.45%) and the retrovirus-related Pol polyprotein from transposon TNT 1-94 (TNT 1-94) (2.80%) were the most pronounced categories.

Genome Wide Association Study

With the FarmCPU method and the selected covariates, we were able to observe satisfactory adherence to the association mapping results (Figure 2A). Four snpsGWAS were identified on chromosomes 2, 5, 8, and 15 (Figure 2B). The MAFs of the snpsGWAS ranged from 10 to 45%, with the proportion of phenotype variance explained (PVE) ranging from 2 to 9% and additive effects ranging from −1 to 0.84 cm (Table 1). To assess all markers associated with SD, we expanded the set of significantly associated markers by means of LD tests on the total set of SNPs. A total of 181 snpsLD were found and showed a correlation greater than 0.7 with the snpsGWAS (Supplementary Figure 4). snpsLD are distributed on the 18 chromosomes of the rubber tree (Figure 3), flanking previously described QTLs (Conson

et al., 2018). We were able to identify SNPs with distances of approximately 40 bp in the QTL regions (Figure 3).

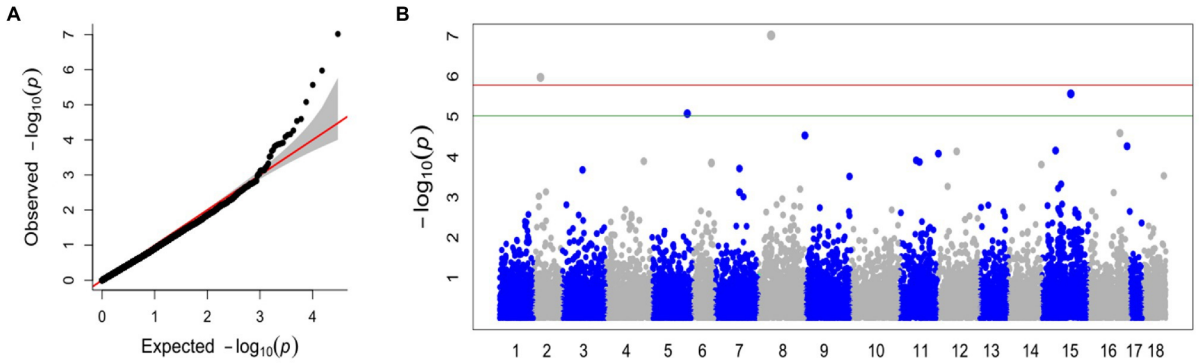

FIGURE 2. (A) Quantile-quantile plot for the broad genomic association model (GWAS), with the inclusion of the first main component (PC1 and PC2) as a covariate. (B) Manhattan plot for the GWAS. The x axis shows the chromosomes containing the discovered markers in their respective positions. The y axis shows the log (value of p) of the association. The green line represents the threshold obtained based on the data, and the red line represents the Bonferroni-corrected threshold of 0.05.

TABLE 1. SNPs identified through the GWAS model.

SNP	Chrom	Position	*p*	MAF	Effect	Va	PVE	Gene
SNP6421	chrom02	14,565,718	1.06E+08	0.10	−1.00	0.18	0.05	SBT4.6
SNP30209	chrom05	75,998,329	8.38E+08	0.17	0.54	0.08	0.02	GEK1
SNP43760	chrom08	26,946,649	9.61E+06	0.45	0.84	0.35	0.09	-
SNP92152	chrom15	50,878,458	2.71E+08	0.29	0.43	0.08	0.02	IQM2

**Chromosome (Chrom), Position on chromosome (Position), value of p for the association (value of p), Frequency of the smallest allele (MAF), Additive effect (Effect), Additive variance (Va), and Proportion of phenotype variance explained (PVE).*

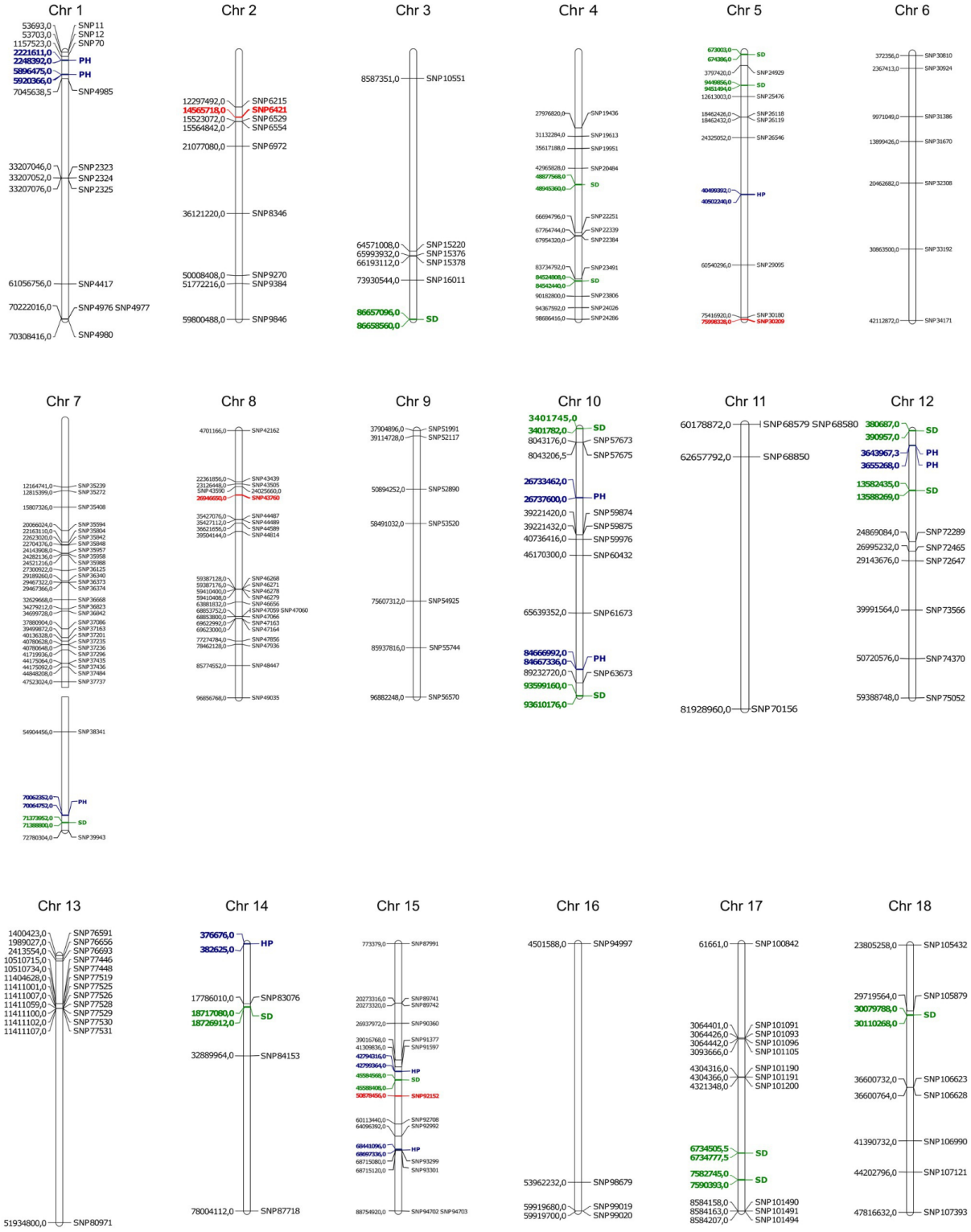

FIGURE 3. Physical position of snpsGWAS in red, snpsLD in black and QTLs discovered by Conson et al. (2018). The QTLs for plant height (PH) are in blue and those for stem diameter (SD) are in green.

To infer the associations between the set of SNPs (snpsGWAS and snpsLD) and expressed genomic regions, we performed comparative alignments of the transcripts assembled to the rubber tree chromosomes. SNPs were assigned to the first genes that were downstream and upstream of their location with an average distance of 7 kbp (Supplementary Table 2). Among the snpsLD, genes related to the transcription of important proteins involved in different stresses were found, such as TNT 1-94, receptor-like protein EIX2, integrin-linked protein kinase 1, U1 small nuclear ribonucleoprotein 70 kDa, histidine-containing phosphotransfer protein 2, rhomboid-like protein 14, and mitochondrial and threonine-protein kinase STN7. The annotation of the set of SNPs putatively associated with SD showed major biological processes related to DNA integration, response to water deprivation, regulation of intracellular pH, proton transmembrane transport, stomatal opening, flavonoid biosynthetic process, pollen sperm cell differentiation, oxidation–reduction process, circadian rhythm, carbohydrate metabolic process, multidimensional cell growth and chromatin organization (Figure 4).

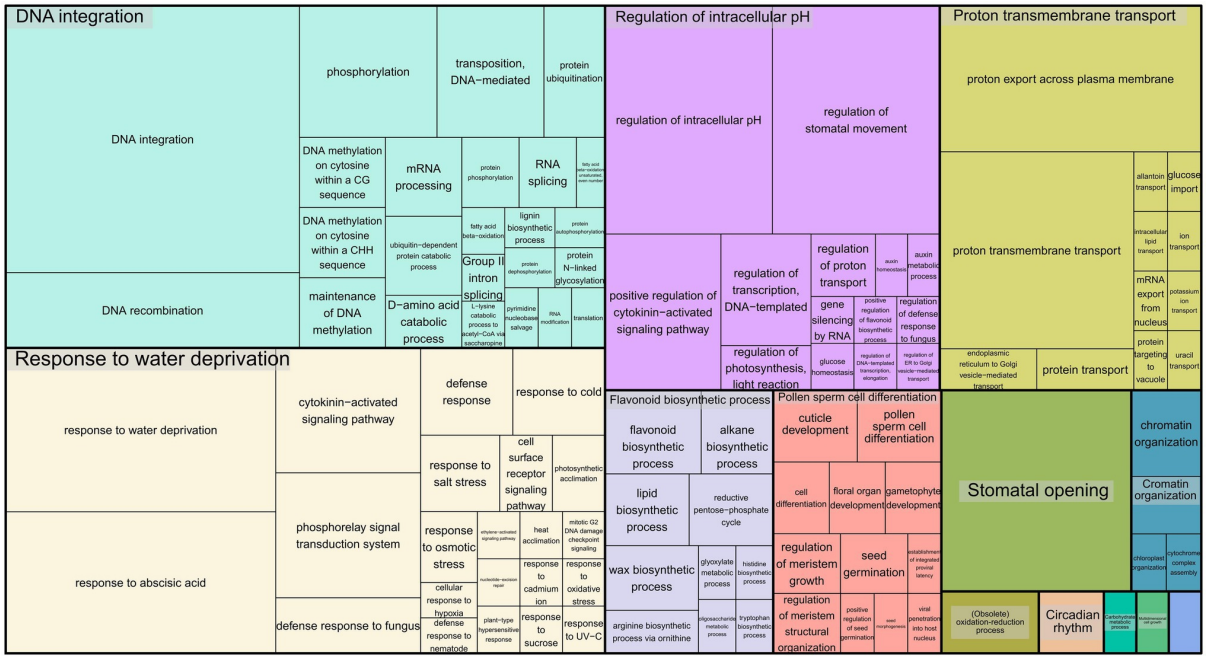

FIGURE 4. Treemap representing the biological processes for the GO terms of the annotated SNPs.

Gene Coexpression Network

Of the 104,738 transcripts, 30,407 were selected for modeling a gene coexpression network using the WGCNA methodology (Zhang and Horvath, 2005). In such a network, pairwise gene interactions are modeled through a similarity measure, such as the Pearson

correlation coefficient employed here. For fitting the network into a scale-free topology, we selected a β power of 9 (scale-free topology model fit with $R^2 > 0.85$ and mean connectivity of ~183.47) and calculated the corresponding dissimilarity matrix through the WGCNA R package. With the network modeled, we combined UPGMA clustering with a variable height pruning technique, enabling the identification of 174 groups, with sizes ranging from 52 to 3,823 genes. The five groups containing the genes potentially related to the snpsGWAS were selected (Supplementary Table 3), and a new coexpression network was built including the genes associated with the snpsLD (Figure 5). All these genes formed a unique interaction network with weaker interactions connecting the found groups, which putatively represents the direct and indirect molecular associations with the SD phenotype. For the analysis of all reactions triggered by the genomic regions associated with GWAS, we evaluated this set of 1,528 genes for related GO terms (Figure 6). From the biological process category, we found new GO terms not associated with the genes related to snpsGWAS and snpsLD. These GO terms included defense response, positive regulation of transcription, cell wall organization, photosynthesis, cell division, mitotic cell cycle phase transition, carbon fixation, cell population proliferation, asymmetric cell division, and stomatal closure.

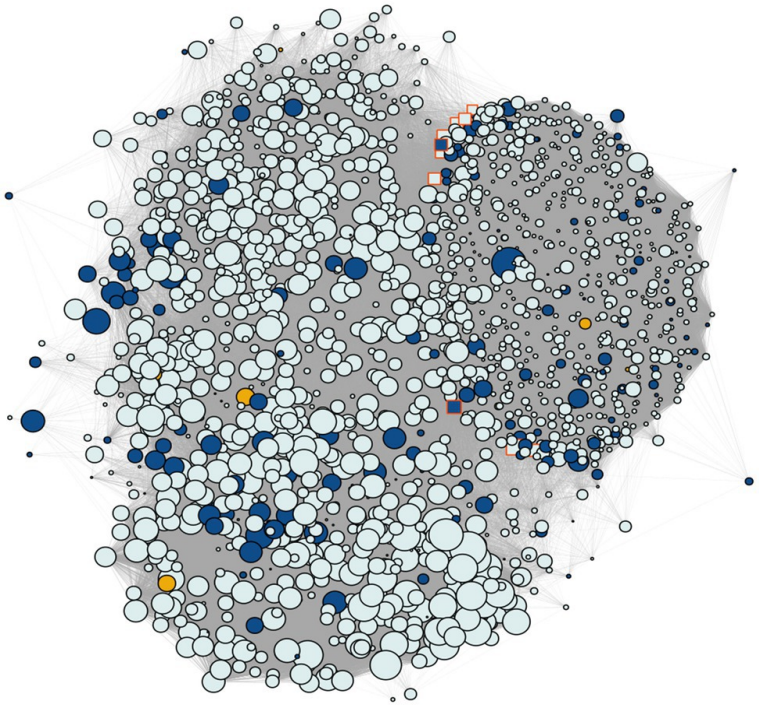

FIGURE 5. Coexpression network containing the SNP gene modules discovered by GWAS. Yellow shows the genes annotated for the snpsGWAS, blue shows the genes annotated for the snpsLD and gray shows the genes identified in the modules. The highlighted genes with a red

border represent the 10 hubs with the most connectivity, while the size of the nodes shows the number of connected genes.

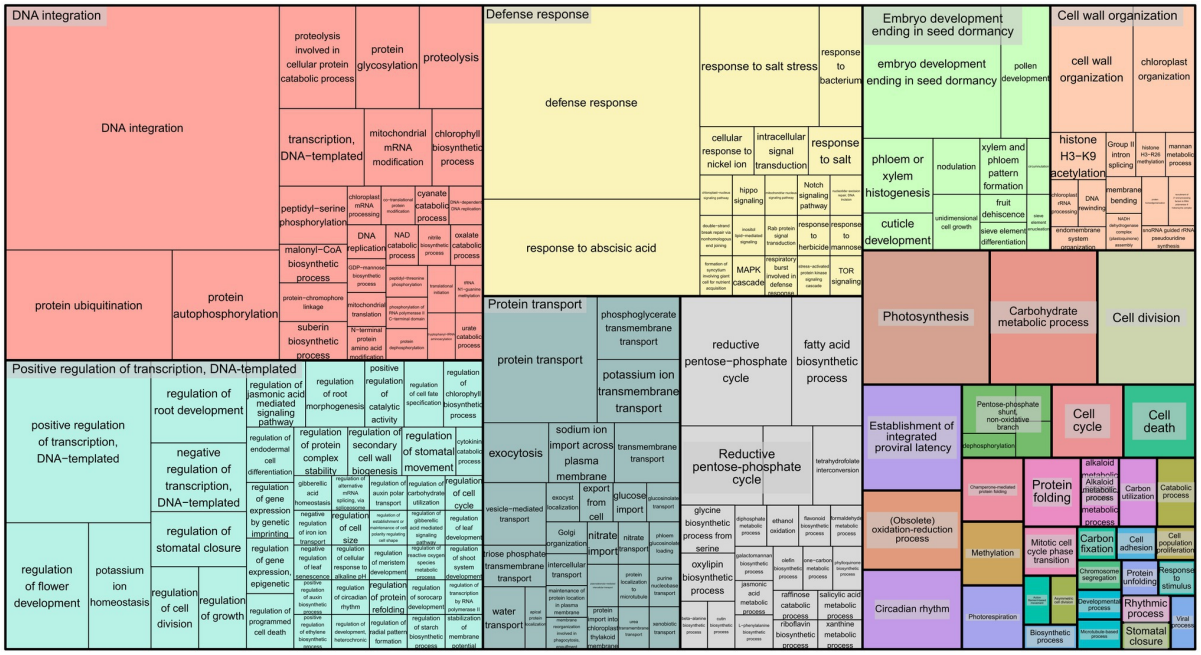

FIGURE 6. Treemap representing the biological processes for the GO terms of the annotated functional modules.

Regarding the genes found in these modules, as also observed in the general transcriptome profile, we observed a predominance of genes related to the protein retrovirus-related Pol polyprotein from transposon 17.6 (TE 17.6) (2.36%) and TNT 1-94 (1.23%). We also found several genes related to proteins involved in (Supplementary Table 3): (i) plant growth (e.g., MEI2-like 4 and threonine-protein kinase GSO1); (ii) the response to biotic and abiotic stress (e.g., abscisic acid-insensitive 5-like protein 6, transcription factor ICE1, abscisic acid receptor PYL4, transcription factor jungbrunnen 1, transcription factor MYB44, and galactinol synthase 2); (iii) root growth (e.g., alkaline/neutral invertase CINV2, threonine protein kinase IREH1, phospholipase D zeta 1, protein arabidillo 1, regulatory-associated protein of TOR 1, agamous-like MADS-box protein AGL12, and omega-hydroxypalmitate O-feruloyl transferase); (iv) the hormone abscisic acid (ABA) pathway; and (v) the light acclimatization process (e.g., GATA transcription factor 7 and malate dehydrogenase [NADP]). However, the great majority of these identified genes were not overexpressed, with a few exceptions (Supplementary Figure 5). To assess the most influential nodes within the network structure, we evaluated the hub scores of each gene within the network. The first hub gene in this network (PASSA_cluster_140395) was among

the snpsLD genes, and the 10 first hubs had many known annotations. The first three hubs that had a known annotation were PASA_cluster_160224, PASA_cluster_87395, and PASA_cluster_140392, showing associations with TEs (Supplementary Table 3).

Metabolic Networks

Due to the clear absence of functional annotations, all the genes identified in the coexpressed modules with a known enzymatic activity relatedness were used for modeling a metabolic network using the KEGG database. In this structure, each enzyme corresponds to a node, and their connections are based on metabolic interactions. Nineteen genes were related to 19 different enzymes present in 28 metabolic pathways (Supplementary Table 4). All these reactions were joined into a unique network structure containing 405 nodes (enzymes) and 1,311 edges (average number of 5.338 neighbors and diameter of 22 nodes; Figure 7A; Supplementary Figure 6), representing a diverse cascade of mechanisms with putative associations with plant growth. Network topology measurements were performed to identify the most important enzymes in the modeled mechanisms.

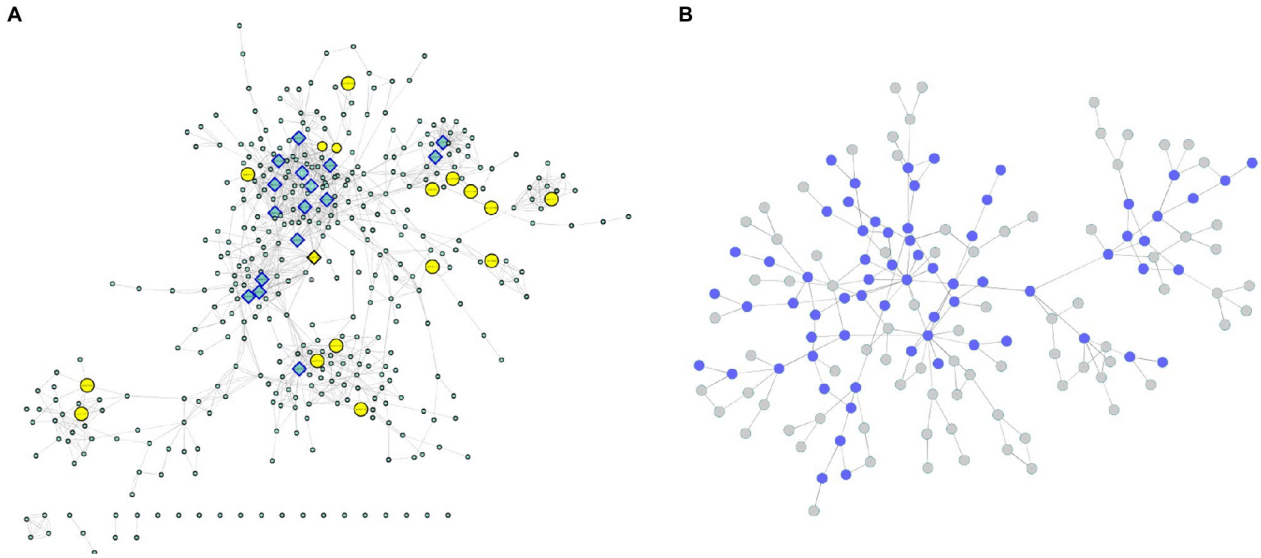

FIGURE 7. (A) Enzyme network. The yellow nodes represent the enzymes discovered in the coexpression modules, and the rectangular nodes indicate the enzymes with the highest centrality values. (B) Communities. The blue nodes are represented by communities containing enzymes discovered in the coexpression modules.

From the degree measures for each node (considering in and out connections), we identified 17 outliers (Figure 7A; Supplementary Figure 6), which were considered network hubs. We found enzymes with diverse roles (Supplementary Table 5), such as UDP-sugar

pyrophosphorylase (ec: 2.7.7.64) (34 connections), ureidoglycolate amidohydrolase (ec:3.5.1.116) (26 connections) and alanine-glyoxylate transaminase (ec:2.6.1.44) (25 connections). Interestingly, these enzymes were also the ones with the highest values of outdegree, stress and betweenness. Considering only the indegree connections, the top four enzymes (also identified among the network hubs) were UDP-sugar pyrophosphorylase (ec: 2.7.7.64) (34 connections), glutamate dehydrogenase (NAD (P) +) (ec: 1.4.1.3) (19 connections), glutamate dehydrogenase (NADP+) (ec: 1.4.1.4) (18 connections) and malate dehydrogenase (oxaloacetate-decarboxylating) (NADP+) (ec: 1.1.1.40) (15 connections). Among the 17 hubs, pyruvate kinase (ec: 2.7.1.40) also presented high values for other centrality measures (betweenness and stress). Additionally, the enzyme threonine synthase (ec: 4.2.3.1) showed the highest short path length value (14.30) and the highest eccentricity value (22), and the glucuronokinase enzyme (ec: 2.7.1.43) showed the highest value for neighborhood connectivity (24).

In addition to these evaluations, the modeled network was also categorized into condensed modules regarding the community structure and enzyme organization (Figure 7B; Supplementary Figure 7; Supplementary Table 6). Using the HiDeF (Zheng et al., 2021) algorithm, 149 communities were identified, containing 4 to 389 enzymes. The community with the highest eccentricity (7) was c1337, which also contained the highest number of enzymes. The community with the highest stress value (255) and betweenness (0.37) was c13340 (with 68 enzymes), and it was among the top three communities with the highest eccentricity value (5; Supplementary Table 6).

Discussion

The genetic improvement of rubber trees requires a long period of time, with more than 30 years estimated for developing an improved genotype (Gonçalves and Fontes, 2012). Despite the specialized labor required for *Hevea* phenotyping, its plantation is only possible in vast areas, making the selection process laborious and financially expensive. In this context, the use of MAS can drastically reduce the time and the cost of genetic improvement, especially if implemented in the first years after obtaining the seeds by selecting the target characteristics indirectly through phenotypically associated markers (Xu and Crouch, 2008). As a way of assisting such initiatives, in this work, we identified SNPs associated with SD,

and this set of markers can be used as high-priority candidates for MAS, with a high potential of providing greater precision and requiring less time in the selection of superior genotypes.

The main abiotic limitations for the productivity of cultivated plants are excessive salinity, adverse temperatures, and water deficit (Zhu, 2016), and for rubber tree production, water stress and cold are widely described as the most impactful limitations (Ding et al., 2020). Although several studies have investigated the molecular mechanisms of *Hevea* in cold resistance for its improvement (Cheng et al., 2018; Deng et al., 2018; Mantello et al., 2019), one of the main characteristics evaluated in *Hevea* breeding programs is SD (Priyadarshan, 2003) due to its versatility in assessing rubber tree productive efficiency (Dijkman, 1951; Goncalves et al., 1984; Chanroj et al., 2017; Conson et al., 2018; Khan et al., 2018; Chen et al., 2020). The use of SD measures can provide insights into phenotypes that can only be measured in specific climate conditions, such as drought resistance (Ohashi et al., 2006; Zhang et al., 2019a), which impacts rubber tree growth (Chandrashekar et al., 1998). Additionally, traits that can only be measured after a certain age of the plant, such as the production of latex and vigor (Dijkman, 1951; Goncalves et al., 1984), can be estimated by SD.

As SD is a quantitative characteristic, the study of the genetic architecture related to this trait is quite complex, considering the high amount of genes and metabolic pathways involved in its definition (Pootakham et al., 2020). Furthermore, the genome of rubber trees encompasses a large number of repetitive regions, reaching approximately 71% of rubber tree genomic content (Tang et al., 2016). The first *Hevea* reference genome at the chromosome level was only recently published in 2020 (Liu et al., 2020a), and most genomic approaches in the species have been based on highly fragmented sequences and biocomputational estimations (Pootakham et al., 2015; Chanroj et al., 2017; Conson et al., 2018; de Souza et al., 2018; Souza et al., 2019). Only with the advent of molecular biology techniques for reducing genomic complexities during sequencing procedures, such as GBS (Elshire et al., 2011; Poland and Rife, 2012), has it been feasible to generate thousands of SNP markers with high frequency in complex plant genomes (Pootakham et al., 2015). By using a GBS approach combined with a rubber tree chromosome-level reference genome, we characterized a large number of high-quality markers regarding their genomic distribution and LD relatedness, which enabled us to compare our findings with the locations of several QTLs for this characteristic, embracing novel possible causal genes explaining this phenotypic variation.

The large number of SNP markers discovered in this work allowed us to assess the LD throughout the genome of the entire population in a very representative way. As in other studies using arboreal and allogamous species (Peláez et al., 2020), our results showed high LD, which is consistent with previous *H. brasiliensis* results (Chanroj et al., 2017; De Souza et al., 2018). Interestingly, such elevated decay is not constant, and regions with a high density of TEs present a lower level of LD compared to the overall genomic LD. TEs are known as mobile elements due to their ability to change positions along the genome and produce copies of themselves (Singh et al., 2019), mainly in genomic regions with low LD (Stuart et al., 2016; Choudhury et al., 2019), as was observed in this work (Supplementary Figure 3). As stated by Choudhury et al. (2019), we also believe that there are two main reasons for this observation: (i) TEs alter the genetic architecture of the chromosome by decreasing the recombination rate in its vicinity and (ii) TEs accumulate in these regions due to the low recombination rate that occurs in these locations.

Several studies have been developed to characterize SD QTLs (Souza et al., 2013; Conson et al., 2018; Rosa et al., 2018); however, these studies are limited to the biparental populations employed (Myles et al., 2009). With the use of genetically diverse populations, GWAS approaches use the historical links between different genotypes, capturing more genetic diversity through a broader set of markers that would be neglected in association maps (Kulwal, 2018). When we are unable to identify the expected segregation ratios in markers from biparental progenies, these regions, even those close to important QTLs, are often discarded along with their associated QTLs (Kulwal, 2018). In this context, GWAS approaches have been suggested as a powerful tool for overcoming such limitations, which are intensified in species such as *H. brasiliensis*, in which there are great difficulties in obtaining mapping populations.

Genome-Wide Association Studies

To date, only one study employing GWAS has been described in the literature for *H. brasiliensis*. Using a population of 170 individuals genotyped with 14,155 SNP markers by capture probes (Shearman et al., 2014), Chanroj et al. (2017) tested four association models. The authors could associate two SNP markers with latex production (one for the rainy season and the other one for the drought season) and two others with SD (also separated by rainy and drought seasons). According to Conson et al. (2018), the rubber tree populations planted in the escape areas are under water stress at all times, despite the differences in water regime across seasons. Due to such observations and the Brazilian climate, we performed our

analyses without making this distinction. In this way, we identified four SNPs associated with SD, which were annotated following an RNA-Seq-based approach. Even though SD is a quantitative characteristic, the discovery of a relatively small number of markers by GWAS is possibly related to the limitation of the technique (Yang et al., 2010; Korte and Farlow, 2013; Tam et al., 2019). To overcome the large number of false negatives caused by the restrictive threshold employed (Tam et al., 2019) and the limitations related to the discovery of SNPs by GBS, we investigated the set of snpsLD. Although the GBS methodology avoids repeated regions of the genome (Elshire et al., 2011), these duplications represent about 70% of the rubber tree genome (Tang et al., 2016; Liu et al., 2020c), and the linkage disequilibrium tests represent an indirect way of assessing putative functional associations with GWAS results, including those ones caused by duplication events (Lee et al., 2012). Additionally, for going further in the establishment of putative genes related to QTLs, we integrated these results with co-expression network analyses based on RNA-Seq data. This approach has been shown to be effective in several species (Calabrese et al., 2017; Schaefer et al., 2018; Yan et al., 2020), but with a restricted use in non-model plant species.

Different from establishing a genomic window surrounding these markers and performing comparative alignments against plant databases (Chanroj et al., 2017; García-Fernández et al., 2021), we used an assembled transcriptome for the association of the snpsGWAS. This step was performed mainly because of the absence of available data for several neglected species in public databases (Schaefer et al., 2018), such as *H. brasiliensis*. Moreover, transcriptome assemblies are a way of categorizing a broader range of important genes found under stress conditions (Valdés et al., 2013; Wei et al., 2021), as already reported by other *Hevea* studies (Ahn et al., 2017; Mantello et al., 2019). Additionally, because of the recent availability of the *Hevea* genome (Liu et al., 2020a), more studies are required for complete and accurate gene categorization. By coupling the transcriptome assembly with GWAS, we could associate the three candidate genes identified by the snpsGWAS, which were annotated and had their expression profile estimated in two different genotypes, including in the population used here, and in specific stages of the plant development and physiology. As pointed out by Schaefer et al. (2018), this type of strategy provides associations not only with growth but also with resistance to abiotic stress. The genes identified flanking the snpsGWAS were interpreted according to their biological function and their metabolic context (Watanabe et al., 2017), suggesting their potential relationships in defining the phenotype.

The SBT4.6 gene (Table 1), identified based on the snpsGWAS, belongs to the subtilisin-like protease family, whose members are involved in general protein turnover and regulatory processes and in mechanisms of resistance to biotic and abiotic stresses (Tian et al., 2005; Budič et al., 2013; Figueiredo et al., 2018). Under normal conditions, mutants for this gene do not show obvious changes in the normal growth of the plant, so there is still a need for further investigations regarding this gene in the development of the plant (Rautengarten et al., 2005). Although this gene is not clearly involved in plant growth under normal conditions, we suggest that it may be indirectly related to this characteristic. Two other genes associated with snpsGWAS show evidence of a relationship with abiotic stresses. In addition to showing an increasing additive effect of 0.54 cm in the SD for a specific genotypic class (Table 1), SNP30209 was in the vicinity of a genomic region containing a candidate gene for GK1. In experiments carried out with *Arabidopsis thaliana*, GK1 showed a behavior of D-aminoacyl-tRNA deacylase, which is important for protecting the plant against the toxicity of D-amino acids (Wydau et al., 2007), which, when present in the soil, can have effects on plant growth in different ecosystems, whether managed or not. These compounds can act in different ways on root and stem growth, with D-serine, D-alanine and D-tyrosine being the strongest growth inhibitors, while others, such as D-lysine, D-isoleucine, D-valine, D-asparagine, and D-glutamine, act as milder inhibitors (Vranova et al., 2012). Another associated gene was IQM2, which contains a domain for the IQM2 protein. Such a protein belongs to a calmodulin-binding family protein and has strict involvement in the response to biotic and abiotic stress (Wan et al., 2012).

Despite the unquestionable importance of GWAS methods, the practical application of these findings in MAS for the selection of several complex characteristics is limited due to the low heritability associated with these markers (Bogardus, 2009). Considering this fact, we also investigated associated genomic regions, which may be jointly involved in phenotype definition (Yuan et al., 2012). Several statistical methods are used to identify genomic associations, such as multifactor dimensionality reduction (Ritchie et al., 2001), LD (Wu et al., 2008) and entropy-based statistics (Dong et al., 2008). In this work, we employed SNP correlations, which led us to already establish QTL positions (Conson et al., 2018), showing the robustness of this method. These newly identified markers may reveal genes that would be overlooked by conventional GWAS approaches.

In addition to MAS, other important tools for the genetic improvement of various plant species have been developed, such as iRNA (Zhang et al., 2017) and CRISPR (Jaganathan et al., 2018), which have shown enormous potential for breeding strategies in

recent years (Kalunke et al., 2020; Liu et al., 2020c). However, these approaches require the definition of target genes and their interactions, which might be estimated through coexpression and metabolic networks. In this way, to provide a deeper investigation into the metabolic activities of the genes associated with the snpsGWAS and snpsLD, we modeled complex networks to investigate their interactions and provide insights into the definition of the SD quantitative trait (Kosová et al., 2015; Tam et al., 2019), decreasing the variability of the indirectly selected phenotype and accessing other omics layers. The multiomics approaches employed here can contribute to a better understanding of the molecular mechanisms that are important to the vegetative growth of rubber trees, opening new perspectives for deeper genomic studies.

Multiomics

Quantitative traits are strongly affected by environment x genotype interactions (Nguyen et al., 2019). Genotypes with a greater capacity to resist these abiotic factors have a greater capacity to grow and develop under these stresses (Mantello et al., 2019). Understanding all the molecular biological levels that confer such a resistance to these specific genotypes requires the integration of multiple omics approaches, such as genomics, transcriptomics, proteomics and metabolomics. Multiomics approaches have as their main objective the integration of data analysis of different biological levels for a better understanding of their relationships and the functioning of a biological system as a whole (Joyce and Palsson, 2006). The use of joint approaches benefits from including all relevant parts that integrate the analyzed biological system (Zhang et al., 2010). Studies that integrate the discovery of QTLs with other omics have used genetically well-studied agricultural crops such as corn (Jiang et al., 2019) and, more recently, tree species such as citrus (Mou et al., 2021).

To provide deeper insights into the molecular basis of the evaluated phenotype, we extended the selected set of SD-associated SNPs with data from transcriptomics using complex network methodologies. These methodologies have revolutionized research in molecular biology because of their capability to simulate complex biological systems (D'haeseleer et al., 2000; Liu et al., 2020b) and infer novel biological associations, such as regulatory relationships, metabolic pathway inferences and annotation transference (Rao and Dixon, 2019). In *H. brasiliensi*s, coexpression network methods have already revealed genes involved in different environmental or stress conditions and are a powerful tool for profiling rubber tree samples (Hurtado Páez et al., 2015; Sathik et al., 2018; Mantello et al., 2019;

Deng et al., 2018; Ding et al., 2020). Such studies in rubber trees are still incipient and have not yet been coupled with breeding strategies for the genetic improvement of the species. Starting from RNA-Seq-based data, we could associate our GWAS results with expression profiles from important *Hevea* genotypes, incorporating our results into a complete set of molecular interactions estimated through the WGCNA approach. Using this strategy, we can infer biological functions for genes present in the same network module, as these genes probably exert correlated functions (Childs et al., 2011). This is the first initiative that proposes the integration of GWAS and coexpression networks in rubber trees to identify genes with great potential to be used in MAS.

The transcriptome used for annotation and construction of the coexpression network showed a large number of TEs, which are indeed present in large amounts in plant genomes (Matsunaga et al., 2015). In addition, these TEs were also found to be abundant in the selected functional modules, with TE 17.6 and TNT 1-94 being the most prominent. These TEs have already been described as being involved in gene expression, responses to external stimuli and plant development (Kashkush et al., 2003; Matsunaga et al., 2015; Traylor-Knowles et al., 2017; Tran and Choi, 2020). In rubber trees, TEs may be related to the differential expression observed in some commercial clones, affecting important processes such as rubber production (Wu et al., 2020a). As pointed out by Wang et al. (2020), the identification of TEs associated with functional genes related to important characteristics suggest that they can be used as molecular markers in MAS, contributing significantly to the genetic improvement of woody trees. In this sense, our findings supply a wide range of genomic resources for breeding. In the selected coexpression network, the most abundant elements were also TE 17.6 and TNT 1-94.

In the coexpression module with the largest number of genes, we were able to identify many genes related to plant growth, such as the protein MEI2-like 4 (ML4), which is a substrate for putative TOR, the main regulator of cell growth in eukaryotes (Anderson et al., 2005), representing an extremely important molecule in meiotic signaling (Watanabe et al., 1988). In this module, we also identified the proteins alkaline/neutral invertase CINV2 (CINV2) and LRR receptor-like serine/threonine-protein kinase GSO1 (GSO1), which are related to root growth and endoderm. The invertase enzyme (INV) is one of only two enzymes capable of catabolizing physiological carbon, together with the sucrose synthase enzyme (SUS); thus, most of the plant biomass is indispensable for normal growth, and the loss of these genes slows plant growth (Barratt et al., 2009). According to Racolta et al. (2014), the GSO1 protein works together with GSO2 for the intracellular signaling of the

plant, positively regulating cell proliferation, the differentiation of root cells and the identity of stem cells.

In the other functional modules, we identified several proteins involved in abiotic stress, such as transcription factor ICE1 (SCRM), an upstream transcription factor that regulates cold CBF gene transcription, improving plant tolerance to freezing (Chinnusamy et al., 2003). The regulatory-associated protein of TOR 1 (RAPTOR1) presents itself as a TOR regulator in response to osmotic stress (Mahfouz et al., 2006). The transcription factor jumgbrunnen 1 (JUB1), which delays senescence, also confers resistance to abiotic stress, such as heat shock, and resistance to high levels of intracellular H2O2 (Wu et al., 2012). Protein galactinol synthase 2 (GOLS2) plays an important role in the response against drought and cold stresses (Taji et al., 2002). The protein E3 ubiquitin-protein ligase PUB23 (PUB23), which responds quickly to water stress (Cho et al., 2008) and biotic stress, and the protein glucan endo-1,3-beta-glucosidase (HGN1) have been reported in *H. brasiliensis* and participate in a defense response against fungi (Galicia et al., 2015). In addition to these proteins produced in response to a given stress, genes involved in the maintenance and development of vegetative parts important for the development of the plant under a given stressful condition, such as constant drought, were identified, including arabidillo 1 protein (FBX5), which is related to the development of the roots (Coates et al., 2006), and Agamous-like MADS-box protein AGL12 (AGL12; Tapia-López et al., 2008). These results confirm the involvement of genes identified by GWAS and other genes identified in functional modules in the investigated characteristic definition. We can also relate the region of the SNP43760 marker, which has no known annotation, to QTLs involved in resistance to environmental factors, since the functional module containing these genes is related to this process.

Finally, a metabolic network for the enzymes found in this data set was constructed to identify the main metabolic pathways involved in the growth process of the rubber tree. The metabolites produced in cells can be understood as a bridge between the genotype and the phenotype. A clearer understanding of the relationship between these enzymes, such as by identifying the main enzymes present in the network, is essential for maintaining the properties of this network and thus preserving these relationships. The network built in this work shows some disconnected enzymes because the reactions that connect them with the other enzymes in the network have not yet been elucidated.

We identified UDP-sugar pyrophosphorylase (USP) as the hub of this enzyme network; this enzyme indicated to be an enzyme of great importance in the network, as it

presented the highest degree value (Barabási and Oltvai, 2004). It also showed the highest out-degree value, which represents the number of connections directed from this node to the other nodes in the network. This enzyme is very conserved in plants (Geserick and Tenhaken, 2013). Evidence indicates a high affinity of USP for acid-1-phosphate (UDP-GlcA-1-P), a substrate of the myo-inositol oxygenase (MIOX) pathway for UDP-GlcA (Geserick and Tenhaken, 2013). USP can also convert different types of sugar-1-phosphatates into the UDP sugars that make up polymers and glycerols in plant cell walls (Geserick and Tenhaken, 2013). USP is found in a single copy in Arabidopsis, and mutants for this gene are lethal (Geserick and Tenhaken, 2013), as the pollen that carries this mutation does not develop normally (Schnurr et al., 2006; Geserick and Tenhaken, 2013). Knock-down mutants also show impaired vegetative growth due to deficiency in sugar recycling (Geserick and Tenhaken, 2013). The enzyme glutamate dehydrogenase (NAD (P) +) (GDH) appeared in the enzymatic network containing a high degree of indegree. GDH catalyzes the deamination of glutamate using NAD as a coenzyme and releases 2-oxoglutarate and ammonia when there is little carbon (Fontaine et al., 2006). Participating in the response to various stresses, including drought and the presence of pathogens, their expression levels are regulated according to the intensity of the stress (Restivo, 2004), increasing the capacity of resistance to stress and the acquisition of biomass by the plant (Qiu et al., 2009; Tercé-Laforgue et al., 2015). The pyruvate kinase enzyme was shown to be central in the integration of its components, presenting a higher value of betweenness centrality (Brandes, 2001), indicating an important control function of this enzyme in the network, since this measure indicates elements in the network that join communities. In addition to this enzyme being important for the integration of the components in the metabolic network, this enzyme also presents itself as important in the dissemination of information among the elements present in the metabolic network, since it presented a higher stress value (Brandes, 2001), which indicates the shortest path between two random nodes in the network. This enzyme is a key element in the regulation and adjustment of the glucose metabolic pathway (Ambasht and Kayastha, 2002; Cai et al., 2018). Pyruvate kinase catalyzes the irreversible transfer of the high-energy phosphate group from phosphoenopyruvate to ADP, synthesizing ATP (Ambasht and Kayastha, 2002). Another important enzyme for the dissemination of information within the network was threonine synthase (thrC), which showed a higher value for the short path length (Watts and Strogatz, 1998) and eccentricity, which indicates the maximum number of nodes necessary for the information to reach all nodes present in the network (Hage and Harary, 1995). Theonine (Thr) enzymes play important roles in the stress response to abiotic factors such as

salinity, cold and drought (Rudrabhatla and Rajasekharan, 2002; Diédhiou et al., 2008), in addition to the different processes related to plant growth, such as cell division and the regulation of several phytohormones (Rudrabhatla and Rajasekharan, 2002) and carbon flux (Zeh et al., 2001).

In this work, we identified many genes involved in the response to drought, showing the importance of this element for the development of rubber trees, as already reported by Conforto (2008). Conson et al. (2018) and Souza et al. (2019) showed that the environments in which the populations used in this work are grown are environments with constant water deficit, which was expected because they are escape areas, which presents different climate of their natural habitat but where the rubber tree has adapted well. In the context of climate change, the discovery of genes involved in responding to water stress is of great value since forecasts show that in the near future areas suitable for planting today may become unsuitable (Ray et al., 2016). Most likely, these changes will occur mainly in the water regime, which can lead to the death of many woody plants (Adams et al., 2009).

Despite the limitations of the GWAS in identifying genes related to quantitative traits, the multiomics strategy employed in this study allowed us to explore the main genes that putatively define this phenotype from a holistic perspective, expanding this investigation and supplying a large reservoir of data. Using the integration of GWAS with coexpression networks and enzyme networks, we were able to elucidate the main relationships of these major genes and their products in a more complete way, mainly considering the limitations of GWAS in the identification of regions of QTLs with small effects. With the functional modules defined, we can gain insight into the genes that work together. In addition to the understanding that the definition of SD is based on the interaction of several processes, we have identified six functional modules. Even with more than one process, all these interactions work together, as we can see in the network shown in Figure 6. In addition, we can see the robustness of these results, which show correlations with previously published QTL maps (Conson et al., 2018). Posttranslational inferences were made regarding the relationships identified in the enzymatic network, which allowed us to identify new and important gene products that were previously unidentified. All these results show the importance of these integrative studies that correct the limitations of each individual technique.

This work is the first initiative that integrates multiomics in the study of QTLs in *H. brasiliensis*. Using this approach, we were able to access all important molecular levels for the definition of SD. Despite the great economic importance of the species, as it is the only

one capable of producing natural rubber in sufficient quantity and quality to supply the world market for this product (Ding et al., 2020), its genetic studies are still quite limited due to the complexity of its genome (Tang et al., 2016), its great genetic variability (De Souza et al., 2018) and the large areas needed for its plantation. Despite all these limitations, this work overcomes these difficulties, producing data, results and new methodological perspectives for future genomic studies in this species and identifying markers and genes useful for genetic improvement.

Data Availability Statement

FF and AA performed all the analyses and wrote the manuscript. CS assisted in the genotypic data analyses. PG, VG, and ES conducted the field experiments. AS, LS, and RF-N conceived the project. All authors contributed to the article and approved the submitted version.

Funding

The authors gratefully acknowledge the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) for Ph.D. fellowship to FF (18/18985-7) and AA (2019/03232-6); the Coordenação de Aperfeiçoamento do Pessoal de Nível Superior (CAPES) for financial support (Computational Biology Program and CAPES-Agropolis Program); and the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for research fellowships to AS and PG.

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.

Publisher's Note

Supplementary Material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2021.768589/full#supplementary-material

Footnotes

1. ^http://hannonlab.cshl.edu/fastx_toolkit/index.html

References

Adams, H. D., Guardiola-Claramonte, M., Barron-Gafford, G. A., Villegas, J. C., Breshears, D. D., Zou, C. B., et al. (2009). Temperature sensitivity of drought-induced tree mortality portends increased regional die-off under global-change-type drought. Proc. Natl. Acad. Sci. U. S. A. 106, 7063–7066. doi: 10.1073/pnas.0901438106

Ahn, H., Jung, I., Shin, S. J., Park, J., Rhee, S., Kim, J. K., et al. (2017). Transcriptional network analysis reveals drought resistance mechanisms of AP2/ERF transgenic rice. Front. Plant Sci. 8:1044. doi: 10.3389/fpls.2017.01044

Ambasht, P. K., and Kayastha, A. M. (2002). Plant pyruvate kinase. Biol. Plant. 45, 1–10. doi: 10.1023/A:1015173724712

Anderson, G. H., Veit, B., and Hanson, M. R. (2005). The Arabidopsis AtRaptor genes are essential for post-embryonic plant growth. BMC Biol. 3:12. doi: 10.1186/1741-7007-3-12

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

Barratt, D. H., Derbyshire, P., Findlay, K., Pike, M., Wellner, N., Lunn, J., et al. (2009). Normal growth of Arabidopsis requires cytosolic invertase but not sucrose synthase. Proc. Natl. Acad. Sci. U. S. A. 106, 13124–13129. doi: 10.1073/pnas.0900689106

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E., et al. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. 31, 365–370. doi: 10.1093/nar/gkg095

Bogardus, C. (2009). Missing heritability and GWAS utility. Obesity 17:209. doi: 10.1038/oby.2008.613

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

Budič, M., Sabotič, J., Meglič, V., Kos, J., and Kidrič, M. (2013). Characterization of two novel subtilases from common bean (Phaseolus vulgaris L.) and their responses to drought. Plant Physiol. Biochem. 62, 79–87. doi: 10.1016/j.plaphy.2012.10.022

Cai, Y., Li, S., Jiao, G., Sheng, Z., Wu, Y., Shao, G., et al. (2018). OsPK2 encodes a plastidic pyruvate kinase involved in rice endosperm starch synthesis, compound granule formation and grain filling. Plant Biotechnol. J. 16, 1878–1891. doi: 10.1111/pbi.12923

Calabrese, G. M., Mesner, L. D., Stains, J. P., Tommasini, S. M., Horowitz, M. C., Rosen, C. J., et al. (2017). Integrating GWAS and co-expression network data identifies bone mineral density genes SPTBN1 and MARK3 and an osteoblast functional module. Cell Syst. 4, 46–59. doi: 10.1016/j.cels.2016.10.014

Campbell, M. S., Holt, C., Moore, B., and Yandell, M. (2014). Genome annotation and curation using MAKER and MAKER-P. Curr. Protoc. Bioinformatics 48, 4.11.1–4.11.39. doi: 10.1002/0471250953.bi0411s48

Challa, S., and Neelapu, N. R. R. (2018). “Genome-wide association studies (GWAS) for abiotic stress tolerance in plants,” in Biochemical, Physiological and Molecular Avenues for Combating Abiotic Stress in Plants. ed. S. H. Wani (London, UK: Academic Press), 125–150.

Chandrashekar, T. R., Nazeer, M. A., Marattukalam, J. G., Prakash, G. P., Annamalainathan, K., and Thomas, J. (1998). An analysis of growth and drought tolerance in rubber during the immature phase in a dry subhumid climate. Exp. Agric. 34, 287–300. doi: 10.1017/S0014479798343045

Chanroj, V., Rattanawong, R., Phumichai, T., Tangphatsornruang, S., and Ukoskit, K. (2017). Genome-wide association mapping of latex yield and girth in Amazonian accessions of Hevea

brasiliensis grown in a suboptimal climate zone. Genomics 109, 475–484. doi: 10.1016/j.ygeno.2017.07.005

Chen, L., Fang, Y., Li, X., Zeng, K., Chen, H., Zhang, H., et al. (2020). Identification of soybean drought-tolerant genotypes and loci correlated with agronomic traits contributes new candidate genes for breeding. Plant Mol. Biol. 102, 109–122. doi: 10.1007/s11103-019-00934-7

Cheng, H., Chen, X., Fang, J., An, Z., Hu, Y., and 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 Physiol. 38, 1409–1423. doi: 10.1093/treephys/tpy014

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

Chinnusamy, V., Ohta, M., Kanrar, S., Lee, B. H., Hong, X., Agarwal, M., et al. (2003). ICE1: a regulator of cold-induced transcriptome and freezing tolerance in Arabidopsis. Genes Dev. 17, 1043–1054. doi: 10.1101/gad.1077503

Cho, S. K., Ryu, M. Y., Song, C., Kwak, J. M., and Kim, W. T. (2008). Arabidopsis PUB22 and PUB23 are homologous U-box E3 ubiquitin ligases that play combinatory roles in response to drought stress. Plant Cell 20, 1899–1914. doi: 10.1105/tpc.108.060699

Choudhury, R. R., Rogivue, A., Gugerli, F., and Parisod, C. (2019). Impact of polymorphic transposable elements on linkage disequilibrium along chromosomes. Mol. Ecol. 28, 1550–1562. doi: 10.1111/mec.15014

Coates, J. C., Laplaze, L., and Haseloff, J. (2006). Armadillo-related proteins promote lateral root development in Arabidopsis. Proc. Natl. Acad. Sci. 103, 1621–1626. doi: 10.1073/pnas.0507575103

Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., et al. (2009). Biopython: freely available python tools for computational molecular biology and bioinformatics. Bioinformatics 25, 1422–1423. doi: 10.1093/bioinformatics/btp163

Conforto, E. D. (2008). Respostas fisiológicas ao déficit hídrico em duas cultivares enxertadas de seringueira (“RRIM 600” e “GT 1”) crescidas em campo. Cienc. Rural. 38, 679–684. doi: 10.1590/S0103-84782008000300013

Conson, A. R. O., Taniguti, C. H., Amadeu, R. R., Andreotti, I. A. A., De Souza, L. M., Dos Santos, L. H. B., et al. (2018). High-resolution genetic map and QTL analysis of growth-related traits of

Hevea brasiliensis cultivated under suboptimal temperature and humidity conditions. Front. Plant Sci. 9:1255. doi: 10.3389/fpls.2018.01255

Csardi, G., and Nepusz, T. (2006). The igraph software package for complex network research. Int. J. Complex Syst. 1695, 1–9. doi: 10.1016/B978-0-12-813066-7.00009-7

D'haeseleer, P., Liang, S., and Somogyi, R. (2000). Genetic network inference: from co-expression clustering to reverse engineering. Bioinformatics 16, 707–726. doi: 10.1093/bioinformatics/16.8.707

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The variant call format and VCFtools. Bioinformatics 27, 2156–2158. doi: 10.1093/bioinformatics/btr330

De Faÿ, E., and Jacob, J. L. (1989). "Anatomical organization of the laticiferous system in the bark," in Physiology of Rubber Tree Latex. eds. J. D'Auzac, J. Jacob, and H. Chrestin (Boca Raton, FL: CRC Press), 3–14.

De Souza, L. M., Dos Santos, L. H. B., Rosa, J., Da Silva, C. C., Mantello, C. C., Conson, A. R. O., et al. (2018). Linkage disequilibrium and population structure in wild and cultivated populations of rubber tree (Hevea brasiliensis). Front. Plant Sci. 9:815. doi: 10.3389/fpls.2018.00815

Deng, X., Wang, J., Li, Y., Wu, S., Yang, S., Chao, J., et al. (2018). Comparative transcriptome analysis reveals phytohormone signalings, heat shock module and ROS scavenger mediate the cold-tolerance of rubber tree. Sci. Rep. 8:4931. doi: 10.1038/s41598-018-23094-y

Diédhiou, C. J., Popova, O. V., Dietz, K. J., and Golldack, D. (2008). The SNF1-type serine-threonine protein kinase SAPK4 regulates stress-responsive gene expression in rice. BMC Plant Biol. 8:49. doi: 10.1186/1471-2229-8-49

Dijkman, M. (1951). Hevea. Thiry Years of Research in the Far East. Waltham, MA: University of Miami Press.

Ding, Z., Fu, L., Tan, D., Sun, X., and 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. Ind. Crop. Prod. 153:112562. doi: 10.1016/j.indcrop.2020.112562

Dong, C., Chu, X., Wang, Y., Wang, Y., Jin, L., Shi, T., et al. (2008). Exploration of gene-gene interaction effects using entropy-based methods. Eur. J. Hum. Genet. 16, 229–235. doi: 10.1038/sj.ejhg.5201921

Dusotoit-Coucaud, A., Brunel, N., Kongsawadworakul, P., Viboonjun, U., Lacointe, A., Julien, J. L., et al. (2009). Sucrose importation into laticifers of Hevea brasiliensis, in relation to ethylene stimulation of latex production. Ann. Bot. 104, 635–647. doi: 10.1093/aob/mcp150

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., et al. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6:e19379. doi: 10.1371/journal.pone.0019379

Figueiredo, J., Sousa Silva, M., and Figueiredo, A. (2018). Subtilisin-like proteases in plant defence: the past, the present and beyond. Mol. Plant Pathol. 19, 1017–1028. doi: 10.1111/mpp.12567

Fonseca, P. A. D. S., Id-Lahoucine, S., Reverter, A., Medrano, J. F., Fortes, M. S., Casellas, J., et al. (2018). Combining multi-OMICs information to identify key-regulator genes for pleiotropic effect on fertility and production traits in beef cattle. PLoS One 13:e0205295. doi: 10.1371/journal.pone.0205295

Fontaine, J. X., Saladino, F., Agrimonti, C., Bedu, M., Tercé-Laforgue, T., Tétu, T., et al. (2006). Control of the synthesis and subcellular targeting of the two GDH genes products in leaves and stems of Nicotiana plumbaginifolia and Arabidopsis thaliana. Plant Cell Physiol. 47, 410–418. doi: 10.1093/pcp/pcj008

Galicia, C., Mendoza-Hernández, G., and Rodríguez-Romero, A. (2015). Impact of the vulcanization process on the structural characteristics and IgE recognition of two allergens, Hev b 2 and Hev b 6.02, extracted from latex surgical gloves. Mol. Immunol. 65, 250–258. doi: 10.1016/j.molimm.2015.01.018

García-Fernández, C., Campa, A., Garzón, A. S., Miklas, P., and Ferreira, J. J. (2021). GWAS of pod morphological and color characters in common bean. BMC Plant Biol. 21:184. doi: 10.1186/s12870-021-02967-x

Gerard, D. (2020). Pairwise linkage disequilibrium estimation for polyploids. Mol. Ecol. Resour. 21, 1230–1242. doi: 10.1111/1755-0998.13349

Geserick, C., and Tenhaken, R. (2013). UDP-sugar pyrophosphorylase is essential for arabinose and xylose recycling, and is required during vegetative and reproductive growth in Arabidopsis. Plant J. 74, 239–247. doi: 10.1111/tpj.12116

Glaubitz, J. C., Casstevens, T. M., Lu, F., Harriman, J., Elshire, R. J., Sun, Q., et al. (2014). TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. PLoS One 9:e90346. doi: 10.1371/journal.pone.0090346

Gonçalves, P. D. S., Bortoletto, N., Ortolani, A. A., Belletti, G. O., and Santos, W. R. D. (1999). Desempenho de novos clones de seringueira: III. seleções promissoras para a região de Votuporanga, Estado de São Paulo. Pesquisa Agropecuária

Gonçalves, P. S., and Fontes, J. R. A. (2012). “Domestication and breeding of rubber tree,” in Domestication and Breeding – Amazonian Species. eds. A. Borém, M. T. G. Lopes, C. R. C. Clement, and H. Noda (Viçosa, Brazil: Suprema Editora Ltda), 393–441.

Goncalves, P. D. S., Rossetti, A. G., Valois, A. C. C., and Viegas, I. D. J. (1984). Genetic and phenotypic correlations between some quantitative traits in juvenile clonal rubber trees (Hevea Spp.). Rev. Bras. Genet. 7, 95–107.

Gonçalves, P. D. S., Silva, M. D. A., Gouvêa, L. R. L., and Scaloppi Junior, E. J. (2006). Genetic variability for girth growth and rubber yield in Hevea brasiliensis. Sci. Agric. 63, 246–254. doi: 10.1590/S0103-90162006000300006

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29, 644–652. doi: 10.1038/nbt.1883

Granato, I., and Fritsche-Neto, R. (2018). snpReady: Preparing Genotypic Datasets in Order to Run Genomic; Analysis. R Package Version 0.9.6. Available at: https://CRAN.R-project.org/package=snpReady. (Acessed July 24, 2020).

Haas, B. J. (2015). Trinotate: Transcriptome Functional Annotation and Analysis. Available at: https://github.com/Trinotate/Trinotate.github.io/wiki. (Acessed July 24, 2020).

Haas, B. J., Delcher, A. L., Mount, S. M., Wortman, J. R., Smith, R. K. Jr., Hannick, L. I., et al. (2003). Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res. 31, 5654–5666. doi: 10.1093/nar/gkg770

Hage, P., and Harary, F. (1995). Eccentricity and centrality in networks. Soc. Networks 17, 57–63. doi: 10.1016/0378-8733(94)00248-9

Hastie, T., Tibshirani, R., Balasubramanian, N., and Chu, G. (2001). Impute: Imputation for Microarray Data. Bioinform. 17, 520–525.

Hurtado Páez, U. A., García Romero, I. A., Restrepo Restrepo, S., Aristizábal Gutiérrez, F. A., and Montoya Castaño, D. (2015). Assembly and analysis of differential transcriptome responses of Hevea brasiliensis on interaction with Microcyclus ulei. PLoS One 10:e0134837. doi: 10.1371/journal.pone.0134837

Jaganathan, D., Ramasamy, K., Sellamuthu, G., Jayabalan, S., and Venkataraman, G. (2018). CRISPR for crop improvement: an update review. Front. Plant Sci. 9:985. doi: 10.3389/fpls.2018.00985

Jamil, I. N., Remali, J., Azizan, K. A., Nor Muhammad, N. A., Arita, M., Goh, H. H., et al. (2020). Systematic multi-omics integration (MOI) approach in plant systems biology. Front. Plant Sci. 11:944. doi: 10.3389/fpls.2020.00944

Jiang, J., Xing, F., Wang, C., Zeng, X., and Zou, Q. (2019). Investigation and development of maize fused network analysis with multi-omics. Plant Physiol. Biochem. 141, 380–387. doi: 10.1016/j.plaphy.2019.06.016

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., and Madden, T. L. (2008). NCBI BLAST: a better web interface. Nucleic Acids Res. 36, W5–W9. doi: 10.1093/nar/gkn201

Joyce, A. R., and Palsson, B. (2006). The model organism as a system: integrating 'omics' data sets. Nat. Rev. Mol. Cell Biol. 7, 198–210. doi: 10.1038/nrm1857

Kalunke, R. M., Tundo, S., Sestili, F., Camerlengo, F., Lafiandra, D., Lupi, R., et al. (2020). Reduction of allergenic potential in bread wheat RNAi transgenic lines silenced for CM3, CM16 and 0.28 ATI genes. Int. J. Mol. Sci. 21:5817. doi: 10.3390/ijms21165817

Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 28, 27–30. doi: 10.1093/nar/28.1.27

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

Khan, M. A., Tong, F., Wang, W., He, J., Zhao, T., and Gai, J. (2018). Analysis of QTL-allele system conferring drought tolerance at seedling stage in a nested association mapping population of soybean [Glycine max (L.) Merr.] using a novel GWAS procedure. Planta 248, 947–962. doi: 10.1007/s00425-018-2952-4

Kim, D., Langmead, B., and Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. Nat. Methods 12, 357–360. doi: 10.1038/nmeth.3317

Kleinberg, J. M. (1999). Hubs, authorities, and communities. ACM Comput. Surv. 31:5–es. doi: 10.1145/345966.345982

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

Kosová, K., Vítámvás, P., Urban, M. O., Klíma, M., Roy, A., and Prášil, I. T. (2015). Biological networks underlying abiotic stress tolerance in temperate crops: a proteomic perspective. Int. J. Mol. Sci. 16, 20913–20942. doi: 10.3390/ijms160920913

Kulwal, P. L. (2018). Trait mapping approaches through linkage mapping in plants. Adv. Biochem. Eng. Biotechnol. 164, 53–82. doi: 10.1007/10_2017_49

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

Langfelder, P., Zhang, B., and Horvath, S. (2008). Defining clusters from a hierarchical cluster tree: the dynamic tree cut package for R. Bioinformatics 24, 719–720. doi: 10.1093/bioinformatics/btm563

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. Nat. Methods 9, 357–359. doi: 10.1038/nmeth.1923

Lee, K. W., San Woon, P., Teo, Y. Y., and Sim, K. (2012). Genome wide association studies (GWAS) and copy number variation (CNV) studies of the major psychoses: what have we learnt? Neurosci. Biobehav. Rev. 36, 556–571. doi: 10.1016/j.neubiorev.2011.09.001

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. Bioinformatics 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

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

Liu, H. J., Jian, L., Xu, J., Zhang, Q., Zhang, M., Jin, M., et al. (2020a). High-throughput CRISPR/Cas9 mutagenesis streamlines trait gene identification in maize. Plant Cell 32, 1397–1413. doi: 10.1105/tpc.19.00934

Liu, C., Ma, Y., Zhao, J., Nussinov, R., Zhang, Y. C., Cheng, F., et al. (2020b). Computational network biology: data, models, and applications. Phys. Rep. 846, 1–66. doi: 10.1016/j.physrep.2019.12.004

Liu, J., Shi, C., Shi, C. C., Li, W., Zhang, Q. J., Zhang, Y., et al. (2020c). The chromosome-based rubber tree genome provides new insights into spurge genome evolution and rubber biosynthesis. Mol. Plant 13, 336–350. doi: 10.1016/j.molp.2019.10.017

Mahfouz, M. M., Kim, S., Delauney, A. J., and Verma, D. P. (2006). Arabidopsis TARGET OF RAPAMYCIN interacts with RAPTOR, which regulates the activity of S6 kinase in response to osmotic stress signals. Plant Cell 18, 477–490. doi: 10.1105/tpc.105.035931

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

Mantello, C. C., Boatwright, L., da Silva, C. C., Scaloppi, E. J., de Souza Goncalves, P., Barbazuk, W. B., et al. (2019). Deep expression analysis reveals distinct cold-response strategies in rubber tree (Hevea brasiliensis). BMC Genomics 20:455. doi: 10.1186/s12864-019-5852-5

Mantello, C. C., Cardoso-Silva, C. B., Da Silva, C. C., De Souza, L. M., Scaloppi Junior, E. J., De Souza Gonçalves, P., 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

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

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

Mou, J., Zhang, Z., Qiu, H., Lu, Y., Zhu, X., Fan, Z., et al. (2021). Multiomics-based dissection of citrus flavonoid metabolism using a Citrus reticulata × Poncirus trifoliata population. Hortic. Res. 8:56. doi: 10.1038/s41438-021-00472-8

Munõz, F., and Sanchez, L. (2017). breedR: Statistical Methods for Forest Genetic; Resources Analysts. R Package Version 0.12–2. Available at: https://github.com/famuvie/breedR (Accessed November 2021).

Myles, S., Peiffer, J., Brown, P. J., Ersoz, E. S., Zhang, Z., Costich, D. E., et al. (2009). Association mapping: critical considerations shift from genotyping to experimental design. Plant Cell 21, 2194–2202. doi: 10.1105/tpc.109.068437

Nebel, A., Kleindorp, R., Caliebe, A., Nothnagel, M., Blanché, H., Junge, O., et al. (2011). A genome-wide association study confirms APOE as the major gene influencing survival in long-lived individuals. Mech. Ageing Dev. 132, 324–330. doi: 10.1016/j.mad.2011.06.008

Nguyen, K. L., Grondin, A., Courtois, B., and Gantet, P. (2019). Next-generation sequencing accelerates crop gene discovery. Trends Plant Sci. 24, 263–274. doi: 10.1016/j.tplants.2018.11.008

Ohashi, Y., Nakayama, N., Saneoka, H., and Fujita, K. (2006). Effects of drought stress on photosynthetic gas exchange, chlorophyll fluorescence and stem diameter of soybean plants. Biol. Plant. 50, 138–141. doi: 10.1007/s10535-005-0089-3

Peláez, P., Ortiz-Martínez, A., Figueroa-Corona, L., Montes, J. R., and Gernandt, D. S. (2020). Population structure, diversifying selection, and local adaptation in Pinus patula. Am. J. Bot. 107, 1555–1566. doi: 10.1002/ajb2.1566

Pérez-Bercoff, Å., McLysaght, A., and Conant, G. C. (2011). Patterns of indirect protein interactions suggest a spatial organization to metabolism. Mol. BioSyst. 7, 3056–3064. doi: 10.1039/c1mb05168g

Pinu, F. R., Beale, D. J., Paten, A. M., Kouremenos, K., Swarup, S., Schirra, H. J., et al. (2019). Systems biology and multi-omics integration: viewpoints from the metabolomics research community. Meta 9:76. doi: 10.3390/metabo9040076

Poland, J. A., and Rife, T. W. (2012). Genotyping-by-sequencing for plant breeding and genetics. Plant Genome 5, 92–102. doi: 10.3835/plantgenome2012.05.0005

Pootakham, W., Ruang-Areerate, P., Jomchai, N., Sonthirod, C., Sangsrakru, D., Yoocha, T., et al. (2015). Construction of a high-density integrated genetic linkage map of rubber tree (Hevea brasiliensis) using genotyping-by-sequencing (GBS). Front. Plant Sci. 6:367. doi: 10.3389/fpls.2015.00367

Pootakham, W., Shearman, J. R., and Tangphatsornruang, S. (2020). “Development of molecular markers in Hevea brasiliensis for marker-assisted breeding,” in The Rubber Tree Genome. eds. M. Matsui and K. S. Chow (Cham: Springer International Publishing), 67–79.

Pootakham, W., Sonthirod, C., Naktang, C., Ruang-Areerate, P., Yoocha, T., Sangsrakru, D., et al. (2017). De novo hybrid assembly of the rubber tree genome reveals evidence of paleotetraploidy in Hevea species. Sci. Rep. 7:41457. doi: 10.1038/srep41457

Priyadarshan, P. (2003). Contributions of weather variables for specific adaptation of rubber tree (Hevea brasiliensis Muell.-Arg) clones. Genet. Mol. Biol. 26, 435–440. doi: 10.1590/S1415-47572003000400006

Priyadarshan, P. M. (2017). Refinements to Hevea rubber breeding. Tree Genet. Genomes 13:20. doi: 10.1007/s11295-017-1101-8

Priyadarshan, P. M., and Clément-Demange, A. (2004). Breeding Hevea rubber: formal and molecular genetics. Adv. Genet. 52, 51–115. doi: 10.1016/s0065-2660(04)52003-5

Qiu, X., Xie, W., Lian, X., and Zhang, Q. (2009). Molecular analyses of the rice glutamate dehydrogenase gene family and their response to nitrogen and phosphorous deprivation. Plant Cell Rep. 28, 1115–1126. doi: 10.1007/s00299-009-0709-z

Racolta, A., Bryan, A. C., and Tax, F. E. (2014). The receptor-like kinases GSO1 and GSO2 together regulate root growth in Arabidopsis through control of cell division and cell fate specification. Dev. Dyn. 243, 257–278. doi: 10.1002/dvdy.24066

Rao, X., and Dixon, R. A. (2019). Co-expression networks for plant biology: why and how. Acta Biochim. Biophys. Sin. 51, 981–988. doi: 10.1093/abbs/gmz080

Rao, G. P., and Kole, P. C. (2016). Evaluation of Brazilian wild Hevea germplasm for cold tolerance: genetic variability in the early mature growth. J. For. Res. 27, 755–765. doi: 10.1007/s11676-015-0188-8

Rautengarten, C., Steinhauser, D., Büssis, D., Stintzi, A., Schaller, A., Kopka, J., et al. (2005). Inferring hypotheses on functional relationships of genes: analysis of the Arabidopsis thaliana subtilase gene family. PLoS Comput. Biol. 1:e40. doi: 10.1371/journal.pcbi.0010040

Ray, D., Behera, M. D., and Jacob, J. (2016). Predicting the distribution of rubber trees (Hevea brasiliensis) through ecological niche modelling with climate, soil, topography and socioeconomic factors. Ecol. Res. 31, 75–91. doi: 10.1007/s11284-015-1318-7

Restivo, F. M. (2004). Molecular cloning of glutamate dehydrogenase genes of Nicotiana plumbaginifolia: structure analysis and regulation of their expression by physiological and stress conditions. Plant Sci. 166, 971–982. doi: 10.1016/j.plantsci.2003.12.011

Ritchie, M. D., Hahn, L. W., Roodi, N., Bailey, L. R., Dupont, W. D., Parl, F. F., et al. (2001). Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am. J. Hum. Genet. 69, 138–147. doi: 10.1086/321276

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139–140. doi: 10.1093/bioinformatics/btp616.

Romain, B., and Thierry, C. (2011). Rubber Clones (Hevea Clonal Descriptions). Available at: http://rubberclones.cirad.fr (Acessed July 24, 2020).

Rosa, J., Mantello, C. C., Garcia, D., De Souza, L. M., Da Silva, C. C., Gazaffi, R., et al. (2018). QTL detection for growth and latex production in a full-sib rubber tree population cultivated under suboptimal climate conditions. BMC Plant Biol. 18:223. doi: 10.1186/s12870-018-1450-y

Rudrabhatla, P., and Rajasekharan, R. (2002). Developmentally regulated dual-specificity kinase from peanut that is induced by abiotic stresses. Plant Physiol. 130, 380–390. doi: 10.1104/pp.005173

Sathik, M. B. M., Luke, L. P., Rajamani, A., Kuruvilla, L., Sumesh, K. V., and Thomas, M. (2018). De novo transcriptome analysis of abiotic stress-responsive transcripts of Hevea brasiliensis. Mol. Breed. 38:32. doi: 10.1007/s11032-018-0782-5

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

Schnurr, J. A., Storey, K. K., Jung, H. J. G., Somers, D. A., and Gronwald, J. W. (2006). UDP-sugar pyrophosphorylase is essential for pollen development in Arabidopsis. Planta 224, 520–532. doi: 10.1007/s00425-006-0240-1

Scossa, F., Alseekh, S., and Fernie, A. R. (2021). Integrating multi-omics data for crop improvement. J. Plant Physiol. 257:153352. doi: 10.1016/j.jplph.2020.153352

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 13, 2498–2504. doi: 10.1101/gr.1239303

Shearman, J. R., Sangsrakru, D., Ruang-Areerate, P., Sonthirod, C., Uthaipaisanwong, P., Yoocha, T., et al. (2014). Assembly and analysis of a male sterile rubber tree mitochondrial genome reveals DNA rearrangement events and a novel transcript. BMC Plant Biol. 14:45. doi: 10.1186/1471-2229-14-45

Singh, G., Ratnaparkhe, M., and Kumar, A. (2019). Comparative analysis of transposable elements from Glycine max, Cajanus cajan and Phaseolus vulgaris. J. Exp. Biol. Agric. Sci. 7, 167–177. doi: 10.18006/2019.7(2).167.177

Sivakumaran, S., Haridas, G., and Abraham, P. D. (1988). Problem of tree dryness with high yielding precocious clones and methods to exploit such clones. Proc. Coll. Hevea 88, 253–267.

Souza, L. M., Francisco, F. R., Gonçalves, P. S., Scaloppi Junior, E. J., Le Guen, V., Fritsche-Neto, R., et al. (2019). Genomic selection in rubber tree breeding: a comparison of models and methods for managing G×E interactions. Front. Plant Sci. 10:1353. doi: 10.3389/fpls.2019.01353

Souza, L. M., Gazaffi, R., Mantello, C. C., Silva, C. C., Garcia, D., Le Guen, V., et al. (2013). QTL mapping of growth-related traits in a full-sib family of rubber tree (Hevea brasiliensis) evaluated in a sub-tropical climate. PLoS One 8:e61238. doi: 10.1371/journal.pone.0061238

Stuart, T., Eichten, S. R., Cahn, J., Karpievitch, Y. V., Borevitz, J. O., and Lister, R. (2016). Population scale mapping of transposable element diversity reveals links to gene regulation and epigenomic variation. elife 5:e20777. doi: 10.7554/eLife.20777

Supek, F., Bošnjak, M., Škunca, N., and Šmuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. PLoS One 6:e21800. doi: 10.1371/journal.pone.0021800

Taji, T., Ohsumi, C., Iuchi, S., Seki, M., Kasuga, M., Kobayashi, M., et al. (2002). Important roles of drought- and cold-inducible genes for galactinol synthase in stress tolerance in Arabidopsis thaliana. Plant J. 29, 417–426. doi: 10.1046/j.0960-7412.2001.01227.x

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

Tang, C., Yang, M., Fang, Y., Luo, Y., Gao, S., Xiao, X., et al. (2016). The rubber tree genome reveals new insights into rubber production and species adaptation. Nat. Plants 2:16073. doi: 10.1038/nplants.2016.73

Tapia-López, R., García-Ponce, B., Dubrovsky, J. G., Garay-Arroyo, A., Pérez-Ruíz, R. V., Kim, S. H., et al. (2008). An AGAMOUS-related MADS-box gene, XAL1 (AGL12), regulates root meristem cell proliferation and flowering transition in Arabidopsis. Plant Physiol. 146, 1182–1192. doi: 10.1104/pp.107.108647

Tenesa, A., Wright, A. F., Knott, S. A., Carothers, A. D., Hayward, C., Angius, A., et al. (2004). Extent of linkage disequilibrium in a Sardinian sub-isolate: sampling and methodological considerations. Hum. Mol. Genet. 13, 25–33. doi: 10.1093/hmg/ddh001

Tercé-Laforgue, T., Clément, G., Marchi, L., Restivo, F. M., Lea, P. J., and Hirel, B. (2015). Resolving the role of plant NAD-glutamate dehydrogenase: III. Overexpressing individually or simultaneously the two enzyme subunits under salt stress induces changes in the leaf metabolic profile and increases plant biomass production. Plant Cell Physiol. 56, 1918–1929. doi: 10.1093/pcp/pcv114

Tian, M., Benedetti, B., and Kamoun, S. (2005). A second kazal-like protease inhibitor from Phytophthora infestans inhibits and interacts with the apoplastic pathogenesis-related protease P69B of tomato. Plant Physiol. 138, 1785–1793. doi: 10.1104/pp.105.061226

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

Trivedi, U. H., Cézard, T., Bridgett, S., Montazam, A., Nichols, J., Blaxter, M., et al. (2014). Quality control of next-generation sequencing data without a reference. Front. Genet. 5:111. doi: 10.3389/fgene.2014.00111

Valdés, A., Ibáñez, C., Simó, C., and García-Cañas, V. (2013). Recent transcriptomics advances and emerging applications in food science. TrAC Trends Anal. Chem. 52, 142–154. doi: 10.1016/j.trac.2013.06.014

Verzegnazzi, A. L., Dos Santos, I. G., Krause, M. D., Hufford, M., Frei, U. K., Campbell, J., et al. (2021). Major locus for spontaneous haploid genome doubling detected by a case-control GWAS in exotic maize germplasm. Theor. Appl. Genet. 134, 1423–1434. doi: 10.1007/s00122-021-03780-8

Voorrips, R. E. (2002). MapChart: software for the graphical presentation of linkage maps and QTLs. J. Hered. 93, 77–78. doi: 10.1093/jhered/93.1.77

Vranova, V., Zahradnickova, H., Janous, D., Skene, K. R., Matharu, A. S., Rejsek, K., et al. (2012). The significance of D-amino acids in soil, fate and utilization by microbes and plants: review and identification of knowledge gaps. Plant Soil 354, 21–39. doi: 10.1007/s11104-011-1059-5

Wan, D., Li, R., Zou, B., Zhang, X., Cong, J., Wang, R., et al. (2012). Calmodulin-binding protein CBP60g is a positive regulator of both disease resistance and drought tolerance in Arabidopsis. Plant Cell Rep. 31, 1269–1281. doi: 10.1007/s00299-012-1247-7

Wang, J., Lu, N., Yi, F., and Xiao, Y. (2020). Identification of transposable elements in conifer and their potential application in breeding. Evol. Bioinforma. 16:1176934320930263. doi: 10.1177/1176934320930263

Wang, Q., Wang, K., Wu, W., Giannoulatou, E., Ho, J. W. K., and Li, L. (2019). Host and microbiome multi-omics integration: applications and methodologies. Biophys. Rev. 11, 55–65. doi: 10.1007/s12551-018-0491-7

Warraich, A. S., Krishnamurthy, S. L., Sooch, B. S., Vinaykumar, N. M., Dushyanthkumar, B. M., Bose, J., et al. (2020). Rice GWAS reveals key genomic regions essential for salinity tolerance at reproductive stage. Acta Physiol. Plant. 42:134. doi: 10.1007/s11738-020-03123-y

Watanabe, Y., Lino, Y., Furuhata, K., Shimoda, C., and Yamamoto, M. (1988). The S.pombe mei2 gene encoding a crucial molecule for commitment to meiosis is under the regulation of cAMP. EMBO J. 7, 761–767. doi: 10.1002/j.1460-2075.1988.tb02873.x

Watanabe, K., Taskesen, E., Van Bochoven, A., and Posthuma, D. (2017). Functional mapping and annotation of genetic associations with FUMA. Nat. Commun. 8:1826. doi: 10.1038/s41467-017-01261-5

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

Wei, Z., Yuan, Q., Lin, H., Li, X., Zhang, C., Gao, H., et al. (2021). Linkage analysis, GWAS, transcriptome analysis to identify candidate genes for rice seedlings in response to high temperature stress. BMC Plant Biol. 21:85. doi: 10.1186/s12870-021-02857-2

Wu, A., Allu, A. D., Garapati, P., Siddiqui, H., Dortay, H., Zanor, M. I., et al. (2012). JUNGBRUNNEN1, a reactive oxygen species-responsive NAC transcription factor, regulates longevity in Arabidopsis. Plant Cell 24, 482–506. doi: 10.1105/tpc.111.090894

Wu, S., Guyot, R., Bocs, S., Droc, G., Oktavia, F., Hu, S., et al. (2020a). Structural and functional annotation of transposable elements revealed a potential regulation of genes involved in rubber biosynthesis by TE-derived siRNA interference in Hevea brasiliensis. Int. J. Mol. Sci. 21:4220. doi: 10.3390/ijms21249440

Wu, L., Han, L., Li, Q., Wang, G., Zhang, H., and Li, L. (2020b). Using interactome big data to crack genetic mysteries and enhance future crop breeding. Mol. Plant 14, 77–94. doi: 10.1016/j.molp.2020.12.012

Wu, X., Jin, L., and Xiong, M. (2008). Composite measure of linkage disequilibrium for testing interaction between unlinked loci. Eur. J. Hum. Genet. 16, 644–651. doi: 10.1038/sj.ejhg.5202004

Wu, S., Zhang, M., Yang, X., Peng, F., Zhang, J., Tan, J., et al. (2018). Genome-wide association studies and CRISPR/Cas9-mediated gene editing identify regulatory variants influencing eyebrow thickness in humans. PLoS Genet. 14:e1007640. doi: 10.1371/journal.pgen.1007640

Wydau, S., Ferri-Fioni, M. L., Blanquet, S., and Plateau, P. (2007). GEK1, a gene product of Arabidopsis thaliana involved in ethanol tolerance, is a D-aminoacyl-tRNA deacylase. Nucleic Acids Res. 35, 930–938. doi: 10.1093/nar/gkl1145

Xia, Z., Liu, K., Zhang, S., Yu, W., Zou, M., He, L., et al. (2018). An ultra-high density map allowed for mapping QTL and candidate genes controlling dry latex yield in rubber tree. Ind. Crop. Prod. 120, 351–356. doi: 10.1016/j.indcrop.2018.04.057

Xu, Y., and Crouch, J. H. (2008). Marker-assisted selection in plant breeding: from publications to practice. Crop Sci. 48, 391–407. doi: 10.2135/cropsci2007.04.0191

Yan, Z., Huang, H., Freebern, E., Santos, D. J., Dai, D., Si, J., et al. (2020). Integrating RNA-Seq with GWAS reveals novel insights into the molecular mechanism underpinning ketosis in cattle. BMC Genomics 21:489. doi: 10.1186/s12864-020-06909-z

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. 42, 565–569. doi: 10.1038/ng.608

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

Zeh, M., Casazza, A. P., Kreft, O., Roessner, U., Bieberich, K., Willmitzer, L., et al. (2001). Antisense inhibition of threonine synthase leads to high methionine content in transgenic potato plants. Plant Physiol. 127, 792–802.

Zhang, H., Chu, Y., Dang, P., Tang, Y., Jiang, T., Clevenger, J. P., et al. (2020). Identification of QTLs for resistance to leaf spots in cultivated peanut (Arachis hypogaea L.) through GWAS analysis. Theor. Appl. Genet. 133, 2051–2061. doi: 10.1007/s00122-020-03576-2

Zhang, B., and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. Stat. Appl. Genet. Mol. Biol. 4:7. doi: 10.2202/1544-6115.1128

Zhang, J., Khan, S. A., Heckel, D. G., and Bock, R. (2017). Next-generation insect-resistant plants: RNAi-mediated crop protection. Trends Biotechnol. 35, 871–882. doi: 10.1016/j.tibtech.2017.04.009

Zhang, W., Li, F., and Nie, L. (2010). Integrating multiple ‘omics’ analysis for microbial biology: application and methodologies. Microbiology 156, 287–301. doi: 10.1099/mic.0.034793-0

Zhang, C., Stratopoulos, L. M. F., Pretzsch, H., and Rötzer, T. (2019a). How do Tilia cordata Greenspire trees cope with drought stress regarding their biomass allocation and ecosystem services? Forests 10:676.

Zheng, F., Zhang, S., Churas, C., Pratt, D., Bahar, I., and Ideker, T. (2021). HiDeF: identifying persistent structures in multiscale ‘omics data. Genome Biol. 22:21. doi: 10.1186/s13059-020-02228-4

Zhu, J. K. (2016). Abiotic stress signaling and responses in plants. Cell 167, 313–324. doi: 10.1016/j.cell.2016.08.029

Resumo dos Resultados

7.1 Capítulo I

- Foram testadas as acurácias preditivas (PA) utilizando duas validações cruzadas (CV1 e CV2). CV1 é construída para um único ambiente, seco (LW), ou úmido (WW), simulando o desenvolvimento de novas gerações, sendo assim predizendo novos indivíduos ainda não avaliados. Enquanto que CV2, indivíduos são testados em uma condição (LW ou WW) e preditos no outro ambiente, portanto simula a predição dos genótipos em novos ambientes. Os modelos de GP em um único ambiente (SM, MM, MDS e MDE) não mostraram diferenças significativas e os valores de PA que variaram em torno de 0.19 em LW, tanto utilizando a matriz de parentesco VanRaden (GB) como a Gaussian Kernel (GK). Já em WW em CV1 mostrou um aumento de PA com valores variando em torno de 0.27. Em relação a CV2, PA tem um significante aumento, mostrando valores acima de 0.80 em todos os modelos testados (MM, MDS e MDE) e ambientes (LW e WW), independente da matriz de parentesco (GB ou GK).
- Comparando os ganhos genéticos esperados (EGG) utilizando GS e métodos tradicionais de melhoramento, em todos os cenários a utilização de GS para seringueira se mostra mais vantajosas, sendo que com a inclusão da interação genótipo x ambiente em CV2 os ganhos genéticos podem ser superiores a 5 vezes quando comparados aos métodos tradicionais de seleção.

7.2 Capítulo II

- Foram identificados 4 SNPs significativamente associados ao crescimento (snpsGWAS) em *H. brasiliensis*. Essas marcas mostraram uma variância fenotípica explicada (PVE) entre 2 a 9% e um efeito aditivo que foi de -1 a 0.84 cm. Essas marcas foram anotadas para os genes SBT 4.6, GK1 e IQM 2, que mostram evidências em outras espécies de estarem relacionados com resistência a estresse biótico e abiótico.
- Foram identificadas outras 181 marcas, que estavam significativamente associadas aos snpsGWAS, esses marcadores foram localizados próximos a QTLs envolvidos com crescimento em um mapa de ligação publicado anteriormente. Além disso, esses

marcadores se mostram como *hubs* nos módulos de co-expressão contendo os genes identificados pelos snpsGWAS.

- A rede de co-expressão construída com 30.407 transcritos, apresentou 174 grupos com tamanho que variou entre 52 a 3.823 genes. 5 módulos altamente correlacionados foram selecionados por conterem os snpsGWAS, assim como os snpsLD. Esses módulos apresentaram uma grande quantidade de genes relacionados com crescimento e estresse abiótico e biótico. Interessantemente, foram identificados uma grande proporção de transposons envolvidos principalmente com estresse abiótico.
- A partir dos genes identificados nesses 5 módulos selecionados, foi construída uma rede enzimática, com base no banco de dados KEGG, que possibilitou a identificação enzimas também importantes para o crescimento da planta como UDP-sugar pyrophosphorylase (ec: 2.7.7.64).

Conclusão

Os modelos de seleção genômica aplicados de corretamente pelos programas de melhoramento de *H. brasiliensis* podem contribuir de forma expressiva na tarefa de selecionar precocemente os genótipos superiores de uma população de melhoramento genético, reduzindo o tempo necessário nas etapas de seleção e portanto os custos para obtenção de clones elite. Apesar da predição de novos indivíduos, como simulado em CV1, se mostra tarefa mais complexa do que a predição do desempenho dos genótipos em novos ambientes, como simulado em CV2, ainda assim tais estratégias parecem vantajosas quando comparado ao melhoramento clássico da espécie.

Destacamos também a importância da utilização de dados obtidos de diferentes ômicas para o aprofundamento da compreensão molecular de características complexas como o crescimento. Essa estratégia foi essencial para a compreensão abrangente dos genes envolvidos com o crescimento da seringueira, uma espécie com limitadas informações genômicas.

Perspectivas

Com o advento das ferramentas de NGS é cada vez mais possível a utilização da biologia molecular no auxílio dos programas de melhoramento genético. Para isso é necessário o desenvolvimento e adaptação de metodologias específicas.

Aqui fizemos uso pela primeira vez da seleção genômica em seringueira incluindo a variação causada pela interação genótipo x ambiente. Apesar de atingirmos uma boa acurácia preditiva, novas adaptações deverão ser feitas como a inclusão de maior diversidade genética e utilizações de novas metodologias de predição como o *machine learning*.

Fizemos uma abordagem inovadora de GWAS integrada com a redes biológicas complexas. Tal abordagem foi a primeira descrita para a espécie arbórea em nosso conhecimento. A integração dessas multiomicas foi de grande importância para identificar um grande número de variantes moleculares com potencial para melhor compreensão das variações fenotípicas observadas nessa população. Tal abordagem se mostrou eficiente para a espécie que detém informações moleculares limitadas. Acreditamos que tal abordagem poderá ser utilizada em outras espécies com as mesmas limitações. Investigações futuras ainda deverão ser realizadas para melhor compreendermos elementos essenciais no genoma da seringueira como os transposons e seu papel na evolução e definição de características complexas na espécie.

Referências

Alika, J. E. (1980). Possibilities of early selection in *Hevea brasiliensis*.

An, Z., Zhao, Y., Zhang, X., Huang, X., Hu, Y., Cheng, H., ... & Huang, H. (2019). A high-density genetic map and QTL mapping on growth and latex yield-related traits in *Hevea brasiliensis* Müll. Arg. Industrial Crops and Products, 132, 440-448.

Atwell, S., Huang, Y. S., Vilhjálmsson, B. J., Willems, G., Horton, M., Li, Y., ... & Nordborg, M. (2010). Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. Nature, 465(7298), 627-631.

Bernardo, R. (1994). Prediction of maize single‐cross performance using RFLPs and information from related hybrids. Crop Science, 34(1), 20-25

Bernardo, R. (2016). Bandwagons I, too, have known. Theoretical and applied genetics, 129(12), 2323-2332.

Bhat, J. A., Ali, S., Salgotra, R. K., Mir, Z. A., Dutta, S., Jadon, V., ... & Prabhu, K. (2016). Genomic selection in the era of next generation sequencing for complex traits in plant breeding. Frontiers in genetics, 7, 221.

Carvalho, Y. M. K. (2020). Tecnologias em Controle Biológico: um estudo a partir da produção paulista de borracha natural.

Cerioli, T., Hernandez, C. O., Angira, B., McCouch, S. R., Robbins, K. R., & Famoso, A. N. (2022). Development and validation of an optimized marker set for genomic selection in southern US rice breeding programs. The Plant Genome, e 20219.

Chanroj, V., Rattanawong, R., Phumichai, T., Tangphatsornruang, S., & Ukoskit, K.(2017). Genome-wide association mapping of latex yield and girth in Amazonian accessions of *Hevea brasiliensis* grown in a suboptimal climate zone. Genomics, 109(5-6), 475-484.

Chen, B., Li, Y., Tian, M., Su, H., Sun, W., & Li, Y. (2022). Linkage mapping and QTL analysis of growth traits in *Rhopilema esculentum*. Scientific reports, 12(1), 1-7.

Chen, Z. Q., Zan, Y., Milesi, P., Zhou, L., Chen, J., Li, L., ... & Wu, H. X. (2021). Leveraging breeding programs and genomic data in Norway spruce (Picea abies L. Karst) for GWAS analysis. Genome biology, 22(1), 1-30.

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, 38(9), 1409-1423.

Chow, K. S., Ghazali, A. K., Hoh, C. C., & Mohd-Zainuddin, Z. (2014). RNA sequencing read depth requirement for optimal transcriptome coverage in Hevea brasiliensis. BMC research notes, 7(1), 1-13.

Clément-Demange, A., Legnaté, H., Seguin, M., Carron, M. P., Le Guen, V., Chapuset, T., & Nicolas, D. (2001). Rubber tree. Tropical plant breeding, 455-480.

Collard, B. C., & Mackill, D. J. (2008). Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. Philosophical Transactions of the Royal Society B: Biological Sciences, 363(1491), 557-572.

Conson, A. R., Taniguti, C. H., Amadeu, R. R., Andreotti, I. A., De Souza, L. M., Dos Santos, L. H., ... & De Souza, A. 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, 9, 1255

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., De Los Campos, G., ... & Varshney, R. K. (2017). Genomic selection in plant breeding: methods, models, and perspectives. Trends in plant science, 22(11), 961-975.

Daetwyler, H. D., Pong-Wong, R., Villanueva, B., & Woolliams, J. A. (2010). The impact of genetic architecture on genome-wide evaluation methods. Genetics, 185(3), 1021-1031.

De Faÿ, E., and Jacob, J. L. (1989). “Anatomical organization of the laticiferous system in the bark,” in Physiology of Rubber Tree Latex. eds. J. D’Auzac, J.Jacob and H. Chrestin (Boca Raton, FL: CRC Press), 3–14

de Souza, L. M., Le Guen, V., Cerqueira-Silva, C. B. M., Silva, C. C., Mantello, C. C., Conson, A. R. O., ... & Souza, A. P. D. (2015). 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, 10(7), e0134607.

Dean, W. (1989). A luta pela borracha no Brasil: um estudo de história ecológica. Studio Nobel.

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.

Duarte Jr, A. M. (2015). Fordlândia e Belterra: as cidades de Henry Ford na Amazônia. Revista Brasileira de Casos de Ensino em Administração, c1-c1.

Fajardo, T. V., & Quecini, V. (2021). Comparative transcriptome analyses between cultivated and wild grapes reveal conservation of expressed genes but extensive rewiring of co-expression networks. Plant Molecular Biology, 106(1), 1-20.

Francisco, F. R., Aono, A. H., Da Silva, C. C., Gonçalves, P. S., Junior, E. J. S., Le Guen, V., ... & de Souza, A. P. (2021). Unravelling rubber tree growth by integrating GWAS and biological network-based approaches. Frontiers in plant science, 12.

Goncalves, P. D. S., Unterpertinger, J. P., & Freire, M. R. C. (1986). Simposio sobre a cultura da seringueira no estado de Sao Paulo

Gonçalves, P. D. S., & Fontes, J. R. A. (2009). Domesticação e melhoramento da seringueira. Domesticação e Melhoramento: Espécies Amazônicas. Viçosa, UFV, 395-423.

Gonçalves, P. D. S., & Marques, J. R. B. (2008). Melhoramento genético da seringueira: passado, presente e futuro. Seringueira.

Gonçalves, P. D. S., Cardoso, M., & Bortoletto, N. (1988). Redução do ciclo de melhoramento e seleção na obtenção de cultivares de seringueira. O Agronômico, Campinas, 40(2), 112-130.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., ... & Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature biotechnology, 29(7), 644-652.

Guo, J., Wang, A., Mao, S., Xu, X., Li, J., & Shen, Y. (2022). Construction of high-density genetic linkage map and QTL mapping for growth performance in black carp (Mylopharyngodon piceus). Aquaculture, 549, 737799.

Hayashi, Y. (2009). Production of natural rubber from Para rubber tree. Plant Biotechnology, 26(1), 67-70.

Hora Júnior, B. T. D., de Macedo, D. M., Barreto, R. W., Evans, H. C., Mattos, C. R. R., Maffia, L. A., & Mizubuti, E. S. (2014). Erasing the past: a new identity for the Damoclean pathogen causing South American leaf blight of rubber. PloS one, 9(8), e104750.

IPEF. Instituto de Pesquisa e Estudos Florestais (2007). *Hevea brasiliensis* (seringueira). Disponível e:http://www.ipef.br/identificacao/hevea.brasiliensi

Isik, F., Kumar, S., Martínez-García, P. J., Iwata, H., & Yamamoto, T. (2015). Acceleration of forest and fruit tree domestication by genomic selection. In Advances in botanical research (Vol. 74, pp. 93-124). Academic Press

Jaimes, Y., Rojas, J., Cilas, C., & Furtado, E. L. (2016). Suitable climate for rubber trees affected by the South American Leaf Blight (SALB): Example for identification of escape zones in the Colombian middle Magdalena. Crop Protection, 81, 99-114.

Jaimes, Y., Rojas, J., Cilas, C., & Furtado, E. L. (2016). Suitable climate for rubber trees affected by the South American Leaf Blight (SALB): Example for identification of escape zones in the Colombian middle Magdalena. Crop Protection, 81, 99-114.

Jain, S. M., & Priyadarshan, P. M. (Eds.). (2009). Breeding plantation tree crops: tropical species (Vol. 84). New York: Springer.

Jannink, J. L. (2010). Dynamics of long-term genomic selection. Genetics Selection Evolution, 42(1), 1-11.

Koenig, D., Jiménez-Gómez, J. M., Kimura, S., Fulop, D., Chitwood, D. H., Headland, L. R., ... & Maloof, J. N. (2013). Comparative transcriptomics reveals patterns of selection in

domesticated and wild tomato. Proceedings of the National Academy of Sciences, 110(28), E2655-E2662.

Korte, A., & Farlow, A. (2013). The advantages and limitations of trait analysis with GWAS: a review. Plant methods, 9(1), 1-9.

Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. BMC bioinformatics, 9(1), 559

Lau, N. S., Makita, Y., Kawashima, M., Taylor, T. D., Kondo, S., Othman, A. S., ... & Matsui, M. (2016). The rubber tree genome shows expansion of gene family associated with rubber biosynthesis. Scientific reports, 6(1), 1-14.

Le Guen, V., Doare, F., Weber, C., & Seguin, M. (2009). Genetic structure of Amazonian populations of Hevea brasiliensis is shaped by hydrographical network and isolation by distance. Tree Genetics & Genomes, 5(4), 673-683.

Le Guen, V., Garcia, D., Mattos, C. R. R., Doare, F., Lespinasse, D., & Seguin, M. (2007). Bypassing of a polygenic Microcyclus ulei resistance in rubber tree, analyzed by QTL detection. New phytologist, 173(2), 335-345.

Le Guen, V., Lespinasse, D., Oliver, G., Rodier-Goud, M., Pinard, F., & Seguin, M. (2003). Molecular mapping of genes conferring field resistance to South American Leaf Blight (Microcyclus ulei) in rubber tree. Theoretical and applied genetics, 108(1), 160-167.

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.

Lespinasse, D., Grivet, L., Troispoux, V., Rodier-Goud, M., Pinard, F., & Seguin, M. (2000). Identification of QTLs involved in the resistance to South American leaf blight (Microcyclus ulei) in the rubber tree. Theoretical and applied genetics, 100(6), 975-984.

Liu, J., Shi, C., Shi, C. C., Li, W., Zhang, Q. J., Zhang, Y., ... & Gao, L. Z. (2020). The chromosome-based rubber tree genome provides new insights into spurge genome evolution and rubber biosynthesis. Molecular plant, 13(2), 336-350.

Ma, D., Ding, Q., Guo, Z., Zhao, Z., Wei, L., Li, Y., ... & Zheng, H. L. (2021). Identification, characterization and expression analysis of lineage-specific genes within mangrove species Aegiceras corniculatum. Molecular Genetics and Genomics, 296(6), 1235-1247.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., ... & Visscher, P. M. (2009). Finding the missing heritability of complex diseases. Nature, 461(7265), 747-753.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., ... & Visscher, P. M. (2009). Finding the missing heritability of complex diseases. Nature, 461(7265), 747-753.

Mantello, C. C., Boatwright, L., da Silva, C. C., Scaloppi, E. J., de Souza Goncalves, P., Barbazuk, W. B., et al. (2019). Deep expression analysis reveals distinct cold-response strategies in rubber tree (*Hevea brasiliensis*). BMC Genomics 20:455. doi: 10.1186/s12864-019-5852-5

Mazzaro, L. G. (2014). Coordenação no sistema agroindustrial da borracha natural: um estudo das relações entre produtores rurais e usinas beneficiadoras paulistas (Doctoral dissertation).

McKown, A. D., Klápště, J., Guy, R. D., Geraldes, A., Porth, I., Hannemann, J., ... & Douglas, C. J. (2014). Genome‐wide association implicates numerous genes underlying ecological trait variation in natural populations of Populus trichocarpa. New Phytologist, 203(2), 535-553.

Montesinos-López, O. A., Montesinos-López, A., Pérez-Rodríguez, P., Barrón-López, J. A., Martini, J. W., Fajardo-Flores, S. B., ... & Crossa, J. (2021). A review of deep learning applications for genomic selection. BMC genomics, 22(1), 1-23.

Montoro, P., Wu, S., Favreau, B., Herlinawati, E., Labrune, C., Martin-Magniette, M. L., ... & Ismawanto, S. (2018). Transcriptome analysis in Hevea brasiliensis latex revealed changes in hormone signalling pathways during ethephon stimulation and consequent Tapping Panel Dryness. Scientific Reports, 8(1), 1-12.

Myles, S., Peiffer, J., Brown, P. J., Ersoz, E. S., Zhang, Z., Costich, D. E., & Buckler, E. S. (2009). Association mapping: critical considerations shift from genotyping to experimental design. The Plant Cell, 21(8), 2194-2202.

Nguyen, K. L., Grondin, A., Courtois, B., and Gantet, P. (2019). Next-generationsequencing accelerates crop gene discovery. Trends Plant Sci. 24, 263–274.doi: 10.1016/j.tplants.2018.11.008

Pires, J. M., Secco, R. D. S., & Gomes, J. I. (2002). Taxonomia e fitogeografia das seringueiras (Hevea spp.) (pp. 103-pp). Empresa Amazônia Oriental.

Pootakham, W., Ruang-Areerate, P., Jomchai, N., Sonthirod, C., Sangsrakru, D., Yoocha, T., ... & Tangphatsornruang, S. (2015). Construction of a high-density integrated genetic linkage map of rubber tree (Hevea brasiliensis) using genotyping-by-sequencing (GBS). Frontiers in plant science, 6, 367.

Pootakham, W., Shearman, J. R., & Tangphatsornruang, S. (2020). Development of molecular markers in Hevea brasiliensis for marker-assisted breeding. In The rubber tree genome (pp. 67-79). Springer, Cham.

Pootakham, W., Sonthirod, C., Naktang, C., Ruang-Areerate, P., Yoocha, T.,Sangsrakru, D., et al. (2017). De novo hybrid assembly of the rubber treegenome reveals evidence of paleotetraploidy in Hevea species. Sci. Rep.7:41457. doi: 10.1038/srep41457

Priyadarshan, P. M. (2011). Biology of Hevea rubber (pp. 1-6). Wallingford, UK: CABI.

Priyadarshan, P. M. (2017). Refinements to Hevea rubber breeding. Tree Genetics & Genomes, 13(1), 1-17.

Priyadarshan, P. M., & Clément-Demange, A. (2004). 3 breeding hevea rubber: formal and molecular genetics. Advances in Genetics, 52, 51-116.

Pszczola, M., Strabel, T., Mulder, H. A., & Calus, M. P. L. (2012). Reliability of direct genomic values for animals with different relationships within and to the reference population. Journal of dairy science, 95(1), 389-400.

Rahman, A. Y. A., Usharraj, A. O., Misra, B. B., Thottathil, G. P., Jayasekaran, K., Feng, Y., ... & Alam, M. (2013). Draft genome sequence of the rubber tree *Hevea brasiliensis*. BMC genomics, 14(1), 1-15.

Rao, X., & Dixon, R. A. (2019). Co-expression networks for plant biology: why and how. Acta biochimica et biophysica Sinica, 51(10), 981-988.

Rosa, J. R. B. F., Mantello, C. C., Garcia, D., de Souza, L. M., Da Silva, C. C., Gazaffi, R., ... & Le Guen, V. (2018). QTL detection for growth and latex production in a full-sib rubber tree population cultivated under suboptimal climate conditions. BMC plant biology, 18(1), 1-16.

Sarig, O., & Sprecher, E. (2017). The molecular revolution in cutaneous biology: Era of next-generation sequencing. Journal of Investigative Dermatology, 137(5), e79-e82.

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.

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.

Sohi, H. S., Gill, M. I. S., Chhuneja, P., Arora, N. K., Maan, S. S., & Singh, J. (2022). Construction of Genetic Linkage Map and Mapping QTL Specific to Leaf Anthocyanin Colouration in Mapping Population ‘Allahabad Safeda’×‘Purple Guava (Local)’of Guava (Psidium guajava L.). Plants, 11(15), 2014.

Souza, L. M., Gazaffi, R., Mantello, C. C., Silva, C. C., Garcia, D., Le Guen, V., ... & Souza, A. P. (2013). QTL mapping of growth-related traits in a full-sib family of rubber tree (Hevea brasiliensis) evaluated in a sub-tropical climate. PLoS One, 8(4), e61238.

Souza, L. M., Francisco, F. R., Gonçalves, P. S., Scaloppi Junior, E. J., Le Guen, V., Fritsche-Neto, R., & Souza, A. P. (2019). Genomic selection in rubber tree breeding: a comparison of models and methods for managing G× E interactions. Frontiers in plant science, 1353.

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., & Meyre, D. (2019). Benefits and limitations of genome-wide association studies. Nature Reviews Genetics, 20(8), 467-484.

Tang, C., Yang, M., Fang, Y., Luo, Y., Gao, S., Xiao, X., ... & Huang, H. (2016). The rubber tree genome reveals new insights into rubber production and species adaptation. Nature plants, 2(6), 1-10.

Vivek, B. S., Krishna, G. K., Vengadessan, V., Babu, R., Zaidi, P. H., Kha, L. Q., ... & Crossa, J. (2017). Use of genomic estimated breeding values results in rapid genetic gains for drought tolerance in maize. The Plant Genome, 10(1), plantgenome2016-07.

Weiss, M., Sniezko, R. A., Puiu, D., Crepeau, M. W., Stevens, K., Salzberg, S. L., ... & De La Torre, A. R. (2020). Genomic basis of white pine blister rust quantitative disease resistance and its relationship with qualitative resistance. The Plant Journal, 104(2), 365-376.

Wu, W., Zhang, X., Deng, Z., An, Z., Huang, H., Li, W., & Cheng, H. (2022). Ultrahigh-density genetic map construction and identification of quantitative trait loci for growth in rubber tree (Hevea brasiliensis). Industrial Crops and Products, 178, 114560.

Xia, Z., Liu, K., Zhang, S., Yu, W., Zou, M., He, L., & Wang, W. (2018). An ultra-high density map allowed for mapping QTL and candidate genes controlling dry latex yield in rubber tree. Industrial Crops and Products, 120, 351-356.

Zargar, S. M., Raatz, B., Sonah, H., Bhat, J. A., Dar, Z. A., Agrawal, G. K., & Rakwal, R. (2015). Recent advances in molecular marker techniques: insight into QTL mapping, GWAS and genomic selection in plants. Journal of crop science and biotechnology, 18(5), 293-308.

Zhao, K., Tung, C. W., Eizenga, G. C., Wright, M. H., Ali, M. L., Price, A. H., ... & McCouch, S. R. (2011). Genome-wide association mapping reveals a rich genetic architecture of complex traits in Oryza sativa. Nature communications, 2(1), 1-10.

ANEXOS

8.1 Anexo

Machine learning for crop science: applications and perspectives in maize breeding

ALEXANDRE HILD AONO, RICARDO JOSÉ GONZAGA PIMENTA, **FELIPE ROBERTO FRANCISCO**, ANETE PEREIRA DE SOUZA, ANA CAROLINA LORENA

http://rbms.cnpms.embrapa.br/index.php/ojs/article/view/1257

Revista Brasileira de Milho e Sorgo, v. 21, e1257, **2022.**

Revista Brasileira de Milho e Sorgo
Brazilian Journal of Maize and Sorghum

ISSN 1980 - 6477
Journal homepage: www.abms.org.br/site/paginas

Alexandre Hild Aono[1](✉), Ricardo José Gonzaga Pimenta[1], Felipe Roberto Francisco[1], Anete Pereira de Souza[1,2] and Ana Carolina Lorena[3]

[1] Molecular Biology and Genetic Engineering Center (CBMEG), University of Campinas (UNICAMP), Campinas, SP, Brazil.
E-mail:alexandre.aono@gmail.com, ricardojgpimenta@gmail.com, felipe.roberto.francisco@gmail.com, anete@unicamp.br.

[2] Department of Plant Biology, Institute of Biology (IB), University of Campinas (UNICAMP), Campinas, SP, Brazil.
E-mail: anete@unicamp.br

[3] Technological Institute of Aeronautics, São José dos Campos, SP, Brazil
E-mail: aclorena@ita.br

✉ Corresponding author

How to cite

AONO, A. H.; PIMENTA, R. J. G.; FRANCISCO, F. R.;SOUZA, A. P.; LORENA, A. C. Machine learning for crop science: applications and perspectives in maize breeding. **Revista Brasileira de Milho e Sorgo**, v. 21, e1257, 2022.

MACHINE LEARNING FOR CROP SCIENCE: APPLICATIONS AND PERSPECTIVES IN MAIZE BREEDING

Abstract – Machine learning (ML) has been a major driver in complex data analysis in recent decades, allowing the mining of large databases. ML techniques allow the creation of computational models for prediction, pattern extraction and recognition, considering the premise that the computer acquires learning skills to perform a given task without being explicitly programmed for such a purpose. Driven by the efficiency of these techniques, several studies have demonstrated their wide range of applications and high potential for maize breeding. From the prediction of genetic values by omic data to applications of high-throughput phenotyping data, ML models have promoted advances in the species comprehension and assisted in the development of more effective tools for its breeding, driving expressive yield gains. In this context, this work presents the main contributions of ML in maize breeding, providing a broad view of the main studies and methodological perspectives in the area.

Keywords: Artificial intelligence, deep learning, high-throughput phenotyping, omics-based prediction

APRENDIZADO DE MÁQUINA NA AGRICULTURA: APLICAÇÕES E PERSPECTIVAS NO MELHORAMENTO DE MILHO

Resumo - O aprendizado de máquina (AM) tem sido um impulsionador na análise de dados complexos nas últimas décadas, permitindo a mineração de amplos bancos de dados. Técnicas de AM permitem a criação de modelos computacionais para predição, extração e reconhecimento de padrões, considerando a premissa de que o computador adquire habilidades de aprendizado para realizar uma dada tarefa sem ser explicitamente programado para tal. Impulsionados pela eficiência de tais técnicas, diversos estudos têm demonstrado a ampla gama de aplicações e elevado potencial no melhoramento de milho. Desde a predição de valores genéticos por dados ômicos a aplicações de tecnologias para fenotipagem de alto desempenho, modelos de AM vêm promovendo avanços no conhecimento da espécie e auxiliando no desenvolvimento de ferramentas mais efetivas para seu melhoramento, impulsionando ganhos produtivos expressivos. Nesse contexto, neste trabalho são apresentadas as principais contribuições do AM no melhoramento de milho, fornecendo uma ampla visão dos principais estudos realizados e perspectivas metodológicas na área.

Palavras-chave: Inteligência artificial, aprendizagem profunda, fenotipagem de alto rendimento, predição baseada em ômicas

DOI: https://doi.org/10.18512/rbms2022vol21e1257

8.2 Anexo

A divide-and-conquer approach for genomic prediction in rubber tree using machine learning.

Alexandre Hild Aono, **Felipe Roberto Francisco**, Livia Moura Souza, Paulo de Souza Gonçalves, Erivaldo J. Scaloppi Junior, Vincent Le Guen, Roberto Fritsche-Neto, Gregor Gorjanc, Marcos Gonçalves Quiles, Anete Pereira de Souza

doi: https://doi.org/10.1101/2022.03.30.486381

Scientific Reports: Published 26 October 2022

https://www.nature.com/articles/s41598-022-20416-z#Sec17

www.nature.com/scientificreports

scientific reports

Check for updates

OPEN

A divide-and-conquer approach for genomic prediction in rubber tree using machine learning

Alexandre Hild Aono[1,2], Felipe Roberto Francisco[1], Livia Moura Souza[1,3], Paulo de Souza Gonçalves[4], Erivaldo J. Scaloppi Junior[4], Vincent Le Guen[5,6], Roberto Fritsche-Neto[7], Gregor Gorjanc[2], Marcos Gonçalves Quiles[8] & Anete Pereira de Souza[1,9]✉

Rubber tree (*Hevea brasiliensis*) is the main feedstock for commercial rubber; however, its long vegetative cycle has hindered the development of more productive varieties via breeding programs. With the availability of *H. brasiliensis* genomic data, several linkage maps with associated quantitative trait loci have been constructed and suggested as a tool for marker-assisted selection. Nonetheless, novel genomic strategies are still needed, and genomic selection (GS) may facilitate rubber tree breeding programs aimed at reducing the required cycles for performance assessment. Even though such a methodology has already been shown to be a promising tool for rubber tree breeding, increased model predictive capabilities and practical application are still needed. Here, we developed a novel machine learning-based approach for predicting rubber tree stem circumference based on molecular markers. Through a divide-and-conquer strategy, we propose a neural network prediction system with two stages: (1) subpopulation prediction and (2) phenotype estimation. This approach yielded higher accuracies than traditional statistical models in a single-environment scenario. By delivering large accuracy improvements, our methodology represents a powerful tool for use in *Hevea* GS strategies. Therefore, the incorporation of machine learning techniques into rubber tree GS represents an opportunity to build more robust models and optimize *Hevea* breeding programs.

Rubber tree (*Hevea brasiliensis*) has an elevated importance in the global economy, being almost the only feedstock for commercial rubber[1,2]. Considering the long perennial vegetative cycle of *Hevea*, breeding programs aim to improve its yield production in order to reach the rapidly increasing rubber demand[1–3]. Therefore, genomic approaches are needed in rubber tree breeding, especially considering its recent domestication history[4]. *H. brasiliensis* is a diploid species ($2n = 36$) with an elevated occurrence of duplicated regions in its genome ($\sim 70\%$)[5–7], and this complex genomic organization has hindered the development of genomic strategies for breeding. However, with the improvement of next-generation sequencing (NGS) technologies and the consequent reduction in genotyping costs, data generation has become more efficient, providing more genomic resources in less time and with lower associated costs[8]. This greater availability of data improved precision in selection with higher genetic gains in various crops[8,9] and, in rubber tree, could complement traditional approaches based on only phenotypic and pedigree information[8,10].

Various rubber tree genomic resources have become available in recent decades, such as a large set of different molecular markers[11–14], draft genomes[5,6], and, more recently, a chromosome-level assembled genome[7]. These data have already allowed the construction of saturated linkage maps with associated quantitative trait loci (QTLs), which were proposed as a tool for marker-assisted selection (MAS)[15]. Although QTLs for several traits have been identified in rubber tree[4,15–20], the amount of phenotypic variance explained by these identified QTLs is usually

[1]Molecular Biology and Genetic Engineering Center (CBMEG), University of Campinas (UNICAMP), Campinas, SP, Brazil. [2]The Roslin Institute and Royal (Dick) School of Veterinary Studies, The University of Edinburgh, Midlothian, UK. [3]São Francisco University (USF), Itatiba, Brazil. [4]Center of Rubber Tree and Agroforestry Systems, Agronomic Institute (IAC), Votuporanga, Brazil. [5]Centre de Coopération Internationale en Recherche Agronomique pour le Développement (CIRAD), UMR AGAP, 34398 Montpellier, France. [6]AGAP, CIRAD, INRAE, Institut Agro, Univ Montpellier, Montpellier, France. [7]Genetics Department, Luiz de Queiroz College of Agriculture (ESALQ), University of São Paulo (USP), Piracicaba, SP, Brazil. [8]Instituto de Ciência e Tecnologia (ICT), Universidade Federal de São Paulo (UNIFESP), São José dos Campos, SP, Brazil. [9]Department of Plant Biology, Institute of Biology (IB), University of Campinas (UNICAMP), Campinas, SP, Brazil. ✉email: anete@unicamp.br

8.3 Anexo

The rubber tree kinome: genome-wide characterization and insights into coexpression patterns associated with abiotic stress responses

Lucas Borges dos Santos, Alexandre Hild Aono, **Felipe Roberto Francisco**, Carla Cristina da Silva, Livia Moura Souza, Anete Pereira de Souza

Front. Plant Sci., 07 February 2023
Sec. Functional and Applied Plant Genomics
Volume 14 - 2023 | https://doi.org/10.3389/fpls.2023.1068202

https://www.frontiersin.org/articles/10.3389/fpls.2023.1068202/full

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 07 February 2023
DOI 10.3389/fpls.2023.1068202

OPEN ACCESS

EDITED BY
Venkateswara Rao,
University of Hyderabad, India

REVIEWED BY
Prafull Salvi,
National Agri-Food Biotechnology Institute, India
Ranay Mohan Yadav,
University of Hyderabad, India

*CORRESPONDENCE
Anete Pereira de Souza
anete@unicamp.br

†These authors contributed equally to this work and share first authorship

SPECIALTY SECTION
This article was submitted to Functional and Applied Plant Genomics, a section of the journal Frontiers in Plant Science

RECEIVED 12 October 2022
ACCEPTED 18 January 2023
PUBLISHED 07 February 2023

CITATION
Santos LB, Aono AH, Francisco FR, da Silva CC, Souza LM and Souza AP (2023) The rubber tree kinome: Genome-wide characterization and insights into coexpression patterns associated with abiotic stress responses.
Front. Plant Sci. 14:1068202.
doi: 10.3389/fpls.2023.1068202

The rubber tree kinome: Genome-wide characterization and insights into coexpression patterns associated with abiotic stress responses

Lucas Borges dos Santos[1†], Alexandre Hild Aono[1†], Felipe Roberto Francisco[1†], Carla Cristina da Silva[1], Livia Moura Souza[1,2] and Anete Pereira de Souza[1,3*]

[1]Center for Molecular Biology and Genetic Engineering, State University of Campinas, Campinas, Brazil, [2]São Francisco University (USF), Itatiba, Brazil, [3]Department of Plant Biology, Biology Institute, University of Campinas (UNICAMP), Campinas, Brazil

The protein kinase (PK) superfamily constitutes one of the largest and most conserved protein families in eukaryotic genomes, comprising core components of signaling pathways in cell regulation. Despite its remarkable relevance, only a few kinase families have been studied in *Hevea brasiliensis*. A comprehensive characterization and global expression analysis of the PK superfamily, however, is currently lacking. In this study, with the aim of providing novel inferences about the mechanisms associated with the stress response developed by PKs and retained throughout evolution, we identified and characterized the entire set of PKs, also known as the kinome, present in the Hevea genome. Different RNA-sequencing datasets were employed to identify tissue-specific expression patterns and potential correspondences between different rubber tree genotypes. In addition, coexpression networks under several abiotic stress conditions, such as cold, drought and latex overexploitation, were employed to elucidate associations between families and tissues/stresses. A total of 1,809 PK genes were identified using the current reference genome assembly at the scaffold level, and 1,379 PK genes were identified using the latest chromosome-level assembly and combined into a single set of 2,842 PKs. These proteins were further classified into 20 different groups and 122 families, exhibiting high compositional similarities among family members and with two phylogenetically close species *Manihot esculenta* and *Ricinus communis*. Through the joint investigation of tandemly duplicated kinases, transposable elements, gene expression patterns, and coexpression events, we provided insights into the understanding of the cell regulation mechanisms in response to several conditions, which can often lead to a significant reduction in rubber yield.

KEYWORDS

coexpression networks, *Hevea brasiliensis*, kinase family, RNA-sequencing, tandem duplications, transposanable elements

8.4 Anexo

Novel insights into the cold resistance of Hevea brasiliensis through coexpression networks

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

doi: https://doi.org/10.1101/2021.11.02.466997

https://www.biorxiv.org/content/10.1101/2021.11.02.466997v2.abstract

Novel insights into the cold resistance of *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

Number of Words:10,851

Number of Figures: 4

Number of Tables: 1

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 source of natural rubber worldwide. Due to the high pressure of fungal diseases in hot, humid regions, rubber plantations have been moved to "escape areas", which are dryer and have lower temperatures during the winter. Here, we combined gene expression data of a primary (GT1) and a secondary (RRIM600) young rubber tree clones, which present different cold tolerance strategies, to analyze rubber tree gene expression regulation during 24 h of cold exposure (10°C). Together with traditional differential expression approaches, a RNA sequencing (RNA-seq) gene coexpression network (GCN) comprising 27,220 genes was established in which the genes were grouped into 832 clusters. In the GCN, most of the rubber tree molecular responses to cold stress were grouped in 26 clusters, which were divided into three GCN modules: a downregulated group comprising 12 clusters and two upregulated groups comprising eleven and three clusters. Considering the three modules identified, the general *Hevea* response to short-term cold exposure involved downregulation of gibberellin (GA) signaling, complex regulation of jasmonic acid (JA) stress responses and programmed cell death (PCD) and upregulation of ethylene responsive genes. The hub genes of the cold-responsive modules were subsequently identified and analyzed. As a result of the GCN strategy applied in this study, we could not only access individual DEGs related to the *Hevea* cold response, but also provide insights into a deeper cascade of associated mechanisms involved in the response to cold stress in young rubber

Material suplementar:

https://www.biorxiv.org/content/10.1101/2021.11.02.466997v1.supplementary-material

8.5 Anexo

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

Stephanie Karenina Bajay[a,†], Alexandre Hild Aono[a,†], **Felipe Roberto Francisco**[a‡], Gary Peter[b], Andrew Sims[b], Matias Kirst[b] and Anete Pereira de Souza[a,c,*]

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

[b]*School of Forest Resources and Conservation (SFRC), University of Florida (UF), Gainesville, FL, United States*

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

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

*Corresponding author

doi: https://doi.org/10.1101/2022.12.06.519371

https://www.biorxiv.org/content/10.1101/2022.12.06.519371v1

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

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

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

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

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

*Corresponding author

Abstract

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

8.6 Anexo: Declaração Bioética e Biossegurança

UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE BIOLOGIA, www.ib.unicamp.br
Rua Monteiro Lobato, 255 CEP: 13083-862 Campinas-SP

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:
"Ferramentas genômicas e moleculares para caracterização genética e predição genômica em seringueira."
Desenvolvida no Programa de Pós-Graduação de Genética e Biologia Molecular do Instituto de Biologia da Unicamp, não versa sobre pesquisa envolvendo seres humanos, animais ou temas afetos a Biossegurança.

Campinas, 20 de Março 2023

Assinatura: Felipe Roberto Francisco
Nome do(a) autor(a): **Felipe Roberto Francisco**
RG nº: 49522230-6

Assinatura:
Nome do(a) orientador(a): **Anete Pereira de Souza**
RG nº: 8680325-6

8.7 Anexo: Declaração de Autoria

DECLARAÇÃO DE AUTORIA

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 Tese de Doutorado, intitulada **"Ferramentas genômicas e moleculares para caracterização genética e predição genômica em seringueira."**, não infringem os dispositivos da Lei nº9. 610/98, nem o direito autoral de qualquer editora.

Campinas, 20 de Março 2023

Assinatura:
Nome do(a) autor(a): **Felipe Roberto Francisco**
RG nº: 49522230-6

Assinatura:
Nome do(a) orientador(a): **Anete Pereira de Souza**
RG nº: 8680325-6