GABRIELA ESTER FERRAZ

POTENTIAL OF ARABICA COFFEE GENOTYPES FOR BEAN QUALITY

LAVRAS – MG
2025

GABRIELA ESTER FERRAZ

POTENTIAL OF ARABICA COFFEE GENOTYPES FOR BEAN QUALITY

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Genética e Melhoramento de Plantas, área de concentração em Genética e Melhoramento de Plantas, para a obtenção do título de Doutor.

Prof.ª Dr(ª). Flávia Maria Avelar Gonçalves
Orientadora

Prof. Dr. Antonio Chalfun Junior
Coorientador

LAVRAS – MG
2025

Ficha catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).

Ferraz, Gabriela Ester.
Potential of Arabica coffee genotypes for bean quality / Gabriela Ester Ferraz. - 2025.
139 p. : il.

Orientador(a): Flávia Maria Avelar Gonçalves.
Coorientador(a): Antonio Chalfun-Junior.
Tese (doutorado) - Universidade Federal de Lavras, 2025.
Bibliografia.

1. diversity. 2. metabolic profile. 3. comparative genomics. I. Gonçalves, Flávia Maria Avelar. II. Chalfun-Junior, Antonio. III. Título.

GABRIELA ESTER FERRAZ

POTENTIAL OF ARABICA COFFEE GENOTYPES FOR BEAN QUALITY

POTENCIAL DE GENÓTIPOS DE CAFÉ ARABICA PARA QUALIDADE DOS GRÃOS

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Genética e Melhoramento de Plantas, área de concentração em Genética e Melhoramento de Plantas, para a obtenção do título de Doutor.

APROVADA em 06 de fevereiro de 2025.
Dr(ª). Flávia Maria Avelar Gonçalves, UFLA
Dr. Antonio Chalfun Junior, UFLA
Dr. Alexsandro Lara Teixeira, EMBRAPA
Dr. Guilherme da Silva Pereira, UFV
Dr. Marcelo Ribeiro Malta, EPAMIG

Prof.ª Dr(ª). Flavia Maria Avelar Gonçalves
Orientadora

Prof. Dr. Antonio Chalfun Junior
Coorientador

LAVRAS-MG
2025

A Deus, pela luz que direcionou o meu caminhar...
Aos meus pai e meu noivo, amores da minha vida...
À cafeicultura, projeto da minha vida!

Dedico

AGRADECIMENTOS

Meus orientadores professores Flávia e Chalfun por todo conhecimento e por acreditar em minha capacidade.

A toda equipe da Empresa de Pesquisa Agropecuária de Minas Gerais (EPAMIG), que colaboraram com o material vegetal para realização de parte das análises moleculares, com as análises sensoriais, e todo conhecimento compartilhado pelos professores Marcelo Ribeiro Malta e Gladyston Rodrigues Carvalho.

Ao Laboratório de Culturas Perenes Sustentáveis, Departamento de Agricultura dos Estados Unidos (USDA – USA), na pessoa do pesquisador Dapeng Zhang.

Ao amigo Ricardo Sebastião dos Santos, extensionista da Empresa de Assistência Técnica e Extensão Rural do Estado de Minas Gerais (EMATER), pelo apoio técnico na condução do experimento implantado na cidade de Inconfidentes, MG.

Todos os colegas do Laboratório de Fisiologia Molecular de Plantas (LFMP) e amigos de classe, pelos ensinamentos, apoio e companheirismo.

Às técnicas Taís Teixeira das Neves do Laboratório de Fisiologia Molecular de Plantas (LFMP), e Lidiany Mendonça Zacaroni Lima da Central de Análises e Prospecção Quimica (CAPQ), por toda atenção e apoio.

Professores e funcionários do Departamento de Biologia que contribuem todos os dias com seu trabalho para manutenção da instituição, me oportunizando as melhores condições para que pudesse adquirir os conhecimentos que tenho hoje.

Ao programa de Pós-Graduação em Genética e Melhoramento de Plantas e a Universidade Federal de Lavras (UFLA) pela oportunidade de me capacitar e adquirir conhecimento.

À Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo apoio financeiro.

RESUMO GERAL

Pesquisas com *Coffea arabica* com foco na obtenção de genótipos promissores para qualidade do grão tem avançado em diversas áreas, incluindo o melhoramento genético, a bioquímica dos frutos, e o estudo de genes relacionados ao desenvolvimento e à qualidade do café. Nesta tese, essas áreas serão abordadas em três artigos. No primeiro artigo, realizou-se a caracterização fenotípica e genotípica de 14 progênies derivadas da cultivar Mundo Novo. Verificou-se alta variabilidade genética e fenotípica, tanto entre quanto dentro das progênies. A variação intracultivar também foi evidente nas cultivares comerciais, destacando que essa característica é comum no café arábica. Essa variabilidade presente em *C. arabica* abre caminho para a identificação de genótipos superiores e para a obtenção de linhagens geneticamente mais homogêneas, contribuindo para maior eficiência nos programas de melhoramento. No segundo artigo, foram realizadas análises de carboidratos e metabólitos diferencialmente abundantes (DAMs), que indicaram variações significativas, especialmente no perisperma inicial, com maior distinção das cultivares no estádio de 30 dias após o florescimento (DS1). Compostos como D-ribose e D-frutose mostraram-se semelhantes entre algumas cultivares apenas em DS1, enquanto diferenças significativas foram observadas em metabólitos relacionados à glicólise e ao ciclo do ácido tricarboxílico, ambos envolvidos na formação de precursores de compostos de qualidade. Por fim, considerando o papel essencial da sacarose nas plantas e sua relevância como componente precursor de sabor do café, o terceiro artigo investigou os genes de invertases, enzimas responsáveis pela quebra irreversível da sacarose. Foram identificados 65 genes em três espécies de *Coffea*, sendo 28 em *C. arabica*, 18 e 19 em representantes modernos de seus parentais, *C. canephora* e *C. eugenioides*, respectivamente. A poliploidização, provavelmente, é a principal causa da expansão no número de genes de invertases em *C. arabica*. Análises de expressão mostraram padrões constitutivos e condicionais em diferentes estádios de desenvolvimento de frutos, fornecendo novas perspectivas para estudos funcionais, especialmente nas espécies de maior relevância econômica, *C. arabica* e *C. canephora*. Portanto, este estudo buscou ampliar o conhecimento sobre fatores que determinam a qualidade dos grãos de *C. arabica*, contribuindo para o desenvolvimento de cultivares promissoras e alinhadas às demandas do mercado global, promovendo a valorização da cadeia produtiva do café.

Palavras-chave: diversidade; qualidade de bebida; GC-MS; metabolômica; perfil de expressão; poliploidia; genômica comparativa.

GENERAL ABSTRACT

Research on *Coffea arabica* aimed at obtaining promising genotypes for grain quality has advanced in various fields, including genetic improvement, fruit biochemistry, and the study of genes related to coffee development and quality. This thesis will address these areas in three manuscripts. The first manuscript conducted a phenotypic and genotypic characterization of 14 progenies derived from the Mundo Novo cultivar. High genetic and phenotypic variability was observed both between and within progenies. Intracultivar variation was also evident in commercial cultivars, highlighting that this characteristic is common in arabica coffee. This variability present in *C. arabica* paves the way for the identification of superior genotypes and the development of genetically more homogeneous lines, contributing to greater efficiency in breeding programs. In the second manuscript, analyses of carbohydrate and differentially abundant metabolites (DAMs) were performed, indicating significant variations, particularly in the early perisperm stage, with greater cultivar distinction at 30 days after flowering (DS1). Compounds such as D-ribose and D-fructose were found to be similar among some cultivars only at DS1, whereas significant differences were observed in metabolites related to glycolysis and the tricarboxylic acid cycle, both of which are involved in the formation of quality-related precursors. Finally, considering the essential role of sucrose in plants and its relevance as a precursor of coffee flavor components, the third manuscript investigated invertase genes, enzymes responsible for the irreversible cleavage of sucrose. A total of 65 genes were identified in three *Coffea* species, including 28 in *C. arabica*, and 18 and 19 in modern representatives of its parent species, *C. canephora* and *C. eugenioides*, respectively. Polyploidization is likely the main cause of the expanded number of invertase genes in *C. arabica*. Expression analyses revealed both constitutive and conditional patterns across different fruit developmental stages, providing new insights for functional studies, particularly in the economically important species *C. arabica* and *C. canephora*. Therefore, this study aimed to expand knowledge about the factors determining the quality of *C. arabica* beans, contributing to the development of promising cultivars aligned with global market demands, thus enhancing the value of the coffee production chain.

Keywords: diversity; cup quality; GC-MS; metabolomics; expression profile; polyploidy; comparative genomics.

INDICADORES DE IMPACTO

Esta tese caracteriza-se por apresentar uma abordagem abrangente sobre *Coffea arabica*, envolvendo diversas áreas de pesquisa, como o melhoramento genético, a bioquímica dos frutos e o estudo de genes relacionados ao desenvolvimento e à qualidade do café. A identificação de genótipos superiores com relação a qualidade dos grãos contribui significativamente para a valorização da cadeia produtiva do café, considerando que este é o produto final de maior relevância comercial. Compreender os mecanismos envolvidos no desenvolvimento dos frutos pode resultar em grãos de qualidade superior, o que, por sua vez, proporciona bebidas de sabores e aromas mais diversificados e atrativos ao mercado consumidor. Entre os impactos sociais, destaca-se a maior sustentabilidade produtiva, alcançada por meio de uma seleção mais eficiente dos genótipos, o que permite otimizar tempo e recursos. Além disso, há benefícios para a saúde humana, visto que, o café é uma bebida reconhecida por suas propriedades antioxidantes, sua capacidade de aumentar o estado de alerta e sua contribuição para a proteção associada a doença de Alzheimer. Dessa forma, a melhoria da qualidade do grão também resulta na preservação de compostos de importância farmacológica que impactam positivamente a qualidade da bebida. No campo tecnológico, os avanços desta pesquisa são significativos, abrindo perspectivas para o desenvolvimento de abordagens interdisciplinares que integrem Genética Quantitativa e Molecular, Genômica Funcional e Comparativa, Biologia Evolutiva, Biologia Sistêmica, Fisiologia Vegetal e Fitotecnia. Do ponto de vista econômico, a valorização da cadeia produtiva do café beneficia todos os seus integrantes, deste o produtor até o consumidor final, uma vez que a maior qualidade dos grãos gera maior rentabilidade ao cafeicultor e agrega valor ao produto final. Os resultados desta pesquisa estão alinhados ás áreas temáticas da Política Nacional de Extensão, incluindo Meio ambiente (5), Saúde (6), Tecnologia e produção (7) e Trabalho (8), Além disso, as contribuições estão em consonância com os Objetivos de Desenvolvimento Sustentável (ODS) da ONU, especialmente os seguintes: ODS 1 (Erradicação da pobreza), ODS 2 (Fome zero a agricultura sustentável), ODS 3 (Saúde e bem estar) e ODS 8 (Trabalho decente e crescimento econômico). Portanto, os resultados apresentados nesta tese, ao aprofundarem o conhecimento sobre o desenvolvimento dos grãos de café, têm o potencial de impactar positivamente a qualidade do produto comercializado. Esse impacto se traduz gerando a possibilidade em maior sustentabilidade na cadeia produtiva, melhor distribuição de renda entre os integrantes e benefícios econômicos, sociais e tecnológicos para o setor cafeeiro como um todo.

IMPACT INDICATORS

This thesis is characterized by its comprehensive approach to *Coffea arabica*, encompassing various fields of research, such as genetic improvement, fruit biochemistry, and the study of genes related to coffee development and quality. The identification of superior genotypes regarding grain quality significantly contributes to the valorization of the coffee production chain, given that this is the most commercially relevant final product. Understanding the mechanisms involved in fruit development can result in higher-quality grains, which, in turn, lead to beverages with more diverse and appealing flavors and aromas for the consumer market. Among the social impacts, greater production sustainability stands out, achieved through more efficient genotype selection, which optimizes time and resources. Furthermore, there are benefits to human health, as coffee is recognized for its antioxidant properties, its ability to enhance alertness, and its role in protection against Alzheimer's disease. Thus, improving grain quality also preserves pharmacologically important compounds, positively impacting beverage quality. In the technological domain, the advances from this research are significant, opening opportunities for the development of interdisciplinary approaches that integrate Quantitative and Molecular Genetics, Functional and Comparative Genomics, Evolutionary Biology, Systems Biology, Plant Physiology, and Agronomy. From an economic perspective, the valorization of the coffee production chain benefits all its stakeholders, from the producer to the final consumer, as higher grain quality generates increased profitability for coffee growers and adds value to the final product. The findings of this research align with the thematic areas of the National Extension Policy, including Environment (5), Health (6), Technology and Production (7), and Work (8). Additionally, its contributions are in accordance with the United Nations Sustainable Development Goals (SDGs), particularly the following: SDG 1 (No Poverty), SDG 2 (Zero Hunger and Sustainable Agriculture), SDG 3 (Good Health and Well-being), and SDG 8 (Decent Work and Economic Growth). Therefore, the results presented in this thesis, by deepening the understanding of coffee grain development, have the potential to positively impact the quality of the marketed product. This impact translates into greater sustainability in the production chain, better income distribution among stakeholders, and economic, social, and technological benefits for the coffee sector as a whole.

SUMÁRIO

PRIMEIRA PARTE..... 11

1 INTRODUÇÃO GERAL..... 11

REFERÊNCIAS..... 14

SEGUNDA PARTE 16

ARTIGO 1 - Phenotypic and Genotypic quantification of intra-cultivar variability in Arabica coffee..... 16

ARTIGO 2 - Metabolic analyses reveal that the early stage of Arabica coffee seed development can be important in determining bean quality.... 40

ARTIGO 3 - Genome-wide characterization of invertases in Arabica coffee and its parents reveals putative genes involved in fruit developmental.. 93

TERCEIRA PARTE..... 138

1 CONSIDERAÇÕES FINAIS..... 138

PRIMEIRA PARTE

INTRODUÇÃO GERAL

O gênero *Coffea* compreende 130 espécies (DAVIS; RAKOTONASOLO, 2021), dentre as quais *Coffea arabica* se destaca como uma espécie alopoliploide (2x = 2n = 44), resultante de uma recente poliploidização originada da fusão de gametas das espécies diploides *C. canephora* e *C. eugenioide*s (SCALABRIN *et al.*, 2020). Apesar de sua base genética ser considerada estreita (SALOJÄRVI *et al.*, 2024; SCALABRIN *et al.*, 2024), a ampla variabilidade fenotípica entre as cultivares (MARTINS *et al.*, 2023; MOREIRA *et al.*, 2022; NADALETI *et al.*, 2022), possibilita avanços significativos no melhoramento genético, visando a atender às demandas atuais para diversas características agronômicas.

Uma outra especificidade de *C. arabica*, é que ela se trata de uma espécie autógama. Entre os métodos mais utilizados para condução de populações segregantes nessa cultura destaca-se o método genealógico, o massal, os retrocruzamentos (CARVALHO, 2008; RAMALHO *et al.*, 2012), e mais recentemente a seleção recorrente (SAAVEDRA *et al.*, 2023). Os processos de melhoramento visam, ao final, à obtenção de linhagens homogêneas com boa estabilidade e adaptabilidade, sendo a produtividade a característica agronômica de maior importância (CARVALHO, 2008). Para aumentar a produção, outras características precisam ser avaliadas como a tolerância a estresses bióticos (SAAVEDRA *et al.*, 2023; SALOJÄRVI *et al.*, 2024), abióticos (BORGO *et al.*, 2025) e a uniformidade de maturação (OLIVEIRA *et al.*, 2024). Outra característica essencial é a qualidade do produto, ou seja, a qualidade do grão.

O conceito de qualidade do grão de café é abrangente e envolve propriedades físicas, químicas e organolépticas. A qualidade física refere-se a granulometria e à ausência de defeitos; a qualidade química, por sua vez, está relacionada aos componentes químicos precursores de sabor; e, por fim, a qualidade organoléptica, envolve aromas e sabores produzidos após os processos de torra e moagem (DE MARIA *et al.*, 1994; LEROY *et al.*, 2006; TORREZ *et al.*, 2023). Adicionalmente, a qualidade do grão também é associada a fatores ecológicos, que incluem características de biodiversidade e funções ecossistêmicas no cultivo do café (TORREZ *et al.*, 2023), que influenciam o desenvolvimento dos frutos e seus componentes. Assim, a qualidade sensorial da bebida é um reflexo da composição química dos grãos, que influencia na diversidade de sabores e aromas após a torra, estando intrinsecamente ligada à qualidade dos grãos.

A qualidade do grão, associada a maiores níveis de produtividade, pode assegurar maior rentabilidade ao produtor. Consequentemente, a qualidade do grão representará uma bebida de melhor qualidade ao consumidor final. A importância comercial do café é grande, sendo *C. arabica* relevante para a economia mundial devido às suas características organolépticas superiores e mais apreciadas pelos consumidores, representando aproximadamente 57% do mercado de produção mundial (ICO, 2023).

Diante desse contexto, os artigos desta tese abordam três objetivos principais. O primeiro artigo, intitulado "*Phenotypic and Genotypic quantification of intra-cultivar variability in Arabica coffee*", objetivou investigar a variabilidade intracultivar em *C. arabica*. Para isso, a variabilidade fenotípica e genotípica de 14 progênies derivadas da cultivar Mundo Novo foram avaliadas, considerando características morfológicas, produtividade total, qualidade sensorial, e análises moleculares por meio de marcadores SNPs (*Single Nucleotide Polymorphism* – Polimorfismos de Nucleotídeo Único). As análises revelaram alta diversidade genética e fenotípica tanto entre quanto dentro das progênies, bem como em cultivares comerciais, evidenciando que a variabilidade intracultivar é uma característica comum no café arábica. Esses resultados possibilitam a identificação de genótipos superiores e o desenvolvimento de linhagens geneticamente mais homogêneas, contribuindo para o melhorando de características agronômicas desejáveis, como a qualidade do grão.

O segundo artigo, "*Metabolic analyses reveal that the early stage of Arabica coffee seed development can be important in determining bean quality*", teve como objetivo aprofundar o entendimento sobre a deposição dos componentes bioquímicos associados a qualidade do grão. Para isso, foi feita a análise da variação de carboidratos essenciais e a identificação de metabólitos diferencialmente abundantes (DAMs) em diferentes estádios de desenvolvimento dos frutos de *C. arabica*, em cultivares de alta qualidade sensorial, porém associadas a categorias distintas. Foram identificados 27 metabólitos, sendo o maior número de DAMs no estádio de 30 dias após o florescimento (DS1), que se mostrou o mais distinto em relação aos demais estádios. As análises revelam diferenças em compostos relacionados a glicólise e ao ciclo do ácido tricarboxílico, envolvidos na formação de precursores para a biossíntese de compostos associados à qualidade final do fruto.

O desenvolvimento fisiológico adequado dos frutos depende do metabolismo eficiente da sacarose, que é o principal fotoassimilado transportado nas plantas, desempenhando um papel importante na relação fonte-dreno (RUAN, 2014). A hidrólise irreversível da sacarose, catalisada pelas enzimas invertases, gera hexoses que estão envolvidas em processos biológicos

essenciais (BRAUN *et al.*, 2014; RUAN, 2014). Nesse contexto, o terceiro artigo, intitulado "*Genome-wide characterization of invertases in Arabica coffee and its parents reveals putative genes involved in fruit development*", teve como objetivo indentificar quais genes de invertases estão relacionados ao desenvolvimento dos frutos que, consequentemente, geram impacto na qualidade desses. Foram caracterizados 28 genes de invertase em *Coffea arabica*, 18 em *C. canephora* e 19 *C.eugenioides*. Em *C. arabica* e *C. canephora* todos os genes estão associados ao desenvolvimento dos frutos. Diante da importância da sacarose tanto como precursores de sabor (CHENG *et al.*, 2016) quanto como fornecedora de esqueletos de carbono para o metabolismo secundário (HUANG *et al.*, 2021), esses resultados abrem perspectivas para compreensão dos mecanismos envolvidos na deposição de componentes essenciais ao adequado desenvolvimento do grão.

Ao longo desta tese, os resultados obtidos destacam que uma abordagem integrada, envolvendo análises genéticas, transcriptômicas, genômica funcional, bioquímicas e moleculares, pode revelar processos-chave que influenciam tanto a qualidade dos grãos quanto a produtividade do café. Esses avanços representam progressos significativos no desenvolvimento de cultivares mais eficientes, resilientes e alinhadas às demandas do mercado global. Além disso, contribuem para a valorização da cadeia produtiva do café e ampliam o conhecimento sobre os fatores que determinam a qualidade dos grãos desta bebida tão apreciada mundialmente.

REFERÊNCIAS

BORGO, L.; RABÊLO, F. H. S.; MARCHIORI, P. E. R. *et al*. Impact of Drought, Heat, Excess Light, and Salinity on Coffee Production: Strategies for Mitigating Stress Through Plant Breeding and Nutrition. **Agriculture**, [*s. l.*], v. 15, n. 1, p. 9, 2025.

BRAUN, D. M.; WANG, L.; RUAN, Y.-L. Understanding and manipulating sucrose phloem loading, unloading, metabolism, and signalling to enhance crop yield and food security. **Journal of Experimental Botany**, [*s. l.*], v. 65, n. 7, p. 1713–1735, 2014.

CARVALHO, C. H. S. **Cultivares de café: origem, características e recomendações**. 1. ed. Brasília, DF: Embrapa Café, 2008. 2008.

CHENG, B.; FURTADO, A.; SMYTH, H. E. *et al*. Influence of genotype and environment on coffee quality. **Trends in Food Science & Technology**, [*s. l.*], v. 57, p. 20–30, 2016.

DAVIS, A. P.; RAKOTONASOLO, F. Six new species of coffee (*Coffea*) from northern Madagascar. **Kew Bulletin**, [*s. l.*], v. 76, n. 3, p. 497–511, 2021.

DE MARIA, C. A. B.; TRUGO, L. C.; MOREIRA, R. F. A. *et al*. Composition of green coffee fractions and their contribution to the volatile profile formed during roasting. **Food Chemistry**, [*s. l.*], v. 50, n. 2, p. 141–145, 1994.

HUANG, L.; HU, L.; FAN, Y. *et al*. Comparative transcriptome analysis revealed the influence of sucrose on flavor and taste quality of *Coffea arabica* and *C. canephora* varieties. **Beverage Plant Research**, [*s. l.*], v. 1, n. 1, p. 1–9, 2021.

ICO. **INTERNATIONAL COFFEE ORGANIZATION** INTERNATIONAL COFFEE ORGANIZATION, , 2023. Disponível em: https://icocoffee.org/documents/cy2023-24/Coffee_Report_and_Outlook_December_2023_ICO.pdf. Acesso em: 21 ago. 2024.

LEROY, T.; RIBEYRE, F.; BERTRAND, B. *et al*. Genetics of coffee quality. **Brazilian Journal of Plant Physiology**, [*s. l.*], v. 18, p. 229–242, 2006.

MARTINS, A. N.; TURCO, P. H. N.; ARAÚJO, H. S. de *et al*. Productivity and physical quality of grains from *Coffea arabica* L. in a tropical high-altitude climate in Brazil. **Coffee Science - ISSN 1984-3909**, [*s. l.*], v. 18, p. e182167–e182167, 2023.

MOREIRA, P. C.; DE REZENDE ABRAHÃO, J. C.; PORTO, A. C. da M. *et al*. Progeny Selection to Develop a Sustainable Arabica Coffee Cultivar. **Agronomy**, [*s. l.*], v. 12, n. 5, p. 1144, 2022.

NADALETI, D. H. S.; DE R. ABRAHÃO, J. C.; ANDRADE, V. T.; *et al*. Sensory quality characterization and selection from a *Coffea arabica* germplasm collection in Brazil. **Euphytica**, [*s. l.*], v. 218, n. 4, p. 35, 2022.

OLIVEIRA, R. R. de; NOMAN, M.; AZEVEDO, L. M. *et al*. Regulation of *Coffea arabica* floral development, flowering and fruit maturation by plant growth regulators. *Em*: ADVANCES IN BOTANICAL RESEARCH. [*S. l.*]: Academic Press, 2024. Disponível em:

https://www.sciencedirect.com/science/article/pii/S0065229624001320. Acesso em: 13 jan. 2025.

RAMALHO, M. A. P.; ABREU, Â. de F. B.; SANTOS, J. B. dos; *et al*. 9. A genética quantitativa e os métodos de condução de populações segregantes em espécies autógamas. *Em*: APLICAÇÕES DA GENÉTICA QUANTITATIVA NO MELHORAMENTO DE PLANTAS AUTÓGAMAS. 1. ed. [*S. l.*]: UFLA, Lavras, 2012. p. 217–262.

RUAN, Y.-L. Sucrose Metabolism: Gateway to Diverse Carbon Use and Sugar Signaling. **Annual Review of Plant Biology**, [*s. l.*], v. 65, n. 1, p. 33–67, 2014.

SAAVEDRA, L. M.; CAIXETA, E. T.; BARKA, G. D.; *et al*. Marker-Assisted Recurrent Selection for Pyramiding Leaf Rust and Coffee Berry Disease Resistance Alleles in *Coffea arabica* L. **Genes**, [*s. l.*], v. 14, n. 1, p. 189, 2023.

SALOJÄRVI, J.; RAMBANI, A.; YU, Z. *et al*. The genome and population genomics of allopolyploid Coffea arabica reveal the diversification history of modern coffee cultivars. **Nature Genetics**, [*s. l.*], v. 56, n. 4, p. 721–731, 2024.

SCALABRIN, S.; MAGRIS, G.; LIVA, M. *et al*. A chromosome-scale assembly reveals chromosomal aberrations and exchanges generating genetic diversity in Coffea arabica germplasm. **Nature Communications**, [*s. l.*], v. 15, n. 1, p. 463, 2024.

SCALABRIN, S.; TONIUTTI, L.; DI GASPERO, G. *et al*. A single polyploidization event at the origin of the tetraploid genome of Coffea arabica is responsible for the extremely low genetic variation in wild and cultivated germplasm. **Scientific Reports**, [*s. l.*], v. 10, n. 1, p. 4642, 2020.

TORREZ, V.; BENAVIDES-FRIAS, C.; JACOBI, J. *et al*. Ecological quality as a coffee quality enhancer. A review. **Agronomy for Sustainable Development**, [*s. l.*], v. 43, n. 1, p. 19, 2023.

SEGUNDA PARTE

ARTIGO 1

Phenotypic and genotypic quantification of intra-cultivar variability in Arabica coffee

(Draft Version)

Manuscript prepared for submission to the Plants Journal

Gabriela Ester Ferraz[1], Marcelo Ribeiro Malta[2], Renato Ribeiro de Lima[3], Carlos Henrique Siqueira de Carvalho[4], Gladyston Rodrigues Carvalho[2], Dapeng Zhang [5,*], Antonio Chalfun-Junior [6,*], Flávia Maria Avelar Gonçalves [1]

[1] Institute of Natural Sciences, Department of Biology, Federal University of Lavras, Lavras 37200-900, MG, Brazil; gabriela.ferraz3@estudante.ufla.br (G.E.F); avelar@ufla.br (F.M.A.G.)
[2] Experimental Field of Lavras, Minas Gerais Agricultural Research Agency (EPAMIG), Lavras 37200-000, MG, Brazil; marcelomalta@epamig.br (M.R.M.); grodriguescarvalho@gmail.com (G.R.C.)
[3] Department of Statistics, Federal University of Lavras, Lavras 37200-900, MG, Brazil; rrlima@ufla.br (R.R.d.L.)
[4] Brazilian Agricultural Research Corporation (Embrapa – Coffee), Varginha 37026-483, MG, Brazil; carlos.carvalho@embrapa.br (C.H.S.d.C.)
[5] Sustainable Perennial Crops Laboratory, United States Department of Agriculture, 10300 Agricultural Research Service, USDA, West Beltsville, Md 20705, USA
[6] Institute of Natural Sciences, Laboratory of Plant Molecular Physiology, Department of Biology, Federal University of Lavras, Lavras 37200-900, MG, Brazil
*** Correspondence:** dapeng.zhang@usda.gov (D.Z.); chalfunjunior@ufla.br (A.C.-J.)

Abstract: In plant breeding programs, the phenotypic and genotypic characterization of populations is essential for defining more effective strategies that optimize resources and time. In this study, (i) the phenotypic variability of 14 progenies of *Coffea arabica* derived from the Mundo Novo cultivar was evaluated based on morphological traits, total yield, and sensory quality over two harvests; and (ii) the genetic variability among and within progenies was investigated using single nucleotide polymorphisms (SNPs), including a comparison with 11 established cultivars. The analysis revealed high genetic and phenotypic diversity both among and within progenies. Intracultivar genetic variation was also evident in commercial cultivars, highlighting that this characteristic is common in *C. arabica*. The intracultivar variability present in Arabica coffee enables the identification of superior genotypes, as well as the development of genetically more homogeneous lines, contributing to greater efficiency in breeding programs.

Keywords: genotypic diversity; phenotypic diversity; progenies; morphological traits; cup quality; productivity

1. Introduction

The genus *Coffea* comprises 130 species [1], among which *Coffea arabica* stands out as an allotetraploid species (2x = 2n = 44), resulting from a recent polyploidization event caused by the fusion of gametes from the diploid species *C. canephora* and *C. eugenioides* [2]. Although recent studies have extensively identified mechanisms generating genetic variability in the genome of *C. arabica*, its genetic basis is considered narrow [2,3]. However, phenotypic variability among cultivars is broad [4–6], enabling advances in breeding for various agronomic traits that meet current demands [7].

C. arabica is predominantly self-pollinating, in contrast to its progenitor species, which exhibit cross-pollination. This aspect directly influences breeding strategies, which are specific to self-pollinating species. After hybridization, generations are advanced in the field until pure lines are obtained, a process that requires approximately five to six generations due to the low crossing rates of these species (< 5%) [8]. For Arabica coffee, the long life cycle, with seed harvests occurring approximately every three years, combined with the necessary steps for progeny evaluation, makes the development of a new cultivar a process that can take up to 30 years [9]. In this context, the phenotypic and genotypic characterization of populations plays a crucial role in breeding programs, enabling the development of more effective strategies that optimize resources and time [10,11].

Since, theoretically, high rates of inbreeding in self-fertilizing species reduce heterozygosity by half with each generation [8], the main goal of generation advancement for the development of pure lines is to achieve homogeneous populations for agronomic traits of interest. However, it is widely recognized that mechanisms of diversity generation occur even in inbred lines, such as residual heterozygosity [12–14], genomic duplications [15–17], and other polymorphisms, including aneuploidies, deletions, and genomic exchanges [18]. If these sources of genetic diversity are not properly understood and managed, they may negatively impact progeny and, consequently, the development of superior cultivars.

Heterogeneity in lines can result in undesired variability in the field, negatively affecting productivity, beverage sensory quality uniformity, stability in tolerance to biotic and abiotic stresses, among other factors. In addition to genetic diversity generation mechanisms arising from natural evolutionary processes, outcrossing rates can vary even in inbred lines [19]. In *C. arabica*, there are few studies quantifying effective outcrossing rates in the population, but the available data indicate substantial variability among cultivars [20]. This underscores the

importance of carefully monitoring the advancement of generations in the process of developing *C. arabica* lines.

Therefore, considering the existing gaps in understanding the process of obtaining *C. arabica* lines, this study aimed to (i) evaluate the phenotypic variability of 14 progenies of *C. arabica* derived from the Mundo Novo cultivar, based on morphological traits, total yield, and sensory quality over two years; and (ii) investigate genetic variability among and within progenies using 96 SNP markers, including a comparison with 11 established cultivars. Our results highlight the potential of molecular markers as essential tools for managing genetic diversity throughout breeding generations, contributing to the development of more homogeneous and promising lines for agronomic traits of interest.

2. Results

2.1. Phenotypic Variation

Over 24 consecutive months, the vegetative growth of 14 progenies and two check cultivars, Catuaí Vermelho IAC144 and Mundo Novo IAC376/4, was evaluated based on plant height and stem diameter (Figure 1). The quarterly values for these variables showed significant variations over time and were more similar to those of the control cultivar Mundo Novo IAC376/4, due to its greater genetic similarity to the progenies [21]. On the other hand, the control cultivar Catuaí Vermelho IAC144 showed significantly lower means for both variables, as expected, due to its characteristic low height [22].

In month 1 (Figure 1a), plant height differed significantly only from the control cultivar Catuaí Vermelho IAC144, which had an average height of 68.9 cm. The progenies exhibited averages ranging from 93.05 cm (progeny 9) to 110.3 cm (progeny 3), but these differences were not significant. From month 4 onward (Figure 1b), significant differences among the progenies became apparent, such as in progeny 3, with an average of 123.73 cm, compared to progeny 9, which had an average of 101.08 cm. As the plants developed, these differences became more pronounced. By month 24 (Figure 1g), progenies 14, 6, 12, 1, and 7, with average heights between 195.53 cm and 198.86 cm, differed significantly from progenies 11, 9, and 10, whose averages ranged from 173.35 cm to 174.8 cm. In this final measurement, all progenies showed significant differences from the control cultivar Catuaí Vermelho IAC144, which maintained the lowest average (129.68 cm).

For stem diameter, differences among progenies were observed as early as the first measurement (Figure 1a). Progeny 3 and the control cultivar Mundo Novo IAC376/4 had

averages of 3.025 cm and 2.975 cm, respectively, differing from progenies 2, 12, and the control cultivar Catuaí Vermelho IAC144, which showed averages of 2.475 cm, 2.45 cm, and 2.275 cm, respectively. The cultivar Catuaí Vermelho IAC144 showed significant differences compared to the other progenies starting from month 12 (Figure 1d), with the smallest average diameter of 2.97 cm. At this same time point, progenies 14 and 4, with averages of 4.175 cm, differed from progenies 2, 8, 5, and 9, whose averages ranged from 3.475 to 3.675 cm. By month 24 (Figure 1g), progenies 2, 3, 5, 8, and 9, along with the control cultivar Mundo Novo IAC376/4, with averages ranging from 4.525 to 4.9 cm, exhibited significantly smaller stem diameters compared to progeny 14 (5.4 cm) and larger diameters than the control cultivar Catuaí Vermelho IAC144, which recorded the smallest average (4 cm).

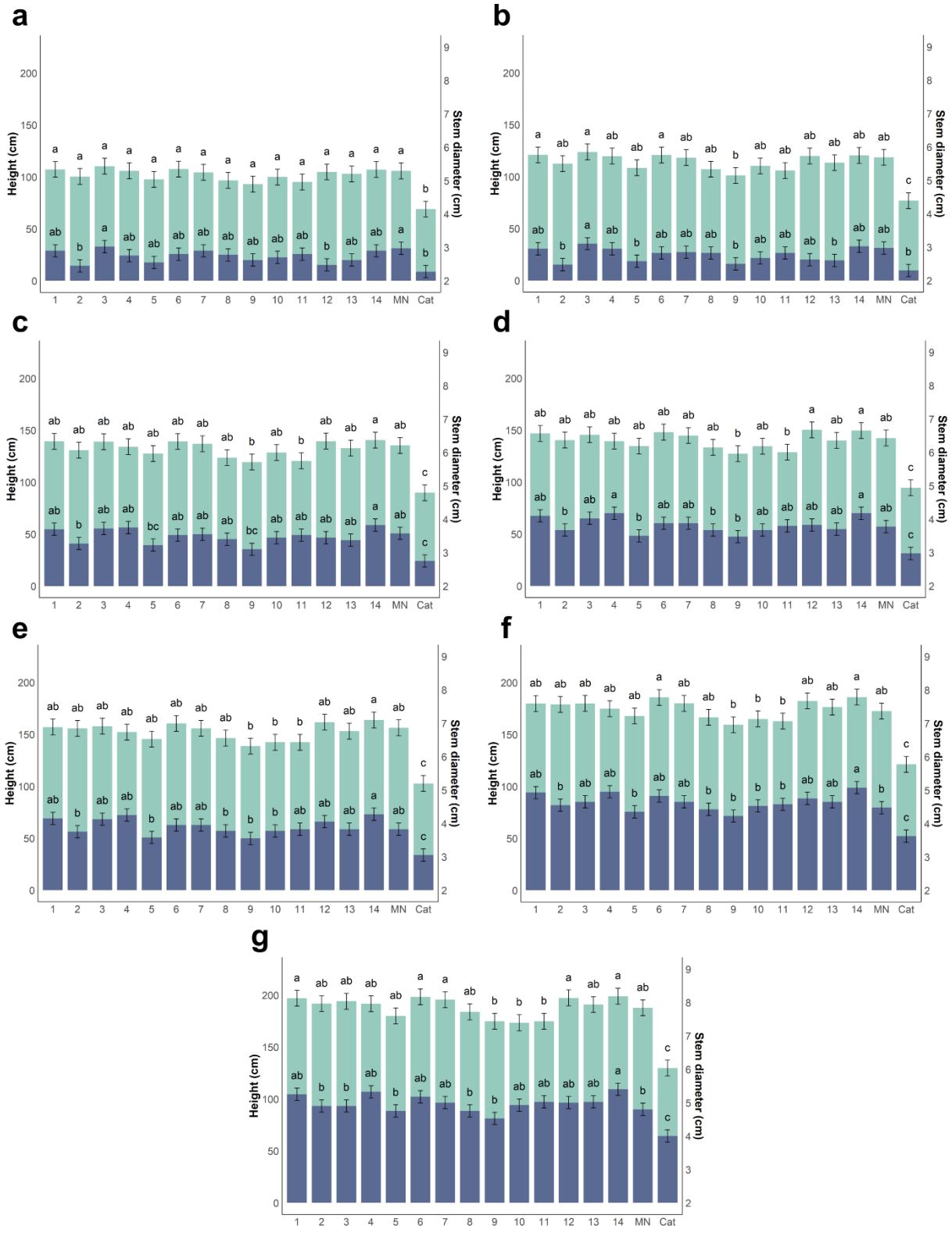


Figure 1. Vegetative growth of 14 progenies (1 – 14) of *Coffea arabica* and two control cultivars, Mundo Novo IAC376/4 (MN) and Catuaí Vermelho IAC144 (Cat), over 24 months. The graphs show the quarterly values for plant height (light green) and stem diameter (blue) measured in months 1 (a), 4 (b), 8 (c), 12 (d), 16 (e), 20 (f), and 24 (g).

Over two harvests, the productivity and sensory quality of 14 progenies and the control cultivars, Catuaí Vermelho IAC144 and Mundo Novo IAC376/4, were quantified (Figure 2). In the first harvest, productivity ranged from 3.63 to 20.90 bags per hectare (bags ha^{-1}) (Figure 1a). The control cultivar Catuaí Vermelho IAC144, with 20.90 bags ha^{-1}, significantly differed from progeny 14, which had an average of 12.71 bags ha^{-1}. This progeny, in turn, showed higher productivity than progenies 9 and 11, which recorded averages of 3.63 bags ha^{-1} and 5.34 bags ha^{-1}, respectively. In the second harvest, the average productivity of the progenies and control cultivars ranged from 4.25 bags ha^{-1} to 13.37 bags ha^{-1}, but no significant differences were observed. This lack of significance may be attributed to the high phenotypic variability in productivity observed in the plots of this harvest (Figure S1), highlighting the broad phenotypic variability among the progenies.

The final total sensory quality score, obtained according to the Specialty Coffee Association (SCA) methodology (Figure 2b), indicated that both the progenies and the control cultivars were categorized as specialty coffees in both harvests, as all achieved scores above 80 points [23].

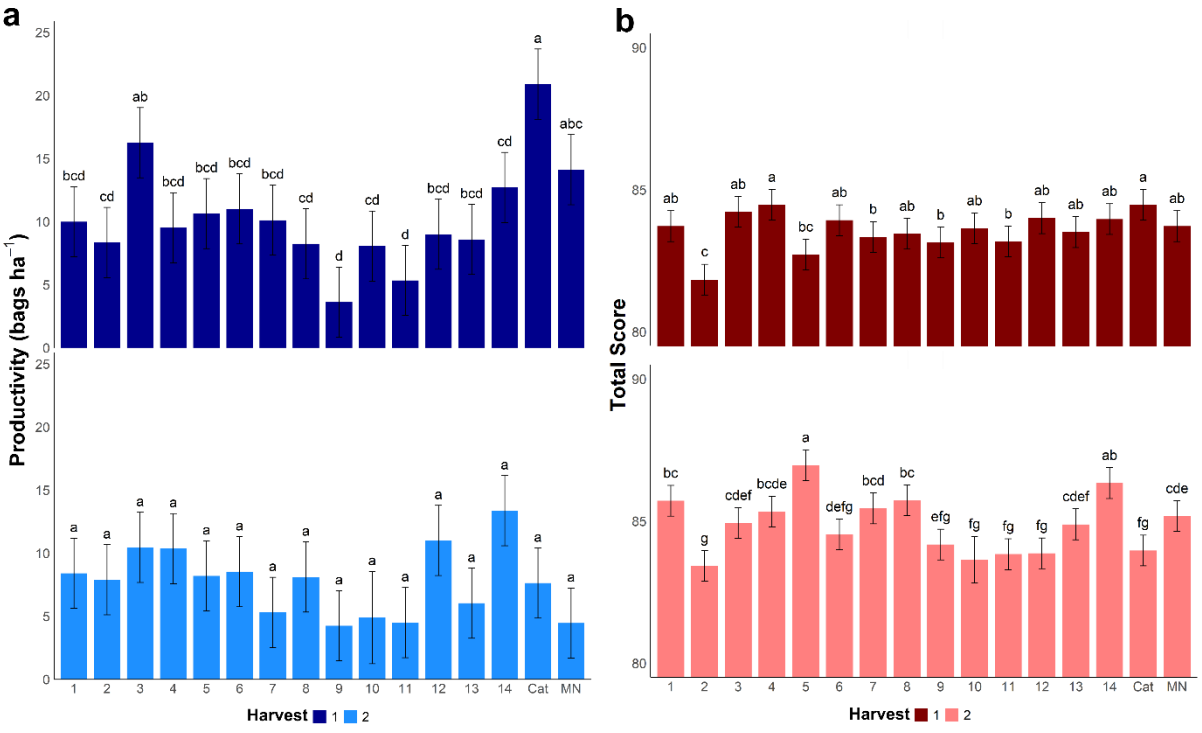


Figure 2. Productivity (a) and total score (b) of 14 progenies (1 – 14) of *Coffea arabica* and two control cultivars, Catuaí Vermelho IAC144 (Cat) and Mundo Novo IAC376/4 (MN), across two harvest.

In the first harvest, progeny 4, along with the control cultivar Catuaí Vermelho IAC144, achieved the highest final score (84.46), significantly differing from progenies 7, 11, 9, 5, and 2. Progeny 5 demonstrated low stability in sensory quality, as it ranked among the progenies with the lowest mean score (82.71) in the first harvest but was among those with the highest mean scores in the second harvest (86.96). Conversely, progeny 2 once again ranked among those with the lowest mean scores (83.42) in the second harvest, significantly differing from progeny 5.

The total score means in the first harvest ranged from 81.83 to 84.46, with all samples classified as very good specialty beverages according to the SCA methodology [24]. In the second harvest, progenies 5, 14, 8, 1, 7, 4, and the control cultivar Mundo Novo IAC376/4 achieved total scores above 85 points, ranging from 85.17 to 86.96, being categorized as excellent. The high sensory quality indices are directly associated with the flavor attributes present in the beverage, particularly in the floral, fruity, and rare sensory note categories [24,25]. Across both harvests, the sensory notes related to the flavor attribute displayed distinct patterns in the progenies and cultivars (Figure 3). Progenies with the highest total score means showed greater frequency in the flavor attribute categories of acidity, body, sweetness, chocolate, and fruitiness. This was evident for progenies 4, 3, 14, 6, 1, and the control cultivar Catuaí Vermelho IAC144 in the first harvest (Figure 3a) and for progenies 5, 14, and 1 in the second harvest (Figure 3b).

Although subtle, the frequency of the floral category was also observed in progenies with the highest total final score means, demonstrating that the presence of sensory notes in this category contributes to elevating the final score. The highest scores, categorized as excellent, were recorded exclusively in the second harvest. In the evaluations from that year, no sensory notes were assigned to the fermented category, and only one defect note was observed, in progeny 6. The absence of these undesirable flavors, which penalize the sensory quality of samples, positively influenced the higher scores.

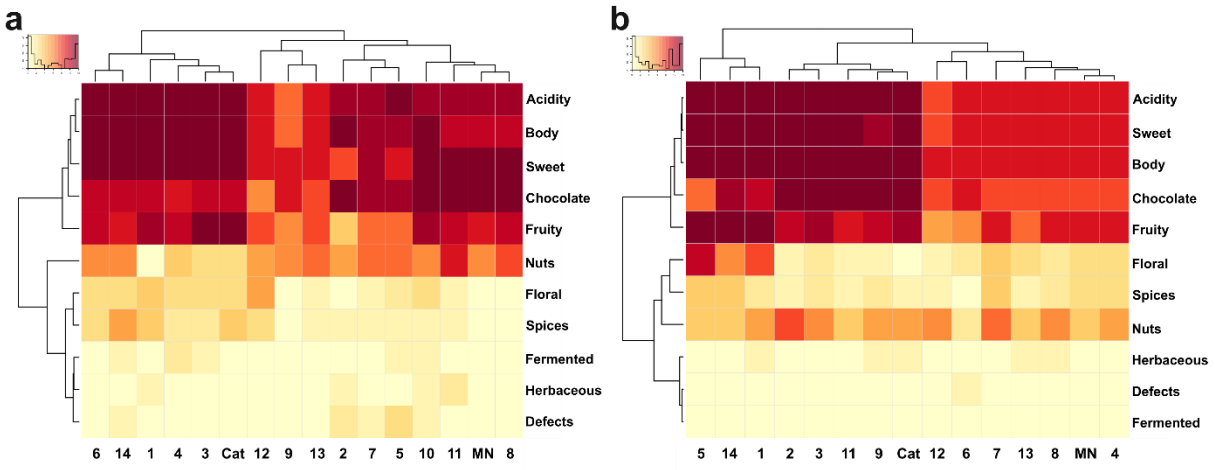


Figure 3. Heatmap of the absolute frequencies of the flavor attribute categories, according to the SCA methodology, for 14 progenies (1 – 14) of *Coffea arabica* and two control cultivars, Catuaí Vermelho IAC144 (Cat) and Mundo Novo IAC376/4 (MN), evaluated across two harvests. (a) Harvest 1; (b) Harvest 2. The categories are sweetness, nuts, chocolate, spices, fermented, floral, fruity, herbaceous, defects, acidity, and body. Dark red represents the highest frequencies, whereas light yellow indicates the lowest frequencies.

Furthermore, the highest total scores coincided with the harvest with lower productivity (Figure 2). One hypothesis for this result is that, during a lower-yielding harvest - such as the second harvest - and considering that the plants had already reached greater vegetative size (Figure 1 a–g), the plants had a higher leaf-to-fruit ratio, ensuring greater availability of photoassimilates [26,27]. Consequently, this may have contributed to better fruit development, positively impacting their final quality. When analyzing the Spearman correlations between productivity and sensory quality, it was observed that in both harvests, the correlations were positive. However, in the first harvest (Figure S2 a), the correlation was moderate and significant (0.59), whereas in the second harvest (Figure S2 b), it was low and not significant (0.35). These results suggest that, for both the progenies and the control cultivars, there was greater independence between these variables in the second year of evaluation, as previously observed in *C. arabica* [28]. The absence of a negative correlation is noteworthy, as it allows for the selection aimed at breeding both traits.

2.1. Genotypic Variation

Using 96 preselected SNPs [29], Bayesian clustering analysis was performed on a set of 250 individual plants derived from 14 *C. arabica* progenies, which were assigned to 12 groups (K=12) (Figure 4). The population structure of the progenies revealed a remarkable degree of

genetic admixture, indicating significant heterogeneity within the progenies. It was observed that some progenies exhibited high levels of genetic admixture among the plants, which was unexpected, considering that *C. arabica* is a predominantly self-pollinating species and the plants analyzed originated from a single mother plant (Figure 7) of the Mundo Novo cultivar [21].

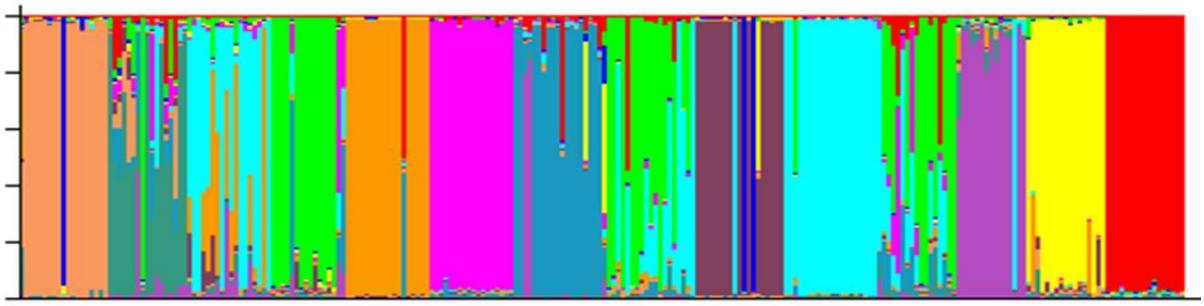


Figure 4. Population structure of 14 selfing progenies of *Coffea arabica*, presented by 250 individual plants using Structure v2.3.4. Multiple colors within each progeny imply admixed individuals under the scenario of K=12.

The overall structure of the 14 progenies was visualized in a neighbor-joining tree (Figure 5). The 250 individual plants were divided into two main groups. Group I included 121 individuals, organized into seven subgroups. Individuals from progeny 4 were included in Group I, distributed across subgroups I and V. Group II comprised 129 individuals, subdivided into four subgroups. Individuals from progenies 3, 12, and 14 were allocated to this group, appearing in subgroups I, III, and IV. In the other progenies (1, 2, 5, 6, 7, 8, 9, 10, 11, and 13), individuals were distributed among subgroups belonging to both major groups, highlighting greater genetic variability within these progenies. In Group II, plants from subgroup I, belonging to progenies 5, 7, 11, and 14, showed greater homogeneity in genetic structure (Figure 4, red color), as did plants from subgroup IV, from progenies 2, 3, 8, 9, 10, and 12 (Figure 4, light blue color). These results highlight that the plants from these progenies exhibit greater genetic similarity among themselves.

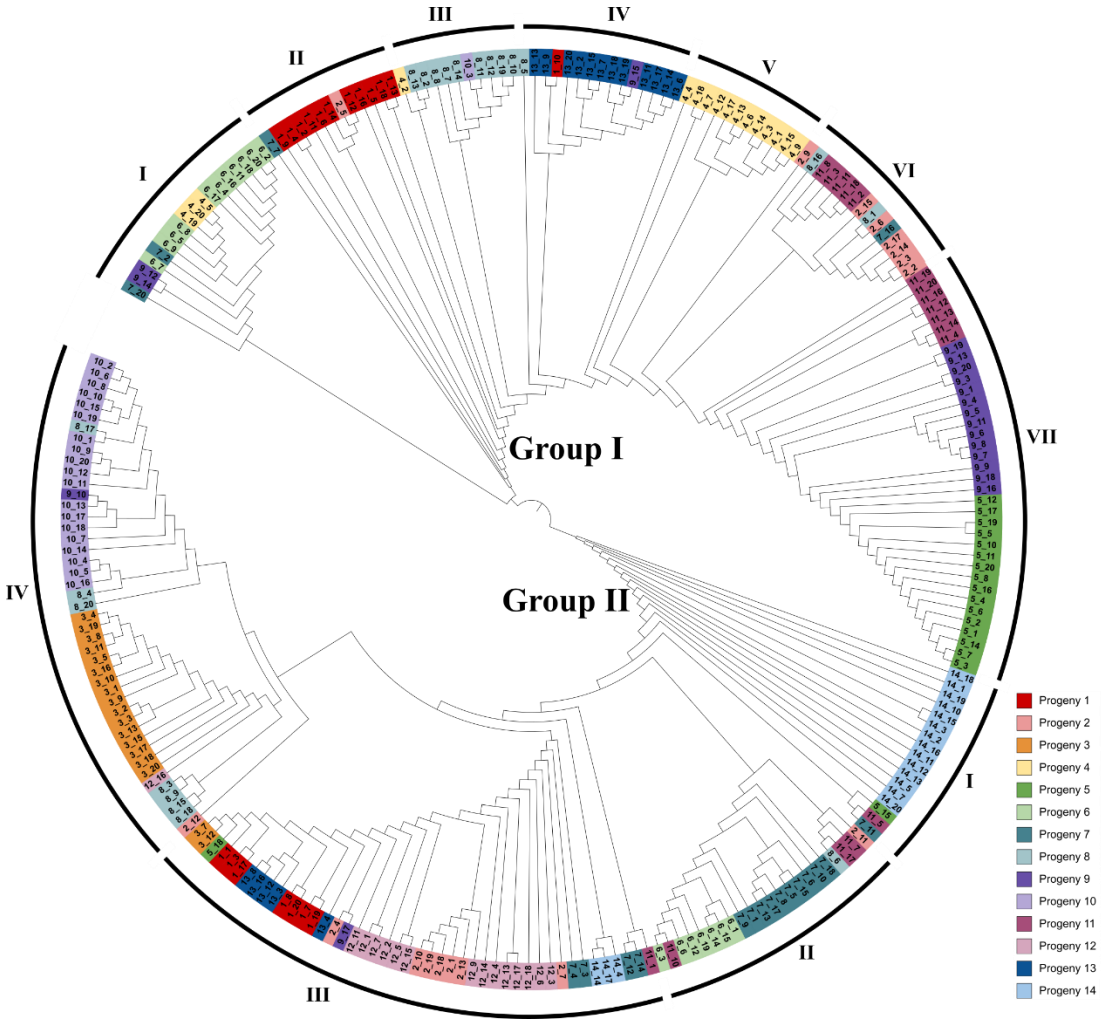


Figure 5. Neighbor-Joining tree depicting the relationships among the 14 selfing progenies derived from *Coffea arabica* represented by 250 individual plants.

To evaluate the intracultivar variation among the 14 progenies, a principal coordinates analysis (PCoA) was performed. Additionally, 55 samples representing 11 cultivars were included in the analysis, providing a comparative reference. The cultivars selected as representatives of genetic variability in *C. arabica* were: Aranãs, Catuaí Vermelho IAC144, Catucaí 2SL, Gueisha, Icatu IAC3282, IPR 100, MGS Paraíso 2, Siriema AS1, Timor Hybrid, Timor Hybrid 442-42, and Typica (Table S1). The PCoA, based on SNP markers, grouped the 250 plants from the progenies and revealed intracultivar variation both within the progenies and the evaluated cultivars (Figure 6). However, the progenies formed a distinct cluster, without overlapping with the cultivars.

The degree of variation appeared to be cultivar-dependent, as some cultivars exhibited greater variation whereas others showed lower diversity. For instance, the native cultivar "Typica" presented very limited variation but was similar to the progenies, positioning itself

close to the cluster that includes them. The degree of diversity also varied among the 14 progenies, with interpopulation variation being greater than intrapopulation variation, consistent with the results observed in the neighbor-joining tree and genetic structure (Figures 4 and 5).

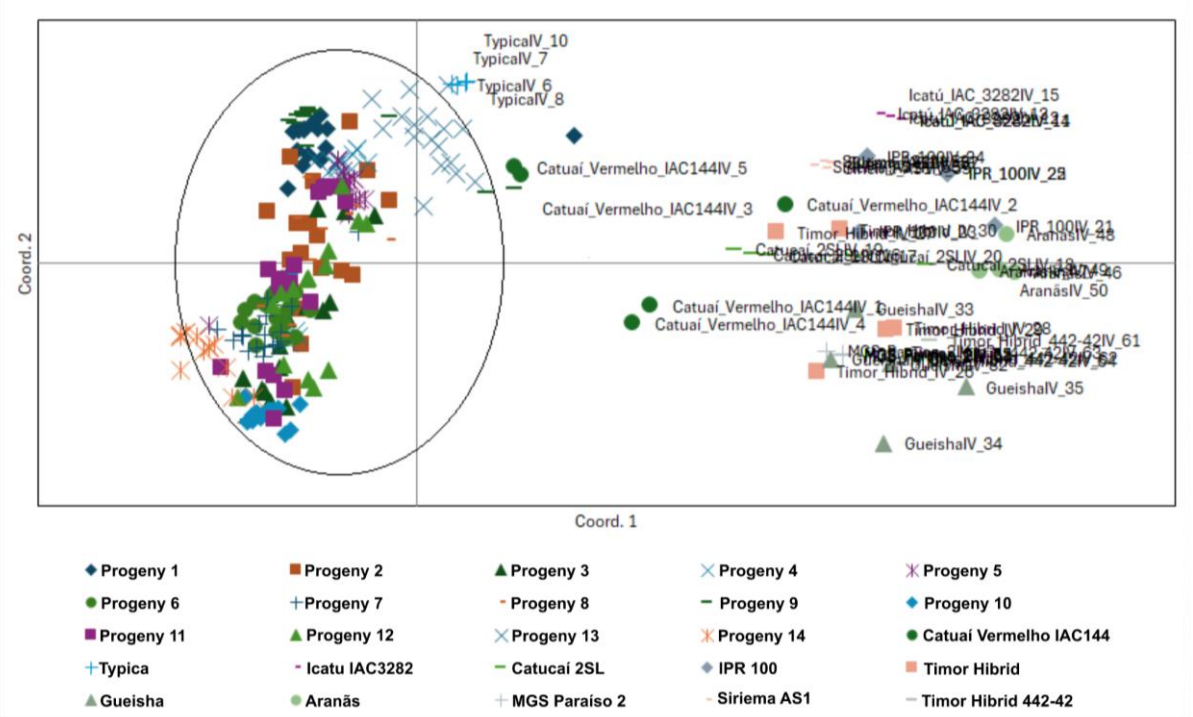


Figure 6. Principal Coordinates Analysis (PCoA) for 14 progenies and 11 cultivars of *Coffea arabica.*

3. Discussion

The phenotypic and genotypic characterization of populations plays a crucial role in breeding programs, allowing for the definition of more effective strategies that optimize resources and time [10,11]. In *Coffea arabica*, a perennial species with a long life cycle, the development of new cultivars can take up to 30 years [9], and detailed knowledge of the genetic material is essential to enhance the efficiency of the breeding process. Despite the narrow genetic base of *C. arabica* [2,3], the observed phenotypic diversity is broad [4–6], enabling progress in agronomic traits of interest, even within the same cultivar, as evidenced by our results. In this study, 14 progenies originating from a single plant of the cultivar Mundo Novo [21], were evaluated, which allowed for the investigation of the phenotypic and genotypic variability within and between the progenies.

The progenies exhibited significant phenotypic variability in morphological traits (Figure 1), showing greater similarity with the Mundo Novo IAC376/4, the reference cultivar

(Figure 1g) due to their relatedness. For the productivity and sensory quality variables (Figures 2 and 3), the progenies demonstrated heterogeneity. As previously observed in *C. arabica* [28], productivity did not show negative correlations with sensory quality (Figure S2), indicating the possibility of selecting promising progenies for both traits. These results highlight a broad selection spectrum within the progenies, similar to the history of the Mundo Novo cultivar, where promising plants were identified on a farm in São Paulo state in 1943, leading to the development of one of the cultivars that most significantly contributed to the expansion of Brazilian coffee cultivation [30]. However, with the advancement of genomic tools and phenotyping strategies it is possible to obtain even more promising cultivars aligned with current demands [7].

The observed phenotypic variation can be explained by the genetic variability found within the progenies, as evidenced both by their heterogeneous genetic structure (Figure 4) and by the distribution of clusters generated by the neighbor-joining tree (Figure 5). These results indicate that the interpopulational variability among the progenies is greater than the intrapopulational variability, confirming the presence of significant genetic variation even in self-pollinating populations. Intracultivar variability has already been observed in other species, such as grapevine clones [31], cherry cultivars [32], olive cultivars [33] and in inbred maize lines [34,35].

Factors such as residual heterozygosity or residual segregation can contribute to intracultivar variability, often interpreted as mutation in the literature. In oilseed rape, residual heterozygosity in lines after seven generations of self-fertilization was extremely high, > 50 %, contrasting with the theoretical estimate of 1.56 % [14]. Similarly, high levels of heterozygosity were observed in European maize lines [12], where the authors suggest that selection against inbreeding depression was crucial in maintaining heterozygosity in the evaluated material. In the inbred diploid clone of *Solanum chacoense*, M6, levels of residual heterozygosity were associated with tuber yield [13]. In contrast, in Arabidopsis and soybean, much of the residual heterozygosity was derived from duplications in the genome [16,17]. Considering that *C. arabica* is a recent polyploid [2], and that genomic instability is still present [36], heterozygosity may be preserved through duplications, adaptive advantages, or other polymorphic events already described in Arabica coffee [18].

In addition to residual heterozygosity, factors such as variation in the rate of cross-pollination can contribute to the observed diversity in the progenies. In Arabidopsis, although cross-pollination rates are estimated at ~1 % [37], variations of up to 14.5 % were observed in

the estimated effective crossbreeding [19]. Whereas it is known that there is variation in cross-pollination rates between *C. arabica* cultivars [20], quantitative studies on this process are scarce. Analyses that identify pollen donors may provide important insights into the possible genetic mixing observed in plants from some progenies (Figure 4).

Although genetic diversity in the early generations for obtaining lines is important, as it enables the selection of the most promising, the principal coordinate analysis (PCoA) reinforced the presence of intracultivar variability in all evaluated cultivars (Figure 6), including commercial ones. However, the degree of this variability proved to be dependent on the cultivar. More ancestral cultivars, such as "Typica," showed low genetic variability, whereas commercial cultivars exhibited greater variation. This increase may reflect the genetic variability resulting from the crosses and hybridizations that originated them, residual heterozygosity, or even polymorphic events in the genomes [18]. However, genetic heterogeneity in commercial cultivars can have negative implications, such as variability in the field, lower productivity, inconsistency in the sensory quality of the beverage, and impacts on tolerance to biotic and abiotic stresses.

The results presented, considering the phenotypic and genotypic evaluation of 14 *C. arabica* progenies, highlight the importance of monitoring genetic quality and seed purity in breeding and propagation programs for the species. Although the objective of breeding programs aiming to generate homogeneous lines is to reduce genetic variability across generations, the outcome of this study indicates that intracultivar variability in Arabica coffee is common. This is the first systematic comparative analysis of intracultivar variation in *C. arabica* based on SNP data, suggesting the need for greater attention in managing the genetic structure of breeding populations. The integration of evaluations with molecular markers can promote greater phenotypic uniformity of progenies over coffee breeding generations, contributing to the development of more homogeneous and production-adapted lines.

4. Materials and Methods

4.1. Genetic Materials and Experimental Design

The field experiment was conducted at “Sítio Gabriela”, located in the municipality of Inconfidentes, in the southwest region of Minas Gerais, Brazil. The site is situated in a section of the “Serra da Mantiqueira”, part of the Atlantic Forest Biome, at an altitude of 1007 meters, with soil classified as clayey-sandy.

Cultivar Mundo Novo IAC376/4, a single mother plant was selected based on its phenotypic variation, which gave rise to 14 progenies. The plants from these progenies stood out for their vigor and the production of large fruits, with a high percentage of beans retained in sieves of size 16 or higher. In 2019, an experiment was established in the field using a randomized complete block design, with 14 progenies and two control cultivars (Catuaí Vermelho IAC144 and Mundo Novo IAC376/4), with four replications and five plants per plot. The spacing used was 0.8 m between plants and 3 m between rows. Over time, some plants did not exhibit good development and died. Therefore, in the evaluations conducted in this study, the experiment included a total of 250 plants derived from the progenies. The implementation and management were carried out following regional technical recommendations for coffee cultivation.

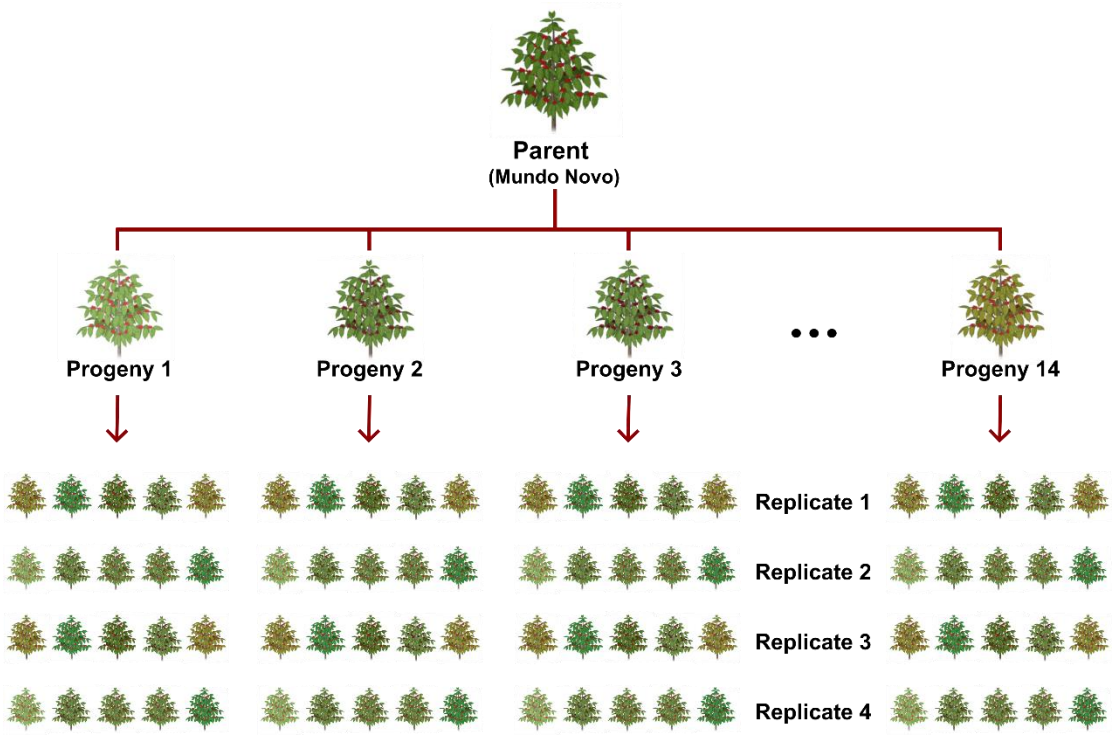


Figure 7. Representative diagram of the experiment involving 14 *Coffea arabica* progenies from which the plant material used in the analyses was obtained. The plants of the check cultivars, Catuaí Vermelho IAC144 and Mundo Novo IAC376/4, followed the same arrangement.

4.2. Morphological Data

As part of the phenotypic characterization, the vegetative growth of the 14 progenies and the two control cultivars, Catuaí Vermelho IAC144 and Mundo Novo IAC376/4, was evaluated over 24 consecutive months. The first evaluation took place six months after the

experiment was implemented. For the evaluations, one plant per plot was selected, including the control cultivars, and the morphological characteristics evaluated were plant height (cm) and stem diameter (cm). Stem diameter was measured at ground level.

4.3. Productivity and Sensory Analyses

Harvests from the 2020/2021 and 2021/2022 seasons were conducted in each experimental plot, including the control cultivars: Catuaí Vermelho IAC144 and Mundo Novo IAC376/4. Through selective harvesting, 2 liters of cherries with complete physiological maturity and no green fruits were used for sensory analyses. Only progeny 10 in the second harvest did not yield a sufficient sample for sensory evaluation. The amount of coffee harvested from each plot was measured in liters and later used to calculate productivity. The samples were dried until they reached a water content of 12 %. After processing, the yield was measured, and the data were used to calculate productivity in bags per hectare. This calculation was based on the volume harvested in each plot relative to the population density, with the result expressed in 60-kilogram bags of processed coffee per hectare (bags ha^{-1}).

The sensory analysis was conducted by three certified tasters, following the methodology proposed by the Specialty Coffee Association (SCA) [24,25], with three repetitions. According to this method, each evaluated attribute (fragrance/aroma, flavor, acidity, body, balance, sweetness, uniformity, cleanliness, and overall score) receives a rating from 0 to 9 based on the intensity observed in the samples, using the terminology presented by Lingle [23]. According to the SCA classification, coffee samples that score below 80 points are not considered specialty coffee; samples scoring between 80 and 84.99 are classified as very good; those scoring between 85 and 89.99 are classified as excellent; and samples scoring above 90 points are considered exceptional.

4.4. Molecular analysis

For each coffee plant, two fully expanded young leaves were collected and inserted into a small zip bag with silica desiccant. After drying, samples were kept in room temperature until further processing. From the dried leaf samples, a total of 8 to 10 leaf disks were collected for each plant using the BioArk Leaf kit (LGC Biosearch Technologies). The prepared BioArk Leaf kits were then submitted to LGC Genomics Hoddesdon, Hertfordshire, UK for DNA extraction and subsequent KASP genotyping.

4.4.1. SNP genotyping

A total of 96 SNPs, which were previously selected and validated by Zhang et al. [29], were used for genotyping. The flanking sequences of the 96 SNPs were submitted to LGC for a KASPar assay design. The KASP chemistry (Biosearch Technologies) was used for genotyping. The KASParTM Genotyping System is a competitive allele-specific dual fluorescence resonance energy transfer (FRET)-based technology for SNP genotyping (Cuppen 2007). The genotyping was conducted following an in-house protocol of LGC. The resulting genotypic data were returned as .csv files and analyzed using an SNP marker analysis software (SNP Viewer version 1.99; Biosearch Technologies).

4.4.2. Molecular Data Analysis

Raw data for the SNP loci and sample calls were organized in Microsoft Excel 2007. Quality control was performed using the Quality Assurance Module from SNP Variation Suite version 8 (SVS8; Golden Helix Inc., Bozeman, Montana). Any SNP having more than a 20% no-call rate was removed from the data set. The filtered data were then used in subsequent analysis.

To assess the population structure in the 14 progenies, the SNP data were analyzed using the model-based Bayesian clustering method [38], as implemented in the software STRUCTURE version 2.3.4 (Stanford University, Stanford, CA, USA). The K value for the genetic clustering was set from 1 to 30. For each given number of clusters (K), 10 independent runs, each with 50,000 iterations following a burn-in of 100,000, were performed. All samples were considered to have unknown origins. The optimum number of clusters (K) was determined using the Delta K method [39,40], as implemented in the online software (STRUCTURE HARVESTER, University of California, Santa Cruz, CA, USA).

To further illustrate the genetic relationships among the individual samples, a cluster analysis using Neighbor-joining method was performed for the 250 plants from the 14 progenies. First, Nei's standard distance [41], was calculated using computer program Microsatellite Analyser (MSA) [42]. The resulting distance matrix was used to generate a dendrogram using the Neighbor-joining algorithm (Saitou and Nei 1987) available in the program PHYLIP [43]. Thereafter, the dendrogram was visualized using the ITOL v6 [44] (https://itol.embl.de/, accessed on 31 December 2024).

To further assess the intra-cultivar variation among the 14 progenies, a Principal coordinate analysis (PCoA) was performed. An additional 55 samples representing 11 cultivars

were added in the analysis, which provided a comparative reference to understand the scope of intra-cultivar variation in the 14 progenies. The cultivars selected as representatives of genetic variability in *C. arabica* were: Aranãs, Catuaí Vermelho IAC144, Catucaí 2SL, Gueisha, Icatu IAC3282, IPR 100, MGS Paraíso 2, Siriema AS1, Timor Hybrid, Timor Hybrid 442-42, and Typica (Table S1). These 55 references plants were genotyped using the same SNP panel and KASP protocol. Pairwise genetic distances were calculated using the "distance" function in the GenAlEx 6.5 program [45]. Based on the distance matrix, Principal coordinate analysis (PCoA) was performed using the same computer program.

4.5. Phenotypic Data Analyses

Split-plot design was used for the variables plant height, stem diameter, productivity, and total score (obtained from sensory analysis), assigning the progeny factor to the main plots arranged in a randomized complete block design (RCBD). For the variables plant height and stem diameter, the month factor was assigned to the subplots, while for productivity and total score, the year factor was assigned to the subplots. A fixed model was used for plant height, stem diameter and productivity, whereas a mixed model was adopted for total score, considering the taster and their interactions as random effects. When significant differences were identified, the Tukey test was applied to compare the progeny means, using SAS software. (https://www.sas.com/pt_br/software/on-demand-for-academics.html, accessed on 01 January 2025).

Spearman correlations were estimated between productivity and total score. Additionally, based on the sensory analysis report conducted following the SCA methodology [24,25], the frequencies of sensory scores for the flavor attribute were calculated and used to generate a heatmap. The analyses and visualizations, including correlations and the heatmap, were performed in R software (version 4.4.0) [46], using the PerformanceAnalytics package [47], for correlations and the gplots package [48] for the heatmap.

5. Conclusions

The analysis of genetic diversity and phenotypic variation in 14 progenies of *C. arabica*, derived from the Mundo Novo cultivar, revealed high variability both among and within the progenies. The presence of intracultivar variation, also observed in commercial cultivars, indicates that this characteristic is common in Arabica coffee. The intracultivar variability present in Arabica coffee cultivars enables the identification of superior genotypes as well as

the development of genetically more homogeneous lines, contributing to greater efficiency in breeding programs.

Supplementary Materials: Figure S1: Interaction plot for block vs. progenies for productivity in 14 progenies of *Coffea arabica* and two control cultivars, Catuaí Vermelho IAC144 and Mundo Novo IAC376/4, across two harvest seasons. Harvest season 1 is shown on the left, and harvest season 2 is shown on the right. Figure S2: Spearman correlation between productivity and total score for 14 progenies of *Coffea arabica* and two control cultivars, Catuaí Vermelho IAC144 and Mundo Novo IAC376/4, across two harvest: 1 (a) and 2 (b). Table S1: Description of the origin of the cultivars used as comparators for the 14 progenies of *Coffea arabica*, derived from the Mundo Novo cultivar, in the Principal Coordinates Analysis (PCoA).

Author Contributions: Conceptualization, data acquisition, and original manuscript writing, G.E.F.; obtaining sensory data, manuscript revision, and contribution to the final version, M.R.M; obtaining data, manuscript revision, and contribution to the final version, C.H.S.d.C; data curation, manuscript revision, and editing, R.R.L.; discussion of data analyses, results, and manuscript revision, G.R.d.C.; bioinformatics and statistical analysis of molecular data, D.Z.; conceptualization, supervision throughout the study, and manuscript revision and editing, F.M.A.G., A.C.-J. and D.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was financially supported by the the “Coordenação de Aperfeiçoamento de Pessoal de Nível Superior” (CAPES) for the doctoral scholarship for G.E.F (88887.568926/2020-00); the “Conselho Nacional de Desenvolvimento Científico e Tecnológico” (CNPq), for scholarship research for A.C.J. under grant number 309005/2022-1, for F.M.A.G under grant number 317001/2021-3, and for G.R.C. under grant number 317634/2021-6; the “Instituto Nacional de Ciência e Tecnologia do Café” (INCT/Café) under the grant of the “Fundação de Amparo à Pesquisa do Estado de Minas Gerais” (FAPEMIG, CAP APQ 03605/17) and scholarship research for A.C.J. (APQ-03406-18); and “Fundação de Amparo à Pesquisa do Estado de São Paulo” (FAPESP, #21/06968-3).

Data Availability Statement: Data are available upon request to the corresponding author.

Acknowledgments: To the Sustainable Perennial Crops Laboratory, United States Department of Agriculture Agricultural Research Service (USDA), for the molecular analysis.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Davis, A. P.; Rakotonasolo, F. Six new species of coffee (*Coffea*) from northern Madagascar. *Kew Bulletin*, **2021**, *76*, 497–511, doi:10.1007/s12225-021-09952-5.
2. Scalabrin, S.; Toniutti, L.; Di Gaspero, G.; Scaglione, D.; Magris, G.; Vidotto, M.; Pinosio, S.; Cattonaro, F.; Magni, F.; Jurman, I.; et al. A Single Polyploidization Event at the Origin of the Tetraploid Genome of *Coffea Arabica* Is Responsible for the Extremely Low Genetic Variation in Wild and Cultivated Germplasm. *Sci Rep* **2020**, *10*, 4642, doi:10.1038/s41598-020-61216-7.
3. Salojärvi, J.; Rambani, A.; Yu, Z.; Guyot, R.; Strickler, S.; Lepelley, M.; Wang, C.; Rajaraman, S.; Rastas, P.; Zheng, C.; et al. The Genome and Population Genomics of Allopolyploid *Coffea arabica* Reveal the Diversification History of Modern Coffee Cultivars. *Nat Genet* **2024**, *56*, 721–731, doi:10.1038/s41588-024-01695-w.
4. Nadaleti, D.H.S.; de R. Abrahão, J.C.; Andrade, V.T.; Malta, M.R.; Botelho, C.E.; Carvalho, G.R. Sensory Quality Characterization and Selection from a *Coffea Arabica* Germplasm Collection in Brazil. *Euphytica* **2022**, *218*, 35, doi:10.1007/s10681-022-02985-2.
5. Moreira, P.C.; de Rezende Abrahão, J.C.; Porto, A.C. da M.; Nadaleti, D.H.S.; Gonçalves, F.M.A.; Carvalho, G.R.; Botelho, C.E. Progeny Selection to Develop a Sustainable Arabica Coffee Cultivar. *Agronomy* **2022**, *12*, 1144, doi:10.3390/agronomy12051144.
6. Martins, A.N.; Turco, P.H.N.; Araújo, H.S. de; Firetti, R. Productivity and Physical Quality of Grains from *Coffea Arabica* L. in a Tropical High-Altitude Climate in Brazil. *Coffee Science - ISSN 1984-3909* **2023**, *18*, e182167–e182167, doi:10.25186/.v18i.2167.
7. Borgo, L.; Rabêlo, F.H.S.; Marchiori, P.E.R.; Guilherme, L.R.G.; Guerra-Guimarães, L.; Resende, M.L.V. de Impact of Drought, Heat, Excess Light, and Salinity on Coffee Production: Strategies for Mitigating Stress Through Plant Breeding and Nutrition. *Agriculture* **2025**, *15*, 9, doi:10.3390/agriculture15010009.
8. Ramalho, M.A.P.; Abreu, Â. de F.B.; Santos, J.B. dos; Nunes, J.A. 4. Estrutura Genética Das Populações de Plantas Autógamas. In *Aplicações da Genética Quantitativa no Melhoramento de Plantas Autógamas*; UFLA, Lavras, 2012; pp. 43–56.
9. Ingelbrecht, I.L.W.; Espinoza, N.A.; Nielen, S.; Jankowicz-Cieslak, J. Mutation Breeding in Arabica Coffee. In *Mutation Breeding in Coffee with Special Reference to Leaf Rust: Protocols*; Ingelbrecht, I.L.W., Silva, M. do C.L. da, Jankowicz-Cieslak, J., Eds.; Springer: Berlin, Heidelberg, 2023; pp. 3–17 ISBN 978-3-662-67273-0.
10. Cobb, J.N.; DeClerck, G.; Greenberg, A.; Clark, R.; McCouch, S. Next-Generation Phenotyping: Requirements and Strategies for Enhancing Our Understanding of Genotype–Phenotype Relationships and Its Relevance to Crop Improvement. *Theor Appl Genet* **2013**, *126*, 867–887, doi:10.1007/s00122-013-2066-0.
11. Kumar, R.; Das, S.P.; Choudhury, B.U.; Kumar, A.; Prakash, N.R.; Verma, R.; Chakraborti, M.; Devi, A.G.; Bhattacharjee, B.; Das, R.; et al. Advances in Genomic

Tools for Plant Breeding: Harnessing DNA Molecular Markers, Genomic Selection, and Genome Editing. *Biological Research* **2024**, *57*, 80, doi:10.1186/s40659-024-00562-6.
12. Brandenburg, J.-T.; Mary-Huard, T.; Rigaill, G.; Hearne, S.J.; Corti, H.; Joets, J.; Vitte, C.; Charcosset, A.; Nicolas, S.D.; Tenaillon, M.I. Independent introductions and admixtures have contributed to adaptation of European maize and its American counterparts. *PLOS Genetics* **2017**, *13*, e1006666, doi:10.1371/journal.pgen.1006666.
13. Marand, A.P.; Jansky, S.H.; Gage, J.L.; Hamernik, A.J.; de Leon, N.; Jiang, J. Residual Heterozygosity and Epistatic Interactions Underlie the Complex Genetic Architecture of Yield in Diploid Potato. *Genetics* **2019**, *212*, 317–332, doi:10.1534/genetics.119.302036.
14. Liang, H.; Ye, J.; Wang, Y.; Wang, X.; Zhou, X.-R.; Batley, J.; King, G.J.; Guo, L.; Tu, J.; Shi, J.; et al. Systematic Trait Dissection in Oilseed Rape Provides a Comprehensive View, Further Insight, and Exact Roadmap for Yield Determination. *Biotechnol Biofuels* **2022**, *15*, 1–19, doi:10.1186/s13068-022-02134-w.
15. Vlad, D.; Rappaport, F.; Simon, M.; Loudet, O. Gene Transposition Causing Natural Variation for Growth in *Arabidopsis thaliana*. *PLOS Genetics* **2010**, *6*, e1000945, doi:10.1371/journal.pgen.1000945.
16. Torkamaneh, D.; Laroche, J.; Tardivel, A.; O'Donoughue, L.; Cober, E.; Rajcan, I.; Belzile, F. Comprehensive Description of Genomewide Nucleotide and Structural Variation in Short-Season Soya Bean. *Plant Biotechnology Journal* **2018**, *16*, 749–759, doi:10.1111/pbi.12825.
17. Jaegle, B.; Pisupati, R.; Soto-Jiménez, L.M.; Burns, R.; Rabanal, F.A.; Nordborg, M. Extensive Sequence Duplication in Arabidopsis Revealed by Pseudo-Heterozygosity. *Genome Biol* **2023**, *24*, 1–19, doi:10.1186/s13059-023-02875-3.
18. Scalabrin, S.; Magris, G.; Liva, M.; Vitulo, N.; Vidotto, M.; Scaglione, D.; Del Terra, L.; Ruosi, M.R.; Navarini, L.; Pellegrino, G.; et al. A Chromosome-Scale Assembly Reveals Chromosomal Aberrations and Exchanges Generating Genetic Diversity in *Coffea Arabica* Germplasm. *Nat Commun* **2024**, *15*, 463, doi:10.1038/s41467-023-44449-8.
19. Bomblies, K.; Yant, L.; Laitinen, R.A.; Kim, S.-T.; Hollister, J.D.; Warthmann, N.; Fitz, J.; Weigel, D. Local-Scale Patterns of Genetic Variability, Outcrossing, and Spatial Structure in Natural Stands of *Arabidopsis thaliana*. *PLOS Genetics* **2010**, *6*, e1000890, doi:10.1371/journal.pgen.1000890.
20. Berecha, G. Genetic Diversity, Pollination Ecology and Organoleptic Characteristics of *Coffea arabica* L., in Ethiopian Moist Forests of Different Management Intensity. Dissertation, Ku Leuven: Leuven, Belgium, 2014.
21. Ferraz, G.E.; López, M.E.; Gonçalves, F.M.A.; Malta, M.R.; Lima, R.R.D.; Carvalho, G.R.D.; Meinhardt, L.W.; Zhang, D.; Chalfun-Junior, A. Progeny Selections of Coffee Cultivar "Mundo Novo" with Potential for the Specialty Coffee Market. *B* **2023**, *3*, 1–11, doi:10.48130/BPR-2023-0001.
22. Carvalho, C.H.S. *Cultivares de Café: Origem, Características e Recomendações*; 1st ed.; Embrapa Café: Brasília, DF, 2008;
23. Lingle, T.R. *The Coffee Cupper's Handbook: A Systematic Guide to the Sensory Evaluation of Coffee's Flavor*; Specialty Coffee Association of America: Long Beach, CA, 2011;
24. Specialty Coffee Association of America (SCAA) SCAA Protocols - Cupping Specialty Coffee. 2015.
25. Rugolo, J.; Alduenda, M.R.F.; Giuliano, P.; Ionescu, K.E. Understanding and Evolving the Specialty Coffee Association Coffee Value Assessment System: Results of the 2020-2021 Cupping Protocol User Perception Study and Proposed Evolution Available online: https://static1.squarespace.com/static/584f6bbef5e23149e5522201/t/62ff6f82e076e71f66

1ca1c6/1660907395782/CVAS+Evolution+Report+2022.pdf (accessed on 31 December 2024).
26. Vaast, P.; Kantem, P. van; Siles, P.; Dzib, B.; Frnack, N.; Harmand, J.M.; Genard, M. Shade: A key factor for coffee sustainability and quality. In Proceedings of the 20th International Conference on Coffee Science; Association Scientifique: Paris : ASIC, 2004; pp. 887–896.
27. Bertrand, B.; Vaast, P.; Alpizar, E.; Etienne, H.; Davrieux, F.; Charmetant, P. Comparison of Bean Biochemical Composition and Beverage Quality of Arabica Hybrids Involving Sudanese-Ethiopian Origins with Traditional Varieties at Various Elevations in Central America. *Tree Physiology* **2006**, *26*, 1239–1248, doi:10.1093/treephys/26.9.1239.
28. Ferreira, D.S.; Canal, G.B.; Nascimento, M.; Nascimento, A.C.C.; Ferreira, J.M.S.; do Amaral, J.F.T.; Pereira, L.L.; Rodrigues, W.N.; Ribeiro, W.R.; Castanheira, D.T.; et al. Exploring the Multivariate Technique in the Discrimination of *Coffea Arabica* L. Cultivars Regarding the Production and Quality of Grains under the Effect of Water Management. *Euphytica* **2021**, *217*, 118, doi:10.1007/s10681-021-02845-5.
29. Zhang, D.; Vega, F.E.; Solano, W.; Su, F.; Infante, F.; Meinhardt, L.W. Selecting a Core Set of Nuclear SNP Markers for Molecular Characterization of Arabica Coffee (*Coffea Arabica* L.) Genetic Resources. *Conservation Genet Resour* **2021**, *13*, 329–335, doi:10.1007/s12686-021-01201-y.
30. Guerreiro Filho, O.; Ramalho, M.A.P.; Andrade, V.T. Alcides Carvalho and the Selection of Catuaí Cultivar: Interpreting the Past and Drawing Lessons for the Future. *Crop Breed. Appl. Biotechnol.* **2018**, *18*, 460–466, doi:https://doi.org/10.1590/1984-70332018v18n4p69.
31. Tortosa, I.; Escalona, J.M.; Douthe, C.; Pou, A.; Garcia-Escudero, E.; Toro, G.; Medrano, H. The Intra-Cultivar Variability on Water Use Efficiency at Different Water Status as a Target Selection in Grapevine: Influence of Ambient and Genotype. *Agricultural Water Management* **2019**, *223*, 105648, doi:10.1016/j.agwat.2019.05.032.
32. Ganopoulos, I.V.; Avramidou, E.; Fasoula, D.A.; Diamantidis, G.; Aravanopoulos, F.A. Assessing inter- and intra-cultivar variation in Greek Prunus avium by SSR markers. *Plant Genetic Resources* **2010**, *8*, 242–248, doi:10.1017/S1479262110000298.
33. Muzzalupo, I.; Chiappetta, A.; Benincasa, C.; Perri, E. Intra-Cultivar Variability of Three Major Olive Cultivars Grown in Different Areas of Central-Southern Italy and Studied Using Microsatellite Markers. *Scientia Horticulturae* **2010**, *126*, 324–329, doi:10.1016/j.scienta.2010.07.014.
34. Gethi, J.G.; Labate, J.A.; Lamkey, K.R.; Smith, M.E.; Kresovich, S. SSR Variation in Important U.S. Maize Inbred Lines. *Crop Science* **2002**, *42*, 951–957, doi:10.2135/cropsci2002.9510.
35. Liu, N.; Liu, J.; Li, W.; Pan, Q.; Liu, J.; Yang, X.; Yan, J.; Xiao, Y. Intraspecific Variation of Residual Heterozygosity and Its Utility for Quantitative Genetic Studies in Maize. *BMC Plant Biol* **2018**, *18*, 1–15, doi:10.1186/s12870-018-1287-4.
36. Adams, K.L.; Wendel, J.F. Polyploidy and Genome Evolution in Plants. *Current Opinion in Plant Biology* **2005**, *8*, 135–141, doi:10.1016/j.pbi.2005.01.001.
37. Bakker, E.G.; Stahl, E.A.; Toomajian, C.; Nordborg, M.; Kreitman, M.; Bergelson, J. Distribution of Genetic Variation within and among Local Populations of *Arabidopsis thaliana* over Its Species Range. *Molecular Ecology* **2006**, *15*, 1405–1418, doi:10.1111/j.1365-294X.2006.02884.x.
38. Pritchard, J.K.; Stephens, M.; Donnelly, P. Inference of Population Structure Using Multilocus Genotype Data. *Genetics* **2000**, *155*, 945–959, doi:10.1093/genetics/155.2.945.

39. Evanno, G.; Regnaut, S.; Goudet, J. Detecting the Number of Clusters of Individuals Using the Software Structure: A Simulation Study. *Molecular Ecology* **2005**, *14*, 2611–2620, doi:10.1111/j.1365-294X.2005.02553.x.
40. Earl, D.A.; vonHoldt, B.M. STRUCTURE HARVESTER: A Website and Program for Visualizing STRUCTURE Output and Implementing the Evanno Method. *Conservation Genet Resour* **2012**, *4*, 359–361, doi:10.1007/s12686-011-9548-7.
41. Nei, M.; Tajima, F.; Tateno, Y. Accuracy of Estimated Phylogenetic Trees from Molecular Data. *J Mol Evol* **1983**, *19*, 153–170, doi:10.1007/BF02300753.
42. Dieringer, D.; Schlötterer, C. Microsatellite Analyser (MSA): A Platform Independent Analysis Tool for Large Microsatellite Data Sets. *Molecular Ecology Notes* **2003**, *3*, 167–169, doi:10.1046/j.1471-8286.2003.00351.x.
43. Felsenstein, J. PHYLIP - phylogeny inference package (v3.2). *Cladistics* 1989, 164–166.
44. Letunic, I.; Bork, P. Interactive Tree Of Life (iTOL) v5: An Online Tool for Phylogenetic Tree Display and Annotation. *Nucleic Acids Research* **2021**, *49*, W293–W296, doi:10.1093/nar/gkab301.
45. Peakall, R.; Smouse, P.E. Genalex 6: Genetic Analysis in Excel. Population Genetic Software for Teaching and Research. *Molecular Ecology Notes* **2006**, *6*, 288–295, doi:10.1111/j.1471-8286.2005.01155.x.
46. R Core Team R: A Language and Environment for Statistical Computing Available online: https://www.R-project.org/.
47. Peterson, B.G.; Carl, P. PerformanceAnalytics: Econometric Tools for Performance and Risk Analysis 2007, 2.0.4.
48. Warnes, G.R.; Bolker, B.; Bonebakker, L.; Gentleman, R.; Huber, W.; Liaw, A.; Lumley, T.; Maechler, M.; Magnusson, A.; Moeller, S.; et al. Gplots: Various R Programming Tools for Plotting Data 2005, 3.2.0.
49. Carvalho, C.H.S.D.; Bartelega, L.; Sera, G.H.; Matiello, J.B.; Almeida, S.R.D.; Santinato, F.; Hotz, A.L. *Catálogo de Cultivares de Café Arábica*; Embrapa Café: Brasília, DF, 2022; ISBN 1678-1694.

Supplementary Materials

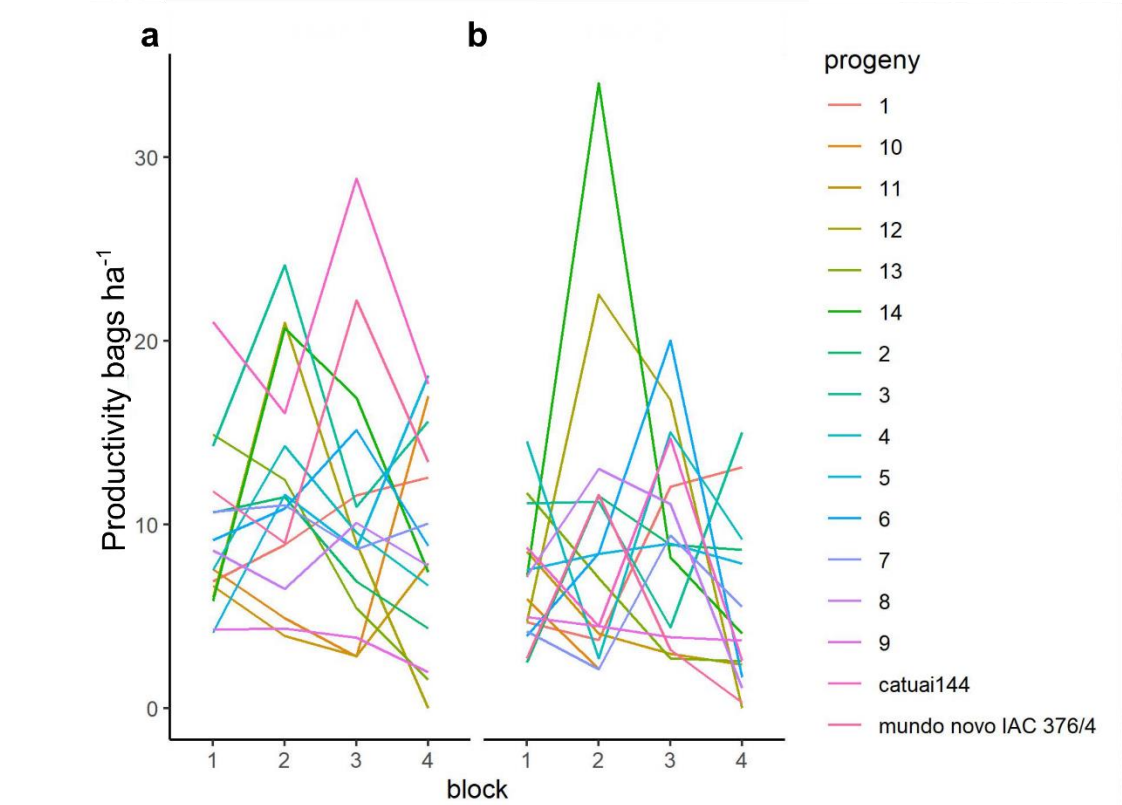


Figure S1. Interaction plot for block vs. progenies for productivity in bags ha^{-1} (60 kg) for 14 progenies of *Coffea arabica* and two control cultivars, Catuaí Vermelho IAC144 and Mundo Novo IAC376/4, across two harvest: 1 (a) and 2 (b).

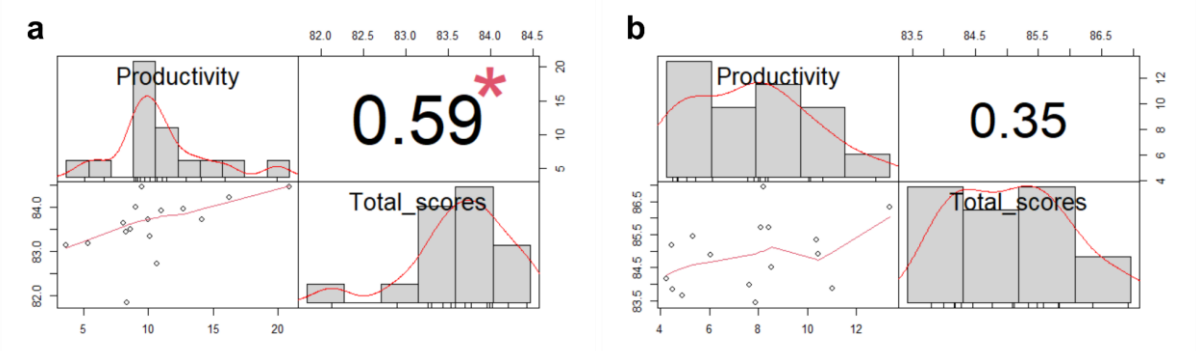


Figure S2. Spearman correlation between productivity and total score for 14 progenies of *Coffea arabica* and two control cultivars, Catuaí Vermelho IAC144 and Mundo Novo IAC376/4, across two harvest: 1 (a) and 2 (b).

Tabela S1. Description of the origin of the cultivars used as comparators for the 14 progenies of *Coffea arabica*, derived from the Mundo Novo cultivar, in the Principal Coordinates Analysis (PCoA).

Cultivars	Genealogy
Aranãs	Icatu Vermelho IAC3851-2 x Catimor 1602-215 UFV
Catuaí Vermelho IAC144	Caturra Amarelo IAC476-11 x Mundo Novo IAC374-19
Catucaí 2SL	Icatu x Catuaí
Gueisha	Landrace from Gesha region, Ethiopia
Icatu precoce IAC3282	Icatu Vermelho x Bourbon Amarelo or Mundo Novo Amarelo
IPR 100	Catuaí Vermelho IAC81 x (Catuaí Vermelho IAC81 x IAC1110-8 (BA10))
MGS Paraíso 2	Catuaí Amarelo IAC30 x Timor Hybrid UFV445-46
Siriema AS1	Selection of Siriema 842 (*C. racemosa* x *C. arabica*)
Timor Hybrid	*C. arabica* x *C. canephora* (Introduced from East Timor)
Timor Hybrid 442-42	*C. arabica* x *C. canephora* (Introduced from East Timor)
Typica	Introduction

Source: Carvalho et al. [49]

ARTIGO 2

Metabolic analyses reveal that the early stage of Arabica coffee seed development can be important in determining bean quality

(Draft Version)

Manuscript prepared for submission to the Agricultural and Food Chemistry Journal

Gabriela Ester Ferraz[1], Robert Márquez Gutiérrez[2], Taís Teixeira das Neves[2], Thales Henrique Cherubino Ribeiro[3], Marili Galvino Santos[1], Raphael Ricon de Oliveira[4], Gladyston Rodrigues de Carvalho[5], Marcelo Ribeiro Malta[5], Renato Ribeiro de Lima[6], Antonio Chalfun-Junior[2*], Flávia Maria Avelar Gonçalves[1]

[1] Institute of Natural Sciences, Department of Biology, Federal University of Lavras, Lavras, 37200900, Brazil

[2] Institute of Natural Sciences, Department of Biology, Laboratory of Plant Molecular Physiology, Plant Physiology Sector, Federal University of Lavras, Lavras, 37200900, Brazil

[3] Genome Center, University of California, Davis, CA 95616, USA

[4] Department of Biology, State University of Santa Cruz, Ilhéus, 5662900, Brazil

[5] Experimental Field of Lavras, Minas Gerais Agricultural Research Agency (EPAMIG), Lavras, 37200-970, Brazil

[6] Department of Statistics, Federal University of Lavras, Lavras, 37200900, Brazil

***E-mail:** chalfunjunior@ufla.br

ABSTRACT

The influence of chemical compounds in coffee seeds on the flavor and aroma precursors present in raw beans remains unclear. This study analyzed the variation of carbohydrates, such as reducing sugars, sucrose, starch, and total sugars, and identified differentially abundant metabolites (DAMs) at different developmental stages of *Coffea arabica* fruits, employing cultivars with specialty sensory quality but categorized into different groups. Biochemical analyses revealed significant variations in flavor and aroma precursors among cultivars. A total of 27 metabolites were identified, with the highest number of DAMs observed at 30 days after flowering (DS1), the most distinct stage according to PLS-DA analysis. Similar levels of D-ribose and D-fructose were found between the MGS Paraíso 2 and Gueisha cultivars, in contrast with MGS Catiguá 3 at DS1. The analyses highlighted differences in compounds related to glycolysis and the tricarboxylic acid cycle, which are involved in forming precursors for the biosynthesis of compounds associated with the quality of coffee beans.

Key words: Metabolic profile, metabolic pathways, chemical composition, coffee quality, GC-MS.

INTRODUCTION

Investigations into the chemical composition of coffee fruits are extensive and aim to elucidate the mechanisms involved in the deposition of their components. These studies often explore the relationship between chemical compounds and the quality of both the fruits and the beverage.[1–9] The relevance of such research stems from the fact that coffee is one of the most consumed beverages globally, with *Coffea arabica* accounting for approximately 57 % and *C. canephora* 43% of global production.[10] The species *C. arabica* produces a beverage that is more appreciated by consumers, whereas *C. canephora* is often used as raw material for instant coffee and in blend compositions in combination with *C.arabica*. This distinction is primarily due to differences in the chemical composition of the fruits among species and cultivars, including sugar levels.[1,11,12]

Coffee quality is a broad concept encompassing various aspects related to the diverse potentialities of the coffee plant. It can be defined by: (i) bean quality, associated with physical characteristics such as size and absence of defects; (ii) chemical quality of the beverage, related to the chemical compounds that act as flavor precursors; and (iii) organoleptic quality, referring to sensory attributes.[13] Therefore, the sensory quality of the beverage reflects the chemical composition of the beans, which influences the diversity of flavors and aromas after roasting, being intrinsically linked to the quality of the beans.

The tissues that make up developing coffee fruits are fundamental in the translocation and storage of components that will constitute the raw bean. In *C. arabica*, the fruit initially consists of the pericarp, which is subdivided into the exocarp (outer skin), mesocarp (pulp), and endocarp (parchment),[2] and the seed (endosperm), which is initially composed of the perisperm - a tissue derived from the nucellus that surrounds the endosperm in the early stages and provides nutrition and support.[4,14] As the fruit develops, the endosperm expands and becomes the main storage tissue, occupying most of the locule.[4] During the initial stages, the pericarp is photosynthetic[1] and contributes to nutrient translocation.

The influence of chemical compounds deposited, stored, and transferred during the development of coffee seeds on the flavor and aroma precursors present in the raw bean remains poorly understood. Among the various identified components, sucrose emerges as a key precursor of beverage quality.[12,15,16] In addition to contributing to the flavor of coffee, sucrose is the primary carbohydrate transported through the phloem of the coffee plant,[5,17] playing a critical role in the developmental stages of the plant and its fruits. Glucose and fructose,

products of sucrose hydrolysis, provide essential carbon frameworks for the synthesis of other structural molecules, which also impact beverage quality.[1,18–20]

Several studies have advanced the understanding of how chemical components are deposited throughout fruit development, encompassing chemical analyses,[22,3,6,13,14,17,21,22] transcriptomic studies,[11,23,24] and proteomic approaches.[7,25] However, most of these investigations have focused on whole coffee fruits (including both pericarp and seed) as well as on later developmental stages. Furthermore, when tissues are analyzed separately, quantifications often commence around 60 days after flowering, limiting the assessment of biochemical and metabolic processes in primordial tissues during early developmental stages.

Given the developmental cycle of coffee fruits, understanding how biochemical compounds are distributed and vary across the different developmental phases is essential. In this context, metabolomic approaches are valuable as they enable the investigation and understanding of complex metabolic and physiological processes. Therefore, to enhance the comprehension of the deposition of biochemical components essential for bean quality during different fruit developmental stages, this study aimed to (i) analyze the variation in essential carbohydrates, including reducing sugars, sucrose, starch, and total sugars, and (ii) identify and investigate differentially abundant metabolites using gas chromatography–mass spectrometry (GC-MS) at different developmental stages of *C. arabica* fruits, focusing on high sensory quality cultivars categorized into different groups across two environments.

MATERIALS AND METHODS

Plant Material, Experimental Design and Fruit Collection. In this experiment, two environments were selected, located in Ibiá and Patrocínio (19º27’31.90”S, 46º18’23.16”W and 18º59’36.20”S, 46º59’13.70”W, respectively), as part of the “Unidades Demonstrativas do Cerrado Mineiro” project implemented by the Minas Gerais Agricultural Research Agency (EPAMIG). This project was implemented with the aim of evaluating the performance of coffee cultivars developed by the EPAMIG Breeding Program and, subsequently, making available cultivars that combine high productivity with other characteristics, including quality.

The location selection was based on differences in altitude (1164 m in Ibiá and 987 m in Patrocínio) and the sensory performance of the cultivars MGS Paraíso 2 and MGS Catiguá 3, analyzed over the 2019 - 2021 triennium. These cultivars, derived from the parents Catuaí Amarelo and Timor Hybrid, presented high sensory quality in both environments. However, they showed distinct total scores according to the SCA methodology,[26] with MGS Paraíso 2

achieving the highest averages, while MGS Catiguá 3 had the lowest in both locations (Figure 1). Therefore, these cultivars are ideal for analyzing the levels of chemical compounds in different environments.

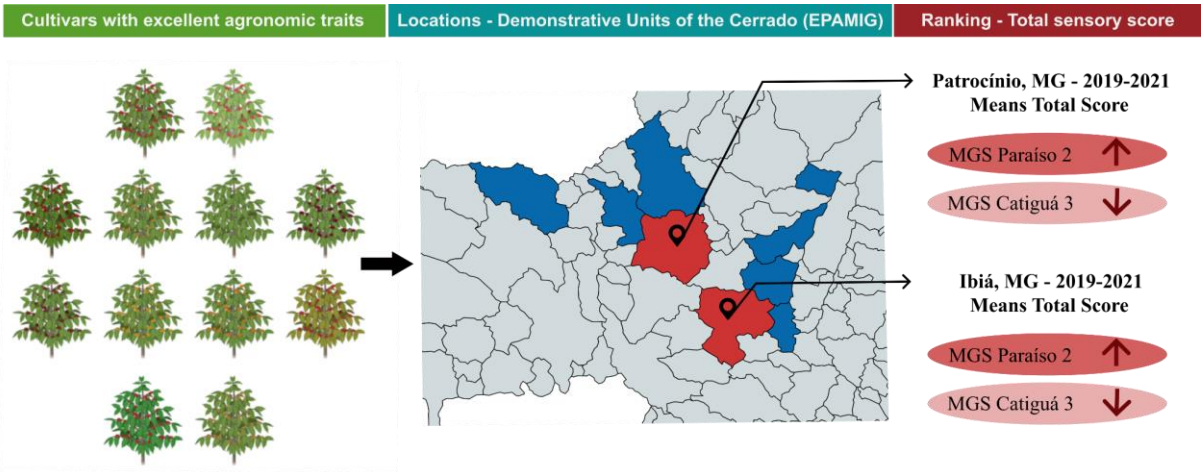


Figure 1. Representative diagram of the selection criteria for the locations Ibiá and Patrocínio, MG, and the cultivars MGS Paraíso 2 and MGS Catiguá 3, which provided plant material for the analyses. In blue, are the other locations of the experiments in the "Unidades Demonstrativas do Cerrado Mineiro" project.

Fruits from the cultivars MGS Paraíso 2, MGS Catiguá 3, and Gueisha (the latter used as a positive control for quality) were collected in Patrocínio, MG (Location 1). In Ibiá (Location 2), fruits were collected only from the cultivars MGS Paraíso 2 and MGS Catiguá 3, as no Gueisha plants were presente at this location. Sampling took place between November 2022 and July 2023. Samples covered seven different developmental stages (D), with collections conducted at 30 (DS1), 60 (DS2), 90 (DS3), 120 (DS4), 150 (DS5), 180 (DS6) and 270 (DS7) days after flowering (DAF), stages considered critical for coffee fruit development, as described by De Castro e Marraccini[4] and Li et al.[2]. For each cultivar, a completely randomized design with three replicates was adopted, each replicate consisting of six plants, and the fruits from each plot were pooled. Sampling was carried out at 10 a.m. in both locations, on the side of the plants opposite to the sun, using branches from the middle third of the plants. Selection criteria for the fruits included the absence of diseases, insect infestation, and uniformity of the fruits.

After each collection, the fruits were immediately frozen in liquid nitrogen and stored at -80 °C. Subsequently, seed tissues (aqueous perisperm at 30 DAF, DS1, to mature endosperm at 270 DAF, DS7) and the pericarp of the fruits were separated and independently used for extractions. Samples of the different tissues of *C. arabica* fruits were separately ground in a mortar and pestle using liquid nitrogen until a fine powder was obtained. Subsequently, these samples were then used for biochemical and metabolic analyses and for dry mass estimating. Dry weight was measured after 48 hours of drying in an oven at 70 °C. Based on the differences observed in the biochemical analyses, five seed developmental stages were selected for metabolic analyses: 30, 90, 120, 150, and 270 DAF (DS1, DS3, DS4, DS5, and DS7, respectively).

Biochemical Analysis. Extractions for the quantification of reducing sugars (RS), sucrose (SUC), starch content (STC), and total soluble sugars (TS) followed the protocol described by López-hidalgo et al.[27], with some adaptations. A total of 350 µL of 100 % ethanol was added to 50 mg of plant material, followed by vortex mixing and incubation in a water bath at 70 °C for 20 minutes. The material was then centrifuged at 22.000 x g and 4 °C for 5 minutes, and the supernatant was transferred to new Eppendorf tubes. This procedure was repeated twice using 80 % and 50 % ethanol. In the end, 1,050 µL of extract was obtained, which was used for the quantification of RS, TS, and SUC extraction. The resulting pellet was used for STC extractions.

For STC extractions, an adapted protocol by Hendricks et al.[28] and Zanandrea et al.[29], was employed. A total of 350 µL of sodium hydroxide 0.1 M was added to the pellet, 70 µL of acetic acid 1 M, followed by vortex mixing for 30 seconds, and 100 µL of amyloglucosidase (1mg/ml of 200 mM potassium acetate buffer pH 4.8), was added with gentle mixing. The tubes were then incubated at 40 °C for 2 hours. The resulting extract was stored at -20 °C until quantification.

For SUC extraction, the protocol by Van Handel,[30] was followed with the following adjustments: 50 µL of ethanolic extract was added to 50 µL of potassium hydroxide (30 %). After homogenization by inversion, the mixture was incubated in a thermocycler (Multigene Gradient – Labnet) at 40 °C for 15 minutes. The extract was then stored at -20 °C. Quantifications of SUC, TS, and STC were performed using the Anthrone method,[31] whereas RS was quantified using the DNS (dinitrosalicylic acid) method.[32] Carbohydrate concentrations were expressed in mmol or µmol glucose g^{-1} DW (dry weight).

GC-MS Analysis. *Pre-treatment.* The extraction was performed following the protocol proposed by Lisec et al.[33], with some adaptations. A 50 mg sample of plant material in DS5 and 25 mg sample of plant material for the other DS ware mixed with 700 µL of 100% methanol, followed by vortex mixing. Then, 30 µL of ribitol (0.4 mg/mL in deionized water) was added. The samples were shaken for 15 minutes at 70 °C in a thermomixer. After mixing, the samples were centrifuged at 4 °C for 10 minutes at 11,000 g, and the supernatant was transferred to a new tube. Subsequently, 375 µL of chloroform and 750 µL of deionized water were added, followed by vortex mixing. The samples were then centrifuged for 15 minutes at 11,000 g, and 150 µL of the supernatant were transferred to a 1.5 mL reaction tube and subjected to vacuum centrifugation to evaporate the extract before the derivatization step.

Derivatization. To the dry residue, 20 µL of methoxyamine reagent (20 mg/mL in pyridine) was added, and the mixture was shaken at 37 °C in a thermomixer for 2 hours for derivatization. Subsequently, 35 µL of N-methyl-N-(trimethylsilyl) trifluoroacetamide (MSTFA) was added, and the sample was shaken again at 37 °C for an additional 30 minutes to promote silylation. Afterward, the samples were stored at -20 °C until analysis by GC-MS.

Analytical Conditions and Compound Identification. A 1 µL aliquot of the derivatized sample was injected into a GCMS-QP2010plus system (SHIMADZU, Kyoto, Japan) in split mode 1:50. The separation of derivatives was performed using a 30 m x 0.25 mm x 0.25 µm internal diameter SLBTM-5MS silica capillary column (Sigma Aldrich), with high-purity helium (≥99.99 %) as the carrier gas at a flow rate of 0.97 mL/min. The initial inlet temperature was set at 80 °C for 5 minutes, then increased to 310 °C at a rate of 10 °C/min. Mass spectra were obtained using a full scan monitoring mode over a mass scan range of 70-600 m/z with a scan event time of 0.20 s. The injector temperature was set to 230 °C, and the ion source and interface temperatures were 200 °C and 250 °C, respectively.

At the GC-MS output, peak integration was performed using the GCMSsolution Update Program v4.41, GCMS Postrun Analysis (SHIMADZU, 2010), focusing on peaks with a slope greater than 1000 min^{-1} and a width greater than 3 s. The identification of target compounds was based on a combination of mass spectrum matching between sample spectra and the spectral libraries available in the GC-MS data system, including FFNSC13 and WILEY139, as well as retention time and mass-to-charge ratio of molecules/fragments. Mass spectra of unknown compounds were interpreted based on their fragmentation patterns. Ribitol was used as an internal standard for data normalization.

Statistical Analysis. For the biochemical data, Generalized Linear Models (GLM) with a Gamma distribution and logarithmic link function were used. The effects evaluated included location, cultivars (within each location), developmental stage (DS), and tissue, as well as the interactions between these factors. When an interaction was significant, its breakdown was performed to compare cultivars at each location. At Location 2, with only two cultivars, the F-test from the deviance analysis was sufficient for the conclusion. At Location 1, after identifying significant differences with the F-test, the t-test was applied to compare the means of the three cultivars, using the lsmeans command from the GLIMMIX procedure in SAS. Spearman's correlation was performed to assess the relationship between RS quantification by the DNS method and the conventionally used calculation, which considers the difference between TS and SUC determinations, as these are different quantification methods for physiologically correlated compounds.

In the semi-quantitative GC-MS data, after analytical processing, the software MetaboAnalyst v6.0[34] was used to perform partial least squares discriminant analysis (PLS-DA)[35] and to obtain the variable importance in projection (VIP) scores. The Kruskal-Wallis test[36] was used to identify significant variances because it is a non-parametric method suitable for comparing multiple groups without assuming a normal distribution. Following this, the Dunnett test[37] was conducted for multiple comparisons to specifically identify which groups showed significant differences. Differentially abundant metabolites (DAMs) were selected based on a p-value < 0.05 and a log2 fold change (log2FC) ≥ 1 in any contrast. Data analysis and visualization were performed using R version 4.3.2, with the corrplot, and PerformanceAnalytics [38] packages for correlation analysis.

RESULTS

Evaluation of Reducing Sugars, Sucrose, Starch, and Total Sugars Contents. The levels of reducing sugars (RS), sucrose (SUC), starch (STC), and total sugars (TS) at different development stages (DS) of *Coffea arabica* fruits were assessed through chemical analyses. The quantification of RS using the dinitrosalicylic acid (DNS) method showed a high correlation with the data obtained by calculating the difference between TS and SUC concentrations (a commonly used method) (Figure S1). This validates the results obtained through different methodologies, considering physiologically related compounds.

Despite the established solute translocation through vascularization between the pericarp and seed[3], the sugar contents in the tissues of the analyzed cultivars did not show

coordinated variation (Figures 2 and S2, respectively). This independent behavior is likely due to the dynamic balance between biosynthesis, import, export, and metabolism of sugars. Whereas these processes occur in both tissues, they may involve intrinsic mechanisms specific to each tissue,[3] that were not evaluated in this study. Therefore, our analyses and inferences focus on the tissues directly involved in fruit development.

Overall, the cultivars exhibited similar sugar content patterns during seed development, regardless of the evaluated location. In the intermediate stages (DS3 and DS4), higher levels of RS, SUC, and TS were observed (Figures 2A-D, G, H), likely due to the critical period of fruit filling and seed expansion at 90 and 120 DAF[2], which requires an increased amount of sugars as precursors for storage compounds.

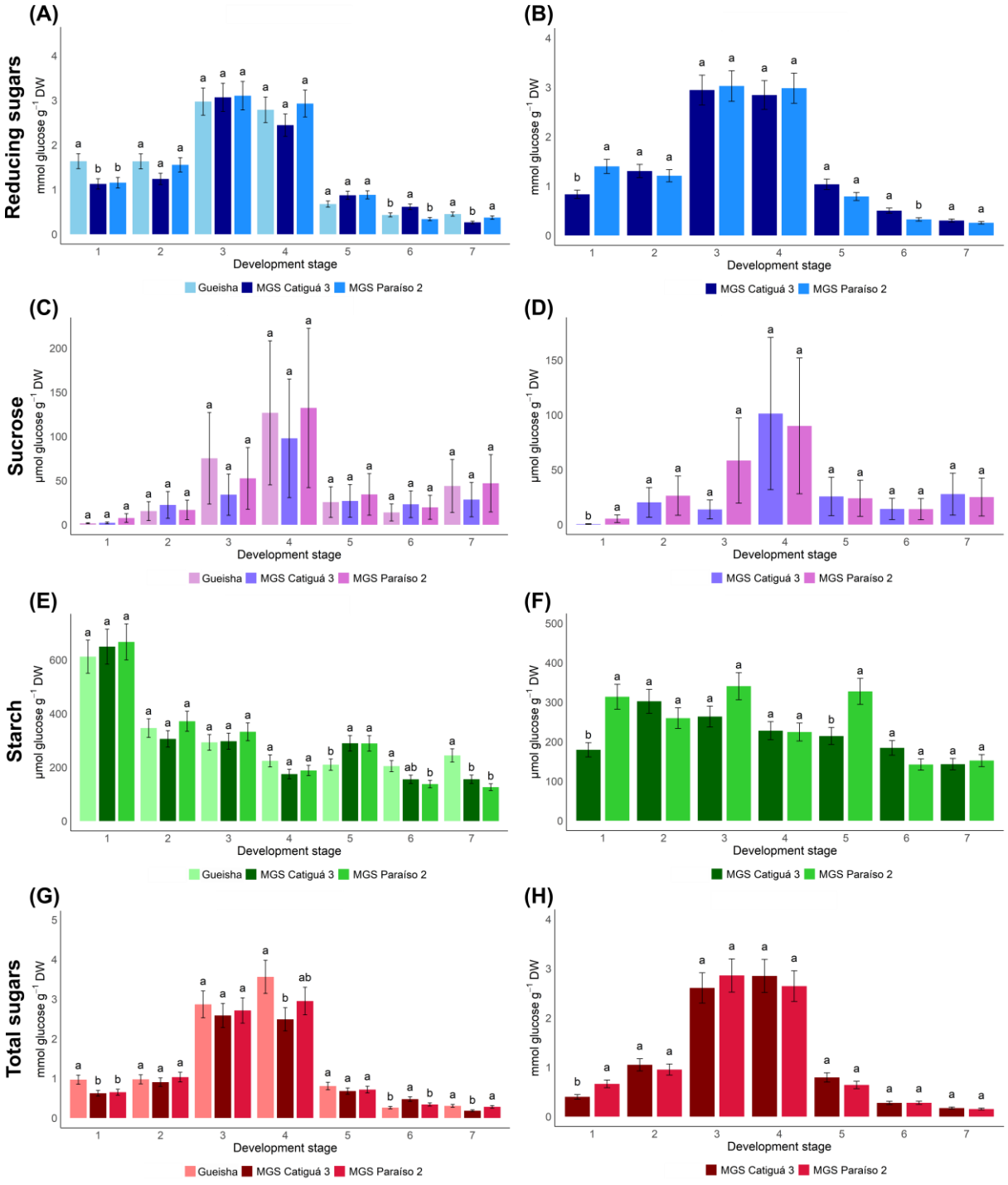


Figure 2. Carbohydrate content in the seeds (separated from pericarp) of fruits from different *Coffea arabica* cultivars: MGS Paraíso 2, MGS Catiguá 3, and Gueisha at location 1 (A, C, E, and G), and MGS Paraíso 2 and MGS Catiguá 3 at location 2 (B, D, F, and H). Evaluations were performed across seven developmental stages (DS), from aqueous perisperm at 30 DAF (DS1) to mature endosperm at 270 DAF (DS7). RS (A and B), SUC (C and D), STC (E and F), and TS (G and H). Different letters indicate significant differences ($p < 0.05$) among cultivars within each DS. The vertical bar represents the standard error of the mean.

For RS (Figure 2A), the cultivars differed in DS1, DS6, and DS7 at location 1. The Gueisha cultivar presented the highest levels in DS1, whereas Gueisha and MGS Paraíso 2 showed higher levels in DS7. At location 2 (Figure 2B), the MGS Paraíso 2 cultivar exhibited the highest levels in DS1. The MGS Catiguá 3 cultivar stood out in DS6, showing the highest RS levels at both locations. Regarding SUC (Figure 2C, D), the cultivars differed significantly only in DS1 at location 2, with MGS Paraíso 2 presenting the highest SUC content. For TS (Figure 2G, H), the most expressive differences occurred at location 1. In DS1, DS6, and DS7, the behavior was the same as observed for RS, with the cultivars differing in the same way. On the other hand, in DS4, the Gueisha cultivar exhibited the highest levels compared to MGS Catiguá 3. At location 2, only in DS1 did the MGS Paraíso 2 cultivar present a higher sugar content.

The highest STC levels were observed in DS1 at location 1 (Figure 2E). The cultivars MGS Paraíso 2 and MGS Catiguá 3 differed from Gueisha in DS5 and DS7. Whereas MGS Paraíso 2 and MGS Catiguá 3 presented higher STC concentrations in DS5, the cultivar Gueisha exhibited a higher content in DS7. In DS6, Gueisha only differed from MGS Paraíso 2. At location 2, differences in STC content occurred in DS1 and DS5, with MGS Paraíso 2 presenting the highest concentrations (Figure 2F). Thus, it is observed that the differences among cultivars were more evident in DS1, considering the set of all analyzed sugars.

Metabolic Profiling of *Coffea arabica* Seeds at Different DS. A total of 27 compounds were identified in the seeds (separated from pericarp) of different *C. arabica* cultivars across five DS, and their relative concentrations were determined (Figure 3, additional details in Table S1). These compounds include seven sugars (D-ribose, D-fructose, galactose, glucose, gulose, melibiose, and sucrose), five organic acids (acetic, citric, malic, quinic, and propanoic acids), four polyols (deoxiinositol, D-mannitol, inositol, and myo-inositol), three amino acids (alanine, glycine, and proline), three glycerol esters (octadecanoic acid, monopalmitin, and monostearin), three polyphenols (caffeic acid, caffeoyl quinic, and silane-hydroquinone), and two fatty acids (stearic and palmitic acids).

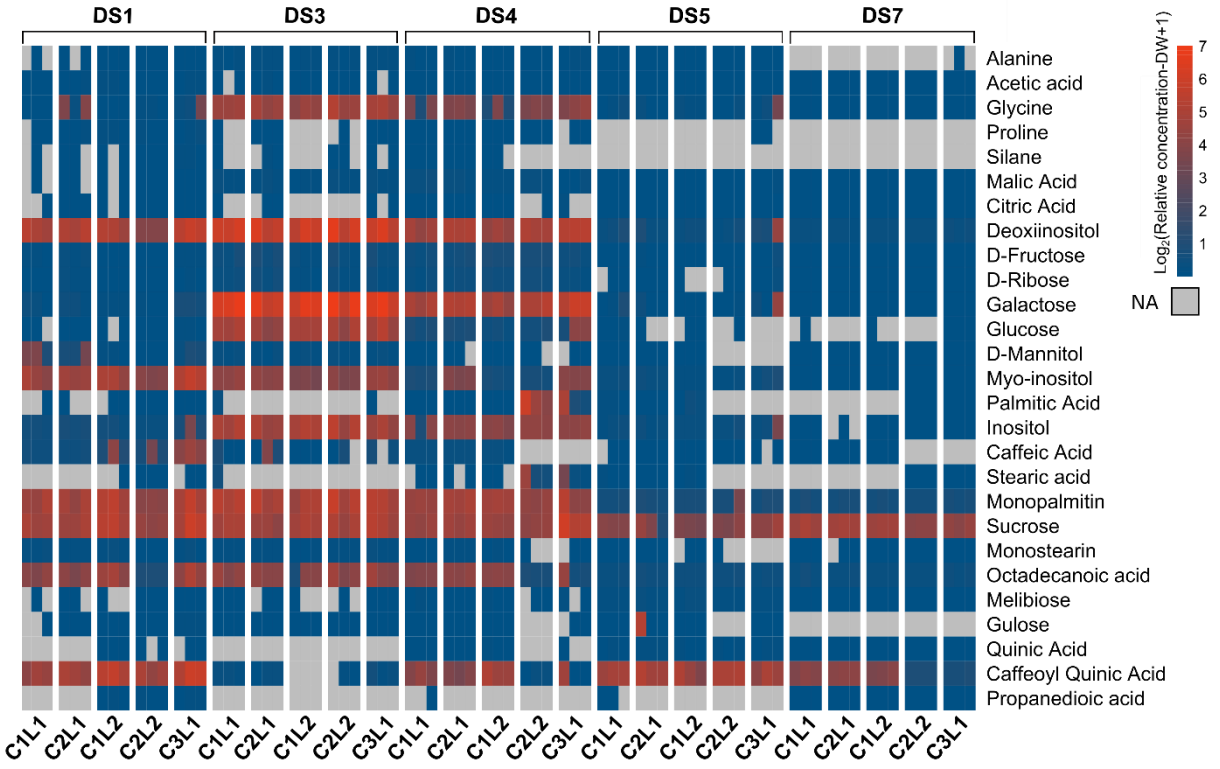


Figure 3. Heatmap of the semiquantitative data for 27 metabolites identified by GC-MS analysis in fruits of different *Coffea arabica* cultivars across five DS. C1: MGS Paraíso 2, C2: MGS Catiguá 3, C3: Gueisha; L1: Patrocínio; L2: Ibiá. The color bar indicates the relative abundance of metabolites according to the Ribitol standard, with higher abundance represented in red and lower abundance in blue. Gray areas (NA) indicate metabolites that were not detected. Each column represents a biological replicate.

Overall, the concentrations of the compounds were higher in the initial stages (DS1, DS3, and DS4), with a declining trend in the later stages (DS5 and DS7). Notable exceptions were quinic and propanoic acids, whose concentrations increased in the final stages. The levels of acetic and malic acids, D-fructose, D-ribose, D-mannitol, melibiose, monostearin, and sucrose remained relatively constant throughout the DS. Silane (hydroquinone) was not detected in DS5 and DS7, whereas proline in these same stages was only observed in the Gueisha cultivar during DS5.

Analysis of Key Metabolites. The set of identified metabolites was used as input for PLS-DA analysis (details of cross-validation for model accuracy and prediction are described in Table S2) and the generation of VIP scores, considering each cultivar at each location: MGS Paraíso 2, MGS Catiguá 3, and Gueisha at location 1 (Figure 4A, B, C) and MGS Paraíso 2 and

MGS Catiguá 3 at location 2 (Figure 4D, E). The first two components explained over 90 % of the observed variation in the metabolites of each cultivar across the different DS.

At location 1 (Figure 4A, B, C), when considering component 1, greater proximity is observed between DS3 and DS4, with overlap in MGS Catiguá 3 and Gueisha. Additionally, DS5 and DS7 also overlap in MGS Paraíso 2 and Gueisha. The overlap of points or confidence intervals indicates that the metabolic profiles do not differ significantly, being more specific in these DS. From the perspective of component 2, clear separation of DS1 is observed, which is more distant from the others, indicating a difference in the metabolic composition related to this DS. At location 2 (Figure 4D, E), the same behavioral trend is observed, except for DS3 in MGS Catiguá 3 (Figure 4E), where the points are more distant and overlap with DS4 and DS7.

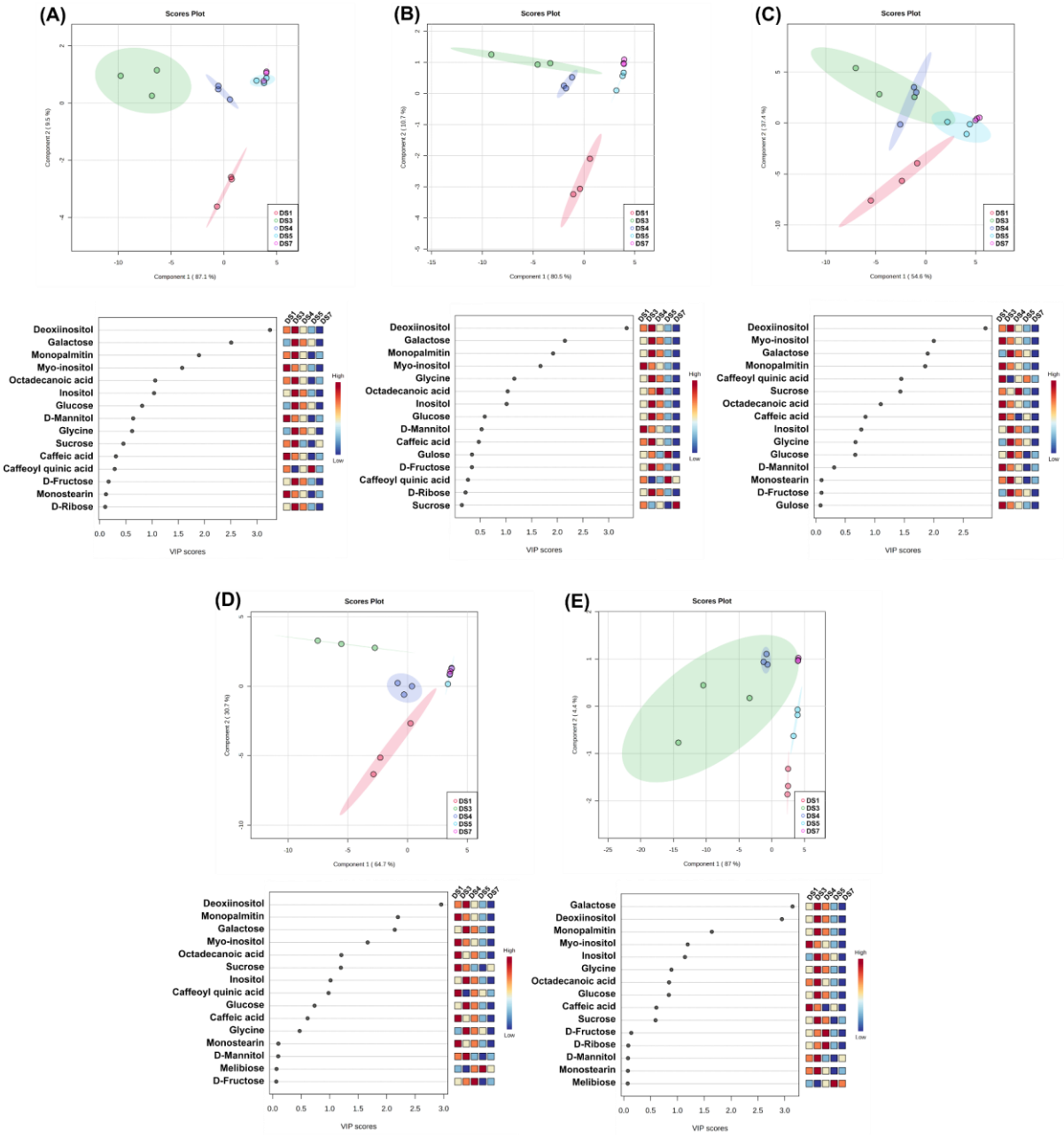


Figure 4. PLS-DA scores plot and VIP scores for 27 metabolites identified in five developmental stages (DS) of fruits from different *Coffea arabica* cultivars in two locations. Cultivars: MGS Paraíso 2 (A), MGS Catiguá 3 (B), and Gueisha (C) at location 1; MGS Paraíso 2 (D) and MGS Catiguá 3 (E) at location 2. DS1: red; DS3: green; DS4: dark blue; DS5: light blue; and DS7: pink. The area around the points in the PLS-DA represents the 95% confidence interval. The colored boxes to the right of the VIP scores indicate the relative concentrations of the corresponding metabolites for each cultivar.

Variables Important in Projection (VIP) were used to identify the metabolites that significantly contributed to the discrimination of the different DS in each cultivar. Compounds with VIP > 1.0 are considered substantial contributors to this discrimination. The main metabolites identified, based on VIP values for all cultivars across the two locations, were deoxyinositol, myo-inositol, monopalmitin, and galactose, which showed the highest contents during the early DS stages. Additionally, inositol and octadecanoic acid also had VIP > 1.0, except for Gueisha (Figure 4C) and MGS Catiguá 3 at location 2 (Figure 4E), respectively. For Gueisha (Figure 4C) and MGS Paraíso 2 (Figure 4D) at location 2, sucrose and caffeoylquinic acid were also substantial contributors, whereas glycine stood out as an important variable for MGS Catiguá 3 at location 1 (Figure 4B).

Correlation Analysis. To better understand the degree of correlation between metabolites in each cultivar across both locations, Spearman correlation analysis was conducted, considering the 27 identified compounds and the five DS (Figure 5). The positive and negative correlations among metabolites showed similar patterns across cultivars in both locations. Notably, strong significant correlations (> 0.8 and < -0.8) were predominantly positive, especially in the group formed by the sugars D-ribose, D-fructose, galactose, glucose, gulose, and melibiose, followed by the esters octadecanoic acid, monopalmitin, and monostearin, as well as the polyols inositol and myo-inositol. Additionally, the amino acids alanine and glycine, caffeic acid, and deoxiinositol also showed strong positive correlations.

On the other hand, the most significant negative correlations involved the organic acids citric, quinic, and propanoic, as well as caffeoylquinic acid and sucrose. Among the cultivars, MGS Catiguá 3 presented the highest number of strong correlations, both positive and negative, at both locations (Figure 5B, E), compared to MGS Paraíso 2 (Figure 5A, D) and Gueisha (Figure 5C). Additionally, sucrose showed more positive correlations in the Geisha cultivar than in the other cultivars in the location 1.

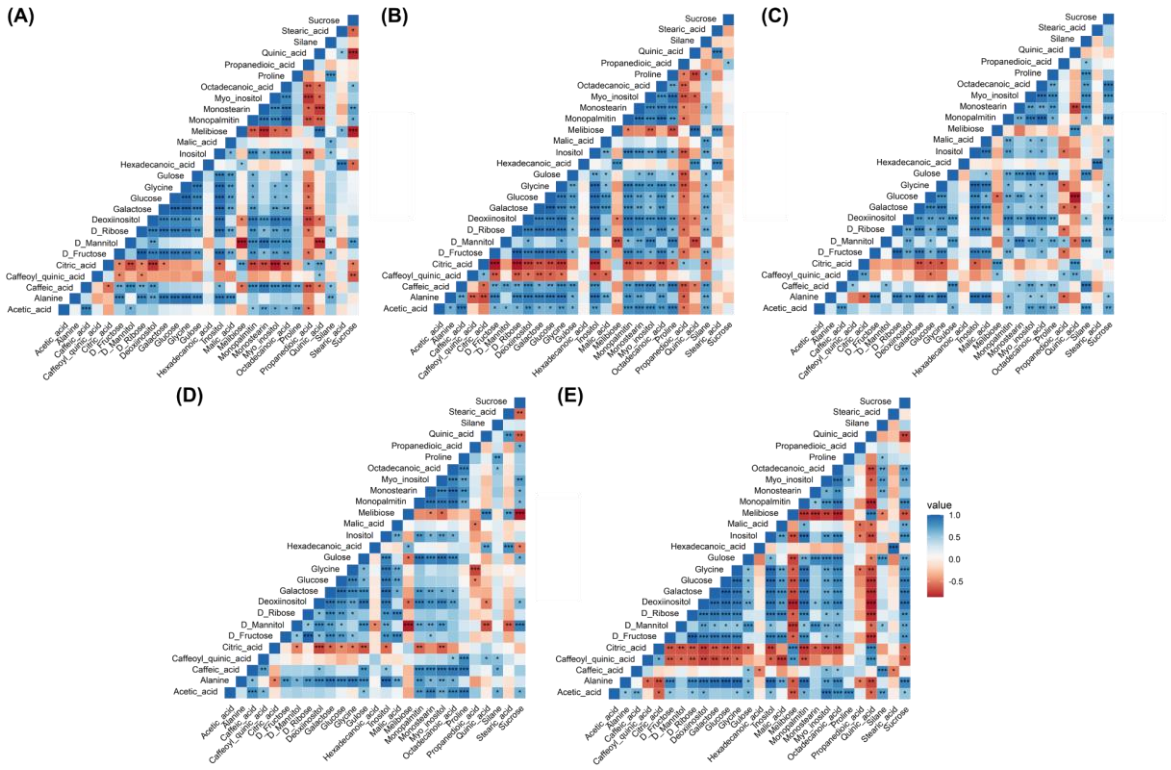


Figure 5. Spearman correlation analysis for 27 metabolites identified in five developmental stages (DS) of seeds from different *Coffea arabica* cultivars in two locations. Cultivars: MGS Paraíso 2 (A), MGS Catiguá 3 (B), and Gueisha (C) at location 1; MGS Paraíso 2 (D) and MGS Catiguá 3 (E) at location 2. According to Spearman correlation coefficient, (*) $p < 0.05$, (**) $p < 0.01$, and (***) $p < 0.001$.

Differentially Abundant Metabolite (DAM) Analyses Within Each DS. To identify DAMs in each DS, comparisons of relative metabolite levels among cultivars were performed, using statistical significance (p-value < 0.05) as the selection criterion for the compounds compared. Considering all the DS investigated at location 1, the results revealed 9 DAMs between MGS Paraíso 2 vs. MGS Catiguá 3 (3 upregulated and 6 downregulated) (Figures 6A, S3A, D, G, J, and Table S3); 25 DAMs significantly different between MGS Paraíso 2 vs. Gueisha (4 upregulated and 21 downregulated) (Figures 6B, S3B, E, H, K, and Table S3); and 22 DAMs were observed between MGS Catiguá 3 vs. Gueisha (5 upregulated and 17 downregulated) (Figures 6C, S3C, F, I, L, and Table S3). At location 2, 15 DAMs were identified, 4 downregulated and 11 upregulated (Figures 6D, S4, and Table S3).

The higher number of downregulated DAMs at location 1 suggests a greater demand for or a less pronounced effect of these compounds in the metabolic activities of the cultivars MGS Paraíso 2 and MGS Catiguá 3 during fruit development at this location, particularly when

compared to the control cultivar Gueisha. This difference is most evident in DS1, where the highest number of DAMs was identified: 14 DAMs between MGS Paraíso 2 vs. Gueisha (1 upregulated and 13 downregulated) and 9 downregulated DAMs between MGS Catiguá 3 vs. Gueisha (Figures 6B, C, respectively). Among these DAMs, noteworthy compounds include the sugars gulose, galactose, glucose, and melibiose, followed by the organic acids malic, acetic, and citric, as well as the polyphenols caffeic acid, silane, and caffeoylquinic acid. Also downregulated in this DS were the amino acids glycine and proline, and the polyol inositol. These results indicate distinct metabolic regulation between the cultivars MGS Paraíso 2, MGS Catiguá 3, and the control cultivar Gueisha at this stage.

In the other DS at location 1, a reduction in DAMs was observed in DS3, followed by an increase in DS4 (Figures S3A–F), and subsequently a decrease in DS5 and DS7 (Figures S3G–L). At location 2, although the differences were less pronounced, DS1 also showed the highest number of DAMs (Figure 6D), with a predominance of positive regulation. These results, highlighting more pronounced differences among metabolites in DS1, align with the biochemical analyses, which also indicated the highest number of significant differences at this stage (Figure 2).

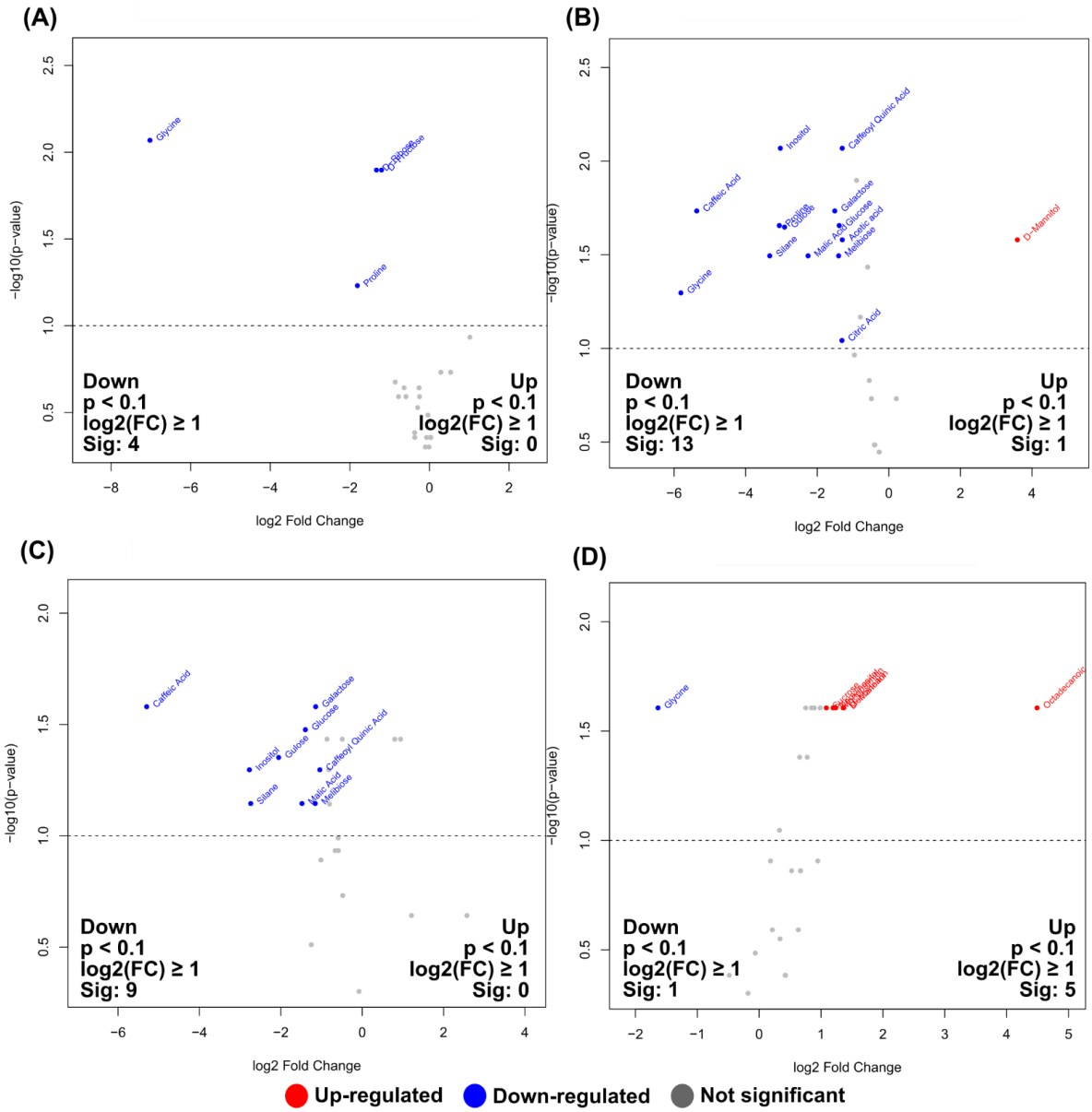


Figure 6. DAM analysis in *Coffea arabica* fruit at DS1. Volcano plots of DAMs in the comparisons: MGS Paraíso 2 vs. MGS Catiguá 3 (A), MGS Paraíso 2 vs. Gueisha (B), and MGS Catiguá 3 vs. Gueisha (C) at location 1; and MGS Paraíso 2 vs. MGS Catiguá 3 (D) at location 2. Each point in the plots represents a metabolite, with the x-axis indicating the log2(FC) value and the y-axis representing the -log10(p-value). Red points correspond to positively regulated metabolites ($p < 0.1$ and FC ≥ 1), blue points represent negatively regulated metabolites ($p < 0.1$ and FC ≤ -1), and gray points indicate metabolites with no significant differences.

Throughout the DS at location 2, a predominance of upregulated DAMs was observed, with an absence of DAMs in DS3 (Figure S4A), suggesting greater synchronization of

metabolic processes at this stage. In DS1, MGS Paraíso 2 showed higher levels of sucrose, octadecanoic acid, deoxiinositol, monopalmitin, and D-mannitol (Figure 6D). The higher sucrose content for MGS Paraíso 2 in DS1 at location 2 was also observed in the biochemical analyses (Figure 2D). The differential levels of these compounds suggest a metabolic adjustment that may favor energy storage[39,40] and the accumulation of protective metabolites[5,41] in MGS Paraíso 2.

Curiously, at location 1, among all differentially expressed compounds (Table S3), only D-fructose and D-ribose (Figures 7A and B, respectively) did not show significant differences between MGS Paraíso 2 and Gueisha, whereas significantly different from MGS Catiguá 3 only in DS1. For other compounds, significant differences were also observed among cultivars, but these variations did not follow this pattern, showing different combinations among the cultivars for other compounds. D-fructose is an important carbon and energy source in plant metabolism,[40] while D-ribose is essential for nucleotide synthesis and energy metabolism. This behavior was exclusively observed in DS1. The similarity between MGS Paraíso 2 and Gueisha for these specific compounds, along with the other differences identified in DS1, suggests that this early stage may involve fine metabolic adjustments and complex regulatory processes that are crucial for the final quality of the fruit.

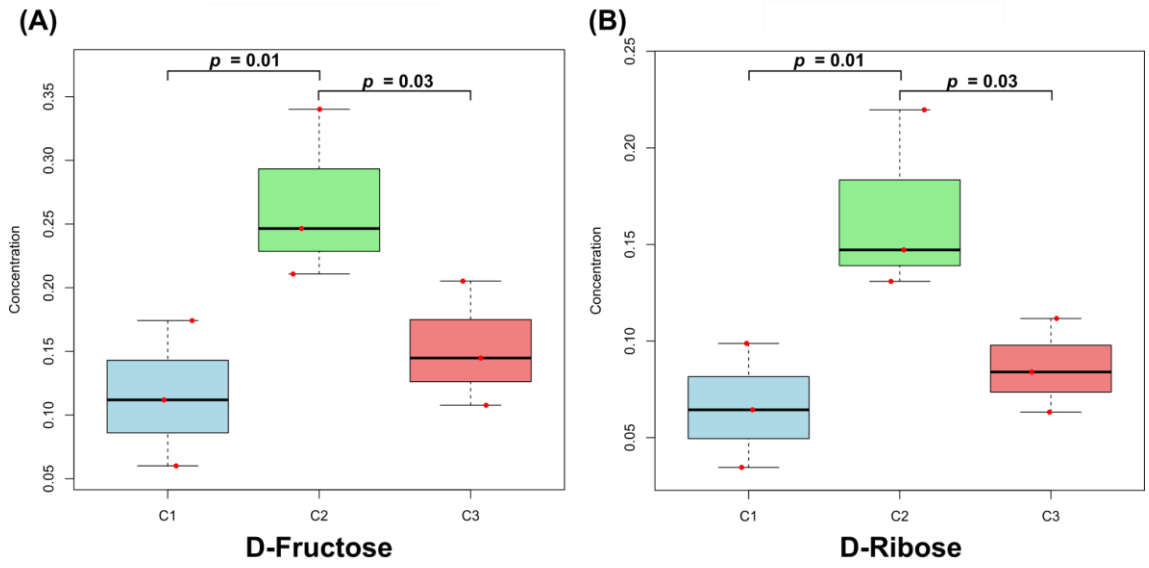


Figure 7. Box plots of the different levels of the metabolites D-fructose and D-ribose ($p < 0.05$) in the fruits of *Coffea arabica* cultivars at DS1 in location 1. C1: MGS Paraíso 2, C2: MGS Catiguá 3, and C3: Gueisha.

DISCUSSION

Investigations into the chemical composition of coffee fruits, aiming to understand the mechanisms of chemical component deposition in the fruits, are extensive.[1–6] These studies include research associating these compounds with fruit and cup quality. [1–4,7,11] Considering the developmental cycle of *Coffea arabica* fruits, this study identified 27 metabolites and demonstrated how different sugars and other differentially abundant metabolites (DAMs) behave throughout developmental phases, highlighting intracultivar variations in high-quality cultivars (Figures 2, 6, 7, S3, S4).

The results indicate a distinct distribution of compounds during seed development (Figure 3), with the identification of metabolites belonging to the classes of carbohydrates, lipids, organic acids, amino acids, and polyphenols. Sucrose, due to its biological relevance in plants,[42,43] was detected at all DS, corroborating previous studies.[12,22] However, curiously, the biochemical data show a decrease in sucrose levels in DS5, DS6, and DS7 (Figure 2C, D), differing from findings reported in the literature. Rogers et al.,[22] observed a different pattern among *Coffea arabica* cultivars, although sucrose levels still exhibited an upward trend during intermediate and late stages of development.

Some hypotheses may explain the distinct pattern observed in the cultivars of this study. Sucrose may have been directed towards the biosynthesis of cell wall polysaccharides, which undergo a critical phase of deposition during intermediate and late stages.[1,2] These polysaccharides exhibit higher levels compared to sucrose during these stages.[5] Additionally, the deposition of 11S storage proteins, which reaches maximum levels at intermediate stages,[7,24] may have also contributed to the reduction in sucrose levels. At these stages, an increase in chlorogenic acid accumulation was also observed (Figure 3), with sucrose serving as a precursor in its biosynthetic pathway.[1] Further studies may reveal the existence of intracultivar variability in sucrose accumulation throughout fruit development, as previously observed in other species.[44,45]

An asynchronous behavior between the pericarp and seed was observed in the biochemical analyses of sugars and starch (STC), reflecting intrinsic yet complementary processes specific to each tissue. The pericarp, with a photosynthetic role, participates in photoassimilate exchange, whereas the seed primarily functions in storage.[4] During the seed developmental stages (DS), an increase in sugars was observed in the DS3 and DS4 (Figure 2), associated with the high energy demand for fruit filling resulting from seed growth.[2] At this stage, the endosperm expands significantly, occupying almost the entire fruit locule.[4,5]

STC, in turn, exhibited a distinct pattern, with an increase in DS1 only at location 1 and relative stability in later stages at both location 1 and location 2 (Figure 2E, F). Higher starch levels at early developmental stages were observed by Joët et al.,[5] due to the higher expression of plastidial enzymes involved in starch biosynthesis. However, this was not observed at location 2, where a smaller amount of sugars was stored as starch in DS1, indicating a strong interaction between genotypes and environments (G x E)[17,46] in the cultivars evaluated.

In the correlation analysis of the 27 metabolites, the cultivar MGS Catiguá 3 displayed the highest magnitude negative correlations, especially among organic acids (citric, quinic, and caffeic), caffeoylquinic acid, and sucrose with other compounds (Figure 5B, E), in contrast to the strong positive correlations between sugars and polyols. This pattern suggests that glycolysis intermediates may be directed toward anaplerotic reactions in other carbohydrate metabolic pathways, balancing sugar levels or even storing these carbon sources as polyols[47] (Figure 8). Polyols, observed in higher amounts in the early stages (Figure 3), may serve antioxidant functions,[48] particularly in the perisperm, due to the high water content in these early stages, which promotes the accumulation of reactive oxygen species (ROS).[4,49]

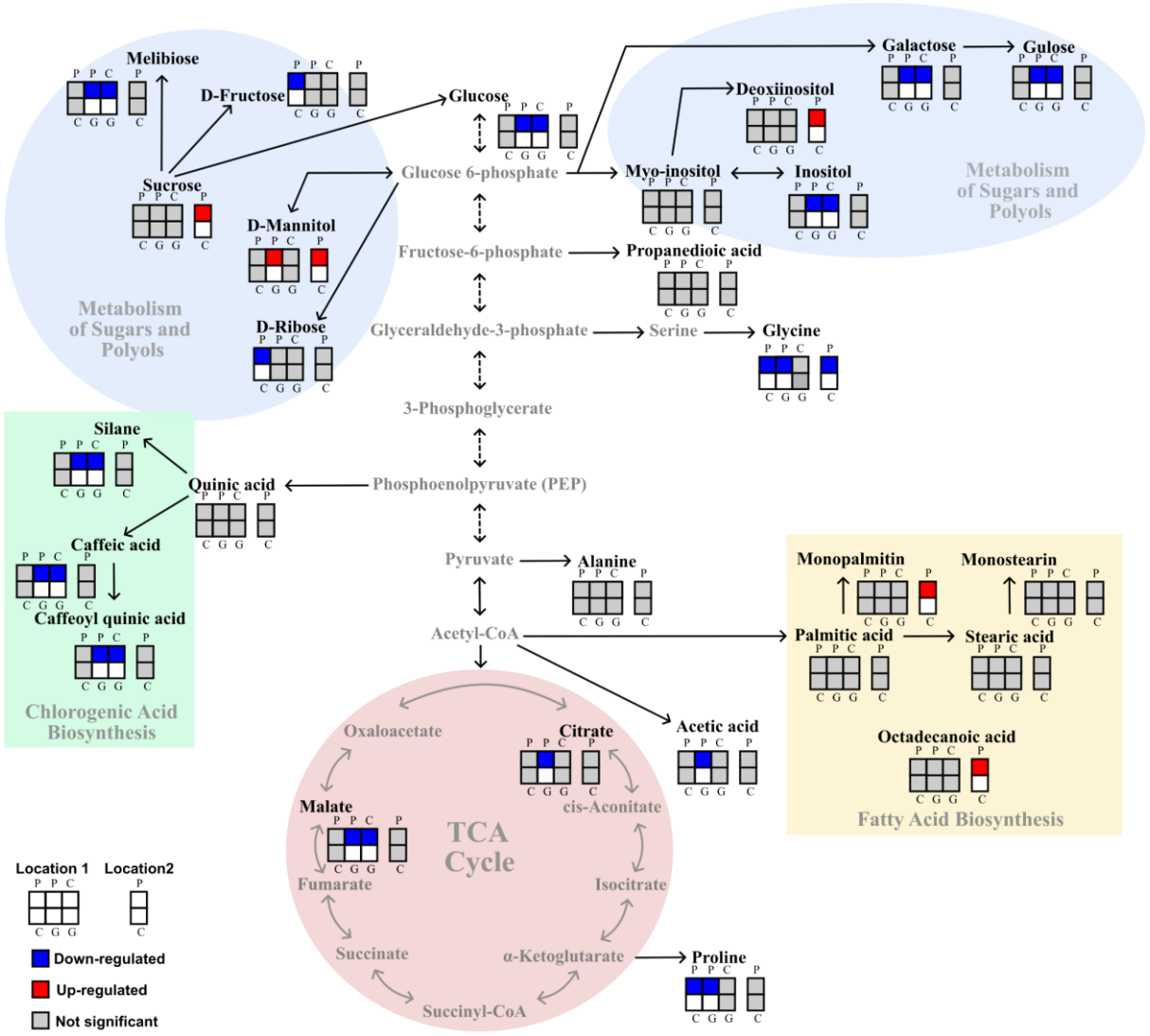


Figure 8. Simplified scheme representing the network of metabolic pathways involving the 27 compounds identified in different *Coffea arabica* cultivars from two locations. Metabolites shown in gray belong to the analyzed pathways but were not identified in the seed samples in this study. Key changes in DAMs in the perisperm at DS1 (30 DAF) are highlighted. For each metabolite, a color-coded table is presented with columns indicating pairwise comparisons between the cultivars: P, MGS Paraíso 2; C, MGS Catiguá 3; and G, Gueisha. Blue represents downregulated DAMs, red represents upregulated DAMs, and gray indicates metabolites that were not differentially abundant.

Interestingly, the cultivars MGS Paraíso 2 (Figure 5A, D) and Gueisha (Figure 5C) exhibited weaker correlations among metabolites. One possible hypothesis for this pattern is the existence of metabolic reprogramming, which generates a greater availability of precursor

compounds for glycolysis and the TCA cycle, thus providing intermediate compounds for both cataplerotic and anaplerotic reactions in MGS Paraíso 2 and Gueisha, instead of favoring one pathway at the expense of the other, as apparently occurs in MGS Catiguá 3. Organic acids that use glycolysis intermediates in their biosynthesis, such as quinic acid, play important roles, as this acid is an intermediate in the biosynthesis of chlorogenic acids, which are essential in combating reactive oxygen species (ROS).[4] Similarly, palmitic acid, involved in lipid biosynthesis, stands out for its importance in forming cellular membranes.[50] The progression of glycolytic pathways facilitates the generation of these compounds through their precursor intermediates (Figure 8).

The investigation of the metabolic profiles in each cultivar, through PLS-DA analysis, allowed for the clear distinction of the different seed developmental stages (DS) (Figure 4). DS1 stood out as the most distinct stage, while DS3 and DS4, as well as DS5 and DS7, showed closer proximity to each other. Compounds such as deoxiinositol, myo-inositol, monopalmitin, and galactose were important for the discrimination of the stages, according to VIP values, showing higher concentrations during the early and intermediate stages (Figure 3).

This early development period is characterized by high rates of cell division, associated with the development of the perisperm in DS1 and the endosperm in DS3 and DS4,[3–5,14] a phase when compounds related to primary metabolism are more actively synthesized in coffee seeds.[11] Therefore, the presence of deoxiinositol and myo-inositol in these stages is justified by the close association of polyols with primary metabolism.[47] Similarly, galactose exhibits higher levels during intermediate stages, playing an important role in grain filling.[22] Monopalmitin, a derivative of palmitic acid (Figure 8), also showed higher concentrations in the early stages, highlighting the role of lipids in the initial development of seeds.[50]

Chlorogenic acids play a fundamental role in the early stages of coffee development. In addition to their well-known antioxidant properties, it is hypothesized that they serve as a source of carbon skeletons for lignin synthesis during intermediate and final stages, contributing to the strengthening of cell walls.[14,51] In the PLS-DA results, caffeoylquinic acid was important for projection only in Gueisha and MGS Paraíso 2 at location 1 (Figure 4C, D), as was sucrose. This, in turn, supports the hypothesis that, in these cultivars, a greater availability of sucrose may have driven the synthesis of chlorogenic acids through anaplerotic reactions involving glycolysis-derived compounds (Figure 8). Stronger cell walls are important during roasting processes, as they help preserve essential compounds for sensory quality, such as organic acids.[52]

As observed in the DAM analyses at location 1 (Figures 6, 8, S3), the highest number of DAMs was identified in comparisons performed in DS1, with a greater number of downregulated DAMs in the MGS Paraíso 2 and MGS Catiguá 3 cultivars compared to the control cultivar Gueisha. It is important to note that MGS Paraíso 2 and MGS Catiguá 3 exhibit high productivity and greater tolerance to diseases such as coffee leaf rust,[53] in contrast to Gueisha. These notable differences in compounds involved in central metabolic pathways, such as energy generation, glycolysis, and the TCA cycle, may reflect optimization and greater synchronization of these pathways in the analyzed cultivars compared to the control cultivar Gueisha.

Curiously, the metabolic profile showed that DAMs were more numerous at locayion 1 compared to locaton 2 (Figures 6, S3, S4). This demonstrates the existence of specific metabolic regulations for each environment, highlighting the strong G × E interaction,[17,46,54] as observed in biochemical analyses. In location 1, located at a lower altitude (~1000 m) compared to location 2 (~1200 m), experiences higher temperatures and lower thermal amplitude, factors that directly affect energy metabolism.

The metabolic profile of seeds has recently been used to distinguish rice cultivars,[55] exemplifying the same intracultivar variation observed in this study. Rogers et al.[22] also reported variations in *C. arabica* cultivars regarding sugars, polyols, and organic acids, although at more advanced stages (84 DAF and 210 DAF). An interesting aspect of the results presented in this study is that the highest number of DAMs was identified at DS1, an early stage, which exerts a significant influence on the final stage of the fruits. Alves et al.[7] reported the presence of storage proteins at this early stage and emphasized that the development of the perisperm has a direct impact on the final grain size. This underscores the need for more in-depth investigations into DS1, especially in the perisperm, which may reveal relevant intracultivar differences for improving agronomic traits.

Regarding the perisperm, its fundamental role in the development of coffee seeds is well-recognized,[3,5,14] with precursors of compounds related to beverage quality being identified in the perisperm at 60 days after flowering (DAF). Other studies have detected these compounds as early as 30 DAF, focusing on specific compounds such as lipids,[50] or on the transcriptional level, as with raffinose,[56] the latter overlooking complementary regulatory levels. However, the results of this study highlight significant differences in metabolites at 30 DAF (DS1), suggesting a complex and synergistic metabolic control that manifests in the early perisperm.

In this context, the similarity in levels of D-ribose and D-fructose between the MGS Paraíso 2 and Gueisha cultivars, contrasting with MGS Catiguá 3, observed exclusively at DS1 (Figure 7), demonstrates significant differences in the early perisperm of higher-quality cultivars. D-ribose, a pentose that forms part of nucleic acids (DNA and RNA) and the nucleotide ATP, in lower quantities, may indicate faster cell division processes and higher energy demand in MGS Paraíso 2 and Gueisha, corroborated by the lower levels of D-fructose. D-fructose is a hexose that, in addition to serve as a substrate for starch and cellulose synthesis, acts as a signaling molecule and is involved in essential biological processes.[18,57] Transcriptomic studies, using whole fruits, have revealed differentially expressed genes primarily involved in nucleic acid synthesis and energy storage up to 180 DAF.[11] These results highlight the importance of further investigations into the perisperm to uncover mechanisms that may influence the future quality of grains and beverage in these cultivars.

The perisperm is a developed form of the nucellus, an active maternal tissue that significantly contributes to the nourishment of the developing seed and, in some species, can serve as the main storage tissue.[58,59] In wheat, genes related to sucrose transport are expressed in the nucellar projection and filial tissues, indicating a significant interaction in nutrient transfer.[60] In maize, the *Meg1* gene (*Maternally expressed gene 1*) is preferentially expressed in maternal tissues and regulates the differentiation of transfer cells, demonstrating the importance of maternal regulation.[61] The MADS29 transcription factor is essential for regulating the degradation of the nucellus and nucellar projection during rice seed development, impacting seed development and viability.[62] Given the above, it is likely that there is intravarietal variation in the primordial tissues of *C. arabica* seeds.

In conclusion, the cultivars MGS Paraíso 2, MGS Catiguá 3, and Gueisha exhibit significant differences in the compounds evaluated throughout the fruits development stages. Through GC-MS analyses, 27 compounds were identified. However, it was in the initial perisperm (DS1) that the differences between the cultivars were most evident in the biochemical analyses, where the highest number of DAMs was observed. These differences during seed development involve fundamental metabolic processes, such as glycolysis and the tricarboxylic acid cycle (TCA), which play crucial roles in energy generation and in the formation of precursors for the biosynthesis of compounds related to the final quality of coffee through anaplerotic and cataplerotic reactions. More detailed investigations of the initial perisperm, including intravarietal variations and evaluations of G x E interactions through integrated approaches involving transcriptomics and functional genomics in *C. arabica*, may reveal key

processes influencing grain quality in the different cultivars. Thus, it will be possible to broaden the understanding of the relationship between the perisperm, fruit development, and the final quality of the coffee bean.

ABBREVIATIONS USED

DAM, differentially abundant metabolites; DS, development stage; GC-MS, gas chromatography–mass spectrometry; RS, reducing sugars; STC, starch; SUC, sucrose; TCA, tricarboxylic acid cycle; TS, total sugars; VIP, variable important in projection.

ACKNOWLEDGMENTS

The authors would like to thank the Central of Analysis and Chemical Prospecting of the Federal University of Lavras, and Finep (Financiadora de Estudos e Projetos), for supplying the equipment and technical support for experiments involving chromatographic analyzes.

FUNDING SOURCES

This work was financially supported by the the “Coordenação de Aperfeiçoamento de Pessoal de Nível Superior” (CAPES) for the doctoral scholarship for G.E.F (88887.568926/2020-00); the “Conselho Nacional de Desenvolvimento Científico e Tecnológico” (CNPq), for scholarship research for A.C.J. under grant number 309005/2022-1, R.R.O. (370434/2024-2, Project 465580/2014-9), for F.M.A.G under grant number 317001/2021-3, and for G.R.C. under grant number 317634/2021-6; the “Instituto Nacional de Ciência e Tecnologia do Café” (INCT/Café) under the grant of the “Fundação de Amparo à Pesquisa do Estado de Minas Gerais” (FAPEMIG, CAP APQ 03605/17) and scholarship research for A.C.J. (APQ-03406-18); and “Fundação de Amparo à Pesquisa do Estado de São Paulo” (FAPESP, #21/06968-3).

REFERENCES

(1) Cheng, B.; Smyth, H. E.; Furtado, A.; Henry, R. J. Slower Development of Lower Canopy Beans Produces Better Coffee. *J. Exp. Bot.* **2020**, *71* (14), 4201–4214. https://doi.org/10.1093/jxb/eraa151.

(2) Li, Z.; Zhang, C.; Zhang, Y.; Zeng, W.; Cesarino, I. Coffee Cell Walls—Composition, Influence on Cup Quality and Opportunities for Coffee Improvements. *Food Qual. Saf.* **2021**, *5*, fyab012. https://doi.org/10.1093/fqsafe/fyab012.

(3) Geromel, C.; Ferreira, L. P.; Guerreiro, S. M. C.; Cavalari, A. A.; Pot, D.; Pereira, L. F.

P.; Leroy, T.; Vieira, L. G. E.; Mazzafera, P.; Marraccini, P. Biochemical and Genomic Analysis of Sucrose Metabolism during Coffee (*Coffea Arabica*) Fruit Development. *J. Exp. Bot.* **2006**, *57* (12), 3243–3258. https://doi.org/10.1093/jxb/erl084.

(4) De Castro, R. D.; Marraccini, P. Cytology, Biochemistry and Molecular Changes during Coffee Fruit Development. *Braz. J. Plant Physiol.* **2006**, *18*, 175–199. https://doi.org/10.1590/S1677-04202006000100013.

(5) Joët, T.; Laffargue, A.; Salmona, J.; Doulbeau, S.; Descroix, F.; Bertrand, B.; de Kochko, A.; Dussert, S. Metabolic Pathways in Tropical Dicotyledonous Albuminous Seeds: *Coffea Arabica* as a Case Study. *New Phytol.* **2009**, *182* (1), 146–162. https://doi.org/10.1111/j.1469-8137.2008.02742.x.

(6) Baptistella, J. L. C.; Assoni, G.; da Silva, M. S.; Mazzafera, P. Variation in Soluble Sugars in Arabica Coffee Cherry Fruits. *Plants* **2024**, *13* (13), 1853. https://doi.org/10.3390/plants13131853.

(7) Alves, L. C.; Magalhães, D. M. D.; Labate, M. T. V.; Guidetti-Gonzalez, S.; Labate, C. A.; Domingues, D. S.; Sera, T.; Vieira, L. G. E.; Pereira, L. F. P. Differentially Accumulated Proteins in Coffea Arabica Seeds during Perisperm Tissue Development and Their Relationship to Coffee Grain Size. *J. Agric. Food Chem.* **2016**, *64* (7), 1635–1647. https://doi.org/10.1021/acs.jafc.5b04376.

(8) Torrez, V.; Benavides-Frias, C.; Jacobi, J.; Speranza, C. I. Ecological Quality as a Coffee Quality Enhancer. A Review. *Agron. Sustain. Dev.* **2023**, *43* (1), 19. https://doi.org/10.1007/s13593-023-00874-z.

(9) Freitas, V. V.; Borges, L. L. R.; Vidigal, M. C. T. R.; dos Santos, M. H.; Stringheta, P. C. Coffee: A Comprehensive Overview of Origin, Market, and the Quality Process. *Trends Food Sci. Technol.* **2024**, *146*, 104411. https://doi.org/10.1016/j.tifs.2024.104411.

(10) ICO. INTERNATIONAL COFFEE ORGANIZATION, 2023. https://icocoffee.org/documents/cy2023-24/Coffee_Report_and_Outlook_December_2023_ICO.pdf (accessed 2024-08-21).

(11) Huang, L.; Hu, L.; Fan, Y.; Wang, X.; Dong, Y.; Li, F.; Long, Y.; Yan, L.; Meng, J.; Huang, L.; Hu, L.; Fan, Y.; Wang, X.; Dong, Y.; Li, F.; Long, Y.; Yan, L.; Meng, J. Comparative Transcriptome Analysis Revealed the Influence of Sucrose on Flavor and Taste Quality of *Coffea Arabica* and *C. Canephora* Varieties. *Beverage Plant Res.* **2021**, *1* (1), 1–9. https://doi.org/10.48130/BPR-2021-0011.

(12) Privat, I.; Foucrier, S.; Prins, A.; Epalle, T.; Eychenne, M.; Kandalaft, L.; Caillet, V.; Lin, C.; Tanksley, S.; Foyer, C.; Mccarthy, J. Differential Regulation of Grain Sucrose Accumulation and Metabolism in *Coffea Arabica* (Arabica) and *Coffea Canephora* (Robusta) Revealed through Gene Expression and Enzyme Activity Analysis. *New Phytol.* **2008**, *178* (4), 781–797. https://doi.org/10.1111/j.1469-8137.2008.02425.x.

(13) Leroy, T.; Ribeyre, F.; Bertrand, B.; Charmetant, P.; Dufour, M.; Montagnon, C.; Marraccini, P.; Pot, D. Genetics of Coffee Quality. *Braz. J. Plant Physiol.* **2006**, *18*, 229–242. https://doi.org/10.1590/S1677-04202006000100016.

(14) Salmona, J.; Dussert, S.; Descroix, F.; de Kochko, A.; Bertrand, B.; Joët, T. Deciphering Transcriptional Networks That Govern *Coffea Arabica* Seed Development Using Combined cDNA Array and Real-Time RT-PCR Approaches. *Plant Mol. Biol.* **2008**, *66* (1), 105–124. https://doi.org/10.1007/s11103-007-9256-6.

(15) Feldman, R. S.; Ryder, W. S.; Kung, J. T. Importance of Nonvolatile Compounds to the Flavor of Coffee. *J. Agric. Food Chem.* **1969**, *17* (4), 733–739. https://doi.org/10.1021/jf60164a018.

(16) De Maria, C. A. B.; Trugo, L. C.; Moreira, R. F. A.; Werneck, C. C. Composition of Green Coffee Fractions and Their Contribution to the Volatile Profile Formed during Roasting. *Food Chem.* **1994**, *50* (2), 141–145. https://doi.org/10.1016/0308-8146(94)90111-2.

(17) Cheng, B.; Furtado, A.; Smyth, H. E.; Henry, R. J. Influence of Genotype and Environment on Coffee Quality. *Trends Food Sci. Technol.* **2016**, *57*, 20–30. https://doi.org/10.1016/j.tifs.2016.09.003.

(18) Braun, D. M.; Wang, L.; Ruan, Y.-L. Understanding and Manipulating Sucrose Phloem Loading, Unloading, Metabolism, and Signalling to Enhance Crop Yield and Food Security. *J. Exp. Bot.* **2014**, *65* (7), 1713–1735. https://doi.org/10.1093/jxb/ert416.

(19) Ruan, Y.-L. Sucrose Metabolism: Gateway to Diverse Carbon Use and Sugar Signaling. *Annu. Rev. Plant Biol.* **2014**, *65* (Volume 65, 2014), 33–67. https://doi.org/10.1146/annurev-arplant-050213-040251.

(20) Roitsch, T.; González, M.-C. Function and Regulation of Plant Invertases: Sweet Sensations. *Trends Plant Sci.* **2004**, *9* (12), 606–613. https://doi.org/10.1016/j.tplants.2004.10.009.

(21) Carvalho, V. D. de; Chagas, S. J. de R.; Chalfoun, S. M.; Botrel, N.; Junior, E. S. G. J. Relação entre a composição físico-química e química do grão beneficiado e a qualidade

de bebida do café 1 - Atividades de polifenoloxidase e peroxidase, índice de coloração de acidez. *Pesqui. Agropecuária Bras.* **1994**, *29* (3), 449–454.

(22) Rogers, W. J.; Michaux, S.; Bastin, M.; Bucheli, P. Changes to the content of sugars, sugar alcohols, myo-inositol, carboxylic acids and inorganic anions in developing grains from different varieties of Robusta (*Coffea canephora*) and Arabica (*C. arabica*) coffees. *Plant Sci.* **1999**, *149* (2), 115–123. https://doi.org/10.1016/S0168-9452(99)00147-8.

(23) Combes, M.-C.; Joët, T.; Stavrinides, A. K.; Lashermes, P. New Cup out of Old Coffee: Contribution of Parental Gene Expression Legacy to Phenotypic Novelty in Coffee Beans of the Allopolyploid Coffea Arabica L. *Ann. Bot.* **2023**, *131* (1), 157–170. https://doi.org/10.1093/aob/mcac041.

(24) Cheng, B.; Furtado, A.; Henry, R. J. The Coffee Bean Transcriptome Explains the Accumulation of the Major Bean Components through Ripening. *Sci. Rep.* **2018**, *8* (1), 11414. https://doi.org/10.1038/s41598-018-29842-4.

(25) Bandil, G. B.; Etto, R. M.; Galvão, C. W.; Ramos, H. J. O.; Souza, E. M.; Pedrosa, F. O.; Chaves, D. F. S.; Huergo, L. F.; Ayub, R. A. Comparative Proteomic Analysis between Early Developmental Stages of the *Coffea Arabica* Fruits. *Genet. Mol. Res.* **2013**, *12* (4), 5102–5110. https://doi.org/10.4238/2013.October.29.4.

(26) Specialty Coffee Association of America (SCAA). SCAA Protocols - Cupping Specialty Coffee, 2015.

(27) López-Hidalgo, C.; Meijón, M.; Lamelas, L.; Valledor, L. The Rainbow Protocol: A Sequential Method for Quantifying Pigments, Sugars, Free Amino Acids, Phenolics, Flavonoids and MDA from a Small Amount of Sample. *Plant Cell Environ.* **2021**, *44* (6), 1977–1986. https://doi.org/10.1111/pce.14007.

(28) Hendriks, J. H. M.; Kolbe, A.; Gibon, Y.; Stitt, M.; Geigenberger, P. ADP-Glucose Pyrophosphorylase Is Activated by Posttranslational Redox-Modification in Response to Light and to Sugars in Leaves of Arabidopsis and Other Plant Species. *Plant Physiol.* **2003**, *133* (2), 838–849. https://doi.org/10.1104/pp.103.024513.

(29) Zanandrea, I.; Alves, J. D.; Deuner, S.; Goulart, P. de F. P.; Henrique, P. de C.; Silveira, N. M. Tolerance of Sesbania Virgata Plants to Flooding. *Aust. J. Bot.* **2010**, *57* (8), 661–669. https://doi.org/10.1071/BT09144.

(30) Van Handel, E. Direct Microdetermination of Sucrose. *Anal. Biochem.* **1968**, *22* (2), 280–283. https://doi.org/10.1016/0003-2697(68)90317-5.

(31) Yemm, E. W.; Willis, A. J. The Estimation of Carbohydrates in Plant Extracts by Anthrone. *Biochem. J.* **1954**, *57* (3), 508–514. https://doi.org/10.1042/bj0570508.

(32) Miller, G. L. Use of Dinitrosalicylic Acid Reagent for Determination of Reducing Sugar. *Anal. Chem.* **1959**, *31* (3), 426–428. https://doi.org/10.1021/ac60147a030.

(33) Lisec, J.; Schauer, N.; Kopka, J.; Willmitzer, L.; Fernie, A. R. Gas Chromatography Mass Spectrometry–Based Metabolite Profiling in Plants. *Nat. Protoc.* **2006**, *1*, 387–396. https://doi.org/10.1038/nprot.2006.59.

(34) Pang, Z.; Lu, Y.; Zhou, G.; Hui, F.; Xu, L.; Viau, C.; Spigelman, A. F.; MacDonald, P. E.; Wishart, D. S.; Li, S.; Xia, J. MetaboAnalyst 6.0: Towards a Unified Platform for Metabolomics Data Processing, Analysis and Interpretation. *Nucleic Acids Res.* **2024**, *52* (W1), W398–W406. https://doi.org/10.1093/nar/gkae253.

(35) Ruiz-Perez, D.; Guan, H.; Madhivanan, P.; Mathee, K.; Narasimhan, G. So You Think You Can PLS-DA? *BMC Bioinformatics* **2020**, *21* (1), 2. https://doi.org/10.1186/s12859-019-3310-7.

(36) Kruskal, W. H.; Wallis, W. A. Use of Ranks in One-Criterion Variance Analysis. *J. Am. Stat. Assoc.* **1952**, *47* (260), 583–621. https://doi.org/10.1080/01621459.1952.10483441.

(37) Dunn, O. J. Multiple Comparisons Using Rank Sums. *Technometrics* **1964**, *6* (3), 241–252. https://doi.org/10.1080/00401706.1964.10490181.

(38) Peterson, B. G.; Carl, P. PerformanceAnalytics: Econometric Tools for Performance and Risk Analysis, 2007, 2.0.4. https://doi.org/10.32614/CRAN.package.PerformanceAnalytics.

(39) Lightner, J.; Wu, J.; Browse, J. A Mutant of Arabidopsis with Increased Levels of Stearic Acid. *Plant Physiol.* **1994**, *106* (4), 1443–1451. https://doi.org/10.1104/pp.106.4.1443.

(40) Ruan, Y.-L. Sucrose Metabolism: Gateway to Diverse Carbon Use and Sugar Signaling. *Annu. Rev. Plant Biol.* **2014**, *65* (Volume 65, 2014), 33–67. https://doi.org/10.1146/annurev-arplant-050213-040251.

(41) Marino, C. L.; Leite, S. M. M.; Farro, A. P. C.; Sassaki, F. T.; Campos, H. L. V. de; Coscrato, V. E. Putative Metabolic Pathway of Mannitol and Sorbitol and in Sugarcane. *Sci. Agric.* **2003**, *60*, 723–728. https://doi.org/10.1590/S0103-90162003000400017.

(42) Ren, Y.; Liao, S.; Xu, Y. An Update on Sugar Allocation and Accumulation in Fruits. *Plant Physiol.* **2023**, *193* (2), 888–899. https://doi.org/10.1093/plphys/kiad294.

(43) Ruan, Y.-L. Sucrose Metabolism: Gateway to Diverse Carbon Use and Sugar Signaling. *Annu. Rev. Plant Biol.* **2014**, *65* (1), 33–67. https://doi.org/10.1146/annurev-arplant-050213-040251.

(44) Hazzouri, K. M.; Gros-Balthazard, M.; Flowers, J. M.; Copetti, D.; Lemansour, A.; Lebrun, M.; Masmoudi, K.; Ferrand, S.; Dhar, M. I.; Fresquez, Z. A.; Rosas, U.; Zhang, J.; Talag, J.; Lee, S.; Kudrna, D.; Powell, R. F.; Leitch, I. J.; Krueger, R. R.; Wing, R. A.; Amiri, K. M. A.; Purugganan, M. D. Genome-Wide Association Mapping of Date Palm Fruit Traits. *Nat. Commun.* **2019**, *10* (1), 4680. https://doi.org/10.1038/s41467-019-12604-9.

(45) Larsen, B.; Migicovsky, Z.; Jeppesen, A. A.; Gardner, K. M.; Toldam-Andersen, T. B.; Myles, S.; Ørgaard, M.; Petersen, M. A.; Pedersen, C. Genome-Wide Association Studies in Apple Reveal Loci for Aroma Volatiles, Sugar Composition, and Harvest Date. *Plant Genome* **2019**, *12* (2), 180104. https://doi.org/10.3835/plantgenome2018.12.0104.

(46) Marie, L.; Abdallah, C.; Campa, C.; Courtel, P.; Bordeaux, M.; Navarini, L.; Lonzarich, V.; Bosselmann, A. S.; Turreira-García, N.; Alpizar, E.; Georget, F.; Breitler, J.-C.; Etienne, H.; Bertrand, B. G × E Interactions on Yield and Quality in *Coffea Arabica*: New F1 Hybrids Outperform American Cultivars. *Euphytica* **2020**, *216* (5), 78. https://doi.org/10.1007/s10681-020-02608-8.

(47) Merchant, A.; Richter, A. A. Polyols as Biomarkers and Bioindicators for 21st Century Plant Breeding. *Funct. Plant Biol.* **2011**, *38* (12), 934. https://doi.org/10.1071/FP11105.

(48) Williamson, J. D.; Jennings, D. B.; Guo, W.-W.; Pharr, D. M.; Ehrenshaft, M. Sugar Alcohols, Salt Stress, and Fungal Resistance: Polyols—Multifunctional Plant Protection? *J. Am. Soc. Hortic. Sci.* **2002**, *127* (4), 467–473. https://doi.org/10.21273/JASHS.127.4.467.

(49) Bailly, C.; El-Maarouf-Bouteau, H.; Corbineau, F. From intracellular signaling networks to cell death: the dual role of reactive oxygen species in seed physiology. *C. R. Biol.* **2008**, *331* (10), 806–814. https://doi.org/10.1016/j.crvi.2008.07.022.

(50) Sant'Ana, G. C.; Pereira, L. F. P.; Pot, D.; Ivamoto, S. T.; Domingues, D. S.; Ferreira, R. V.; Pagiatto, N. F.; da Silva, B. S. R.; Nogueira, L. M.; Kitzberger, C. S. G.; Scholz, M. B. S.; de Oliveira, F. F.; Sera, G. H.; Padilha, L.; Labouisse, J.-P.; Guyot, R.; Charmetant, P.; Leroy, T. Genome-Wide Association Study Reveals Candidate Genes Influencing Lipids and Diterpenes Contents in *Coffea Arabica* L. *Sci. Rep.* **2018**, *8* (1),

465. https://doi.org/10.1038/s41598-017-18800-1.

(51) Li, Z.; Zhang, C.; Zhang, Y.; Zeng, W.; Cesarino, I. Coffee Cell Walls—Composition, Influence on Cup Quality and Opportunities for Coffee Improvements. *Food Qual. Saf.* **2021**, *5*, fyab012. https://doi.org/10.1093/fqsafe/fyab012.

(52) Yeager, S. E.; Batali, M. E.; Guinard, J.-X.; Ristenpart, W. D. Acids in Coffee: A Review of Sensory Measurements and Meta-Analysis of Chemical Composition. *Crit. Rev. Food Sci. Nutr.* **2023**, *63* (8), 1010–1036. https://doi.org/10.1080/10408398.2021.1957767.

(53) Carvalho, C. H. S. D.; Bartelega, L.; Sera, G. H.; Matiello, J. B.; Almeida, S. R. D.; Santinato, F.; Hotz, A. L. Catálogo de Cultivares de Café Arábica, 2022.

(54) Figueiredo, L. P.; Borém, F. M.; Ribeiro, F. C.; Giomo, G. S.; Malta, M. R.; Taveira, J. H. da S. Sensory Analysis and Chemical Composition of 'Bourbon' Coffees Cultivated in Different Environments. **2018**. https://doi.org/10.25186/cs.v13i1.1400.

(55) Zhou, D.; Jing, H.; Yuan, J.; Zhou, M.; Liu, L.; Fu, J.; Ouyang, L.; Xu, J.; Bian, J.; Fu, H.; He, H. Non-Targeted GC–MS Metabolomics-Based Differences in Indica Rice Seeds of Different Varieties. *BMC Plant Biol.* **2024**, *24* (1), 1–11. https://doi.org/10.1186/s12870-024-05255-6.

(56) Ivamoto, S. T.; Júnior, O. R.; Domingues, D. S.; Santos, T. B. dos; Oliveira, F. F. de; Pot, D.; Leroy, T.; Vieira, L. G. E.; Carazzolle, M. F.; Pereira, G. A. G.; Pereira, L. F. P. Transcriptome Analysis of Leaves, Flowers and Fruits Perisperm of *Coffea Arabica* L. Reveals the Differential Expression of Genes Involved in Raffinose Biosynthesis. *PLOS ONE* **2017**, *12* (1), e0169595. https://doi.org/10.1371/journal.pone.0169595.

(57) Ruan, Y.-L.; Jin, Y.; Yang, Y.-J.; Li, G.-J.; Boyer, J. S. Sugar Input, Metabolism, and Signaling Mediated by Invertase: Roles in Development, Yield Potential, and Response to Drought and Heat. *Mol. Plant* **2010**, *3* (6), 942–955. https://doi.org/10.1093/mp/ssq044.

(58) Souto, L. S.; Oliveira, D. M. T. Seed development in Malpighiaceae species with an emphasis on the relationships between nutritive tissues. *C. R. Biol.* **2014**, *337* (1), 62–70. https://doi.org/10.1016/j.crvi.2013.11.001.

(59) Lu, J.; Magnani, E. Seed Tissue and Nutrient Partitioning, a Case for the Nucellus. *Plant Reprod.* **2018**, *31* (3), 309–317. https://doi.org/10.1007/s00497-018-0338-1.

(60) Bagnall, N.; Wang, X.-D.; Scofield, G. N.; Furbank, R. T.; Offler, C. E.; Patrick, J. W. Sucrose Transport-Related Genes Are Expressed in Both Maternal and Filial Tissues of

Developing Wheat Grains. *Funct. Plant Biol.* **2000**, *27* (11), 1009–1020. https://doi.org/10.1071/pp00012.

(61) Costa, L. M.; Yuan, J.; Rouster, J.; Paul, W.; Dickinson, H.; Gutierrez-Marcos, J. F. Maternal Control of Nutrient Allocation in Plant Seeds by Genomic Imprinting. *Curr. Biol.* **2012**, *22* (2), 160–165. https://doi.org/10.1016/j.cub.2011.11.059.

(62) Yin, L.-L.; Xue, H.-W. The MADS29 Transcription Factor Regulates the Degradation of the Nucellus and the Nucellar Projection during Rice Seed Development. *Plant Cell* **2012**, *24* (3), 1049–1065. https://doi.org/10.1105/tpc.111.094854.

Supporting information for:

Metabolic analyses reveal that the early stage of Arabica coffee seed development can be crucial in determining bean quality

Gabriela Ester Ferraz[1], Robert Márquez Gutiérrez[2], Taís Teixeira das Neves[2], Thales Henrique Cherubino Ribeiro[3], Marili Galvino Santos[1], Raphael Ricon de Oliveira[4], Gladyston Rodrigues de Carvalho[5], Marcelo Ribeiro Malta[5], Renato Ribeiro de Lima[6], Flávia Maria Avelar Gonçalves[1], Antonio Chalfun-Junior[2*]

[1] Institute of Natural Sciences, Department of Biology, Federal University of Lavras, Lavras, 37200900, Brazil

[2] Institute of Natural Sciences, Department of Biology, Laboratory of Plant Molecular Physiology, Plant Physiology Sector, Federal University of Lavras, Lavras, 37200900, Brazil

[3] Genome Center, University of California, Davis, CA 95616, USA

[4] Department of Biology, State University of Santa Cruz, Ilhéus, 5662900, Brazil

[5] Experimental Field of Lavras, Minas Gerais Agricultural Research Agency (EPAMIG), Lavras, 37200-970, Brazil

[6] Department of Statistics, Federal University of Lavras, Lavras, 37200900, Brazil

***Corresponding Author:**

Antonio Chalfun-Júnior, chalfunjunior@ufla.br

Figure S1. Spearman correlation ($p < 0.01$) for RS concentration data obtained via the DNS method (A) and data based on the difference between TS and SUC concentrations (B).

Figure S2. Carbohydrate content in the pericarp of fruits from different *Coffea arabica* cultivars: MGS Paraíso 2, MGS Catiguá 3, and Gueisha at location 1 (A, C, E, and G), and MGS Paraíso 2, MGS Catiguá 3 at location 2 (B, D, F, and H), across seven DS, from aqueous perisperm at 30 DAF (DS1) to mature endosperm at 270 DAF (DS7). RS (A and B), SUC (C and D), STC (E and F), and TS (G and H). Different letters indicate significant differences ($p < 0.05$) within each DS. The vertical bar represents the standard error of the mean.

Figure S3. DAM analysis in seeds of *Coffea arabica* cultivars at location 1. Multiple comparisons were performed between cultivars: MGS Paraíso 2 vs. MGS Catiguá 3 (A, D, G, and J); MGS Paraíso 2 vs. Gueisha (B, E, H, and K); and MGS Catiguá 3 vs. Gueisha (C, F, I, and L). Volcano plots display DAMs in four DS: DS3 (A–C), DS4 (D–F), DS5 (G–I), and DS7 (J–L). Each point in the plots represents a metabolite; with the x-axis indicating the log2(FC) value, and the y-axis representing the -log10 (p-value). Red points correspond to positively regulated metabolites ($p < 0.1$ and $FC \geq 1$), blue points represent negatively regulated metabolites ($p < 0.1$ and $FC \leq -1$), and gray points indicate metabolites with no significant difference.

Figure S4. DAM analysis in seeds of *Coffea arabica* cultivars at location 2. Comparisons were performed between cultivars MGS Paraíso 2 and MGS Catiguá 3. Volcano plots display DAMs in four DS: DS3 (A), DS4 (B), DS5 (C), and DS7 (E). Each point in the plots represents a metabolite; with the x-axis indicates the log2(FC) value, and the y-axis representing the -log10 (p-value). Red points correspond to positively regulated metabolites ($p < 0.1$ and $FC \geq 1$), blue points represent negatively regulated metabolites ($p < 0.1$ and $FC \leq -1$), and gray points indicate metabolites with no significant difference.

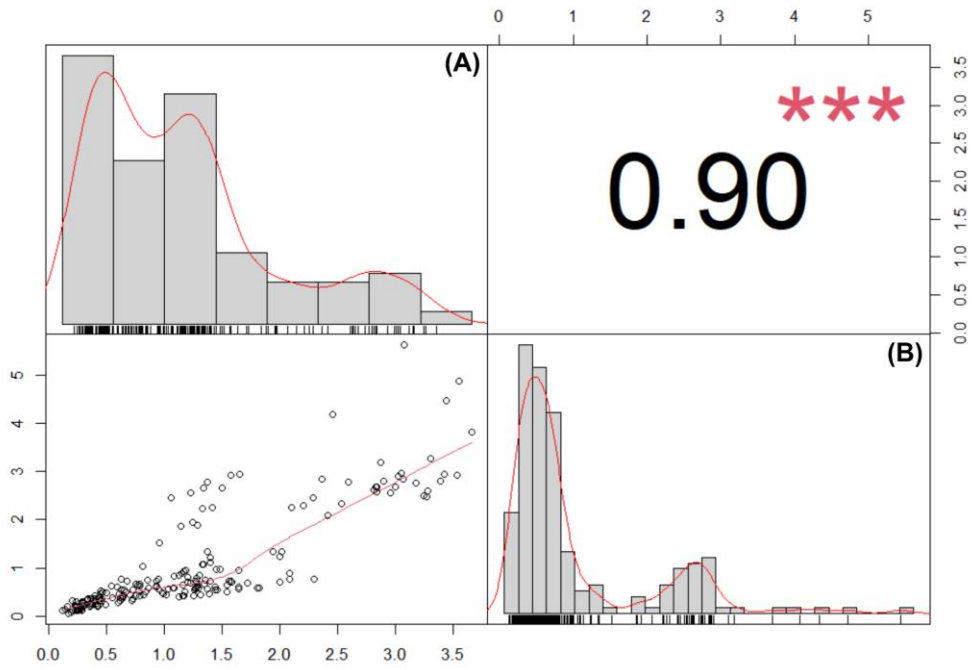


Figure S1. Spearman correlation ($p < 0.01$) for RS concentration data obtained via the DNS method (A) and data based on the difference between TS and SUC concentrations (B).

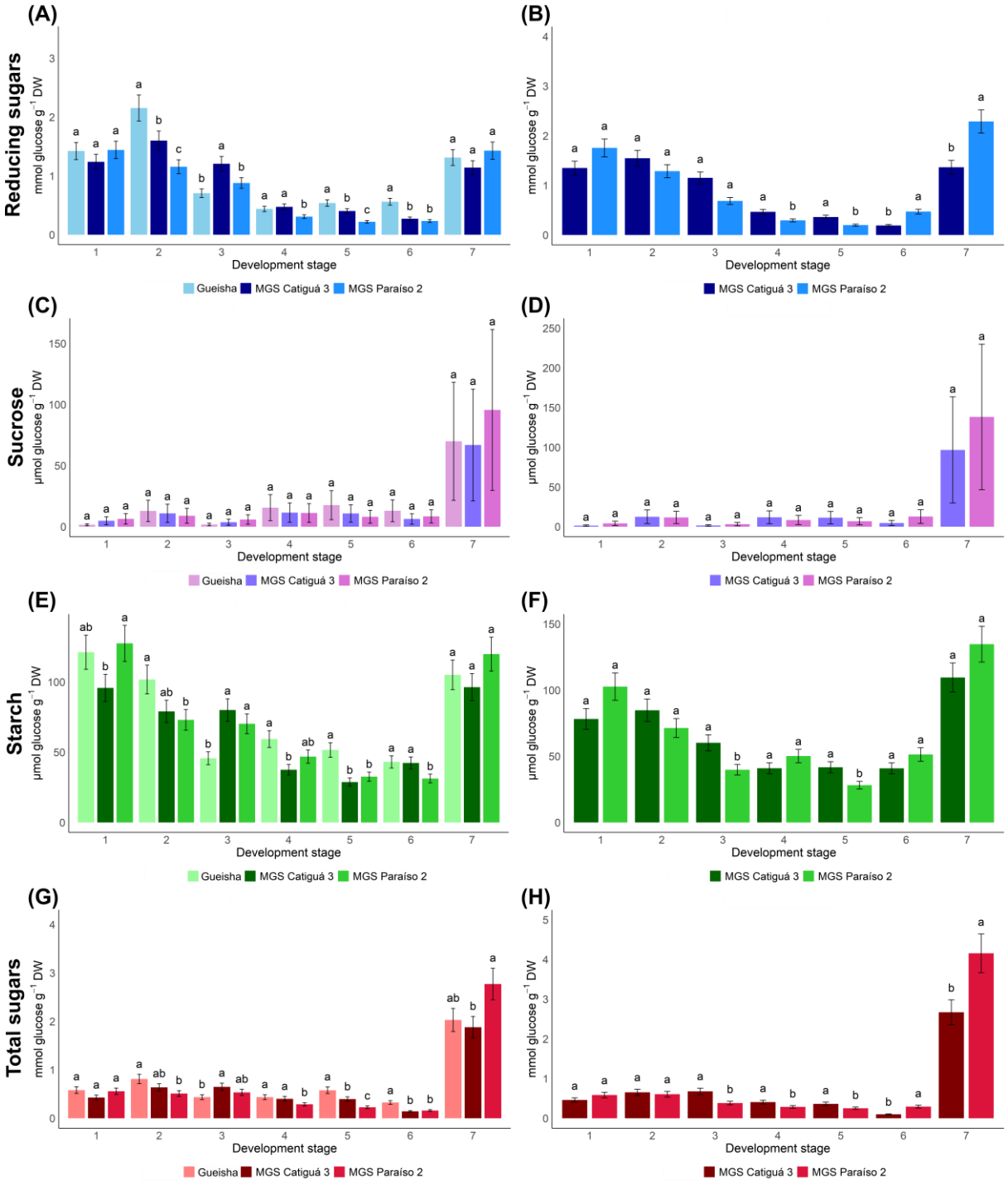


Figure S2. Carbohydrate content in the pericarp of fruits from different *Coffea arabica* cultivars: MGS Paraíso 2, MGS Catiguá 3, and Gueisha at location 1 (A, C, E, and G), and MGS Paraíso 2, MGS Catiguá 3 at location 2 (B, D, F, and H), across seven DS, from aqueous perisperm at 30 DAF (DS1) to mature endosperm at 270 DAF (DS7). RS (A and B), SUC (C and D), STC (E and F), and TS (G and H). Different letters indicate significant differences ($p < 0.05$) within each DS. The vertical bar represents the standard error of the mean.

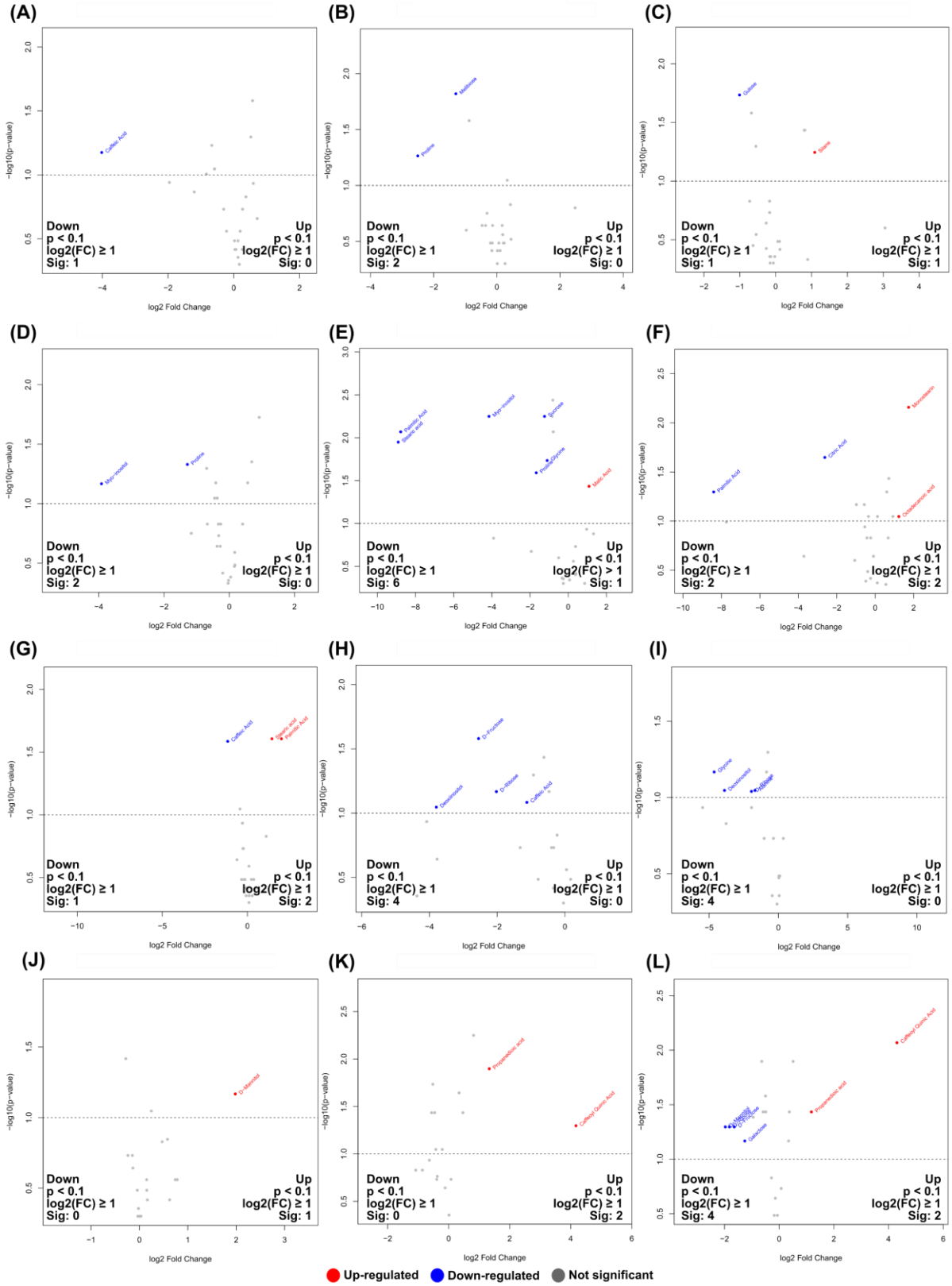


Figure S3. DAM analysis in seeds of *Coffea arabica* cultivars at location 1. Multiple comparisons were performed between cultivars: MGS Paraíso 2 vs. MGS Catiguá 3 (A, D, G,

and J); MGS Paraíso 2 vs. Gueisha (B, E, H, and K); and MGS Catiguá 3 vs. Gueisha (C, F, I, and L). Volcano plots display DAMs in four DS: DS3 (A–C), DS4 (D–F), DS5 (G–I), and DS7 (J–L). Each point in the plots represents a metabolite; with the x-axis indicating the log2(FC) value, and the y-axis representing the -log10 (p-value). Red points correspond to positively regulated metabolites (p < 0.1 and FC ≥ 1), blue points represent negatively regulated metabolites (p < 0.1 and FC ≤ -1), and gray points indicate metabolites with no significant difference.

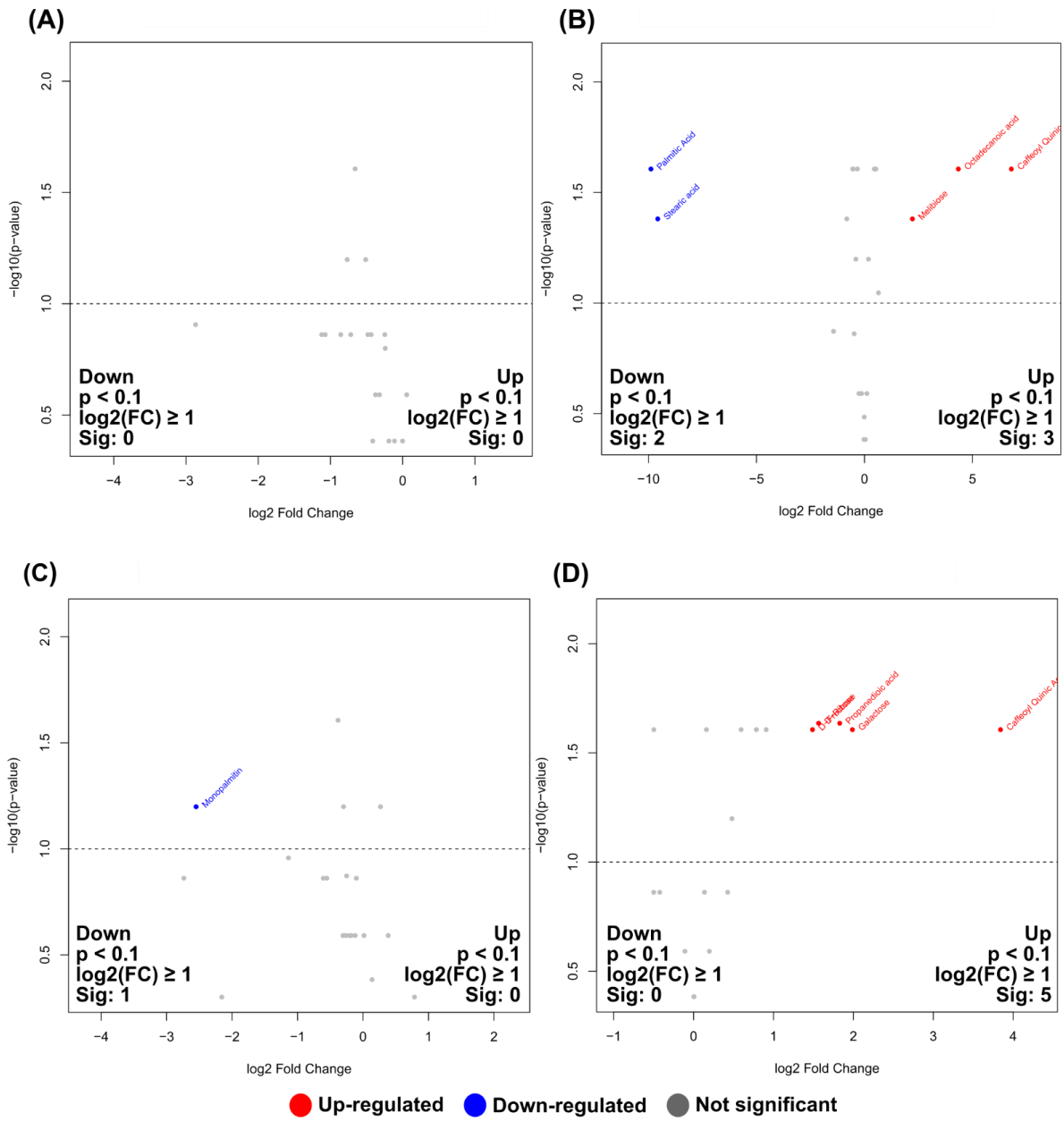


Figure S4. DAM analysis in seeds of *Coffea arabica* cultivars at location 2. Comparisons were performed between cultivars MGS Paraíso 2 and MGS Catiguá 3. Volcano plots display DAMs

in four DS: DS3 (A), DS4 (B), DS5 (C), and DS7 (E). Each point in the plots represents a metabolite; with the x-axis indicating the log2(FC) value, and the y-axis representing the -log10 (p-value). Red points correspond to positively regulated metabolites ($p < 0.1$ and $FC \geq 1$), blue points represent negatively regulated metabolites ($p < 0.1$ and $FC \leq -1$), and gray points indicate metabolites with no significant difference.

Table S1. Characteristics of compounds identified in seeds of *Coffea arabica* cultivars at five developmental stages.

Retention times	Chemical name of the according library in the GS-MS data system	Common name	m/z
8.914	L-Alanine, N-(trimethylsilyl)-, trimethylsilyl ester (CAS)	Alanine	116, 73, 147
9.717	Acetic acid, [(trimethylsilyl)oxy]-, trimethylsilyl ester (CAS) GLYCOLIC ACID-DITMS	Acetic acid	73, 147, 133
12.016	Glycine, N,N-bis(trimethylsilyl)-, trimethylsilyl ester (CAS) GLYCINE-TRITMS	Glycine	174, 73, 100
12.542	PROLINE 2TMS	Proline	142, 73, 147
14.121	Silane, [1,4-phenylenebis(oxy)]bis[trimethyl- (CAS) HYDROQUINONE DITMS	Silane	239, 254, 73
15.212	MALIC ACID, TRIS(TRIMETHYLSILYL) ESTER	Malic Acid	73, 147, 233
18.040*	RIBITOL-1,2,3,4,5-PENTATMS	Ribitol*	73, 217, 147
19.095	CITRIC ACID-TETRATMS	Citric acid	73, 273, 147
19.500	DEOXIINOSITOL, PENTAKIS-O-(TRIMETHYLSILYL)-	Deoxiinositol	73, 345, 255
19.618	D-Fructose, 1,3,4,5,6-pentakis-O-(trimethylsilyl)-, O-methyloxime (CAS) PENTAKISTRIMETHYLSILYL ETHER FRUCTOSE (D) METHOXIME	D-Fructose	73, 103, 217
19.717	D-Ribose, 2,3,4,5-tetrakis-O-(trimethylsilyl)-, O-methyloxime (CAS) TETRAKISTRIMETHYLSILYL ETHER RIBOSE (D) METHOXIME	D-Ribose	73, 217, 103
19.889	GALACTOSE OXIME 6TMS	Galactose	73, 319, 205
20.102	GLUCOSE OXIME 6TMS	Glucose	73, 319, 147
20.225	D-Glucitol, 1,2,3,4,5,6-hexakis-O-(trimethylsilyl)- (CAS) GLUCITOL-HEXATMS	D-Mannitol	73, 147, 319
20.974	MYO-INOSITOL, HEXAKIS-O-(TRIMETHYLSILYL)-	Myo-inositol	73, 217, 147
21.491	Hexadecanoic acid, trimethylsilyl ester (CAS) PALMITIC ACID-MONOTMS	Palmitic acid	117, 73, 75
21.848	Inositol, 1,2,3,4,5,6-hexakis-O-(trimethylsilyl)-, epi- (CAS) EPI-INOSITOL TMS ETHER	Inositol	73, 217, 305
22.338	CAFFEIC ACID-TMS-ETHER	Caffeic acid	219, 73, 396
23.296	Octadecanoic acid, trimethylsilyl ester (CAS) STEARIC ACID-MONOTMS	Stearic acid	117, 341, 73
26.041	MONOPALMITIN 2TMS	Monopalmitin	371, 73, 147
26.305	SUCROSE-OCTATMS	Sucrose	361, 73, 217
27.201	2-Monostearin trimethylsilyl ether	Monostearin	129, 73, 218
27.460	Octadecanoic acid, 2,3-bis[(trimethylsilyl)oxy]propyl ester (CAS) 1-MONOSTEARIN-DITMS	Octadecanoic acid	399, 73, 147
28.853	MELIBIOSE 8TMS	Melibiose	204, 73, 217
28.953	Gulose, 2,3,4,5,6-pentakis-O-(trimethylsilyl)- (CAS) GULOSE-PENTATMS	Gulose	204, 73, 147

Continuation...

29.771	QUINIC ACID-PENTATMS	Quinic acid	73, 345, 255
29.915	HEXATRIMETHYLSILYL-TRANS-5-O-CAFFEOYL-D-QUINIC ACID	Caffeoyl quinic acid	73, 345, 255
30.675	Propanedioic acid, bis[(trimethylsilyl)oxy]-, bis(trimethylsilyl) ester (CAS) MESOXALIC ACID-TETRATMS KETOMALONIC ACID-TETRATMS	Propanedioic acid	307, 73, 345

* Internal standard

Table S2. Cross-validation parameters of PLS-DA model*. Location 1: Patrocinio, MG; Location 2: Ibiá, MG.

Location	Cultivars	Measure	1 comps	2 comps	3 comps
Location 1	MGS Paraíso 2	Accuracy	0.21333	0.48667	0.69333
		R2	0.38482	0.92146	0.96185
		Q2	0.18806	0.86046	0.88524
	MGS Catiguá 3	Accuracy	0.19667	0.45333	0.60333
		R2	0.44651	0.90814	0.97717
		Q2	0.19498	0.82313	0.86598
	Guesha	Accuracy	0.14667	0.45333	0.49333
		R2	0.72324	0.84478	0.9006
		Q2	0.61301	0.64499	0.69565
Location 2	MGS Paraíso 2	Accuracy	0.32917	0.60833	0.80417
		R2	0.60023	0.84205	0.94608
		Q2	0.40128	0.75141	0.83448
	MGS Catiguá 3	Accuracy	0.21667	0.53333	0.81667
		R2	0.18192	0.67361	0.81322
		Q2	-0.041909	0.28461	0.37528

* Q2 is an estimate of the predictive ability of the model, and is calculated via cross-validation (CV). In each CV, the predicted data are compared with the original data, and the sum of squared errors is calculated. The prediction error is then summed over all samples (Predicted Residual Sum of Squares or PRESS). For convenience, the PRESS is divided by the initial sum of squares and subtracted from 1 to resemble the scale of the R2. Good predictions will have low PRESS or high Q2. (Commentary from the MetaboAnalyst[34])

Table S3. Multiple comparisons of 27 metabolites quantified in *Coffea arabica* seeds at five developmental stages. Location 1: Patrocínio, MG, Location 2: Ibiá, MG; C1: MGS Paraíso 2, C2: MGS Catiguá 3, C3: Guesha; log2FC: log2 fold change.

Metabolite	Developmental Stage	Location	Cultivar	Cultivar	log2FC	p_value
Alanine	1	2	C1	C2	-0.36976	0.41363
Alanine	1	2	C1	C3	-0.96167	0.108522
Alanine	1	2	C2	C3	-0.59192	0.102279
Acetic acid	1	2	C1	C2	-0.63622	0.228028
Acetic acid	1	2	C1	C3	-1.30085	0.026316
Acetic acid	1	2	C2	C3	-0.66463	0.116519
Glycine	1	2	C1	C2	-7.02162	0.008536
Glycine	1	2	C1	C3	-5.81021	0.050525
Glycine	1	2	C2	C3	1.211403	0.228028
Proline	1	2	C1	C2	-1.80957	0.058762
Proline	1	2	C1	C3	-3.05982	0.022086
Proline	1	2	C2	C3	-1.25025	0.308538
Silane	1	2	C1	C2	-0.5889	0.256345
Silane	1	2	C1	C3	-3.32744	0.032039
Silane	1	2	C2	C3	-2.73854	0.071617
Malic Acid	1	2	C1	C2	-0.77649	0.256345
Malic Acid	1	2	C1	C3	-2.25634	0.032039
Malic Acid	1	2	C2	C3	-1.47985	0.071617
Citric Acid	1	2	C1	C2	-0.29562	0.29649
Citric Acid	1	2	C1	C3	-1.3102	0.090725
Citric Acid	1	2	C2	C3	-1.01457	0.12842
Deoxiinositol	1	2	C1	C2	-0.10286	0.5
Deoxiinositol	1	2	C1	C3	-0.59301	0.036819
Deoxiinositol	1	2	C2	C3	-0.49015	0.036819
D-Fructose	1	2	C1	C2	-1.20388	0.012674
D-Fructose	1	2	C1	C3	-0.4028	0.32736
D-Fructose	1	2	C2	C3	0.801082	0.036819
D-Ribose	1	2	C1	C2	-1.33152	0.012674
D-Ribose	1	2	C1	C3	-0.38835	0.32736
D-Ribose	1	2	C2	C3	0.943171	0.036819
Galactose	1	2	C1	C2	-0.36794	0.440749
Galactose	1	2	C1	C3	-1.51124	0.018444
Galactose	1	2	C2	C3	-1.1433	0.026316
Glucose	1	2	C1	C2	0.009391	0.354694
Glucose	1	2	C1	C3	-1.38623	0.022086
Glucose	1	2	C2	C3	-1.39562	0.033377
D-Mannitol	1	2	C1	C2	1.01713	0.116519
D-Mannitol	1	2	C1	C3	3.590791	0.026316
D-Mannitol	1	2	C2	C3	2.57366	0.228028
Myo-inositol	1	2	C1	C2	-0.0373	0.32736
Myo-inositol	1	2	C1	C3	-0.90224	0.012674
Myo-inositol	1	2	C2	C3	-0.86493	0.036819

Continuation...

Palmitic Acid	1	2	C1	C2	0.532874	0.185547
Palmitic Acid	1	2	C1	C3	-0.26862	0.3575
Palmitic Acid	1	2	C2	C3	-0.8015	0.072064
Inositol	1	2	C1	C2	-0.25921	0.228028
Inositol	1	2	C1	C3	-3.03077	0.008536
Inositol	1	2	C2	C3	-2.77156	0.050525
Caffeic Acid	1	2	C1	C2	-0.07103	0.440749
Caffeic Acid	1	2	C1	C3	-5.36751	0.018444
Caffeic Acid	1	2	C2	C3	-5.29648	0.026316
Monopalmitin	1	2	C1	C2	-0.00848	0.5
Monopalmitin	1	2	C1	C3	-0.48748	0.185547
Monopalmitin	1	2	C2	C3	-0.479	0.185547
Sucrose	1	2	C1	C2	0.023828	0.440749
Sucrose	1	2	C1	C3	-0.79571	0.068019
Sucrose	1	2	C2	C3	-0.81954	0.050525
Monostearin	1	2	C1	C2	0.287291	0.185547
Monostearin	1	2	C1	C3	0.208648	0.185547
Monostearin	1	2	C2	C3	-0.07864	0.5
Octadecanoic acid	1	2	C1	C2	0.036993	0.440749
Octadecanoic acid	1	2	C1	C3	-0.54642	0.148359
Octadecanoic acid	1	2	C2	C3	-0.58341	0.116519
Melibiose	1	2	C1	C2	-0.24793	0.256345
Melibiose	1	2	C1	C3	-1.40259	0.032039
Melibiose	1	2	C2	C3	-1.15466	0.071617
Gulose	1	2	C1	C2	-0.86045	0.211339
Gulose	1	2	C1	C3	-2.91336	0.02251
Gulose	1	2	C2	C3	-2.05291	0.044487
Caffeoyl Quinic Acid	1	2	C1	C2	-0.26105	0.228028
Caffeoyl Quinic Acid	1	2	C1	C3	-1.30239	0.008536
Caffeoyl Quinic Acid	1	2	C2	C3	-1.04134	0.050525
Alanine	1	1	C1	C2	0.668989	0.137617
Acetic acid	1	1	C1	C2	0.891329	0.024767
Glycine	1	1	C1	C2	-1.63438	0.024767
Proline	1	1	C1	C2	0.213577	0.256345
Silane	1	1	C1	C2	-0.17988	0.5
Malic Acid	1	1	C1	C2	0.947225	0.124107
Citric Acid	1	1	C1	C2	0.776949	0.041632
Deoxiinositol	1	1	C1	C2	1.194166	0.024767
D-Fructose	1	1	C1	C2	0.420468	0.41363
D-Ribose	1	1	C1	C2	0.426349	0.41363
Galactose	1	1	C1	C2	0.524514	0.137617
Glucose	1	1	C1	C2	0.337289	0.281851
D-Mannitol	1	1	C1	C2	1.364188	0.024767
Myo-inositol	1	1	C1	C2	0.987265	0.024767
Palmitic Acid	1	1	C1	C2	0.654721	0.041632
Inositol	1	1	C1	C2	0.845652	0.024767
Caffeic Acid	1	1	C1	C2	0.633401	0.256345

Continuation...

Stearic acid	1	1	C1	C2	0.328623	0.089856
Monopalmitin	1	1	C1	C2	1.237429	0.024767
Sucrose	1	1	C1	C2	1.086511	0.024767
Monostearin	1	1	C1	C2	1.356924	0.024767
Octadecanoic acid	1	1	C1	C2	4.490819	0.024767
Melibiose	1	1	C1	C2	-0.06274	0.32736
Gulose	1	1	C1	C2	0.667488	0.137617
Quinic Acid	1	1	C1	C2	0.183066	0.124107
Caffeoyl Quinic Acid	1	1	C1	C2	0.75269	0.024767
Propanedioic acid	1	1	C1	C2	-0.48342	0.41363
Alanine	3	2	C1	C2	-0.30723	0.185547
Alanine	3	2	C1	C3	-0.20171	0.32736
Alanine	3	2	C2	C3	0.105519	0.32736
Acetic acid	3	2	C1	C2	-1.20172	0.135949
Acetic acid	3	2	C1	C3	-0.31047	0.17727
Acetic acid	3	2	C2	C3	0.891252	0.466323
Glycine	3	2	C1	C2	0.022535	0.32736
Glycine	3	2	C1	C3	-0.14582	0.32736
Glycine	3	2	C2	C3	-0.16836	0.185547
Proline	3	2	C1	C2	-1.95807	0.114551
Proline	3	2	C1	C3	-2.5025	0.054405
Proline	3	2	C2	C3	-0.54443	0.285375
Silane	3	2	C1	C2	-0.82956	0.098353
Silane	3	2	C1	C3	0.259298	0.5
Silane	3	2	C2	C3	1.088861	0.056923
Malic Acid	3	2	C1	C2	0.066449	0.382797
Malic Acid	3	2	C1	C3	0.184908	0.275492
Malic Acid	3	2	C2	C3	0.118459	0.382797
Citric Acid	3	2	C1	C2	0.712267	0.219289
Citric Acid	3	2	C1	C3	0.45012	0.302788
Citric Acid	3	2	C2	C3	-0.26215	0.375915
Deoxiinositol	3	2	C1	C2	0.193941	0.275492
Deoxiinositol	3	2	C1	C3	0.169189	0.228028
Deoxiinositol	3	2	C2	C3	-0.02475	0.440749
D-Fructose	3	2	C1	C2	-0.58958	0.089856
D-Fructose	3	2	C1	C3	0.20556	0.32736
D-Fructose	3	2	C2	C3	0.795136	0.036819
D-Ribose	3	2	C1	C2	-0.58126	0.089856
D-Ribose	3	2	C1	C3	0.231186	0.32736
D-Ribose	3	2	C2	C3	0.812444	0.036819
Galactose	3	2	C1	C2	0.370259	0.148359
Galactose	3	2	C1	C3	0.104836	0.382797
Galactose	3	2	C2	C3	-0.26542	0.228028
Glucose	3	2	C1	C2	0.570986	0.026316
Glucose	3	2	C1	C3	0.0195	0.382797
Glucose	3	2	C2	C3	-0.55149	0.050525
D-Mannitol	3	2	C1	C2	0.59528	0.116519

Continuation...

D-Mannitol	3	2	C1	C3	0.429964	0.148359
D-Mannitol	3	2	C2	C3	-0.16532	0.440749
Myo-inositol	3	2	C1	C2	0.03472	0.382797
Myo-inositol	3	2	C1	C3	-0.13479	0.228028
Myo-inositol	3	2	C2	C3	-0.16951	0.148359
Palmitic Acid	3	2	C1	C3	2.47703	0.158655
Inositol	3	2	C1	C2	0.267711	0.185547
Inositol	3	2	C1	C3	0.32815	0.089856
Inositol	3	2	C2	C3	0.060438	0.32736
Caffeic Acid	3	2	C1	C2	-4.01651	0.066807
Caffeic Acid	3	2	C1	C3	-0.96847	0.251167
Caffeic Acid	3	2	C2	C3	3.048039	0.251167
Monopalmitin	3	2	C1	C2	0.128199	0.32736
Monopalmitin	3	2	C1	C3	0.063575	0.32736
Monopalmitin	3	2	C2	C3	-0.06462	0.5
Sucrose	3	2	C1	C2	0.519489	0.050525
Sucrose	3	2	C1	C3	-0.15633	0.382797
Sucrose	3	2	C2	C3	-0.67582	0.026316
Monostearin	3	2	C1	C2	-0.2197	0.275492
Monostearin	3	2	C1	C3	-0.35653	0.228028
Monostearin	3	2	C2	C3	-0.13683	0.440749
Octadecanoic acid	3	2	C1	C2	0.169975	0.5
Octadecanoic acid	3	2	C1	C3	0.01444	0.5
Octadecanoic acid	3	2	C2	C3	-0.15554	0.5
Melibiose	3	2	C1	C2	-0.66795	0.058762
Melibiose	3	2	C1	C3	-1.29768	0.01513
Melibiose	3	2	C2	C3	-0.62973	0.354694
Gulose	3	2	C1	C2	0.130043	0.440749
Gulose	3	2	C1	C3	-0.87761	0.026316
Gulose	3	2	C2	C3	-1.00765	0.018444
Caffeoyl Quinic Acid	3	2	C1	C2	0.256994	0.382797
Caffeoyl Quinic Acid	3	2	C1	C3	-0.46756	0.228028
Caffeoyl Quinic Acid	3	2	C2	C3	-0.72455	0.148359
Alanine	3	1	C1	C2	-0.43741	0.137617
Acetic acid	3	1	C1	C2	-0.76833	0.063315
Glycine	3	1	C1	C2	-0.51264	0.063315
Malic Acid	3	1	C1	C2	-0.71931	0.137617
Deoxiinositol	3	1	C1	C2	-0.37674	0.256345
D-Fructose	3	1	C1	C2	-1.07317	0.137617
D-Ribose	3	1	C1	C2	-1.12053	0.137617
Galactose	3	1	C1	C2	-0.32128	0.256345
Glucose	3	1	C1	C2	0.055218	0.256345
D-Mannitol	3	1	C1	C2	-0.19144	0.41363
Myo-inositol	3	1	C1	C2	-0.24768	0.137617
Inositol	3	1	C1	C2	-0.11365	0.41363
Caffeic Acid	3	1	C1	C2	-2.86649	0.124107
Monopalmitin	3	1	C1	C2	-0.48196	0.137617

Continuation...

Sucrose	3	1	C1	C2	-0.00253	0.41363
Monostearin	3	1	C1	C2	-0.65902	0.024767
Octadecanoic acid	3	1	C1	C2	-0.85751	0.137617
Melibiose	3	1	C1	C2	-0.24207	0.158655
Gulose	3	1	C1	C2	-0.41535	0.41363
Alanine	4	2	C1	C2	-0.02255	0.440749
Alanine	4	2	C1	C3	-0.26148	0.440749
Alanine	4	2	C2	C3	-0.23893	0.382797
Acetic acid	4	2	C1	C2	-0.27964	0.228028
Acetic acid	4	2	C1	C3	0.395809	0.382797
Acetic acid	4	2	C2	C3	0.675448	0.148359
Glycine	4	2	C1	C2	-0.67349	0.148359
Glycine	4	2	C1	C3	-1.10203	0.018444
Glycine	4	2	C2	C3	-0.42854	0.148359
Proline	4	2	C1	C2	-1.2854	0.046803
Proline	4	2	C1	C3	-1.67525	0.025614
Proline	4	2	C2	C3	-0.38985	0.326391
Silane	4	2	C1	C2	0.165325	0.256345
Malic Acid	4	2	C1	C2	0.170231	0.32736
Malic Acid	4	2	C1	C3	1.092064	0.036819
Malic Acid	4	2	C2	C3	0.921833	0.089856
Citric Acid	4	2	C1	C2	0.688539	0.044487
Citric Acid	4	2	C1	C3	-1.9333	0.211339
Citric Acid	4	2	C2	C3	-2.62184	0.02251
Deoxiinositol	4	2	C1	C2	-0.45138	0.089856
Deoxiinositol	4	2	C1	C3	-0.80401	0.003645
Deoxiinositol	4	2	C2	C3	-0.35263	0.089856
D-Fructose	4	2	C1	C2	-0.33567	0.148359
D-Fructose	4	2	C1	C3	0.267969	0.275492
D-Fructose	4	2	C2	C3	0.603643	0.050525
D-Ribose	4	2	C1	C2	-0.32457	0.185547
D-Ribose	4	2	C1	C3	0.384092	0.185547
D-Ribose	4	2	C2	C3	0.708661	0.036819
Galactose	4	2	C1	C2	-0.27593	0.148359
Galactose	4	2	C1	C3	-0.84266	0.005635
Galactose	4	2	C2	C3	-0.56673	0.068019
Glucose	4	2	C1	C2	-0.20117	0.382797
Glucose	4	2	C1	C3	-3.91011	0.148359
Glucose	4	2	C2	C3	-3.70895	0.228028
D-Mannitol	4	2	C1	C2	0.157771	0.336302
D-Mannitol	4	2	C1	C3	-0.32829	0.432886
D-Mannitol	4	2	C2	C3	-0.48606	0.408481
Myo-inositol	4	2	C1	C2	-3.91829	0.068019
Myo-inositol	4	2	C1	C3	-4.15122	0.005635
Myo-inositol	4	2	C2	C3	-0.23293	0.148359
Palmitic Acid	4	2	C1	C2	-0.37567	0.228028
Palmitic Acid	4	2	C1	C3	-8.77685	0.008536

Continuation...

Palmitic Acid	4	2	C2	C3	-8.40118	0.050525
Inositol	4	2	C1	C2	-0.69368	0.050525
Inositol	4	2	C1	C3	-0.78233	0.008536
Inositol	4	2	C2	C3	-0.08865	0.228028
Caffeic Acid	4	2	C1	C2	0.042051	0.41363
Stearic acid	4	2	C1	C2	-1.16559	0.17727
Stearic acid	4	2	C1	C3	-8.90547	0.011247
Stearic acid	4	2	C2	C3	-7.73988	0.102447
Monopalmitin	4	2	C1	C2	-0.37446	0.089856
Monopalmitin	4	2	C1	C3	-0.25604	0.5
Monopalmitin	4	2	C2	C3	0.118417	0.089856
Sucrose	4	2	C1	C2	-0.26331	0.148359
Sucrose	4	2	C1	C3	-1.24115	0.005635
Sucrose	4	2	C2	C3	-0.97784	0.068019
Monostearin	4	2	C1	C2	-0.41375	0.066807
Monostearin	4	2	C1	C3	1.327105	0.131776
Monostearin	4	2	C2	C3	1.74085	0.006953
Octadecanoic acid	4	2	C1	C2	-0.36749	0.089856
Octadecanoic acid	4	2	C1	C3	0.862827	0.5
Octadecanoic acid	4	2	C2	C3	1.230318	0.089856
Melibiose	4	2	C1	C2	0.562922	0.066807
Melibiose	4	2	C1	C3	-0.49974	0.251167
Melibiose	4	2	C2	C3	-1.06266	0.251167
Gulose	4	2	C1	C2	-0.03674	0.466594
Gulose	4	2	C1	C3	0.071955	0.455223
Gulose	4	2	C2	C3	0.108692	0.425651
Quinic Acid	4	2	C1	C2	0.916935	0.018818
Quinic Acid	4	2	C1	C3	0.376206	0.394634
Quinic Acid	4	2	C2	C3	-0.54073	0.114551
Caffeoyl Quinic Acid	4	2	C1	C2	0.408542	0.148359
Caffeoyl Quinic Acid	4	2	C1	C3	0.964879	0.116519
Caffeoyl Quinic Acid	4	2	C2	C3	0.556337	0.440749
Alanine	4	1	C1	C2	0.110846	0.256345
Acetic acid	4	1	C1	C2	0.529559	0.024767
Glycine	4	1	C1	C2	-1.43262	0.134143
Proline	4	1	C1	C2	-0.25539	0.256345
Malic Acid	4	1	C1	C2	-0.40051	0.063315
Citric Acid	4	1	C1	C2	-0.00961	0.32736
Deoxiinositol	4	1	C1	C2	-0.54502	0.024767
D-Fructose	4	1	C1	C2	-0.02481	0.41363
D-Ribose	4	1	C1	C2	0.039382	0.41363
Galactose	4	1	C1	C2	-0.52105	0.024767
Glucose	4	1	C1	C2	-0.47716	0.137617
D-Mannitol	4	1	C1	C2	-0.81876	0.041632
Myo-inositol	4	1	C1	C2	-0.13654	0.256345
Palmitic Acid	4	1	C1	C2	-9.88776	0.024767
Inositol	4	1	C1	C2	-0.32506	0.024767

Continuation...

Stearic acid	4	1	C1	C2	-9.57237	0.041632
Monopalmitin	4	1	C1	C2	0.455143	0.024767
Sucrose	4	1	C1	C2	0.186546	0.063315
Monostearin	4	1	C1	C2	0.647474	0.089856
Octadecanoic acid	4	1	C1	C2	4.347405	0.024767
Melibiose	4	1	C1	C2	2.219473	0.041632
Caffeoyl Quinic Acid	4	1	C1	C2	6.802655	0.024767
Alanine	5	2	C1	C2	0.369871	0.32736
Alanine	5	2	C1	C3	-1.32091	0.185547
Alanine	5	2	C2	C3	-1.69078	0.089856
Acetic acid	5	2	C1	C2	-0.2645	0.185547
Acetic acid	5	2	C1	C3	-0.61973	0.036819
Acetic acid	5	2	C2	C3	-0.35523	0.185547
Glycine	5	2	C1	C2	0.554575	0.382797
Glycine	5	2	C1	C3	-4.0804	0.116519
Glycine	5	2	C2	C3	-4.63498	0.068019
Malic Acid	5	2	C1	C2	0.235696	0.32736
Malic Acid	5	2	C1	C3	-0.79053	0.32736
Malic Acid	5	2	C2	C3	-1.02623	0.185547
Citric Acid	5	2	C1	C2	0.10488	0.440749
Citric Acid	5	2	C1	C3	-0.33885	0.382797
Citric Acid	5	2	C2	C3	-0.44373	0.440749
Deoxiinositol	5	2	C1	C2	0.091046	0.5
Deoxiinositol	5	2	C1	C3	-3.79687	0.089856
Deoxiinositol	5	2	C2	C3	-3.88791	0.089856
D-Fructose	5	2	C1	C2	-0.61419	0.228028
D-Fructose	5	2	C1	C3	-2.54973	0.026316
D-Fructose	5	2	C2	C3	-1.93554	0.116519
D-Ribose	5	2	C1	C2	-0.0794	0.382797
D-Ribose	5	2	C1	C3	-2.02376	0.068019
D-Ribose	5	2	C2	C3	-1.94436	0.091211
Galactose	5	2	C1	C2	1.100388	0.148359
Galactose	5	2	C1	C3	-4.36832	0.440749
Galactose	5	2	C2	C3	-5.46871	0.116519
Glucose	5	2	C1	C2	-0.341	0.32736
D-Mannitol	5	2	C1	C2	0.089399	0.256345
Myo-inositol	5	2	C1	C2	-0.07518	0.440749
Myo-inositol	5	2	C1	C3	-0.9292	0.050525
Myo-inositol	5	2	C2	C3	-0.85402	0.068019
Palmitic Acid	5	2	C1	C2	2.000624	0.024767
Inositol	5	2	C1	C2	0.000657	0.382797
Inositol	5	2	C1	C3	-3.77444	0.228028
Inositol	5	2	C2	C3	-3.7751	0.148359
Caffeic Acid	5	2	C1	C2	-1.15859	0.025956
Caffeic Acid	5	2	C1	C3	-1.12405	0.082457
Caffeic Acid	5	2	C2	C3	0.034543	0.336302
Stearic acid	5	2	C1	C2	1.440573	0.024767

Continuation...

Monopalmitin	5	2	C1	C2	-0.23481	0.185547
Monopalmitin	5	2	C1	C3	-0.33359	0.185547
Monopalmitin	5	2	C2	C3	-0.09878	0.5
Sucrose	5	2	C1	C2	0.288307	0.440749
Sucrose	5	2	C1	C3	-0.46751	0.068019
Sucrose	5	2	C2	C3	-0.75582	0.050525
Monostearin	5	2	C1	C2	-0.18876	0.41363
Octadecanoic acid	5	2	C1	C2	-0.29066	0.116519
Octadecanoic acid	5	2	C1	C3	-0.23115	0.148359
Octadecanoic acid	5	2	C2	C3	0.059517	0.440749
Melibiose	5	2	C1	C2	-0.14867	0.32736
Melibiose	5	2	C1	C3	-0.04339	0.5
Melibiose	5	2	C2	C3	0.105286	0.32736
Gulose	5	2	C1	C2	-9.85666	0.382797
Gulose	5	2	C1	C3	0.046485	0.275492
Gulose	5	2	C2	C3	9.903142	0.382797
Quinic Acid	5	2	C1	C2	-0.43377	0.089856
Quinic Acid	5	2	C1	C3	-0.39446	0.185547
Quinic Acid	5	2	C2	C3	0.039313	0.32736
Caffeoyl Quinic Acid	5	2	C1	C2	-0.19044	0.32736
Caffeoyl Quinic Acid	5	2	C1	C3	0.169678	0.32736
Caffeoyl Quinic Acid	5	2	C2	C3	0.360122	0.185547
Alanine	5	1	C1	C2	-0.18012	0.256345
Acetic acid	5	1	C1	C2	-0.25059	0.134143
Glycine	5	1	C1	C2	0.384276	0.256345
Malic Acid	5	1	C1	C2	-0.30636	0.256345
Citric Acid	5	1	C1	C2	-0.60614	0.137617
Deoxiinositol	5	1	C1	C2	0.015496	0.256345
D-Fructose	5	1	C1	C2	-2.7361	0.137617
D-Ribose	5	1	C1	C2	-2.15669	0.5
Galactose	5	1	C1	C2	-0.1197	0.256345
Glucose	5	1	C1	C2	-1.13828	0.110336
Myo-inositol	5	1	C1	C2	-0.38097	0.024767
Inositol	5	1	C1	C2	-0.26605	0.256345
Caffeic Acid	5	1	C1	C2	0.137872	0.41363
Monopalmitin	5	1	C1	C2	-2.54921	0.063315
Sucrose	5	1	C1	C2	-0.29762	0.063315
Monostearin	5	1	C1	C2	0.786221	0.5
Octadecanoic acid	5	1	C1	C2	-0.21061	0.256345
Melibiose	5	1	C1	C2	0.266561	0.063315
Quinic Acid	5	1	C1	C2	-0.10211	0.137617
Caffeoyl Quinic Acid	5	1	C1	C2	-0.55279	0.137617
Acetic acid	7	2	C1	C2	-0.23733	0.185547
Acetic acid	7	2	C1	C3	-0.41981	0.089856
Acetic acid	7	2	C2	C3	-0.18248	0.32736
Glycine	7	2	C1	C2	0.031909	0.5
Glycine	7	2	C1	C3	-0.45943	0.036819

Continuation...

Glycine	7	2	C2	C3	-0.49134	0.036819
Malic Acid	7	2	C1	C2	0.469886	0.148359
Malic Acid	7	2	C1	C3	0.814534	0.005635
Malic Acid	7	2	C2	C3	0.344648	0.068019
Citric Acid	7	2	C1	C2	-0.14526	0.185547
Citric Acid	7	2	C1	C3	-0.20984	0.089856
Citric Acid	7	2	C2	C3	-0.06458	0.32736
Deoxiinositol	7	2	C1	C2	0.24641	0.089856
Deoxiinositol	7	2	C1	C3	-0.39167	0.185547
Deoxiinositol	7	2	C2	C3	-0.63808	0.012674
D-Fructose	7	2	C1	C2	0.780936	0.275492
D-Fructose	7	2	C1	C3	-0.86065	0.148359
D-Fructose	7	2	C2	C3	-1.64159	0.050525
D-Ribose	7	2	C1	C2	0.741467	0.275492
D-Ribose	7	2	C1	C3	-1.07871	0.148359
D-Ribose	7	2	C2	C3	-1.82018	0.050525
Galactose	7	2	C1	C2	0.62511	0.382797
Galactose	7	2	C1	C3	-0.63283	0.116519
Galactose	7	2	C2	C3	-1.25794	0.068019
Glucose	7	2	C1	C3	-1.14032	0.32736
D-Mannitol	7	2	C1	C2	1.980891	0.068019
D-Mannitol	7	2	C1	C3	0.011588	0.440749
D-Mannitol	7	2	C2	C3	-1.9693	0.050525
Myo-inositol	7	2	C1	C2	0.025046	0.5
Myo-inositol	7	2	C1	C3	-0.55992	0.036819
Myo-inositol	7	2	C2	C3	-0.58496	0.036819
Inositol	7	2	C1	C2	0.580337	0.142525
Inositol	7	2	C1	C3	-0.37239	0.172352
Inositol	7	2	C2	C3	-0.95273	0.041176
Caffeic Acid	7	2	C1	C2	-0.28141	0.038261
Monopalmitin	7	2	C1	C2	0.149493	0.275492
Monopalmitin	7	2	C1	C3	0.011064	0.440749
Monopalmitin	7	2	C2	C3	-0.13843	0.228028
Sucrose	7	2	C1	C2	-0.04428	0.32736
Sucrose	7	2	C1	C3	0.466266	0.036819
Sucrose	7	2	C2	C3	0.51055	0.012674
Monostearin	7	2	C1	C2	-0.03262	0.5
Monostearin	7	2	C1	C3	0.340037	0.02275
Monostearin	7	2	C2	C3	0.372659	0.036819
Octadecanoic acid	7	2	C1	C2	0.164213	0.382797
Octadecanoic acid	7	2	C1	C3	-0.11691	0.228028
Octadecanoic acid	7	2	C2	C3	-0.28113	0.148359
Melibiose	7	2	C1	C2	-0.01977	0.440749
Melibiose	7	2	C1	C3	-0.51916	0.018444
Melibiose	7	2	C2	C3	-0.49939	0.026316
Quinic Acid	7	2	C1	C2	0.001688	0.5
Quinic Acid	7	2	C1	C3	0.076223	0.185547

Continuation...

Quinic Acid	7	2	C2	C3	0.074534	0.185547
Caffeoyl Quinic Acid	7	2	C1	C2	-0.13457	0.228028
Caffeoyl Quinic Acid	7	2	C1	C3	4.170581	0.050525
Caffeoyl Quinic Acid	7	2	C2	C3	4.305153	0.008536
Propanedioic acid	7	2	C1	C2	0.151838	0.32736
Propanedioic acid	7	2	C1	C3	1.33087	0.012674
Propanedioic acid	7	2	C2	C3	1.179032	0.036819
Acetic acid	7	1	C1	C2	0.908965	0.024767
Glycine	7	1	C1	C2	-0.4971	0.024767
Malic Acid	7	1	C1	C2	0.197522	0.256345
Citric Acid	7	1	C1	C2	0.78631	0.024767
Deoxiinositol	7	1	C1	C2	0.136452	0.137617
D-Fructose	7	1	C1	C2	1.488619	0.024767
D-Ribose	7	1	C1	C2	1.564971	0.023151
Galactose	7	1	C1	C2	1.987652	0.024767
D-Mannitol	7	1	C1	C2	-0.42336	0.137617
Myo-inositol	7	1	C1	C2	0.426887	0.137617
Inositol	7	1	C1	C2	0.00264	0.41363
Monopalmitin	7	1	C1	C2	0.161481	0.024767
Sucrose	7	1	C1	C2	0.594735	0.024767
Monostearin	7	1	C1	C2	0.480166	0.063315
Octadecanoic acid	7	1	C1	C2	0.006254	0.41363
Melibiose	7	1	C1	C2	-0.49942	0.137617
Quinic Acid	7	1	C1	C2	-0.10798	0.256345
Caffeoyl Quinic Acid	7	1	C1	C2	3.841426	0.024767
Propanedioic acid	7	1	C1	C2	1.829014	0.023151

ARTIGO 3

Genome-wide characterization of invertases in Arabica coffee and its progenitors reveals putative genes involved in fruit development

(Draft Version)

Manuscript prepared for submission to the Plant Cell Reports Journal

Gabriela Ester Ferraz[1], Robert Márquez Gutiérrez[2], Tiago Yukio Inoue[1], Gabriel de Campos Rume[2], Thais Aparecida Sales[3], Teodorico Castro Ramalho[3], Evandro Novaes[1], Dapeng Zhang[4], Antonio Chalfun-Junior[2*], Flavia Maria Avelar Gonçalves[1]

[1] Institute of Natural Sciences, Department of Biology, Federal University of Lavras, Lavras, Minas Gerais, Brazil, 37200900, MG, Brazil.
[2] Institute of Natural Sciences, Department of Biology, Laboratory of Plant Molecular Physiology, Plant Physiology Sector, Federal University of Lavras, Lavras, Minas Gerais, Brazil, 37200900, MG, Brazil
[3] Institute of Natural Sciences, Department of Chemistry, Federal University of Lavras, Lavras, Minas Gerais, Brazil, 37200900, MG, Brazil.
[4] Sustainable Perennial Crops Laboratory, United States Department of Agriculture, Agricultural Research Service, USDA, USA. 10300, Baltimore Avenue Bldg 001 Barc, West Beltsville, Md 20705

***Corresponding Author:** chalfunjunior@ufla.br

Abstract: Invertases are involved in essential physiological processes, as they irreversibly degrade sucrose, generating hexoses that play a role in plant growth and development. Few invertase genes have been characterized in *Coffea*, and consequently, the role of these enzymes in coffee plant development is not well understood. The present study conducted a genome-wide characterization of invertases in Arabica coffee and its diploid progenitors. A total of 65 invertase genes were identified, with 28 in *C. arabica* (*CaINV*), 18 in *C. canephora* (*CcINV*), and 19 in *C. eugenioides* (*CeINV*), with polyploidization being the primary cause of the expansion of the *CaINV* gene family. Synteny analyses and chromosomal mapping showed that the *CaINV* genes are mainly derived from *C. eugenioides*. The cis-acting elements suggested the involvement of invertase genes in three main categories: growth and development, responses to biotic and abiotic stress, and hormonal response. Transcriptome analyses showed that, except for the neutral/alkaline invertase gene *CaN/AINV1*, all invertase genes of *C. arabica* and *C. canephora* were expressed at different stages of fruit development, some constitutively and others conditionally. Although *CaN/AINV1* was not expressed in the evaluated tissues, the presence of two GCN4 cis-elements suggests its involvement in specific processes in tissues or stages not analyzed in this study. Our results open many avenues for future studies in *Coffea*, especially for *C. arabica* and *C. canephora*, which are economically most significant and exhibit particularities regarding disease resistance, tolerance to abiotic stresses, and fruit chemical composition.

Key words: *Coffea*, Sucrose, Glycoside hydrolase, Polyploidy, Expression profiles.

Key message **Sixty-five invertase genes were characterized, with 28 genes in *Coffea arabica*, 18 in *C. canephora*, and 19 in *C. eugenioides*. In *C. arabica* and *C. canephora*, all genes are related to fruit development.**

Introduction

Sucrose is the most widely distributed photoassimilate in plants, playing an important role in the source-sink relationship. Sucrose metabolism generates hexoses that are involved in growth, development, synthesis of essential compounds, signaling, and regulation of the expression of microRNAs, transcription factors, and hormonal and defense signaling genes (Ruan 2014; Bergareche et al. 2018). One way these sugars become available is through the irreversible breakdown of sucrose by invertase enzymes, which generate the hexoses glucose and fructose. These hexoses serve as substrates for starch and cellulose production, act as signaling molecules, and participate in essential biological processes and responses to biotic and abiotic stresses (Ruan et al. 2010; Braun et al. 2014).

Phylogenetic and biochemical data classify invertases into two orthologs groups, meaning they have converged to the same function: (i) neutral/alkaline invertases (N/AINV) with the conserved glycosyl hydrolase 100 (GH100) domain, which have a more ancestral origin and are found in algae and cyanobacteria, divided into plastidial and cytosolic (α and β groups, respectively), and (ii) acid invertases of the cell wall and vacuole (CWINV and VINV), which contain the glycosyl hydrolase 32 (GH32) domain and likely originated within the vascular plants group (Wan et al. 2018; Pang et al. 2019). GH100 invertases hydrolyze only sucrose, while the GH32 group are beta-fructofuranosidases, meaning they can also hydrolyze fructose-containing oligosaccharides such as raffinose, stachyose, and 1-kestose (Ruan et al. 2010). GH32 invertases also exhibit fine transcriptional and post-translational regulation through their inhibitors (Coculo and Lionetti 2022).

Acid invertases contain crucial amino acid residues necessary for their catalytic activity and sucrose hydrolysis. Three Trp residues are important for forming the aromatic zone involved in stabilizing the sugar-enzyme complex. An Asp residue acts as the nucleophile of the B-fructofuranosidase motif in sucrose hydrolysis contained in the NDPD/NG domain, and a Glu residue in the WECP domain (Le Roy et al. 2007; Van den Ende et al. 2009; Pang et al. 2019). Another key residue is Asp239 in the enzyme's active site, which, along with Lys-242, ensures efficient catalysis (Le Roy et al. 2007, 2013). Asp239 is used as a discriminant between invertases and exohydrolases of the GH32 family, where this residue is absent (Van den Ende et al. 2009).

Coffea arabica is a recent allopolyploid resulting from the hybridization of its diploid parents *C. canephora* and *C. eugenioides* (Scalabrin et al. 2020; Salojärvi et al. 2024). Polyploidization enables the organism to acquire particular phenotypic and genotypic

characteristics, resulting in specific adaptations in the speciation process (Saul et al. 2023). This phenomenon involves various genetic alterations such as neofunctionalization and subfunctionalization of duplicated genes, reversion of genes to the diploid state, and gene silencing through the insertion of transposable elements (Soltis and Soltis 1999; Adams and Wendel 2005). The characterization of the *C. arabica* genome has already enabled the verification of several of these processes (Scalabrin et al. 2024; Salojärvi et al. 2024).

Genomic studies in the *Coffea* genus are increasing, as *C. arabica* and *C. canephora* are of great importance to the world economy, unlike *C. eugenioides*, which has low productivity. Coffee is one of the most consumed beverages in the world, with *C. arabica* representing about 57%, and *C. canephora* 43 % of the global production market (ICO 2023). This is due to the fact that *C. arabica* is more appreciated by consumers, while *C. canephora* is mainly used as a raw material by the soluble coffee industry, due to differences in the chemical composition of the fruits, including distinct sugar levels (Privat et al. 2008; Cheng et al. 2020; Huang et al. 2021).

In terms of adaptability and responses to biotic and abiotic stresses, *C. arabica* is better adapted to higher altitude regions (above 1000 meters), more susceptible to diseases, and less resistant to water deficit and high temperatures (Carvalho 2008). On the other hand, *C. canephora* is a more robust plant, resistant to various diseases and better developed in lowland regions (Ferrão et al. 2017). Another striking difference is the time of fruit development and ripening after fertilization in *C. arabica*, it ranges from six to nine months, whereas in *C. canephora*, the fruits complete their development in about nine to eleven months (Wintgens 2012). However, with genetic breeding programs, there are C. canephora cultivars with an early fruit development time, similar to that of C. arabica (Partelli et al., 2014).

Some studies on coffee have demonstrated the importance of certain invertases in fruit development (Cheng et al. 2020; Li et al. 2021), and even a differential behavior between *C. arabica* and *C. canephora*, in invertase activity influencing sugar deposition (Privat et al. 2008), although these finding have been very specific. Therefore, given the peculiarities of *Coffea* species and the essential role of invertases in plants, this study aimed to (i) characterize invertase enzymes in *C. arabica* and in modern representatives of its parents *C. canephora* and *C. eugenioides*; (ii) examine the impacts of polyploidization on invertase genes; and (iii) investigate which invertase genes are related to fruit development.

Methods

Identification of invertase genes

The protein sequences of *C. arabica* and modern representatives of its parents, *C. canephora* and *C. eugenioides*, were obtained from the genome and proteome annotations available at the National Center for Biotechnology Information (NCBI) (GCF_003713225.1, GCF_003713205.1 and GCA_900059795.1, respectively). For *C. canephora*, the annotation from the Coffea Genome Hub database was checked (https://coffee-genome-hub.southgreen.fr/organism/1, accessed on 15 May 2024). A sequence of each conserved domain, GH32 (Pfam: PF00251) and GH100 (Pfam: PF12899), annotated in *Arabidopsis thaliana* (AT1G62660 and AT4G09510, respectively), were retrieved from the TAIR database (Berardini et al. 2015) and used as query against the proteome of each species using the BLASTp v2.10.1 tool (Altschul et al. 1990). Proteins with a high degree of similarity and more than 300 amino acids were selected for further analyses. Different isoforms of the same locus were manually removed, and only one representative per invertase gene (unigene) was retained. Conserved domains were confirmed in all selected sequences using the Conserved Domain tool (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi, accessed on 13 February 2025).

Parent invertase sequences from *C. canephora* and *C. eugenioides* absent in the respective subgenomes of *C. arabica* were used in tBLASTn. After exhaustive manual filtering for candidate locus selection, two structurally reliable protein sequences within the annotated genome region of *C. arabica* were identified, which were inserted into the phylogenetic trees: *CaCWINV7* (GH32) and *CaN/AINV9* (GH100). The retrieval of protein sequences belonging to these loci was done through the Phytozome database (https://phytozome-next.jgi.doe.gov/, accessed on 14 May 2024). The Conserved Domain tool was used to confirm the presence of the GH100 and GH32 conserved domains. The same approach was used for the parental genomes, resulting in the absence of new sequences.

Finally, 65 non-redundant protein sequences of invertases were identified in *C. arabica* (CaINV), *C. canephora* (CcINV), and *C. eugenioides* (CeINV). Their physicochemical properties (amino acid sequence length, molecular weight, and isoelectric point) were determined using the ExPASy Proteomics tool (https://web.expasy.org/cgi-bin/protparam/protparam, accessed on 7 May 2024). The Plant-mPloc server (http://www.csbio.sjtu.edu.cn/bioinf/plant-multi/, accessed on 15 May 2024) was used to predict the subcellular localization of the invertases (details in Table S1).

Phylogenetic analysis

Reference protein sequences from the plant species *Arabidopsis thaliana*, *Citrus sinensis*, *Oryza sativa*, *Populus trichocarpa*, *Solanum lycopersicum*, *Vitis vinifera*, and *Zea mays* were obtained from the NCBI database and used for phylogenetic analyses. The same Arabidopsis query sequences mentioned earlier and used for the *Coffea* species were employed to retrieve invertase sequences in the other species. Global alignments of the invertase proteins were performed using the MAFFT v7.475 program (Katoh et al. 2002), and the alignment quality was assessed by the Guidance v2.02 algorithm (Sela et al. 2015). Phylogenetic trees were generated using the PHYLIP v3.696 package (Felsenstein 1989) with 1000 bootstrap replicates generated by the seqboot algorithm. The protdist algorithm was used to calculate similarity matrices between the sequences, and the trees were constructed using the Jones-Taylor-Thornton matrix and the Neighbor-Joining method. The tree was then visualized using the ITOL v6 tool (Letunic and Bork 2021). The subgroups formed by the Arabidopsis sequences were used to classify the invertase proteins in the three *Coffea* species. The same methods were applied to generate the trees containing only the three *Coffea* species.

Genetic structure, promoter regions, and motif composition

The genomic and corresponding coding sequences (CDS) of each invertase gene were used to generate the exon-intron structure. The structures were visualized using the online program Gene Structure Display Server GSDS2.0 (Hu et al. 2015). Sequences of 2000 bp upstream of the transcription start site of the *CaINV*, *CcINV*, and *CeINV* genes were extracted, and the cis-acting elements were predicted using PlantCARE (Lescot et al. 2002). The prediction results were selected and visualized using R language v4.4.0.

The identification of putative motifs was performed with the online tool Multiple EM for Motif Elicitation (MEME) (Bailey et al. 2015), with parameters set to identify 10 motifs and a width between 6 and 50. The remaining parameters were kept at default values. For normalization, sequences from *A. thaliana*, *C. sinensis*, *O. sativa*, *P. trichocarpa*, *S. lycopersicum*, *V. vinifera*, and *Z. mays* were used. The resulting conserved motifs were verified in the Classification of Protein Families database in InterPro (https://www.ebi.ac.uk/interpro/, accessed on 26 June 2024) and Conserved Domain.

Syntenic analysis and chromosomal mapping

The synteny analysis was performed using SyMap v5.4.6 (Soderlund et al. 2011), where maps and alignments are generated by MUMmer v4 (Marçais et al. 2018). The GFF and genomic sequence files obtained from the NCBI annotation of *C. arabica*, *C. canephora*, and *C. eugenioides* were converted into the required input format using the ConvertNCBI script from SyMap. The analysis was conducted using the program's default parameters. The location data corresponding to each gene were obtained from the coffee genome annotation.

Expression profile of invertase genes in different fruit tissues

The gene expression profile of invertase genes was generated from 16 RNA-Seq libraries from different tissues of *C. arabica* and *C. canephora* related to fruit development (Fig. S2). Of these libraries, 12 are from *C. arabica* and 4 from *C. canephora*. For *C. eugenioides*, there are no available RNA-seq libraries from fruit tissues. The accession numbers and additional data for the libraries are available on Table S2. The transcriptomes were obtained from the Sequence Read Archive (SRA) database. The quality of the libraries was analyzed using the FastQC v.0.11.9 program (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and the sequences were trimmed with the Trimmomatic v0.39 tool (Bolger et al. 2014). Alignments and fragment counting were performed using the STAR v2.7.11a program (Dobin et al. 2013), with default parameters. Two genes found in tBLASTn, *CaCWINV7* and *CaN/AINV9*, despite having complete gene structures, are described as pseudogenes in the genome annotation. Therefore, during genome indexing, the STAR tool does not recognize these regions as genes, since the GTF annotation file is used as the basis for indexing. Consequently, these genes were not included in the gene expression profile analysis. Expression values were calculated in log 10 (FPKM +1) using the edgeR package (Robinson et al. 2010). The heatmap of gene expression was generated in R v4.4.0 using the gplot package.

Alignment of protein sequences and interaction analysis of CWINVs with sucrose

To verify the presence or absence of essential amino acids for hydrolysis, catalytic activity, binding, stability, and specificity in the recognition of sucrose by acid invertases, global protein alignments performed using the MAFFT v7.475 program were visualized using the NCBI

Multiple Sequence Alignment Viewer v1.25.0 (https://www.ncbi.nlm.nih.gov/projects/msaviewer/, accessed on 26 June 2024).

The interaction between CWINVs and sucrose was investigated using the primary sequences of the invertases CaCWINV1 and CaCWINV13 from *C. arabica*. Initially, a search was conducted for existing three-dimensional structures that could serve as templates for modeling the two enzymes. The search for templates was performed in three ways: the first was through the BLAST tool from the Uniprot database, where searches were conducted by aligning the primary sequence of the targets with primary sequences present in the database. The second search tool used was the Swiss Model bioinformatics server (Bienert et al. 2017), where homology modeling was performed. Finally, a manual search was conducted in the PDB, followed by sequence alignment of the candidate structures. To select the best template, the primary criterion used was the sequence identity between the model and the template, that is, the presence of identical residues between the two structures, especially in the regions corresponding to the active site. Subsequently, for tie-breaking criteria, factors such as taxonomy, the resolution of the 3D structure of the template, and the presence of active ligands in its site were also considered.

From the selected template, homology modeling was performed using the SWISS-MODEL server (Bienert et al. 2017), with the quality of the models assessed through Global Model Quality Estimation (GMQE) (Waterhouse et al. 2018) and Quality Model Energy Analysis with ensembles of distance constraints (QMEANDisco) (Studer et al. 2020) scores. Additionally, the percentage of sequence identity was considered as a primary sequence-level analysis, and Ramachandran plots (Ramachandran et al. 1963; Sobolev et al. 2020), of the models were analyzed as a secondary structure-level quality assessment.

After modeling the enzymes and evaluating the obtained models, the docking stage, or molecular docking, was conducted. For molecular docking calculations, Autodock software (Morris et al. 2009; Trott and Olson 2010) was used. Initially, the four molecular targets and the two enzymes were prepared using Autodock Tools (Morris et al. 2009), and Discovery Studio 3.5 (https://discover.3ds.com/discovery-studio-visualizer-download, asseced 23 august 2024), with adjustments to electrostatic charges, insertion of hydrogen atoms, and a general check of all residues. A grid-box was constructed in the active site region, measured based on the enzyme's natural substrate, sucrose. To validate the results of this program, ROC curve analysis was performed to predict whether the software could distinguish between active compounds and false positives, known as decoys (Piccirillo and Amaral 2018). After these validations, the selected ligands (sucrose, 1-kestose, raffinose, and stachyose) were docked into

the binding site of each of the two invertases. The interaction energy obtained was assessed, and 2D intermolecular interaction diagrams were created and analyzed.

Results

Identification and Analysis of Basic Physicochemical Properties of Invertases

A total of 65 invertase genes were identified, with 28 from *C. arabica*, 18 from *C. canephora* and 19 from *C. eugenioides*. The number of neutral/alkaline and cell wall invertases varies among species, with *C. arabica* having 13 cell wall invertases and 13 neutral/alkaline invertases; *C. canephora* having nine cell wall invertases and seven neutral/alkaline invertases; and *C. eugenioides* having nine cell wall invertases and eight neutral/alkaline invertases. However, all species have the same number of vacuolar invertases: two.

For convenience, the genes were named with the prefixes Ca, Cc, and Ce, according to their location: VINV for vacuolar invertases, CWINV for cell wall invertases, and N/AINV for neutral/alkaline invertases (details available in Table S1). The size of invertase genes in the species ranged from 1737 to 7229 bp, with CDS ranging from 960 to 2100 nucleotides. Accordingly, the size of the proteins varied from 319 to 699 amino acids, with molecular weights ranging from 35329.5 to 79141.1 Da (Daltons). The theoretical isoelectric points ranged from 4.85 to 9.28, indicating a wide range of stability under different pH conditions.

Phylogenetic Analysis

To understand the evolutionary relationship of invertases in *C. arabica* and its parental species *C. canephora* and *C. eugenioides*, protein sequences from seven other plant species were used to construct phylogenetic trees with GH32 and GH100 domains (Fig. 1). The phylogenetic relationship was analyzed separately for each conserved domain, as they are structurally unrelated domains, i.e from convergent evolution. The seven other species included in the phylogenetic trees were *A. thaliana*, *C. sinensis*, *O. sativa*, *P. trichocarpa*, *S. lycopersicum*, *V. vinifera*, and *Z. mays*. These are model species that have been extensively characterized genetically.

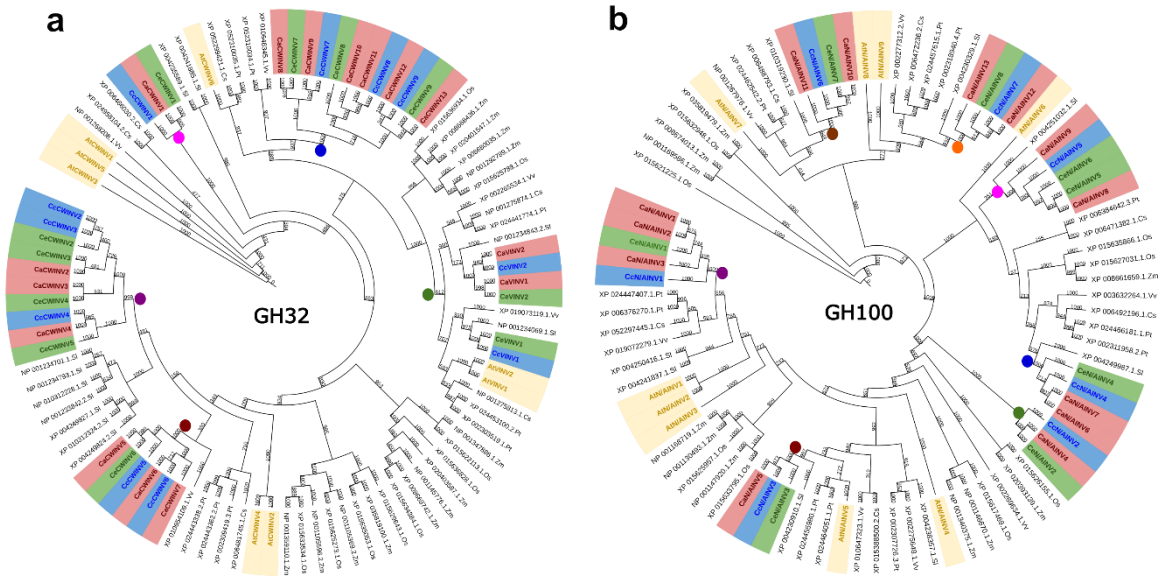


Fig. 1 Phylogenetic relationships of invertase genes in *Arabidopsis thaliana* (At), *Citrus sinensis* (Cs), *Coffea arabica* (Ca), *Coffea canephora* (Cc), *Coffea eugenioides* (Ce), *Oryza sativa* (Os), *Populus trichocarpa* (Pt), *Solanum lycopersicum* (Sl), *Vitis vinifera* (Vv), and *Zea mays* (Zm). (a) Unrooted phylogenetic tree of the glycosyl hydrolase 32 (GH32) domain; (b) Phylogenetic tree of the glycosyl hydrolase 100 (GH100) domain. Genes from *C. arabica*, *C. canephora*, and *C. eugenioides* are colored red, blue, and green, respectively; genes from *A. thaliana* are colored yellow. The other species can be identified by a suffix at the end of the accession number for each sequence. The classification of invertase genes in the three *Coffea* species was based on the subgroups formed by the Arabidopsis sequences. Clades containing the *Coffea* genes are indicated with circles in the following colors: purple, dark red, dark green, dark blue, pink, orange, and brown. The images were visualized using the ITOL tool

The results show consistent groupings (evaluated by 1000 bootstraps) that reflect the distribution of the species used to investigate the phylogenetic relationships of the *CaINV*, *CcINV*, and *CeINV* genes. The clades coherently group the grasses *O. sativa* and *Z. mays*, as well as the tree and shrub species *C. sinensis*, *P. trichocarpa*, and *V. vinifera*. Additionally, in most clades, it can be observed that the *S. lycopersicum* genes are closely related to the genes of the *Coffea* genus, as expected due to the phylogenetic proximity between *S. lycopersicum* and coffee species.

Regarding the acid invertases, the phylogenetic tree of the GH32 domain (Fig. 1a) consistently grouped the CWINV and VINV sequences. The clade marked with a dark green circle, which includes the vacuolar invertases, contains the two VINVs from each *Coffea* species, along with the VINVs from Arabidopsis. Similarly, in the clades marked with purple and pink circles, there is a lower number of *CaINV* genes than expected for an allopolyploid compared to its parental species. In the tree of the GH100 domain (Fig. 1b), the patterns of differential gene numbers among the *Coffea* species are also noted in the clades marked with

dark red, dark green, and pink circles. However, the clade marked with the purple circle contains one additional *CaINV* gene, *CaN/AINV1*.

Gene structure and conserved motifs of invertase genes

The 65 protein sequences of CaINV, CcINV, and CeINV were aligned to construct phylogenetic trees for the GH32 and GH100 domains (Fig. 2a). The structural data of the genes revealed a variation in the number of exons, ranging between four and eight. It was observed that genes with four exons are present only in the GH100 domain. The N/AINVs can be classified as plastidial in the α group, generally with six exons, and cytosolic in the β group, with four exons (Pan et al. 2019). The cluster formed by the genes *CaN/AINV1-5*, *CcN/AINV1-3*, and *CeN/AINV1-3* has four exons and, phylogenetically grouped with Arabidopsis invertases *AtN/AINV1-5*, which belong to the β group (Wan et al. 2018). Therefore, it can be inferred that this cluster of *Coffea* invertases contains cytosol-specific proteins, differing from the results obtained with the Plant-mPloc tool, which characterized all N/AINVs as plastidial (Table S1).

In the GH32 domain, the genes exhibit greater structural variability (Fig. 2b). As expected, the distribution of exons and introns is consistent with the organization of the genes in the phylogenetic tree (Fig. 2a, b).

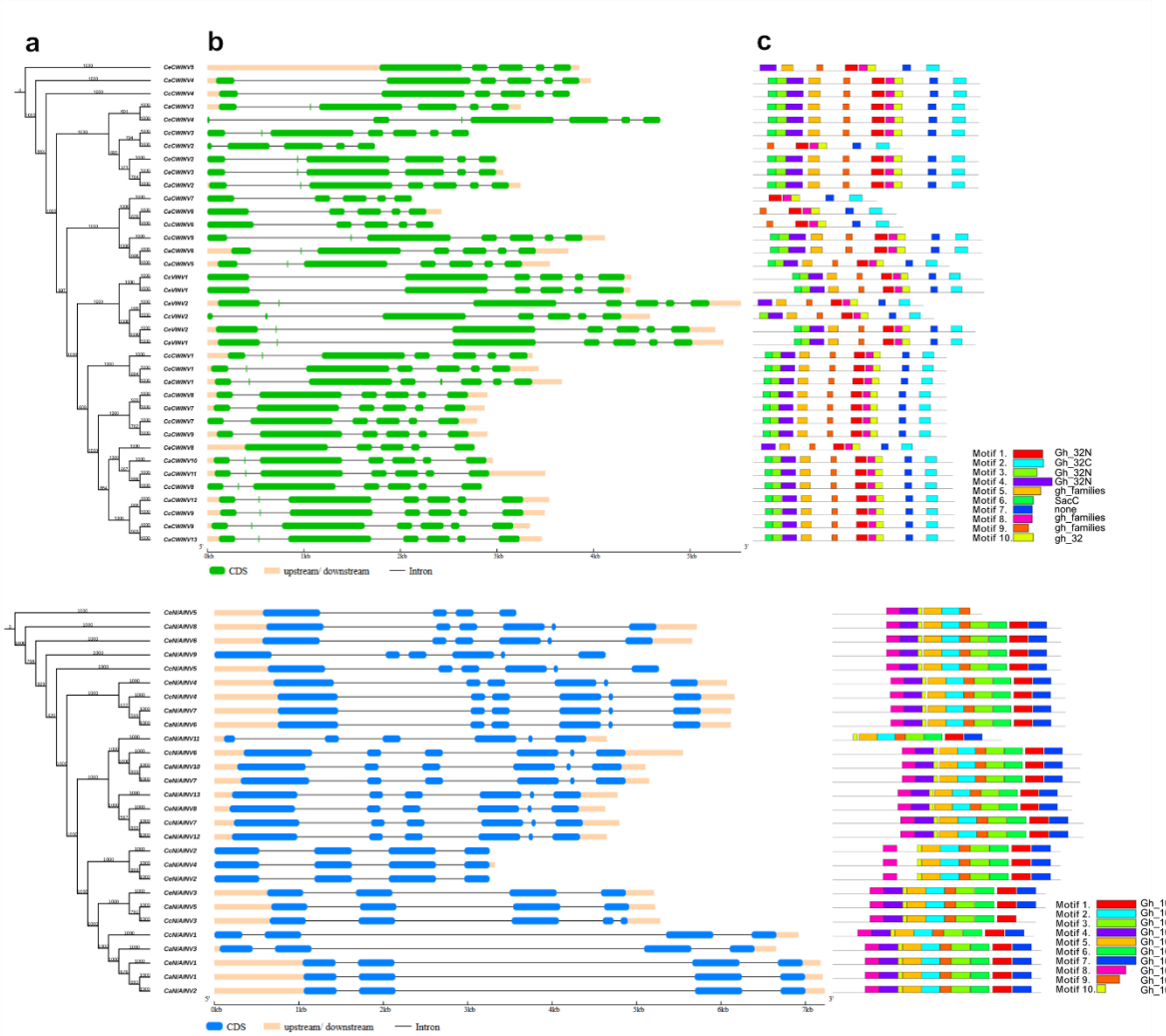


Fig. 2 Structures and protein motifs of *CaINV*, *CcINV*, and *CeINV* genes. (a) Phylogenetic tree of invertase genes in *Coffea arabica*, *C. canephora*, and *C. eugenioides*; (b) organization of exons and introns, where green and blue solid boxes represent coding sequences (CDS) of the GH32 and GH100 domains, respectively, beige solid boxes represent upstream/downstream regions, and black solid lines represent introns; (c) conserved protein motifs, with colored solid bars representing the 10 identified and characterized motifs. The images were generated using the ITOL (a), GSDS2.0 (b), and MEME (c) tools

Analyses of protein motifs revealed the presence of 10 conserved motifs in each of the GH32 and GH100 domains (Fig. 2c). Highly conserved motifs related to hydrolysis and catalytic activity of the acid invertases, NDPD/NG and WECV/P (Van den Ende et al. 2009), were identified in motifs 6 and 9, respectively. Three tryptophan residues involved in stabilizing the enzyme complex, WM/INDPNG, VWGNI, and WSGSAT, are located in domains 6, 3, and 4, respectively (Fig. S1). These domains, along with domain 5, are absent in the protein encoded by the CaCWINV7 gene. It was also observed that CaCWINV6, CcCWINV2, and CcCWINV6 do not contain motifs 3, 4, 5, and 6. Additionally, proteins CeCWINV5 and 8, and CaVINV2

are missing motifs 3 and 6. Finally, motif 6 is also absent in CcVINV2. The absence of these motifs suggests low affinity and reduced efficacy in sucrose hydrolysis.

In GH100 domain invertases, the protein structures exhibited a more consistent pattern, which may be attributed to the earlier emergence of this domain compared to GH32 (Wan et al. 2018). In GH100 proteins, the 10 identified motifs are present in most of the sequences (Fig. 2c). However, proteins CaN/AINV2, CaN/AINV4, and CeN/AINV2 lack domain 4, which is associated with carbohydrate metabolism. Additionally, CaN/AINV11 is missing both domain 4 and domain 8. Domain 7, also related to carbohydrate metabolism, is absent in CcN/AINV3. The most divergent protein sequence is that encoded by the *CeN/AINV5* gene, which lacks domains 1, 3, 6, and 7. Domains 1, 3, and 6 are associated with the typical glycosyl hydrolase 100 domain, sucrose cleavage activity, and carbohydrate metabolism, respectively.

Prediction of cis-elements

Due to their crucial role in transcriptional regulatory networks, cis-elements of invertase genes were investigated to assess possible variation among the three *Coffea* species. A total of 92, 89, and 94 unique cis-elements were identified in the promoter regions of 65 genes from *CaINV* (Fig. 3a), *CcINV* (Fig. 3b), and *CeINV* (Fig. 3c), respectively. Apart from the elements common to all genes, 47 cis-elements were categorized into groups: response to biotic and abiotic stresses, responsive to phytohormones, and plant growth and development (Fig. 3).

In the category of response to biotic and abiotic stresses, all genes from the three species exhibit the MYC regulatory cis-elements, which are associated with responses to low temperatures (Zhang et al. 2016). A greater number of these elements were observed in *CaINV* and *CeINV* genes. Additionally, other regulatory elements were identified: LTR, also involved in cold response; TC-rich, in defense and stress mechanisms; STRE, related to heat shock proteins; MBS, induced by drought; DRE, associated with both drought and salinity responses (Zhang et al. 2009); and ARE, involved in anaerobic induction.

Among cis-elements responsive to phytohormones, P-box and GARE, TGA and AuxRR-core, ABRE, TCA, and ERE, related to responses to gibberellin, auxin, abscisic acid, salicylic acid, and ethylene, respectively, were identified. A higher ratio of cis-elements per gene was observed for ABRE. Regulatory elements associated with response to methyl jasmonate, named as-1, TGACG, and CGTCA, showed the same quantity and location in the promoter region. In this case, as-1 corresponds to the group containing TGACG elements,

differing only in nomenclature (Xu and Timko 2004). TGACG and CGTCA sequences represent palindromic cis-element sequences.

Finally, the roles of *CaINV*, *CcINV*, and *CeINV* genes are so diverse that regulatory cis-elements related to circadian control (circadian), meristem expression (CAT-box), and seed development (RY) were detected in the plant growth and development category. Additionally, the presence of GCN4 (Fig. 3, red asterisks) is notable, with a duplicated presence in the genes *CaN/AINV1* and *CaN/AINV13*. It is important to highlight that GCN4 is a specific cis-element involved in endosperm expression. Another relevant observation is that the *CcN/AINV1* gene is the only one among the three species to present the cis-regulatory element AACA (Fig. 3, red asterisk), also expressed in the endosperm. Most other elements are associated with light responses, with GT1 being notably present eight times in *CaCWINV10*.

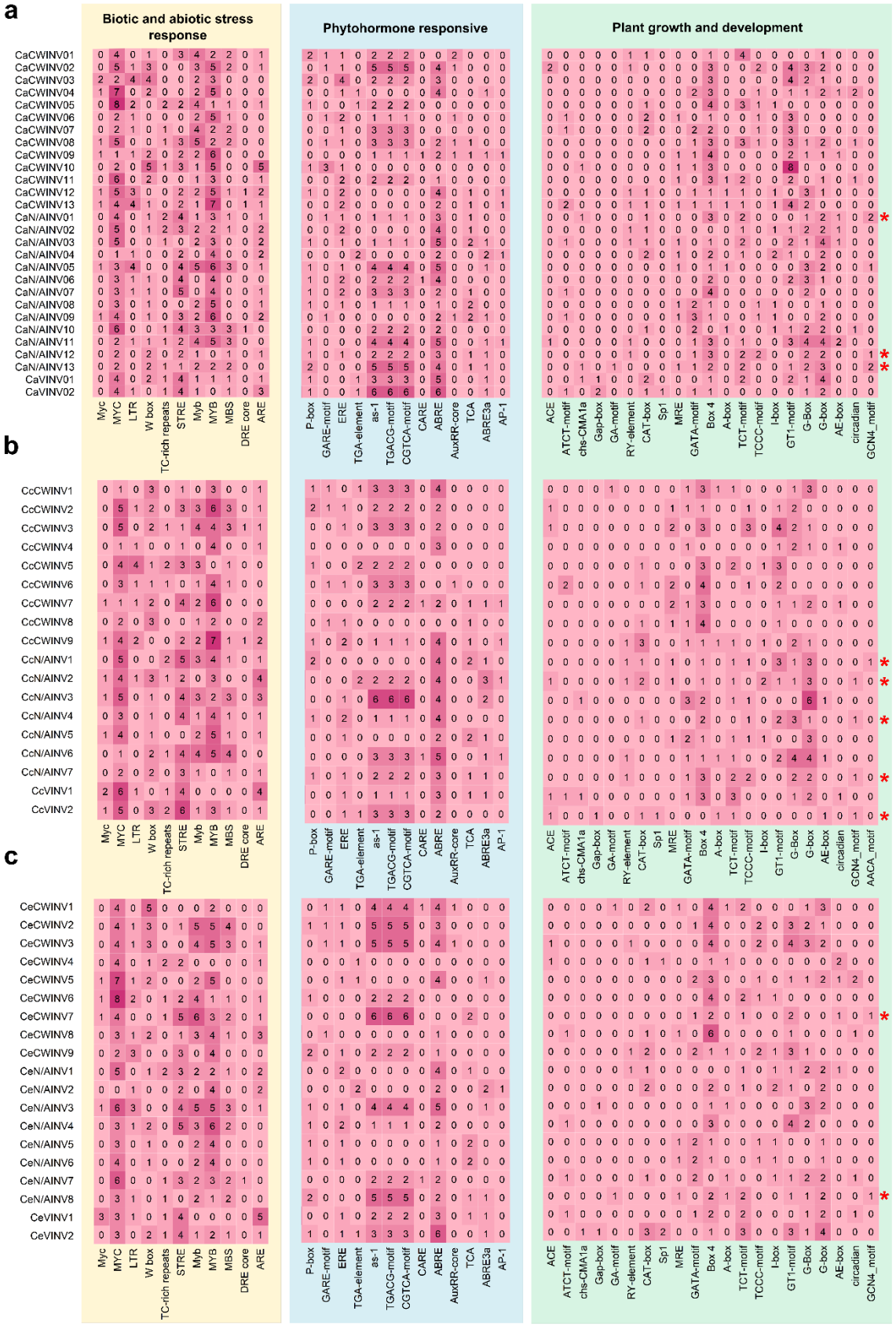


Fig. 3 Analysis of the number of cis-elements in the 65 Genes of *CaINV* (a), *CcINV* (b), and *CeINV* (c). The pink gradient represents the varying numbers of cis-elements in invertases. The elements were categorized into three groups: response to biotic and abiotic stresses, phytohormones responsive, and plant growth and development, represented by yellow, blue,

and green boxes, respectively. Red asterisks highlight genes containing endosperm-specific cis-elements. The image was generated using the R language

Syntenic analysis and chromosomal mapping

The syntenic analysis revealed 379 synteny blocks, corresponding to 78.6% of the *C. canephora* genome and 81.9% of the corresponding subgenome in *C. arabica* (Cc x subCe), respectively. Of these blocks, approximately 34.4% and 36.9% are homologous to two or more blocks, indicating possible genomic duplications. The total number of syntenic blocks larger than 10 Mb (megabases) is 3.16%. In the analysis comparing *C. eugenioides* with the corresponding subgenome in *C. arabica* (Ce x subCc), 222 synteny blocks were identified, covering 97.1% and 92.7% of the genomes, respectively, with 6.36% of these syntenic blocks being larger than 10 Mb. The percentage of homologous sequences by two or more blocks was 31.3% for *C. eugenioides* and 32.7% for subCe.

Regarding collinear sets, 3,211 sets were found for Ce x subCe and 3,303 sets for Cc x subCc, with a higher number of collinear sets (>=10-20 sets) observed in 10.6% of Ce x subCe and 4.7% of Cc x subCc. Thus, high collinearity and synteny were observed, with no evidence of major rearrangements in the chromosomes of the Ce x subCa alignment (Fig.4). Significant differences in the number of transposable elements were noted between *C. canephora* and subCc, likely caused by homeologous exchanges after polyploidization (Salojärvi et al. 2024), which contributed to the higher differentiation found in Cc x subCc compared with Ce x subCe.

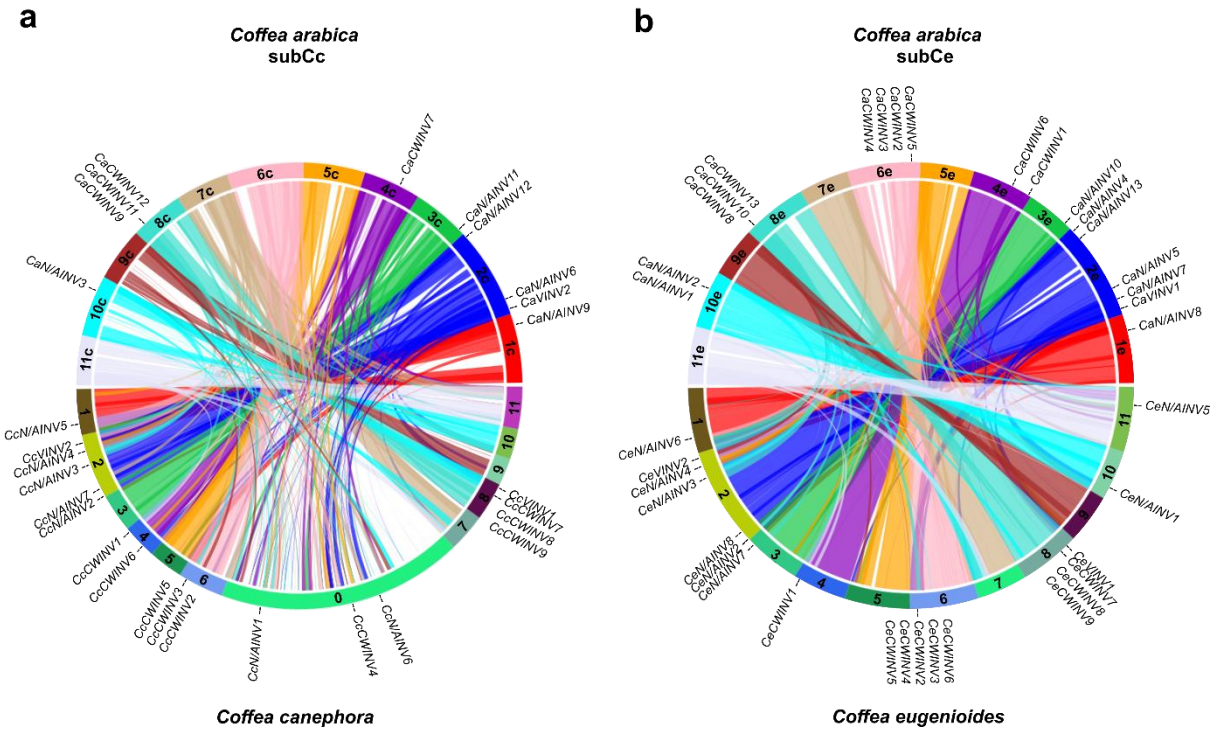


Fig. 4 Synteny analysis and schematic representations of chromosomal distribution of the 65 *CaINV*, *CcINV*, and *CeINV* Genes. Two circular maps were generated using SyMAP, where synteny blocks were scaled by physical position and represented by different colored bands. (a) Synteny alignment of the *C. canephora* genome with subCc. (b) Synteny alignment of the *C. eugenioides* genome with subCe. Chromosome numbers are indicated inside the colored bars

A total of 28 invertase genes were identified in *C. arabica*. Of these, 10 are found in subCc, including one VINV, four CWINVs, and five N/AINVs. In subCe, invertase genes are divided into one VINV, nine CWINVs, and eight N/AINVs, totaling 18 (Fig. 4 and Table S1). The *C. canephora* genome annotation includes chromosome 0, which consists of sequences that have not been assigned to a specific chromosome during the genomic assembly process. This chromosome contains three invertase genes: *CcN/AINV1*, *CcN/AINV6*, and *CcCWINV4*. According to the phylogenetic tree, it can be inferred that *CcN/AINV1* and *CcN/AINV6* are orthologous to *CaN/AINV3* and *CaN/AINV11* (Fig. 1b), likely located on chromosomes 10 and 3, respectively. The gene *CcCWINV4* is phylogenetically close to *CaCWINV4* and *CeCWINV5*, which are located on chromosomes 6 and 6e, respectively (Fig. 1a). Thus, it is highly likely that *CcCWINV4* is also located on chromosome 6 of the *C. canephora* genome.

In the *C. eugenioides* genome (Fig. 4b), the gene *CeN/AINV5* is located on chromosome 11, unlike *C. canephora* and *C. arabica*, which do not have invertase genes on this chromosome. In the phylogenetic tree, *CeN/AINV5* shows high similarity with *CeN/AINV6*, with both genes positioned side by side in the same cluster. The absence of orthologs in the two genomes and their proximity in the phylogeny suggest a possible segmental duplication of

CeN/AINV5 and *CeN/AINV6* after the polyploidization event in *C. arabica*, considering that these genes are located on different chromosomes (1 and 11), this was confirmed by syntenic analyses (Fig. 4b). Asymmetry in gene numbers can also be observed on chromosome 4, where the chromosome 4e of *C. arabica* and the chromosome 4 of *C. canephora* have an additional copy compared to chromosome 4c of *C. arabica* and chromosome 4 of *C. eugenioides*.

Considering the number of orthologous genes, it is noted that *CaINV* genes are mostly derived from *C. eugenioides*. Polyploidization is likely the main cause of the expansion in the number of invertase genes in *C. arabica*. However, due to the smaller number of orthologous gene pairs in *C. canephora*, it can be hypothesized that genome restructuring following polyploidization impacted *CaINV* genes in subCc. Tandemly arranged *CcINV* and *CeINV* genes were observed on chromosomes 6 and 8 of both species, unlike *CaINV* genes, where orthologs were observed on 8c, 6e, and 8e, but not on chromosome 6c. Additionally, two tandemly arranged genes are observed on chromosome 10e (*CaN/AINV1* and *CaN/AINV2*), which is a specificity of *C. arabica*. Orthologs of *CcVINV1* and *CeVINV1* genes are absent in the homeologous chromosomes of *C. arabica*, possibly due to gene deletions or insertions reverting the genes to a diploid state, a phenomenon common in subgenome dominance previously observed in *C. arabica* (Salojärvi et al., 2024).

Expression profile analysis of invertase genes in fruits of commercial cultivars at different developmental stages

To investigate the role of invertase genes in various tissues and developmental stages of *C. arabica* and *C. canephora* fruits, 16 public RNA-Seq libraries were selected from the NCBI-SRA database (Table S2 and Fig. S2). Of these libraries, 12 are from *C. arabica* and 4 are from *C. canephora*. However, libraries of fruit tissues from *C. eugenioides* are not available. The expression levels of 24 *CaINV* genes and 18 *CcINV* genes are summarized in the heatmaps shown in Figure 5 (a, b).

Transcriptome results show that, with the exception of *CaN/AINV1*, invertase genes were expressed in all tissues and fruit stages assessed, but with different spatiotemporal expression patterns. For *C. arabica* (Fig. 5a), clusters of genes with constitutive expression across different developmental stages (*CaN/AINV3*, *CaN/AINV6-8*, and *CaN/AINV10-13*) are observed. A second group of genes, including *CaCWINV10*, *CaCWINV13*, *CaN/AINV5*, *CaVINV1*, and *CaVINV2*, show expression in all tissues but with higher levels in specific developmental stages. Notably, *CaVINV1* shows high expression in the perisperm of early fruit

development stages (30, 60, and 90 DAF - Days After Flowering, Fig. S2), in small grains (2mm), and in the cherry fruit stage (Red drupes).

The cell wall invertase *CaCWINV13* also demonstrated high expression levels, especially in intermediate and advanced developmental stages (Fig. S2), such as perisperm at 120 and 150 DAF, green fruits, and cherry fruits (Green drupes and Red drupes). In the pulp and grain of cherry fruits (Red pulp and Red bean), the highest expression levels were observed for genes *CaN/AINV10* and *CaN/AINV8*, respectively, with *CaN/AINV8* being highly expressed in the grain also at the final ripening stage, Pink bean (Fig. S2). On the other hand, genes *CaCWINV1-6*, *CaCWINV8*, *9*, *11*, *12*, and *CaN/AINV4* exhibited lower and more conditional expression profiles. The gene *CaN/AINV2* showed a slight positive regulation in the grain at the Pink bean ripening stage, while *CaN/AINV1* did not show gene expression in the assessed tissues.

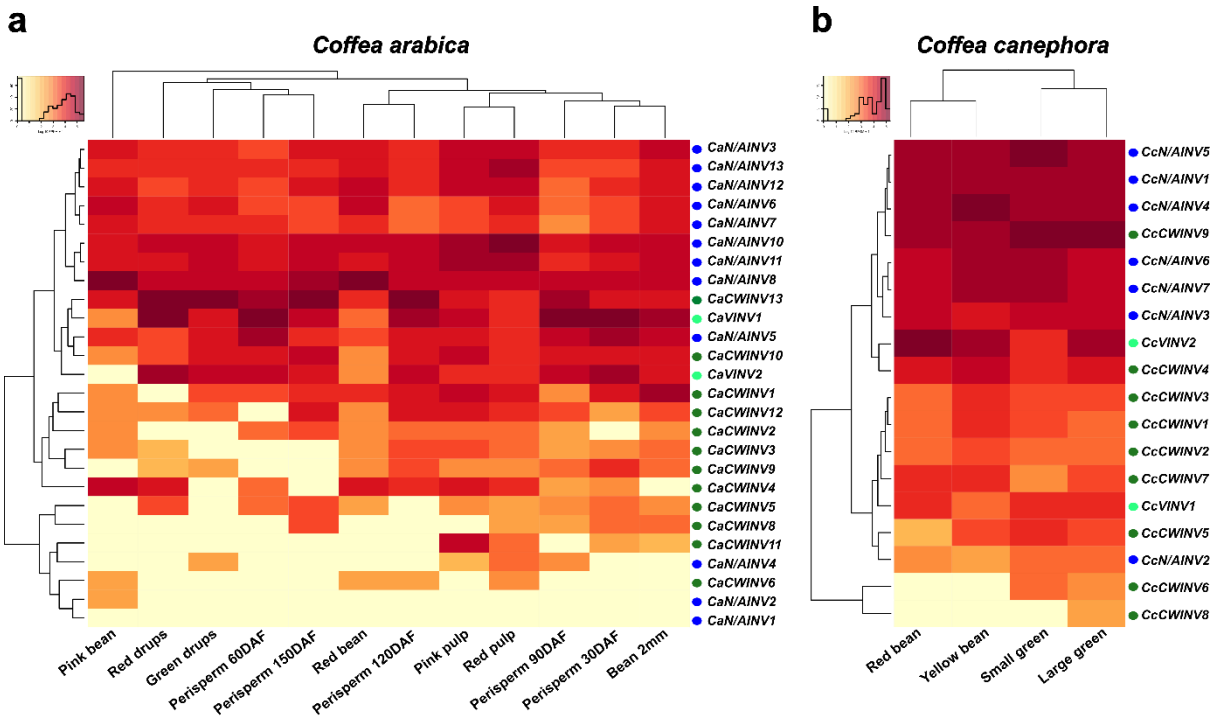


Fig. 5 RNA-Seq expression profile of invertase genes in *Coffea arabica* and *C. canephora* fruits. Hierarchical clustering of expression profiles in different tissues and developmental stages of the fruits is shown. The dendrogram of clustering is displayed at the top and left of each heatmap, and gene names are located on the right side. Dark red indicates high expression levels, while light yellow indicates low levels. (a) The public libraries used for *C. arabica* refer to the following tissues: perisperm 30, 60, 90, 120, and 150 DAF (days after flowering), bean 2mm, green drupes, pink bean, pink pulp, red bean, red pulp, and red drupes. (b) For *C. canephora*, the evaluated tissues were: small green, large green, yellow bean, and red bean. Blue circles indicate neutral/acid invertases, dark green circles indicate cell wall invertases, and light green circles indicate vacuolar invertases. Expression values were calculated as $\log_{10}$ FPKM +1 (upper left graphs). The data were visualized using the R pachage gplots

Similarly, the heatmap for *C. canephora* (Fig.5b) shows clusters with high, medium, and low levels of expression. Among the constitutively expressed genes, N/AINV-encoding genes such as *CcN/AINV1*, *CcN/AINV3-6*, and *CcN/AINV7* predominate, along with *CcCWINV9*, which is positively regulated in the green fruit stages (Small and Large green, Fig. S2), as well as *CcN/AINV5* in the Small green stage. Interestingly, the gene *CcVINV2* shows the highest expression in the red bean stage of the cherry fruit (Red bean), in contrast to its ortholog *CaVINV2*. Conditional expression genes *CcCWINV6* and *CcCWINV8* were observed in the beans at the yellow and cherry fruit stages (Red bean and Yellow bean, Fig. S2).

Alignment of acid invertases and the importance of Asp239 in the interaction with sucrose

The comparison of conserved regions involved in hydrolysis (NDPD/NG), catalytic activity (WECP/V), substrate recognition (Asp239), and enzyme-substrate complex stability (WM/INDPNG, VWGNI, and WSGSAT) among 37 CaINV, CcINV, and CeINV sequences from the GH32 family, through protein sequence alignment, revealed incomplete and absent conserved protein motifs in some sequences (Fig. S3). The seven Arabidopsis acid invertase sequences were used as a reference due to their demonstrated enzymatic activity. For example, AtCWINV1 was recognized as a functional enzyme (Le Roy et al. 2007), while AtCWINV3 and 6, lack invertase activity (De Coninck et al. 2005).

The three tryptophan residues, the essential catalytic motif WECP/V, and the NDPD/NG domain are absent in eight proteins: CaCWINV6 and 7, CcCWINV2 and 6, CeCWINV5, CaVINV2, CcVINV1, and CcVINV2, as previously observed in Fig. 2c. Additionally, concerning the WECP/V domain, the Valine residue was observed in all VINVs, while the Proline residue common in CWINVs was not detected in the CaCWINV6 and CcCWINV6 sequences. It is also important to note that in the CaCWINV7 sequence, these domains are completely absent (Fig. S3 and Fig. 2c).

Sequence comparisons revealed significant differences in the active site regions. For instance, the residue Asp-239 is crucial for invertase activity (Le Roy et al. 2013), along with Lys-242, ensuring efficient sucrose catalysis, as observed in AtCWINV1 (Le Roy et al. 2007). This residue is present in both CWINVs and VINVs; however, 12 sequences lack Asp-239. In the CaCWINV8-11, CcCWINV8, and CeCWINV7-8 sequences, we find Asn instead of Asp-239, which also contributes to the catalytic activity of these enzymes with sucrose, albeit with reduced efficiency (Le Roy et al. 2007).

On the other hand, the CaCWINV12-13, CcCWINV9, and CeCWINV9 sequences do not contain Asp-239, similar to AtCWINV3 and AtCWINV6, whose activities have not been fully clarified; likewise, closely related enzymes such as fructan exohydrolases (FEHs) are characterized by the absence of Asp-239 (Van den Ende et al. 2009). These data suggest that the CaCWINV12-13, CcCWINV9, and CeCWINV9 proteins sequences may have additional functions beyond sucrose breakdown. Interestingly, sequences where Asp-239 is absent or replaced by Asn at the same position are found on the homologous chromosomes 8, 8c, and 8e of the three *Coffea* species (Fig. 4a, b).

To investigate the interaction between CWINVs and sucrose, interaction analyses were performed with CaCWINV1 sequences containing Asp-239 and CaCWINV13 in which this residue is absent. Based on these searches, the selected template for constructing the three-dimensional models of *C. arabica* invertases was the invertase enzyme from *A. thaliana* (details of the protein sequence alignment are shown in Fig. S4), with a crystallographic structure registered in the Protein Data Bank under the code 2AC1. The GMQE and QMEAN scores indicated good model quality (Table S3). For the first generated model, the percentage of identity was 65.73 %, while the second model showed a percentage of identity of 59.69%.

The quality of the models, indicated by GMQE and QMEAN Discrete scores, was assessed to determine if they are reliable and suitable for molecular docking simulations. The GMQE and QMEAN values estimate the accuracy of the obtained tertiary structures and are a measure of model quality. These values range from 0 to 1, with values closer to 1 considered high quality. The GMQE and QMEAN Discrete scores obtained for both models were 0.84 and 0.84 ± 0.05, 0.78 and 0.80 ± 0.05 for the CaCWINV1 and CaCWINV13 enzymes, respectively. From Table S3 and the Ramachandran plot (Fig. S5), it can be observed that almost all amino acid residues are located in allowed regions, with 92.82% and 91.75% favorable residues (Table S3).

To validate the parameters used in docking, a ROC curve was constructed (Fig. S6). The obtained value was above 0.7, indicating that the parameters are satisfactory, with a good balance of sensitivity and specificity for detecting active compounds in the target. After validation, docking calculations were performed for the four molecules in the two studied enzymes: sucrose, 1-kestose, raffinose and stachyose. The results of the interaction energies obtained can be viewed in Table S4 and Figures 6a and 6b.

Figures 6a and 6b show the interaction diagrams obtained for sucrose in both enzymes. As can be seen, sucrose forms two hydrogen bonds with CaCWINV1, while in CaCWINV13, this number decreases to only one bond. The absence of this interaction with residue Arg182

makes the molecule less stable within the site. As observed, the absence of residue Asp239 leads to a structural modification in the active site region, increasing the size of the cavity (Fig. 6c), which affects the conformational accommodation of sucrose within the active sites of the enzymes (Fig. 6d and 6e). It can be observed that sucrose is more snugly fitted in the cavity of CaCWINV1, whereas in CaCWINV13 it has greater conformational freedom and fails to interact with Asp182.

It is noted that the interaction energy of sucrose with the protein containing Asp239 (CaCWINV1) is higher compared to proteins where this residue is absent (CaCWINV13) in the overlay of the two invertases. For CaCWINV1, the interaction energy obtained was -5.7 kcal/mol, while for CaCWINV13, the interaction energy dropped to -5.2 kcal/mol. Since the interaction of CaCWINV13 with sucrose is lower, it may not preferentially bind to sucrose compared to the other tested ligands, as both 1-kestose, raffinose, and stachyose showed higher interaction energy with CaCWINV13 compared to sucrose.

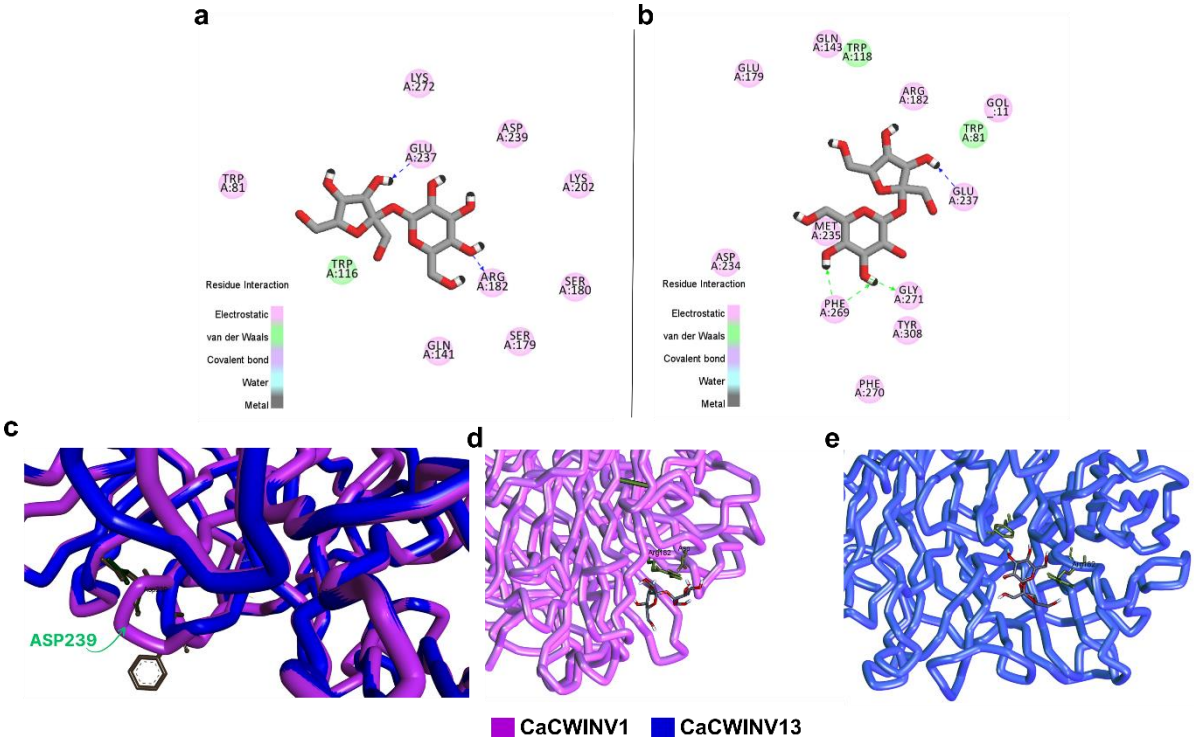


Fig. 6 Representation of the predicted 3D structure models for the active sites of the acid invertases CaCWINV1 and CaCWINV13. (a) and (b) Interaction diagrams obtained for sucrose in CaCWINV1 and CaCWINV13, respectively. (c) Overlay of the invertases CaCWINV1 (light purple) and CaCWINV13 (dark blue). (d) and (e) Conformational accommodation of sucrose within the active sites of the enzymes. Blue arrows indicate hydrogen bonds; green arrows indicate Van der Waals interactions

Discussion

Invertases are involved in essential physiological processes in plants and their mechanisms of action have been increasingly studied. The genomic characterization of these enzymes has been carried out in several species, such as sugarcane (Wang et al. 2017), millet (Zhao et al. 2024), maize (Juárez-Colunga et al. 2018), cucumber (Qi et al. 2023), tobacco (Cheng et al. 2023), tomato (Ahiakpa et al. 2022), wheat (Wang et al. 2023), and in detail in Arabidopsis (De Coninck et al. 2005; Le Roy et al. 2007; Van den Ende et al. 2009), demonstrating the wide conservation of these genes. However, information about these enzymes in coffee is scarce, which motivated our investigation. This study characterized invertase genes in three closely related species: *C. arabica* and modern representatives of its progenitors, *C. canephora* and *C. eugenioides*. This allowed us to investigate the impacts of polyploidization on *C. arabica* invertases compared to its diploid progenitors.

Bioinformatics analyses identified a total of 65 invertase genes, with 28 in *C. arabica*, 18 in *C. canephora*, and 19 in *C. eugenioides*. These genes were classified into acid invertases of the cell wall and vacuole (CWINV and VINV) from the GH32 family, and neutral/alkaline invertases (N/AINV) from the GH100 family, based on phylogenetic analyses. The investigation of genetic structure revealed high conservation in GH100 proteins, as well as in the number of exons and introns, while GH32 proteins demonstrated greater diversity, including the absence of key motifs related to hydrolysis and catalytic activity of invertases on sucrose. Therefore, genes with missing domains (Fig. 2c) may be considered incapable of hydrolyzing sucrose, although the expression of some sequences indicates they might have other functions (Fig. 5a, b). On the other hand, proteins sequences CaCWINV7 and CeN/AINV5 exhibit multiple missing structures, indicating the likely non-functionality of these enzymes (De Coninck et al. 2005; Le Roy et al. 2007; Wan et al. 2018).

Invertase genes are located in highly syntenic regions in the comparisons of subgenomes with their corresponding genomes (Fig. 4a, b), which was expected, as invertases are involved in basal physiological processes of extreme importance in plants (Ruan et al. 2010). The distribution of invertase genes in the three *Coffea* species was uneven across chromosomes, with tandem clusters on chromosomes 6 and 8 in *C. canephora* and *C. eugenioides*, and on chromosomes 6e, 8c and 8e in *C. arabica* (Fig. 4b). Consequently, these regions also showed high collinearity and conservation of CWINVs located there (Fig. 2a), highlighting the importance of further investigations into gene duplication in these regions, a common

phenomenon in invertases (Yao et al. 2014; Qi et al. 2023; Wang et al. 2023; Cheng et al. 2023), as well as the impact of the absence of these homeologs on chromosome 6c of *C. arabica*.

Moreover, we identified a possible segmental duplication of *CeN/AINV5*, which shows high similarity to *CeN/AINV6* (Fig. 1b), as they are located on different chromosomes (11 and 1, respectively). We also observed a tandem duplication of the *CaN/AINV1* gene, which is adjacent to the *CaN/AINV2* gene on chromosome 10e. *CaN/AINV1* has no orthologs in *C. canephora* or *C. eugenioides* (Fig. 4a, b). Thus, these duplications likely occurred after the polyploidization process. Future studies are important to investigate whether these duplications resulted in neofunctionalizations or subfunctionalizations (Adams and Wendel 2005; Ganko et al. 2007; Saul et al. 2023).

Interestingly, *CaN/AINV1* showed no expression in the fruit tissues evaluated (Fig. 5a), indicating that its function may be associated with fruit development stages not investigated in this study, or even with other tissues that were not assessed. This is supported by the fact that *CaN/AINV1* contains two GCN4 cis elements in its promoter region (Fig. 3a), which are involved in regulating the expression of genes coding for storage proteins in grain endosperm (Qu et al. 2023). Thus, this element may be related to divergent expression of *CaN/AINV1* (Haberer et al. 2004).

Due to their potential importance in maintaining cytosolic sugar homeostasis, N/AINV genes show constitutive expression and high conservation in Arabidopsis, rice, maize, and tomato (Wan et al. 2018), which is also observed in N/AINVs of *C. arabica* and *C. canephora* at different fruit development stages. Moreover, the genes *CcN/AINV1* and *CcN/AINV4*, which show higher expression in different fruit stages, contain the specific endosperm cis elements AACA, RY, and GCN4 (Fig. 3b), demonstrating their significant role in the development of *C. canephora* fruits.

Analyzing gene expression, *CaCWINV13* and *CaVINV1* stand out for their higher expression in various fruit development stages of *C. arabica* (Fig. 5a). Similar results were observed in *C. arabica* grains for the expression of *VINV* genes and one *CWINV* (Cheng et al. 2020). In other species, such as tomato, the balance between CWINV and VINV is strongly related to fruit ripening processes (Qin et al. 2016). In wheat, the expression of the *TaCWI50* and *TaVI27* genes are highly related to grain size characteristics (Wang et al. 2023), as is *OsINV3* in rice (Huang et al. 2024). In millet (*Setaria italica*), the *SiCIN7* gene (CWINV) is involved in grain filling was found (Zhao et al. 2024). The similarities observed in these species suggest a conserved role for *CaCWINV13* and *CaVINV1* in fruit development and contribute to

the hypothesis of CWINV selection during the domestication process of *C. arabica* and *C. canephora* (Ruan 2022).

In *C. arabica*, we observed a higher number of invertase genes in the subCe (Fig. 1, Fig. 4, and Table S1), possibly due to the association of these enzymes with basal processes, where the preferential homeologous expression of subCe is common (Vidal et al. 2010). This homeologous expression pattern was also observed in other coffee gene families, such as mannitol genes (de Carvalho et al. 2014), and even in genes coding for proteins integrated into the transcription machinery involved in regulatory networks of essential biological processes such as photosynthesis, cell wall biogenesis, translation, transcription, and catabolism (Ribeiro et al. 2023).

However, the results show that this is not just about gene silencing but also about the absence of homeologous genes in subCc. Since polyploidization generates extensive genomic redundancy that is followed by restructuring, differential gene loss is common in the evolution of polyploids (Adams and Wendel 2005). One hypothesis for the absence of invertase genes in *C. arabica* is extensive genetic modifications caused by transposable elements. Even though duplicated genes are more protected from the deleterious effects of transposition (Soltis and Soltis 1999). This reversion to a diploid state is common in polyploids during the evolutionary process; however, due to the recent polyploidization of *C. arabica*, the percentage of gene deletion is not large (Salojärvi et al. 2024).

The investigation of polymorphisms in genomic sequences associated with specific invertase genes may impact their functionality, affecting sugar content in fruits, both for higher sucrose accumulation (Hazzouri et al. 2019) and reducing sugars (Larsen et al. 2019; Nishio et al. 2021), potentially guiding crop breeding programs. Based on gene expression, *CaCWINV13* and *CaVINV1*, which are orthologs of *C. eugenioides*, are strong candidates for functional studies, given that sugar content in *C. arabica* is highly related to fruit quality (Privat et al. 2008; Cheng et al. 2020; Huang et al. 2021). The fact that these genes are derived from *C. eugenioides* may provide valuable insights into how *C. arabica* developed specificities in chemical compound storage, differentiating itself from *C. canephora*.

In a set of invertases located on homologous chromosomes 8, 8c, and 8e in the three *Coffea* species (CaCWINV12-13, CcCWINV9, and CeCWINV9), the Asp-239 residue is either absent or replaced by Asn. The 3D modeling of the active site of CaCWINV13, which lacks Asp239 (Fig. 6c, d), and CaCWINV1, which contains Asp239 (Fig. 6c, e), revealed that both enzymes interact with raffinose, stachyose, and 1-kestose. This result was expected due to their

function as beta-fructofuranosidases (Ruan et al., 2010), but it is noteworthy that this behavior is only observed in *C. arabica* but also in other non-coffea species, such as Arabidopsis.

As observed, the absence of the Asp239 residue leads to a structural modification in the active site region, increasing the size of its cavity (Fig. 6c). While this residue does not participate in the catalytic reaction, the modification of the cavity space caused by its absence may alter the stability of substrates within it. To quantitatively analyze the change in sucrose stability within the active sites, interaction energies were calculated and presented in Table S4. It is noted that the interaction energy of sucrose with the isoform containing Asp239 (CaCWINV1) is higher compared to the isoform lacking this residue (CaCWINV13).

However, among the four tested substrates, the enzyme CaCWINV1 does not show the best interaction energy value with sucrose (Table S4), but rather raffinose. Raffinose may serve as a transient source of galactose, which is highly demanded in intermediate developmental stages for the synthesis of cell wall storage polysaccharides (Joët et al. 2009). Further investigations into CaCWINV1 could clarify the role of this enzyme in cell wall structuring in coffee fruits, which is important for beverage quality (Li et al. 2021).

Given the above, it can be concluded that few invertase genes may be involved in dynamic and complex processes, as already verified in other species (Yan et al. 2019), since they are key enzymes in sucrose metabolism, essential for metabolic and biosynthetic processes in plant growth and development (Ruan 2014). All invertase genes identified in *C. arabica* and *C. canephora* showed were expressed during fruit development or contained specific endosperm cis elements, broadening the possibilities of studying the protein functionalities of these enzymes in both species.

Studies on invertases could contribute to various research areas in *Coffea*. Investigations in other species has revealed the essential role of invertase genes in drought tolerance (Zhu et al. 2021; Qiao et al. 2024), cold response (Liu et al. 2011; Qian et al. 2018; Feng et al. 2021), pollination, fertilization, and ovule development (Liao et al. 2020; Liu et al. 2024), as well as in increased biomass yield (Zhong et al. 2021). Therefore, our results open many avenues for future studies in *Coffea*, especially for *C. arabica* and *C. canephora*, which have significant economic relevance and particularities regarding disease resistance, tolerance to abiotic stresses, and fruit chemical composition.

Conclusion

In this study, we characterized 65 invertase genes, with 28 in *C. arabica*, 18 in *C. canephora*, and 19 in *C. eugenioides*. Polyploidization is likely the main cause of the expansion in the number of invertase genes in *C. arabica*. However, considering the number of orthologous genes, it is observed that *CaINV* genes are more closely related to those from *C. eugenioides*. Transcriptome results showed that all invertase genes in *C. arabica* and *C. canephora*, except for *CaN/AINV1*, were expressed at different stages of fruit development. However, even though *CaN/AINV1* was not expressed in the evaluated tissues, it contains two GCN4 cis elements, which are involved in regulating the expression of genes related to protein storage in the endosperm. These results suggest that *CaN/AINV1* may participate in tissues or developmental stages not assessed in this study.

Acknowledgements

Special thanks to Luis Carlos da Silva Soares for his support with the use of the R language, and to the members of the Laboratory of Plant Molecular Physiology (LFMP) at the Federal University of Lavras.

References

Adams KL, Wendel JF (2005) Polyploidy and genome evolution in plants. Curr Opin Plant Biol 8:135–141. https://doi.org/10.1016/j.pbi.2005.01.001

Ahiakpa JK, Magdy M, Karikari B, et al (2022) Genome-Wide Identification and Expression Profiling of Tomato Invertase Genes Indicate Their Response to Stress and Phytohormones. J Plant Growth Regul 41:1481–1498. https://doi.org/10.1007/s00344-021-10384-5

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

Bailey TL, Johnson J, Grant CE, Noble WS (2015) The MEME Suite. Nucleic Acids Res 43:W39–W49. https://doi.org/10.1093/nar/gkv416

Berardini TZ, Reiser L, Li D, et al (2015) The arabidopsis information resource: Making and mining the "gold standard" annotated reference plant genome. genesis 53:474–485. https://doi.org/10.1002/dvg.22877

Bergareche D, Royo J, Muñiz LM, Hueros G (2018) Cell wall invertase activity regulates the expression of the transfer cell-specific transcription factor ZmMRP-1. Planta 247:429–442. https://doi.org/10.1007/s00425-017-2800-y

Bienert S, Waterhouse A, de Beer TAP, et al (2017) The SWISS-MODEL Repository—new features and functionality. Nucleic Acids Res 45:D313–D319. https://doi.org/10.1093/nar/gkw1132

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

Braun DM, Wang L, Ruan Y-L (2014) Understanding and manipulating sucrose phloem loading, unloading, metabolism, and signalling to enhance crop yield and food security. J Exp Bot 65:1713–1735. https://doi.org/10.1093/jxb/ert416

Carvalho CHS (2008) Cultivares de café: origem, características e recomendações, 1st edn. Embrapa Café, Brasília, DF

Cheng B, Smyth HE, Furtado A, Henry RJ (2020) Slower development of lower canopy beans produces better coffee. J Exp Bot 71:4201–4214. https://doi.org/10.1093/jxb/eraa151

Cheng L, Jin J, He X, et al (2023) Genome-wide identification and analysis of the invertase gene family in tobacco (*Nicotiana tabacum*) reveals NtNINV10 participating the sugar metabolism. Front Plant Sci 14:. https://doi.org/10.3389/fpls.2023.1164296

Coculo D, Lionetti V (2022) The Plant Invertase/Pectin Methylesterase Inhibitor Superfamily. Front Plant Sci 13:. https://doi.org/10.3389/fpls.2022.863892

de Carvalho K, Petkowicz CLO, Nagashima GT, et al (2014) Homeologous genes involved in mannitol synthesis reveal unequal contributions in response to abiotic stress in *Coffea arabica.* Mol Genet Genomics 289:951–963. https://doi.org/10.1007/s00438-014-0864-y

De Coninck B, Le Roy K, Francis I, et al (2005) Arabidopsis AtcwINV3 and 6 are not invertases but are fructan exohydrolases (FEHs) with different substrate specificities. Plant Cell Environ 28:432–443. https://doi.org/10.1111/j.1365-3040.2004.01281.x

Dobin A, Davis CA, Schlesinger F, et al (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29:15–21. https://doi.org/10.1093/bioinformatics/bts635

Felsenstein J (1989) PHYLIP - phylogeny inference package (v3.2). Cladistics 164–166

Feng Z, Zheng F, Wu S, et al (2021) Functional Characterization of a Cucumber (*Cucumis sativus* L.) Vacuolar Invertase, CsVI1, Involved in Hexose Accumulation and Response to Low Temperature Stress. Int J Mol Sci 22:9365. https://doi.org/10.3390/ijms22179365

Ferrão F, Fonseca AFA da, Ferrão MAG, Muner LHD (2017) Café Conilon, 2nd edn. Incaper, Vitória, ES

Ganko EW, Meyers BC, Vision TJ (2007) Divergence in Expression between Duplicated Genes in Arabidopsis. Mol Biol Evol 24:2298–2309. https://doi.org/10.1093/molbev/msm158

Haberer G, Hindemitt T, Meyers BC, Mayer KFX (2004) Transcriptional Similarities, Dissimilarities, and Conservation of cis-Elements in Duplicated Genes of Arabidopsis. Plant Physiol 136:3009–3022. https://doi.org/10.1104/pp.104.046466

Hazzouri KM, Gros-Balthazard M, Flowers JM, et al (2019) Genome-wide association mapping of date palm fruit traits. Nat Commun 10:4680. https://doi.org/10.1038/s41467-019-12604-9

Hu B, Jin J, Guo A-Y, et al (2015) GSDS 2.0: an upgraded gene feature visualization server. Bioinformatics 31:1296–1297. https://doi.org/10.1093/bioinformatics/btu817

Huang J, Zhou Z, Wang Y, et al (2024) SMS2, a Novel Allele of OsINV3, Regulates Grain Size in Rice. Plants 13:1219. https://doi.org/10.3390/plants13091219

Huang L, Hu L, Fan Y, et al (2021) Comparative transcriptome analysis revealed the influence of sucrose on flavor and taste quality of em *Coffea arabica* em and em *C. canephora* em varieties. Beverage Plant Res 1:1–9. https://doi.org/10.48130/BPR-2021-0011

ICO - INTERNATIONAL COFFEE ORGANIZATION (2023) Coffee Report And Outlook https://icocoffee.org/documents/cy2023-24/Coffee_Report_and_Outlook_December_2023_ICO.pdf Accessed 26 June 2024

Joët T, Laffargue A, Salmona J, et al (2009) Metabolic pathways in tropical dicotyledonous albuminous seeds: *Coffea arabica* as a case study. New Phytol 182:146–162. https://doi.org/10.1111/j.1469-8137.2008.02742.x

Juárez-Colunga S, López-González C, Morales-Elías NC, et al (2018) Genome-wide analysis of the invertase gene family from maize. Plant Mol Biol 97:385–406. https://doi.org/10.1007/s11103-018-0746-5

Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30:3059–3066. https://doi.org/10.1093/nar/gkf436

Larsen B, Migicovsky Z, Jeppesen AA, et al (2019) Genome-Wide Association Studies in Apple Reveal Loci for Aroma Volatiles, Sugar Composition, and Harvest Date. Plant Genome 12:180104. https://doi.org/10.3835/plantgenome2018.12.0104

Le Roy K, Lammens W, Verhaest M, et al (2007) Unraveling the Difference between Invertases and Fructan Exohydrolases: A Single Amino Acid (Asp-239) Substitution Transforms Arabidopsis Cell Wall Invertase1 into a Fructan 1-Exohydrolase. Plant Physiol 145:616–625. https://doi.org/10.1104/pp.107.105049

Le Roy K, Vergauwen R, Struyf T, et al (2013) Understanding the Role of Defective Invertases in Plants: Tobacco Nin88 Fails to Degrade Sucrose. Plant Physiol 161:1670–1681. https://doi.org/10.1104/pp.112.209460

Lescot M, Déhais P, Thijs G, et al (2002) PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. Nucleic Acids Res 30:325–327. https://doi.org/10.1093/nar/30.1.325

Letunic I, Bork P (2021) Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res 49:W293–W296. https://doi.org/10.1093/nar/gkab301

Li Z, Zhang C, Zhang Y, et al (2021) Coffee cell walls—composition, influence on cup quality and opportunities for coffee improvements. Food Qual Saf 5:fyab012. https://doi.org/10.1093/fqsafe/fyab012

Liao S, Wang L, Li J, Ruan Y-L (2020) Cell Wall Invertase Is Essential for Ovule Development through Sugar Signaling Rather Than Provision of Carbon Nutrients1 [OPEN]. Plant Physiol 183:1126–1144. https://doi.org/10.1104/pp.20.00400

Liu H, Yao X, Fan J, et al (2024) Cell wall invertase 3 plays critical roles in providing sugars during pollination and fertilization in cucumber. Plant Physiol 195:1293–1311. https://doi.org/10.1093/plphys/kiae119

Liu X, Zhang C, Ou Y, et al (2011) Systematic analysis of potato acid invertase genes reveals that a cold-responsive member, StvacINV1, regulates cold-induced sweetening of tubers. Mol Genet Genomics 286:109–118. https://doi.org/10.1007/s00438-011-0632-1

Marçais G, Delcher AL, Phillippy AM, et al (2018) MUMmer4: A fast and versatile genome alignment system. PLOS Comput Biol 14:e1005944. https://doi.org/10.1371/journal.pcbi.1005944

Morris GM, Huey R, Lindstrom W, et al (2009) AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. J Comput Chem 30:2785–2791. https://doi.org/10.1002/jcc.21256

Nishio S, Hayashi T, Shirasawa K, et al (2021) Genome-wide association study of individual sugar content in fruit of Japanese pear (*Pyrus* spp.). BMC Plant Biol 21:378. https://doi.org/10.1186/s12870-021-03130-2

Pan L, Guo Q, Chai S, et al (2019) Evolutionary Conservation and Expression Patterns of Neutral/Alkaline Invertases in *Solanum*. Biomolecules 9:763. https://doi.org/10.3390/biom9120763

Pang WC, Ramli ANM, Johari ND (2019) Structural Properties, Production, and Commercialisation of Invertase. Sains Malays 48:523–531. https://doi.org/10.17576/jsm-2019-4803-04

Partelli FL, Espindula MC, Marré WB, Vieira HD (2014) Dry matter and macronutrient accumulation in fruits of Conilon coffee with different ripening cycles. Rev Bras Ciênc Solo 38:214–222. https://doi.org/10.1590/S0100-06832014000100021

Piccirillo E, Amaral AT do (2018) VIRTUAL SCREENING OF BIOACTIVE COMPOUNDS: CONCEPTS AND APLICATIONS. Quím Nova 41:662–677. https://doi.org/10.21577/0100-4042.20170210

Privat I, Foucrier S, Prins A, et al (2008) Differential regulation of grain sucrose accumulation and metabolism in Coffea arabica (Arabica) and Coffea canephora (Robusta)

revealed through gene expression and enzyme activity analysis. New Phytol 178:781–797. https://doi.org/10.1111/j.1469-8137.2008.02425.x

Qi C, Xv L, Xia W, et al (2023) Genome-Wide Identification and Expression Patterns of Cucumber Invertases and Their Inhibitor Genes. Int J Mol Sci 24:13421. https://doi.org/10.3390/ijms241713421

Qian W, Xiao B, Wang L, et al (2018) CsINV5, a tea vacuolar invertase gene enhances cold tolerance in transgenic Arabidopsis. BMC Plant Biol 18:228. https://doi.org/10.1186/s12870-018-1456-5

Qiao K, Zeng Q, Lv J, et al (2024) Exploring the role of GhN/AINV23: implications for plant growth, development, and drought tolerance. Biol Direct 19:1–10. https://doi.org/10.1186/s13062-024-00465-2

Qin G, Zhu Z, Wang W, et al (2016) A Tomato Vacuolar Invertase Inhibitor Mediates Sucrose Metabolism and Influences Fruit Ripening. Plant Physiol 172:1596–1611. https://doi.org/10.1104/pp.16.01269

Qu G, Wang K, Mu J, et al (2023) Identifying cis-Acting Elements Associated with the High Activity and Endosperm Specificity of the Promoters of Genes Encoding Low-Molecular-Weight Glutenin Subunits in Common Wheat (Triticum aestivum). J Agric Food Chem 71:17432–17441. https://doi.org/10.1021/acs.jafc.3c04209

Ramachandran GN, Ramakrishnan C, Sasisekharan V (1963) Stereochemistry of polypeptide chain configurations. J Mol Biol 7:95–99. https://doi.org/10.1016/S0022-2836(63)80023-6

Ribeiro THC, Oliveira RR de, Antonio C-J (2023) Differential universal ortholog composition of *Coffea arabica* L. sub-genomes and its contribution to regulatory networks governing essential biological processes. 2023.03.08.531780

Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26:139–140. https://doi.org/10.1093/bioinformatics/btp616

Ruan Y-L (2014) Sucrose Metabolism: Gateway to Diverse Carbon Use and Sugar Signaling. Annu Rev Plant Biol 65:33–67. https://doi.org/10.1146/annurev-arplant-050213-040251

Ruan Y-L (2022) CWIN-sugar transporter nexus is a key component for reproductive success. J Plant Physiol 268:153572. https://doi.org/10.1016/j.jplph.2021.153572

Ruan Y-L, Jin Y, Yang Y-J, et al (2010) Sugar Input, Metabolism, and Signaling Mediated by Invertase: Roles in Development, Yield Potential, and Response to Drought and Heat. Mol Plant 3:942–955. https://doi.org/10.1093/mp/ssq044

Salojärvi J, Rambani A, Yu Z, et al (2024) The genome and population genomics of allopolyploid *Coffea arabica* reveal the diversification history of modern coffee cultivars. Nat Genet 56:721–731. https://doi.org/10.1038/s41588-024-01695-w

Saul F, Scharmann M, Wakatake T, et al (2023) Subgenome dominance shapes novel gene evolution in the decaploid pitcher plant Nepenthes gracilis. Nat Plants 9:2000–2015. https://doi.org/10.1038/s41477-023-01562-2

Scalabrin S, Magris G, Liva M, et al (2024) A chromosome-scale assembly reveals chromosomal aberrations and exchanges generating genetic diversity in *Coffea arabica* germplasm. Nat Commun 15:463. https://doi.org/10.1038/s41467-023-44449-8

Scalabrin S, Toniutti L, Di Gaspero G, et al (2020) A single polyploidization event at the origin of the tetraploid genome of *Coffea arabica* is responsible for the extremely low genetic variation in wild and cultivated germplasm. Sci Rep 10:4642. https://doi.org/10.1038/s41598-020-61216-7

Sela I, Ashkenazy H, Katoh K, Pupko T (2015) GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. Nucleic Acids Res 43:W7–W14. https://doi.org/10.1093/nar/gkv318

Sobolev OV, Afonine PV, Moriarty NW, et al (2020) A Global Ramachandran Score Identifies Protein Structures with Unlikely Stereochemistry. Structure 28:1249-1258.e2. https://doi.org/10.1016/j.str.2020.08.005

Soderlund C, Bomhoff M, Nelson WM (2011) SyMAP v3.4: a turnkey synteny system with application to plant genomes. Nucleic Acids Res 39:e68. https://doi.org/10.1093/nar/gkr123

Soltis DE, Soltis PS (1999) Polyploidy: recurrent formation and genome evolution. Trends Ecol Evol 14:348–352. https://doi.org/10.1016/S0169-5347(99)01638-9

Studer G, Rempfer C, Waterhouse AM, et al (2020) QMEANDisCo—distance constraints applied on model quality estimation. Bioinformatics 36:1765–1771. https://doi.org/10.1093/bioinformatics/btz828

Trott O, Olson AJ (2010) AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem 31:455–461. https://doi.org/10.1002/jcc.21334

Van den Ende W, Lammens W, Van Laere A, et al (2009) Donor and acceptor substrate selectivity among plant glycoside hydrolase family 32 enzymes. FEBS J 276:5788–5798. https://doi.org/10.1111/j.1742-4658.2009.07316.x

Vidal RO, Mondego JMC, Pot D, et al (2010) A High-Throughput Data Mining of Single Nucleotide Polymorphisms in *Coffea* Species Expressed Sequence Tags Suggests Differential Homeologous Gene Expression in the Allotetraploid Coffea arabica. Plant Physiol 154:1053–1066. https://doi.org/10.1104/pp.110.162438

Wan H, Wu L, Yang Y, et al (2018) Evolution of Sucrose Metabolism: The Dichotomy of Invertases and Beyond. Trends Plant Sci 23:163–177. https://doi.org/10.1016/j.tplants.2017.11.001

Wang C, Wang G, Wen X, et al (2023) Characteristics and Expression Analysis of Invertase Gene Family in Common Wheat (Triticum aestivum L.). Genes 14:41. https://doi.org/10.3390/genes14010041

Wang L, Zheng Y, Ding S, et al (2017) Molecular cloning, structure, phylogeny and expression analysis of the invertase gene family in sugarcane. BMC Plant Biol 17:109. https://doi.org/10.1186/s12870-017-1052-0

Waterhouse A, Bertoni M, Bienert S, et al (2018) SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res 46:W296–W303. https://doi.org/10.1093/nar/gky427

Wintgens JN (2012) Coffee - Growing, Processing, Sustainable Production: A Guidebook for Growers, Processors, Traders and Researchers. John Wiley & Sons

Xu B, Timko M (2004) Methyl jasmonate induced expression of the tobacco putrescine N-methyltransferase genes requires both G-box and GCC-motif elements. Plant Mol Biol 55:743–761. https://doi.org/10.1007/s11103-004-1962-8

Yan W, Wu X, Li Y, et al (2019) Cell Wall Invertase 3 Affects Cassava Productivity via Regulating Sugar Allocation From Source to Sink. Front Plant Sci 10:. https://doi.org/10.3389/fpls.2019.00541

Yao Y, Geng M-T, Wu X-H, et al (2014) Genome-Wide Identification, 3D Modeling, Expression and Enzymatic Activity Analysis of Cell Wall Invertase Gene Family from Cassava (*Manihot esculenta* Crantz). Int J Mol Sci 15:7313–7331. https://doi.org/10.3390/ijms15057313

Zhang Y, Chen C, Jin X-F, et al (2009) Expression of a rice *DREB1* gene, *OsDREB1D*, enhances cold and high-salt tolerance in transgenic Arabidopsis. BMB Rep 42:486–492. https://doi.org/10.5483/BMBRep.2009.42.8.486

Zhang Y, Sun T, Liu S, et al (2016) MYC cis-Elements in PsMPT Promoter Is Involved in Chilling Response of Paeonia suffruticosa. PLOS ONE 11:e0155780. https://doi.org/10.1371/journal.pone.0155780

Zhao Y, Wang T, Wan S, et al (2024) Genome-wide identification and functional analysis of the *SiCIN* gene family in foxtail millet (*Setaria italica* L.). Gene 921:148499. https://doi.org/10.1016/j.gene.2024.148499

Zhong H-L, Liu Y, Nie Y-D, et al (2021) Change of soluble acid invertase gene (SAI-1) haplotype in hybrid sorghum breeding program in China. Mol Breed 41:37. https://doi.org/10.1007/s11032-021-01231-2

Zhu C, Yang K, Li G, et al (2021) Identification and Expression Analyses of Invertase Genes in Moso Bamboo Reveal Their Potential Drought Stress Functions. Front Genet 12:. https://doi.org/10.3389/fgene.2021.696300

Statements and Declarations

Funding

This work was financially supported by the the “Coordenação de Aperfeiçoamento de Pessoal de Nível Superior” (CAPES) for the doctoral scholarship for G.E.F (88887.568926/2020-00); the “Conselho Nacional de Desenvolvimento Científico e Tecnológico” (CNPq), for scholarship research for A.C.J. under grant number 309005/2022-1, and for F.M.A.G under grant number 317001/2021-3; the “Instituto Nacional de Ciência e Tecnologia do Café” (INCT/Café) under the grant of the “Fundação de Amparo à Pesquisa do Estado de Minas Gerais” (FAPEMIG, CAP APQ 03605/17) and scholarship research for A.C.J. (APQ-03406-18); and “Fundação de Amparo à Pesquisa do Estado de São Paulo” (FAPESP, #21/06968-3).

Declaration

The authors declare that they have no conflict of interest.

Author contributions

GEF Conceptualization, conducted the analyses of the data and bioinformatics analyses, wrote the original manuscript. RMG Conceptualization and supported the bioinformatics analyses revised and edited the original manuscript. TYI Conceptualization, revised and edited the original manuscript. GCR Conducted the syntenic analyses and revised the final manuscript. TS and TCR Conducted homology modeling and structure prediction analyses. EN Supported the bioinformatics analyses and contributed to the revision of the manuscript. DZ Revised and edited the final manuscript. FMAG and ACJ Provided supervision throughout the study, revised and edited the final manuscript. All authors read and approved the manuscript.

Data availability

Main data supporting the findings of this study are available in the manuscript and supplementary materials published online.

Abbreviations

GH32	Glycoside hydrolase 32
GH100	Glycoside hydrolase 100
CaINV	*Coffea arabica* invertase
CcINV	*C. canephora* invertase

Cc x subCc	Genome of *C. canephora* compared to the corresponding subgenome in *C. arabica*
CDS	Coding sequences
CeINV	*C. eugenioides* invertase
Ce x subCe	Genome of *C. eugenioides*compared to the corresponding subgenome in *C. arabica*
CWINV	Cell wall invertases
N/AINV	Neutral/alkaline invertases
SubCc	Subgenome derived from *C. canephora* in *C. arabica*
SubCe	Subgenome derived from *C. eugenioides* in *C. arabica*
VINV	Vacuolar invertases

Genome-wide characterization of invertases in Arabica coffee and your parents reveals putative genes involved in fruit development

Plant Cell Reports

Gabriela Ester Ferraz, Robert Márquez Gutiérrez, Tiago Yukio Inoue, Gabriel de Campos Rume, Thais Sales, Teodorico de Castro Ramalho, Evandro Novaes, Flavia Maria Avelar Gonçalves, Dapeng Zhang, Antonio Chalfun-Junior[1*]

*Corresponding Author:

[1] Institute of Natural Sciences, Department of Biology, Laboratory of Plant Molecular Physiology, Plant Physiology Sector, Federal University of Lavras, Lavras, Minas Gerais, Brazil, 37200900, MG, Brazil, chalfunjunior@ufla.br

Supplementary Material

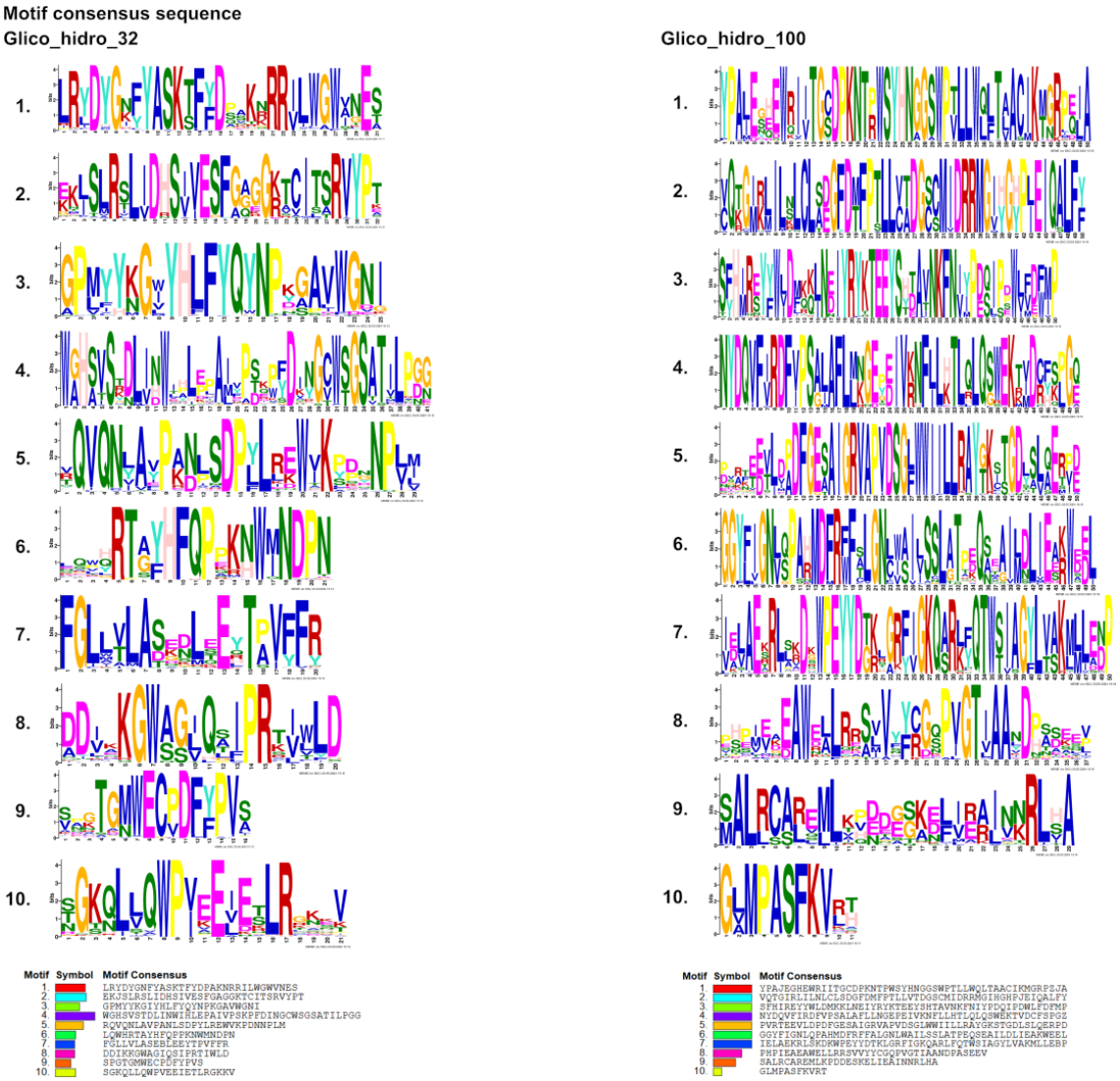


Fig. S1 Conserved amino acid sequences of invertases from *Coffea arabica*, *C. canephora*, and *C. eugenioides*. Glycosyl hydrolase 32 motifs (10) are located on the left side of the figure, while glycosyl hydrolase 100 motifs (10) are found on the right side. The images were generated using the MEME tool

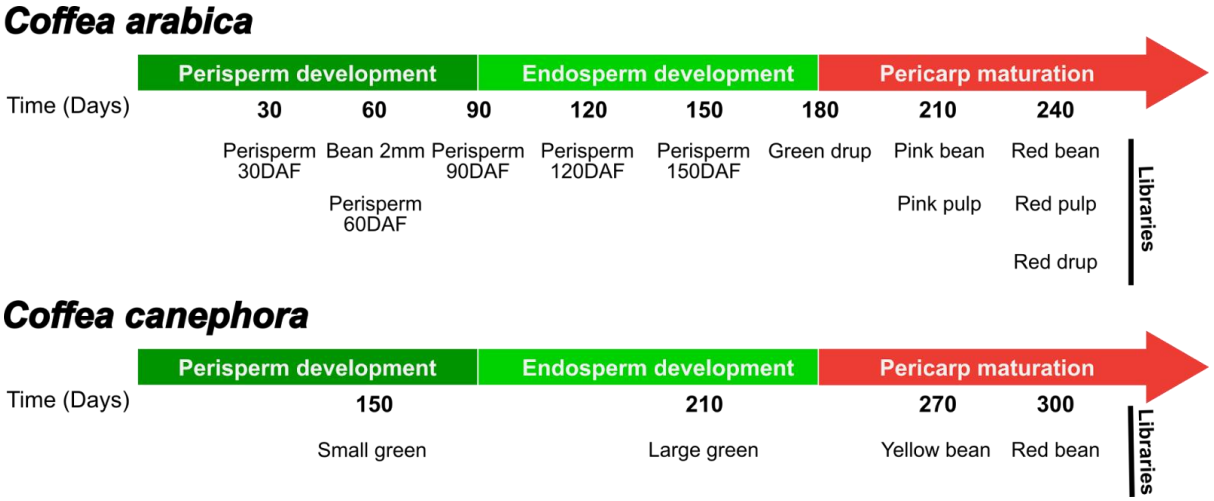


Fig. S2 Distribution of selected RNA-Seq libraries for the expression profiling of invertase genes at different fruit development stages in *Coffea arabica* and *C. canephora*. Public libraries available in the NCBI-SRA database. (Development stages are distributed according to Privat et al., 2008). DAF: days after flowering. The diagram was created using Inkscape software

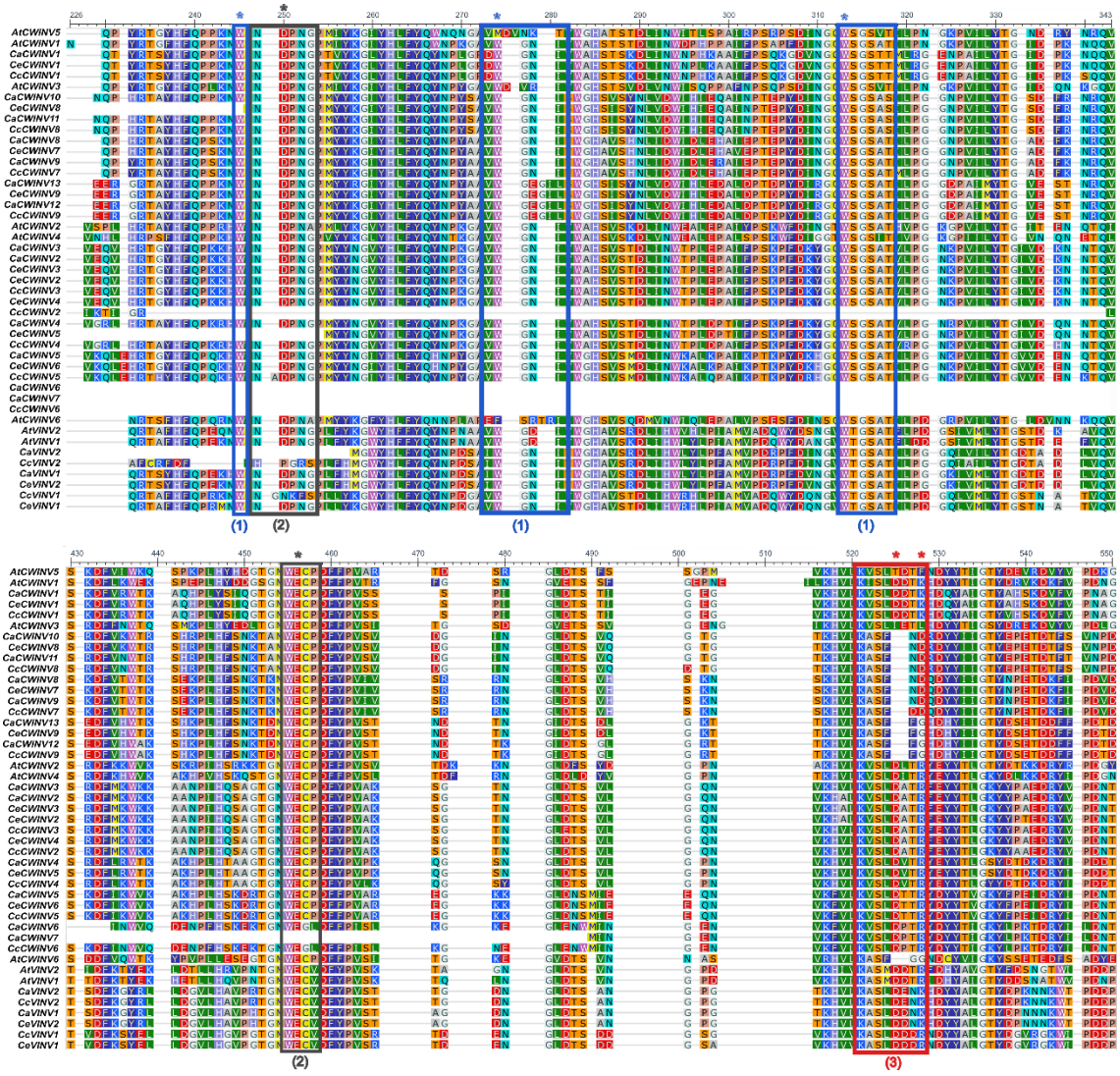


Fig. S3 Alignment of highly conserved regions in acidic invertase proteins from *Arabidopsis thaliana*, *Coffea arabica*, *C. canephora*, and *C. eugenioides*, visualized with Multiple Sequence Alignment Viewer - NCBI. (1) Tryptophan (Trp) residues stabilizing the enzyme-substrate complex (WM/INDPNG, VWGNI, and WSGSAT). (2) Motifs related to hydrolysis and catalytic activity of acidic invertases (NDPD/NG and WECV/P). (3) Aspartate (Asp-239) and lysine (Lys-242) residues related to substrate recognition. Asterisks indicate the predicted active residues for the invertases. The image was visualized using the NCBI Multiple Sequence Alignment Viewer

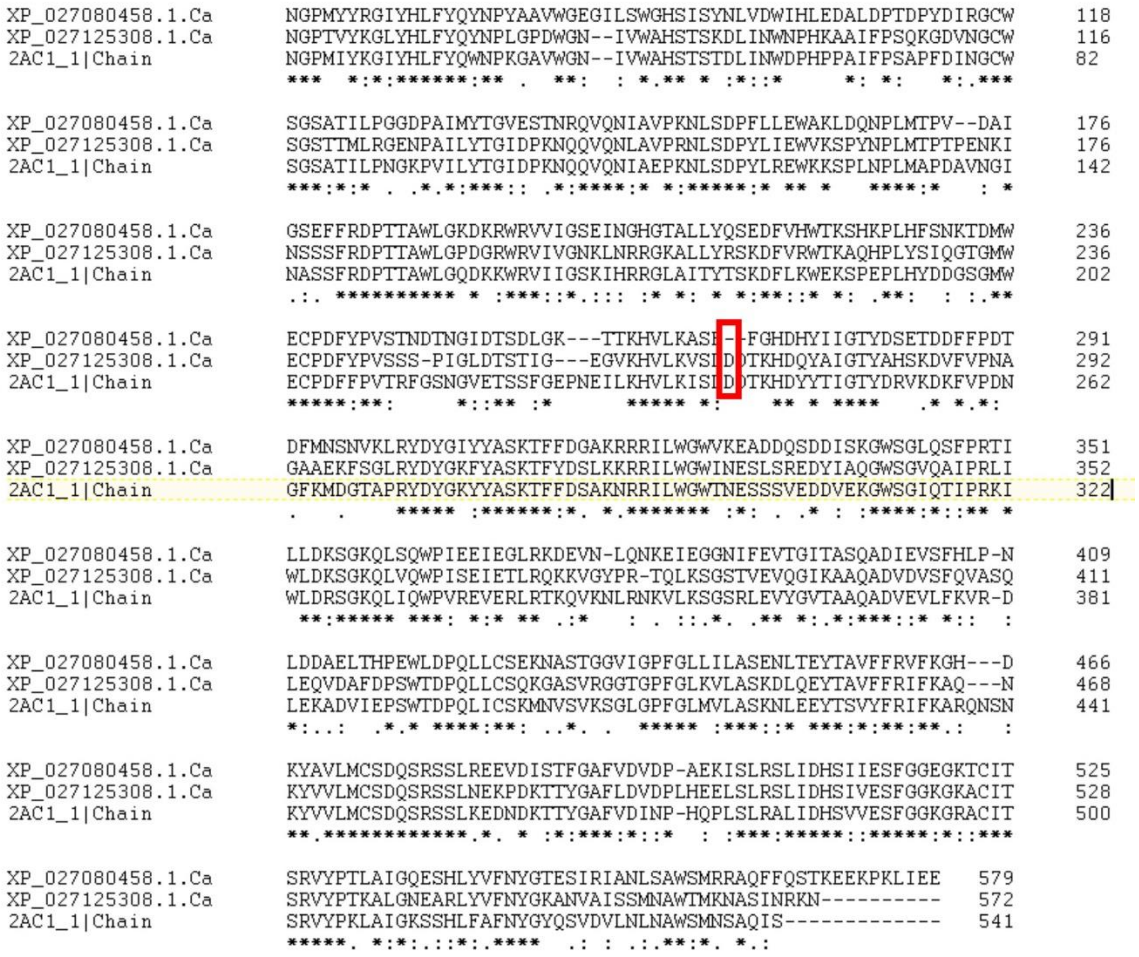


Fig. S4 Alignment of the protein sequences of CaCWINV1 (XP_027125308.1.Ca) and CaCWINV13 (XP_027080458.1.Ca) with the *Arabidopsis thaliana* template sequence used for enzyme modeling. The red square highlights Asp239 present in CaCWINV1, in the *Arabidopsis* sequence, and absent in CaCWINV13

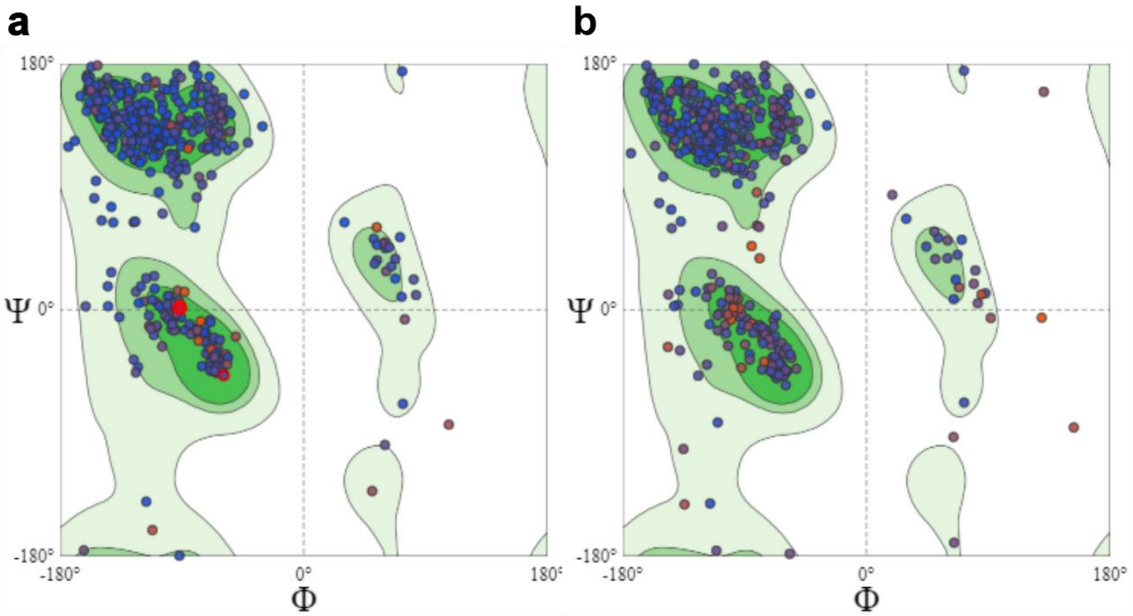


Fig. S5 Ramachandran plot for the generated models. (a) Model 1 - CaCWINV1, with Asp239 and Model 2 - CaCWINV13, without Asp239

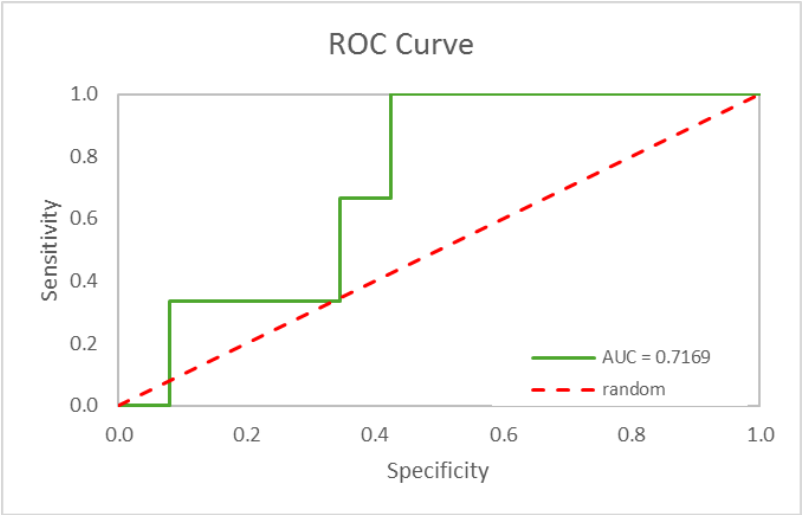


Fig. S6 ROC curve obtained for target 1.

Table S1. Physico-chemical characteristics and properties of invertase gene family in *Coffea arabica*, *C. canephora* and *C. eugenioides*.

Species	Gene name [a]	gen_ID [b]	Chromosome location [c]	transcript_ID [b]	protein_ID [b]	Size (bp)	CDS length	Number of exons	Strand	Protein length (aa)	Mw [d]	pI [e]	Conserved domain	Plant-mPloc
C. arabica	*CaVINV1*	113730173	Chr2e: 3533562..3538907	XM027254686.1	XP027110487.1	5346	1986	7	(-)	661	72756	5.27	Glyco_ hydro_32	Vacuole
	CaVINV2	113725148	Chr2c: 3339187..3344713	XM027248144.1	XP027103945.1	5527	1527	7	(-)	508	55741.9	5.22	Glyco_ hydro_32	Vacuole
	CaCWINV1	113741863	Chr4e: 1981490..1985160	XM027269507.1	XP027125308.1	3671	1719	8	(-)	572	64391.1	9.24	Glyco_ hydro_32	Cell wall
	CaCWINV2	113697414	Chr6e: 5050950..5054193	XM027217018.1	XP027072819.1	3244	1728	7	(+)	575	64719.7	9.1	Glyco_ hydro_32	Cell wall
	CaCWINV3	113695374	Chr6e: 5062010..5065256	XM027214453.1	XP027070254.1	3247	1728	6	(+)	575	64928.8	9.2	Glyco_ hydro_32	Cell wall
	CaCWINV4	113695416	Chr6e: 5066610..5070580	XM027214511.1	XP027070312.1	3971	1740	7	(+)	579	65103.5	7.24	Glyco_ hydro_32	Cell wall
	CaCWINV5	113697412	Ch6e:: 5046562..5050109	XM027217016.1	XP027072817.1	3548	1755	7	(+)	584	66197.7	8.86	Glyco_ hydro_32	Cell wall
	CaCWINV6	113740840	Chr4e: 15166807..15169231	XM027268358.1	XP027124159.1	2425	1107	5	(+)	368	41752.8	8.48	Glyco_ hydro_32	Cell wall
	CaCWINV7	113739049	Chr4c: 14302966..14305866	-	-	2119	960	6	(+)	319	35329.5	9.09	Glyco_ hydro_32	Cell wall
	CaCWINV8	113703453	Chr8e: 40107424..40110324	XM027224821.1	XP027080622.1	2901	1728	7	(+)	575	65478.5	9.06	Glyco_ hydro_32	Cell wall
	CaCWINV9	113706320	Chr8c: 34385109..34388009	XM027228201.1	XP027084002.1	2901	1728	7	(+)	575	65576.7	9.05	Glyco_ hydro_32	Cell wall
	CaCWINV10	113702893	Chr8e: 40098044..40101000	XM027224053.1	XP027079854.1	2957	1725	7	(+)	574	65789.3	6.42	Glyco_ hydro_32	Cell wall
	CaCWINV11	113706424	Chr8c: 34368115..34371614	XM027228326.1	XP027084127.1	3500	1725	7	(+)	574	65779.4	6.23	Glyco_ hydro_32	Cell wall
	CaCWINV12	113706360	Chr8c: 34362267..34365807	XM027228250.1	XP027084051.1	3541	1743	7	(+)	580	65729.7	4.85	Glyco_ hydro_32	Cell wall
	CaCWINV13	113703335	Chr8e: 40091733..40095196	XM027224657.1	XP027080458.1	3464	1740	7	(+)	579	65743.6	4.85	Glyco_ hydro_32	Cell wall
	CaN/AINV1	113712501	Chr10e: 12387294..12394499	XM027235963.1	XP027091764.1	7206	1677	5	(-)	558	63852.5	5.87	Glyco_ hydro_100	Chloroplastic
	CaN/AINV2	113712468	Chr10e: 12362885..12370113	XM027235909.1	XP027091710.1	7229	1677	5	(-)	558	63852.5	5.87	Glyco_ hydro_100	Chloroplastic
	CaN/AINV3	113714619	Chr10c: 9976294..9982945	XM027238555.1	XP027094356.1	6652	1677	4	(-)	558	63824.4	5.87	Glyco_ hydro_100	Chloroplastic
	CaN/AINV4	113729518	Chr2e: 71110714..71114035	XM027253807.1	XP027109608.1	3322	1833	4	(+)	610	68301.3	7.9	Glyco_ hydro_100	Chloroplastic
	CaN/AINV5	113731178	Chr2e: 14016405..14021622	XM027256280.1	XP027112081.1	5218	1713	5	(+)	570	65132.8	5.96	Glyco_ hydro_100	Chloroplastic
	CaN/AINV6	113725687	Chr2c: 7886514..7892630	XM027248984.1	XP027104785.1	6117	1962	7	(-)	653	73227.8	7.49	Glyco_ hydro_100	Chloroplastic
	CaN/AINV7	113730692	Chr2e: 8054560..8060679	XM027255551.1	XP027111352.1	6120	1962	8	(-)	653	73269.9	7.49	Glyco_ hydro_100	Chloroplastic
	CaN/AINV8	113706295	Chr1e: 38504385..38510098	XM027228171.1	XP027083972.1	5714	1923	6	(-)	640	72033.7	5.97	Glyco_ hydro_100	Chloroplastic
	CaN/AINV9	113742072	Chr1c: 41803600..41809871	-	-	4634	1704	8	(-)	567	63677.9	8.02	Glyco_ hydro_100	Chloroplastic
	CaN/AINV10	113738022	Chr3e: 361353..366456	XM027265172.1	XP027120973.1	5104	2085	6	(+)	694	78609.6	5.78	Glyco_ hydro_100	Chloroplastic
	CaN/AINV11	113735307	Chr3c: 345796..350442	XM027262327.1	XP027118128.1	4647	1425	7	(+)	474	53940.8	5.93	Glyco_ hydro_100	Chloroplastic

Continuation ...

	CaN/AINV12	113727822	Chr2c: 62048995..62053642	XM027252183.1	XP027107984.1	4648	2016	6	(-)	671	76172.5	6.63	Glyco_ hydro_100	Chloroplastic
	CaN/AINV13	113732993	Chr2e: 67523839..67528613	XM027259085.1	XP027114886.1	4775	2016	6	(-)	671	76103.4	6.81	Glyco_ hydro_100	Chloroplastic
C. canephora	*CcVINV1*	GSCOCG00030409001	Chr8: 28193689.. 28198078	GSCOCT00030409001	Cc08t12930.1	4390	2049	6	(+)	682	75180.7	5.52	Glyco_ hydro_32	Vacuole
	CcVINV2	GSCOCG00020322001	Chr2: 3313467.. 3318049	GSCOCT00020322001	Cc02t04180.1	4583	1620	7	(-)	539	59349.2	5.57	Glyco_ hydro_32	Vacuole
	CcCWINV1	GSCOCG00021847001	Chr4: 1407049.. 1410416	GSCOCT00021847001	Cc04t01850.1	3368	1731	7	(-)	576	64661.6	9.17	Glyco_ hydro_32	Cell wall
	CcCWINV2	GSCOCG00043182001	Chr6:4807368.. 4809104	GSCOCT00043182001	Cc06t06110.1	1737	1158	5	(+)	385	43124.1	9.25	Glyco_ hydro_32	Cell wall
	CcCWINV3	GSCOCG00043180001	Chr6: 4794793.. 4797501	GSCOCT00043180001	Cc06t06090.1	2709	1728	7	(+)	575	64836.7	9.05	Glyco_ hydro_32	Cell wall
	CcCWINV4	GSCOCG00000738001	Chr0:108046114..108049867	GSCOCT00000738001	Cc00t15520.1	3754	1743	7	(-)	580	65178.6	7.28	Glyco_ hydro_32	Cell wall
	CcCWINV5	GSCOCG00043179001	Chr6: 4789900.. 4794018	GSCOCT00043179001	Cc06t06080.1	4119	1758	7	(+)	585	66561	8.88	Glyco_ hydro_32	Cell wall
	CcCWINV6	GSCOCG00011372001	Chr4: 15552644.. 15554986	GSCOCT00011372001	Cc04t13060.1	2343	1158	5	(+)	385	43720.9	8.8	Glyco_ hydro_32	Cell wall
	CcCWINV7	GSCOCG00027054001	Chr8: 25577059.. 25579852	GSCOCT00027054001	Cc08t10760.1	2794	1728	7	(+)	575	65547.6	8.79	Glyco_ hydro_32	Cell wall
	CcCWINV8	GSCOCG00027052001	Chr8: 25559243.. 25562087	GSCOCT00027052001	Cc08t10750.1	2845	1725	7	(+)	574	65849.4	6.15	Glyco_ hydro_32	Cell wall
	CcCWINV9	GSCOCG00027051001	Chr8: 25551014.. 25554508	GSCOCT00027051001	Cc08t10740.1	3495	1740	7	(+)	579	65658.6	4.88	Glyco_ hydro_32	Cell wall
	CcN/AINV1	GSCOCG00005837001	Chr0: 51965037.. 51971952	GSCOCT00005837001	Cc00t06590.1	6916	1617	4	(-)	538	61513.8	5.91	Glyco_ hydro_100	Chloroplastic
	CcN/AINV2	GSCOCG00027695001	Chr2: 54095168.. 54098428	GSCOCT00027695001	Cc02t39590.1	3261	1933	4	(+)	610	68295.3	6.49	Glyco_ hydro_100	Chloroplastic
	CcN/AINV3	GSCOCG00014151001	Chr2: 15303680.. 15308958	GSCOCT00014151001	Cc02t16580.1	5279	1632	6	(+)	543	62078.2	5.68	Glyco_ hydro_100	Chloroplastic
	CcN/AINV4	GSCOCG00029592001	Chr2: 8052190.. 8058349	GSCOCT00029592001	Cc02t09920.1	6160	1959	7	(-)	652	73140.7	7.17	Glyco_ hydro_100	Chloroplastic
	CcN/AINV5	GSCOCG00024019001	Chr1: 28266297.. 28271562	GSCOCT00024019001	Cc01t09590.1	5266	1923	7	(-)	640	72008.6	5.96	Glyco_ hydro_100	Chloroplastic
	CcN/AINV6	GSCOCG00009601001	Chr0: 135297384.. 135302934	GSCOCT00009601001	Cc00t21200.1	5551	2100	6	(-)	699	79141.1	5.91	Glyco_ hydro_100	Chloroplastic
	CcN/AINV7	GSCOCG00042866001	Chr2: 48999886.. 49004681	GSCOCT00042866001	Cc02t35310.1	4796	2016	6	(-)	671	76202.5	6.63	Glyco_ hydro_100	Chloroplastic
C. eugenioides	*CeVINV1*	113780858	Chr8: 43876705..43881084	XM027326641.1	XP027182442.1	4380	2064	7	(+)	687	75800.5	5.38	Glyco_ hydro_32	Vacuole
	CeVINV2	113762384	Chr2: 4857769..4863029	XM027305825.1	XP027161626.1	5261	1986	7	(-)	661	72759	5.28	Glyco_ hydro_32	Vacuole
	CeCWINV1	113767641	Chr4: 1506148..1509579	XM027311799.1	XP027167600.1	3432	1728	7	(-)	575	64785.6	9.28	Glyco_ hydro_32	Cell wall
	CeCWINV2	113775492	Chr6: 4340899..4343905	XM027320394.1	XP027176195.1	3007	1728	6	(-)	575	64872.7	9.16	Glyco_ hydro_32	Cell wall
	CeCWINV3	113775009	Chr6: 4354553..4357618	XM027319713.1	XP027175514.1	3066	1728	6	(-)	575	64704.5	9.1	Glyco_ hydro_32	Cell wall

Continuation...

CeCWINV4	113776412	Chr6: 4332626..4337317	XM027321560.1	XP027177361.1	4692	1734	7	(-)	577	65124	9.2	Glyco_ hydro_32	Cell wall
CeCWINV5	113776413	Chr6: 4327214..4331068	XM027321562.1	XP027177363.1	3855	1536	7	(-)	511	57347.6	6.61	Glyco_ hydro_32	Cell wall
CeCWINV6	113775008	Chr6: 4358408..4362146	XM027319712.1	XP027175513.1	3739	1755	7	(-)	584	66197.7	8.86	Glyco_ hydro_32	Cell wall
CeCWINV7	113779026	Chr8: 41725351..41728221	XM027324434.1	XP027180235.1	2871	1728	7	(+)	575	65592.5	8.8	Glyco_ hydro_32	Cell wall
CeCWINV8	113780805	Chr8: 41714036..41716805	XM027326583.1	XP027182384.1	2770	1566	7	(+)	521	59224.7	5.86	Glyco_ hydro_32	Cell wall
CeCWINV9	113781073	Chr8: 41708930..41712269	XM027326914.1	XP027182715.1	3340	1740	7	(+)	579	65715.6	4.85	Glyco_ hydro_32	Cell wall
CeN/AINV1	113749519	Chr10: 13913549..13920726	XM027293277.1	XP027149078.1	7178	1677	5	(-)	558	63836.5	5.87	Glyco_ hydro_100	Chroroplastic
CeN/AINV2	113760361	Chr2: 79307366..79310624	XM027302909.1	XP027158710.1	3259	1833	4	(+)	610	68301.3	7.9	Glyco_ hydro_100	Chroroplastic
CeN/AINV3	113763438	Chr2: 18116003..18121208	XM027307257.1	XP027163058.1	5206	1713	5	(-)	570	65148.8	5.87	Glyco_ hydro_100	Chroroplastic
CeN/AINV4	113760393	Chr2: 9410038..9416107	XM027302952.1	XP027158753.1	6070	1962	7	(-)	653	73244.9	7.49	Glyco_ hydro_100	Chroroplastic
CeN/AINV5	113752301	Chr11: 32627781..32631357	XM027296419.1	XP027152220.1	3577	1257	5	(-)	419	46730.4	5.18	Glyco_ hydro_100	Chroroplastic
CeN/AINV6	113750710	Chr1: 42376411..42382068	XM027294906.1	XP027150707.1	5658	1923	7	(-)	640	72065.7	5.97	Glyco_ hydro_100	Chroroplastic
CeN/AINV7	113765079	Chr3: 407117..412265	XM027309129.1	XP027164930.1	5149	2085	6	(+)	694	78605.7	5.85	Glyco_ hydro_100	Chroroplastic
CeN/AINV8	113761853	Chr2: 75284923..75289550	XM027305009.1	XP027160810.1	4628	2016	6	(-)	671	76091.3	6.42	Glyco_ hydro_100	Chroroplastic

[a] Name in this work; [b] Locus identity used in this work; [c] Start and end chromosome position of genes; [d] Molecular weight (Da); [e] Theoretical isoelectric point; - Not reported in the consulted databases.

Table S2. Access number and additional data from the RNA-seq libraries selected from *Coffea arabica* and *C. canephora*.

Organism	Tissue	Cultivar	Development stage	Submitted by	Instrument	Strategy	Source	Selection	Layout	Run	Spots	Of Bases	Size	Published
C. canephora	coffee bean	robusta	Small green fruit	Yunnan University	BGISEQ-500	RNA-Seq	TRANSCRIPTOMIC	cDNA	PAIRED	SRR10019318	70,075,294	7G	4Gb	16/08/2020
C. canephora	coffee bean	robusta	Large green fruit	Yunnan University	BGISEQ-500	RNA-Seq	TRANSCRIPTOMIC	cDNA	PAIRED	SRR10019314	65,468,078	6.5G	3.7Gb	16/08/2020
C. canephora	coffee bean	robusta	Yellow fruit	Yunnan University	BGISEQ-500	RNA-Seq	TRANSCRIPTOMIC	cDNA	PAIRED	SRR10019311	67,630,972	6.8G	3.8Gb	16/08/2020
C. canephora	coffee bean	robusta	Red fruit	Yunnan University	BGISEQ-500	RNA-Seq	TRANSCRIPTOMIC	cDNA	PAIRED	SRR10019308	69,774,652	7G	4Gb	16/08/2020
C. arabica	fruit pulp	geisha	Red pulp	UC Davis	Illumina HiSeq 3000	RNA-Seq	TRANSCRIPTOMIC	cDNA	PAIRED	SRR26586534	33,042,542	9.9G	3.6Gb	31/10/2023
C. arabica	fruit pulp	geisha	Pink pulp	UC Davis	Illumina HiSeq 3000	RNA-Seq	TRANSCRIPTOMIC	cDNA	PAIRED	SRR26586545	31,709,933	9.5G	3.4Gb	31/10/2023
C. arabica	fruit	geisha	Red bean	UC Davis	Illumina HiSeq 3000	RNA-Seq	TRANSCRIPTOMIC	cDNA	PAIRED	SRR26586567	33,885,723	10.2G	3.7Gb	31/10/2023
C. arabica	fruit	geisha	Developing berry bean 2mm	UC Davis	Illumina HiSeq 3000	RNA-Seq	TRANSCRIPTOMIC	cDNA	PAIRED	SRR26586579	26,286,819	7.9G	2.9Gb	31/10/2023
C. arabica	fruit	geisha	Pink bean	UC Davis	Illumina HiSeq 3000	RNA-Seq	TRANSCRIPTOMIC	cDNA	PAIRED	SRR26586623	32,454,995	9.7G	3.6Gb	31/10/2023
C. arabica	fruits (perisperm)	iapar59	30 DAF	Embrapa Cafe	Illumina HiSeq 2000	RNA-Seq	TRANSCRIPTOMIC	RANDOM	SINGLE	SRR4046866	16,324,779	1.6G	1.1Gb	01/10/2016
C. arabica	fruits (perisperm)	iapar59	60 DAF	Embrapa Cafe	Illumina HiSeq 2000	RNA-Seq	TRANSCRIPTOMIC	RANDOM	SINGLE	SRR4046867	2,530,615	253.1M	174Mb	01/10/2016
C. arabica	fruits (perisperm)	iapar59	90 DAF	Embrapa Cafe	Illumina HiSeq 2000	RNA-Seq	TRANSCRIPTOMIC	RANDOM	SINGLE	SRR4046868	9,207,895	920.8M	639.8Mb	01/10/2016
C. arabica	fruits (perisperm)	iapar59	120 DAFC2	Embrapa Cafe	Illumina HiSeq 2000	RNA-Seq	TRANSCRIPTOMIC	RANDOM	SINGLE	SRR4046861	17,444,494	1.7G	1,021Mb	01/10/2016
C. arabica	fruits (perisperm)	iapar59	150 DAF	Embrapa Cafe	Illumina HiSeq 2000	RNA-Seq	TRANSCRIPTOMIC	RANDOM	SINGLE	SRR4046860	2,333,189	233.3M	161.3Mb	01/10/2016
C. arabica	green drupe rep1	not available	Green drupe rep1	IGA Technology Services Srl	Illumina HiSeq 2500	RNA-Seq	TRANSCRIPTOMIC	cDNA	SINGLE	SRR9964222	27,206,783	6.8G	2.2Gb	13/03/2020
C. arabica	red drupe rep1	not available	Red drupe rep1	IGA Technology Services Srl	Illumina HiSeq 2500	RNA-Seq	TRANSCRIPTOMIC	cDNA	SINGLE	SRR9964221	22,612,568	5.7G	1.8Gb	13/03/2020

Table S3. Percentage of sequence identity and the GMQE, QMEAN, and Ramachandran scores obtained for the model based on the *Arabidopsis thaliana* template sequence.

Enzyme	Template (PDB ID)	Organism	Model parameters			
			GMQE	QMEAN Discrete	Sequence identity	Favorable Ramachandran
CaCWINV1	2ac1	*Arabidopsis thaliana*	0.84	0.84±0.05	65.73 %	92.82 %
CaCWINV13	2ac1	*Arabidopsis thaliana*	0.78	0.80±0.05	59.69 %	91.75 %

Table S4. Interaction energies obtained from molecular docking.

Enzyme	Interaction energy (kcal.mol^{-1})			
	Sucrose	1-kestose	Raffinose	Stachyose
CaCWINV1	-5.7	-5.5	-6.4	-5.2
CaCWINV13	-5.2	-5.4	-5.6	-5.9

TERCEIRA PARTE

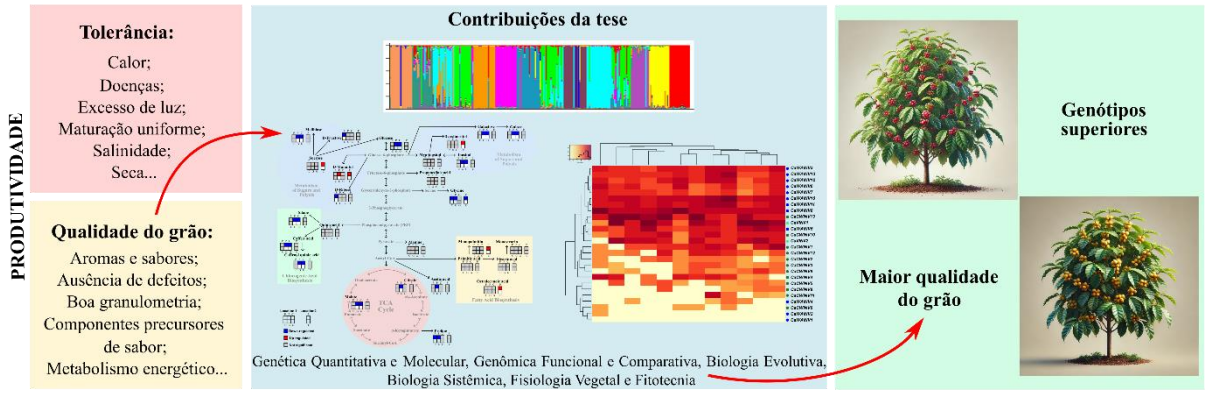
CONSIDERAÇÕES FINAIS

Pesquisas que envolvem Genética e Melhoramento de Plantas têm como principais objetivos a identificação de genótipos superiores, maior eficiência na seleção e otimização de tempo e recursos. No caso do café arábica, a identificação de genótipos superiores se traduz no desenvolvimento de linhagens geneticamente mais homogêneas, produtivas e com alta qualidade do produto final, especialmente no que diz respeito à qualidade do grão. Estudos voltados para obtenção de genótipos de *Coffea arabica* promissores para qualidade do grão, tem avançado em diversas áreas, incluindo o melhoramento genético, a bioquímica dos frutos e o estudo de genes relacionados ao desenvolvimento e à qualidade do café.

Esta tese caracteriza-se por apresentar uma abordagem holística sobre *C. arabica*, envolvendo análises genéticas, transcriptômicas, genômicas, bioquímicas e moleculares. Esses estudos contribuem para compreensão dos mecanismos responsáveis pela formação de fenótipos desejáveis, facilitando a identificação de genótipos superiores e o desenvolvimento de linhagens que apresentam maior qualidade dos grãos. Ademais, os resultados aqui apresentados abrem muitas perspectivas para pesquisas futuras em *Coffea*, incorporando abordagens interdisciplinares como Genética Quantitativa e Molecular, Genômica Funcional e Comparativa, Biologia Evolutiva, Biologia Sistêmica, Fisiologia Vegetal e Fitotecnia.

A integração de diferentes áreas do conhecimento contribui para aumentar a eficiência na seleção de genótipos promissores, algo especialmente relevante para o café arabica, cultura de ciclo de vida longo, com estimativas de até 30 anos para o desenvolvimento de uma nova cultivar. Assim, esta tese representa avanços significativos no desenvolvimento de cultivares mais eficientes, resilientes e alinhadas às demandas do mercado global, promovendo a valorização da cadeia produtiva do café. As contribuições desta tese estão resumidas no esquema simplificado a seguir.

Esquema simplificado das contribuições desta tese para o melhoramento da qualidade dos grãos do cafeeiro.



Fonte: Da autora (2025).