UNIVERSIDADE ESTADUAL PAULISTA – UNESP

CÂMPUS DE JABOTICABAL

ESTUDO GENÔMICO COMPARATIVO EM *Theobroma* AMAZÔNICOS: IDENTIFICAÇÃO DE GENES ASSOCIADOS À QUALIDADE DOS FRUTOS EM *T. grandiflorum, T. cacao* E *T. umbraticum*

Rafaely Pantoja Oliveira

Engenheira Florestal

2025

UNIVERSIDADE ESTADUAL PAULISTA – UNESP

CÂMPUS DE JABOTICABAL

ESTUDO GENÔMICO COMPARATIVO EM *Theobroma* AMAZÔNICOS: IDENTIFICAÇÃO DE GENES ASSOCIADOS À QUALIDADE DOS FRUTOS EM *T. grandiflorum, T. cacao* E *T. umbraticum*

Rafaely Pantoja Oliveira

Orientador: Prof. Dr. Alessandro de Mello Varani

Tese apresentada à Faculdade de Ciências Agrárias e Veterinárias – Unesp, Câmpus de Jaboticabal, como parte das exigências para a obtenção do título de Doutora em Agronomia (Genética e Melhoramento de Plantas)

2025

O48e

Oliveira, Rafaely Pantoja

Estudo genômico comparativo em theobroma amazônicos: Identificação de genes associados à qualidade dos frutos em T. grandiflorum, T. cacao e T. umbraticum / Rafaely Pantoja Oliveira. -- Jaboticabal, 2025

142 p.

Tese (doutorado) - Universidade Estadual Paulista (UNESP), Faculdade de Ciências Agrárias e Veterinárias, Jaboticabal

Orientador: Alessandro de Mello Varani

1. Biodiversidade amazônica. 2. Sequenciamento genômico. 3. Melhoramento genético. 4. Cobra-likes. 5. Domesticação de espécies. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Dados fornecidos pelo autor(a).

Registro de Impacto

Este estudo genômico comparativo em espécies amazônicas de *Theobroma* contribui para o avanço da agricultura sustentável, ao identificar genes associados à qualidade dos frutos, com potencial aplicação em programas de melhoramento genético.

UNIVERSIDADE ESTADUAL PAULISTA

Câmpus de Jaboticabal

CERTIFICADO DE APROVAÇÃO

TÍTULO DA TESE: ESTUDO GENÔMICO COMPARATIVO EM *Theobroma* AMAZÔNICOS: IDENTIFICAÇÃO DE GENES ASSOCIADOS À QUALIDADE DOS FRUTOS EM *T. grandiflorum, T. cacao E T. umbraticum*

AUTORA: RAFAELY PANTOJA OLIVEIRA
ORIENTADOR: ALESSANDRO DE MELLO VARANI

Aprovada como parte das exigências para obtenção do Título de Doutora em Agronomia (Genética e Melhoramento de Plantas), pela Comissão Examinadora:

Prof. Dr. ALESSANDRO DE MELLO VARANI (Participaçao Virtual)
Departamento de Biotecnologia Agropecuaria e Ambiental / FCAV UNESP Jaboticabal

gov.br Documento assinado digitalmente
VINICIUS AUGUSTO CARVALHO DE ABREU
Data: 28/02/2025 14:49:18-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. VINICIUS AUGUSTO CARVALHO DE ABREU (Participaçao Virtual)
Departamento de Ciências Exatas e Naturais / Universidade Federal do Pará (UFPA) - Belém/PA

gov.br Documento assinado digitalmente
VITOR FERNANDES OLIVEIRA DE MIRANDA
Data: 06/03/2025 21:27:56-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. VITOR FERNANDES OLIVEIRA DE MIRANDA (Participaçao Virtual)
Departamento de Biologia / FCAV UNESP Jaboticabal

gov.br Documento assinado digitalmente
ALESSANDRA JACKELINE GUEDES DE MORAES
Data: 28/02/2025 14:21:59-0300

Profa. Dra. ALESSANDRA JACKELINE GUEDES DE MORAES (Participaçao Virtual)
Departamento de Tecnologia de Madeira / Universidade do Estado do Pará - Belém/PA

gov.br Documento assinado digitalmente
SAURA RODRIGUES DA SILVA
Data: 06/03/2025 21:02:16-0300
Verifique em https://validar.iti.gov.br

Pós-Doutoranda SAURA RODRIGUES DA SILVA (Participaçao Virtual)
Departamento de Biologia Aplicada à Agropecuária / FCAV UNESP Jaboticabal

Jaboticabal, 26 de fevereiro de 2025

gov.br Documento assinado digitalmente
ALESSANDRO DE MELLO VARANI
Data: 28/02/2025 13:50:03-0300
Verifique em https://validar.iti.gov.br

Dados Curriculares do Autor

Rafaely Pantoja Oliveira, nascida em Castanhal, PA. É graduada em Engenharia Florestal pela Universidade Federal Rural da Amazônia (UFRA), onde concluiu seu curso em 2018. Durante a graduação, realizou estágio no Laboratório de Biotecnologia e Recursos Vegetais da Embrapa Amazônia Oriental, em Belém, atuando na área de propagação de plantas. Desenvolveu trabalhos com *Piper nigrum* L. por meio de cultura de tecidos, utilizando métodos de micropropagação. Após a graduação, ingressou no programa de pós-graduação em Ciências Florestais na mesma instituição, UFRA, onde obteve o título de mestra em 2021 com a pesquisa "Estudo Comparativo da Anatomia e Densidade Básica da Madeira das Variedades de *Schizolobium parahyba* (Vell.) S. F. Blake em um Plantio na Amazônia". Em termos de publicações científicas, é coautora do artigo de alto impacto intitulado "Genomic decoding of *Theobroma grandiflorum* (*cupuassu*) at chromosomal scale: evolutionary insights for horticultural innovation," publicado na GigaScience, v. 13, p. 1-19, em 2024, ao lado de importantes pesquisadores. Ao longo da sua trajetória acadêmica, contribuiu para projetos de pesquisa voltados para a sustentabilidade e o manejo florestal, promovendo técnicas e práticas que visam a conservação dos recursos naturais na região amazônica. Além disso, sua atuação em pesquisas de genômica comparativa entre espécies do gênero *Theobroma* tem auxiliado na identificação de genes relacionados à qualidade dos frutos e características agronômicas, ampliando o conhecimento sobre a biodiversidade e o potencial produtivo dessas espécies amazônicas.

Dedico esta tese à minha avó, Maria Conceição, que, em seus momentos de lucidez, sonhou junto comigo esta conquista.

Seu amor, força e fé silenciosa seguem sendo minha inspiração mais profunda.

Agradecimentos

Agradeço à **Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES)** pelo apoio financeiro por meio da concessão da bolsa de estudos, fundamental para a realização da pesquisa.

À **Universidade Estadual Paulista "Júlio de Mesquita Filho" (UNESP)** e ao **Programa de Pós-Graduação em Agronomia (Genética e Melhoramento de Plantas)**, pela sólida formação acadêmica e pelas oportunidades que contribuíram para meu crescimento profissional e intelectual.

À minha família, minha base e meu alicerce, deixo minha gratidão mais profunda: à minha mãe **Natalina** e ao meu pai **Antônio**, por todo o amor, apoio incondicional e incentivo em cada etapa da minha vida acadêmica; aos meus irmãos, **Rafaela** e **Rafael**, pelo carinho, parceria e palavras de encorajamento; e às minhas sobrinhas, **Maria Cecília** e **Júlia**, que alegram meus dias com afeto e doçura.

Ao meu orientador, **Dr. Alessandro Varani**, agradeço imensamente pela orientação dedicada, pela confiança depositada em meu trabalho e por todo o aprendizado compartilhado ao longo desta jornada. Sou profundamente grata por ter sido acolhida em seu laboratório como sua aluna de doutorado, pelas trocas de conhecimentos valiosas e, principalmente, pela paciência e humanidade com que sempre me guiou — demonstrando empatia, sensibilidade e respeito em cada etapa do caminho.

Agradeço aos meus amigos e colega de laboratório, **João Victor e Mauro**, pela colaboração direta neste trabalho, pela parceria diária e pelo companheirismo. A todos os demais amigos e colegas de laboratório, deixo meu sincero reconhecimento pelo convívio, pelo apoio mútuo e pelas trocas que tornaram este percurso mais leve e significativo.

Agradeço profundamente a todos os **professores da UNESP-FCAV** pelo conhecimento compartilhado, pela dedicação ao ensino e pelo ambiente acadêmico

inspirador que proporcionaram ao longo da minha formação. Em especial, expresso minha sincera gratidão ao Prof. Dr. Vitor Fernandes Oliveira de Miranda, por suas valiosas contribuições acadêmicas, e à Pós-Doutoranda Saura Rodrigues da Silva, pelas trocas enriquecedoras de conhecimento.

Agradeço à **Embrapa Amazônia Oriental**, em nome do Dr. Rafael Moysés Alves, e à **Universidade Federal do Pará (UFPA)**, em nome do Prof. Dr. Vinicius Augusto Carvalho de Abreu, pelas contribuições fundamentais no desenvolvimento do projeto do genoma do cupuaçu. Suas colaborações científicas, apoio institucional e compromisso com a pesquisa foram essenciais para a realização do sequenciamento genômico de uma das espécies frutíferas mais importantes da Amazônia, o cupuaçu, com grande relevância para a biodiversidade e o desenvolvimento sustentável da região.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

SUMÁRIO

CAPÍTULO 1 - Considerações Gerais 8

1. INTRODUÇÃO 8

2. REVISÃO DE LITERATURA 14

2.1. Introdução à família Malvaceae e a importância dos frutos de *Theobroma grandiflorum, T. cacao* e *T. umbraticum* 14

2.2. Aspectos de doença e controle: *Moniliophthora perniciosa* e *Moniliophthora roreri* 18

2.2.1. Mecanismos de resistência e suscetibilidade das plantas do gênero *Theobroma* em relação a *M. perniciosa* e *M. roreri* 23

2.3. Domesticação de *Theobroma* 28

2.4. Genômica do *Theobroma* 33

2.5. Genes relacionados à qualidade dos frutos e sua importância 49

2.6. A família de genes COBRA-LIKE (COBLs) e seu provável papel na regulação da parede celular 61

2.7. Justificativa e hipótese 66

3 REFERÊNCIAS 68

CAPÍTULO 2 - Diversificação Funcional dos Genes COBRA-LIKE (COBLs) em *Theobroma* Amazônicos: Estudo Comparativo entre *T. grandiflorum*, *T. cacao* **e** *T. umbraticum* 85

1. INTRODUÇÃO 87

2. MATERIAIS E MÉTODOS 91

2.1 Genoma selecionado para estudo 91

2.2. Busca por genes e mapeamento dos genes nos cromossomos 92

2.3. Expressão gênica em transcritos por milhão (TPM) 93

2.4. Visualização da estrutura gênica e análise de domínios 94

2.5. Análises Filogenéticas 94

2.6. Modelagem molecular e análise estrutural das proteínas COBRA-Like nas espécies estudadas 95

3. RESULTADOS E DISCUSSÃO 96

3.1. Padrões de distribuição e expressão (TPM) de genes COBRA-Like em *T. grandiflorum, T. cacao* e *T. umbraticum*: predominância de duplicações em tandem sugerindo adaptação funcional 96

3.2. Domínio e estrutura da família gênica COBRA-Likes: conservação evolutiva e variação no tamanho das proteínas 105

3.3. Análises filogenéticas indicam que genes COBRA-Like exclusivos de *Theobroma* apresentam padrões estruturais distintos e possíveis adaptações funcionais 108

4. REFERÊNCIAS 115

CAPÍTULO 3 – Considerações finais e perspectivas 122

APÊNDICES 124

ESTUDO GENÔMICO COMPARATIVO EM *Theobroma* AMAZÔNICOS: IDENTIFICAÇÃO DE GENES ASSOCIADOS À QUALIDADE DOS FRUTOS EM *T. grandiflorum, T. cacao* E *T. umbraticum*

RESUMO - O cupuaçuzeiro (*Theobroma grandiflorum*), membro da família Malvaceae, é uma das espécies mais populares da Amazônia brasileira. A polpa dos frutos do cupuaçu é muito apreciada para o preparo de sucos, doces, licores, geleias, sorvetes e outros produtos alimentícios. As sementes são utilizadas na fabricação de chocolate em pó, tabletes (cupulate) e na indústria de cosméticos. Apesar dos produtores da fruta aproveitarem cada vez mais de seus recursos, incentivando uma cadeia produtiva na região Amazônica, a cultura do cupuaçu ainda é pouco explorada e seus produtos são pouco conhecidos em abrangência nacional. Em adição, diversas doenças afetam a cultura, dentre elas a vassoura-de-bruxa, ocasionada pelo fungo *Moniliophthora perniciosa* e a monilíase, causada pelo fungo *Moniliophthora roreri,* causando grandes perdas de produtividade. Nesse sentido, programas de melhoramento já alcançaram genótipos resistentes e mais produtivos, porém a base molecular de resistência a estes patógenos, bem como a exploração de genes relacionados a qualidade dos frutos ainda são incipientes. Por exemplo, embora o cupuaçu pertença ao mesmo gênero do cacaueiro (*Theobroma cacao*), espécie da qual se obtém o chocolate, e que foi amplamente estudada sob o ponto de vista molecular, genômico e funcional no decorrer da última década, o cupuaçu ainda carece de estudos aprofundados que explorem sua base molecular e genômica, sendo esses conhecimentos fundamentais para subsidiar estratégias de melhoramento genético mais eficientes e direcionadas para a qualidade dos frutos e valorização de características agronômicas e industriais. O cacaueiro, por outro lado, é amplamente cultivado e sua importância econômica global está atrelada à produção de cacau, fundamental para a indústria do chocolate. Os esforços científicos no estudo do *Theobroma cacao* resultaram no sequenciamento completo de seu genoma, revelando uma riqueza de informações sobre genes relacionados à qualidade do fruto, resistência a doenças, e processos de maturação, o que tem fornecido uma base sólida para o melhoramento genético dessa espécie. Isso torna o cupuaçu um excelente candidato para estudos moleculares e de genômica comparativa, principalmente frente à sua espécie congênere, o cacaueiro. Além dessas espécies, outro membro relevante da família Malvaceae é *Theobroma umbraticum,* uma espécie menos conhecida e que, assim como o cupuaçu e o cacau, tem potencial para estudos comparativos. *Theobroma umbraticum* também pertence à subfamília Byttnerioideae, e seus frutos compartilham características semelhantes com *T. cacao*, embora a espécie seja subutilizada e pouco explorada em termos econômicos e científicos. A análise comparativa entre *Theobroma grandiflorum, Theobroma cacao* e *Theobroma umbraticum* pode contribuir para a compreensão da evolução e diversificação dos genes responsáveis por características importantes referentes a qualidade de frutos, podendo oferecer novas informações para a conservação e exploração sustentável da biodiversidade amazônica.

Palavras-chave: Biodiversidade amazônica, Sequenciamento genômico, Melhoramento genético, Cobra-likes, Domesticação de espécies.

COMPARATIVE GENOMIC STUDY OF AMAZONIAN *Theobroma* SPECIES: IDENTIFICATION OF GENES ASSOCIATED WITH FRUIT QUALITY IN *T. grandiflorum*, *T. cacao*, AND *T. umbraticum*

ABSTRACT - The cupuaçu tree (*Theobroma grandiflorum*), a member of the Malvaceae family, is one of the most popular species in the Brazilian Amazon. The pulp of cupuaçu fruits is highly valued for the preparation of juices, sweets, liqueurs, jams, ice creams, and other food products. The seeds are used to produce cocoa powder, chocolate bars (cupulate), and in the cosmetics industry. Despite the increasing utilization of its resources by fruit producers, encouraging a productive chain in the Amazon region, the cultivation of cupuaçu remains underexplored, and its products are little known on a national scale. In addition, various diseases affect the crop, including witches' broom, caused by the fungus *Moniliophthora perniciosa*, and moniliasis, caused by the fungus *Moniliophthora roreri*, leading to significant productivity losses. In this context, breeding programs have already achieved more resistant and productive genotypes. However, the molecular basis of resistance to these pathogens, as well as the exploration of genes related to fruit quality, remains in its infancy. For example, although cupuaçu belongs to the same genus as *cacao* (*Theobroma cacao*), a species from which chocolate is derived and which has been extensively studied from molecular, genomic, and functional perspectives over the past decade, cupuaçu still lacks in-depth studies exploring its moleculares and genomic bases. Such knowledge is crucial for supporting more efficient genetic improvement strategies aimed at enhancing fruit quality and valuing agronomic and industrial traits. *Cacao*, on the other hand, is widely cultivated, and its globais economic importance is tied to cocoa production, which is fundamental to the chocolate industry. Scientific efforts in studying *Theobroma cacao* have resulted in the complete sequencing of its genome, revealing a wealth of information on genes related to fruit quality, disease resistance, and ripening processes, providing a solid foundation for the genetic improvement of this species. This makes cupuaçu an excellent candidate for molecular and comparative genomic studies, particularly in relation to its congener, *cacao*. In addition to these species, another notable member of the Malvaceae family is *Theobroma umbraticum*, a lesser-known species that, like cupuaçu and *cacao*, has potential for comparative studies. *Theobroma umbraticum* also belongs to the subfamily Byttnerioideae, and its fruits share similar characteristics with *T. cacao*, although the species is underutilized and remains economically and scientifically underexplored. Comparative analyses of *Theobroma grandiflorum*, *Theobroma cacao*, and *Theobroma umbraticum* could contribute to understanding the evolution and diversification of genes responsible for important traits related to fruit quality, offering new insights for the conservation and sustainable use of Amazonian biodiversity.

biodiversity.Keywords: Amazonian biodiversity, genomic sequencing, genetic improvement, Cobra-likes, species domestication.

CAPÍTULO 1 - Considerações Gerais

1. INTRODUÇÃO

Theobroma (Malvaceae: Byttnerioideae), um gênero Neotropical de ampla distribuição nas florestas tropicais, especialmente na Amazônia, representa um componente significativo da biodiversidade regional. Atualmente, são descritas cerca de 40 espécies para o gênero, compreendendo 23 espécies de árvores de sub-bosque nativas de florestas neotropicais (Colli-Silva; Richardson; Pirani, 2023; Colli-Silva *et al*., 2024), das quais 17 espécies são oficialmente reconhecidas em território brasileiro, conforme a revisão taxonômica mais recente (Flora e Funga do Brasil, 2024). Dois trabalhos taxonômicos clássicos no estudo do grupo dos cacaus são a revisão de *Theobroma* (Cuatrecasas, 1964) e a sinopse de *Herrania* (Schultes, 1958), tradicionalmente considerado seu gênero-irmão (Figura 1), com 17 espécies descritas. Dentre elas, destacava-se *Herrania umbratica*, recentemente reclassificada como *Theobroma umbraticum* (R.E.Schult.) Colli-Silva (Colli-Silva *et al.*, 2024), a espécie que compartilha a maioria das proteínas e genes com *T. cacao*, apresentando mais de 96% de similaridade (Villamar-Torres *et al*., 2024).

O gênero *Herrania* distingue-se de *Theobroma* por apresentar folhas compostas palmadas (ao contrário das folhas simples de *Theobroma*), um cálice predominantemente trímero (diferindo do cálice pentâmero usual em *Theobroma*) e uma porção superior da pétala unguiculada consideravelmente mais longa em *Herrania* do que em seu gênero-irmão (Bossa-Castro *et al.*, 2024; Colli-Silva *et al.*, 2024; Flora e Funga do Brasil, 2024). Porém, estudos recentes propõem uma reavaliação taxonômica, sugerindo a sinonimização do gênero *Herrania* em *Theobroma* com base em análises morfológicas e filogenéticas (Bossa-Castro *et al.*, 2024; Colli-Silva *et al.*, 2024). Com base em dados moleculares, os estudos demonstram que o gênero *Theobroma* é parafilético, uma vez que não inclui todos os descendentes de um ancestral comum (Figura 1). Esse padrão decorre da posição filogenética de *Herrania*, que emerge como um grupo monofilético intimamente relacionado a *Theobroma* e compartilha com *Theobroma* importantes características

morfológicas, como o hábito arbóreo, a morfologia das inflorescências e os tipos de frutos. Esses resultados levam à proposta de reclassificar *Herrania* como uma seção dentro de *Theobroma*, restaurando uma circunscrição clássica do gênero e promovendo uma melhor compreensão das relações evolutivas dentro da família Malvaceae (Bossa-Castro *et al.*, 2024; Colli-Silva *et al.*, 2024).

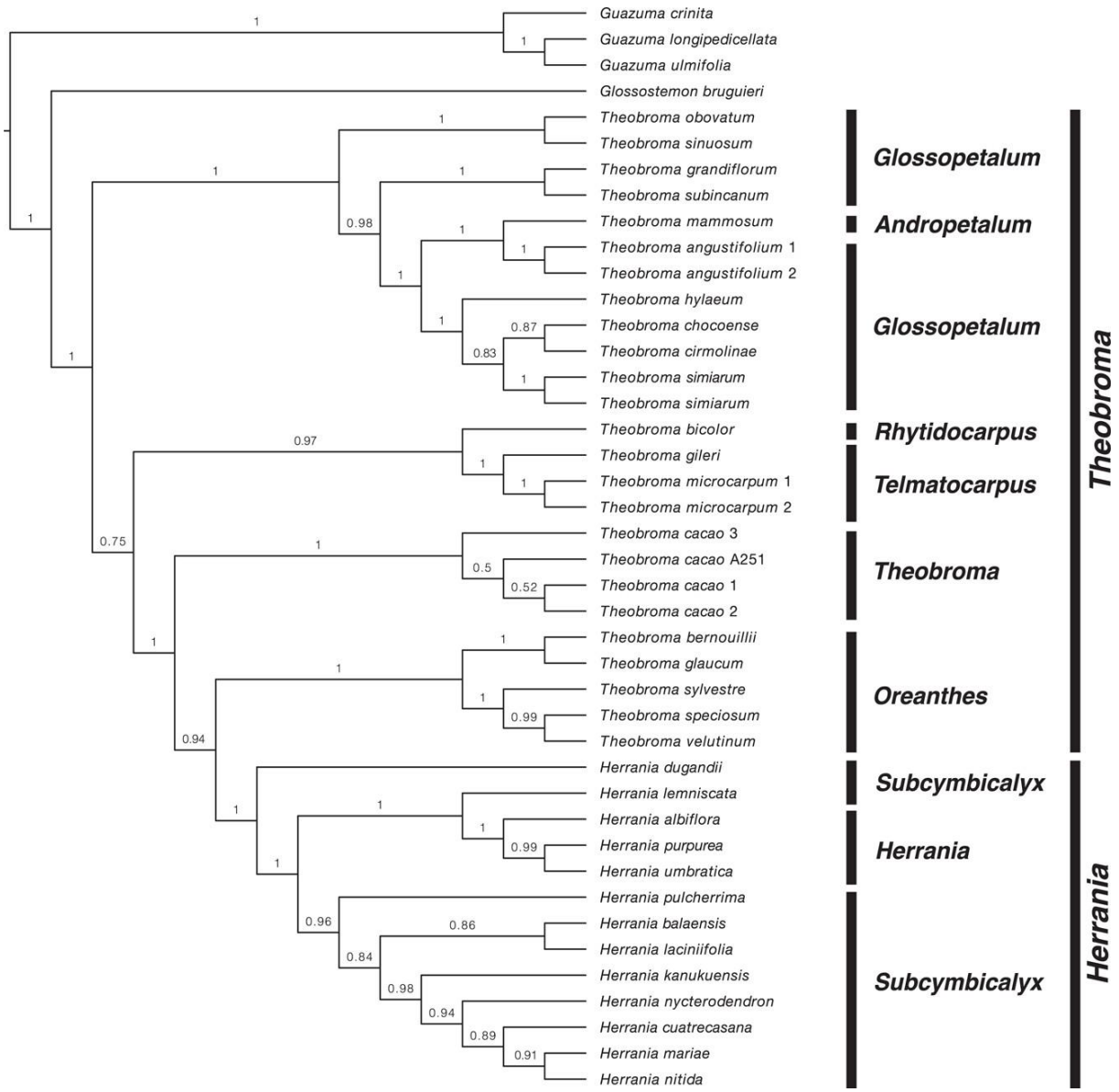

Figura 1. Árvore filogenética inferida para os gêneros *Theobroma* e *Herrania* (Malvaceae), com destaque para a parafilia de *Theobroma*. O gênero *Theobroma* é parafilético, com *Herrania* posicionado em um ramo interno de *Theobroma*. Isso indica que a definição tradicional de *Theobroma* exclui descendentes críticos de seu ancestral comum. Fonte: Bossa-Castro *et al.*, 2024.

Muitas espécies de *Theobroma*, especialmente as espécies comercializadas (*Theobroma cacao* L. – o cacau comercial – e *Theobroma grandiflorum* (Willd. ex Spreng.) K.Schum – o cupuaçu) são amplamente cultivadas no Brasil (Mar *et al.*, 2024).

A fruticultura e a diversidade de cultivares na região amazônica são muito ricas, sendo o cacau a espécie mais conhecida e cultivada, considerada a principal fonte de matéria-prima para a indústria de chocolate. Por outro lado, o cupuaçuzeiro uma das suas frutas mais populares da região norte e nordeste do Brasil, apresenta diversas aplicações em sua utilização, desde fins alimentícios, a usos industriais. Vários produtos são fabricados a partir da polpa do cupuaçu, como sucos, sorvetes, licores, geleias, doces e cosméticos (Calzavara, 1987; Venturieri; Aguiar, 1988; Nazare; Barbosa; Viegas, 1990; Silva; Silva, 1999; Salgado *et al.*, 2011; Costa *et al.*, 2015; Alves *et al.*, 2017). Porém, ao contrário do cacau, que é cultivado em outros continentes, o cupuaçu tem cultivo restrito à região Amazônica, sendo tradicionalmente plantado com maior frequência em pequenas propriedades, ocupando mão-de-obra familiar e, geralmente, consorciado com outras culturas (Alves; Chaves, 2023).

A produção de cupuaçuzeiro (Figura 2-A) está concentrada principalmente no estado do Pará, que se destaca como o maior produtor, seguido pelos estados do Amazonas, Rondônia, Acre e Bahia (Idam, 2021). A produção anual de cupuaçu no estado do Pará alcançou, em média, 1,6 milhão de toneladas em uma área de 224 mil hectares. Esses dados ressaltam a importância econômica da cultura na região Norte do Brasil, particularmente no Pará, que lidera a produção nacional da espécie (SEDAP, 2024).

Dentro do gênero *Theobroma*, o fruto do cupuaçu é o que apresenta o maior tamanho, sendo a polpa o principal produto (Figura 2-C). As análises de polpa revelam excelentes características e teores médios de fósforo e de vitamina C, superiores aos de muitas outras polpas de frutas (Calzavara, 1987; Mar *et al.*, 2024). As características organolépticas de sua polpa e propriedades favoráveis como matéria-prima para industrialização, têm sido responsáveis por um interesse cada vez maior na sua exploração por parte dos diversos segmentos da cadeia produtiva. Dentre as possibilidades de ampla comercialização, a melhor opção para a comercialização em

larga escala encontra-se principalmente na polpa, o que pode ser viabilizado por meio de ações estratégicas do setor privado e da esfera governamental (SUFRAMA, 2017; Mar *et al.*, 2024).

Figura 2. (A) Árvore de *T. grandiflorum* com frutos em desenvolvimento. (B) Detalhe de um fruto de cupuaçu. (C) Fruto de cupuaçu aberto, mostrando a polpa interna. Extraído de Alves *et al.*, 2024

Por outro lado, a partir das sementes do cupuaçu são fabricados chocolate branco, ao leite e meio amargo de ótima qualidade (também chamado de “cupulate”) (Nazare; Barbosa; Viegas, 1990). As sementes também são utilizadas para produção da manteiga de cupuaçu feita a partir de seu óleo, proveniente de processos mais econômicos em relação à manteiga de cacau, podendo ser empregada na produção de cupulate em tabletes e na indústria de cosméticos; portanto, trazendo benefícios

sociais e econômicos interessantes, além do desenvolvimento de produtos para a área de saúde estética devido à riqueza de compostos bioativos presentes em sua polpa e sementes (Nazare; Barbosa; Viegas, 1990; Purcell, 2019; Mar *et al.*, 2024). As sementes do cupuaçu também são ricas em gordura e, quando fermentadas, secas e torradas adequadamente, podem ser utilizadas na elaboração de produtos análogos aos oriundos das amêndoas de cacau (Mar *et al.*, 2024).

Nos últimos anos, os produtores da fruta vêm aproveitando cada vez mais seus recursos, incentivando uma cadeia produtiva na região Amazônica. Como consequência, o fruto e sua diversidade de produtos derivados se encontram cada vez mais presentes no mercado nacional e internacional (Mar *et al.*, 2024). O aumento da industrialização da polpa de cupuaçu, tem proporcionado também um incremento significativo no volume de sementes, que correspondem a 20% do peso do fruto (Cohen; Jackix, 2005). Porém, o interesse do setor industrial pela polpa e sementes do cupuaçu é acompanhado da necessidade de se obter novas formas para sua produção, melhores estratégias de conservação e redução das perdas pós-colheita, geralmente provocadas pelo tempo demandado no transporte, visto que muitas regiões potencialmente produtoras de cupuaçu estão afastadas de polos industriais, e por doenças que afetam a sua produtividade (Cohen; Jackix, 2005; Gondim, *et al.,* 2021).

As principais doenças que acometem o cupuaçuzeiro e o cacaueiro são causadas por fungos, com destaque para a vassoura-de-bruxa, provocada por *Moniliophthora perniciosa*, e a monilíase, causada por *Moniliophthora roreri* (Bailey *et al.*, 2018). A vassoura-de-bruxa induz hipertrofia dos meristemas ativos, seguida pelo secamento de frutos, ramos e folhas, sendo dispersada principalmente pelo vento e pela chuva (de Souza *et al*., 2014). Já a monilíase é uma doença devastadora que compromete severamente a produção e a qualidade dos frutos caracterizada pelo escurecimento, enrugamento e murcha dos frutos, que acabam secando e tornando-se inadequados para consumo ou processamento (Bailey *et al.*, 2018). Ambos os patógenos são intimamente relacionados e representam uma ameaça significativa às culturas de cupuaçu e cacau (Ploetz, 2016; Bailey *et al.*, 2018).

Com a finalidade de contornar esses problemas e melhorar a produtividade, programas de melhoramento genético do cupuaçuzeiro, conduzidos pela EMBRAPA Amazônia Oriental em Belém, PA, Brasil, desenvolveram diversos genótipos para seleção de clones elites de cupuaçu, sendo, por exemplo o C174 resistente a vassoura-de-bruxa, e C1074 suscetível a vassoura-de-bruxa (Alves *et al.*, 2017). Porém, apesar da sua importância comercial e estratégica, principalmente para a Região Norte do Brasil, ainda são incipientes os estudos nas áreas de genética molecular e genômica; como também em áreas de pesquisa mais aplicadas, como, nutrição, seleção de cultivares adaptados às condições edafoclimáticas das regiões produtoras e estudo de mercados.

Porém, os esforços científicos têm se concentrado principalmente no cacaueiro, devido ao seu impacto econômico global na produção de chocolate, evidenciando uma lacuna no desenvolvimento de tecnologias para espécies como o cupuaçu. Portanto, para aproveitar plenamente o potencial do cupuaçu, é essencial intensificar pesquisas em genética molecular e genômica, com foco na identificação de genes ligados a características agronômicas e qualidade dos frutos. Esses estudos podem viabilizar uma seleção mais eficiente, práticas de manejo aprimoradas e maior resistência a doenças, promovendo a sustentabilidade e competitividade do cupuaçu no mercado nacional e internacional como uma alternativa viável ao cacau.

A caracterização de genes relacionados à qualidade dos frutos em cupuaçu, que possam ser identificados através de estudos genômicos é de extrema importância para o melhor conhecimento e aproveitamento da espécie. Por exemplo, estudos de Alves *et al*. (2024) revelam que o genoma do cupuaçu possui genes exclusivos e padrões únicos de expansão e contração em famílias de genes relacionados a características de qualidade dos frutos e mecanismos de defesa. Em *T. grandiflorum*, foram identificadas 282 famílias de genes exclusivas. Além disso, 730 famílias de genes são compartilhadas entre *T. grandiflorum* e *T. cacao*, enquanto 297 são comuns a *T. grandiflorum* e *T. umbraticum*. Indicando que entre as famílias de genes exclusivas e compartilhadas e singletons, muitas estão ligadas à qualidade da fruta, maturação, desenvolvimento de características organolépticas, desenvolvimento geral da planta e resistência a patógenos.

Essas variações genômicas identificadas podem explicar as diferenças morfológicas, de amadurecimento e resistência a doenças observadas entre o cupuaçu e outras espécies da família Malvaceae, como o cacau e *T. umbraticum*. Portanto, o sequenciamento do genoma do cupuaçu pode apresentar grande potencial para contribuir com o aprimoramento das características agrícolas e industriais do cupuaçu, como produtividade e resistência a patógenos, sendo um recurso valioso para futuros estudos e inovações no setor (Alves *et al*., 2024).

Partindo desse contexto, o objetivo principal desta tese de doutorado foi de realizar uma análise genômica comparativa do cupuaçu em relação aos genomas sequenciados de outras espécies da família Malvaceae, especificamente cacau e *T. umbraticum,* com foco em identificar, caracterizar, comparar e avaliar filogeneticamente genes relacionados às qualidades comerciais dos produtos derivados do fruto do cupuaçu.

2. REVISÃO DE LITERATURA

2.1. Introdução à família Malvaceae e a importância dos frutos de *Theobroma grandiflorum, T. cacao* e *T. umbraticum*

A família Malvaceae é considerada uma das linhagens botânicas mais diversificadas (Christenhusz e Byng, 2016). Essa família apresenta mais de 200 gêneros e mais de 4.000 espécies amplamente encontradas nas regiões tropicais do planeta, com destaque para as Américas, África e Ásia (Bayer e Craene, 2003; Flora e Funga do Brasil, 2024). Essas plantas estão globalmente distribuídas, exceto no Ártico, na Antártida e no Deserto de Gobi (Christenhusz *et al*., 2017). No entanto, apresentam predominância em regiões tropicais e subtropicais, como a Amazônia (Veldkornet, 2023).

Historicamente, a família Malvaceae é conhecida por plantas de grande importância econômica e ecológica, incluindo herbáceas, arbustos e árvores (Xu e Deng, 2017). Dentre os vários representantes de importância econômica, destacam-

se espécies como o quiabo (*Abelmoschus* spp.), o algodão (*Gossypium* spp.), o cacau e cupuaçu (*Theobroma* spp.), além de hibiscos e malvaviscos (gêneros *Hibiscus* e *Malvaviscus*), e árvores emblemáticas como os baobás e as barrigudas, pertencentes à subfamília Bombacoideae, incluindo os gêneros *Ceiba* Mill., *Adansonia* L. e *Pseudobombax* Dugand (Flora e Funga do Brasil, 2024). Uma característica notável dessa família é a presença de tricomas estelares na superfície de muitas espécies, que auxiliam na proteção contra a perda de água e herbivoria (Silva, Gallão e De Arruda, 2023).

Morfologicamente, as Malvaceae apresentam caules robustos, frequentemente com fibras de líber e cavidades mucilaginosas, funcionando na defesa contra herbívoros e retenção hídrica (Solereder, 1908; Craene, 2003). As folhas são geralmente simples, alternadas, com margens serradas ou palmatamente divididas, e apresentam uma venação frequentemente palmada (Rocha e Machado, 2009). Outro aspecto distintivo é a presença de estípulas e pecíolos que suportam as lâminas foliares, flexibilidade na captação de luz solar (Celka, 2005).

As flores são, em sua maioria, bissexuais e actinomórficas, podendo ser encontradas tanto isoladas quanto cimas, panículas ou fascículos (Silva *et al.,* 2018). Um traço comum é a presença de um epicálice, que forma um envoltório ao redor do cálice, frequentemente dividido em três ou mais lóbulos, adicionando uma camada extra de proteção à flor (Pereira e Souza, 2020). As sépalas são cinco, podendo ser livres ou parcialmente fundidas, com prefoliação valvar (Costa *et al.,* 2017). As pétalas também são cinco, dispõem-se de forma rotativa e frequentemente ligadas à base da coluna estaminal, destacando a forma característica das flores dessa família (Almeida e Ferreira, 2019).

Os estames são numerosos e possuem os filamentos se fundem para formar um tubo, uma característica conhecida como condição adelfo (Chalk, 1950). O ovário é superior, variando de dois a muitos lóculos, sendo mais comum a presença de cinco lóculos formados por dois ou mais carpelos (Chalk, 1950; Walks, 1984; Bayer e Craene, 2003). Os estiletes são únicos, mas frequentemente ramificados na parte superior, com ramificações correspondentes ou múltiplas ao número de carpelos presentes (Rocha e Machado, 2009).

O fruto das Malvaceae pode variar significativamente entre as espécies, indo desde cápsulas loculicidas até esquizocarpos, e, em raros casos, podendo ser semelhantes a bagas (El-khiat *et al.,* 2023). As sementes são geralmente reniformes, podendo ser glabras ou cobertas por tricomas, e apresentam endosperma abundante que nutre o embrião (Heywood, 1971; Fawzi, 2018). O embrião é frequentemente curvado, com cotilédones planos, dobrados ou convolutos, adaptando-se ao espaço dentro da semente (Masullo *et al.,* 2020).

Na Amazônia, a diversidade da família Malvaceae é notável, abrangendo gêneros como *Theobroma*, *Ceiba*, *Hibiscus* e *Helicteres*. No entanto, espécies como *Theobroma cacao* e *Theobroma grandiflorum* se destacam por sua importância econômica e cultural (Richardson *et al*., 2015). *T. cacao* é amplamente cultivado para a produção de chocolate, enquanto *T. grandiflorum* (cupuaçu) é valorizado por seus frutos, utilizados na produção de sucos, doces e cosméticos (Purcell, 2019; Matos Andrade *et al.*, 2023). Além disso, *T. umbraticum*, uma espécie menos conhecida, contribui para a biodiversidade regional e representa a variabilidade genética do grupo (Schultes, 1943). Tanto *Herrania* quanto *Theobroma* apresentam frutos que podem ser usados para fins alimentícios (Figura 3).

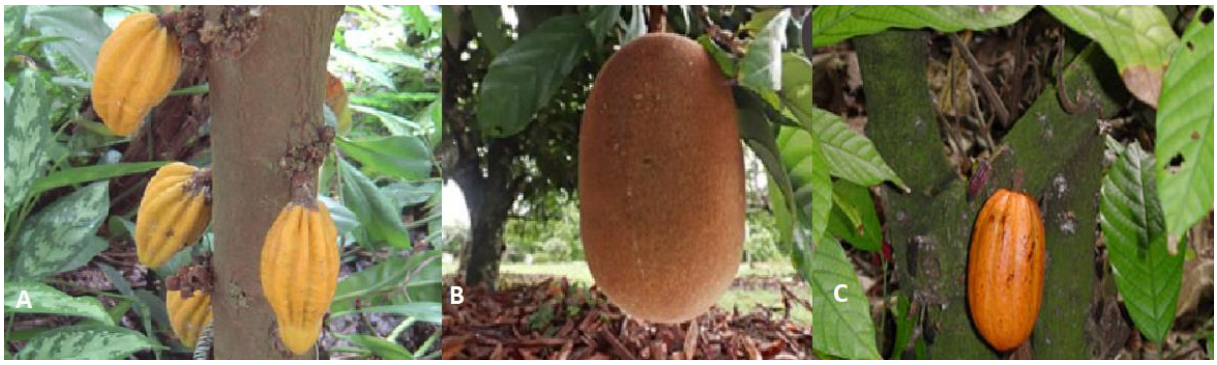

Figura 3. A: Fruto de *Theobroma umbraticum*. B: *Theobroma grandiflorum* e C: *Theobroma cacao.* Fonte: Adaptado de trade winds fruit, 2025 e Alves *et al*., 2024

O cacau e o cupuaçu são árvores perenes, com alturas variáveis, sendo o cacau de 4–8 m de altura e raramente cresce até 20 m, com casca marrom-acinzentada escura, enquanto o cupuaçu pode crescer 5–15 m de altura e tem um tronco de casca marrom. Morfologicamente próximas, suas folhas são simples e alternas, variando de estreitamente ovadas a obovadas-elípticas. As nervuras são pinadas em direção à extremidade com uma nervura principal bem marcada e

nervuras secundárias proeminentes. Elas variam de 20 a 35 cm de comprimento e 6 a 10 cm de largura (Jean-Marie *et al.*, 2022).

Para o cacau, as sementes são ovoides, elipsoides, amigdaloides, mais ou menos complanados ou redondos, variando de 20 a 40 mm de comprimento e 12 a 20 mm de largura com uma polpa envolvida branca ou amarelo-clara. Enquanto o cupuaçu, as sementes são ovoides ou elipsoide-ovoides, mais ou menos achatadas e variam de 20 a 30 mm de comprimento e 20 a 25 mm de largura com uma polpa amarelada (Cuatrecasas, 1964).

No contexto econômico o cupuaçu e o cacau têm um papel importante para as comunidades locais na Amazônia e para o mercado global. O cacau é a base da produção de chocolate, cuja demanda mundial tem crescido de forma constante. Em 2023, estima-se que a produção global de cacau alcançou 5,3 milhões de toneladas, gerando receitas superiores a 131 bilhões de dólares, segundo a Organização Internacional do Cacau (ICCO, 2023). No Brasil, a produção de cacau movimentou cerca de 3 bilhões de reais em 2022, com a maior parte concentrada na região norte e no sul da Bahia, configurando-se como uma importante atividade econômica para pequenos agricultores e cooperativas regionais (IBGE, 2022).

Além do cacau, o cupuaçu vem ganhando relevância tanto nos mercados de alimentos quanto de cosméticos. Os frutos do cupuaçu são amplamente utilizados para a produção de polpas, sucos, geleias e chocolates alternativos, além de serem apreciados na indústria cosmética por suas propriedades antioxidantes (Calzavara, 1987; Venturieri; Aguiar, 1988; Nazare; Barbosa; Viegas, 1990; Silva; Silva, 1999; Salgado *et al.*, 2011; Costa *et al.*, 2015; Alves *et al.*, 2017). Em 2021, a estimativa de produção de cupuaçu no Brasil atingiu cerca de 13 mil toneladas, provenientes em grande parte de pequenos agricultores que adotam Sistemas Agroflorestais (SAFs) na Amazônia, onde o cultivo do cupuaçu é frequentemente combinado com outras espécies nativas, como açaí e castanha-do-pará (AgroMap, 2021).

Enquanto cacau e o cupuaçu já são amplamente cultivados, *T. umbraticum* apresenta um potencial ainda a ser explorado (Whitlock e Baum, 1999). Presente em áreas de floresta na Amazônia, essa espécie possui frutos que se assemelham aos frutos do cacau, sendo importantes na alimentação da fauna local (Nousias *et al.,*

2024). O cultivo de *Herrania* em SAFs poderia seguir os modelos bem-sucedidos já aplicados para cacau e cupuaçu, diversificando ainda mais a produção agrícola na Amazônia e fornecendo novas fontes de renda para agricultores familiares (Rocha *et al.,* 2020).

2.2. Aspectos de doença e controle: *Moniliophthora perniciosa* e *Moniliophthora roreri*

A sustentabilidade das culturas tropicais, especialmente aquelas suscetíveis a doenças fúngicas, enfrenta sérios desafios. Entre essas doenças, a vassoura-de-bruxa, provocada pelo fungo *Moniliophthora perniciosa* e monilíase, causada pelo fungo *Moniliophthora roreri,* destacam-se como as mais devastadoras (Meinhardt *et al.,* 2008; Ploetz, 2016; Bailey *et al.*, 2018). Estes patógenos são responsáveis por perdas significativas na produção de culturas como cacaueiro e cupuaçuzeiro, devido ao seu impacto direto na morfologia da planta e, consequentemente, na sua produtividade (Evans, 1981; Ploetz, 2016; Bailey *et al.*, 2018).

A vassoura-de-bruxa é caracterizada pelo crescimento anômalo de brotações, levando à formação de estruturas que se assemelham a vassouras, o que resulta em uma drástica redução na qualidade e na quantidade da produção (Figura 4) (Mendonça *et al.,* 2019), enquanto a monilíase provoca sintomas conhecidos como "inchaços ou caroços, ilhas verdes, mancha chocolate, pontos oleosos", levando ao escurecimento, enrugamento e murcha dos frutos, que acabam por secar e se tornam inapropriados para o consumo e para o processamento (Figura 5) (Leite *et al.*, 2023). Afetando tanto *Theobroma* quanto *Herrania* (Figura 4; Figura 5 e Figura 6).

Figura 4. Sintomas da vassoura-de-bruxa (*Moniliophthora perniciosa*) em *Theobroma cacao*. A: Frutos afetados pela doença. B: Frutos infectados e necrosados no solo. C: Estruturas reprodutivas do fungo (basidiocarpos). D: Fruto cortado mostrando infecção interna e sementes necrosadas. E: ramos e folhas secas com presença do fungo. F: Fruto com sintomas característicos da doença. Fonte: Adaptado de Mendonça *et al.,* 2019.

Figura 5. Frutos de cacau e cupuaçu infectados por *M. roreri*. A: sintomas conhecidos como caroços decorrentes da infecção em cacau. B: ilhas verdes causadas pela infecção em cacau. C: mancha chocolate e formação de pseudoestroma em cacau. D: sintomas denominados pontos oleosos em cacau. E, F: fruto recobertos por pseudoestroma em cacau (massa de hifas). G: lesões necróticas em cupuaçu. H: formação de psoudoestroma em cupuaçu. I: cupuaçu recoberto por pseudoestroma (massa de hifas). Fonte: Adaptado de Leite *et al*., 2023.

Figura 6. Frutos de *Theobroma umbraticum* afetados por *Moniliophthora roreri*. A: Detalhe da frutificação da espécie com flores e frutos jovens. B: Fruto saudável aderido ao tronco. C: Fruto infectado por *M. roreri*, apresentando os sintomas característicos da monilíase, como necrose e recoberto por pseudoestroma (massa de hifas). Fonte: Leite *et al.,* 2023.

A *M. perniciosa* é um basidiomiceto que se reproduz tanto de forma assexuada quanto sexuada, dependendo das condições ambientais (Tavares *et al.,* 2021). O ciclo de vida desse fungo é complexo e envolve a produção de esporos que se dispersam pelo vento, água ou insetos, estabelecendo novas infecções em plantas suscetíveis. A infecção ocorre predominantemente em condições de alta umidade, que favorecem a germinação dos esporos (Rocha *et al.,* 2020). À medida que a infecção progride, as plantas afetadas apresentam não apenas deformidades nas folhas e ramos, mas também a morte das partes afetadas, o que pode resultar em perdas de até 70% na produção (Cruz *et al.,* 2021).

Os mecanismos de patogenicidade do *M. perniciosa* envolvem a indução de estresses fisiológicos nas plantas hospedeiras, incluindo a inibição do crescimento e a alteração da estrutura da planta. Estudos demonstram que a infecção por *M. perniciosa* resulta em um aumento na produção de fitoalexinas e na ativação de respostas de defesa, embora essas respostas muitas vezes não sejam suficientes para conter a progressão da doença (Almeida *et al.,* 2021; Silva *et al.,* 2022). A interação entre a planta e o patógeno é complexa e multifatorial, envolvendo a regulação de genes relacionados ao estresse e à defesa, além de alterações na composição química das folhas infectadas, que podem afetar a herbivoria e a interação com outros patógenos (Nascimento *et al.,* 2023).

O manejo integrado da vassoura-de-bruxa requer uma abordagem multifacetada, combinando práticas culturais, controle biológico e químico. A remoção de ramos infectados, podas regulares, densidades adequadas de plantio e melhoria da circulação de ar reduzem a umidade, desfavorecendo o fungo (Tavares *et al*., 2021). Além disso, a utilização de variedades resistentes, aplicação de fungicidas e estratégias de manejo cultural têm demonstrado eficácia para reduzir a incidência da doença (Cruz *et al*., 2020). Contudo, a resistência genética das plantas é desafiada pela evolução do patógeno, exigindo contínua pesquisa e desenvolvimento de novas estratégias (Kato *et al*., 2021). O controle biológico, empregando fungos antagonistas, também é uma alternativa promissora (Rocha *et al*., 2020). No que diz respeito ao controle químico, o uso de fungicidas específicos pode ser necessário em situações de surtos severos. No entanto, a resistência a fungicidas é uma preocupação crescente, o que torna fundamental a rotação de produtos químicos para mitigar esse risco e garantir a eficácia do controle (Mendonça *et al.,* 2019). Esse aspecto é particularmente relevante considerando a importância econômica do cacau, que não só impulsiona a economia local, mas também desempenha um papel significativo no mercado global de chocolates. Com receitas anuais que ultrapassam bilhões de dólares, o Brasil se destaca como o quarto maior produtor mundial de cacau (FAO, 2022).

Estudos indicam que a vassoura-de-bruxa pode causar perdas de até 50% na produção em áreas afetadas, levando a uma redução significativa na oferta de grãos de cacau (Pereira *et al.,* 2020). Essas perdas não se limitam à quantidade; a qualidade dos grãos também é comprometida, resultando em um declínio nos preços de mercado. Em alguns casos, as plantações severamente afetadas podem apresentar grãos de qualidade inferior, fazendo com que os agricultores enfrentem uma queda de 30% a 40% nos preços que recebem por seus produtos (Marcelino *et al.,* 2021).

Além das perdas diretas, o controle da vassoura-de-bruxa impõe custos adicionais aos agricultores. O investimento em fungicidas, práticas de manejo integrado de doenças (MID) e a adoção de variedades resistentes exigem um aporte financeiro significativo, que pode ser inviável para pequenos agricultores. Estudos apontam que o custo médio de controle da doença pode variar de 300 a 1.000 dólares

por hectare, dependendo da severidade da infestação e das práticas adotadas (Cruz *et al.,* 2021).

Outra preocupação diz respeito a *Moniliophthora roreri,* que é um fungo hemibiotrófico, assim como *M. perniciosa,* ou seja, passa por duas fases no ciclo de infecção: primeiro como biotrófico, convivendo no hospedeiro sem causar danos aparentes, e depois como necrotrófico, matando o tecido vegetal e se alimentando dele, com hospedeiros limitada a todas as espécies dos gêneros intimamente relacionados *Herrania* e *Theobroma* (Evans; Holmes; Reid, 2003). Nativo da América do Sul e com centro de origem na Colômbia, este fungo é o agente causador da monilíase, uma das doenças mais devastadoras para o cacaueiro em regiões tropicais da América Latina (Phillips-Mora; Aime; Wilkinson, 2007).

Em áreas onde o fungo foi introduzido, como em países da América Central e outros da América do Sul, as perdas podem chegar a 90% da produção, com sérios impactos econômicos e sociais para as comunidades que dependem do cacau (Alvarez; Martínez; Coy, 2014). Em 2021, a situação se agravou no Brasil quando o Laboratório Federal de Defesa Agropecuária de Goiânia (LFDA/GO) confirmou a presença de *M. roreri* no município de Cruzeiro do Sul, no estado do Acre (https://www.gov.br/agricultura). Esta confirmação indicou que a doença ultrapassou as barreiras fitossanitárias do país, possivelmente devido à proximidade com regiões infestadas no Peru e Bolívia. A disseminação do fungo no Brasil representa uma ameaça não apenas para o estado do Acre, mas também para toda a produção cacaueira nacional, especialmente no Pará, estados que são grandes produtores de cacau e cupuaçu (Idam, 2021; Alves *and* Chaves 2023).

Os danos econômicos causados pela monilíase têm motivado o desenvolvimento de estudos moleculares sobre a interação *T. cacao* × *M. roreri* (Zugaib *et al*., 2023). Estudos transcriptômicos dessa interação identificaram genes que podem estar relacionados à defesa e ao estresse do cacau e à biossíntese, regulação e ação de hormônios vegetais possivelmente envolvidos na via de patogenicidade do fungo durante as diferentes fases da doença (Bailey *et al.*, 2014; Zugaib *et al*., 2023).

A doença pode ser controlada por uma combinação de métodos culturais, por exemplo, manutenção da altura da árvore (o que facilita o manejo, a aplicação de produtos fitossanitários e melhora o arejamento e a incidência de luz, condições que inibem o desenvolvimento de fungos) e remoção dos frutos infectadas (essencial para reduzir a fonte de inóculo, interrompendo o ciclo do patógeno e limitando a propagação da doença), além de controle químicos (uso de fungicidas específicos) (Leite *et al.*, 2023). Esses métodos se beneficiam da aplicação regional, mas podem ter um custo proibitivo. A reprodução para resistência a doenças oferece o maior potencial para o gerenciamento monilíase e novos materiais tolerantes estão se tornando disponíveis (Hebbar, 2007; Torres de la Cruz *et al.*, 2011; Zugaib *et al*., 2023).

2.2.1. Mecanismos de resistência e suscetibilidade das plantas do gênero *Theobroma* em relação a *M. perniciosa* e *M. roreri*.

A interação entre os fungos do gênero *Moniliophthora* e as plantas do gênero *Theobroma*, como o cacau e o cupuaçu, exemplifica uma relação patógeno-hospedeiro altamente complexa, marcada por uma constante batalha evolutiva, onde os fungos desenvolvem mecanismos de virulência para superar as defesas das plantas, enquanto estas, por sua vez, ativam respostas de defesa para limitar a infecção e minimizar os danos aos tecidos vegetais (Falcão *et al.*, 2022).

As plantas do gênero *Theobroma* possuem uma série de mecanismos de defesa sofisticados para se proteger contra infecções fúngicas. Um dos primeiros passos na defesa imunológica das plantas é o reconhecimento de PAMPs (Pathogen-Associated Molecular Patterns), que são moléculas conservadas presentes nos patógenos, como flagelina, lipopolissacarídeos e componentes das paredes celulares fúngicas, como a quitina. Esses PAMPs são detectados por PRRs (Pattern Recognition Receptors) localizados na superfície celular da planta, desencadeando a PTI (PAMP-Triggered Immunity), uma resposta imune basal que inclui a produção de espécies reativas de oxigênio (EROs), o reforço da parede celular e a ativação de genes relacionados com a defesa (Junior *et al.*, 2012; dos Santos *et al.*, 2020).

No entanto, *Moniliophthora perniciosa* e *Moniliophthora roreri* desenvolveram estratégias para evadir essa resposta inicial. Eles secretam efetores, que são proteínas capazes de interferir nos processos celulares da planta hospedeira, suprimindo a resposta imune basal e facilitando a colonização (Falcão *et al.*, 2022; da Silva Estrela Junior *et al.*, 2023). Em resposta a essa tentativa de evasão, as plantas ativam a ETI (Effector-Triggered Immunity), uma resposta mais robusta e específica mediada por NLRs (Nucleotide-binding Leucine-rich repeat Receptors). Os NLRs reconhecem diretamente ou indiretamente os efetores dos patógenos, levando à ativação de uma série de respostas que incluem a resposta hiperssistêmica (HR), caracterizada pela morte celular localizada para impedir a disseminação do patógeno.

Além dessas respostas imunológicas, estas plantas utilizam também uma variedade de mecanismos fisiológicos e moleculares para se defender. A produção de quitinase e osmotina desempenha um papel crucial na defesa contra infecções fúngicas (Santana Silva *et al.*, 2020). A quitinase é uma enzima que degrada a quitina, um componente estrutural das paredes celulares fúngicas, dificultando a proliferação do patógeno, enquanto a osmotina é uma proteína que contribui para a resistência ao estresse osmótico e possui atividade antifúngica, ajudando a planta a inibir o crescimento do fungo e a fortalecer as suas defesas celulares (Gomes *et al.*, 2016; Hakim *et al.*, 2018; Santana Silva *et al.*, 2020).

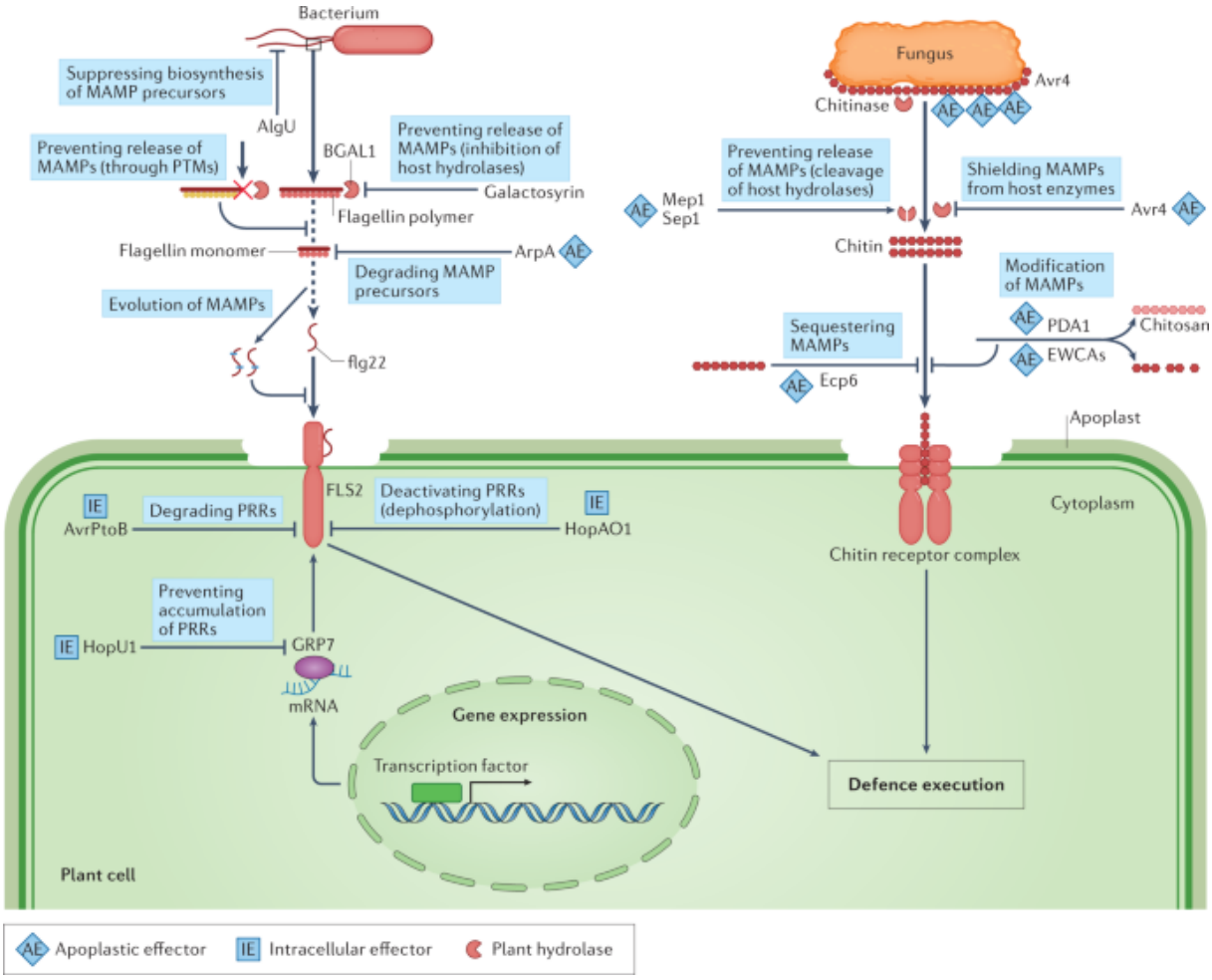

Figura 7. Estratégias de evasão do reconhecimento imunológico de plantas por bactérias e fungos. As plantas detectam padrões moleculares associados a microrganismos (Microbe-Associated Molecular Patterns – MAMPs), também chamados de padrões moleculares associados a patógenos (Pathogen-Associated Molecular Patterns – PAMPs), por meio de receptores de reconhecimento de padrões (Pattern Recognition Receptors – PRRs), desencadeando uma resposta imune. Na parte superior direita, um fungo adota diversas estratégias para evitar a detecção de sua quitina, como: Prevenção da liberação de MAMPs, através de enzimas como Mep1 e Sep1, que degradam hidrolases vegetais responsáveis por expor quitina na parede celular fúngica. Sequestro de MAMPs, onde a proteína Ecp6 captura moléculas de quitina antes que possam ser reconhecidas pelos PRRs vegetais. Proteção dos MAMPs contra degradação, utilizando efetores como Avr4, que se liga à quitina e a protege de enzimas vegetais que poderiam modificá-la. Avr4 é um efetor de avirulência (Avr), ou seja, pode desencadear uma resposta imune se a planta possuir genes de resistência correspondentes. Modificação de MAMPs, por meio de enzimas como PDA1, que convertem quitina em quitosana, uma forma estruturalmente diferente e menos imunogênica, reduzindo sua detecção pelos receptores de quitina da planta. Os efetores representados na figura estão divididos em duas categorias: Efetores apoplásticos (Apoplastic Effectors – AE) e Efetores intracelulares (Intracellular Effectors – IE) Fonte: Wang *et al*., 2022.

Os fungos, utilizam estratégias altamente sofisticadas para evitar o sistema imunológico das plantas durante o processo de infecção (Figura 7). Um dos principais mecanismos envolve uma modulação de moléculas associadas à sua parede celular,

como a quitina, que é um componente estrutural crucial e um padrão molecular associado a patógenos (MAMP). Normalmente, fragmentos de quitina liberados durante o crescimento ou manipulação fúngica são reconhecidos pelos receptores de quitina nas plantas, ativando respostas de defesa. Para evitar esse reconhecimento, os fungos adotam diferentes táticas (Wang *et al*., 2022).

Uma estratégia importante é a proteção de proteínas como Avr4, sendo um efetor de avirulência (Avr), ou seja, pode desencadear uma resposta imune se a planta possuir genes de resistência correspondentes, se ligando à quitina, protegendo-a da ação das enzimas hidrolíticas produzidas pela planta. Isso impede que fragmentos de quitina sejam liberados e reconhecidos pelos receptores. Além disso, os fungos produzidos por enzimas como PDA1, que convertem quitina em quitosana, são uma forma modificada que é menos imunogênica e, portanto, mais difícil de ser detectada pelos receptores da planta.

Outra tática crucial é o sequestro de moléculas de quitina por proteínas como Ecp6. Essas proteínas "capturam" os fragmentos de quitina antes que possam interagir com os receptores da planta, neutralizando a ativação da resposta imune. Em conjunto, os fungos podem modificar ou evoluir seus MAMPs, alterando sua estrutura química para dificultar ainda mais o reconhecimento pelas plantas.

Essas estratégias permitem que os fungos contornem a imunidade basal das plantas, conhecida como imunidade disparada por padrões (PTI), criando condições adequadas para sua colonização. Esse tipo de interação é central em patossistemas como os causados por *Moniliophthora perniciosa* e *M. roreri*, que infectam espécies de *Theobroma* (Santana Silva *et al.*, 2020). Ao manipular os sistemas de defesa da planta hospedeira, esses fungos promovem infecções bem-sucedidas, causando sérios danos às culturas agrícolas (Gomes *et al.*, 2016; Hakim *et al.*, 2018; Santana Silva *et al.*, 2020).

No caso do cupuaçu, estudos têm demonstrado que a expressão de genes codificadores de quitinase e osmotina é significativamente aumentada durante a infecção por *M. perniciosa* (Santana Silva *et al.*, 2020). Essas proteínas atuam de forma sinérgica para degradar a parede celular do fungo e reforçar a integridade das células vegetais, limitando assim a progressão da doença. Além disso, a sinalização

hormonal mediada por hormônios como o ácido salicílico (SA), o ácido jasmônico (JA) e o etileno (ET) é fundamental para regular a resposta de defesa (Van Loon *et al*., 2006; Kilaru; Bailey; Hasenstein, 2007; Anil Kumar *et al*., 2015; Falcão *et al.*, 2022). O SA está geralmente associado à defesa contra patógenos biotróficos (Kilaru; Bailey; Hasenstein, 2007), enquanto o JA e o ET estão mais relacionados à defesa contra patógenos necrotróficos e herbívoros (Van Loon *et al*., 2006; Anil Kumar *et al*., 2015; Falcão *et al.*, 2022).

Moniliophthora perniciosa e *Moniliophthora roreri* possuem estratégias sofisticadas para superar as defesas das plantas hospedeiras. Além da penetração inicial facilitada pela formação de estruturas especializadas e pela secreção de enzimas degradativas como celulases, pectinases e lignases, esses patógenos manipulam o metabolismo da planta para redirecionar recursos que favorecem seu próprio crescimento (Barbosa *et al.*, 2018; da Silva Estrela Junior *et al.*, 2023). A *M. perniciosa*, por exemplo, estabelece uma fase biotrófica inicial onde mantém as células vegetais vivas para obtenção de nutrientes, transitando posteriormente para uma fase necrotrófica que causa a morte celular e libera nutrientes adicionais para sustentar seu crescimento patogênico (Mondego *et al.*, 2008). A *M. roreri,* por sua vez, apresenta uma estratégia semelhante em termos de fases biotrófica e necrotrófica, mas com características distintas em sua interação com o hospedeiro. Este fungo se especializa em infectar diretamente os frutos em desenvolvimento, iniciando uma fase biotrófica em que coloniza os tecidos sem causar danos visíveis imediatos, enquanto manipula o metabolismo do fruto para favorecer sua nutrição. Com a transição para a fase necrotrófica, *M. roreri* induz a morte celular, resultando em necroses severas, deformações e o colapso completo do tecido afetado. Essa fase é acompanhada por uma intensa produção de esporos, o que contribui para sua alta capacidade de disseminação (Meinhardt *et al.*, 2014).

A interação dinâmica entre *Theobroma* e *Moniliophthora* é resultado de uma coevolução contínua, onde a planta desenvolve novas estratégias de defesa e o patógeno adapta suas táticas de ataque (Falcão *et al.*, 2022). A eficácia das defesas da planta depende de fatores genéticos, ambientais e do estágio da infecção, tornando o manejo dessas doenças um desafio complexo para os agricultores. A compreensão detalhada dos mecanismos de reconhecimento e resposta imune, bem como das

estratégias de ataque dos patógenos, é essencial para o desenvolvimento de variedades de *Theobroma* mais resistentes e para a implementação de métodos de controle mais eficazes.

2.3. Domesticação de *Theobroma*

A domesticação de plantas exerceu uma influência determinante na composição florística atual das florestas amazônicas (Levis *et al*., 2017). Espécies vegetais domesticadas e manejadas por populações indígenas da Amazônia ao longo de milênios apresentam uma probabilidade cinco vezes superior de serem encontradas em altas densidades – caracterizando-se como hiperdominantes – em relação às espécies não domesticadas (Levis *et al*., 2017). A maior abundância e riqueza dessas espécies ocorrem em regiões próximas a sítios arqueológicos, particularmente no sudoeste e leste da Amazônia, áreas associadas a antigas ocupações humanas (Levis *et al.,* 2017). A distribuição atual dessas espécies domesticadas sugere que o manejo e a domesticação humana deixaram uma marca duradoura na estruturação das comunidades arbóreas amazônicas, refletindo práticas culturais e agrícolas de povos antigos.

A distribuição geográfica e a provável origem da domesticação do cacau na Amazônia (Figura 8-E), representada pelo símbolo “+++” na região noroeste aponta para uma origem conhecida de domesticação do cacau, sugerindo que essa área foi fundamental no início do processo de domesticação, abrangendo uma parte da região norte do Brasil, Amazonas e Acre, além de áreas fora do Brasil, como Colômbia e Peru. As sombras em cinza, indica a estimativa da presença do cacau, sendo que áreas mais escuras refletem maior concentração ou densidade da espécie em determinadas regiões. Os pontos pretos representam parcelas de amostragem onde o cacau foi registrado, e o tamanho desses pontos está relacionado à abundância relativa da espécie nesses locais. Por outro lado, os pontos vermelhos indicam parcelas onde o cacau não foi detectado, destacando áreas da Amazônia em que a espécie está ausente. A Amazônia foi subdividida em seis regiões geológicas (Figura 8) — NWA (Amazônia Noroeste), SWA (Amazônia Sudoeste), SA (Amazônia Sul), CA (Amazônia Central), GS (Escudo das Guianas) e EA (Amazônia Oriental).

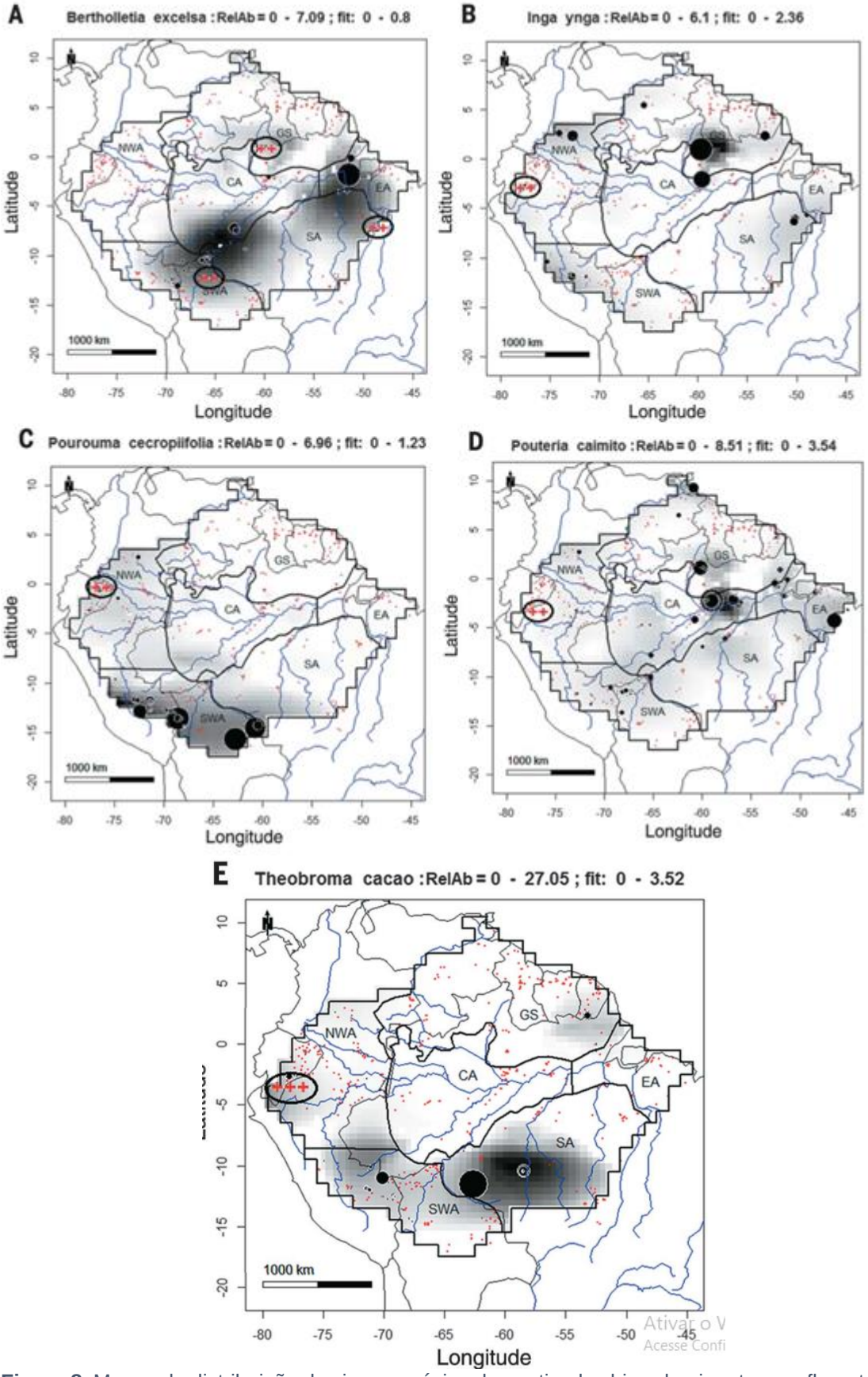

Figura 8. Mapas de distribuição de cinco espécies domesticadas hiperdominantes em florestas amazônicas e suas prováveis origens de domesticação (A a E). "+++" para origem conhecida. "++" para origem hipotética. pontos pretos: abundância. Pontos vermelhos: parcelas onde cada espécie domesticada não foi registrada. Fonte: Levis *et al.,* 2017

O processo inicial de domesticação de plantas envolve a seleção e o cultivo de variedades com características morfológicas desejáveis, como o aumento do tamanho dos frutos, visando o aprimoramento das gerações subsequentes, além de favorecer a dispersão das populações vegetais de seus habitats selvagens originais para novas paisagens antropogênicas, resultando frequentemente em um evento fundador – um fenômeno no qual novas populações se originam a partir de uma amostra reduzida da população parental, ocasionando uma redução na variabilidade genética e morfológica (Clement, 1999; Zohary, 2004). Em populações arbóreas, as mudanças genéticas e morfológicas derivadas desse processo são tipicamente sutis, especialmente em cenários de manejo dentro de florestas (Clement, 1999).

A domesticação do cacau é um processo complexo que se estende por milênios. O cacau passou por forte domesticação ~3.600 anos atrás, com a introdução de alguns indivíduos da Amazônia Ocidental na Mesoamérica (Cornejo *et al.*, 2018). Há também evidências arqueogenômicas do consumo de cacau na Amazônia Ocidental por volta de 5.300 anos por diferentes grupos indígenas (Zarrillo *et al.*, 2018).

Inicialmente, o cacau foi domesticado em florestas úmidas localizadas no noroeste da Amazônia, uma região caracterizada por solos ricos em nutrientes, que são ideais para o crescimento da planta. No entanto, ao longo do tempo, a área de cultivo e ocorrência do cacau mudou, e hoje a planta é mais abundante nas florestas do sudoeste e sul da Amazônia (Thomas *et al.*, 2012). O autor destaca ainda que embora o cacau seja endêmico da bacia amazônica, acredita-se que seu processo de domesticação visando a produção de bebida de chocolate ocorreu na Mesoamérica. Contudo, a distribuição da diversidade genética do cacau na América do Sul é influenciada não apenas por processos naturais de diferenciação genética, mas também por intervenções humanas pré-colombianas (Levis *et al.,* 2017).

Tradicionalmente, o cacau foi classificado em dois grandes grupos genéticos: “Criollo” e “Forastero”, baseados em características morfológicas e origens geográficas. Um terceiro grupo, “Trinitario”, representa híbridos “Criollo × Forastero” (Cheesman, 1944). Análises moleculares identificaram 10 grandes grupos genéticos no cacau: Amelonado, Contamana, Criollo, Curaray, Guiana, Iquitos, Maraňón, Nanay,

Nacional e Purús (Motamayor *et al*., 2002), refletindo com mais precisão a diversidade genética útil para melhoristas.

Recentemente, a pesquisa genômica identificou mais grupos, com destaque para as populações de cacau selvagens do Alto Amazonas, como Contamana, Iquitos e Nanay, que não são domesticados, mas desempenham um papel crucial ao fornecer novos recursos genômicos para o melhoramento da espécie. Estudos de sequenciamento de genomas de alta qualidade dessas populações selvagens, utilizando a tecnologia PacBio Hi-Fi, revelaram que a divergência entre os grupos genéticos de cacau ocorreu entre 1,83 e 0,69 milhões de anos atrás, durante o período Pleistoceno inicial e médio (Nousias *et al*., 2024).

O cacau Criollo é amplamente considerado um dos primeiros genótipos domesticados, possivelmente originado na América Central, onde há registros arqueológicos do uso e do consumo de cacau em culturas como a dos maias e astecas (Motamayor *et al.*, 2002). Este genótipo é conhecido por sua alta qualidade e sabor refinado, com uma baixa presença de taninos, o que resulta em um sabor menos amargo, e foi intensamente selecionado por essas características, que eram desejáveis para a produção de bebidas e produtos rituais e de prestígio, contudo por ser um genótipo altamente domesticado, o cacau Criollo apresenta menos diversidade genética e uma maior vulnerabilidade a doenças, uma consequência de sua seleção intensiva (Bartley, 2005; Motamayor *et al.*, 2008).

O cacau Matina, originário da América do Sul, pertence ao grupo Forastero, caracterizado por maior resistência e produtividade. O cultivar Matina 1-6 exibe fenótipo Amelonado, com menos diversidade genética, sendo amplamente cultivado no mundo (Motamayor *et al*., 2003). Enquanto o Criollo foi selecionado pelo sabor, o Matina foi favorecido por resistência e produtividade, refletindo contextos culturais e ecológicos distintos (Motamayor *et al*., 2002; Motamayor *et al*., 2003). Este grupo apresenta pouca diversidade genética em comparação com os outros nove grupos sendo o tipo de cacau mais comumente cultivado em todo o mundo (Motamayor *et al.*, 2003). As árvores Forastero também são classificadas de acordo com a morfologia do fruto, como o tipo Amelonado, caracterizado pelo formato de melão da fruta e que possui um sabor mais intenso e amargo devido a níveis mais altos de compostos

fenólicos e taninos, características que o tornam adequado para o cultivo em larga escala e para misturas com outros cacaus (Motamayor *et al.*, 2003).

A domesticação dos dois genótipos mostra diferentes estratégias de manejo e seleção, com o Criollo tendo surgido cerca de 1,83 milhões de anos atrás e o Matina por volta de 1,34 milhões de anos atrás (Nousias *et al.*, 2024). Isso indica o momento em que cada uma dessas variedades se separou geneticamente de suas populações ancestrais. Enquanto o Criollo foi selecionado principalmente pelo sabor e qualidade, o Matina foi favorecido por sua resistência e produtividade (Motamayor *et al.*, 2002; Motamayor *et al.*, 2003). Essa seleção diferencial reflete os contextos culturais e ecológicos de cada região, onde as populações humanas moldaram geneticamente as variedades de cacau com base em suas necessidades e preferências locais.

Quanto ao cupuaçu, estudos genômicos indicam que esta espécie é uma variante domesticada de *Theobroma subincanum* (cupuí), encontrado na bacia do Médio-Alto Rio Negro (Colli-Silva *et al*., 2023). A domesticação do cupuaçu remonta a 5.000-8.000 anos, durante o Holoceno, refletindo manejo intensivo por populações indígenas (Colli-Silva *et al*., 2023). Esse período apresentou gargalo genético significativo, possivelmente resultante da seleção de frutos maiores e polpa de melhor qualidade. Uma segunda fase de domesticação mais recente, relacionada à era moderna dos séculos XIX e XX, e à expansão comercial do cultivo, promoveu nova redução na diversidade genética (Colli-Silva *et al*., 2023).

O cupuaçu foi domesticado antes do cacau (Cornejo *et al.*, 2018; Colli-Silva *et al.*, 2023), sendo o cacau usado pela primeira vez por sua polpa por sociedades indígenas (Zarrillo *et al.*, 2018) e só mais tarde foi domesticado para produzir uma bebida semelhante ao chocolate a partir das sementes. O uso inicial da polpa em ambas as espécies teria tornado o cupuaçu mais atraente do que o cacau se uma maior quantidade e/ou qualidade de polpa fosse encontrada no primeiro ou em representantes de seu progenitor cupuí (Colli-Silva *et al.*, 2023).

O processo de domesticação do cupuaçu é extraordinário porque a seleção artificial aplicada ao seu parente silvestre, o cupuí, foi tão intensa e eficaz que resultou em mudanças significativas nas características morfológicas da planta (Cuatrecasas, 1964; Colli-Silva *et al.*, 2023). Essas alterações foram tão pronunciadas que os

taxonomistas, ao estudarem o cupuaçu, o classificaram como uma espécie distinta do seu ancestral selvagem (Cuatrecasas, 1964). Isso indica um grau elevado de intervenção e seleção pelos humanos, que moldaram o cupuaçu para ter características desejáveis, como frutos maiores e com mais polpa.

Já *T. umbraticum* não apresenta processo significativo de domesticação, mantendo-se predominantemente em estado selvagem (Colli-Silva; Richardson; Pirani, 2023). A especiação de *Theobroma cacao* e *Theobroma umbraticum* ocorreu há cerca de 10 milhões de anos, conforme sugerido por estudos genéticos. *T. cacao* é amplamente cultivado ao contrário de *T. umbraticum* (Nousias *et al*., 2024).

2.4. Genômica do *Theobroma*

Um dos primeiros passos para se estudar a estrutura genômica de uma espécie de planta é primeiramente obter-se a sequência genômica em escala cromossômica. Este processo vem se tornando cada vez mais simplificado através da utilização de técnicas de sequenciamento de leituras longas e Hi-C. As leituras longas, fornecidas por tecnologias como PacBio e Oxford Nanopore, permitem a geração de sequências contínuas que cobrem grandes regiões do genoma sem as quebras típicas das leituras curtas. Isso facilita a montagem de genomas mais completos e precisos, reduzindo as lacunas e melhorando a resolução das regiões complexas, como aquelas ricas em repetições ou estruturas estruturais intricadas (Soto-Serrano *et al.*, 2024; Alves *et al.,*2024).

Por outro lado, a técnica de Hi-C é uma metodologia de sequenciamento que captura a interação física entre diferentes regiões do genoma dentro do núcleo celular. Basicamente, Hi-C envolve a fixação das interações cromossômicas com formaldeído, seguida pela digestão do DNA, ligadura dos fragmentos proximais e sequenciamento das ligações resultantes. Isso gera um mapa tridimensional das interações genômicas, permitindo a organização das sequências de DNA em suas posições cromossômicas corretas. A combinação de leituras longas com dados de Hi-C possibilita a montagem de genomas em escala cromossômica, onde cada cromossomo é representado de

forma completa e linear, refletindo sua verdadeira organização dentro da célula (Alves *et al.,*2024; Pladevall *et al.*, 2025).

Somente com genomas em escala cromossômica é possível conhecer sua história de paleopoliploidia, ou seja, os eventos evolutivos que moldaram seu genoma ao longo do tempo. Assim como a uva (*Vitis vinifera*), o gênero *Theobroma* passou por apenas um evento de duplicação de genoma inteiro (WGD), correspondente à triplificação gama (gamma triplication), que foi originada há aproximadamente 125 milhões de anos (mya) (Doyle; Hotton, 1991; Jiao *et al.*, 2011). Este evento de triplificação é compartilhado por todas as eudicotiledoneas e é fundamental para a diversificação e complexidade dos genomas dessas plantas (Jiao *et al.*, 2011; Jiao *et al.*, 2012). Com os genomas completos, em escala cromossômica, também é possível realizar análises detalhadas de sintenia e colinearidade, que são essenciais para comparar a organização genômica entre diferentes espécies (Richardson *et al.,* 2015; Kumar *et al.,* 2022; Alves *et al.,* 2024).

A bioinformática desempenha um papel crucial nesse processo, fornecendo as ferramentas e métodos necessários para a montagem, anotação e análise dos genomas sequenciados. Com genomas em escala cromossômica devidamente montados e anotados, a bioinformática permite a realização de análises comparativas que elucidam a evolução genômica, identificam regiões de conservação e divergência, e detectam eventos de duplicação e perda gênica ao longo da história evolutiva das espécies. Essas análises são fundamentais para entender como diferentes espécies de plantas, se relacionam e divergem em termos de genética e funcionalidade.

Portanto, o estudo genômico por métodos de bioinformática é reconhecido como uma das principais ferramentas para o avanço do conhecimento da biologia de uma espécie cultivável, podendo levar a médio e longo prazo a inúmeras aplicações na agroindústria, desde as relacionadas à conservação, melhoramento e biotecnologias (Diniz; Canduri, 2017). Diversas pesquisas envolvendo genômica e bioinformática foram aplicados nas últimas décadas, levando a um grande avanço do conhecimento de diversos cultivares de interesse comercial, como por exemplo o abacate (Rendón-Anaya *et al.*, 2019), morango (Edger *et al.*, 2019), uva (Zharkikh *et al.*, 2008), milho (Haberer *et al.*, 2005), abacaxi (Chen *et al.*, 2019), maça (Zhang *et al.*, 2019), melancia (Guo *et al.*, 2013), banana (D'Hont *et al.*, 2012), mamão (Ming *et*

al., 2008), laranja (Xu *et al.*, 2013), pêssego (Verde *et al.*, 2013), arroz (Eckardt, 2000), feijão (Schmutz *et al.*, 2014), cana-de-açucar (Garsmeur *et al.*, 2018; Zhang *et al.*, 2018), eucalipto (Myburg *et al.*, 2014), algodão (Huang *et al.*, 2020) e outros. Todos esses projetos geraram conhecimentos que são aplicados em programas de melhoramento, tanto para favorecer a produtividade destes cultivares, melhorar a resistência a doenças, como também de criar frutas com novos sabores e texturas (Tabela 1).

Tabela 1. Exemplos de pesquisas genômicas e bioinformáticas realizadas em diversas espécies de interesse comercial, incluindo sequenciamento do genoma e estudos relacionados a melhorias na produtividade, resistência a doenças e características sensoriais, como sabor e textura. A tabela inclui a espécie, a abordagem realizada e a referência de cada trabalho.

Espécie	Trabalho realizado	Referência
Abacate	Sequenciou e analisou os genomas de abacates mexicanos, Hass e espécies relacionadas, identificando sua diversidade genética, eventos evolutivos como poliploidia e introgressão, além de características adaptativas relevantes para o melhoramento genético, resistência a doenças e otimização de características desejáveis. Ele também avançou na compreensão das relações filogenéticas dos magnoliídeos com outros clados de angiospermas e na função diferencial de genes duplicados.	Rendón-Anaya *et al.,* 2019
Morango	Decifrou a origem e evolução do genoma octoplóide do morango cultivado, revelando a origem norte-americana, a dinâmica entre seus subgenomas e a dominância de um subgenoma associado à expressão gênica e características chave, fornecendo uma base para estudos evolutivos e melhoramento molecular.	Edger *et al.,* 2019
Uva	Sequenciou e montou o genoma altamente heterozigoto de *Vitis vinifera*, fornecendo informações sobre estrutura genômica, polimorfismos, SNPs, inserções e deleções, além de insights sobre a dinâmica evolutiva de elementos repetitivos em plantas.	Zharkikh *et al.,* 2008
Milho	Analisou sistematicamente sequências do genoma do milho, destacando a alta densidade de genes, a predominância de retroelementos repetitivos (66%) e a contribuição desses elementos para o aumento do genoma em relação a outras plantas modelo, como arroz e *Arabidopsis* .	Haberer *et al.,* 2005
Abacaxi	O trabalho sequenciou e analisou o genoma do abacaxi bracteatus, elucidando a base genômica de características como produção de fibras e acúmulo de açúcar, e forneceu evidências de uma domesticação híbrida com mistura genética e uma operação de uma etapa em cultivares como 'Singapore Spanish'.	Chen *et al.,* 2019
Maçã	trabalho montou um genoma de alta qualidade da maçã, revelando variações genômicas significativas atribuídas a elementos transponíveis e identificando uma inserção de retrotransposon associada à coloração vermelha da fruta, proporcionando resultados importantes sobre os mecanismos moleculares da coloração.	Zhang *et al.,* 2019

Melancia	Sequenciou o genoma de alta qualidade da melancia e realizou resequenciamento de 20 acessos para mapear a diversidade genética e estrutura populacional, identificando regiões selecionadas durante a domesticação e genes importantes para características de qualidade da fruta.	Guo *et al*., 2013
Banana	sequenciou o genoma de *Musa acuminata*, revelando duplicações genômicas e proporcionando uma base para o melhoramento genético da banana, além de esclarecer as relações filogenéticas das monocotiledôneas e identificar sequências não codificantes conservadas antes da divergência monocotiledônea-eudicotiledônea.	D'Hont *et al.,* 2012
Mamão	O trabalho sequenciou o genoma do mamão, uma árvore frutífera transgênica resistente a vírus, e identificou características genômicas únicas, como a ausência de duplicações recentes e amplificações de genes relacionados à adaptação tropical, oferecendo uma base para estudos de genômica funcional e melhoria da fruta.	Ming *et al.,* 2008
Laranja	O trabalho sequenciou o genoma da laranja, identificando a origem híbrida entre pummelo e tangerina, e destacando genes envolvidos no metabolismo da vitamina C, fornecendo uma base genômica valiosa para estudos e melhoramento das características da fruta.	Xu *et al*., 2013
Pêssego	O estudo sequenciou o genoma do pêssego, revelando padrões de diversidade genética e processos de domesticação, além de indicar a ausência de duplicações genômicas recentes e identificar blocos ancestrais triplicados fragmentados, proporcionando novas perspectivas sobre sua evolução.	Verde *et al.,* 2013
Arroz	O sequenciamento do genoma do arroz revelou a complexidade e o tamanho relativamente compacto de sua estrutura, com aproximadamente 430 Mb. Considerado um modelo ideal devido ao seu tamanho menor em comparação com outros cereais como o trigo, o arroz compartilha alta sintenia com outras gramíneas, facilitando estudos genéticos e melhoramento molecular.	Eckardt, 2000
Feijão	O sequenciamento do genoma do feijão, revelou uma estrutura genômica de 473 Mb ancorada em 11 cromossomos, com 98% da sequência mapeada. A análise identificou duas domesticações independentes no feijão, uma nas regiões mesoamericanas e outra nas andinas, e revelou genes relacionados ao aumento do tamanho das folhas e sementes, fundamentais para o melhoramento das variedades cultivadas.	Schmutz *et al.,* 2014
Cana-de-açúcar	O estudo do genoma da cana-de-açúcar produziu uma sequência de referência monoploide de 382 Mb, gerada a partir de 4660 BACs, para superar os desafios de seu genoma altamente complexo. A análise revelou 25.316 genes, com 17% apresentando divergências significativas em relação ao sorgo, e forneceu insights sobre os rearranjos cromossômicos e elementos transponíveis nas linhagens S. officinarum e S. spontaneum, explicando as diferenças em seus tamanhos genômicos e número de cromossomos.	Garsmeur *et al*., 2018; Zhang *et al.,* 2018
Eucalipto	O sequenciamento do genoma do *Eucalyptus grandis* cobriu mais de 94% de seu genoma de 640 Mb, identificando 36.376 genes codificadores de proteínas, dos quais 34% estão em duplicações em tandem. A análise também revelou uma diversidade significativa de genes relacionados a terpenos, essenciais para a defesa química e	Myburg *et al.,* 2014

	produção de óleos farmacêuticos. Comparações com o genoma de *E. globulus* e genomas consanguíneos de *E. grandis* mostraram uma evolução dinâmica do genoma e pontos críticos de depressão por endogamia. Este estudo fornece uma base sólida para o melhoramento genético e biotecnologia do eucalipto.	
Algodão	O sequenciamento do genoma de *Gossypium herbaceum* (A1) e as atualizações dos genomas de *Gossypium arboreum* (A2) e *Gossypium hirsutum* ((AD)1) revelaram que todos os genomas A podem ter se originado de um ancestral comum, mais próximo de A1 do que de A2. A formação alotetraploide precedeu a especiação de A1 e A2, com variações estruturais abundantes nas regiões gênicas que impactaram a expressão de genes essenciais para a evolução das células de fibra em (AD)1. Este estudo resolve questões sobre as origens do genoma A e oferece recursos para o melhoramento genético do algodão.	Huang *et al.,* 2020

Em relação ao gênero *Theobroma*, um consórcio coordenado por cientistas franceses sequenciou em 2011, utilizando tecnologias de sequenciamento de alto rendimento de segunda geração, com curtos fragmentos de sequências de DNA. A principal tecnologia empregada foi a plataforma Illumina (Solexa), que era uma das mais avançadas da época. Ela usava o sequenciamento por sintese (sequencing by synthesis), que gerava fragmentos de DNA curtos (em torno de 100-150 pares de bases). A plataforma produz grande quantidade de dados, mas com um limite de leitura de fragmentos curtos, o que dificultava a montagem do genoma completo de organismos complexos. Esse sequenciamento do genoma do cultivar B97-61/B2 - Criollo de cacau, gerou 326 megabases (Mb), abrangendo 76% do tamanho total do genoma estimado por citometria de fluxo (Argout *et al.*, 2011). Durante o processo, foram identificados 10 pseudomoléculas, que representam cada um dos cromossomos da planta, além de uma grande quantidade de genes e outros elementos genéticos, como 25.912 genes, 6 rRNAs (RNA ribossomal), 473 tRNAs (RNA transportador), 83 ncRNAs (RNA não codificante) e 65.575 regiões associadas a elementos transponíveis (TEs), que são sequências de DNA móveis no genoma (Argout *et al.*, 2011). Este mesmo estudo revelou informações importantes sobre a evolução genética do genoma Criollo de cacau.

Por exemplo, uma descoberta relevante foi a expansão de uma família de genes ligados aos flavonoides, compostos que estão potencialmente relacionados à qualidade do fruto, como sabor e propriedades antioxidantes (Argout *et al.*, 2011).

Além disso, este estudo identificou genes que conferem resistência da planta a patógenos, especialmente fungos e oomicetos, que são organismos prejudiciais à saúde da planta, como exemplo, NBS-LRR (Nucleotide-Binding Site - Leucine-Rich Repeat), NBS (Nucleotide-Binding Site), LRR-LRK (Leucine-Rich Repeat - Leucine-Rich Repeat Kinase) e NPR1-like (Nonexpresser of PR genes 1-like) e mapeamento de QTLs (Quantitative Trait Locus). Por fim, foram encontrados genes e vias metabólicas que influenciam diretamente a qualidade das sementes do cacau, envolvendo compostos como terpenos aromáticos (responsáveis por aromas), teobromina (um alcaloide presente no cacau, conhecido por suas propriedades estimulantes), além de outros metabólitos secundários, como flavonoides, ácidos fenólicos, cacáoína, lipídios e ácidos graxos, alcaloides, glutamato, asparagina e alanina de importância para o sabor e as características do cacau. Isso tudo contribui para entender melhor a base genética que afeta tanto a resistência da planta quanto a qualidade dos frutos e sementes (Argout *et al.*, 2011).

Posteriormente, em 2017, um grupo de cientistas franceses realizou uma nova versão do sequenciamento do genoma do cacau – cultivar B97-61/B2 Criollo, aprimorando o trabalho anterior de 2011. Usando tecnologias de terceira geração, com leituras longas proporcionadas pelo sequenciamento SMRT (Single Molecule, Real-Time), a nova abordagem permitiu uma análise mais detalhada e precisa do genoma, superando limitações das tecnologias de segunda geração. As leituras longas possibilitaram a montagem de um genoma mais completo, especialmente em regiões repetitivas e complexas do DNA, que eram difíceis de sequenciar com tecnologias mais antigas. Essa inovação não só melhorou a precisão da anotação dos genes, mas também facilitou a identificação de variantes genéticas associadas a características importantes, como sabor, resistência a doenças e produtividade do cacau. A combinação de sequenciamento de terceira geração com outras tecnologias complementares, como o sequenciamento de alta precisão da Illumina, também ajudou a corrigir erros, garantindo uma cobertura de genoma mais robusta e aprofundada (Argout *et al.*, 2017).

O destaque deste aperfeiçoamento foi a reconstrução completa de telômero a telômero dos 10 cromossomos do cacau, o que representou um avanço significativo na compreensão detalhada da organização genética da planta (Argout *et al.*, 2017).

Isso significa que os cientistas conseguiram mapear com mais precisão o material genético, o que está sendo considerado crucial para estudos genéticos e programas de melhoramento do cacau.

Paralelamente a estes estudos, em 2013, um consórcio de cientistas norte-americanos sequenciou o genoma do genótipo Matina 1-6 de cacau, utilizando tecnologias ilumina de segunda geração, assim como foi feito anteriormente com o genótipo Criollo. O cultivar Matina é um dos mais amplamente cultivados, o que torna seu sequenciamento relevante para a agricultura e o melhoramento genético (Motamayor *et al.*, 2013). Esse estudo, além de mapear o genoma, também focou em aspectos semelhantes aos explorados no genoma do Criollo, como a evolução genômica, a identificação de genes relacionados à qualidade dos frutos, a presença de elementos transponíveis (TEs), e as expansões ou contrações de famílias gênicas (Motamayor *et al.*, 2013). O estudo também realizou comparações entre os genomas dos cultivares Matina 1-6 e B97-61/B2 - Criollo, oferecendo resultados sobre as diferenças genéticas que podem influenciar características como a qualidade dos frutos e a resistência da planta. Estas comparações genômicas revelaram que os dois cultivares são muito parecidos, apresentando longos trechos de macrosintenia, e apenas pequenas variações em escala microsintênica (Figura 9), sendo essas diferenças relacionadas a características exclusivas de cada cultivar (Motamayor *et al.*, 2013). Portanto, assim como para outros cultivares, o estudo genômico também avançou os conhecimentos em relação à biologia de cacau, gerando diversos resultados aplicados em programas de melhoramento, como para estudos de associação genômica e seleção genômica (McElroy *et al.*, 2018).

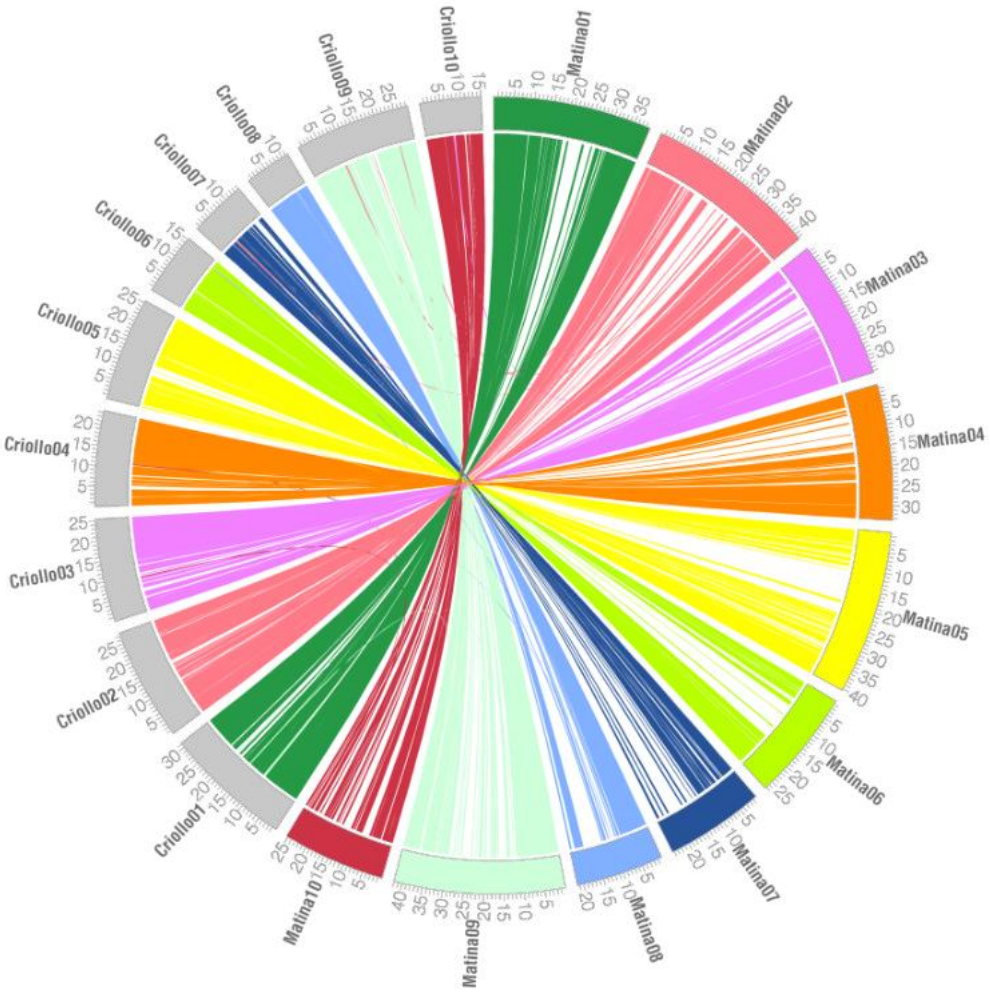

Figura 9. Relação de alta sintenia entre os genomas de *T. cacao*, cultivar B97-61/B2 - Criollo e Matina 1-6. Cada cromossomo de Matina recebeu uma cor e as regiões ortólogas correspondentes no Criollo são conectadas com linhas da mesma cor. As linhas coloridas ligam regiões ortólogas. Extraído de Motamayor *et al.*, 2013.

Em 2024, foi publicado o estudo que reconstruiu o genoma, sem a utilização de uma sequência de referência pré-existente, de três genomas de cacau selvagem de alta qualidade no Alto Amazonas, utilizando tecnologias modernas como leituras longas HiFi e Hi-C, permitindo alcançar resolução cromossômica (Nousias *et al.*, 2024). Esses genomas em escala cromossômica representaram três grupos genéticos fundamentais (Contamana, Iquitos e Nanay) de cacau selvagem e foram comparados com genomas domesticados (Criollo - B97-61/B2 e Matina - Matina 1–6). Foi inferido que os cinco grupos genéticos de cacau, criollo, forasteiro, trinitário, amelonado e

nacional divergiram durante o Pleistoceno, há 1,83–0,69 milhões de anos, e o estudo forneceu recursos cruciais para compreender a diversidade genética e avançar no melhoramento genético (Nousias *et al.*, 2024).

Ainda em 2024, uma pesquisa arqueogenômica analisou resíduos cerâmicos de culturas pré-colombianas para entender a domesticação do cacau em tempos pré-colombianos (Lanaud *et al.*, 2024). Esses resíduos cerâmicos designam traços de substâncias orgânicas, como alimentos, bebidas ou outros materiais que foram absorvidos ou ficaram incrustados nas paredes das cerâmicas utilizadas por culturas pré-colombianas. Foi utilizada a extração de DNA antigo com o Qiagen DNeasy PowerLyzer PowerSoil Kit, seguida de construção de bibliotecas com o kit NEXTflex™ Rapid DNA-Sequencing, e sequenciamento em plataforma Illumina NovaSeq 6000. Além disso, foi aplicada a captura direcionada com Mybaits para sequenciamento de SNPs específicos do cacau.

O estudo revelou o uso do cacau fora de sua área nativa amazônica, especificamente na América do Sul e Central, abrangendo cerca de 5.000 anos. Foi identificada uma mistura genética entre populações de cacau distantes já no Holoceno médio (Criollo, o Amelonado e o Forasteiro, entre outros), impulsionada por interações humanas, o que favoreceu a adaptação do cacau a novos ambientes (Lanaud *et al.*, 2024). Este trabalho destacou a complexa história de domesticação do cacau e sua importância para a formação das populações de cacau atuais.

Em relação ao cupuaçu, na última década, o conhecimento sobre a sua biologia avançou significativamente por meio de estudos moleculares e genômicos. Um exemplo é o trabalho de da Silva *et al.* (2017), o estudo revelou semelhanças entre o cupuaçu e o cacau. Este estudo indicou que o cupuaçu, assim como o cacau, possui 10 cromossomos, medindo aproximadamente 1,5 micrômetros de comprimento, sendo em sua maioria meta ou submetacêntricos (tipos de cromossomos com base na posição do centrômero). Embora a coloração na imagem não evidencie diretamente os centrômeros (Figura 10 - e), a disposição dos cromossomos dispersos na prometáfase sugere essa organização característica. De modo geral, tanto o cacau quanto o cupuaçu são espécies diploides (2n=20) e majoritariamente meta ou submetacêntricos, com regiões CMA+ conservadas,

indicando estabilidade cariotípica no gênero *Theobroma* (Figura 10). Esse padrão conservado no gênero *Theobroma* indica uma estabilidade cariotípica ao longo da evolução, o que pode ter implicações na regulação genômica e nos processos adaptativos dessas espécies amazônicas.

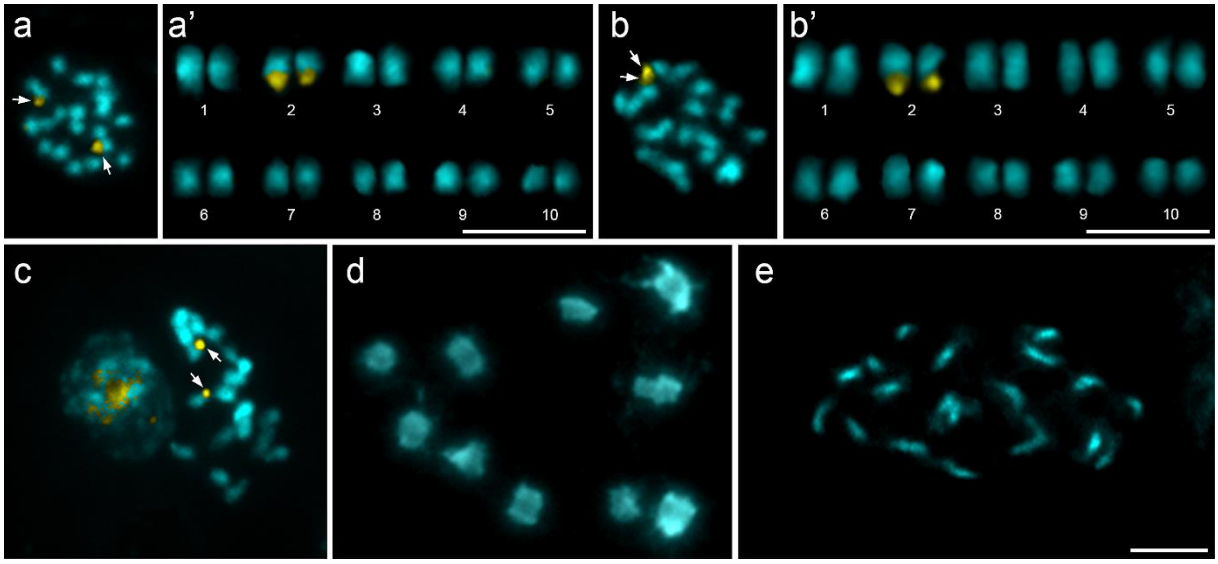

Figura 10. Análise citogenética de espécies do gênero *Theobroma* utilizando coloração com CMA (marcação em amarelo, indicando regiões ricas em GC) e DAPI (marcação em azul, indicando regiões ricas em AT). (a-a') Metáfase e cariograma de *Theobroma cacao ssp leucocarpum*, evidenciando bandas CMA+ (setas). (b-b') Metáfase e cariograma de *T. cacao* cv. Scavina 6, com padrões similares de marcação CMA+ (setas). (c) Metáfase de *Theobroma* bicolor com bandas CMA+ (setas). (d) Diacinese de *T. bicolor*, mostrando a organização dos 10 bivalentes cromossômicos. (e) Prometáfase de *Theobroma grandiflorum*, destacando a organização cromossômica característica do gênero. Barras de escala = 5 µm. Fonte: da Silva *et al.* 2017

O estudo mencionado também revelou que o genoma do cupuaçu tem um tamanho estimado de 448 megabases (Mb), bastante próximo ao genoma do cacau, que foi estimado em 456 Mb. Além disso, por meio de análises de marcadores moleculares do tipo EST (Expressed Sequence Tag) -SSR (Simple Sequence Repeat), que são marcadores moleculares baseados em repetições curtas dentro de sequências expressas, foi constatado que o cupuaçu apresenta baixa heterozigosidade, ou seja, uma baixa diversidade genética em termos de variabilidade alélica (Na: 1,1, He: 0,04, HO: 0), quando comparado ao cacau (Na: 2,4, He: 0,31, HO: 0,12) (da Silva *et al.*, 2017), onde Na (número de alelos), He (heterozigosidade

esperada), HO (heterozigosidade observada). Isso significa que o cupuaçu possui menos variação genética, o que pode impactar sua capacidade de adaptação e resistência, enquanto o cacau apresenta uma maior diversidade genética. Portanto, a baixa diversidade genética observada no cupuaçu pode representar um desafio significativo para programas de melhoramento genético da espécie, especialmente no que se refere à sua capacidade de adaptação a mudanças ambientais e resistência a pragas e doenças. Essa baixa variabilidade genética pode limitar a disponibilidade de alelos favoráveis que poderiam ser utilizados para introduzir características desejáveis na planta. Em contraste, o cacau, com sua maior diversidade genética, oferece um estoque gênico mais amplo, permitindo maior flexibilidade nos programas de melhoramento e maior potencial de desenvolvimento de cultivares adaptados a diferentes condições ou com resistência aprimorada (da Silva *et al.*, 2017).

De uma forma geral os resultados de da Silva *et al.* (2017) sugerem que os genomas do cupuaçu e cacau apresentam muita similaridade, portanto, permitindo que abordagens comparativas sejam utilizadas para aprofundamento do conhecimento entre essas espécies. Recentemente, o sequenciamento do genoma do cupuaçu, em nível de cromossomos, trouxe novos conhecimentos sobre sua evolução e oferece oportunidades para inovações na horticultura. Ao decodificar o genoma do cupuaçu, os pesquisadores identificaram informações genéticas detalhadas que permitem entender melhor a biologia dessa planta, suas características de frutos e sementes, bem como sua resistência a doenças (Alves *et al.*, 2024).

Os autores exploram como as variações genômicas específicas do cupuaçu, como expansões e contrações de famílias gênicas, além de retrotransposons (que são um tipo de TEs) e duplicações de genes, moldaram sua evolução. Essas variações genéticas ajudam a explicar as diferenças entre o cupuaçu e outras espécies de Malvaceae relacionadas, como o cacau e *T. umbraticum*, em aspectos como a forma e o tamanho dos frutos, o amadurecimento e as defesas contra doenças (Alves *et al.*, 2024). Além disso, o estudo destaca como a compreensão desses processos evolutivos pode ser usada para melhorar o cultivo e o manejo do cupuaçu, abrindo caminho para novas estratégias de melhoramento genético e inovações na produção agrícola. A pesquisa, portanto, oferece uma base sólida para aplicar esses

conhecimentos em práticas hortícolas e no desenvolvimento de novas variedades com características desejáveis para o mercado e para os produtores.

A exemplo disso, a identificação de dois loci de autoincompatibilidade no cacau, CH1 e CH4, sendo este último fortemente associado à queda de frutos, e a comparação desses loci em *T. grandiflorum* e *T. umbraticum* revelou padrões distintos: CH1 é altamente conservado entre as três espécies, exceto pela ausência de um gene regulador em *T. umbraticum* (histona metilase H3K4 semelhante ao COMPASS), enquanto CH4 apresenta maior variação, por exemplo, no cacau, esse locus contém duas cópias truncadas adicionais do gene GEX1, além de elementos transponíveis (TEs) e um retrotransposon de LTR-RT da linhagem Copia/Tork, próximo à GEX1, o que pode afetar sua expressão e a formação dos frutos. A hipótese proposta pelo autor é que as múltiplas cópias truncadas do gene GEX1 no cacau, somadas à presença de elementos transponíveis próximos ao locus CH4, podem influenciar a expressão desse gene ou resultar em proteínas GEX1 não funcionais. Esse efeito poderia estar ligado à ausência de abscisão natural nos frutos do cacau. No entanto, essa hipótese ainda precisa ser confirmada experimentalmente (Alves *et al.*, 2024).

Contudo, apesar de algumas diferenças, os autores relatam que as análises genômicas comparativas revelaram que o genoma do cupuaçu compartilha uma alta sintenia e similaridade de nucleotídeos com o cacau, aproximadamente 65% de sintenia (Figura 11), indicando que existem regiões no genoma dessas duas espécies que são conservadas ao longo da evolução (Alves *et al.*, 2024). A análise genômica revelou que *T. grandiflorum* possui 282 famílias de genes exclusivas e 1.160 singletons (genes que não possuem cópias semelhantes em outras espécies analisadas), enquanto compartilha 730 famílias de genes com *T. cacao* e 297 com *T. umbraticum* (Alves *et al.*, 2024). No total, as três espécies dividem 1.816 famílias de genes, das quais apenas duas categorias funcionais apresentaram enriquecimento significativo: uma relacionada ao reconhecimento de pólen (GO:0048544) e outra à localização de proteínas na superfície celular (GO:0034394). As demais famílias de genes exclusivas e compartilhadas incluem muitas associadas à qualidade dos frutos, maturação, características organolépticas, desenvolvimento da planta e resistência a patógenos (Alves *et al.*, 2024).

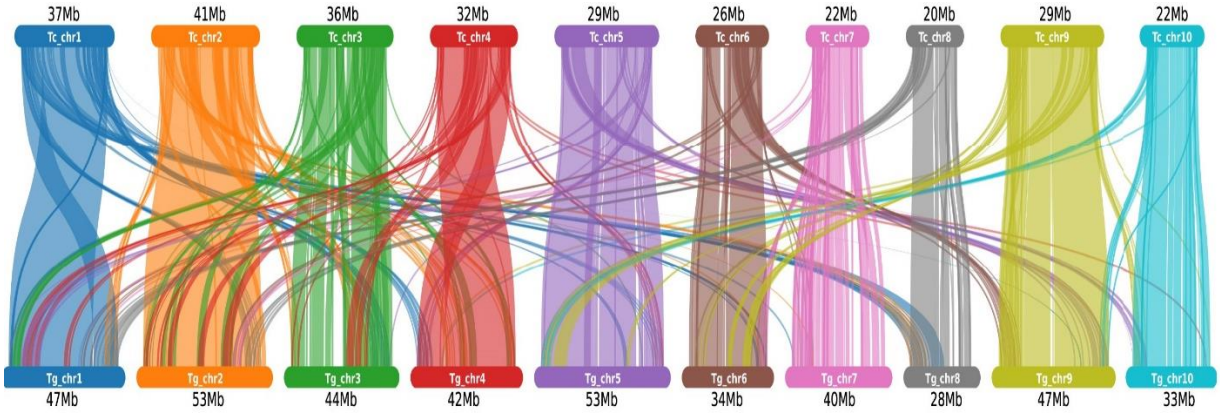

Figura 11. Padrões macrossintênicos entre cupuaçu e cacau, revelando estruturas genômicas conservadas nos 10 cromossomos. Extraído de Alves *et al.*, 2024

Porém, o estudo de Alves *et al*. (2024) investigou apenas o genoma do clone C1074 suscetível à doença vassoura-de-bruxa, restando ainda futuras análises para o clone C174 (cultivar BRS Coari) resistente à mesma doença. Análises comparativas entre esses clones estão previstas como próximos passos para futuros estudos. Dentre as perguntas que se pretendem responder, destacam-se quais são os genes específicos que conferem resistência ao *Moniliophthora perniciosa* e *M. roreri* no clone C174, como diferem os padrões de expressão gênica entre os clones C1074 e C174 durante a infecção pelo fungo, quais vias metabólicas e mecanismos de defesa estão ativados no clone resistente e ausentes ou menos ativos no clone suscetível, se existem variantes genéticas ou polimorfismos associados à resistência que possam ser utilizados como marcadores genéticos para a seleção de cultivares resistentes e como a resistência observada no clone C174 pode ser integrada em programas de melhoramento para desenvolver novas variedades de cupuaçu com maior resistência a doenças (comunicação pessoal).

Essas perguntas visam elucidar os mecanismos genéticos e moleculares subjacentes à resistência, proporcionando subsídios valiosos para estratégias de melhoramento genético mais eficazes e sustentáveis. Com a resposta a essas questões, será possível identificar e manipular genes-chave que aumentem a resistência das plantas às doenças, otimizar práticas de manejo e desenvolver cultivares que não apenas mantenham a produtividade, mas também garantam a

qualidade dos frutos em condições adversas. Além disso, a integração dessas descobertas em programas de melhoramento pode acelerar o desenvolvimento de variedades mais robustas, reduzindo a dependência de métodos químicos de controle e promovendo uma agricultura mais sustentável e ambientalmente amigável.

Em relação a *T. umbraticum*, o estudo sobre a sequência do seu genoma revela informações importantes sobre sua relação com o gênero *Theobroma*. Sendo *T. umbraticum* um gênero irmão de *Theobroma* dentro da família Malvaceae, mais especificamente na subfamília Byttnerioideae e na tribo *Theobroma*teae (Colli-Silva; Richardson; Pirani, 2023). Essa espécie, que exibe uma morfologia única, apresenta uma rica composição em ácidos graxos insaturados e de cadeia longa, similar ao que é observado em cupuaçu (Alves *et al.*, 2024).

Em análises comparativas de genomas de plantas, uma técnica que pode ser utilizada está relacionada à análise topográfica ou espacial dos genes, que podem estar distribuídos na forma de duplicados dispersos, tandem, próximos ou derivados de eventos WGD (segmentais). Essas duplicações podem originar-se de diferentes eventos evolutivos, como duplicações de genoma inteiro (WGD), onde todo o conjunto de cromossomos é duplicado, resultando em múltiplas cópias de cada gene; duplicações tandem, onde genes são duplicados adjacentes uns aos outros no mesmo cromossomo; duplicações proximais, onde genes duplicados estão próximos, mas não necessariamente em sequência direta; e duplicações dispersas, onde genes duplicados estão espalhados por diferentes regiões do genoma (Hanada *et al.*, 2008; Zou *et al.*, 2009; Wang *et al.*, 2016; Sangi *et al.*, 2021; Alves *et al.*, 2024).

Essa diversidade de padrões de duplicação permite a expansão e diversificação das famílias gênicas, contribuindo para a adaptação e evolução das espécies. No caso do cupuaçu, a identificação desses padrões de duplicação é crucial para compreender como determinados genes relacionados à resistência a patógenos, foram mantidos e diversificados ao longo do tempo (Alves *et al.*, 2024). Por exemplo, genes duplicados através de WGD podem fornecer redundância funcional, permitindo que algumas cópias possam desenvolver novas funções, pelo mecanismo de divergência funcional, sem comprometer a sobrevivência da planta. Já duplicações

tandem e proximais podem facilitar a co-regulação de genes envolvidos em respostas específicas a estresses bióticos e abióticos (Hanada *et al.,* 2008; Zou *et al.,* 2009).

Neste sentido, análises baseadas em Ontologias Gênicas (Gene Ontology do inglês, ou GO) podem contribuem para elucidar como eventos de duplicação e seleção podem influenciar rotas metabólicas fundamentais, revelando potenciais alvos para estudos de melhoramento e aplicações biotecnológicas. O GO apresenta um vocabulário padronizado que permite descrever funções gênicas de maneira consistente, agrupando-as em três categorias principais: processos biológicos, funções moleculares e componentes celulares. Nesse contexto, o enriquecimento do termo “processo biossintético de flavonoides” (GO:0009813) em *T. grandiflorum* está vinculado a uma duplicação gênica em tandem de um gene que codifica a enzima naringenina 2-oxoglutarato 3-dioxigenase, evidenciada em dois loci (*TgrandC1074G00000004751* e *TgrandC1074G00000004753*), ambos sob seleção positiva (Alves *et al.*, 2024). Duplicações em tandem, como essa, envolvem genes adjacentes no mesmo cromossomo e são frequentemente associadas à co-regulação de respostas específicas, como a adaptação a estresses biológicos. Essa duplicação possivelmente desempenha papel essencial na síntese de flavonoides, antocianidinas, catequinas e proantocianidinas, sugerindo sua relevância para a adaptação e diversificação funcional da espécie (Alves *et al.*, 2024). No caso do cupuaçu, diversas outras duplicações foram identificadas podendo estar relacionadas ao desenvolvimento das características organolépticas da fruta, como seu sabor e aroma, além de desempenharem um papel nas propriedades antioxidantes. A presença dessas duplicações em *T. grandiflorum* sugere uma estratégia evolutiva que pode ter permitido a diversificação e a especialização dessas vias metabólicas, enquanto em *T. umbraticum* não se observa o mesmo enriquecimento, o que indica uma diferença significativa no padrão genético entre as duas espécies. (Alves *et al.*, 2024).

Paralelamente, o estudo realizado Colli-Silva, Richardson, Pirani, (2023), apresenta um conjunto de dados taxonômicos sobre ocorrências de espécimes preservados dos gêneros *Theobroma* e *Herrania*. Esses dados são fundamentais para a compreensão da biodiversidade e distribuição geográfica dessas espécies, que incluem plantas de grande relevância econômica, como o cacau e o cupuaçu. O artigo,

destaca a importância dos dados de biodiversidade para a pesquisa e conservação, enfatizando a riqueza e diversidade dos gêneros. O estudo sugere que as informações coletadas podem facilitar análises filogenéticas, estudos sobre a ecologia e a biogeografia dessas plantas, além de contribuírem para a conservação das espécies.

As informações foram integradas em um sistema conhecido como *speciesLink*, que tem como objetivo facilitar o acesso e o compartilhamento de dados sobre a biodiversidade. Este sistema é uma ferramenta crucial para a promoção da pesquisa científica e para a preservação do conhecimento sobre a flora brasileira, especialmente em relação às espécies menos estudadas como *T. umbraticum*.

Atualmente, há 11 genomas sequenciados e publicados de *T. cacao*, enquanto *T. grandiflorum* e *T. umbraticum* possuem apenas um genoma publicado cada, destacando o estágio incipiente das publicações sobre espécies amazônicas como o cupuaçu, em contraste com o cacau (Tabela 2). Esses genomas completos sequenciados e publicados estão disponíveis em: https://www.plabipd.de/plant_genomes_pa.ep e https://www.ncbi.nlm.nih.gov/.

Tabela 2. Genomas publicados de espécies do gênero *Theobroma* e *Herrania*, com informações sobre o número de genomas, tamanho do genoma (em Mb), cultivares e respectivos bioprojetos (NCBI).

Espécie	**Nº de Genomas Publicados**	**Tamanho do Genoma (Mb)**	**Cultivar**	**Bioprojeto (NCBI)**
Theobroma cacao	11	324.7 380,7 380.9 380,9 382.4 382,4 405,5 405.5 374,3 380.7 346	B97-61/B2 Sca6 IMC 67 IMC_67 SCA 6 SCA_6 Pound 7 Pound 7 Clone 26 Clone 26 Matina 1-6	PRJEB14326 PRJNA558793 PRJNA982528 PRJEB64280 PRJNA982528 PRJEB64280 PRJNA982528 PRJEB64280 PRJNA1083235 PRJNA1086984 PRJNA51633
Theobroma grandiflorum	1	423,9	C1074	PRJNA691024
Theobroma umbraticum	1	234	Fairchild	PRJNA383741

2.5. Genes relacionados à qualidade dos frutos e sua importância

Estudar, sob uma perspectiva genômica, a qualidade dos frutos em espécies do gênero *Theobroma* é de suma importância, pois suas sementes possuem alto valor alimentício e são amplamente utilizadas na produção de chocolates e outros produtos de interesse econômico. O cacau, por exemplo, possui óleo volátil que confere sabor característico, bem como 1,5 a 3% de teobromina, alcaloide estimulante semelhante à cafeína (Cuatrecasas, 1964; Argout *et al*., 2011). Além disso, a composição nutricional e sensorial dos frutos, como a polpa do cupuaçu, apresenta propriedades únicas (ésteres e aldeídos) que agregam valor aos produtos derivados, como sucos, chocolates e cosméticos. Além disso, a polpa é rica em compostos bioativos, como flavonoides e polifenóis, que conferem propriedades antioxidantes benéficas à saúde, porém, baixo teor de Teobromina comparado ao cacau (Genovese; Lannes, 2009). Seu alto teor de fibras contribui para a saúde digestiva, enquanto os ácidos graxos presentes na semente, como o ácido oleico e o ácido linoleico, são valorizados na formulação de chocolates, manteigas vegetais e cosméticos hidratantes (Genovese; Lannes, 2009; Argout *et al*., 2011). A identificação e caracterização de genes relacionados a essas características são fundamentais para o desenvolvimento de estratégias biotecnológicas que visam aprimorar a qualidade dos frutos através do melhoramento genético.

Por exemplo, técnicas avançadas de biotecnologia, como a seleção assistida por marcadores, a edição genética utilizando CRISPR-Cas9 e a transgenia, permitem a modulação precisa de genes de interesse, acelerando o processo de desenvolvimento de cultivares superiores (Etaware, 2021). A utilização de sequenciamento de alto rendimento e análises genômicas comparativas facilita a identificação de loci genéticos associados a características desejáveis, como resistência a doenças, teores elevados de compostos bioativos e melhoria das características organolépticas dos frutos (Alves *et al.,* 2024).

Além disso, a integração de abordagens genômicas, como a genômica funcional e a transcriptômica, possibilita uma compreensão mais profunda dos mecanismos moleculares que regulam essas características (Junior *et al.*, 2012;

Falcão *et al.*, 2022). Isso permite a identificação de redes de genes e vias metabólicas que podem ser alvo de intervenções para otimizar a produção e a qualidade dos frutos. Neste sentido, a aplicação de bioinformática na análise e interpretação dos dados genômicos é essencial para mapear a diversidade genética existente e identificar variantes genéticas que conferem vantagens adaptativas e comerciais (Diniz; Canduri, 2017).

Por outro lado, o melhoramento genético assistido por dados genômicos não apenas pode auxiliar na melhora da produtividade e a resistência das plantas a patógenos e estresses ambientais, mas também pode assegurar a consistência e a qualidade dos frutos, atendendo às demandas do mercado e promovendo a sustentabilidade econômica dos produtores. Por exemplo, a incorporação de genes de resistência a fungos do gênero *Moniliophthora* pode reduzir significativamente as perdas de produção, enquanto a otimização da biossíntese de flavonoides e terpenoides pode aprimorar o sabor e as propriedades antioxidantes dos frutos (Kilaru; Bailey; Hasenstein, 2007; Barbosa *et al.*, 2018; Falcão *et al.*, 2022).

Além do mais, a biotecnologia facilita a conservação da diversidade genética das espécies cultivadas, permitindo a criação de bancos de genes e germoplasma que preservam variantes genéticas importantes para futuras melhorias. A clonagem de genes e a transferência de genes de espécies relacionadas também representam estratégias viáveis para introduzir novas funcionalidades nos cultivares de cacau e cupuaçu, ampliando seu potencial comercial e nutricional (Doyle; Hotton, 1991; Farrokhi *et al.*, 2006).

Portanto, a caracterização de genes relacionados à qualidade dos frutos no gênero *Theobroma*, através de estudos genômicos e de bioinformátoca, é crucial para o avanço do melhoramento genético. Essas iniciativas não apenas podem promover o desenvolvimento de cultivares mais resistentes e de alta qualidade, mas também contribuem para a sustentabilidade e a competitividade da cultura no mercado global, beneficiando tanto os agricultores quanto os consumidores finais.

2.5.1. Ácidos graxos e lipídeos

Os lipídios de armazenamento nas sementes, sobretudo os triacilgliceróis, fornecem reservas de carbono e energia na germinação (Figueira; Luccas, 2022). A manteiga de cacau, representando cerca de 50% do peso da semente seca, possui alto teor de estearato (30-37%), um ácido graxo saturado que garante um ponto de fusão elevado (34-38°C), importante para a textura do chocolate (Griffiths; Harwood, 1991; Argout *et al*., 2011). Esse ponto de fusão relativamente alto faz com que a manteiga de cacau permaneça sólida à temperatura ambiente, mas derreta facilmente quando consumida, característica crucial para a textura e sensação no paladar do chocolate.

O perfil único de ácidos graxos da manteiga de cacau melhora as qualidades olfativas de chocolates e confeitos, além de torná-la valiosa para produtos cosméticos e farmacêuticos (Argout *et al*., 2011). Além dos triacilgliceróis, outros lipídios presentes nas sementes, como fosfolipídios e esteróis, desempenham papéis importantes na integridade das membranas celulares e na sinalização intracelular, influenciando diretamente a qualidade e a estabilidade dos produtos derivados (Matos *et al*., 2020).

Estudos genômicos, como o de Argout *et al*. (2011), identificaram 84 genes ortólogos de *Arabidopsis* em cacau Criollo que estão potencialmente envolvidos na biossíntese de lipídios, indicando uma complexa regulação genética responsável pela síntese dos ácidos graxos que compõem a manteiga de cacau. Esses genes incluem aqueles que codificam enzimas chave no caminho biossintético dos ácidos graxos, como acetil-coa carboxilase (ACC), ácido graxo sintase (FAS) e desaturases. A expressão diferencial desses genes em diferentes estágios de desenvolvimento da semente sugere uma regulação altamente controlada, que pode ser manipulada para otimizar a composição lipídica (Argout *et al*., 2011). Uma visão detalhada de alguns dos principais genes envolvidos na regulação e síntese dos lipídios são detalhados (Tabela 3), contribuindo para um entendimento mais amplo da base genética responsável por este fator de qualidade dos frutos em *Theobroma*.

Tabela 3. Lista de genes de *Theobroma cacao* que codificam enzimas-chave envolvidas na via de biossíntese de lipídios de armazenamento. Homólogos em *Arabidopsis thaliana.*

Gene	Proteína	Função	Referência
CAC2	Proteína transportadora de carboxila de biotina da acetil-CoA carboxilase	O gene CAC2 é expresso ao máximo em órgãos e tecidos que sintetizam ativamente ácidos graxos para lipídios de membrana ou deposição de óleo. O padrão de expressão observado de CAC2 espelha ao relatado para o gene CAC1. Acumula-se nos tecidos que sintetizam ácidos graxos. Expressa-se principalmente em síliquas, folhas em desenvolvimento e flores, presentes em raízes e embriões e, em menor grau, em folhas maduras.	Sun *et al.*, 1997
TT2	Fator de transcrição TT2	o gene TT2 regula a biossíntese de proantocianidinas no tegumento da semente de *Arabidopsis* , que são compostos fenólicos importantes. Recentemente, foi descoberto que o TT2 também desempenha um papel na inibição da biossíntese de ácidos graxos (FA) no embrião da semente. No entanto, o mecanismo exato pelo qual o TT2 suprime o acúmulo de ácidos graxos ainda não está claro. Isso sugere que TT2 pode ter funções além da síntese de PAs, influenciando outros processos metabólicos na semente, mas as interações moleculares subjacentes à inibição dos ácidos graxos permanecem desconhecidas.	Wang *et al.*, 2014
KASI	beta-cetoacil-[proteína transportadora de acil] sintase I	A enzima β-cetoacil-[proteína transportadora de acil] sintase I (KASI), que catalisa o alongamento na síntese de ácidos graxos (de novo). A deficiência de KASI interrompe o desenvolvimento embrionário antes do estágio globular e reduz drasticamente os níveis de ácidos graxos nas sementes (cerca de 33,6% dos níveis presentes no tipo selvagem). Esses resultados indicam que a síntese de ácidos graxos é essencial para o crescimento das plantas, pois afeta múltiplos processos (efeitos pleiotrópicos). Além disso, o suprimento adequado de lipídios polares é fundamental para a divisão celular e o desenvolvimento dos cloroplastos, revelando a importância crucial da síntese de ácidos graxos no desenvolvimento dos plastídios e, consequentemente, no crescimento e desenvolvimento das plantas como um todo.	Wu; Xue, 2010
ACP1	Proteína transportadora de acila 1	A proteína transportadora de acil (ACP) é um cofator essencial para a biossíntese de novos ácidos graxos em todos os organismos vivos, pois desempenha um papel fundamental no transporte de intermediários durante o processo de síntese de ácidos graxos. Essa função torna a ACP crucial para a produção de lipídios nas células. A isoforma ACP-1, em particular, é a mais predominante nas sementes, sugerindo que ela tem um papel importante na síntese de ácidos graxos durante o desenvolvimento das sementes, quando a produção de lipídios é especialmente ativa.	Branen; Chiou; Engeseth, 2001
FATA	Proteína	As tioesterases ligadas à proteína transportadora de acil (ACP) são enzimas responsáveis por finalizar a síntese de ácidos graxos dentro dos plastídeos das	Moreno-Pérez *et al.*, 2012

	transportadora de oleoil-acil tioesterase 1, cloroplástica	plantas, através da quebra (hidrólise) dos complexos acil-ACP. Esse processo é crucial para liberar os ácidos graxos que foram formados, permitindo que eles sejam usados em outras funções celulares. Entre as várias famílias de tioesterases encontradas em plantas, a FatA desempenha um papel importante na exportação dos ácidos graxos de cadeia C18, que são os principais componentes usados para formar glicerolipídios, essenciais para a formação das membranas celulares e outros lipídios estruturais das plantas.	
FAD2	Delta (12)-dessaturase de ácido graxo	Os ácidos graxos poliinsaturados, como linoleato e alfa-linolenato, são componentes essenciais das membranas celulares e fundamentais para a nutrição humana, pois desempenham papéis críticos em várias funções fisiológicas. A enzima chave responsável pela síntese desses ácidos graxos é a oleato dessaturase presente no retículo endoplasmático das células vegetais. Em *Arabidopsis thaliana*, a atividade dessa enzima é regulada pelo gene FAD2 (locus de dessaturação de ácidos graxos 2), que controla a conversão de ácido oleico em linoleato, um precursor dos ácidos graxos essenciais. Isso destaca a importância do controle genético sobre a síntese de lipídios essenciais tanto para a estrutura celular quanto para a saúde humana.	Okuley *et al.*, 1994
mccA	Sulfito redutase dissimilatória MccA	A enzima biotina, chamada 3-metilcrotonil-CoA carboxilase (MCCase), é essencial para o metabolismo da leucina e dos isoprenóides, que são compostos importantes tanto no catabolismo de aminoácidos quanto na produção de moléculas como hormônios e pigmentos nas plantas. Essa enzima catalisa uma reação necessária para o catabolismo da leucina, sendo formada por duas subunidades: uma que contém biotina (MCC-A) e outra que não contém biotina (MCC-B). Embora a sequência da subunidade MCC-A já tenha sido determinada, a estrutura primária da subunidade MCC-B ainda é desconhecida. Isso destaca que, apesar do progresso na compreensão de uma parte da enzima, ainda existem lacunas no conhecimento sobre sua composição total.	McKean *et al.*, 2000
KCS1	3-cetoacil-CoA sintase 1	Contribui para a biossíntese da cera cuticular e da suberina. Envolvido em ambas as vias de síntese de cera de descarbonilação e redução de acil. Enzima de condensação elongase principalmente ativa com ácidos graxos saturados	Todd; Post-Beittenmiller; Jaworski, 1999
LACS2	Acil-CoA sintetase 2 de cadeia longa	As atividades da acil-CoA sintetase de cadeia longa (LACS) são codificadas por uma família de pelo menos nove genes em *Arabidopsis* (*Arabidopsis thaliana*). Essas enzimas têm papéis na síntese de lipídios, no catabolismo de ácidos graxos e no transporte de ácidos graxos entre os compartimentos subcelulares. Os genes LACS2 (At1g49430) é expresso em tecidos jovens, em rápida expansão, e nas folhas a expressão é limitada a células das camadas epidérmicas adaxial e abaxial, sugerindo	Schnurr; Shockey; Browse, 2004

		que a enzima LACS2 pode atuar na síntese de cutina ou cuticular ceras.	

2.5.2. Metabólitos secundários e flavonoides

Os flavonóides são um grupo diversificado de metabólitos secundários de plantas que desempenham muitos papéis importantes durante o desenvolvimento da planta e destacam-se nos frutos de cacau e cupuaçu (Silva *et al*., 2020). Eles estão envolvidos na defesa da planta contra insetos, patógenos e micro-organismos, na absorção de radicais livres e luz ultravioleta e na atração de simbiontes e polinizadores benéficos (Owens *et al.*, 2008b). As proantocianidinas são polímeros flavonoides que estão presentes em grandes quantidades nas sementes de cacau (Argout *et al*., 2011). *Arabidopsis* contém três classes principais de flavonóides: as antocianinas (pigmentos vermelhos a roxos), os flavonóis (pigmentos incolores a amarelo pálido) e as proantocianidinas (pigmentos incolores que se tornam marrons), que também são conhecidos como taninos condensados e que também estão presentes em grandes quantidades nas sementes de cacau (Argout *et al*., 2011). As proantocianidinas demonstraram ser os principais determinantes das propriedades agronômicas e nutricionais das sementes (Shirley, 1998; Nesi *et al*., 2001).

Além disso, os flavonóides desempenham um papel crucial na resistência a doenças e na qualidade dos frutos (Owens *et al.*, 2008b). A biossíntese de flavonóides é regulada por uma complexa rede de genes que inclui fatores de transcrição, como MYB e bHLH, que controlam a expressão de enzimas-chave no caminho biossintético (Nesi *et al.*, 2001). Em *Theobroma cacao*, por exemplo, genes como F3H (flavanone 3-hydroxilase) e DFR (dihidroflavonol 4-reductase) são essenciais para a produção de flavonóides que conferem propriedades antioxidantes e melhoram a qualidade sensorial do chocolate (Owens *et al.*, 2008a; Owens *et al.*, 2008b).

No cupuaçu, estudos indicam que a diversidade de flavonóides está associada não apenas à qualidade dos frutos e atividade antioxidante, mas também à resistência a patógenos como *M. perniciosa*, o agente causador da vassoura-de-bruxa (Pugliese *et al.*, 2013; Silva *et al*., 2022). A manipulação genética de vias biossintéticas

de flavonóides pode, portanto, ser uma estratégia promissora para desenvolver cultivares mais resistentes e com melhor qualidade de frutos. Esses genes estão envolvidos nas etapas cruciais da biossíntese de flavonoides em *T. cacao* e seus homólogos em *A. thaliana* fornecem uma comparação importante para entender a função desses genes em plantas relacionadas (Tabela 4).

Tabela 4. Lista dos genes de *T. cacao* envolvidos na codificação de enzimas chave da via de biossíntese de flavonoides. Homólogos em *A. thaliana.*

Gene	Proteína	Função	Referência
4CL	4-coumarate--CoA ligase	Produz tioésteres CoA de uma variedade de ácidos cinâmicos hidroxi e metoxi-substituídos, que são usados para sintetizar vários compostos derivados de fenilpropanóides, incluindo antocianinas, flavonóides, isoflavonóides, cumarinas, lignina, suberina e fenólicos de parede. Quatro isoformas de 4CL foram identificadas em *Arabidopsis* (*Arabidopsis thaliana*). A análise filogenética revela que 4CL1, 4CL2 e 4CL4 estão mais intimamente relacionados entre si do que com 4CL3, sugerindo que os dois grupos podem ter diferentes funções biológicas. A análise do promotor-GUS mostra que 4CL1 e 4CL2 são expressos em células lignificantes. Em contraste, 4CL3 é expresso em uma ampla gama de tipos de células, e 4CL3 adquiriu um papel distinto no metabolismo de flavonóides.	Li *et al.*, 2015
F3H	Naringenina,2-oxoglutarato 3-dioxigenase	O gene F3H (flavanona-3-hidroxilase) é amplamente utilizado para manipular a síntese de flavonoides, tanto em microrganismos quanto em plantas, muitas vezes em conjunto com outros genes que codificam enzimas da via dos flavonoides. A ausência de F3H em algumas linhagens de plantas, como *Arabidopsis* , leva à produção de flavonoides incomuns, como os flavonoides 3-desoxi. Esses compostos, que não contêm um grupo hidroxila no carbono 3, podem se acumular em sementes. Além disso, a enzima antocianidina sintase pode parcialmente compensar a função da F3H em plantas, o que sugere um nível de redundância funcional entre esses genes no metabolismo de flavonoides.	Owens *et al.*, 2008a
FLS1	Flavonol sintase/flavanona 3-hidroxilase	A biossíntese de antocianinas e flavonóis, que são dois tipos de flavonoides produzidos pela via metabólica dos fenilpropanóides. Ambos são derivados do diidroflavonol, mas seguem caminhos diferentes. Para a produção de antocianinas, o diidroflavonol é convertido em antocianinas pela enzima diidroflavonol 4-redutase (DFR), enquanto para a produção de flavonóis, o diidroflavonol é convertido pela flavonol sintase (FLS). No genoma de *Arabidopsis thaliana*, existem seis genes FLS diferentes, nomeados FLS1 a FLS6, que estão	Owens *et al.*, 2008b

		envolvidos na produção de flavonóis. Esses genes desempenham um papel central na regulação da síntese de flavonoides, influenciando a cor, defesa da planta e outros aspectos relacionados ao metabolismo secundário.	
UFGT	Antocianidina 3-O-glicosiltransferase UFGT	a importância da enzima UDP-glicose óide 3-O-glucosiltransferase (UFGT) no processo de biossíntese de antocianinas na uva, é essencial para a adição de um grupo de glicose às antocianinas, o que é um passo crucial para a estabilização e armazenamento desses pigmentos na planta. A expressão do gene UFGT foi restrita à casca da baga, a única parte onde as antocianinas estavam acumuladas. Isso indica que a UFGT desempenha um papel específico e localizado na produção de antocianinas, que são responsáveis pela coloração da fruta.	Kobayashi *et al.*, 2001

2.5.3. Terpenóides

Os terpenóides constituem uma grande família de compostos naturais e desempenham diversos papéis nas plantas como hormônios, pigmentos e nas interações e defesa planta-ambiente (Moses *et al.*, 2013). São componentes principais de resinas, óleos essenciais e aromas (Liu *et al.*, 2005). Dentre eles, duas subclasses de terpenóides estão particularmente envolvidas nos aromas: monoterpenos (C10), que representam compostos aromáticos que são a base de essências florais e óleos essenciais, e sesquiterpenos (C15), que também podem constituir uma resposta de defesa das plantas contra microorganismos ou agressão de insetos. Em comparação ao cacau a granel, as variedades de cacau de sabor fino, como Criollo e Nacional, apresentam níveis mais altos de monoterpenos, como o linalol. O linalol é um álcool monoterpeno acíclico encontrado no aroma floral de diversas plantas, incluindo *Clarkia breweri*, que se destaca por sua concentração notável desse composto. Os monoterpenos, como o linalol, são responsáveis pelas notas florais e frutadas características dessas variedades de cacau de sabor fino (Argout *et al*., 2011).

Os terpenóides compreendem uma classe diversa de metabólitos vegetais; a maioria deles são conhecidos como compostos, sabores ou aromas relacionados à defesa. Além disso, metabólitos primários como carotenóides, clorofila, ácido

abscísico, transportadores de elétrons de quinona, esteróides e giberelinas também são derivados de terpeno (Lange; Ghassemian, 2003; McGarvey; Croteau, 1995).

Os terpenóides agrupados sob metabólitos primários são os componentes essenciais para as funções celulares básicas, e incluem cadeias prenil de quinonas (ubiquinona e plastoquinona), pigmentos fotossintéticos (fitol na clorofila e pigmentos carotenóides), esteróis (a maioria dos quais é responsável pela estabilidade da membrana), hormônios de crescimento (giberelinas) e proteínas preniladas (Bohlmann; Keeling, 2008; Moses *et al.*, 2013). A prenilação de proteínas, um tipo de modificação pós-traducional, facilita sua associação com membranas celulares, o que é altamente essencial para sua função, por exemplo, GTPases (Bracha-Drori *et al.*, 2008).

Os terpenóides, que desempenham um papel crucial na realização de várias funções ecológicas em resposta a fatores bióticos e abióticos, são classificados como produtos do metabolismo secundário, por exemplo, atrativos de polinização, inibidores de herbívoros, inseticidas/repelentes, compostos antibacterianos, moléculas alelopáticas e tóxicas, antioxidantes, moléculas de termotolerância e fotoproteção, entre outras funções (Gershenzon; Dudareva, 2007). A Tabela 5 apresenta uma lista de genes de *Theobroma cacao* que codificam enzimas-chave na via de biossíntese de terpenoides, essenciais para a produção desses compostos bioativos.

Tabela 5. Lista de genes de *Theobroma cacao* que codificam enzimas-chave envolvidas na via de biossíntese de terpenoides.

Gene	Proteína	Função	Referência
GPS1	Citocromo P450 705A22	Codificam enzimas importantes para a síntese de precursores de terpeno. Os terpenóides compreendem uma classe diversa de metabólitos vegetais; a maioria deles são conhecidos como compostos, sabores ou aromas relacionados à defesa. Geranil difosfato sintase (GPS) é geralmente considerado responsável pela biossíntese de precursores de monoterpeno apenas. O o geranil difosfato sintase é necessária para a biossíntese de giberelinas.	Van Schie *et al.*, 2007
TPS14	S-(+)-linalol sintase, cloroplástico	codificar proteínas que catalisam a formação dos principais voláteis florais.	Chen *et al.*, 2003
TPS03	Terpenóide sintase 3, cloroplástico	Envolvida na biossíntese de sesquiterpenos (compostos com 15 carbonos, como o (E,E)-alfa-farneseno) quando usa FPP (farnesil difosfato) como substrato. O produto principal dessa	Fäldt *et al.*, 2003

		reação é o (E,E)-alfa-farneseno, com pequenas quantidades de isômeros como (Z,E)-alfa-farneseno e (E,E)-beta-farneseno. Quando a enzima utiliza GPP (geranil difosfato) como substrato, também pode sintetizar monoterpenos (compostos com 10 carbonos), como o (E)-beta-ocimeno, com produtos secundários como (Z)-beta-ocimeno e mirceno. Isso indica que a enzima tem a capacidade de atuar em diferentes substratos para produzir compostos terpenoides de diversas cadeias de carbonos, sugerindo sua versatilidade na biossíntese de terpenos.	
TPS10	Provável alfa,alfa-trealose-fosfato sintase [formadora de UDP] 10	Codifica uma enzima supostamente envolvida na biossíntese da trealose. A proteína tem um domínio semelhante à trealose sintase (TPS) que pode ou não estar ativo, bem como um domínio semelhante à trealose fosfatase (TPP).	Fäldt *et al.*, 2003
TPS27	1,8-cineol sintase 1, cloroplástico	Está envolvida na biossíntese de monoterpenos, compostos com 10 carbonos. O principal produto dessa via biossintética é o 1,8-cineol, que representa 52% da produção total. Outros compostos produzidos em menores quantidades incluem sabineno (14,5%), mirceno (13,3%), beta-pineno (7,8%), limoneno (4,0%), beta-ocimeno (2,7%), alfa-terpineol (2,4%), alfa-pineno (1,9%), terpinoleno (0,8%) e alfa-tujeno (0,6%). Isso destaca a diversidade de compostos voláteis gerados pela ação dessa enzima, muitos dos quais têm significância em termos de aroma e funções ecológicas, como atração de polinizadores ou defesa contra herbívoros.	Chen *et al.*, 2004
TPS02	(E)-beta-ocimeno sintase, cloroplástico	Responsável pela biossíntese de monoterpenos, compostos com 10 átomos de carbono. Quando utiliza o geranil difosfato (GPP) como substrato, o produto principal é o (E)-beta-ocimeno, que é acompanhado por menores quantidades de (Z)-beta-ocimeno e mirceno. Além disso, ao usar o farnesil difosfato (FPP) como substrato, a enzima também consegue sintetizar sesquiterpenos, que têm 15 átomos de carbono, sendo o (E,E)-alfa-farneseno o produto principal, com (Z,E)-alfa-farneseno e (E,E)-beta-farneseno como produtos secundários. Essa informação destaca a versatilidade da enzima em produzir diferentes tipos de terpenos a partir de diferentes substratos, refletindo a complexidade da biossíntese de terpenoides nas plantas.	Huang *et al.*, 2010

Um estudo realizado por Ferraz dos Santos *et al*. (2022) investigou a expressão de genes candidatos (TgSTP1, TgWRKY33 e TgUBA1) nas polpas e sementes de frutos de diferentes estágios de maturação (jovens, em maturação e maduros) de genótipos de *T. grandiflorum* com diferentes níveis de resistência à vassoura-de-bruxa. Os resultados mostraram diferenças significativas associadas à

qualidade dos frutos, que podem estar relacionadas ao metabolismo de compostos bioativos, como os terpenóides. Esses genes foram selecionados com base em estudos anteriores que indicaram sua relevância em processos ligados à qualidade do fruto e à resistência a patógenos, atuando em diferentes níveis das vias de sinalização celular. Além disso, os genes de *T. grandiflorum* envolvidos no metabolismo primário e na resposta ao estresse foram comparados com seus ortólogos em *T. cacao* e *Arabidopsis thaliana*. Esses resultados sugerem que a produção de terpenóides, que desempenham um papel importante na defesa contra patógenos, como o agente da vassoura-de-bruxa, pode estar diretamente ligada à expressão desses genes, influenciando tanto a resistência quanto a qualidade do fruto em *T. grandiflorum* (Ferraz dos Santos *et al.*, 2022).

Além dos genes identificados por Ferraz dos Santos *et al.* (2022), a literatura revela a presença de outros genes com funções semelhantes à aqueles envolvidos no metabolismo primário e na resposta ao estresse em *T. grandiflorum* em diferentes espécies vegetais (Tabela 6). Esses genes adicionais também estão relacionados à adaptação ao estresse biótico e abiótico, além de desempenharem papéis cruciais na regulação de processos metabólicos (Tabela 6).

Tabela 6. Genes de *Theobroma grandiflorum* envolvidos no metabolismo primário e na resposta ao estresse, e seus ortólogos identificados em *Theobroma cacao* e *Arabidopsis thaliana*.

Gene	**Proteína**	**Função**	**Referência**
STP1	Proteína transportadora de açúcar 1	Na polpa de *T. grandiflorum,* STP1 mostrou alta expressão nos estágios de maturação e maduro, particularmente no genótipo 174, no qual o estágio maduro era o mais doce. O transportador de hexose responsável por realizar a captação ativa de açúcares simples, chamados hexoses, nas células. Esse processo ocorre por meio de um mecanismo de simporte, onde o transportador move as hexoses juntamente com íons hidrogênio (H+) para dentro da célula, usando o gradiente de H+ para impulsionar o transporte. Além de glicose, o transportador consegue movimentar uma variedade de outras hexoses e derivados, como 3-O-metilglicose, frutose, xilose, manose, galactose, fucose, 2-desoxiglicose e arabinose. Esse transportador também tem um papel em conferir sensibilidade à galactose, especialmente em mudas, sugerindo uma resposta fisiológica ou um efeito regulador no crescimento jovem diante da presença desse açúcar.	Ferraz dos Santos *et al.*, 2022
PAB2	Proteína de ligação a	A proteína PAB2 tem um papel mais significativo no crescimento vegetativo das plantas de *Arabidopsis*	Gallie, 2017

	poliadenilato 2	*thaliana*. Quando a expressão de PAB2 e PAB4 é perdida, há uma redução notável no crescimento da planta, além de uma transição antecipada do estágio vegetativo para o floral, sugerindo que ambas as proteínas são importantes para a regulação desses processos. No entanto, a ausência de efeitos significativos no crescimento ou na aparência das inflorescências ao perder apenas uma das proteínas (PAB2 ou PAB4) indica que elas têm funções que se sobrepõem, ou seja, uma pode compensar a falta da outra. Além disso, foi notada uma redução semelhante na fertilidade de mutantes que combinam perdas nas expressões de PAB2 com outras proteínas, como pab8 e pab4. Por fim, os dados sugerem que PAB4 é crucial para manter níveis estáveis de certos mRNAs, o que, por sua vez, impacta a expressão desses mRNAs na forma de proteínas, indicando que PAB4 desempenha um papel importante na regulação da expressão gênica. O papel das proteínas PAB2 e PAB4 está diretamente relacionado ao crescimento vegetativo e à transição para a floração, processos que são essenciais para o desenvolvimento das plantas e podem impactar o metabolismo primário.	
WRKY33	fator de transcrição WRKY 33	TgWRKY33, pertencentes à família de fatores de transcrição WRKY de *T. grandiflorum*, desempenha um papel potencial em diversos processos biológicos. Isso inclui a resposta a estresses bióticos, como infecções por patógenos, e estresses abióticos, como variações ambientais (ex. seca, salinidade). Além disso, essas proteínas podem estar envolvidas na regulação da transcrição, ou seja, no controle da expressão de outros genes, o que pode influenciar diretamente a adaptação e defesa da planta frente a diferentes tipos de estresse.	Ferraz dos Santos *et al.*, 2022
PRF1	Profilina-1	o papel das profilinas AtPFN1 e AtPFN2 de *A. thaliana* como proteínas antifúngicas, detalhando como essas proteínas podem inibir o crescimento de cepas fúngicas patogênicas. Quando expressas em Escherichia coli, essas proteínas recombinantes foram capazes de inibir o crescimento celular de diversas linhagens fúngicas em concentrações que variam de 10 a 160 µg/mL, indicando sua eficácia na inibição do desenvolvimento fúngico. Além disso, essas proteínas acumulam-se de forma significativa no interior das células e têm uma afinidade de ligação com as células fúngicas. Isso sugere que as profilinas podem se ligar a estruturas fúngicas e exercer sua ação inibitória. Um aspecto interessante é a capacidade das proteínas AtPFN de penetrar na parede celular e na membrana das células fúngicas. Essa penetração permite que elas atuem como inibidores do crescimento fúngico, gerando espécies reativas de oxigênio (ROS) e superóxido mitocondrial, que são conhecidos por serem tóxicos para os patógenos.	Park *et al.*, 2018
RBP45B	Proteína de ligação a poliadenilato	Semelhante à proteína RBP45 do tabaco, foi demonstrado que a AtRBP45b de *Arabidopsis* está associada ao RNA poli-A. Análise do banco de dados de interação de proteínas indicou que a AtRBP45b interage	Peal; Jambunathan; Mahalingam,

	RBP45B	com a proteína de ligação ao cap CBP20 (Tabela Suplementar S3). Especula-se que a AtRBP45b provavelmente se liga a sequências ricas em uracila nas 3'-UTRs de genes-alvo e, em coordenação com a CBP20, pode estar envolvida na regulação da expressão gênica durante o desenvolvimento normal e em resposta a perturbações ambientais.	2011
ALA1	ATPase 1 transportadora de fosfolipídios	ATPases são responsáveis pela desintoxicação celular de micotoxinas, que são substâncias tóxicas produzidas por fungos. O mecanismo de ação envolve o transporte de vesículas mediado por AtALA1 e AtALA7, no qual as toxinas são sequestradas em vacúolos para degradação. Além disso, a superexpressão de AtALA1 e AtALA7 aumenta significativamente a resistência de plantas transgênicas a F. graminearum e V. dahliae, sugerindo que essas proteínas desempenham um papel crucial na defesa das plantas contra esses patógenos em *A. thaliana*. Mudanças na fluidez local poderiam afetar a atividade de proteínas de membrana importantes, o que, por sua vez, poderia influenciar a capacidade da planta de se adaptar a temperaturas frias.	Wang *et al.*, 2021
UBA1	Proteínas associadas a UBP1	o gene TgUBA1 em *T. grandiflorum* apresentou níveis de expressão mais altos especificamente nas sementes do genótipo 174. Isso sugere que TgUBA1 pode ter um papel relevante nas funções ou processos biológicos que ocorrem nas sementes desse genótipo, possivelmente relacionado ao desenvolvimento, maturação ou resposta a estímulos específicos.	Ferraz dos Santos *et al.*, 2022

2.6. A família de genes COBRA-LIKE (COBLs) e seu provável papel na regulação da parede celular

A parede celular da planta tem muitas funções: regula a expansão celular, contribui para a adesão celular, atua como uma barreira para potenciais pragas e patógenos e determina as propriedades físicas da planta (Braam, 1999; Jones; Takemoto, 2004; Farrokhi *et al*., 2006). As diferentes funções da parede celular são refletidas na grande variação na composição da parede celular entre os diferentes tipos de células e durante a diferenciação celular. Uma estimativa sugere que até 15% dos genes codificados por um genoma de planta podem estar relacionados com a síntese, remodelação ou renovação da parede celular (Carpita; Tierney; Campbell, 2001).

A presença da família COBL é notável em diversas espécies do reino vegetal, sendo registrada no musgo *Physcomitrella patens* (Roudier *et al*., 2002; Ching *et al*., 2006; Brady *et al*., 2007; Liu *et al*., 2013; Niu *et al*., 2015; Li *et al*., 2019; Sangi *et al*., 2021) e em plantas com flores como o milho (*Zea mays*), tomate (*Solanum lycopersicum*) e algodão (*Gossypium* spp.) (Smith *et al*., 2020; Garcia *et al*., 2019), além de *Euphorbia esula* (Eu), *Glycine max* (Gm, soja), *Gossypium hirsutum* (Gh, algodão), *Hordeum vulgare* (Hv, cevada), *Lycopersicon esculentum* (Le, tomate), *Lotus japonicus* (Lj, lótus japonês), *Medicago sativa* (Ms, alfafa), *Medicago truncatula* (Mt, medicago-do-truncado), *Oryza sativa* (Os, arroz), *Populus tremula* (Pt, álamo-tremedor), *Sorghum bicolor* (Sb, sorgo), *Solanum tuberosum* (St, batata), *Zea mays* (Zm, milho) (Figura 12). Essa observação ressalta a ampla distribuição e conservação dessa família ao longo da evolução, sugerindo implicações significativas em processos biológicos fundamentais. Em *Arabidopsis thaliana*, planta modelo e amplamente estudada, foram identificados 12 membros desta família, denominados AtCOBLs (Roudier *et al*., 2002).

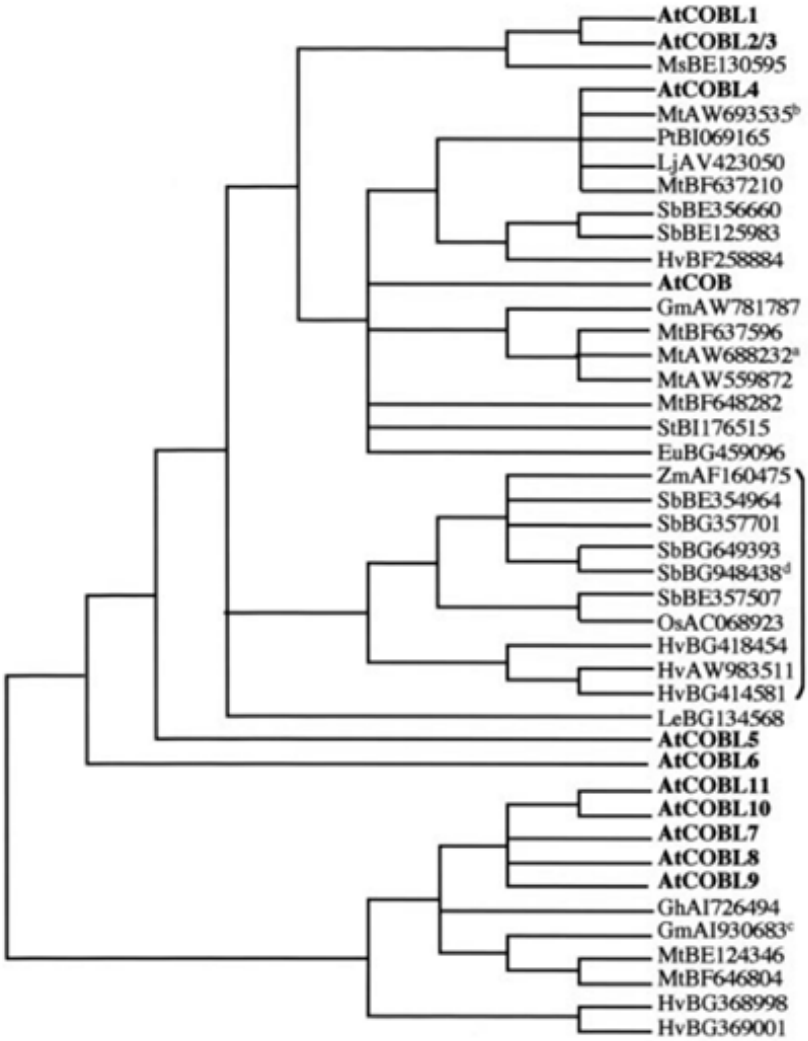

Figura 12. Identificação de ortólogos dos genes COB e COBL em outras espécies de plantas (**At** – *Arabidopsis thaliana;* **Ms** – *Medicago sativa;* **Mt** – *Medicago truncatula;* ***Pt*** *– Populus trichocarpa;* **Lj** – *Lotus japonicus;* **Sb** – *Sorghum bicolor;* Hv – *Hordeum vulgare;* **Gm** – *Glycine max;* **St** – *Solanum tuberosum;* **Eu** – *Eucalyptus sp;* **Zm** – *Zea mays;* **Os** – *Oryza sativa (arroz);* **Le** – *Lycopersicon esculentum;* **Gh** – *Gossypium hirsutum;* **Hb** – *Hevea brasiliensis.* Um clado específico das gramíneas é destacado com um colchete. A raiz da árvore foi posicionada na separação entre os genes dos grupos COB/COBL1-6 e COBL7-11, sugerindo uma divergência evolutiva entre esses dois subgrupos. Fonte: Roudier *et al*., 2002

A localização e distribuição dos genes em um genoma são aspectos fundamentais para compreender sua função e evolução. No caso da família COBL em *Arabidopsis thaliana*, a organização cromossômica desses genes fornece indícios sobre seus padrões evolutivos e possíveis relações funcionais (Figura 13). Além disso, os eventos de duplicação gênica, indicados por setas tracejadas, sugere que esses

genes podem ter surgido a partir de duplicações durante a evolução da planta (Roudier *et al.,* 2002).

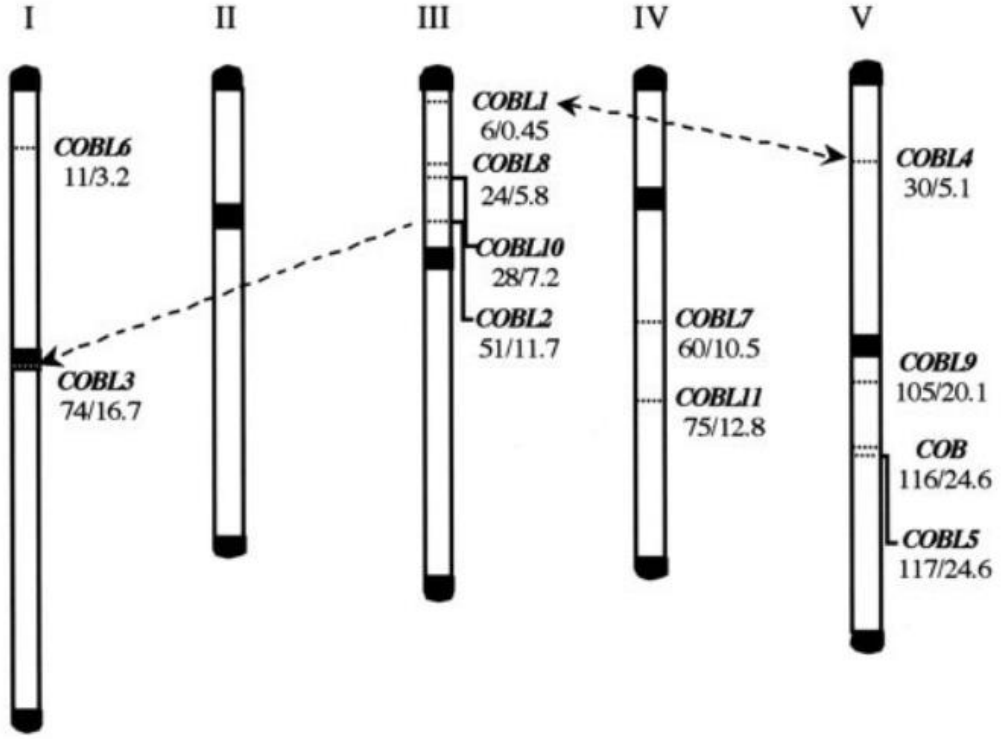

Figura 13. Localização dos COBL nos cromossomos de *Arabidopsis thaliana*. Os cinco cromossomos são representados em algarismos romanos (I a V). As regiões teloméricas e centroméricas são destacadas em preto. setas tracejadas destaca eventos de duplicação gênica. Para cada gene, há duas informações ao lado: a posição genética, expressa em centiMorgans (cm, número à esquerda), e a posição física, expressa em megabases (Mb, número à direita). Isso indica a localização exata de cada homólogo nos cromossomos. Fonte: Roudier *et al*., 2002

A família COBL tem sido objeto de estudo em diversas plantas, fornecendo análises valiosas sobre a função e a regulação desses genes em diferentes contextos biológicos. Por exemplo, em *A. thaliana*, os genes COBL são essenciais para direcionar a expansão celular (AtCOBRA) e controlar a cristalinidade da celulose e a reconstrução da parede celular secundária (AtCOBL4) (Schindelman *et al*., 2001; Brady *et al*., 2007). No milho, mutações no gene Bk2, ortólogo funcional de COBL4 em *A. thaliana*, afetam a resistência do caule ao interferir na deposição de celulose na parede celular secundária em células fibrosas (Ching *et al*., 2006).

Em tomate (*Solanum lycopersicum*), o gene SlCOBRA homólogo a AtCOBRA desempenha um papel fundamental na biologia de frutas carnosas (Cao *et al.*, 2012).

Este gene é altamente expresso em órgãos vegetativos e no desenvolvimento inicial da fruta, mas sua expressão na fruta diminui drasticamente durante os estágios de amadurecimento, implicando em um papel primário no desenvolvimento inicial da fruta (Cao *et al*., 2012). No algodão (*Gossypium* spp.), a análise abrangente da família de genes COBL identifica dois COBLs (GhCOBL9 e GhCOBL13) potencialmente associados ao espessamento da parede celular secundária da fibra, sendo muito mais expressos em fibras durante os estágios da biossíntese da parede celular secundária (Niu *et al*., 2015). Assim como no sorgo, o gene BC1, ortólogo do AtCOBL4, desempenha um papel crucial no controle da resistência mecânica da planta, influenciando diretamente o teor de celulose na parede celular secundária (Li *et al*., 2019). Já em madeiras de *Populus*, os genes PtrCOBL modulam a formação do lenho de tração e estão envolvidos em respostas a estresses abióticos (Sajjad *et al*., 2023).

Esses achados sugerem a relevância funcional dos genes COBL na modulação da resistência e integridade estrutural das espécies domesticadas estudadas. Por exemplo, o comprimento da fibra do lenho do cupuaçu e do cacau é classificado como de tamanho médio (900–1600 µm), de acordo com a IAWA (1989). Nesse contexto, apesar da semelhança na classificação do tamanho das fibras de cupuaçu e cacau, um estudo realizado por Reis e da Cruz (2021) demonstra, através de análises biométricas de fibras secundárias do xilema de cupuaçu e cacau, uma diferença significativa na espessura da parede celular entre as espécies, com as fibras do cupuaçu sendo mais espessas. Além disso, os autores destacam que a fração de parede celular e o índice de Runkel (razão entre a espessura da parede celular e do diâmetro do lúmen) são maiores no cupuaçu. Esses resultados sugerem que a família de genes COBLs pode influenciar essas características distintas.

No entanto, é importante observar que até o momento não foram conduzidos estudos específicos sobre os genes COBL em *Theobroma grandiflorum*, *Theobroma cacao* e *Theobroma umbraticum*, plantas que apresentam uma diferença notável em seus frutos. A ausência de investigações nesse sentido destaca uma lacuna no conhecimento científico, deixando em aberto a compreensão das características específicas dessa família de genes nessas espécies. Esta investigação pode contribuir significativamente para o conhecimento da biologia molecular e do desenvolvimento dessas plantas, proporcionando uma visão abrangente das

interações genéticas e dos processos fisiológicos envolvidos nas qualidades dos frutos dessas espécies amazônicas. Além disso, a caracterização dos genes COBL em *Theobroma* e *Herrania* pode revelar novos possíveis alvos para o melhoramento genético, visando a melhoria da resistência a patógenos e a otimização das propriedades estruturais das fibras, o que seria benéfico tanto para a produção agrícola quanto para a indústria de derivados.

2.7. Justificativa e hipótese

O cupuaçu é uma espécie frutífera de grande importância econômica para a região Amazônica, destacando-se como uma alternativa promissora ao cacau na fabricação de doces, chocolates e uma variedade de produtos cosméticos e farmacêuticos. Além de seu valor comercial, o cupuaçu possui propriedades nutricionais e sensoriais únicas que agregam valor aos produtos derivados, como sucos, geleias e manteiga de cupuaçu. Entretanto, a produtividade e a qualidade dos frutos de cupuaçu ainda enfrentam desafios significativos, incluindo a resistência a patógenos e a necessidade de melhorar as características organolépticas dos frutos.

Apesar de sua relevância, os estudos nas áreas de genética molecular e genômica do cupuaçu ainda são incipientes, especialmente quando comparados ao cacau, que possui maior reconhecimento e investimentos devido ao seu impacto econômico no mercado internacional. A caracterização de genes relacionados à qualidade dos frutos, como aqueles envolvidos na biossíntese de lipídeos e flavonoides, bem como a família de genes COBRA-LIKE (COBL), é fundamental para o desenvolvimento de programas de melhoramento genético mais eficientes e direcionados. A utilização de técnicas avançadas de sequenciamento genômico em escala cromossômica, aliada a ferramentas bioinformáticas, pode permitir a identificação e a análise de genes de interesse que conferem resistência a doenças e melhoram as características sensoriais e nutricionais dos frutos.

Além disso, a compreensão da história de paleopoliploidia do cupuaçu e a análise topográfica dos genes duplicados são essenciais para elucidar os eventos

evolutivos que moldaram seu genoma. Esse conhecimento não apenas enriquece a biologia molecular da espécie, mas também poderá fornecer *insights* valiosos para estratégias de melhoramento baseadas em dados genômicos. Por exemplo, a análise comparativa dos genomas de *Theobroma grandiflorum, Theobroma cacao* e *Theobroma umbraticum* poderá permitir a identificação de padrões de conservação genômica, facilitando a transferência de conhecimento entre espécies e promovendo o desenvolvimento de cultivares mais resilientes e de alta qualidade.

Adicionalmente, a caracterização dos genes da família COBL poderá contribuir para a compreensão das variações na espessura da parede celular das fibras, influenciando diretamente a resistência mecânica das plantas. Isso é particularmente relevante para estas espécies, onde estudos indicam que a espessura da parede celular das fibras pode estar associada à presença e à expressão de genes COBL, impactando tanto a resistência a patógenos quanto as propriedades físicas dos frutos.

Por fim, esta tese postula a hipótese de que a caracterização e análise comparativa dos genes relacionados à biossíntese de lipídeos e flavonoides, bem como da família de genes COBRA-LIKE (COBL), em *Theobroma grandiflorum* e suas espécies relacionadas, permitirá a identificação de prováveis alvos genéticos responsáveis pela resistência a patógenos e pela melhoria da qualidade dos frutos. Supõe-se que a manipulação desses genes, através de técnicas de biotecnologia e melhoramento genético, poderá resultar no desenvolvimento de cultivares de cupuaçu com maior resistência à doença e aprimoramento das propriedades sensoriais e nutricionais dos frutos, garantindo maior eficiência na produção e qualidade dos derivados de cupuaçu.

3 REFERÊNCIAS

ALVAREZ, Javier Correa; MARTÍNEZ, Sergio Castro; COY, Jairo. Estado de la moniliasis del *cacao* causada por *Moniliophthora roreri* en Colombia. **Acta Agronómica**, [*s. l.*], v. 63, n. 4, p. 388–399, 2014.

ALVES, Rafael Moysés *et al.* Genomic decoding of *Theobroma grandiflorum* (*cupuaçu*) at chromosomal scale: evolutionary insights for horticultural innovation. **GigaScience**, [*s. l.*], v. 13, p. giae027, 2024.

ALVES, Rafael Moysés *et al.* Phenotypic and genotypic characterization and compatibility among genotypes to select elite clones of *cupuaçu*. **Acta Amazonica**, [*s. l.*], v. 47, n. 3, p. 175–184, 2017.

ALVES, Rafael Moysés; CHAVES, Saulo Fabrício da Silva. Selection of *Theobroma grandiflorum* clones adapted to agroforestry systems using an additive index. **Acta Scientiarum. Agronomy**, [*s. l.*], v. 45, p. e57519, 2023.

ANIL KUMAR, S. *et al*. Osmotin: a plant sentinel and a possible agonist of mammalian adiponectin. Frontiers in Plant Science, [s. l.], v. 6, 2015. Disponível em: https://www.frontiersin.org/journals/plant-cience/articles/10.3389/fpls.2015.00163/full. Acesso em: 14 jan. 2025.

ARGOUT, X. *et al.* The *cacao* Criollo genome v2.0: an improved version of the genome for genetic and functional genomic studies. **BMC Genomics**, [*s. l.*], v. 18, n. 1, p. 730, 2017.

ARGOUT, Xavier *et al.* The genome of *Theobroma cacao*. **Nature Genetics**, [*s. l.*], v. 43, n. 2, p. 101–108, 2011.

BAILEY, Bryan A. *et al.* Differential gene expression by oniliophthora roreri while overcoming *cacao* tolerance in the field. **Molecular Plant Pathology**, [*s. l.*], v. 15, n. 7, p. 711–729, 2014.

BAILEY, Bryan A. *et al. Moniliophthora roreri*, causal agent of *cacao* frosty pod rot. **Molecular Plant Pathology**, [*s. l.*], v. 19, n. 7, p. 1580–1594, 2018.

BARTLEY, B. G. D. **The Genetic Diversity of *Cacao* and Its Utilization**. [*S. l.*]: CABI, 2005.

BAYER, C. Sarcolaenaceae. *In*: KUBITZKI, Klaus; BAYER, Clemens (org.). **Flowering Plants · Dicotyledons**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003. p. 345–352. Disponível em: http://link.springer.com/10.1007/978-3-662-07255-4_38. Acesso em: 30 out. 2024.

BOSSA-CASTRO, Ana M. *et al.* A phylogenetic framework to study desirable traits in the wild relatives of *Theobroma cacao* (Malvaceae). **Journal of Systematics and Evolution**, [*s. l.*], v. 62, n. 5, p. 963–978, 2024.

BARBOSA, Ceslaine Santos *et al.* Genome sequence and effectorome of *Moniliophthora perniciosa* and *Moniliophthora roreri* subpopulations. **BMC genomics**, [*s. l.*], v. 19, n. 1, p. 509, 2018.

BOHLMANN, Jörg; KEELING, Christopher I. Terpenoid biomaterials. **The Plant Journal**, [*s. l.*], v. 54, n. 4, p. 656–669, 2008.

BRAAM, Janet. If walls could talk: Commentary. **Current Opinion in Plant Biology**, [*s. l.*], v. 2, n. 6, p. 521–524, 1999.

BRACHA-DRORI, Keren *et al.* Functional Analysis of *Arabidopsis* Postprenylation CaaX Processing Enzymes and Their Function in Subcellular Protein Targeting. **Plant Physiology**, [*s. l.*], v. 148, n. 1, p. 119–131, 2008.

BRADY, Siobhan M. *et al.* Combining Expression and Comparative Evolutionary Analysis. The COBRA Gene Family. **Plant Physiology**, [*s. l.*], v. 143, n. 1, p. 172–187, 2007.

BRANEN, Jill K.; CHIOU, Tzyy-Jen; ENGESETH, Nicki J. Overexpression of Acyl Carrier Protein-1 Alters Fatty Acid Composition of Leaf Tissue in *Arabidopsis* . **Plant Physiology**, [*s. l.*], v. 127, n. 1, p. 222–229, 2001.

CALZAVARA, B. B. G. Cupuaçuzeiro. [*s. l.*], 1987. Disponível em: http://www.infoteca.cnptia.embrapa.br/handle/doc/381210. Acesso em: 16 nov. 2024.

CAO, Ying *et al.* Functional characterization of a tomato COBRA-likegene functioning in fruit development and ripening. **BMC Plant Biology**, [*s. l.*], v. 12, n. 1, p. 211, 2012.

CARPITA, Nick; TIERNEY, Mary; CAMPBELL, Malcolm. Molecular biology of the plant cell wall: searching for the genes that define structure, architecture and dynamics. **Plant Molecular Biology**, [*s. l.*], v. 47, n. 1, p. 1–5, 2001.

CHALK, L. **Anatomy of the Dicotyledons: Leaves, Stem, and Wood in Relation to Taxonomy with Notes on Economic Uses**. [s.l.] Clarendon Press, 1950. v. II

CHEESMAN E (1944) Notas sobre a nomenclatura, classificação e possíveis relações de populações de cacau. Tropical Agriculture 21: 144–159.

CHEN, Feng *et al.* Biosynthesis and Emission of Terpenoid Volatiles from *Arabidopsis* Flowers. **The Plant Cell**, [*s. l.*], v. 15, n. 2, p. 481–494, 2003a.

CHEN, Feng *et al.* Biosynthesis and Emission of Terpenoid Volatiles from *Arabidopsis* Flowers. **The Plant Cell**, [*s. l.*], v. 15, n. 2, p. 481–494, 2003b.

CHEN, Feng *et al.* Characterization of a Root-Specific *Arabidopsis* Terpene Synthase Responsible for the Formation of the Volatile Monoterpene 1,8-Cineole. **Plant Physiology**, [*s. l.*], v. 135, n. 4, p. 1956–1966, 2004.

CHEN, Li-Yu *et al.* The bracteatus pineapple genome and domestication of clonally propagated crops. **Nature Genetics**, [*s. l.*], v. 51, n. 10, p. 1549–1558, 2019.

CHING, Ada *et al.* Brittle stalk 2 encodes a putative glycosylphosphatidylinositol-anchored protein that affects mechanical strength of maize tissues by altering the composition and structure of secondary cell walls. **Planta**, [*s. l.*], v. 224, n. 5, p. 1174–1184, 2006.

CHRISTENHUSZ, Maarten J.M.; BYNG, James W. The number of known plants species in the world and its annual increase. **Phytotaxa**, [*s. l.*], v. 261, n. 3, p. 201, 2016.

CLEMENT, Charles R. 1492 and the loss of amazonian crop genetic resources. I. The relation between domestication and human population decline. **Economic Botany**, [*s. l.*], v. 53, n. 2, p. 188–202, 1999.

COHEN, Kelly de Oliveira; JACKIX, Marisa de Nazaré Hoelz. Estudo do liquor de cupuaçu. **Food Science and Technology**, [*s. l.*], v. 25, p. 182–190, 2005.

COLLI-SILVA, Matheus *et al.* Domestication of the Amazonian fruit tree cupuaçu may have stretched over the past 8000 years. **Communications Earth & Environment**, [*s. l.*], v. 4, n. 1, p. 1–8, 2023.

COLLI-SILVA, Matheus *et al.* Phylogenetic evidence reshapes the taxonomy of *Cacao* and its allies (*Theobroma* and *Herrania*; Malvaceae, Byttnerioideae). **Brittonia**, [*s. l.*], v. 76, n. 1, p. 53–61, 2024.

COLLI-SILVA, Matheus; RICHARDSON, James; PIRANI, José. A taxonomic dataset of preserved specimen occurrences of *Theobroma* and *Herrania* (Malvaceae, Byttnerioideae) stored in 2020. **Biodiversity Data Journal**, [*s. l.*], v. 11, p. e99646, 2023.

CORNEJO, Omar E. *et al.* Population genomic analyses of the chocolate tree, *Theobroma cacao* L., provide insights into its domestication process. **Communications Biology**, [*s. l.*], v. 1, n. 1, p. 1–12, 2018.

COSTA, Marion P. *et al. Cupuaçu* (*Theobroma grandiflorum*) pulp, probiotic, and prebiotic: Influence on color, apparent viscosity, and texture of goat milk yogurts. **Journal of Dairy Science**, [*s. l.*], v. 98, n. 9, p. 5995–6003, 2015a.

COSTA, Marion P. *et al.* Polpa de cupuaçu (*Theobroma grandiflorum*), probiótico e prebiótico: Influência na cor, viscosidade aparente e textura de iogurtes de leite de cabra. **Journal of Dairy Science**, [*s. l.*], v. 98, n. 9, p. 5995–6003, 2015b.

CRAENE, Louis Ronse De. Review Reviewed Work(s): Flowering plants. Dicotyledons. Capparales, Malvales and Nonbetalain Caryophyllales. (The Families and Genera of Vascular Plants, vol. 5) by K. Kubitzki and C. Bayer. **Plant Systematics and Evolution**, [*s. l.*], v. 242, n. 1/4, p. 217–219, 2003.

CUATRECASAS, José. ***Cacao* and Its Allies: A Taxonomic Revision of the Genus *Theobroma***. [*S. l.*]: Smithsonian Institution, 1964.

DA SILVA ESTRELA JUNIOR, Ailton *et al.* Viability of *Moniliophthora roreri* on Cocoa Beans Under Microfermentation and Long-Term Survival on Carrier Materials. **Plant Disease**, [*s. l.*], v. 107, n. 11, p. 3497–3505, 2023a.

DA SILVA ESTRELA JUNIOR, Ailton *et al.* Viability of *Moniliophthora roreri* on Cocoa Beans Under Microfermentation and Long-Term Survival on Carrier Materials. **Plant Disease**, [*s. l.*], v. 107, n. 11, p. 3497–3505, 2023b.

DE SOUZA, Maria Geralda; DE ALMEIDA, Olívia Cordeiro. (*Theobroma grandiflorum* Willd. Spend.) Schum. [*s. l.*],

D'HONT, Angélique *et al.* The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. **Nature**, [*s. l.*], v. 488, n. 7410, p. 213–217, 2012.

DOS SANTOS, Everton Cruz *et al.* The pathogen *Moniliophthora perniciosa* promotes differential proteomic modulation of *cacao* genotypes with contrasting resistance to witches´ broom disease. **BMC Plant Biology**, [*s. l.*], v. 20, n. 1, p. 1, 2020.

DOYLE, James A; HOTTON, Carol L. Diversification of early angiosperm pollen in a cladistic context. *In*: BLACKMORE, S; BARNES, S H (org.). **Pollen and Spores**. [*S. l.*]: Oxford University PressOxford, 1991. p. 169–196. Disponível em: https://academic.oup.com/book/53649/chapter/422151855. Acesso em: 15 jan. 2025.

ECKARDT, Nancy A. Sequencing the Rice Genome. **The Plant Cell**, [*s. l.*], v. 12, n. 11, p. 2011–2017, 2000.

ETAWARE, Peter Mudiaga. The effects of the phytochemistry of cocoa on the food chemistry of chocolate(s) and how disease resistance in cocoa can be improved using CRISPR/Cas9 technology. **Food Chemistry: Molecular Sciences**, [*s. l.*], v. 3, p. 100043, 2021.

EDGER, Patrick P. *et al.* Origin and evolution of the octoploid strawberry genome. **Nature Genetics**, [*s. l.*], v. 51, n. 3, p. 541–547, 2019.

EL-KHIAT, Sarah *et al.* Taxonomic characters of fruit, mericarp and seed of some Malvaceae species. **International Journal of Theoretical and Applied Research**, [*s. l.*], v. 0, n. 0, p. 97–109, 2023.

EVANS, H. C.; HOLMES, K. A.; REID, A. P. Phylogeny of the frosty pod rot pathogen of cocoa. **Plant Pathology**, [*s. l.*], v. 52, n. 4, p. 476–485, 2003.

FÄLDT, Jenny *et al.* Functional identification of AtTPS03 as (E)-β-ocimene synthase: a monoterpene synthase catalyzing jasmonate- and wound-induced volatile formation in *Arabidopsis thaliana*. **Planta**, [*s. l.*], v. 216, n. 5, p. 745–751, 2003.

FALCÃO, Loeni Ludke *et al.* Comparative transcriptomics of *cupuaçu* (*Theobroma grandiflorum*) offers insights into the early defense mechanism to Moniliophthora perniciosa, the causal agent of witches' broom disease. **Journal of Plant Interactions**, [*s. l.*], v. 17, n. 1, p. 991–1005, 2022.

FARROKHI, Naser *et al.* Plant cell wall biosynthesis: genetic, biochemical and functional genomics approaches to the identification of key genes. **Plant Biotechnology Journal**, [*s. l.*], v. 4, n. 2, p. 145–167, 2006.

FERRAZ DOS SANTOS, Lucas *et al. Cupuaçu* (*Theobroma grandiflorum* [Willd. ex Sprengel] Schumann) Fruit Development: Key Genes Involved in Primary Metabolism and Stress Response. **Agronomy**, [*s. l.*], v. 12, n. 4, p. 763, 2022.

FIGUEIRA, Aline Cássia; LUCCAS, Valdecir. Physicochemical characterization of national and commercial cocoa butter used in Brazil to make chocolate. **Brazilian Journal of Food Technology**, [*s. l.*], v. 25, p. e2022033, 2022.

FLORA E FUNGA DO BRASIL. [*S. l.*], [*s. d.*]. Disponível em: https://floradobrasil.jbrj.gov.br/FB23617. Acesso em: 17 nov. 2024.

GALLIE, Daniel R. Class II members of the poly(A) binding protein family exhibit distinct functions during *Arabidopsis* growth and development. **Translation**, [*s. l.*], v. 5, n. 1, p. e1295129, 2017.

GARSMEUR, Olivier *et al.* A mosaic monoploid reference sequence for the highly complex genome of sugarcane. **Nature Communications**, [*s. l.*], v. 9, n. 1, p. 2638, 2018.

GERSHENZON, Jonathan; DUDAREVA, Natalia. The function of terpene natural products in the natural world. **Nature Chemical Biology**, [*s. l.*], v. 3, n. 7, p. 408–414, 2007.

GOMES, Dayane Santos *et al.* O desenvolvimento micelial que precede a formação do basidioma em *Moniliophthora perniciosa* está associado a alterações no metabolismo de quitina, açúcar e nutrientes envolvendo autofagia. **Fungal Genetics and Biology**, [*s. l.*], v. 86, p. 33–46, 2016.

GONDIM, T. M. de S. *et al.* Aspectos da produção de cupuaçu. [*s. l.*], 2001. Disponível em: http://www.infoteca.cnptia.embrapa.br/handle/doc/498481. Acesso em: 16 nov. 2024.

GRIFFITHS, Gareth; HARWOOD, John L. The regulation of triacylglycerol biosynthesis in cocoa (*Theobroma cacao*) L. **Planta**, [*s. l.*], v. 184, n. 2, p. 279–284, 1991.

GRIFFITHS, Gareth; HARWOOD, John L. The regulation of triacylglycerol biosynthesis in cocoa (*Theobroma cacao*) L. **Planta**, [*s. l.*], v. 184, n. 2, p. 279–284, 1991.

GRIFFITHS, G.; HARWOOD, J. L. The Regulation of Triacylglycerol Biosynthesis in Cocoa (*Theobroma Cacao*) L. *Planta*, v. 184, n. 2, p. 279–284, 1 maio 1991.

GUO, Shaogui *et al.* The draft genome of watermelon (Citrullus lanatus) and resequencing of 20 diverse accessions. **Nature Genetics**, [*s. l.*], v. 45, n. 1, p. 51–58, 2013.

HABERER, Georg *et al.* Structure and Architecture of the Maize Genome. **Plant Physiology**, [*s. l.*], v. 139, n. 4, p. 1612–1624, 2005.

HAKIM *et al.* Osmotin: A plant defense tool against biotic and abiotic stresses. **Plant Physiology and Biochemistry**, [*s. l.*], v. 123, p. 149–159, 2018.

HEBBAR, Prakash K. *Cacao* Diseases: A Global Perspective from an Industry Point of View. **Phytopathology®**, [*s. l.*], v. 97, n. 12, p. 1658–1663, 2007.

HUANG, Gai *et al.* Genome sequence of Gossypium herbaceum and genome updates of Gossypium arboreum and Gossypium hirsutum provide insights into cotton A-genome evolution. **Nature Genetics**, [*s. l.*], v. 52, n. 5, p. 516–524, 2020.

HUANG, Mengsu *et al.* Variation of Herbivore-Induced Volatile Terpenes among *Arabidopsis* Ecotypes Depends on Allelic Differences and Subcellular Targeting of Two Terpene Synthases, TPS02 and TPS03. **Plant Physiology**, [*s. l.*], v. 153, n. 3, p. 1293–1310, 2010.

INSTITUTO DE DESENVOLVIMENTO AGROPECUÁRIO E FLORESTAL SUSTENTÁVEL DO ESTADO DO AMAZONAS – IDAM. *Boas práticas de produção sustentável no Amazonas: cupuaçu*. Manaus: IDAM, 2021. Disponível em:

http://www.idam.am.gov.br/wp-content/uploads/2021/01/Cupua%C3%A7u.pdf. Acesso em: 23 abr. 2025.

JONES, David A; TAKEMOTO, Daigo. Plant innate immunity – direct and indirect recognition of general and specific pathogen-associated molecules. **Current Opinion in Immunology**, [*s. l.*], v. 16, n. 1, p. 48–62, 2004.

JIAO, Yuannian *et al.* Ancestral polyploidy in seed plants and angiosperms. **Nature**, [*s. l.*], v. 473, n. 7345, p. 97–100, 2011.

JUNIOR, Braz Tavares da Hora *et al.* Transcriptomics and systems biology analysis in identification of specific pathways involved in *cacao* resistance and susceptibility to witches' broom disease. **Molecular BioSystems**, [*s. l.*], v. 8, n. 5, p. 1507–1519, 2012.

KILARU, Aruna; BAILEY, Bryan A.; HASENSTEIN, Karl H. *Moniliophthora perniciosa* produces hormones and alters endogenous auxin and salicylic acid in infected cocoa leaves. **FEMS Microbiology Letters**, [*s. l.*], v. 274, n. 2, p. 238–244, 2007.

KOBAYASHI, S. *et al.* Comparison of UDP-glucose:flavonoid 3-*O*-glucosyltransferase (UFGT) gene sequences between white grapes (*Vitis vinifera*) and their sports with red skin. **Plant Science**, [*s. l.*], v. 160, n. 3, p. 543–550, 2001.

LANGE, B. Markus; GHASSEMIAN, Majid. Genome organization in *Arabidopsis thaliana*: a survey for genes involved in isoprenoid and chlorophyll metabolism. **Plant Molecular Biology**, [*s. l.*], v. 51, n. 6, p. 925–948, 2003.

LANAUD, Claire *et al.* A revisited history of *cacao* domestication in pre-Columbian times revealed by archaeogenomic approaches. **Scientific Reports**, [*s. l.*], v. 14, n. 1, p. 2972, 2024.

LEVIS, C. *et al.* Persistent effects of pre-Columbian plant domestication on Amazonian forest composition. **Science**, [*s. l.*], v. 355, n. 6328, p. 925–931, 2017.

LEITE, R. S. V. *et al.* Monilíase do cacaueiro e cupuaçuzeiro. [*s. l.*], 2023. Disponível em: http://www.infoteca.cnptia.embrapa.br/handle/doc/1157539. Acesso em: 22 dez. 2024.

LI, Pan *et al.* Brittle Culm 1 Encodes a COBRA-Like Protein Involved in Secondary Cell Wall Cellulose Biosynthesis in Sorghum. **Plant and Cell Physiology**, [*s. l.*], v. 60, n. 4, p. 788–801, 2019.

LI, Zhanwen *et al.* FATCAT 2.0: towards a better understanding of the structural diversity of proteins. **Nucleic Acids Research**, [*s. l.*], v. 48, n. W1, p. W60–W64, 2020.

LI, Yi *et al.* Four Isoforms of *Arabidopsis* 4-Coumarate:CoA Ligase Have Overlapping yet Distinct Roles in Phenylpropanoid Metabolism. **Plant Physiology**, [*s. l.*], v. 169, n. 4, p. 2409–2421, 2015.

LIU, Yan *et al.* Advances in the Plant Isoprenoid Biosynthesis Pathway and Its Metabolic Engineering. **Journal of Integrative Plant Biology**, [*s. l.*], v. 47, n. 7, p. 769–782, 2005.

LOON, L. C. van; REP, M.; PIETERSE, C. M. J. Significance of Inducible Defense-related Proteins in Infected Plants. **Annual Review of Phytopathology**, [*s. l.*], v. 44, n. Volume 44, 2006, p. 135–162, 2006.

MAR, Josiana Moreira *et al. Theobroma spp.:* Uma revisão do seu potencial químico e de inovação para a indústria alimentícia. **Food Chemistry Advances**, [*s. l.*], v. 4, p. 100683, 2024.

MASULLO, Fernanda de Araújo *et al.* Fruits of neotropical species of the tribe Malveae (Malvoideae - Malvaceae): macro- and micromorphology. **Acta Botanica Brasilica**, [*s. l.*], v. 34, p. 301–311, 2020.

MATOS ANDRADE, Geisa Maria *et al.* Plantas alimentícias não convencionais (PANC) nos levantamentos etnobotânicos do Brasil. **Novos Cadernos NAEA**, [*s. l.*], v. 26, n. 2, 2023. Disponível em: https://periodicos.ufpa.br/index.php/ncn/article/view/13003. Acesso em: 30 out. 2024.

MEINHARDT, Lyndel W. *et al.* Genome and secretome analysis of the hemibiotrophic fungal pathogen, *Moniliophthora roreri*, which causes frosty pod rot disease of *cacao*: mechanisms of the biotrophic and necrotrophic phases. **BMC Genomics**, [*s. l.*], v. 15, n. 1, p. 164, 2014.

MCELROY, Michel S. *et al.* Prediction of *Cacao* (*Theobroma cacao*) Resistance to Moniliophthora spp. Diseases via Genome-Wide Association Analysis and Genomic Selection. **Frontiers in Plant Science**, [*s. l.*], v. 9, 2018. Disponível em: https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2018.00343/full. Acesso em: 23 out. 2024.

MONDEGO, Jorge MC *et al.* A genome survey of *Moniliophthora perniciosa* gives new insights into Witches' Broom Disease of *cacao*. **BMC Genomics**, [*s. l.*], v. 9, p. 548, 2008.

MCGARVEY, D J; CROTEAU, R. Terpenoid metabolism. **The Plant Cell**, [*s. l.*], v. 7, n. 7, p. 1015–1026, 1995.

MCKEAN, Angela L. *et al.* Molecular Characterization of the Non-biotin-containing Subunit of 3-Methylcrotonyl-CoA Carboxylase *. **Journal of Biological Chemistry**, [*s. l.*], v. 275, n. 8, p. 5582–5590, 2000.

MEINHARDT, Lyndel W. *et al.* Moniliophthora perniciosa, the causal agent of witches' broom disease of *cacao*: what's new from this old foe? **Molecular Plant Pathology**, [*s. l.*], v. 9, n. 5, p. 577–588, 2008.

MING, Ray *et al.* The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus). **Nature**, [*s. l.*], v. 452, n. 7190, p. 991–996, 2008.

MORENO-PÉREZ, Antonio J. *et al.* Reduced expression of FatA thioesterases in *Arabidopsis* affects the oil content and fatty acid composition of the seeds. **Planta**, [*s. l.*], v. 235, n. 3, p. 629–639, 2012.

MOSES, Tessa *et al.* Bioengineering of plant (tri)terpenoids: from metabolic engineering of plants to synthetic biology in vivo and in vitro. **New Phytologist**, [*s. l.*], v. 200, n. 1, p. 27–43, 2013.

MOTAMAYOR, J. C. *et al. Cacao* domestication I: the origin of the *cacao* cultivated by the Mayas. **Heredity**, [*s. l.*], v. 89, n. 5, p. 380–386, 2002.

MOTAMAYOR, J. C. *et al. Cacao* domestication II: progenitor germplasm of the Trinitario *cacao* cultivar. **Heredity**, [*s. l.*], v. 91, n. 3, p. 322–330, 2003.

MOTAMAYOR, Juan C. *et al.* Geographic and Genetic Population Differentiation of the Amazonian Chocolate Tree (*Theobroma cacao* L). **PLOS ONE**, [*s. l.*], v. 3, n. 10, p. e3311, 2008.

MOTAMAYOR, Juan C. *et al.* The genome sequence of the most widely cultivated *cacao* type and its use to identify candidate genes regulating pod color. **Genome Biology**, [*s. l.*], v. 14, n. 6, p. r53, 2013.

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

NAZARE, R. F. R. de; BARBOSA, W. C.; VIEGAS, R. M. F. Processamento das sementes de cupuaçu para a obtenção de cupulate. [*s. l.*], 1990. Disponível em: http://www.infoteca.cnptia.embrapa.br/handle/doc/381906. Acesso em: 16 nov. 2024.

NOUSIAS, Orestis *et al.* Three de novo assembled wild *cacao* genomes from the Upper Amazon. **Scientific Data**, [*s. l.*], v. 11, n. 1, p. 369, 2024.

NESI, Nathalie *et al.* The *Arabidopsis* TT2 Gene Encodes an R2R3 MYB Domain Protein That Acts as a Key Determinant for Proanthocyanidin Accumulation in Developing Seed. **The Plant Cell**, [*s. l.*], v. 13, n. 9, p. 2099–2114, 2001.

NIU, Erli *et al.* Comprehensive Analysis of the COBRA-Like (COBL) Gene Family in Gossypium Identifies Two COBLs Potentially Associated with Fiber Quality. **PLOS ONE**, [*s. l.*], v. 10, n. 12, p. e0145725, 2015.

NOUSIAS, Orestis *et al.* Three de novo assembled wild *cacao* genomes from the Upper Amazon. **Scientific Data**, [*s. l.*], v. 11, n. 1, p. 369, 2024.

OKULEY, J *et al. Arabidopsis* FAD2 gene encodes the enzyme that is essential for polyunsaturated lipid synthesis. **The Plant Cell**, [*s. l.*], v. 6, n. 1, p. 147–158, 1994.

OWENS, Daniel K. *et al.* Caracterização bioquímica e genética da flavanona 3β-hidroxilase *de Arabidopsis* . **Plant Physiology and Biochemistry**, [*s. l.*], v. 46, n. 10, p. 833–843, 2008a.

OWENS, Daniel K. *et al.* Functional Analysis of a Predicted Flavonol Synthase Gene Family in *Arabidopsis* . **Plant Physiology**, [*s. l.*], v. 147, n. 3, p. 1046–1061, 2008b.

PARK, Seong-Cheol *et al.* Mecanismo molecular das profilinas de *Arabidopsis thaliana como proteínas antifúngicas*. **Biochimica et Biophysica Acta (BBA) - General Subjects**, [*s. l.*], v. 1862, n. 12, p. 2545–2554, 2018.

PEAL, Lila; JAMBUNATHAN, Niranjani; MAHALINGAM, Ramamurthy. Phylogenetic and Expression Analysis of RNA-binding Proteins with Triple RNA Recognition Motifs in Plants. **Molecules and Cells**, [*s. l.*], v. 31, n. 1, p. 55–64, 2011.

PLADEVALL, Clara *et al.* **The genome sequence of the Violet Copper, *Lycaena helle* (Denis & Schiffermüller, 1775)**. [*S. l.*]: F1000Research, 2025. Disponível em: https://f1000research.com/articles/14-60. Acesso em: 15 jan. 2025.

PUGLIESE, Alexandre G. *et al.* Flavonoids, proanthocyanidins, vitamin C, and antioxidant activity of *Theobroma grandiflorum* (*Cupuaçu*) pulp and seeds. **Journal of Agricultural and Food Chemistry**, [*s. l.*], v. 61, n. 11, p. 2720–2728, 2013.

PHILLIPS-MORA, W.; AIME, M. C.; WILKINSON, M. J. Biodiversity and biogeography of the *cacao* (*Theobroma cacao*) pathogen *Moniliophthora roreri* in tropical America. **Plant Pathology**, [*s. l.*], v. 56, n. 6, p. 911–922, 2007.

PLOETZ, Randy. The Impact of Diseases on *Cacao* Production: A Global Overview. *In*: BAILEY, Bryan A.; MEINHARDT, Lyndel W. (org.). ***Cacao* Diseases: A History of Old Enemies and New Encounters**. Cham: Springer International Publishing, 2016. p. 33–59. Disponível em: https://doi.org/10.1007/978-3-319-24789-2_2. Acesso em: 7 nov. 2024.

PURCELL, Thomas F. 'Hot chocolate': financialized global value chains and cocoa production in Ecuador. **The Journal of Peasant Studies**, [*s. l.*], v. 45, n. 5–6, p. 904–926, 2018.

PURCELL, Thomas F. 'Hot chocolate': financialized global value chains and cocoa production in Ecuador. *In*: AGRARIAN MARXISM. [*S. l.*]: Routledge, 2019.

REIS, Alisson Rodrigo Souza; DA CRUZ, Cassiele Fonseca. Biometry and verification of papermaking potential of secondary xylem fibers of two *Theobroma* species. **Environment, Development and Sustainability**, [*s. l.*], v. 23, n. 7, p. 11090–11097, 2021.

RENDÓN-ANAYA, Martha *et al.* The avocado genome informs deep angiosperm phylogeny, highlights introgressive hybridization, and reveals pathogen-influenced gene space adaptation. **Proceedings of the National Academy of Sciences**, [*s. l.*], v. 116, n. 34, p. 17081–17089, 2019.

RICHARDSON, James E. *et al.* The age of chocolate: a diversification history of *Theobroma* and Malvaceae. **Frontiers in Ecology and Evolution**, [*s. l.*], v. 3, 2015. Disponível em: https://www.frontiersin.org/journals/ecology-and-evolution/articles/10.3389/fevo.2015.00120/full. Acesso em: 30 out. 2024.

ROCHA, Joecildo Francisco; MACHADO, Silvia Rodrigues. Anatomy, ultrastructure and secretion of Hibiscus pernambucensis Arruda (Malvaceae) extrafloral nectary. **Brazilian Journal of Botany**, [*s. l.*], v. 32, p. 489–498, 2009.

ROUDIER, François *et al.* The COBRA Family of Putative GPI-Anchored Proteins in *Arabidopsis* . A New Fellowship in Expansion. **Plant Physiology**, [*s. l.*], v. 130, n. 2, p. 538–548, 2002.

SAJJAD, Muhammad *et al.* Recent genome resequencing paraded COBRA-Like gene family roles in abiotic stress and wood formation in Poplar. **Frontiers in Plant Science**, [*s. l.*], v. 14, 2023. Disponível em: https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2023.1242836/full. Acesso em: 10 jul. 2024.

SALGADO, Jocelem Mastrodi *et al. Cupuaçu* (*Theobroma grandiflorum*) Peel as Potential Source of Dietary Fiber and Phytochemicals in Whole-Bread Preparations. **Plant Foods for Human Nutrition**, [*s. l.*], v. 66, n. 4, p. 384–390, 2011.

SANGI, Sara *et al.* Genome-Wide Analysis of the COBRA-Like Gene Family Supports Gene Expansion through Whole-Genome Duplication in Soybean (Glycine max). **Plants**, [*s. l.*], v. 10, n. 1, p. 167, 2021.

SANTANA SILVA, Raner José *et al.* Involvement of structurally distinct *cupuaçu* chitinases and osmotin in plant resistance to the fungus Moniliophthora perniciosa. **Plant physiology and biochemistry: PPB**, [*s. l.*], v. 148, p. 142–151, 2020.

SOTO-SERRANO, Axel *et al.* Matching excellence: Oxford Nanopore Technologies' rise to parity with Pacific Biosciences in genome reconstruction of non-model bacterium with high G+C content. **Microbial Genomics**, [*s. l.*], v. 10, n. 11, p. 001316, 2024.

SEDAP, ecretaria de Estado de Desenvolvimento Agropecuária e da Pesca. (2022). Disponível em: < https://www.sedap.pa.gov.br/boletim-cvis >.

SEDAP, ecretaria de Estado de Desenvolvimento Agropecuária e da Pesca. (2024). Disponível em: < https://www.sedap.pa.gov.br/node/347>.

SOLEREDER, H. **Systematic anatomy of the dicotyledons: a handbook for laboratories of pure and applied botany**. [s.l.] Clarendon Press, 1908. v. 2

SUFRAMA. (2017). Potencialidades – Cupuaçu. Disponível em: https://www.gov.br/suframa/pt-br/assuntos/potencialidades-cupuacu

SCHINDELMAN, Gary *et al.* COBRA encodes a putative GPI-anchored protein, which is polarly localized and necessary for oriented cell expansion in *Arabidopsis* . **Genes & Development**, [*s. l.*], v. 15, n. 9, p. 1115–1127, 2001.

SCHMUTZ, Jeremy *et al.* A reference genome for common bean and genome-wide analysis of dual domestications. **Nature Genetics**, [*s. l.*], v. 46, n. 7, p. 707–713, 2014.

SCHNURR, Judy; SHOCKEY, Jay; BROWSE, John. The Acyl-CoA Synthetase Encoded by LACS2 Is Essential for Normal Cuticle Development in *Arabidopsis* . **The Plant Cell**, [*s. l.*], v. 16, n. 3, p. 629–642, 2004.

SCHULTES, Richard Evans. Plantae Colombianae, V: Species nova generis *Herrania*. **Caldasia**, [*s. l.*], n. 8, p. 261–264, 1943.

SHIRLEY, Brenda Winkel. Flavonoids in seeds and grains: physiological function, agronomic importance and the genetics of biosynthesis. **Seed Science Research**, [*s. l.*], v. 8, n. 4, p. 415–422, 1998.

SILVA, Rangeline Azevedo da *et al.* Genome size, cytogenetic data and transferability of EST-SSRs markers in wild and cultivated species of the genus *Theobroma* L. (Byttnerioideae, Malvaceae). **PLOS ONE**, [*s. l.*], v. 12, n. 2, p. e0170799, 2017.

SILVA, Pâmela Clemente de Meneses; GALLÃO, Maria Izabel; DE ARRUDA, Emília Cristina Pereira. Comparative anatomical study of two Malvaceae species from a seasonally dry forest. **Flora**, [*s. l.*], v. 303, p. 152290, 2023.

SILVA, Filipa M.; SILVA, Cristina L. M. Colour changes in thermally processed cupuaçu (*Theobroma grandiflorum*) puree: critical times and kinetics modelling. **International Journal of Food Science & Technology**, [*s. l.*], v. 34, n. 1, p. 87–94, 1999.

SUN, J. *et al.* Biochemical and Molecular Biological Characterization of CAC2, the *Arabidopsis thaliana* Gene Coding for the Biotin Carboxylase Subunit of the Plastidic Acetyl-Coenzyme A Carboxylase. **Plant Physiology**, [*s. l.*], v. 115, n. 4, p. 1371–1383, 1997.

THOMAS, Evert *et al.* Present Spatial Diversity Patterns of *Theobroma cacao* L. in the Neotropics Reflect Genetic Differentiation in Pleistocene Refugia Followed by Human-Influenced Dispersal. **PLOS ONE**, [*s. l.*], v. 7, n. 10, p. e47676, 2012.

TODD, James; POST-BEITTENMILLER, Dusty; JAWORSKI, Jan G. encodes a fatty acid elongase 3-ketoacyl-CoA synthase affecting wax biosynthesis in. **The Plant Journal**, [*s. l.*], v. 17, n. 2, p. 119–130, 1999.

TORRES DE LA CRUZ, M. *et al.* Temporal Progress and Integrated Management of Frosty Pod Rot (*Moniliophthora roreri*) of Cocoa in Tabasco, Mexico. **Journal of plant pathology : Rivista di patologia vegetale : an international journal of the Italian society of plant pathology : 93, 1, 2011**, [*s. l.*], p. 31–36, 2011.

VAN SCHIE, Chris C. N. *et al.* Geranyl diphosphate synthase is required for biosynthesis of gibberellins. **The Plant Journal**, [*s. l.*], v. 52, n. 4, p. 752–762, 2007.

VILLAMAR-TORRES, Ronald Oswaldo *et al. Herrania* species: A potential genetic resource for cocoa breeding revealed by a comparative bioinformatic study among close species. **Bragantia**, [*s. l.*], v. 84, p. e20230298, 2024.

VELDKORNET, Dimitri Allastair. The Influence of Macroclimatic Drivers on the Macrophyte Phylogenetic Diversity in South African Estuaries. **Diversity**, [*s. l.*], v. 15, n. 9, p. 986, 2023.

VENTURIERI, Giorgini Augusto; AGUIAR, Jaime Paiva Lopes. Composição do chocolate caseiro de amêndoas de cupuaçu (*Theobroma grandiflorum* (Willd ex Spreng) Schum). **Acta Amazonica**, [*s. l.*], v. 18, n. 1–2, p. 3–8, 1988.

VERDE, Ignazio *et al.* The high-quality draft genome of peach (Prunus persica) identifies unique patterns of genetic diversity, domestication and genome evolution. **Nature Genetics**, [*s. l.*], v. 45, n. 5, p. 487–494, 2013.

WANG, Yan *et al.* Evasion of plant immunity by microbial pathogens. **Nature Reviews Microbiology**, [*s. l.*], v. 20, n. 8, p. 449–464, 2022.

WANG, Zhong *et al.* 2 regulates embryonic fatty acid biosynthesis by targeting 3 during the early developmental stage of *Arabidopsis* seeds. **The Plant Journal**, [*s. l.*], v. 77, n. 5, p. 757–769, 2014.

WHITLOCK, Barbara A.; BAUM, David A. Phylogenetic Relationships of *Theobroma* and *Herrania* (Sterculiaceae) Based on Sequences of the Nuclear Gene Vicilin. **Systematic Botany**, [*s. l.*], v. 24, n. 2, p. 128, 1999.

WALKS, B. Metcalfe, C. R. & Chalk, L. 1983. Anatomy of the dicotyledons 2, 2. **Nordic Journal of Botany**, v. 4, n. 5, p. 668–668, dez. 1984.

WANG, Fanlong *et al. Arabidopsis* P4 ATPase-mediated cell detoxification confers resistance to Fusarium graminearum and Verticillium dahliae. **Nature Communications**, [*s. l.*], v. 12, n. 1, p. 6426, 2021.

WHITLOCK, Barbara A.; BAUM, David A. Phylogenetic Relationships of *Theobroma* and *Herrania* (Sterculiaceae) Based on Sequences of the Nuclear Gene Vicilin. **Systematic Botany**, [*s. l.*], v. 24, n. 2, p. 128–138, 1999.

WU, Guo-Zhang; XUE, Hong-Wei. *Arabidopsis* β-Ketoacyl-[Acyl Carrier Protein] Synthase I Is Crucial for Fatty Acid Synthesis and Plays a Role in Chloroplast Division and Embryo Development. **The Plant Cell**, [*s. l.*], v. 22, n. 11, p. 3726–3744, 2010.

XU, Z.; DENG, M. Malvaceae. *Em*: XU, Z.; DENG, M. (Eds.). . **Identification and Control of Common Weeds: Volume 2**. Dordrecht: Springer Netherlands, 2017. p. 717–735.

XU, Qiang *et al.* The draft genome of sweet orange (Citrus sinensis). **Nature Genetics**, [*s. l.*], v. 45, n. 1, p. 59–66, 2013.

ZARRILLO, Sonia *et al.* The use and domestication of *Theobroma cacao* during the mid-Holocene in the upper Amazon. **Nature Ecology & Evolution**, [*s. l.*], v. 2, n. 12, p. 1879–1888, 2018.

ZHANG, Liyi *et al.* A high-quality apple genome assembly reveals the association of a retrotransposon and red fruit colour. **Nature Communications**, [*s. l.*], v. 10, n. 1, p. 1494, 2019.

ZHANG, Jisen *et al.* Allele-defined genome of the autopolyploid sugarcane Saccharum spontaneum L. **Nature Genetics**, [*s. l.*], v. 50, n. 11, p. 1565–1573, 2018.

ZHARKIKH, Andrey *et al.* Sequencing and assembly of highly heterozygous genome of *Vitis vinifera* L. cv Pinot Noir: Problems and solutions. **Journal of Biotechnology**, [*s. l.*], v. 136, n. 1, Genome Research in the Light of Ultrafast Sequencing Technologies, p. 38–43, 2008.

ZOHARY, Daniel. Unconscious selection and the evolution of domesticated Plants. **Economic Botany**, [*s. l.*], v. 58, n. 1, p. 5–10, 2004.

ZUGAIB, Maria *et al.* Comparative proteome profile of ungerminated spores and mycelium of the fungus *Moniliophthora roreri*, causal agent of frosty pod rot disease in *cacao*. **Journal of Phytopathology**, [*s. l.*], v. 171, n. 6, p. 242–257, 2023.

CAPÍTULO 2 - Diversificação Funcional dos Genes COBRA-LIKE (COBLs) em *Theobroma* Amazônicos: Estudo Comparativo entre *T. grandiflorum*, *T. cacao* e *T. umbraticum*

RESUMO - A expansão da parede celular é essencial para o desenvolvimento e a morfologia das plantas, influenciando diretamente a estrutura dos frutos. A família multigênica COBRA-LIKE (COBL) desempenha um papel crucial nesse processo, codificando proteínas ancoradas em glicosilfosfatidilinositol (GPI) que regulam a organização e deposição da parede celular. Este estudo investiga os genes COBLs em clones e cultivares de *Theobroma grandiflorum* (cupuaçu), *T. cacao* (cacau) e *Theobroma umbraticum*, espécies amazônicas de importância econômica e cultural, visando compreender melhor suas funções na expansão da parede celular e deposição de celulose. Foram identificados 62 genes COBLs distribuídos em dois subgrupos: 12 em *T. grandiflorum* (clone C1074 suscetível à vassoura-de-bruxa), 10 em *T. grandiflorum* (clone C174 resistente à vassoura-de-bruxa), 16 em *T. cacao* (cultivar Matina 1-6), 12 em *T. cacao* (cultivar B97-61/B2 - Criollo) e 12 em *T. umbraticum*. A análise estrutural dos genes revelou que o subgrupo I contém seis éxons, enquanto o subgrupo II possui dois, indicando uma significativa conservação estrutural. Os COBLs apresentam distribuição conforme padrões evolutivos distintos, derivados de eventos de duplicação como duplicações segmentares em grande escala (WGD), implicando uma origem mais antiga, e duplicações menores (tandem, proximal e dispersas), representando eventos mais recentes. Hipotetizamos que esses eventos possivelmente influenciaram características agronômicas nas espécies domesticadas de cacau e cupuaçu. Análises filogenéticas indicam que cinco genes (um para cada cultivar) não se agruparam com membros de *Arabidopsis thaliana* e outras monocotiledôneas e *eudicotiledôneas*, sugerindo a presença de novos genes exclusivos nas espécies de Malvaceae. A descoberta de genes COBLs exclusivos das Malvaceae sugere adaptações evolutivas e diversificação funcional. Este estudo destaca possíveis correlações entre genes COBLs e a morfologia, crescimento e qualidade dos frutos dessas plantas, ressaltando sua importância para avanços futuros no melhoramento de plantas e na agricultura.

Palavras-chave: Genômica Comparativa, Filogenômica, Parede celular, Qualidade de Frutos, Diversidade Amazônica

Functional Diversification of COBRA-LIKE (COBL) Genes in Amazonian *Theobroma*: A Comparative Study of *T. grandiflorum, T. cacao,* **and** *T. umbraticum*

ABSTRACT - Cell wall expansion is essential for plant development and morphology, directly influencing fruit structure. The COBRA-LIKE (COBL) multigene family plays a crucial role in this process by encoding glycosylphosphatidylinositol (GPI)-anchored proteins that regulate cell wall organization and deposition. This study investigates COBL genes in clones and cultivars of *Theobroma grandiflorum* (cupuaçu), *T. cacao* (*cacao*), and *Theobroma umbraticum*, Amazonian species of economic and cultural importance, aiming to better understand their functions in cell wall expansion and cellulose deposition. A total of 62 COBL genes were identified, distributed into two subgroups: 12 in *T. grandiflorum* (clone C1074, susceptible to witches' broom), 10 in *T. grandiflorum* (clone C174, resistant to witches' broom), 16 in *T. cacao* (cultivar Matina 1-6), 12 in *T. cacao* (cultivar B97-61/B2 - Criollo), and 12 in *T. umbraticum*. Structural gene analysis revealed that subgroup I contains six exons, while subgroup II has two, indicating significant structural conservation. COBLs exhibit distribution patterns aligned with distinct evolutionary trajectories, derived from duplication events such as large-scale segmental duplications (WGD), suggesting an ancient origin, as well as smaller duplications (tandem, proximal, and dispersed), representing more recent events. We hypothesize that these events may have influenced agronomic traits in domesticated *cacao* and cupuaçu species. Phylogenetic analyses indicate that five genes (one per cultivar) did not cluster with members of *Arabidopsis thaliana* or other monocotyledonous and dicotyledonous species, suggesting the presence of novel genes exclusive to Malvaceae species. The discovery of COBL genes unique to Malvaceae suggests evolutionary adaptations and functional diversification. This study highlights possible correlations between COBL genes and fruit morphology, growth, and quality in these plants, emphasizing their importance for future advancements in plant breeding and agriculture.

Keywords: Comparative Genomics, Phylogenomics, Cell wall, Fruit Quality, Amazonian Diversity.

1. INTRODUÇÃO

Theobroma grandiflorum (Willd. ex Spreng.) K.Schum (cupuaçu) e *Theobroma cacao* L. (cacau) são espécies arbóreas nativas da Amazônia, pertencentes à família Malvaceae. No panorama das 23 espécies catalogadas dentro do gênero *Theobroma* L., destacam-se a significativa importância econômica compartilhada entre o cupuaçu e o cacau, enquanto as outras espécies têm menor relevância (Colli-Silva; Richardson; Pirani, 2023; Colli-Silva *et al*., 2024; Flora e Funga do Brasil, 2024). Por exemplo, *Theobroma umbraticum*, apresenta frutos baciformes com polpa doce consumida por humanos e pela vida selvagem (Bletter & Daly, 2006; Colli-Silva *et al*., 2024). O gênero *Theobroma* é monofilético, sendo um grupo irmão do gênero *Herrania*, mas a monofilia de *Theobroma* é fracamente suportada (Sousa Silva & Figueira, 2005). Estudos recentes propõem a sinonimização do gênero *Herrania* no gênero *Theobroma*, fundamentando-se em uma análise filogenética. Esses estudos utilizam dados moleculares para revisar as relações evolutivas entre essas espécies, revelando uma proximidade genética mais estreita do que se supunha anteriormente. Essa revisão taxonômica sugere uma reorganização dentro da família, alinhando *Herrania* e *Theobroma* em um único gênero, o que pode ter implicações significativas tanto para a classificação botânica quanto para estudos futuros envolvendo essas espécies.

Os produtos alimentícios derivados do grão de cacau, como o chocolate, desfrutam de ampla popularidade global, abrangendo todas as faixas etárias. Essa demanda mundial por produtos à base de cacau tem um impacto significativo na economia da região amazônica, que se destaca por sua concentração na atividade alimentar, constituindo-se como uma das principais esferas econômicas locais. Os frutos do cupuaçu também têm suscitado um interesse crescente no âmbito do desenvolvimento de subprodutos, tanto no setor agroalimentar, por meio da produção de sucos, geléias, manteigas, doces, sorvetes e chocolate conhecido como "cupulate" (Cohen & Jackix, 2005; McNeil, 2006; Alves *et al*., 2007; Genovese & Lannes, 2009) quanto na produção de cosméticos (Silva da Costa *et al*., 2022). Porém, sua notoriedade ainda é limitada na esfera mundial quando comparado ao cacau.

O comércio e o cultivo do cupuaçu estão vinculados à produção comunitária local, sendo sua importância econômica crescente nas últimas décadas.

Dessa forma, a produção de cupuaçu desempenha um papel fundamental na economia regional, registrando um volume anual estimado de 1,6 milhões de toneladas em 2024, em uma área de 224 mil hectares, segundo dados da Secretaria de Estado de Desenvolvimento Agropecuária e da Pesca (SEDAP, 2024). No Brasil, especialmente na região norte, essa cultura representa uma atividade agrícola de grande relevância para pequenos produtores em sistemas agroflorestais e cooperativas regionais (Alves & Chaves, 2023).

Nesse cenário, para desenvolver plantas mais produtivas e resilientes, é importante aprofundar nossa compreensão acerca dos processos de seleção de clones ou cultivares superiores (Alves *et al*., 2017). Além disso, é necessário investigar a base molecular através das funções de genes e dos mecanismos evolutivos inerentes a cada espécie, estabelecendo paralelos biológicos e realizando comparações com culturas de referência (Niu *et al*., 2019; Mournet *et al*., 2020; de Abreu *et al*., 2023).

Nesse contexto, o advento das técnicas de sequenciamento de nova geração revolucionou o cenário de melhoramento de diversos cultivares, disponibilizando informações com possível aplicabilidade em campo. Atualmente, é possível realizar análises mais precisas de todo o genoma devido à disponibilidade de dados genômicos de *T. grandiflorum*, *T. cacao* e *T. umbraticum* de alta qualidade, recentemente sequenciados e disponíveis em bases de dados públicas (Argout *et al*., 2011; Motamayor *et al*., 2013; Alves *et al*., 2024).

Foram identificados 31.381, 33.046, 30.623, 28.896 e 27.905 genes codificadores de proteínas nos genomas do cupuaçu nos exemplares suscetível e resistente à vassoura-de-bruxa, no cacau para os cultivares B97-61/B2 - Criollo e Matina 1-6, e em *Theobroma umbraticum*, respectivamente (Alves *et al*., 2024). Dentre essa extensa coleção de genes, destaca-se uma família gênica de relevância agronômica, conhecida como COBRA-LIKEs (COBLs). Essa família gênica faz parte de uma classe multigênica específica de plantas, similar à COB de *Arabidopsis thaliana*, e codifica proteínas ancoradas em glicosilfosfatidilinositol (GPI) que incorporam a região hidrofílica no terminal C, normalmente apresentando um local de ligação ω para modificação de GPI, acompanhado por um sinal de secreção N-

terminal (Liu *et al*., 2013; Yang *et al*., 2021). As proteínas COBLs apresentam o domínio COBRA [identificador PFAM: PF04833.18] (Roudier *et al*., 2002).

A ligação das proteínas COBL com o grupo GPI (glicosilfosfatidilinositol) é um processo específico e importante para a função dessas proteínas. Dentro do retículo endoplasmático, a proteína COBL recebe uma âncora GPI, um tipo de lipídio que se liga à proteína através de um aminoácido específico na extremidade C-terminal, frequentemente glicina, serina ou asparagina. Assim, o grupo GPI se liga covalentemente à proteína COBL por meio da enzima transamidase, que reconhece o sinal específico na proteína, corta a extremidade C-terminal e liga o GPI no local apropriado (Brady *et al*., 2007; Ye *et al*., 2009). Após a adição do GPI, a proteína COBL ancorada é transportada para o complexo de Golgi dentro de vesículas e, posteriormente, é transportada para a superfície da célula, onde o GPI ajuda a ancorá-la na membrana ou na parede celular (Schindelman *et al*., 2001; Roudier *et al*., 2002).

Os integrantes da família COBL foram reconhecidos por desempenharem papéis em diversos processos fisiológicos e de desenvolvimento da planta, abrangendo a resistência do caule, o crescimento do tubo polínico, a resistência a patógenos e o desenvolvimento de pelos radiculares (Ko *et al*., 2006; Li *et al*., 2013, 2022; Yang *et al*., 2021).

A família COBL é conservada em monocotiledôneas e eudicotiledôneas (Roudier *et al*., 2002; Brady *et al*., 2007), sendo identificada em várias espécies como tomate (*Solanum lycopersicum*) (Cao *et al*., 2012), milho (*Zea mays*) (Brady *et al*., 2007; Ching *et al*., 2006), arroz (*Oryza sativa*) (Liu *et al*., 2013), algodão (*Gossypium* spp.) (Niu *et al*., 2015), *Populus* (Ye *et al*., 2009) e *Arabidopsis thaliana* (Roudier *et al*., 2002), apresentando diferentes funções (Tabela 1). *Arabidopsis* , milho e arroz possuem, respectivamente, 12, 11 e 9 genes COBLs, os quais podem ser agrupados em dois subgrupos principais. Um desses subgrupos é caracterizado pela presença de aminoácidos adicionais na região N-terminal (Roudier *et al*., 2002).

Tabela 1. Genes da família COBRA-LIKE (COBL) presentes no genoma de *Arabidopsis thaliana*, desempenhando funções cruciais na biossíntese, cristalização e integração da celulose nas paredes celulares.

Família gênica	Nome da proteína	Função	IDs UniProt	Referência
Subfamília 1	COB:COBRA	Facilita a cristalização da celulose na membrana plasmática. Em *Oryza sativa* (arroz), o gene BC1 apresentou 60,7% homólogo ao gene COB de *Arabidopsis* ; plantas *bc1* apresentaram parede celular reduzida, menor conteúdo de celulose e rigidez do colmo reduzida	Q94KT8 PMID: 25331944	Li *et al*., 2003; Dai *et al*., 2009; Sorek *et al.,* 2014
	COBL1:COBRA-LIKE1	Organização de microfibrilas de celulose. Homólogo ao gene BC1L1 do arroz; altamente expresso em células epidérmicas da raiz; mutantes apresentaram desenvolvimento anômalo de tricoma radiculares e rendimento de campo reduzido.	Q9SRT7	(Hochholdinger *et al.*, 2008)
	COBL2:COBRA-LIKE2 COBL3:COBRA-LIKE3	Envolvido na deposição de celulose em células do tegumento da semente	Q8L8Q7 PMID: 25583925 PMID: 29446495	Ben-Tov *et al.*, 2015; 2018
	COBL4:COBRA-LIKE4	Envolvido na biossíntese da parede celular secundária. Os mutantes produzem plantas menores com níveis reduzidos de celulose e açúcares da parede celular.	Q9LFW3 PMID:15980264 Mutantes: PMID: 29869696	Brown *et al.*, 2005; Niu *et al*. 2018
	COBL5:COBRA-LIKE5	Respostas imunológicas das plantas aos oomicetos. Ausência funcional afetou crescimento e desenvolvimento; induziu produção de substâncias responsivas ao estresse; expressão gênica relacionada à defesa regulada positivamente	Q9FME5 PMID:21711359	Ko *et al*., 2006; Hok *et al*., 2011
	COBL6:COBRA-LIKE6	Suposta fitoquelatina sintase importante para a detoxificação de metais	O04500 PMID: 16790430	Schaaf *et al.*, 2006

		pesados nas células vegetais.		
Subfamília2	COBL7:COBRA-LIKE7	Formação estomática através da regulação da deposição de celulose e regulam o crescimento dos pêlos radiculares.	Q8GZ17 PMID: 37853545; 35104663	Li *et al.*, 2022; Ge *et al.*, 2023
	COBL8:COBRA-LIKE8	Formação estomática através da regulação da deposição de celulose	Q9LIB6 PMID: 37853545	Ge *et al.*, 2023
	COBL9:COBRA-LIKE9	Regulam o crescimento dos pêlos radiculares. Deleção do gene resultou em tricoma radiculares mais curtos e menos numerosos.	Q9FJ13 PMID: 35104663	Jones;Raymond ;Smirnoff, 2006; Li *et al.*, 2022
	COBL10:COBRA-LIKE10	Crescimento direcional de tubos polínicos	Q9LJU0 PMID:23384085	Li *et al.*, 2013
	COBL11:COBRA-LIKE11	Regula a integridade do tubo polínico	Q9T045 PMID: 38007687	Li *et al.*, 2023

Este trabalho apresenta como hipótese central que os genes COBL no cupuaçu, cacau e *Theobroma umbraticum* podem apresentar adaptações evolutivas e diversificação funcional específicas dentro da família das Malvaceae, e que essas diversificações podem influenciar diretamente características agronômicas como morfologia, crescimento e qualidade dos frutos. Portanto, a identificação e caracterização desses genes em nível genômico, através de análises filogenéticas, estrutura gênica, identificação e rotulação de motivos conservados e disposição cromossômica nos três organismos-alvo deste estudo, poderão fornecer informações valiosas sobre a biologia e função destes genes, e que em última instância poderá levar ao desenvolvimento de estratégias de melhoramento genético, visando frutos de melhor qualidade, especialmente no caso do cupuaçu.

2. MATERIAIS E MÉTODOS

2.1 Genoma selecionado para estudo

O presente estudo se baseou na análise de dados genômicos de dois clones de cupuaçu, designadamente o clone C1074, suscetível à vassoura-de-bruxa, e o

clone C174, denominado Coari, reconhecido por sua resistência à vassoura-de-bruxa (Alves *et al.*, 2024). Ambas as amostras fazem parte do acervo genético da Embrapa Amazônia Oriental, situada em Belém, Pará (Mournet *et al.*, 2020). Os dados genômicos foram obtidos a partir dos registros disponíveis no GenBank, acessados por meio dos números de referência PRJNA691024.

Para fins de referência e rastreabilidade, as amostras vegetais foram catalogadas no Herbário JABU, vinculado à Unesp-FCAV, sob o acesso JABU1369 e JABU1370 para os clones C174 e C1074, respectivamente.

Adicionalmente, os genomas de referência das espécies *Theobroma umbraticum* cultivar*: Fairchild (draft)* e *Theobroma cacao* (cultivares B97-61/B2 - Criollo e Matina 1-6) em escalas cromossômicas foram utilizadas (Argout *et al.*, 2011; Motamayor *et al.*, 2013). Os dados genômicos, incluindo os arquivos de anotação em formato FASTA, foram obtidos a partir dos registros disponíveis no GenBank, acessados por meio dos números de referência PRJNA383741, PRJEB14326 e PRJNA239715, respectivamente.

Os dados genômicos das espécies desse estudo também estão disponíveis na plataforma de dados Plant Genomics (https://plantgenomics.ncc.unesp.br).

2.2. Busca por genes e mapeamento dos genes nos cromossomos

As proteínas correspondentes às sequências codificadas pelos genes membros da família COBRA-Like em *Arabidopsis thaliana* foram obtidos a partir da base de dados UniProt (The UniProt Consortium 2023): COB [Q94KT8]; COBL1 [Q9SRT7]; COBL2 [Q8L8Q7]; COBL4 [Q9LFW3]; COBL5 [Q9FME5]; COBL6 [O04500]; COBL7 [Q8GZ17]; COBL8 [Q9LIB6]; COBL9 [Q9FJ13]; COBL10 [Q9LJU0]; COBL11 [Q9T045] (Tabela 1). Estas sequências foram empregadas na identificação de proteínas similares preditas nos arquivos de anotações nos clones C174 e C1074 de *T. grandiflorum*, bem como nas cultivares B97-61/B2 - Criollo e Matina1-6 de *T. cacao*, e em *T. umbraticum*. Essa análise foi executada utilizando a ferramenta BLASTP na plataforma web Plant Genomic. As proteínas candidatas foram

selecionadas de acordo com seu E-value de 1e-5 e similaridade e cobertura de alinhamento superior a 60%.

Adicionalmente, foi conduzida uma análise complementar utilizando a ferramenta HmmerWeb versão 2.41.2 (Potter *et al.*, 2018) em conjunto com a base de dados Pfam (Mistry *et al.*, 2021). Tal procedimento visou a confirmação da presença do domínio COBRA [PF04833.18] na família de genes COBRA-Like previamente identificadas. As sequências desprovidas do domínio COBRA foram excluídas das análises subsequentes.

Todos os genes foram mapeados nos cromossomos de *T. grandiflorum* e *T. cacao* e montagem *draft* de *T. umbraticum*, seguindo as informações de distribuição, tamanho inicial e terminal dos genes e origem estrutural, pelas informações de anotação genômica previamente disponibilizadas pela base de dados Plant Genomics DB (Alves *et al.*, 2024). Para efetuar esse procedimento, utilizou-se a ferramenta MG2C (V2.1) (Chao *et al.*, 2021).

2.3. Expressão gênica em transcritos por milhão (TPM)

No caso específico de *Theobroma grandiflorum*, as análises de TPM foram conduzidas em folhas jovens, onde o RNA foi extraído de folhas frescas usando o protocolo modificado de CTAB. Os dados de sequenciamento de RNAseq e Iso-Seq foram disponibilizados em nosso estudo prévio (Alves *et al.*, 2024). Por outro lado, em *Theobroma umbraticum* e *Theobroma cacao*, a análise de TPM foi realizada a partir de diferentes tecidos da planta, a partir de todos os dados públicos disponíveis GenBank usados para montagem do transcriptoma. Os dados de RNA-seq de *Theobroma umbraticum* utilizados foram: SRR5630802 flores abertas; SRR5481331 flores abertas; SRR5481332 gemas laterais; SRR5481333 botões florais fechados; SRR5481334 caules apicais; SRR5462401 folhas jovens. Enquanto *Theobroma cacao* incluem tecidos de folhas, casca de baga, caules, raiz, flor e sementes. Os números de acesso SRA do GenBank utilizados para todos os dados de RNA-seq de *Theobroma cacao* utilizados estão disponíveis na Tabela suplementar 3.

2.4. Visualização da estrutura gênica e análise de domínios

A informação estrutural referente aos éxon, íntrons e regiões não traduzidas (UTR) dos genes COBLs nas espécies em análise foi extraída do arquivo GFF3 da anotação disponibilizada na base de dados Plant Genomics DB e recentemente publicada por nosso grupo de pesquisa (Alves *et al.*, 2024). A análise estrutural se baseou também no programa MCScanX (Wang *et al.*, 2012), onde os genes são organizados conforme sua topologia no genoma, em genes derivados de grandes duplicações (segmental ou WGD), duplicados em tandem, duplicados proximais (proximal) e duplicados dispersos.

O software Tbtools - v2.019 (Chen *et al.*, 2023) foi empregado para a análise e visualização dos domínios protéicos e estrutura gênica, utilizando as funcionalidades Visualize Domain Pattern (from Pfam Search) e Gene Structure View (Advanced).

2.5. Análises Filogenéticas

No contexto da investigação das relações filogenéticas entre *Theobrama grandiflorum, Theobroma cacao* e *Theobroma umbraticum*, uma árvore filogenética foi elaborada utilizando o método de Máxima Verossimilhança (Trifinopoulos *et al.,* 2016). Nesse sentido, para garantir a precisão e a abrangência da análise, foram necessárias a aquisição e a seleção de sequências proteicas completas de *Arabidopsis thaliana*, disponíveis na base de dados UniProt (The UniProt Consortium 2023). Especificamente, foram escolhidas todas as sequências revisadas com os seguintes números de acesso: Q9LJU0, Q8L8Q7, Q9T045, Q9FJ13, Q8GZ17, Q9LFW3, Q9FME5, Q9SRT7, Q9LIB6 e O04500 (Tabela 1).

Através do alinhamento de múltiplas sequências utilizando o programa MAFFT versão 7 (Katoh *and* Standley 2013), foi viabilizada a construção de uma árvore filogenética representativa da família COBRA-Likes pelo programa IQ-TREE versão 2 (Nguyen *et al.*, 2015). A inferência filogenética foi realizada utilizando o modelo de substituição JTT+G4, selecionado como o de melhor ajuste com base no

critério de informação bayesiano (BIC) por meio do ModelFinder (Kalyaanamoorthy *et al.,* 2017). Os valores de suporte foram estimados com 1000 réplicas de bootstrap ultrarrápido (UFBoot2) (Hoang *et al*., 2018) e estão indicados ao longo dos ramos. A árvore de consenso foi gerada a partir dessas réplicas de bootstrap. A visualização e edição da árvore filogenética foi realizada utilizando a plataforma iTOL, versão 6 (Letunic & Bork, 2024).

A análise filogenética das isoformas de proteínas codificadas pelos genes COBRA-Like foi expandida para incluir uma ampla variedade de espécies pertencentes às ordens Asterales, Brassicales, Cucurbitales, Fabales, Fagales, Lamiales, Malpighiales, Malvales, Poales, Rosales e Solanales com a finalidade de verificar se as isoformas exclusivas de *Theobroma cacao*, *Theobroma grandiflorum* e *Theobroma umbraticum* se agrupam com as isoformas de genes COBRA-Like de outras espécies. Nesse sentido, o alinhamento foi realizado com o MAFFT versão 7 (Katoh *and* Standley 2013), e a análise filogenética foi conduzida utilizando o programa FastTree v2.1 (Price, Dehal e Arkin, 2010), configurado com os parâmetros padrão para sequências de aminoácidos, baseando-se no modelo evolutivo JTT+CAT. Os genomas analisados foram obtidos das bases de dados públicas NCBI RefSeq e GenBank, e a identificação da presença do domínio COBRA no proteoma das espécies foi realizada por meio do programa InterProScan v5.67-99.0 (Quevillon *et al.*, 2005). Após a obtenção das isoformas únicas, aplicou-se um script específico disponível em https://www.zybuluo.com/lakesea/note/1743213 para o refinamento dos dados e validação final, garantindo a qualidade dos dados, pois remove isoformas redundantes, padroniza sequências e valida a presença do domínio COBRA, cruzando informações do InterProScan. Além disso, formata os dados para análises filogenéticas, assegurando sua consistência.

2.6. Modelagem molecular e análise estrutural das proteínas COBRA-Like nas espécies estudadas

A predição das estruturas tridimensionais das proteínas COBRA-Like codificadas pelos genes das espécies estudadas foi realizada utilizando a plataforma

AlphaFold v.3 (Abramson *et al*., 2024). Para garantir a precisão das predições, a qualidade das estruturas foi avaliada com a plataforma SAVES 6.0, que inclui uma análise detalhada do gráfico de Ramachandran, permitindo a visualização da conformação estereoquímica das proteínas modeladas (Laskowski *et al*., 1993).

Para realizar a comparação estrutural das proteínas COBRA-Like identificadas como exclusivas nas espécies estudadas, bem como das proteínas COBL6, escolhida por ser a mais próxima das proteínas exclusivas em termos de sequência e estrutura, foram utilizados arquivos no formato PDB, gerados pela modelagem molecular do AlphaFold. Os modelos inicialmente gerados no formato CIF foram convertidos para o formato PDB utilizando o software PyMOL (v2.3.1). A sobreposição estrutural foi então conduzida usando o programa FATCAT (https://fatcat.godziklab.org/fatcat/fatcat_pair.html), que permite o alinhamento flexível das estruturas (Li *et al.*, 2020). Esse método de alinhamento flexível permite uma comparação abrangente das estruturas das proteínas, sem impor restrições rígidas, facilitando a identificação de similaridades e diferenças ao longo de toda a cadeia proteica.

A análise de sobreposição estrutural foi realizada especificamente para as proteínas com os identificadores *Tgrandc1074g00000008413*.1 e *Tgrandc1074g00000008646*.1, uma vez que as demais proteínas analisadas seguiram o mesmo modelo estrutural, o que tornava desnecessário incluir todas elas na comparação. Os resultados estão acessíveis através da plataforma FATCAT no link: https://fatcat.godziklab.org/fatcat/result/1733691421.2239/index.html.

3. RESULTADOS E DISCUSSÃO

3.1. Padrões de distribuição e expressão (TPM) de genes COBRA-Like em *T. grandiflorum, T. cacao* e *T. umbraticum*: predominância de duplicações em tandem sugerindo adaptação funcional

Os genomas montados em escala cromossômica de *T. grandiflorum* clones C1074 e C174, *T. cacao* cultivar B97-61/B2 - Criollo e Matina 1-6 são constituídos por uma coleção de 10 cromossomos, cujas dimensões variam de 28 a 53 megabases (Mb) para o clone C1074, de 25 a 52 Mb para o clone C174, 19 a 41 Mb para a cultivar B97-61/B2 - Criollo e de 21 a 43 Mb para a cultivar Matina 1-6 culminando em uma

extensão total de 423.916.809, 415.767.609, 324.761.211 e 341.709.215 pares de bases (pb), respectivamente (Argout *et al.*, 2011; Motamayor *et al.*, 2013; Alves *et al.*, 2024). O genoma de *T. umbraticum* apresenta 621 *scaffolds* com dimensões variáveis de 7 a 22 Mb e extensão total de 249.829.655 pb (Alves *et al.*, 2024).

Foram identificados 62 genes que codificam proteínas COBRA-Like. Estes genes encontram-se distribuídos entre os diversos membros da família COBL nas espécies *T. grandiflorum, T. cacao* e *T. umbraticum* (Tabela 2). Destaca-se que a cultivar *T. cacao*-Matina 1-6 apresentou a maior representatividade entre os genes COBL identificados. Onde, 16 membros COBLs foram encontrados em *T. cacao*-Matina 1-6, 12 em *T. cacao* B97-61/B2 - Criollo, 10 em *T. grandiflorum* C174, 12 em *T. grandiflorum* C1074 e 12 em *T. umbraticum.* Notavelmente, um representante de cada espécie e clone analisada não se agrupou com nenhum membro COBL da espécie modelo *Arabidopsis thaliana*, sugerindo a existência potencial de novos genes COBL exclusivos das espécies de Malvaceae investigadas (Tabela 2).

Tabela 2. Quantidade de genes da família COBRA-Like identificados em *Theobroma cacao* (Criollo e Matina), *Theobroma grandiflorum* (174 e 1074) e *Theobroma umbraticum*. A tabela inclui informações sobre os grupos, subfamília, identificação dos genes, tamanho total das proteínas (em resíduos de aminoácidos, aa) e a região correspondente ao alinhamento com o domínio COBRA (PFAM), indicando as posições inicial e final dentro da sequência proteica.

Grupo	Subfamíla	Identificação	Tamanho (aa)	Alinhamento COBL (aa)	
				Início	Fim
1		sp|Q9SRT7|COBL1_ARATH	452	57	220
		sp|Q8L8Q7|COBL2_ARATH	441	52	215
		sp|Q9FME5|COBL5_ARATH	204	46	201
		Tcacao-CriolloG00000015522.1	455	60	223
		Tcacao-CriolloG00000011662.2	456	60	223
		Tcacao-MatinaG00000014208.1	455	60	223
		Tcacao-MatinaG00000000122.1	456	60	223
	COBL1,2,5	*Tcacao-MatinaG00000010512.1*	348	1	129
		HumbraticaG00000019843.1	455	60	223
		HumbraticaG00000014817.1	456	60	223
		TgrandC174G00000014592.1	455	60	223
		TgrandC174G00000010808.1	491	95	258
		TgrandC1074G00000015762.1	473	60	241
		TgrandC1074G00000011891.1	491	95	258
		sp|Q9LFW3|COBL4_ARATH	431	44	207
		Tcacao-CriolloG00000015523.3	432	43	206

	Tcacao-CriolloG00000011663.1	437	49	212
	Tcacao-CriolloG00000011667.1	388	1	162
	Tcacao-CriolloG00000011671.1	160	1	129
	Tcacao-MatinaG00000014209.2	432	43	206
	Tcacao-MatinaG00000000123.2	463	75	238
	Tcacao-MatinaG00000010516.1	438	49	212
	Tcacao-MatinaG00000000126.1	384	8	158
COBL4	*Tcacao-MatinaG00000010513.1*	150	1	124
	Tcacao-MatinaG00000010518.1	160	1	129
	HumbraticaG00000019844.3	432	43	206
	HumbraticaG00000014818.1	463	75	238
	HumbraticaG00000014819.1	438	49	212
	TgrandC174G00000014593.1	432	43	206
	TgrandC174G00000010809.1	463	75	238
	TgrandC174G00000010810.1	438	49	212
	TgrandC1074G00000015761.1	432	43	206
	TgrandC1074G00000011890.1	463	75	238
	TgrandC1074G00000011889.1	438	49	212
	TgrandC1074G00000011886.1	149	21	109
	sp|O04500|COBL6_ARATH	454	59	222
	Tcacao-CriolloG00000006609.1	443	57	216
	Tcacao-MatinaG00000005908.1	443	57	216
COBL6	*HumbraticaG00000015638.1*	443	57	216
	TgrandC174G00000008356.1	443	57	216
	TgrandC1074G00000008413.1	443	57	216
	Tcacao-CriolloG00000006352.1	**443**	**51**	**213**
	Tcacao-MatinaG00000005665.1	**443**	**51**	**213**
COBL EXCLUSIVO	***HumbraticaG00000015867.1***	**443**	**51**	**213**
	TgrandC174G00000008604.1	**443**	**51**	**213**
	TgrandC1074G00000008646.1	**443**	**51**	**213**
	sp|Q8GZ17|COBL7_ARATH	661	233	412
	sp|Q9LIB6|COBL8_ARATH	653	226	405
	sp|Q9FJ13|COBL9_ARATH	663	232	411
	Tcacao-CriolloG00000017670.1	652	227	405
	Tcacao-CriolloG00000004346.1	655	231	410
	Tcacao-MatinaG00000016291.1	652	227	405
COBL7,8,9	*Tcacao-MatinaG00000003724.1*	655	231	410
	HumbraticaG00000021654.1	652	227	405
	HumbraticaG00000010520.1	655	231	410
	HumbraticaG00000010515.1	655	231	410
	TgrandC174G00000016764.1	652	227	405
	TgrandC174G00000003636.1	655	231	410
	TgrandC1074G00000013516.1	652	227	405

2

	TgrandC1074G00000001217.1	655	231	410
COBL10, 11	*sp|Q9LJU0|CBL10_ARATH*	672	245	424
	sp|Q9T045|CBL11_ARATH	668	235	414
	Tcacao-CriolloG00000025455.2	662	234	413
	Tcacao-CriolloG00000024510.1	547	229	408
	Tcacao-MatinaG00000023553.2	662	234	413
	Tcacao-MatinaG00000022663.1	314	1	78
	Tcacao-MatinaG00000022662.1	302	219	302
	HumbraticaG00000018922.1	662	234	413
	HumbraticaG00000025413.1	548	184	306
	TgrandC174G00000021411.1	687	262	441
	TgrandC1074G00000027067.1	662	234	413
	TgrandC1074G00000021804.1	617	192	371

O arranjo espacial e a distribuição dos genes nos genomas de *T. grandiflorum, T. cacao* e *Theobroma umbraticum* mostram um padrão semelhante de acordo com seus laços evolutivos intimamente relacionados (Figura 1). Este padrão inclui genes de duplicações de eventos em grande escala originários do evento WGD (Whole Genome Duplication) como segmental e eventos de duplicação em pequena escala como tandem, proximal e dispersos. Os genes resultantes dessas duplicações em pequena escala, em particular, apresentam uma natureza geralmente mais jovem em termos evolutivos, frequentemente emergindo como resposta imune a patógenos, estresse abiótico e biótico (Wang *et al.*, 2016).

Entretanto, notou-se uma distribuição desigual em seis, dentre os dez cromossomos, evidenciando uma peculiaridade na organização genômica dos genes COBL no gênero *Theobroma*. Nenhum gene da família COBL foi identificado nos cromossomos 5, 8, 9 e 10 para *T. cacao* Criollo e Matina, bem como para *T. grandiflorum* clone C1074. Esta ausência também foi observada em *T. grandiflorum* clone C174, abrangendo os cromossomos 5, 8, 9 e 10, assim como o cromossomo 7. Em *H. umabratica*, apesar de não montado em nível de cromossomos, houve presença de COBL em apenas cinco *scaffolds*, sendo esses, *scaffold* 5, 8, 9, 11, 14 (Figura 1).

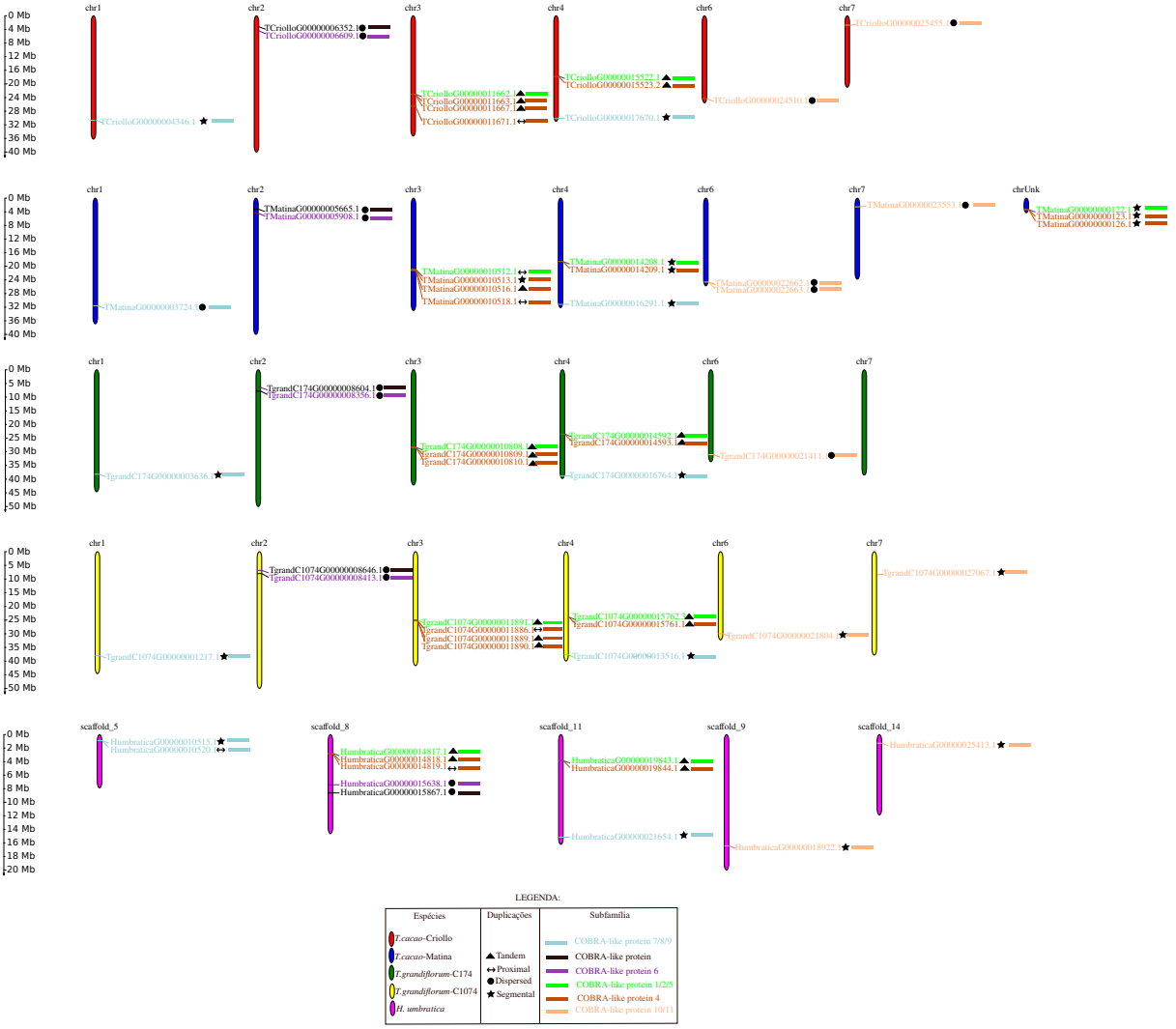

Figura 1. Mapeamento cromossômico e genético dos genes COBL identificados, delineando suas localizações específicas nos genomas estudados.

A distribuição dos genes COBL foi observada exclusivamente em seis cromossomos (1º, 2º, 3º, 4º, 6º, 7º) nas cultivar *T.cacao* B97-61/B2-Criollo e Matina 1-6, assim como na *T.grandiflorum* C1074. Destaca-se, contudo, uma exceção notável em *T.grandiflorum* C174, onde não houve presença de COBL cromossomo 7.

A predominância dos genes COBL foi constatada nos cromossomos 3 e 4, que, conjuntamente, englobam 54% de todos os genes homólogos a COBRA-Like presentes no genoma do gênero *Theobroma*. Para a espécie *T. umbraticum*, os scaffolds 8 e 11 destacaram-se como os que apresentaram maior predominância de COBL sendo 66,7%.

Essas duplicações distintas incluem derivados de duplicações tandem, proximal, dispersa, sendo essas, originárias de eventos de menor escala. E derivados de duplicações em grande escala como segmental em *T. grandiflorum* clone C174 e C1074*, T. cacao* cultivar Matina 1-6 e B97-61/B2 - Criollo e *T. umbraticum* (Tabela 3).

Tabela 3. Número de genes derivados de diferentes tipos de duplicação gênica em *Theobroma grandiflorum* (clones C174 e C1074), *Theobroma cacao* (cultivares Matina 1-6 e B97-61/B2 - Criollo) e *Theobroma umbraticum*. Os tipos de duplicação incluem tandem, proximal, dispersa e em grande escala (segmental).

Duplicação	***T. grandiflorum* (C174)**	***T. grandiflorum* (C1074)**	***T. cacao* (Matina 1-6)**	***T. cacao* (B97-61/B2 - Criollo)**	***T. umbraticum***
Tandem	5	5	1	5	4
Proximal	0	1	2	1	2
Dispersa	3	2	6	4	2
Segmental	2	4	7	2	4

No contexto do nosso estudo, observamos que o arranjo espacial e a distribuição dos genes COBLs nos genomas de *T. grandiflorum, T. cacao,* e *T. umbraticum* foram mais predominantes em duplicações tandem. Esses resultados sugerem implicações significativas para as plantas domesticadas, como o cupuaçu e o cacau (Colli-Silva *et al.,* 2023b), onde a exposição a condições severas de estresse biótico e abiótico devido à domesticação dessas espécies podem ter desencadeado estas amplificações contíguas (Hanada *et al.,* 2008; Zou *et al.,* 2009).

Essas alterações genéticas podem ser interpretadas como respostas adaptativas ao estresse ambiental e podem culminar em modificações estruturais no genoma, nesse caso com o ganho de segmentos de DNA contíguos, resultando ou não em funções distintas nos genes (Hanada *et al.,* 2008; Zou *et al.,* 2009). Além disso, o estresse biótico pode aumentar ainda mais a variação em genes tandem, uma vez que a taxa de recombinação é elevada sob ataque de patógenos (Lucht *et al.,* 2002).

Notavelmente, observou-se que as duplicações em tandem foram mais proeminentes nas espécies domesticadas em comparação com *T. umbraticum*, sugerindo um possível papel da domesticação na diversificação genômica e na adaptação dessas espécies de Malvaceae a ambientes variáveis, podendo sugerir

também que a domesticação pode ter ocasionado em um impacto significativo nas características dos frutos de cupuaçu e cacau, tornando-os mais adaptados às necessidades humanas e possivelmente mais desejáveis em termos de sabor, valor nutricional e resistência a doenças quando comparados aos frutos da espécie selvagem *T. umbraticum*. A domesticação do cupuaçu pelos povos indígenas amazônicos ao longo dos últimos 8.000 anos (Colli-Silva *et al.,* 2023b) pode ter influência nessas características.

As duplicações segmentares, ao contrário das demais, representam a segunda forma mais predominante de duplicação nos genes *COBLs* analisados neste estudo. Esse padrão pode estar relacionado à divergência gênica resultante da triplicação gama, um evento compartilhado por todas as angiospermas e datado em aproximadamente 117 milhões de anos (Jiao *et al*., 2012). No entanto, em *Theobroma*, essa diversificação ocorreu de forma mais recente, há cerca de 14 milhões de anos, e entre os gêneros *Theobroma* e *Herrania*, há 18 milhões de anos (Richardson *et al*., 2015; Kumar *et al*., 2022). Como as duplicações segmentares estão associadas a eventos de grande escala e reorganizações cromossômicas mais antigas, os genes derivados desse processo podem ter sido preservados independentemente das pressões seletivas impostas pela domesticação. Dessa forma, a predominância relativa desse tipo de duplicação pode refletir uma divergência a longo prazo, em vez de um impacto direto dos processos de domesticação.

Todos os genes são expressos nas espécies de Malvaceae estudas, exceto *TgrandC174G00000021411.1; TgrandC174G00000010810.1 TgrandC1074G00000011889*.1; *Tcacao-CriolloG00000011671*.1; *Tcacao-MatinaG00000010518*.1; em *T. grandiflorum* e *T*. *cacao* (Tabela 4). Essa ausência de expressão sugere que esses genes específicos podem não ser funcionais, ou que sua função pode estar condicionada a fatores específicos não presentes nas condições do estudo.

Tabela 4. Valores de TPM (*Transcripts Per Million*) para genes COBRA-Like em *T. cacao* (Criollo e Matina), *T. grandiflorum* (174 e 1074) e *Theobroma umbraticum.*

Subfamíla	**Identificação**	**TPM**	**TPM**		
			Não inoculado	**24 horas após inoculação**	**48 horas após inoculação**
COBL1,2,5	*Tcacao-CriolloG00000015522.1*	50.593.000	-	-	-
	Tcacao-CriolloG00000011662.2	40.234.000	-	-	-
	Tcacao-MatinaG00000014208.1	57.919.700	-	-	-
	Tcacao-MatinaG00000000122.1	42.113.600	-	-	-
	Tcacao-MatinaG00000010512.1	25.496.800	-	-	-
	HumbraticaG00000019843.1	54.641.200	-	-	-
	HumbraticaG00000014817.1	69.620.400	-	-	-
	TgrandC174G00000010808.1	-	198.312.958	179.218.979	191.107.803
	TgrandC174G00000014592.1	-	104.270.325	90.776.268	125.250.557
	TgrandC1074G00000015762.1	-	111.605.759	114.439.583	91.317.940
	TgrandC1074G00000011891.1	-	161.124.313	157.237.015	166.291.687
COBL4	*Tcacao-CriolloG00000015523.3*	41.876.900	-	-	-
	Tcacao-CriolloG00000011663.1	2.539.700	-	-	-
	Tcacao-CriolloG00000011667.1	0.227.700	-	-	-
	Tcacao-CriolloG00000011671.1	0.0	-	-	-
	Tcacao-MatinaG00000014209.2	41.970.800	-	-	-
	Tcacao-MatinaG00000000123.2	43.943.700	-	-	-
	Tcacao-MatinaG00000010516.1	0.199.400	-	-	-
	Tcacao-MatinaG00000000126.1	0.192.600	-	-	-
	Tcacao-MatinaG00000010513.1	4.312.500	-	-	-
	Tcacao-MatinaG00000010518.1	0.0	-	-	-
	HumbraticaG00000019844.3	36.212.600	-	-	-
	HumbraticaG00000014818.1	49.473.400	-	-	-
	HumbraticaG00000014819.1	0.307.300	-	-	-
	TgrandC174G00000010809.1	-	39.029.919	32.167.343	42.282.715
	TgrandC174G00000014593.1	-	18.118.376	18.979.977	16.390.179
	TgrandC174G00000010810.1	-	0.0	0.0	0.0
	TgrandC1074G00000015761.1	-	15.880.945	23.048.134	14.081.443
	TgrandC1074G00000011890.1	-	33.405.003	45.773.148	35.938.641
	TgrandC1074G00000011889.1	-	0.0	0.0	0.0
	TgrandC1074G00000011886.1	-	0.054.246	0.0	0.0
COBL6	*Tcacao-CriolloG00000006609.1*	9.423.900	-	-	-
	Tcacao-MatinaG00000005908.1	9.731.200	-	-	-
	HumbraticaG00000015638.1	2.027.900	-	-	-
	TgrandC174G00000008356.1	-	0.062.911	0.170.018	0.112.788
	TgrandC1074G00000008413.1	-	0.186.137	0.0	0.051375
COBL	***Tcacao-CriolloG00000006352.1***	**2.541.600**	**-**	**-**	**-**

EXCLUSIVO	***Tcacao-MatinaG00000005665.1***	**2.615.600**	**-**	**-**	**-**
	HumbraticaG00000015867.1	**10.368.900**	**-**	**-**	**-**
	TgrandC174G00000008604.1	**-**	**0.815666**	**0.302273**	**0.733600**
	TgrandC1074G00000008646.1	**-**	**0.996881**	**0.710089**	**0.512646**
COBL7,8,9	*Tcacao-CriolloG00000017670.1*	40.657.300	-	-	-
	Tcacao-CriolloG00000004346.1	11.557.000	-	-	-
	Tcacao-MatinaG00000016291.1	26.783.100	-	-	-
	Tcacao-MatinaG00000003724.1	15.400.900	-	-	-
	HumbraticaG00000021654.1	3.488.700	-	-	-
	HumbraticaG00000010520.1	11.365.800	-	-	-
	HumbraticaG00000010515.1	11.773.800	-	-	-
	TgrandC174G00000016764.1	-	115.690.308	55.794.567	39.470.226
	TgrandC174G00000003636.1	-	6.358.520	4.029.745	7.170.976
	TgrandC1074G00000013516.1	-	83.022.560	96.272.377	48.259.354
	TgrandC1074G00000001217.1	-	8.147.701	6.392.522	6.198.557
COBL10, 11	*Tcacao-CriolloG00000025455.2*	39.898.500	-	-	-
	Tcacao-CriolloG00000024510.1	5.084.300	-	-	-
	Tcacao-MatinaG00000023553.2	53.494.500	-	-	-
	Tcacao-MatinaG00000022663.1	6.398.100	-	-	-
	Tcacao-MatinaG00000022662.1	3.191.400	-	-	-
	HumbraticaG00000018922.1	67.516.800	-	-	-
	HumbraticaG00000025413.1	0.635.400	-	-	-
	TgrandC174G00000021411.1	-	0.0	0.0	0.0
	TgrandC1074G00000027067.1	-	0.098740	0.0	0.0
	TgrandC1074G00000021804.1	-	1.244.708	0.710190	1.396.095

O grupo de genes COBL4 apresentou o maior número de representantes, com destaque para *Theobroma cacao* cultivar Matina 1-6, que possui a maior quantidade desses genes. Esses genes estão diretamente associados ao controle da cristalinidade da celulose e à reorganização da parede celular secundária como evidenciado em *A. thaliana* (Schindelman *et al.,* 2001; Brady *et al.,* 2007). No sorgo, o gene BC1, ortólogo do COBL4 em *A. thaliana*, desempenha um papel crucial no controle da resistência mecânica da planta, influenciando diretamente o teor de celulose na parede celular secundária (Li *et al.,* 2019).

A elevada quantidade de genes COBL4 em *T. cacao* cultivar Matina 1-6 pode indicar uma importância na regulação da estrutura e resistência da parede celular, evidenciando a relevância desse grupo de genes em diferentes espécies vegetais (Schindelman *et al.,* 2001; Brady *et al.,* 2007; Li *et al.,* 2019). Além disso, o cultivar Matina 1-6, originária da América Central, geralmente apresenta maior resistência a

doenças, incluindo a vassoura-de-bruxa, em comparação ao Criollo (Motamayor *et al.,* 2008). Essa resistência relativa faz do Matina uma opção mais robusta para os produtores de cacau que enfrentam altos índices de infecção por doenças (Motamayor *et al.*, 2013).

Em contraste, foi observada uma representatividade menor de genes COBL4 em *T. umbraticum* em comparação com *T. grandiflorum* e *T. cacao*, o que sugere que a domesticação pode ter influenciado o aumento na quantidade desses genes em *Theobroma*. Esse padrão indica possíveis adaptações específicas durante o processo de domesticação, contribuindo para aprimorar características como resistência a doenças e outras propriedades desejáveis nas qualidades dos frutos nas espécies domesticadas, como mencionado anteriormente.

Para o grupo COBL7, 8 e 9, *T. umbraticum* apresentou um gene a mais em comparação com as espécies de *Theobroma*. Esses genes regulam sinergicamente o crescimento dos pelos radiculares por meio do controle da deposição de celulose apical em *A. thaliana* (Li *et al.,* 2022). Os tricomas radiculares são geralmente mais desenvolvidos em *Herrania* do que em *Theobroma*, sendo uma adaptação que permite uma melhor absorção de nutrientes e água (Bletter *and* Daly 2006). Nesse contexto, a maior quantidade de genes COBL7, 8 e 9 em *T. umbraticum* pode, talvez, justificar essas características.

3.2. Domínio e estrutura da família gênica COBRA-Likes: conservação evolutiva e variação no tamanho das proteínas

As ferramentas de verificação de domínio revelaram que todos os genes homólogos ao COBRA-Like em *Theobroma grandiflorum* clone C174 e C1074, *Theobroma cacao* cultivar B97-61/B2 - Criollo e Matina 1-6 e *Theobroma umbraticum* apresentam o domínio COBRA (Figura 2C).

Os domínios conservados representam padrões de sequência que se repetem em diferentes proteínas e que são preservados ao longo da evolução devido à sua importância funcional (Zhang *et al.,* 2012). No caso dos genes COBL, o domínio conservado identificado foi o COBRA (PF04833.18), que desempenha um papel importante em plantas, especialmente relacionado à parede celular vegetal (Sangi *et*

al., 2021). A análise revelou que, em média, o domínio COBRA é composto por 160 aminoácidos no clado I e por 177 aminoácidos no clado II, e os tamanhos das proteínas variam de 149 a 687 aminoácidos, respectivamente. Assim, as proteínas do clado II são mais longas, o que corrobora os achados de Sangi *et al.* (2021).

Este domínio COBRA, característico nos membros do Grupo I, apresenta uma proximidade notável com o peptídeo sinal de secreção N-terminal, denotando uma organização estrutural particular. Por outro lado, nos membros do Grupo II, esse domínio funcional se localiza no interior das sequências proteicas, conferindo uma distribuição diferenciada em relação ao Grupo I. Importante ressaltar, contudo, que apesar dessa tendência distintiva, observam-se exceções a essa regra estabelecida, indicando a complexidade e a plasticidade da arquitetura molecular subjacente a esses domínios COBRA (Figura 2C).

A estrutura gênica e os domínios conservados representam mecanismos propulsores no processo evolutivo das famílias gênicas. Para caracterizar estruturalmente os COBLs, foram realizadas análises de estrutura gênica e domínio (Figura 2).

O número total de éxons nos referidos genes apresentou variações entre dois e seis, sendo que tal intervalo demonstrou divergências entre os grupos. Para o grupo COBL 7, 8, 9 todos os genes apresentaram número de éxons uniforme: 2 éxons. No grupo COBL10, 11, houve distribuição similar ao grupo 7, 8, 9, porém dois genes *TgrandC174G00000022328.1* e *HumbraticaG00000025413.1* obtiveram 3 e 5 éxons, respectivamente. No grupo COBL exclusivo, COBL6 e COBL5, 2, 1, com exceção do gene *Tcacao-MatinaG00000010512.1* do grupo COBL5, 2, 1 que apresentou 4 éxon, todos os demais genes apresentaram 6 éxons. Por fim, o grupo COBL4, com a maior quantidade de genes, apresentou quantidades de éxons diversificada entre os genes agrupados nesse conjunto. Entretanto, a frequência de éxons nesse grupo foi predominantemente de 6 éxons (Figura 2).

Os resultados observados anteriormente podem suportar a classificação dos genes COBL em dois clados nitidamente diferentes (Figura 2). O primeiro clado compreende genes com seis éxons, com os seguintes grupos identificados: COBL1, COBL2, COBL4, COBL5, COBL6 e COBL exclusivo. Já o segundo, consiste em genes

que possuem apenas dois éxons, designadamente COBL7, COBL8, COBL9, COBL10, COBL11. Estruturalmente, os COBLs nestes dois grupos exibiram a divergência no N-terminal. Os membros do Grupo Ⅱ continham um trecho extra de aminoácidos após o peptídeo sinal N-terminal e levaram à grande diversidade no tamanho dos aminoácidos variando de 302 a 687 (Tabela 2). O clado I difere do clado II devido aos mais de 196 aminoácidos adicionais na região N-terminais no clado II (Figura 2C). A diferença na estrutura do gene, bem como no número de éxons e íntrons dá suporte à divisão em dois clados (Figura 3). O clado II (COBL10,11; e COBL7,8,9) possuem apenas um íntron. Isso indica que esses grupos evoluíram recentemente com menos íntrons, sugerindo uma diminuição no número de íntrons ao longo do tempo. Pesquisas em milho mostram que o número de íntrons diminuiu ao longo do tempo em um gene semelhante ao COBL, Bk2 (Ching *et al.*, 2006). Além disso, como mencionado anteriormente, eventos significativos de inserção/deleção podem afetar a estrutura exon-íntron.

A análise da estrutura gênica revelou que, apesar das variações no tamnho dos genes e na posição dos éxons, os genes dentro de um mesmo clado apresentam um número conservado de éxons. Isso indica uma forte preservação estrutural entre a maioria dos genes COBL (Figura 2B). A organização e conservação dos éxons reforçam os padrões evolutivos identificados na análise filogenética (Figura 2A).

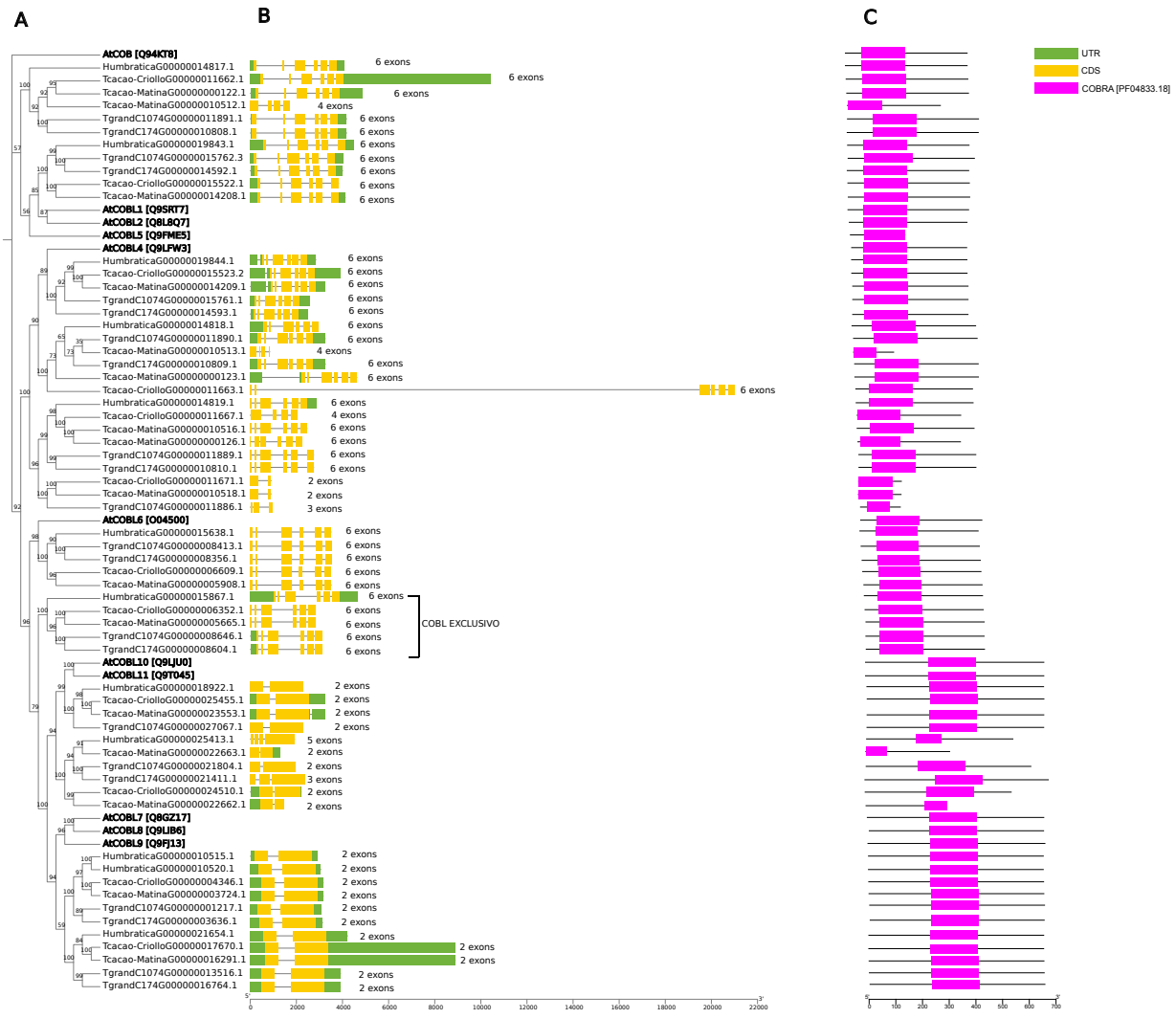

Figura 2. Caracterização da estrutura e domínio dos genes COBRA-Likes. (A) Árvore filogenética construída por método de Máxima Verossimilhança a partir de 1000 árvores de bootstrap (B) estrutura do gene e (C) análise de domínio COBRA conservado

Os genes COBLs deste estudo têm uma estrutura genética comparável com os COBLs de outras espécies (Roudier *et al.,* 2002; Ching *et al.,* 2006; Brady *et al*., 2007; Liu *et al.,* 2013; Niu *et al.,* 2015; Li *et al.,* 2019; Sangi *et al.*, 2021). Indicando que esses genes foram substancialmente conservados.

3.3. Análises filogenéticas indicam que genes COBRA-Like exclusivos de *Theobroma* apresentam padrões estruturais distintos e possíveis adaptações funcionais

A família gênica COBRA-Like, associada à regulação do crescimento celular e ao desenvolvimento de plantas, demonstra uma complexidade intrigante ao ser dividida em duas subfamílias distintas, clado I e II (Figura 3). A primeira subfamília,

composta por COBL1, COBL2, COBL4, COBL5, COBL6, destaca-se por apresentar cinco membros, cada um caracterizado por sua composição de genes contendo seis exons. Essa estrutura genética sugere uma possível correlação funcional entre esses membros, o que justifica a classificação conjunta na subfamília 1. Por outro lado, a subfamília 2, abrangendo COBL7, COBL8, COBL9, COBL10, COBL11, revela uma estrutura genômica mais simplificada, com cada membro composto por apenas dois exons. Essa configuração gênica distinta sugere uma especialização funcional diferente em comparação com a subfamília 1. Em geral, houve alta consistência de sequência de COBLs entre as Malvaceae analisadas e *A. thaliana*, porém algumas exceções foram detectadas (Figura 3). Por exemplo, um clado de COBL inicialmente identificado como exclusivo das Malvaceae, com base na análise filogenética de *Theobroma cacao*, *Theobroma grandiflorum*, *Theobroma umbraticum* e *Arabidopsis thaliana*, sugere uma exclusividade de *Theobroma* amazônicos (Figura 3). No entanto, essa aparente exclusividade foi confirmada somente após a realização de uma análise filogenética ampliada, incluindo espécies de diversas ordens botânicas, o que permitiu verificar que esse conjunto de genes realmente não se agrupa com ortólogos de outras linhagens, reforçando sua exclusividade em Malvaceae (Figura 4).

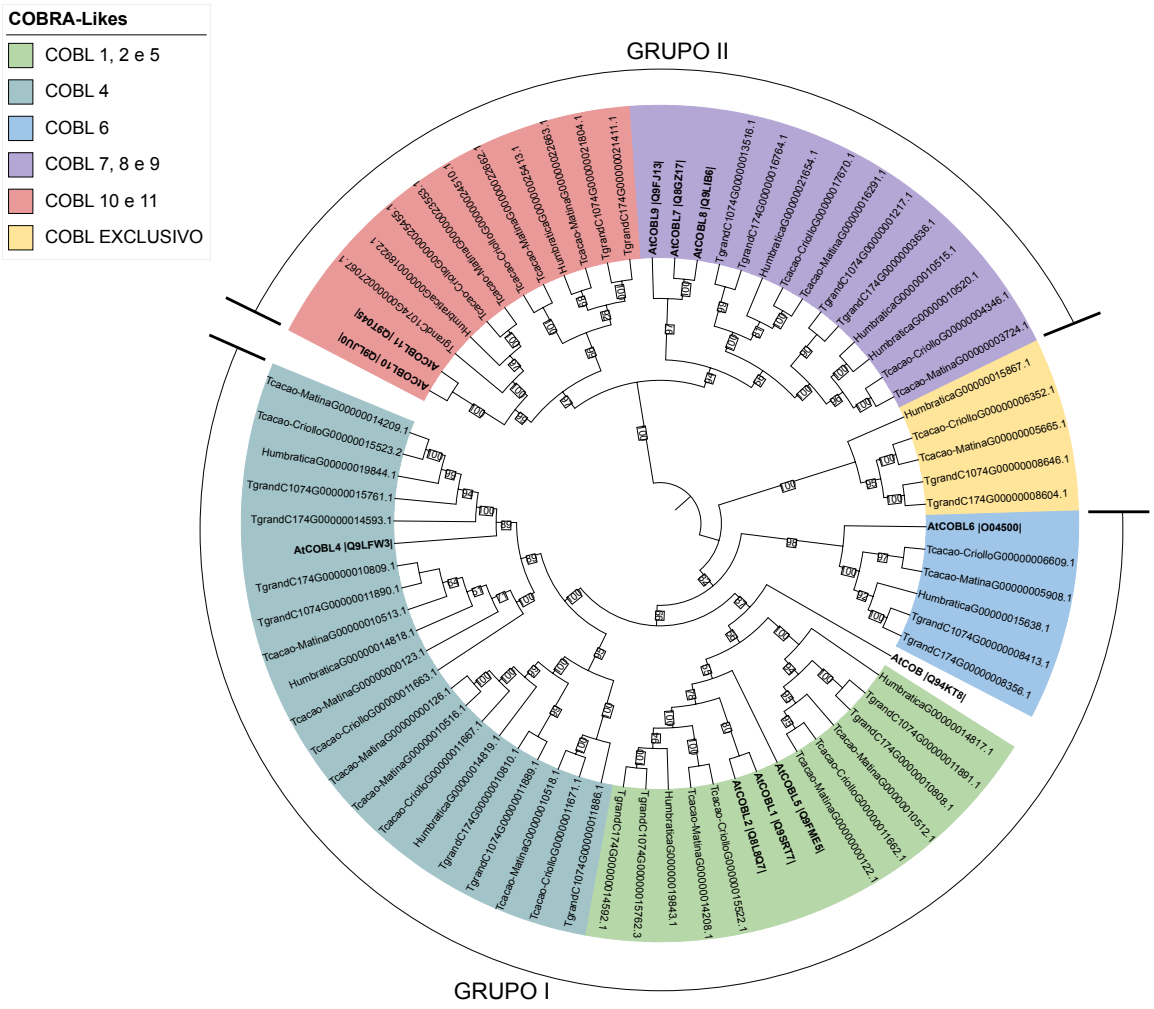

Figura 3. Árvore filogenética da Família COBRA-Likes construída utilizando o método de Máxima Verossimilhança.

Consistentes com os ortólogos em *A. thaliana*, os membros da família COBL foram agrupados em dois clados: clado I e clado II, filogeneticamente relacionados a AtCOBRA e AtCOBL7 em *A. thaliana* (Roudier *et al.,* 2002). A análise filogenética indicou que os COBLs são descendentes de uma antiga duplicação (gamma triplication) ocorrida antes mesmo da separação de monocotiledôneas e *eudicotiledôneas* (Niu *et al.*, 2015). A organização em dois clados e a conservação estrutural dos genes evidenciam suporte aos resultados da análise filogenética.

No que concerne aos COBLs, o clado I apresentou uma contagem intrônica superior àquela observada no clado II. Adicionalmente, a análise de motivos revelou a presença de domínios conservados característicos da família de genes COBL (Figura 2-C), dando suporte a divisão filogenética em dois clados semelhante a *Arabidopsis* (Figura 3). Em *Glycine max*, de acordo com a distribuição filogenética,

as proteínas GmCOBL podem ser classificadas em dois clados semelhantes a *Arabidopsis* , milho, algodão e arroz (Sangi *et al.*, 2021), e também em *Theobroma* e *Herrania.*

Notavelmente, um clado de COBL exclusivo das Malvaceae não se agrupou com nenhum COBL de *A. thaliana* (Figura 3), apesar das semelhanças estruturais com o clado I (Figura 2-B). Essa discrepância pode ser atribuída a um evento de duplicação adicional que ocorreu nas Malvaceae analisadas. Especulamos que tal duplicação gênica pode ter sido fundamental na geração de novidades evolutivas, proporcionando uma maior adaptação a novos ambientes por meio da diversificação funcional desses genes exclusivos das Malvaceae.

Adicionalmente, uma análise filogenética extensiva entre monocotiledôneas e *eudicotiledôneas* revelou a presença de genes COBLs em uma ampla variedade de espécies vegetais de diferentes ordens (Figura 4-A). Notavelmente, um clado de genes COBL exclusivo das Malvaceae deste estudo não se agrupou com nenhum outro COBL das monocotiledôneas e *eudicotiledôneas* analisadas, sendo classificado na subfamília Byttnerioideae da família Malvaceae (Figura 4-B). Além disso, os modelos tridimensionais revelam diferenças estruturais entre as proteínas, destacando as particularidades do COBL exclusivo, que apresenta alterações em domínios funcionais e conformações únicas em relação às proteínas conservadas do seu grupo mais próximo COBL6, sugerindo possíveis adaptações específicas associadas à evolução e à qualidade dos frutos em *Theobroma* e *Herrania* (Figura Suplementar S1).

Os modelos tridimensionais (Figura Suplementar S1) que apesar de possuírem uma alta similaridade por sobreposição das estruturas entre os COBLs exclusivo e COBL6 (Figura Suplementar S2), possuem alterações nos aminoácidos que compõem seu domínio funcional (Figura Suplementar S3).

Os resultados obtidos pelo alinhamento estrutural utilizando o FATCAT indicam que as estruturas de *tgrandc1074g00000008413*.1 e *tgrandc1074g00000008646*.1 (Figura Suplementar S2) apresentam uma semelhança significativa, com um valor de P igual a 0, evidenciando uma similaridade estatisticamente relevante (P-valor < 0,05). O alinhamento resultou em 402 posições equivalentes, com um RMSD de 1,96 Å, sugerindo um alto grau de sobreposição

estrutural entre os dois modelos. Adicionalmente, foi identificado 1 "twist", o que pode refletir uma leve diferença conformacional ou flexibilidade em uma região específica, que pode estar associada a variações funcionais sutis. Esses dados são indicativos de uma possível diferença funcional ou evolutiva entre as proteínas analisadas. Além disso, a construção da árvore filogenética utilizando apenas os aminoácidos que compõem a região do domínio revelou que o grupo COBL exclusivo de *Theobroma* e *Herrania* forma um clado distinto. Essa separação sugere que variações nos resíduos do domínio funcional podem estar associadas a um processo de especialização ou a uma modulação funcional diferenciada. Tais alterações podem influenciar a interação dessas proteínas com componentes da parede celular, impactando características como a firmeza do fruto, o metabolismo de pectinas e hemiceluloses e, possivelmente, associados a aspectos da domesticação destas espécies.

Essas análises podem indicar que os genes COBLs descendem de uma duplicação ancestral anterior à separação entre monocotiledôneas e *eudicotiledôneas*. E que essas exclusividades de um clado COBL exclusivo em *Theobroma* e *Herrania* desse estudo pode indicar possíveis diversificações funcionais.

Esse clado de genes COBL exclusivo apresenta duplicação dispersa no genoma de *T. grandiflorum, T. cacao* e *T. umbraticum*. Essas duplicações dispersas podem ter sido cruciais para a diversificação e especialização desses gêneros, fornecendo material genético adicional para a evolução de características específicas e a adaptação ambiental entre os gêneros *Theobroma* e *Herrania*. Embora o gênero *Theobroma* seja grupo irmão do gênero *Herrania* (Colli-Silva *et al*., 2023a), o que significa que eles divergiram de um ancestral comum, o que resultou no compartilhamento de características genéticas e fenotípicas semelhantes entre eles.

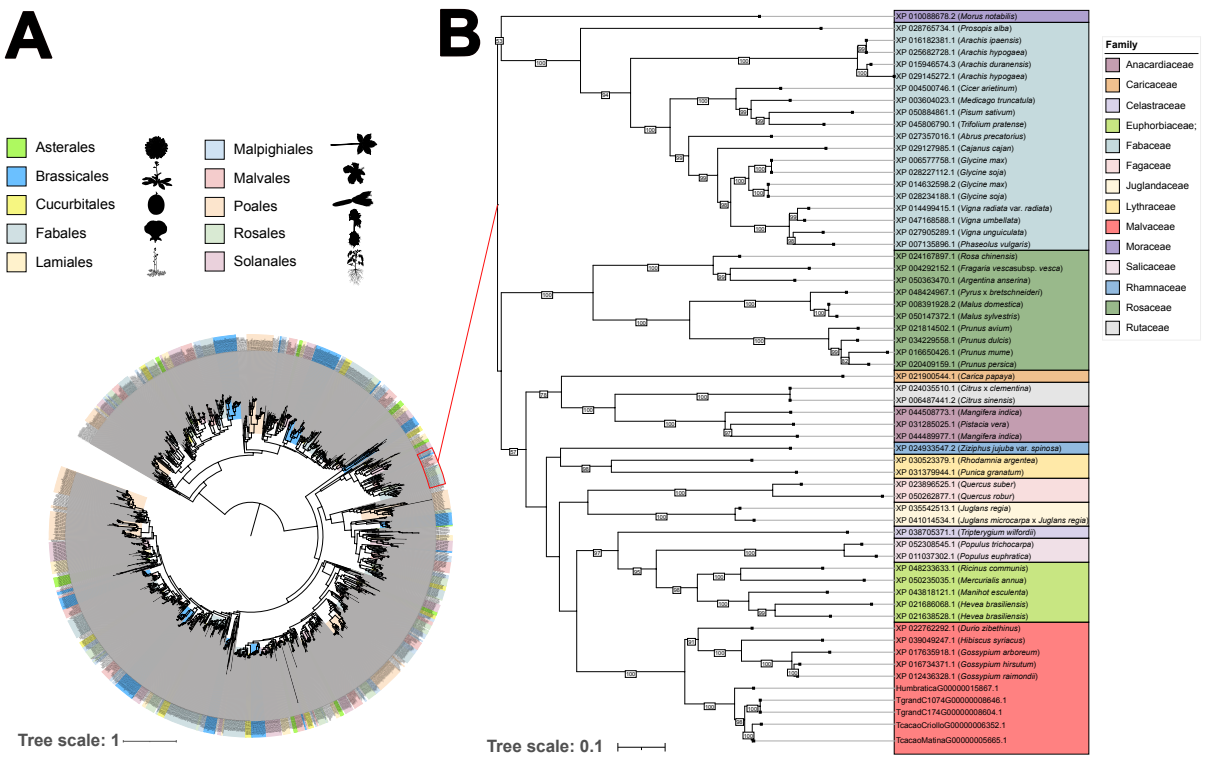

Figura 4. Análise filogenética dos genes COBLs em diferentes espécies das ordens Asterales, Brassicales, Cucurbitales, Fabales, Fagales, Lamiales, Malpighiales, Malvales, Poales, Rosales e Solanales.

Os dados sugerem que as duplicações gênicas observadas em *Theobroma grandiflorum* e *Theobroma cacao* podem estar associadas ao processo de domesticação, possivelmente desempenhando um papel na adaptação, diversificação genética e aprimoramento das qualidades agronômicas dos frutos dessas espécies. Além disso, a identificação de um grupo exclusivo de genes COBL nas Malvaceae analisadas, com uma divergência notável entre monocotiledôneas e *eudicotiledôneas*, pode indicar potenciais diversificações funcionais. Essas observações levantam a hipótese de que tais diversificações poderiam ter contribuído para adaptações evolutivas e diversificação morfológica nas espécies estudadas.

Agradecimentos

Agradecemos à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela bolsa de doutorado, à UNESP, e ao Programa de Pós-Graduação em Agronomia (Genética e Melhoramento de Plantas). Agradecemos também à Embrapa Amazônia Oriental pelo fornecimento das amostras. Este estudo foi financiado pela Fundação de Amparo à Pesquisa do Estado de São Paulo - FAPESP, Processo: 2019/25176-0, e pela Fundação Amazônia de Amparo a Estudos e Pesquisas do Estado do Pará – FAPESPA, processo: 075/2020. Adicionalmente, a extração de DNA HMW e o suporte ao sequenciamento PacBio Sequel IIe foram fornecidos pelo Arizona Genomics Institute (Tucson, AZ – EUA).

Conflito de Interesses:

Os autores declaram que não há conflito de interesses que possa ser percebido como prejudicial à imparcialidade da pesquisa relatada.

Contribuições dos Autores:

RPO formulou o estudo, procedeu à análise dos dados, produção de figuras e tabelas, redigiu o manuscrito e, subsequentemente, efetuou a revisão crítica e edição; JVAA buscas de sequências fastas e GFF3, edição de figuras, auxílio nas ferramentas de bioinformática; AMV curadoria dos dados geral do projeto, análise dos programas, metodologia, administração dos recursos e supervisão do projeto.

Todos os autores leram e aprovaram a versão final.

4. REFERÊNCIAS

Alfonso VA, Paulo S, Baas P, Carlquist S, Chimelo JP, Paulo S, Coradin VTR, Détienne P, Gasson PE, Grosser D *et al*. with an Appendix on non-anatomical information.

Alves RM and Chaves SF da S (2023) Selection of *Theobroma grandiflorum* clones adapted to agroforestry systems using an additive index. Acta Sci, Agron 45:e57519.

Alves RM, de Abreu VAC, Oliveira RP, Almeida JV dos A, de Oliveira M de M, Silva SR, Paschoal AR, de Almeida SS, de Souza PAF, Ferro JA *et al*. (2024) Genomic decoding of *Theobroma grandiflorum* (*cupuaçu*) at chromosomal scale: evolutionary insights for horticultural innovation. GigaScience 13:giae027.

Alves RM, Sebbenn AM, Artero AS, Clement C and Figueira A (2007) High levels of genetic divergence and inbreeding in populations of *cupuaçu* (*Theobroma grandiflorum*). Tree Genetics & Genomes 3:289–298.

Alves RM, Silva CR de S, Albuquerque PSB de and Santos VS dos (2017) Phenotypic and genotypic characterization and compatibility among genotypes to select elite clones of *cupuaçu*. Acta Amaz 47:175–184.

Argout X, Salse J, Aury J-M, Guiltinan MJ, Droc G, Gouzy J, Allegre M, Chaparro C, Legavre T, Maximova SN *et al*. (2011) The genome of *Theobroma cacao*. Nat Genet 43:101–108.

Ben-Tov D, Abraham Y, Stav S, Thompson K, Loraine A, Elbaum R, de Souza A, Pauly M, Kieber JJ and Harpaz-Saad S (2015) COBRA-LIKE2, a member of the glycosylphosphatidylinositol-anchored COBRA-LIKE family, plays a role in cellulose deposition in *Arabidopsis* seed coat mucilage secretory cells. Plant Physiol 167:711–724.

Ben-Tov D, Idan-Molakandov A, Hugger A, Ben-Shlush I, Günl M, Yang B, Usadel B and Harpaz-Saad S (2018) The role of COBRA-LIKE 2 function, as part of the complex network of interacting pathways regulating *Arabidopsis* seed mucilage polysaccharide matrix organization. Plant J 94:497–512.

Bletter N and Daly DC (2006) *Cacao* and Its Relatives in South America: An Overview of Taxonomy, Ecology, Biogeography, Chemistry, and Ethnobotany. In: McNeil (ed) Chocolate in Mesoamerica: A Cultural History of *Cacao*. University Press of Florida, p 0

Brady SM, Song S, Dhugga KS, Rafalski JA and Benfey PN (2007) Combining Expression and Comparative Evolutionary Analysis. The COBRA Gene Family. Plant Physiology 143:172–187.

Brown DM, Zeef LAH, Ellis J, Goodacre R and Turner SR (2005) Identification of Novel Genes in *Arabidopsis* Involved in Secondary Cell Wall Formation Using Expression Profiling and Reverse Genetics. The Plant Cell 17:2281–2295.

Cao Y, Tang X, Giovannoni J, Xiao F and Liu Y (2012) Functional characterization of a tomato COBRA-likegene functioning in fruit development and ripening. BMC Plant Biology 12:211.

Chao J, Li Z, Sun Y, Aluko OO, Wu X, Wang Q and Liu G (2021) MG2C: a user-friendly online tool for drawing genetic maps. Molecular Horticulture 1:16.

Chen C, Wu Y, Li J, Wang X, Zeng Z, Xu J, Liu Y, Feng J, Chen H, He Y *et al*. (2023) TBtools-II: A “one for all, all for one” bioinformatics platform for biological big-data mining. Molecular Plant 16:1733–1742.

Ching A, Dhugga KS, Appenzeller L, Meeley R, Bourett TM, Howard RJ and Rafalski A (2006) Brittle stalk 2 encodes a putative glycosylphosphatidylinositol-anchored protein that affects mechanical strength of maize tissues by altering the composition and structure of secondary cell walls. Planta 224:1174–1184.

Cohen K de O and Jackix M de NH (2005) Estudo do liquor de cupuaçu. Food Sci Technol 25:182–190.

Colli-Silva M, Richardson J and Pirani J (2023a) A taxonomic dataset of preserved specimen occurrences of *Theobroma* and *Herrania* (Malvaceae, Byttnerioideae) stored in 2020. Biodiversity Data Journal 11:e99646.

Colli-Silva M, Richardson JE, Neves EG, Watling J, Figueira A and Pirani JR (2023b) Domestication of the Amazonian fruit tree cupuaçu may have stretched over the past 8000 years. Commun Earth Environ 4:1–8.

Cuatrecasas J (1964) *Cacao* and Its Allies: A Taxonomic Revision of the Genus *Theobroma*. Smithsonian Institution

DAI, Xiaoxia *et al.* Molecular characterization, expression pattern, and function analysis of the OsBC1L family in rice. **Plant Molecular Biology**, [*s. l.*], v. 71, n. 4, p. 469, 2009.

de Abreu VAC, Moysés Alves R, Silva SR, Ferro JA, Domingues DS, Miranda VFO and Varani AM (2023) Análises comparativas dos mitogenomas de *Theobroma cacao* e *T. grandiflorum* revelam conteúdo gênico conservado incorporado em estruturas complexas e plásticas. Gene 849:146904.

Ge S, Sun P, Wu W, Chen X, Wang Y, Zhang M, Huang J and Liang Y-K (2024) COBL7 is required for stomatal formation via regulation of cellulose deposition in *Arabidopsis* . New Phytologist 241:227–242.

Genovese M and Lannes S (2009) Comparison of total phenolic content and antiradical capacity of powders and “chocolates” from cocoa and *cupuaçu*. Ciencia E Tecnologia De Alimentos - CIENCIA TECNOL ALIMENT. doi: 10.1590/S0101-20612009000400017

Hanada K, Zou C, Lehti-Shiu MD, Shinozaki K and Shiu S-H (2008) Importance of Lineage-Specific Expansion of Plant Tandem Duplicates in the Adaptive Response to Environmental Stimuli. Plant Physiology 148:993–1003.

HOCHHOLDINGER, Frank *et al.* The maize (Zea mays L.) roothairless3 gene encodes a putative GPI-anchored, monocot-specific, COBRA-like protein that significantly affects grain yield. **The Plant Journal**, [*s. l.*], v. 54, n. 5, p. 888–898, 2008.

Hoang DT, Chernomor O, von Haeseler A, Minh BQ and Vinh LS (2018) UFBoot2: Improving the Ultrafast Bootstrap Approximation. Molecular Biology and Evolution 35:518–522.

Hok S, Danchin EGJ, Allasia V, Panabières F, Attard A and Keller H (2011) An *Arabidopsis* (malectin-like) leucine-rich repeat receptor-like kinase contributes to downy mildew disease. Plant, Cell & Environment 34:1944–1957.

Idam (2021). Boas práticas de produção sustentável no amazonas – cupuaçu. Disponível em: http://www.idam.am.gov.br/wp-content/uploads/2021/01/Cupua%C3%A7u.pdf.

IAWA Committee. (1989). List of microscope features for hardwood identification. IAWA Bulletin New Series, 10, 1–332.

Jiao Y, Leebens-Mack J, Ayyampalayam S, Bowers JE, McKain MR, McNeal J, Rolf M, Ruzicka DR, Wafula E, Wickett NJ *et al*. (2012) A genome triplication associated with early diversification of the core eudicots. Genome Biology 13:R3.

JONES, Mark A.; RAYMOND, Marjorie J.; SMIRNOFF, Nicholas. Analysis of the root-hair morphogenesis transcriptome reveals the molecular identity of six genes with roles in root-hair development in *Arabidopsis* . **The Plant Journal**, [*s. l.*], v. 45, n. 1, p. 83–100, 2006.

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A and Jermiin LS (2017) ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods 14:587–589.

KO, Jae-Heung *et al.* Loss of function of COBRA, a determinant of oriented cell expansion, invokes cellular defence responses in *Arabidopsis thaliana*. **Journal of Experimental Botany**, [*s. l.*], v. 57, n. 12, p. 2923–2936, 2006.

Katoh K and Standley DM (2013) MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Molecular Biology and Evolution 30:772–780.

Ko J-H, Kim JH, Jayanty SS, Howe GA and Han K-H (2006) Loss of function of COBRA, a determinant of oriented cell expansion, invokes cellular defence responses in *Arabidopsis thaliana*. Journal of Experimental Botany 57:2923–2936.

Kumar S, Suleski M, Craig JM, Kasprowicz AE, Sanderford M, Li M, Stecher G and Hedges SB (2022) TimeTree 5: An Expanded Resource for Species Divergence Times. Molecular Biology and Evolution 39:msac174.

Letunic I and Bork P (2024) Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. Nucleic Acids Research gkae268.

LI, Yunhai *et al.* BRITTLE CULM1, Which Encodes a COBRA-Like Protein, Affects the Mechanical Properties of Rice Plants. **The Plant Cell**, [*s. l.*], v. 15, n. 9, p. 2020–2031, 2003.

Li H and Durbin R (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25:1754–1760.

Li H, Yang Y, Zhang H, Li C, Du P, Bi M, Chen T, Qian D, Niu Y, Ren H *et al*. (2023) The *Arabidopsis* GPI-anchored protein COBL11 is necessary for regulating pollen tube integrity. Cell Reports 42:113353.

Li P, Liu Y, Tan W, Chen J, Zhu M, Lv Y, Liu Y, Yu S, Zhang W and Cai H (2019) Brittle Culm 1 Encodes a COBRA-Like Protein Involved in Secondary Cell Wall Cellulose Biosynthesis in Sorghum. Plant and Cell Physiology 60:788–801.

Li S, Ge F-R, Xu M, Zhao X-Y, Huang G-Q, Zhou L-Z, Wang J-G, Kombrink A, McCormick S, Zhang XS *et al*. (2013) *Arabidopsis* COBRA-LIKE 10, a GPI-anchored protein, mediates directional growth of pollen tubes. The Plant Journal 74:486–497.

Li Z, Zhou T, Sun P, Chen X, Gong L, Sun P, Ge S and Liang Y-K (2022) *COBL9* and *COBL7* synergistically regulate root hair tip growth via controlling apical cellulose deposition. Biochemical and Biophysical Research Communications 596:6–13.

Lucht JM, Mauch-Mani B, Steiner H-Y, Metraux J-P, Ryals J and Hohn B (2002) Pathogen stress increases somatic recombination frequency in *Arabidopsis* . Nat Genet 30:311–314.

McNeil CL (2006) Chocolate in Mesoamerica: a cultural history of *cacao*. University Press of Florida, Gainesville

Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ *et al*. (2021) Pfam: The protein families database in 2021. Nucleic Acids Research 49:D412–D419.

Motamayor JC, Lachenaud P, Mota JW da S e, Loor R, Kuhn DN, Brown JS and Schnell RJ (2008) Geographic and Genetic Population Differentiation of the Amazonian Chocolate Tree (*Theobroma cacao* L). PLOS ONE 3:e3311.

Motamayor JC, Mockaitis K, Schmutz J, Haiminen N, III DL, Cornejo O, Findley SD, Zheng P, Utro F, Royaert S *et al*. (2013) The genome sequence of the most widely cultivated *cacao* type and its use to identify candidate genes regulating pod color. Genome Biology 14:r53.

Mournet P, de Albuquerque PSB, Alves RM, Silva-Werneck JO, Rivallan R, Marcellino LH and Clément D (2020) A reference high-density genetic map of *Theobroma grandiflorum* (Willd. ex Spreng) and QTL detection for resistance to witches' broom disease (Moniliophthora perniciosa). Tree Genetics & Genomes 16:89.

Nguyen L-T, Schmidt HA, von Haeseler A and Minh BQ (2015) IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Molecular Biology and Evolution 32:268–274.

Niu E, Fang S, Shang X and Guo W (2018) Ectopic expression of GhCOBL9A, a cotton glycosyl-phosphatidyl inositol-anchored protein encoding gene, promotes cell elongation, thickening and increased plant biomass in transgenic *Arabidopsis* . Mol Genet Genomics 293:1191–1204.

Niu E, Shang X, Cheng C, Bao J, Zeng Y, Cai C, Du X and Guo W (2015) Comprehensive Analysis of the COBRA-Like (COBL) Gene Family in Gossypium Identifies Two COBLs Potentially Associated with Fiber Quality. PLOS ONE 10:e0145725.

Niu Y-F, Ni S-B and Liu J (2019) The complete chloroplast genome of *Theobroma grandiflorum*, an important tropical crop. Mitochondrial DNA Part B 4:4157–4158.

Potter SC, Luciani A, Eddy SR, Park Y, Lopez R and Finn RD (2018) HMMER web server: 2018 update. Nucleic Acids Research 46:W200–W204.

Price MN, Dehal PS and Arkin AP (2010) FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. PLOS ONE 5:e9490.

QUEVILLON, E. *et al.* InterProScan: protein domains identifier. **Nucleic Acids Research**, [*s. l.*], v. 33, n. Web Server issue, p. W116-120, 2005.

Reis ARS and da Cruz CF (2021) Biometry and verification of papermaking potential of secondary xylem fibers of two *Theobroma* species. Environ Dev Sustain 23:11090–11097.

Richardson JE, Whitlock BA, Meerow AW and Madriñán S (2015) The age of chocolate: a diversification history of *Theobroma* and Malvaceae. Front Ecol Evol. doi: 10.3389/fevo.2015.00120

Roudier F, Schindelman G, DeSalle R and Benfey PN (2002) The COBRA Family of Putative GPI-Anchored Proteins in *Arabidopsis* . A New Fellowship in Expansion. Plant Physiology 130:538–548.

Sajjad M, Ahmad A, Riaz MW, Hussain Q, Yasir M and Lu M-Z (2023) Recent genome resequencing paraded COBRA-Like gene family roles in abiotic stress and wood formation in Poplar. Front Plant Sci. doi: 10.3389/fpls.2023.1242836

Sangi S, Araújo PM, Coelho FS, Gazara RK, Almeida-Silva F, Venancio TM and Grativol C (2021) Genome-Wide Analysis of the COBRA-Like Gene Family Supports Gene Expansion through Whole-Genome Duplication in Soybean (Glycine max). Plants (Basel) 10:167.

Schaaf G, Honsbein A, Meda AR, Kirchner S, Wipf D and von Wirén N (2006) *AtIREG2* Encodes a Tonoplast Transport Protein Involved in Iron-dependent Nickel Detoxification in *Arabidopsis thaliana* Roots*. Journal of Biological Chemistry 281:25532–25540.

Schindelman G, Morikami A, Jung J, Baskin TI, Carpita NC, Derbyshire P, McCann MC and Benfey PN (2001) COBRA encodes a putative GPI-anchored protein, which is polarly localized and necessary for oriented cell expansion in *Arabidopsis* . Genes Dev 15:1115–1127.

Silva da Costa R, Pinheiro WB de S, Arruda MSP, Costa CEF, Converti A, Ribeiro Costa RM and Silva Júnior JOC (2022) Thermoanalytical and phytochemical study of the *cupuaçu* (*Theobroma grandiflorum* Schum.) seed by-product in different processing stages. J Therm Anal Calorim 147:275–284.

Sorek N, Sorek H, Kijac A, Szemenyei HJ, Bauer S, Hématy K, Wemmer DE and Somerville CR (2014) The *Arabidopsis* COBRA Protein Facilitates Cellulose Crystallization at the Plasma Membrane*. Journal of Biological Chemistry 289:34911–34920.

Sousa Silva CR and Figueira A (2005) Phylogenetic analysis of *Theobroma* (Sterculiaceae) based on Kunitz-like trypsin inhibitor sequences. Plant Syst Evol 250:93–104.

The Angiosperm Phylogeny Group, Chase MW, Christenhusz MJM, Fay MF, Byng JW, Judd WS, Soltis DE, Mabberley DJ, Sennikov AN, Soltis PS *et al*. (2016) An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. Botanical Journal of the Linnean Society 181:1–20.

The UniProt Consortium (2023) UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Research 51:D523–D531.

Trifinopoulos J, Nguyen L-T, von Haeseler A and Minh BQ (2016) W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. Nucleic Acids Research 44:W232–W235.

Wang J, Tao F, Marowsky NC and Fan C (2016) Evolutionary Fates and Dynamic Functionalization of Young Duplicate Genes in *Arabidopsis* Genomes. Plant Physiology 172:427–440.

Wang Y, Tang H, DeBarry JD, Tan X, Li J, Wang X, Lee T, Jin H, Marler B, Guo H *et al*. (2012) MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Research 40:e49.

Yang Q, Wang S, Chen H, You L, Liu F and Liu Z (2021) Genome-wide identification and expression profiling of the COBRA-like genes reveal likely roles in stem strength in rapeseed (Brassica napus L.). PLOS ONE 16:e0260268.

Ye X, Kang B, Osburn LD and Cheng Z-M (2009) The COBRA gene family in Populus and gene expression in vegetative organs and in response to hormones and environmental stresses. Plant Growth Regul 58:211–223.

Zhang X-C, Wang Z, Zhang X, Le MH, Sun J, Xu D, Cheng J and Stacey G (2012) Evolutionary dynamics of protein domain architecture in plants. BMC Evolutionary Biology 12:6.

Zou C, Lehti-Shiu MD, Thomashow M and Shiu S-H (2009) Evolution of Stress-Regulated Gene Expression in Duplicate Genes of *Arabidopsis thaliana*. PLOS Genetics 5:e1000581.

Recursos da Internet

Plant Genomics DB, https://plantgenomics.ncc.unesp.br/ (26 de janeiro de 2024)

HmmerWeb versão 2.41.2, https://www.ebi.ac.uk/Tools/hmmer/search/hmmscan (26 de janeiro de 2024)

iTol, https://itol.embl.de/# (26 de janeiro de 2024)

MG2C (V2.1), http://mg2c.iask.in/mg2c%5Fv2.1/ (26 de janeiro de 2024)

NCBI - PRJNA691024, https://www.ncbi.nlm.nih.gov/bioproject/691024 (26 de janeiro de 2024)

extract_kraken_reads.py, https://github.com/jenniferlu717/KrakenTools (26 de janeiro de 2024)

PlusPFP, https://benlangmead.github.io/aws-indexes/k2 (26 de janeiro de 2024)

UniProt, https://www.uniprot.org/ (26 de janeiro de 2024)

BLAST, https://plantgenomics.ncc.unesp.br/Genomes/*Theobroma*DB/BLAST/BLAST.html (26 de janeiro de 2024)

CAPÍTULO 3 – Considerações finais e perspectivas

O estudo comparativo dos genes COBRA-LIKE (COBL) em *Theobroma grandiflorum, T. cacao* e *Theobroma umbraticum* forneceu evidências valiosas sobre a filogenia, a estrutura e as possíveis funções desses genes em espécies amazônicas de Malvaceae. Ao longo desta pesquisa, foi possível identificar, pela primeira vez, genes COBLs exclusivos nessas espécies, bem como eventos de duplicação que moldaram sua diversidade genômica. Tais descobertas sugerem adaptações evolutivas potencialmente ligadas a características agronômicas, como resistência a doenças e qualidade dos frutos — tópico de grande relevância, dada a escassez de estudos sobre genes associados à qualidade do fruto do cupuaçu.

No caso específico do cupuaçu, a análise detalhada desses genes oferece avanços significativos para entender os mecanismos moleculares envolvidos na resistência a doenças como vassoura-de-bruxa e monilíase, além de subsidiar programas de melhoramento genético. Padrões estruturais distintos observados nos genes COBRA-LIKE de *Theobroma*, em comparação a outras espécies, reforçam a hipótese de adaptações funcionais relacionadas à resposta a estresses patogênicos. Essa divergência representa uma oportunidade para o desenvolvimento de estratégias de seleção genética que fortaleçam a resistência do cupuaçu.

Contudo, elucidar as funções exatas dos genes COBRA-LIKE recém-identificados requer experimentos de expressão gênica e manipulação genética que confirmem seu papel nos mecanismos de defesa. Tais estudos futuros podem não apenas elevar a produtividade e o valor comercial do cupuaçu, mas também contribuir para a conservação dessa espécie, um recurso valioso na bioeconomia amazônica, dado seu uso nos mercados de alimentos e cosméticos.

O presente trabalho abre novas perspectivas para o aumento da produtividade do cupuaçu, ampliando seu potencial econômico e promovendo sua preservação. Além disso, reforça a importância da genômica comparativa como ferramenta essencial para o uso sustentável da biodiversidade amazônica. Investigações futuras sobre a caracterização funcional desses genes, abrangendo seus papéis na resistência a patógenos, na qualidade dos frutos e na regulação da biossíntese

celular, poderão revelar mecanismos-chave para melhorar variedades de cupuaçu e contribuir para a preservação de espécies nativas.

Recomenda-se também a realização de análises anatômicas quantitativas das fibras para investigar, de modo mais detalhado, a ação das proteínas COBLs na deposição de celulose, aspecto fundamental para a estrutura e a expansão celular. Somado a isso, a ampliação do sequenciamento genômico e transcriptômico, bem como o uso de ferramentas de edição genética em genótipos de cupuaçu e espécies correlatas, tem potencial para acelerar o melhoramento genético e consolidar o cupuaçu como uma cultura estratégica na bioeconomia da Amazônia.

APÊNDICES

Figura Suplementar S1. Modelagem molecular das proteínas COBLs dos grupos COB6, COB7,8,9 e COBL10,11, comparadas ao COBL exclusivo identificado neste estudo em espécies de Malvaceae amazônicas. Gráfico de Ramachandran gerado pelo programa PROCHECK para as proteínas. O gráfico apresenta a distribuição dos resíduos da proteína nas diferentes regiões conformacionais baseadas nos ângulos φ e ψ: regiões mais favorecidas (vermelho), adicionalmente permitidas (amarelo), generosamente permitidas (amarelo claro) e não permitidas (branco). Essas regiões refletem a estabilidade e a viabilidade energética das conformações peptídicas.

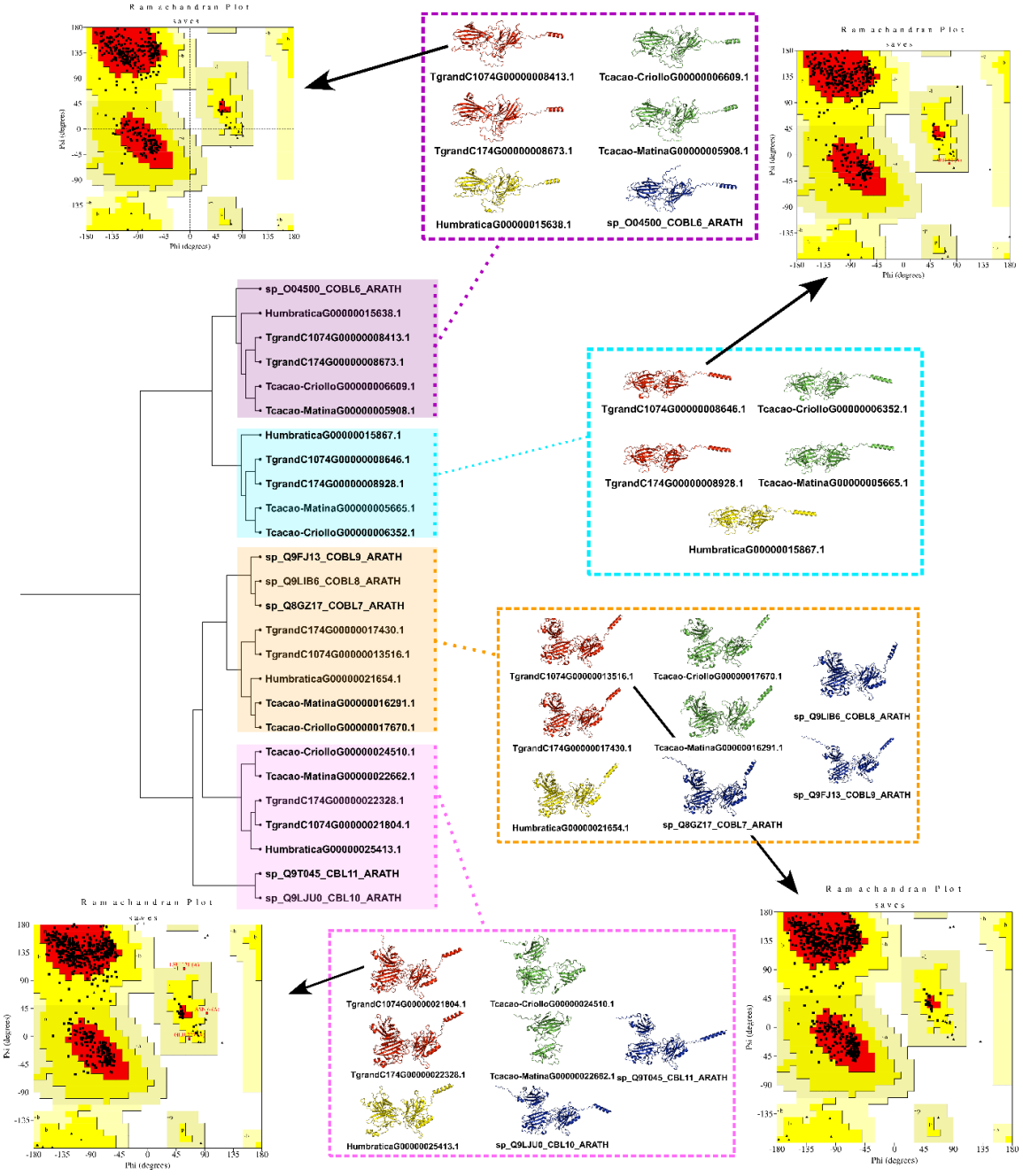

Tabela Suplementar S1. Avaliação da qualidade estrutural das proteínas do grupo COBL6, COBL7,8,9, COBL10,11 de *T. grandiflorum, T. cacao e T. umbraticum*. A tabela apresenta os valores de ERRAT (Overall Quality Factor), que indica a qualidade geral do modelo estrutural, e os resultados do PROCHECK (Ramachandran plot, %), que mostram a distribuição dos resíduos nas regiões conformacionalmente permitidas: vermelho (mais favorecida), amarelo (adicionalmente permitida), amarelo claro (generosamente permitida) e branco (não permitida).

Grupo ARATH	ID	ERRAT	PROCHECK (Ramachandran plot, %)			
		Fator de Qualidade Geral	vermelho	amarelo	amerelo claro	branco
COBL6	TgrandC1074G00000008413.1	93.0748	89.3	10.7	0	0
	TgrandC174G00000008673.1	91.5068	87.9	11.9	0.3	0
	Tcacao-CriolloG00000006609.1	89.7297	87.7	12.3	0	0
	Tcacao-MatinaG00000005908.1	92.3706	87.2	12.8	0	0
	HumbraticaG00000015638.1	89.3048	85.8	13.6	0.6	0
	TgrandC1074G00000008646.1	93.7158	88.5	11.3	0.3	0
	TgrandC174G00000008928.1	93.9560	87.9	12.1	0	0
	Tcacao-CriolloG00000006352.1	93.1429	86.3	13.4	0.3	0
	Tcacao-MatinaG00000005665.1	93.6986	87.7	12.0	0.3	0
	HumbraticaG00000015867.1	94.6524	89.6	9.9	0.5	0
COBL7, 8 e 9	TgrandC1074G00000013516.1	93.0973	92.8	7.2	0	0
	TgrandC174G00000017430.1	92.2942	91.5	8.5	0	0
	Tcacao-CriolloG00000017670.1	92.9825	92.2	7.6	0.2	0
	Tcacao-MatinaG00000016291.1	88.9474	91.8	8.2	0.	0
	HumbraticaG00000021654.1	91.5641	91.5	8.5	0.	0
COBL10 e 11	TgrandC1074G00000021804.1	87.3122	90.3	9.1	0.6	0
	TgrandC174G00000022328.1	88.5522	90.5	9.1	0.4	0
	Tcacao-CriolloG00000024510.1	88.5776	87.3	12.7	0.	0
	Tcacao-MatinaG00000022662.1	93.8525	89.3	10.3	0.4	0
	HumbraticaG00000025413.1	83.6066	91.1	8.7	0.2	0

Figura Suplementar S2. Alinhamento estrutural das proteínas *T. grandiflorum* - c1074g000000008413.1 e *T. grandiflorum* - c1074g00000008646.1 utilizando o FATCAT. As análises indicam uma similaridade estrutural significativa entre as proteínas, com valor de P igual a 0, confirmando relevância estatística (P-valor < 0,05).

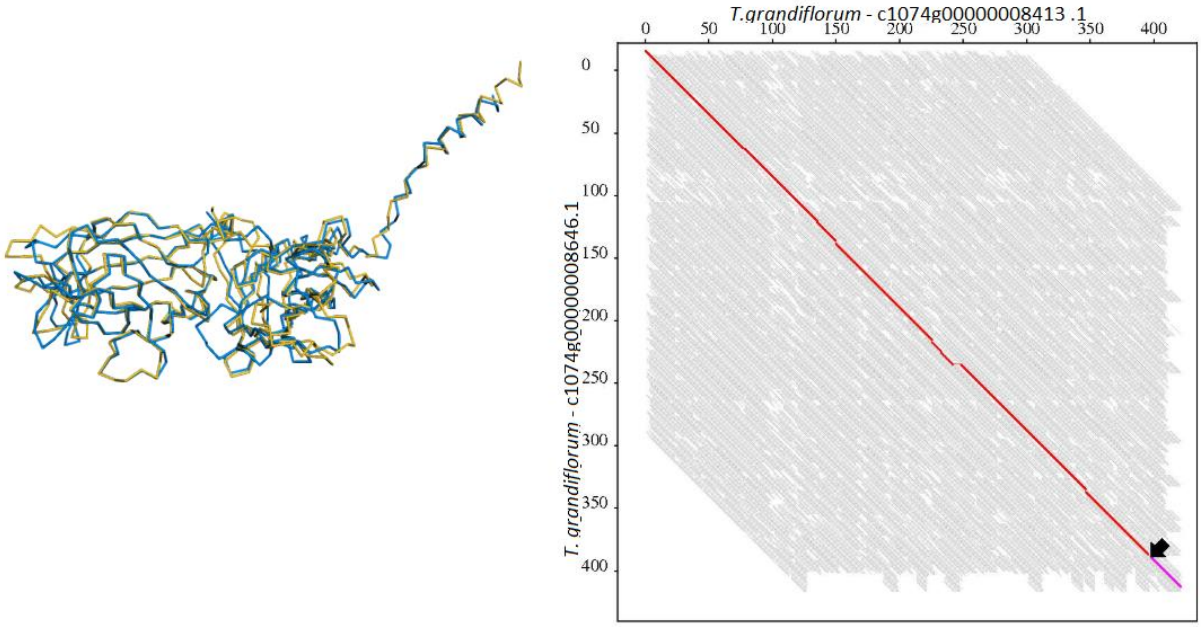

Figura Suplementar S3. Alinhamento de sequência de aminoácidos do domínio COBRA nas proteínas COBL6 (retângulo azul) e COBL exclusivo (retângulo vermelho). A análise evidencia alterações específicas nos aminoácidos que compõem o domínio funcional, destacando variações estruturais que podem estar associadas a funções adaptativas nas Malvaceae amazônicas.

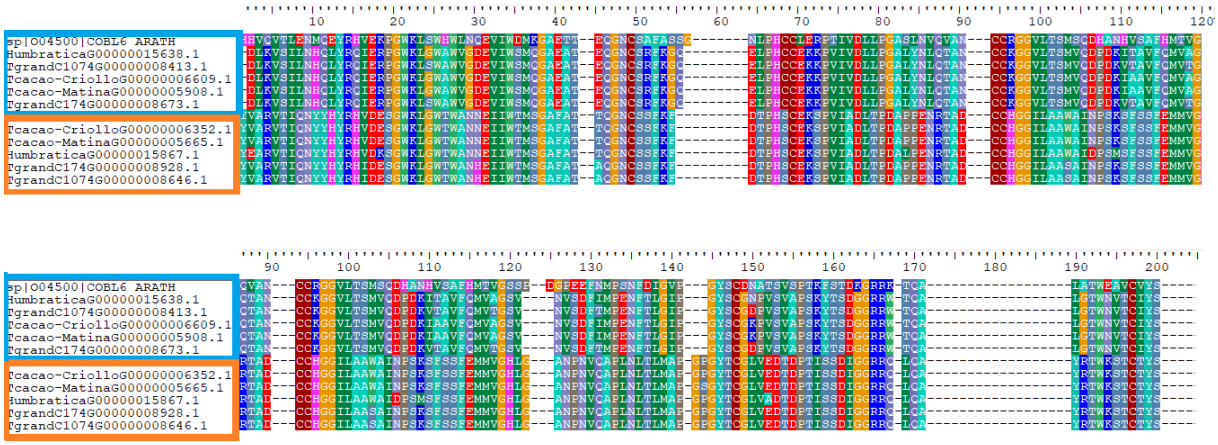

Figura Suplementar S4. Alinhamento de sequência de aminoácidos do domínio COBRA nas proteínas COBL1, COBL2, COBL3, COBL4, COBL5, COBL6, COBL7, COBL8, COBL9, COBL10, COBL11, COBL12 e COBL exclusivo. A análise evidencia alterações específicas nos aminoácidos que compõem o domínio funcional, destacando variações estruturais que podem estar associadas a funções adaptativas nas Malvaceae amazônicas.

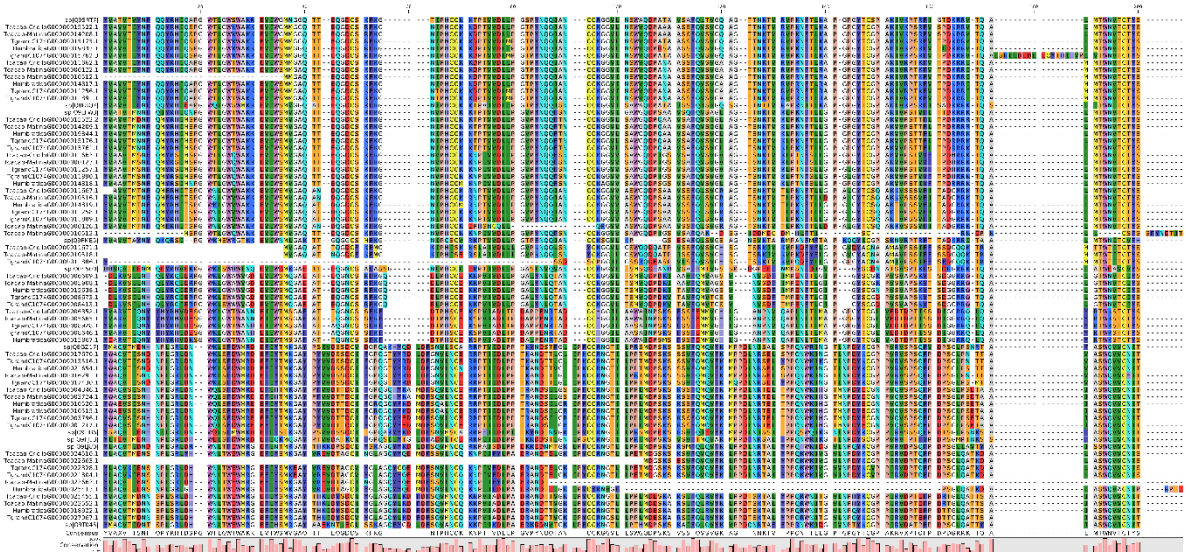

Figura Suplementar S4. Script específico disponível em https://www.zybuluo.com/lakesea/note/1743213, utilizado para identificar a maior isoforma de cada gene nos genomas analisados das espécie da Tabela suplementar S2.

@lakesea 2020-10-29 12:46 字数 2609 阅读 1458

How to extract the longest transcript/protein from gff3

how_to

Use this script:

```
###get_the_longest_transcripts.py
#!/usr/bin/env python3

# v0.1:fixing the problem that the feature of exon/mRNA is before gene
# v0.2: Use CDS to decide which transcript is primary
# v0.3: Fix the bug that if the CDS_parent include '\r' , the lst_len will be 0

# GFF must have CDS feature
# GFF must have ID and Parent in column 9

import re
import fileinput
from collections import defaultdict

gene_dict = defaultdict(list)
tx_pos_dict = defaultdict(list)
CDS_dict = defaultdict(list)

for line in fileinput.input():
    if line.startswith("#"):
        continue
    content = line.split("\t")
    if len(content) <= 8:
        continue
    #print(content)
    '''
    if content[2] == 'gene':
        gene_id = re.search(r'ID=(.*?)[;\n]', content[8]).group(1)
        #print(gene_id)
        if gene_id not in gene_dict:
            gene_dict[gene_id] = []
    '''
    if content[2] == "transcript" or content[2] == "mRNA":

        # use regual expression to extract the gene ID
        # match the pattern ID=xxxxx; or ID=xxxxx

        tx_id = re.search(r'ID=(.*?)[;\n]',content[8]).group(1)
        tx_parent = re.search(r'Parent=(.*?)[;\n]',content[8]).group(1)
        tx_parent = tx_parent.strip() # remove the 'r' and '\n'

        # if the parent of transcript is not in the gene_dict, create it rather than append
        if tx_parent in gene_dict:
            gene_dict[tx_parent].append(tx_id)
        else:
            gene_dict[tx_parent] = [tx_id]
        tx_pos_dict[tx_id] = [content[0],content[3], content[4], content[6] ]
    # GFF must have CDS feature
    if content[2] == 'CDS':
        width = int(content[4]) - int(content[3])
        CDS_parent = re.search(r'Parent=(.*?)[;\n]',content[8]).group(1)
        CDS_parent = CDS_parent.strip() # strip the '\r' and '\n'
        CDS_dict[CDS_parent].append(width)

for gene, txs in gene_dict.items():
    tmp = 0
    for tx in txs:
        tx_len = sum(CDS_dict[tx])
        if tx_len > tmp:
            lst_tx = tx
            tmp = tx_len
    tx_chrom = tx_pos_dict[lst_tx][0]
    tx_start = tx_pos_dict[lst_tx][1]
    tx_end   = tx_pos_dict[lst_tx][2]
    tx_strand = tx_pos_dict[lst_tx][3]
    print("{gene}\t{tx}\t{chrom}\t{start}\t{end}\t{strand}".format(gene=gene,tx=lst_tx,chrom=tx_chro
m,start=tx_start,end=tx_end, strand=tx_strand))
```

get the longest transcript ID:

```
zcat GWHACEE00000000.gff.gz|python3 /scratch/pawsey0149/hhu/train/GeneFamily20200803/script/get_longe
st_transcript.py|cut -f2
```

Transfer the protein id into original one:

```
for i in `cat file.list`;do cut -f 19 ${i}.feature |grep -v ID > ${i}.rename.list;done

for i in `cat file.list`;do awk 'NR==FNR{names[NR]=$0; next} /^>/{$1=">"names[++c]}1' ${i}.rename.lis
t ${i}.Protein.faa > newname_${i}.Protein.fa;done

for i in `cat file.list`;do ~/biosoft/seqtk/seqtk subseq newname_${i}.Protein.fa ${i}.longest_transcr
ipt.id.list >${i}.longest_pro.fa;done
```

blast to find the oil gene:

```
blastp -db AT.oil.pro.fa -max_target_seqs 2 -outfmt "6 qseqid sseqid pident length qlen slen mismatch
gapopen gaps qstart qend sstart send stitle evalue bitscore qcovs" -num_threads 24 -evalue 1e-10 -que
ry Wm82.longest_pro.fa -out Wm82_vs_AToil.blastp.out
```

Tabela Suplementar S2. Espécies utilizadas na análise filogenética dos diferentes grupos com base em sequências cobra (PF04833.1).

ESPÉCIE	**GRUPO**	**IDS**	**Número Sequência COBRA (**PF04833.1**)**
Arabidopsis thaliana	eudicotiledôneas	GCF_000001735.4	12
Physcomitrium patens	musgos	GCF_000002425.4	11
Populus trichocarpa	eudicotiledôneas	GCF_000002775.5	14
Sorghum bicolor	monocotiledôneas	GCF_000003195.3	10
Vitis vinífera	eudicotiledôneas	GCF_000003745.3	11
Cucumis sativus	eudicotiledôneas	GCF_000004075.3	9
Arabidopsis lyrata subsp. lyrata	eudicotiledôneas	GCF_000004255.2	13
Glycine max	eudicotiledôneas	GCF_000004515.6	21
Brachypodium distachyon	monocotiledôneas	GCF_000005505.3	10
Selaginella moellendorffii	musgo	GCF_000143415.4	8
Oryza glaberrima	monocotiledôneas	GCF_000147395.1	11
Carica papaya	eudicotiledôneas	GCF_000150535.2	9
Fragaria vesca subsp. vesca	eudicotiledôneas	GCF_000184155.1	13
Solanum lycopersicum	eudicotiledôneas	GCF_000188115.5	15
Theobroma cacao	eudicotiledôneas	GCF_000208745.1	10
Solanum tuberosum	eudicotiledôneas	GCF_000226075.1	18
Oryza brachyantha	monocotiledôneas	GCF_000231095.2	10

Setaria italica	monocotiledôneas	GCF_000263155.2	9
Brassica rapa	eudicotiledôneas	GCF_000309985.2	19
Musa acuminata subsp. malaccensis	monocotiledôneas	GCF_000313855.2	11
Cicer arietinum	eudicotiledôneas	GCF_000331145.1	12
Cajanus cajan	eudicotiledôneas	GCF_000340665.2	12
Prunus persica	eudicotiledôneas	GCF_000346465.2	11
Prunus mume	eudicotiledôneas	GCF_000346735.1	11
Nelumbo nucifera	eudicotiledôneas	GCF_000365185.1	9
Capsella rubella	eudicotiledôneas	GCF_000375325.1	12
Nicotiana tomentosiformis	eudicotiledôneas	GCF_000390325.2	16
Nicotiana sylvestris	eudicotiledôneas	GCF_000393655.1	17
Morus notabilis	eudicotiledôneas	GCF_000414095.1	11
Elaeis guineensis	monocotiledôneas	GCF_000442705.1	10
Tarenaya hassleriana	eudicotiledôneas	GCF_000463585.1	15
Amborella trichopoda	eudicotiledôneas	GCF_000471905.2	7
Eutrema salsugineum	eudicotiledôneas	GCF_000478725.1	13
Citrus x clementina	eudicotiledôneas	GCF_000493195.1	10
Populus euphratica	eudicotiledôneas	GCF_000495115.1	14
Phaseolus vulgaris	eudicotiledôneas	GCF_000499845.1	14
Erythranthe guttata	eudicotiledôneas	GCF_000504015.1	12
Sesamum indicum	eudicotiledôneas	GCF_000512975.1	12
Gossypium arboreum	eudicotiledôneas	GCF_000612285.1	18
Camelina sativa	eudicotiledôneas	GCF_000633955.1	37
Brassica oleracea var. oleracea	eudicotiledôneas	GCF_000695525.1	18

Nicotiana tabacum	eudicotiledôneas	GCF_000715135.1	29
Vigna radiata var. radiata	eudicotiledôneas	GCF_000741045.1	14
Raphanus sativus	eudicotiledôneas	GCF_000801105.1	24
Arachis ipaensis	eudicotiledôneas	GCF_000816755.2	15
Arachis duranensis	eudicotiledôneas	GCF_000817695.3	14
Phalaenopsis equestris	monocotiledôneas	GCF_001263595.1	9
Solanum pennellii	eudicotiledôneas	GCF_001406875.1	15
Juglans regia	eudicotiledôneas	GCF_001411555.2	12
Cynara cardunculus var. scolymus	eudicotiledôneas	GCF_001531365.2	13
Ananas comosus	monocotiledôneas	GCF_001540865.1	10
Dendrobium catenatum	monocotiledôneas	GCF_001605985.1	6
Dendrobium catenatum	monocotiledôneas	GCF_001605985.2	6
Daucus carota subsp. sativus	eudicotiledôneas	GCF_001625215.1	12
Quercus lobata	eudicotiledôneas	GCF_001633185.2	13
Hevea brasiliensis	eudicotiledôneas	GCF_001654055.1	15
Manihot esculenta	eudicotiledôneas	GCF_001659605.2	13
Chenopodium quinoa	eudicotiledôneas	GCF_001683475.1	19
Lupinus angustifolius	*eudicotiledôneas*	GCF_001865875.1	16
Asparagus officinalis	monocotiledôneas	GCF_001876935.1	9
Nicotiana attenuata	eudicotiledôneas	GCF_001879085.1	16
Ipomoea nil	eudicotiledôneas	GCF_001879475.1	12
Momordica charantia	eudicotiledôneas	GCF_001995035.1	9

Spinacia oleracea	eudicotiledôneas	GCF_002007265.1	11
Malus domestica	eudicotiledôneas	GCF_002114115.1	16
Helianthus annuus	eudicotiledôneas	GCF_002127325.2	15
Triticum dicoccoides	monocotiledôneas	GCF_002162155.2	24
Prunus avium	eudicotiledôneas	GCF_002207925.1	12
Panicum hallii	monocotiledôneas	GCF_002211085.1	9
Durio zibethinus	eudicotiledôneas	GCF_002303985.1	18
Aegilops tauschii	monocotiledôneas	GCF_002575655.2	13
Cucurbita maxima	eudicotiledôneas	GCF_002738345.1	14
Cucurbita moschata	eudicotiledôneas	GCF_002738365.1	14
Olea europaea var. sylvestris	eudicotiledôneas	GCF_002742605.1	13
Cucurbita pepo subsp. pepo	eudicotiledôneas	GCF_002806865.1	14
Lactuca sativa	eudicotiledôneas	GCF_002870075.4	11
Capsicum annuum	eudicotiledôneas	GCF_002878395.1	15
Quercus suber	eudicotiledôneas	GCF_002906115.1	10
Rosa chinensis	eudicotiledôneas	GCF_002994745.2	12
Triticum urartu	monocotiledôneas	GCF_003073215.2	13
Arachis hypogaea cultivar:Tifrunner	eudicotiledôneas	GCF_003086295.2	28
Medicago truncatula	eudicotiledôneas	GCF_003473485.1	12
Papaver somniferum	eudicotiledôneas	GCF_003573695.1	19
Ipomoea triloba	eudicotiledôneas	GCF_003576645.1	13
Coffea eugenioides	eudicotiledôneas	GCF_003713205.1	12
Coffea arabica	eudicotiledôneas	GCF_003713225.1	22
Abrus precatorius	eudicotiledôneas	GCF_003935025.1	12

Vigna unguiculata	eudicotiledôneas	GCF_004118075.2	14
Camellia sinensis	eudicotiledôneas	GCF_004153795.1	18
Glycine soja	eudicotiledôneas	GCF_004193775.1	21
Vitis riparia	eudicotiledôneas	GCF_004353265.1	11
Salvia splendens	eudicotiledôneas	GCF_004379255.1	21
Juglans microcarpa x Juglans regia	eudicotiledôneas	GCF_004785595.1	12
Prosopis alba	eudicotiledôneas	GCF_004799145.1	13
Populus alba	eudicotiledôneas	GCF_005239225.1	12
Setaria viridis	monocotiledôneas	GCF_005286985.1	10
Hibiscus syriacus	eudicotiledôneas	GCF_006381635.1	15
Punica granatum	eudicotiledôneas	GCF_007655135.1	12
Gossypium hirsutum	eudicotiledôneas	GCF_007990345.1	30
Pistacia vera	eudicotiledôneas	GCF_008641045.1	15
Nymphaea colorata	eudicotiledôneas	GCF_008831285.2	5
Phoenix dactylifera	monocotiledôneas	GCF_009389715.1	11
Benincasa hispida	eudicotiledôneas	GCF_009727055.1	9
Dioscorea cayenensis subsp. rotundata	monocotiledôneas	GCF_009730915.1	8
Andrographis paniculata	eudicotiledôneas	GCF_009805555.1	11
Erigeron canadensis	eudicotiledôneas	GCF_010389155.1	11
Mangifera indica	eudicotiledôneas	GCF_011075055.1	17
Macadamia integrifolia	eudicotiledôneas	GCF_013358625.1	8
Tripterygium wilfordii	eudicotiledôneas	GCF_013401445.1	18
Diospyros lotus	eudicotiledôneas	GCF_014633365.1	9
Panicum virgatum	monocotiledôneas	GCF_016808335.1	18

Triticum aestivum	monocotiledôneas	GCF_018294505.1	34
Zingiber officinale	monocotiledôneas	GCF_018446385.1	27
Carya illinoinensis	eudicotiledôneas	GCF_018687715.1	11
Vigna umbellata	eudicotiledôneas	GCF_018835915.1	13
Telopea speciosissima	eudicotiledôneas	GCF_018873765.1	7
Solanum stenotomum	eudicotiledôneas	GCF_019186545.1	15
Lollum perenne	monocotiledôneas	GCF_019359855.1	9
Pyrus x bretschneideri	eudicotiledôneas	GCF_019419815.1	15
Ricinus communis	eudicotiledôneas	GCF_019578655.1	12
Trifolium pratense	eudicotiledôneas	GCF_020283565.1	15
Brassica napus	eudicotiledôneas	GCF_020379485.1	34
Ziziphus jujuba var. spinosa	eudicotiledôneas	GCF_020796205.1	10
Rhodamnia argentea	eudicotiledôneas	GCF_020921035.1	14
Citrus sinensis	eudicotiledôneas	GCF_022201045.2	9
Lolium rigidum	monocotiledôneas	GCF_022539505.1	14
Salvia hispanica	eudicotiledôneas	GCF_023119035.1	12
Pisum sativum	eudicotiledôneas	GCF_024323335.1	13
Cucumis melo	eudicotiledôneas	GCF_025177605.1	10
Gossypium raimondii	eudicotiledôneas	GCF_025698545.1	19
Solanum verrucosum	eudicotiledôneas	GCF_900185275.1	14
Cannabis sativa	eudicotiledôneas	GCF_900626175.2	0
Zea mays	monocotiledôneas	GCF_902167145.1	10
Prunus dulcis	eudicotiledôneas	GCF_902201215.1	11
Hordeum vulgare subsp. vulgare	monocotiledôneas	GCF_904849725.1	12

Impatiens glandulifera	eudicotiledôneas	GCF_907164915.1	13
Malus sylvestris	eudicotiledôneas	GCF_916048215.2	16
Quercus robur	eudicotiledôneas	GCF_932294415.1	9
Argentina anserina	eudicotiledôneas	GCF_933775445.1	11
Mercurialis annua	eudicotiledôneas	GCF_937616625.1	11

Tabela Suplementar S3. Números de acesso do GenBank SRA utilizados para a montagem do transcriptoma. A. Todos os dados de RNA-Seq de *Theobroma cacao* utilizados. B. Dados de RNA-Seq de *Theobroma umbraticum* utilizados. Fonte: Alves *et al.*, 2024

A

SRR7388543	SRR7388568	SRR7388593	SRR11389079	SRR7172375	SRR3217284
SRR7388544	SRR7388569	SRR7388594	SRR11389080	SRR7172376	SRR3217292
SRR7388545	SRR7388570	SRR7388595	SRR11389081	SRR7172377	SRR3217294
SRR7388546	SRR7388571	SRR7388596	SRR11389082	SRR7172378	SRR3217297
SRR7388547	SRR7388572	SRR7388597	SRR11389083	SRR7172379	SRR3217298
SRR7388548	SRR7388573	SRR7388598	SRR11389084	SRR7172380	SRR3217299
SRR7388549	SRR7388574	SRR7388599	SRR7172356	SRR7172381	SRR3217301
SRR7388550	SRR7388575	SRR7388600	SRR7172357	SRR7172382	SRR3217304
SRR7388551	SRR7388576	SRR7388601	SRR7172358	SRR7172383	SRR3217315
SRR7388552	SRR7388577	SRR7388602	SRR7172359	SRR7172384	SRR3217317
SRR7388553	SRR7388578	SRR7388603	SRR7172360	SRR7172385	SRR3217318
SRR7388554	SRR7388579	SRR7388604	SRR7172361	SRR7172386	SRR3217319
SRR7388555	SRR7388580	SRR7388605	SRR7172362	SRR7172387	SRR1034656
SRR7388556	SRR7388581	SRR7388606	SRR7172363	SRR7172388	SRR1036616
SRR7388557	SRR7388582	SRR7388607	SRR7172364	SRR7172389	SRR747762
SRR7388558	SRR7388583	SRR7388608	SRR7172365	SRR7172390	SRR747765
SRR7388559	SRR7388584	SRR7388609	SRR7172366	SRR7172391	SRR747772
SRR7388560	SRR7388585	SRR7388610	SRR7172367	SRR3217276	SRR747773
SRR7388561	SRR7388586	SRR7388611	SRR7172368	SRR3217277	SRR747774
SRR7388562	SRR7388587	SRR7388612	SRR7172369	SRR3217278	SRR747775
SRR7388563	SRR7388588	SRR7388613	SRR7172370	SRR3217279	SRR747776
SRR7388564	SRR7388589	SRR7388614	SRR7172371	SRR3217280	SRR747777
SRR7388565	SRR7388590	SRR11389076	SRR7172372	SRR3217281	SRR747778
SRR7388566	SRR7388591	SRR11389077	SRR7172373	SRR3217282	SRR747779
SRR7388567	SRR7388592	SRR11389078	SRR7172374	SRR3217283	SRR13960577
					SRR13960578
					SRR13960579

B

SRA Accession	Description
SRR5630802	open flowers
SRR5481331	open flowers
SRR5481332	lateral buds
SRR5481333	closed flower buds
SRR5481334	apical stems
SRR5462401	young leaves