UNIVERSIDADE FEDERAL DE LAVRAS

THALES HENRIQUE CHERUBINO RIBEIRO

GENE REGULATORY NETWORKS: CO-EXPRESSION MODULES OF PROTEIN CODING GENES AND SMALL RNAS GOVERNING ESSENTIAL BIOLOGICAL PROCESSES IN *Coffea arabica* L.

LAVRAS-MG
2023

THALES HENRIQUE CHERUBINO RIBEIRO

GENE REGULATORY NETWORKS: CO-EXPRESSION MODULES OF PROTEIN CODING GENES AND SMALL RNAS GOVERNING ESSENTIAL BIOLOGICAL PROCESSES IN *Coffea arabica* L.

Thesis submitted for the Degree of Doctor of Philosophy in Agronomy concentration area of Plant Physiology Postgraduate Program of Federal University of Lavras

Prof. Antonio Chalfun Junior, Ph.D
Supervisor

Ph.D. Blake C. Meyers
Co-Supervisor

Ph.D. Raphael Ricon de Oliveira
Co-Supervisor

LAVRAS-MG
2023

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

CHERUBINO RIBEIRO, THALES HENRIQUE.
Gene regulatory networks: co-expression modules of protein coding genes and small RNAs governing essential biological processes in *Coffea arabica* L. / THALES HENRIQUE CHERUBINO RIBEIRO. - 2023.
122 p. : il.

Orientador(a): Antonio Chalfun-junior.
Coorientador(a): Blake C. Meyers, Raphael Ricon de Oliveira.
Tese (doutorado) - Universidade Federal de Lavras, 2023.
Bibliografia.

1. Coffea arabica. 2. Genome. 3. Transcriptome. I. Chalfun-junior, Antonio. II. Meyers, Blake C.. III. de Oliveira, Raphael Ricon. IV. Título.

THALES HENRIQUE CHERUBINO RIBEIRO

GENE REGULATORY NETWORKS: CO-EXPRESSION MODULES OF PROTEIN CODING GENES AND SMALL RNAS GOVERNING ESSENTIAL BIOLOGICAL PROCESSES IN *Coffea arabica* L.

REDES DE REGULAÇÃO GÊNICA: MÓDULOS DE CO-EXPRESSÃO DE GENES CODIFICADORES DE PROTEÍNAS E PEQUENOS RNAS GOVERNAM PROCESSOS BIOLÓGICOS ESSENCIAIS EM *Coffea arabica* L.

Thesis submitted for the Degree of Doctor of Philosophy in Agronomy concentration area of Plant Physiology Postgraduate Program of Federal University of Lavras

Approved on January 30th 2023
Blake C. Meyers, Ph.D. Donald Danforth Plant Science Center, USA
Marie Mirouze, Ph.D. Institut de recherche pour le développement, France
Prof. Michael Eric Schranz, Ph.D. Wageningen University, Netherlands
Prof. Teodorico de Castro Ramalho, Ph.D. UFLA, Brazil

Prof. Antonio Chalfun Junior, Ph.D.
Supervisor

Ph.D.Blake C. Meyers
Co-Supervisor

Ph.D. Raphael Ricon de Oliveira
Co-Supervisor

LAVRAS-MG
2023

To Sarah Cherubino-Jaramillo

ACKNOWLEDGMENTS

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001 and in part by the Instituto Brasileiro de Ciência e Tecnologia do Café (INCT/Café), under FAPEMIG grant (CAG APQ 03605/17) and São Paulo Research Foundation (FAPESP) grant #2021/06968-3

In addition, I would like to tanks to individuals and institutions that helped me along this journey:

Andrea Jaramillo Mesa, Adelir, Adriano, Alan, Alex, Alexandre, Alice, Alódio, Amanda, Amaral, Ana, Anakin, André, Andressa, Artur, Atilio, Atul, Aurora, Balbinay, Bárbara, Beatriz, Bethania, Beto, Brian, Bruno, BrunoM, Caique, Caio, Capa, Cardon, Carina, Carlos, Caroline, Cássio, Cecilio, Chalfun, Charles, Christopher, Ciarán, Cris, Cleber, Cleusa, Cleverson, Cris, Dana, Danforth, Daniel, Daniela, Dawyson, Débora, Denise, Dimas, Edmond, Eliza, Elizabete, Elza, Eric, Eron, Eunice, Eustaquio, Evandro, Everton, Fabio, FAPEMIG, Fatinha, Fausto, Felipe, Fernanda, Fernando, FIOCRUZ, Fiorita, Flávio, Flávia, Frodo, Gabriel, Gabriela, Gavilanes, Géssica, Gil, Gilson, Giovani, Glória, Gregory, Gregorin, Gumercindo, Hamilton, Hélio, Henrique, Iam, Iasminy, Ícaro, Igor, Ildo, Isabela, Isadora, Ivanei, Ivo, Jacqui, James, Jane, JB, John, Jhonathan, Joel, Jonas, José, Josi, Joyce, Juliana, Katia, Kellen, Kesia, LABMAI, LAE, Laís, Laurence, Lena, Leila, Larissa, Leonardo, Letícia, LFMP, Lillian, Livia, Louise, Lourdes, Lucas, Lucia, Lucio, Luísa, Luiza, Luke, Magrelo, Maick, Manoel, Marcelo, Marcus, Margot, Maria, Mariana, Marie, Marina, Mario, Marllon, Marlon, Marta, Martin, Matheus, Mayra, Merry, Micaele, Milange, Mirian, Moisés, Monique, Monstrão, Muhammed, Murilo, Nara, Natália, Nathália, Neto, Noelly, Noman, NUIG-Galway, Obi, Pablo, Pâmela, Patricia, Paulo, Passamani, Pedro, Peregrin, Priscila, PPGFV, Priscilla, Query, Rafael, Rafaella, Ramon, Raphael, Renata, Renato, Rick, Robert, Rochele, Rodrigo, Roger, Ronildo, Rose, Rosildo, Salete, Samwise, Sandra, Sara, Sayron, Sibele, Sueli, Tamires, Tarick, Tassio, Taynan, Teodorico, Teofilo, Terri, Thaís, (Thales-1)!, Thiago, Toli, Túlio, UFLA, Ulysses,Vânia, Victor, Vinicius, Vitor, Vitória, Vivianny, Vøn, Wallace, Wanchana, Wanda, Wenceslau, Wesley, Wilder, Yasmin, Yoda, Yudai, Zardo, Zélia, Zé Leite, Zezito.

Last but not least, special thanks to Blake C Meyers for receiving me at the Donald Danforth Center where a substantial part of this work was performed; Raphael Ricon de Oliveira for countless hours of discussion and Antônio Chalfun-junior for showing me the marvels of the Molecular Biology.

RESUMO

O cafeeiro é a fonte de uma das commodities mais negociadas no mundo. Da colheita, processamento e comercialização, o grão de café movimenta um mercado internacional do qual depende a vida de milhões de pessoas. O progressivo entendimento de como as plantas funcionam vem promovendo sucessivas evoluções tecnológicas que garantiram uma maior oferta de alimentos. Estas sucessivas quebras de fronteiras no conhecimento agrícola estão ocorrendo ao longo de séculos de civilização. Uma das fronteiras mais imediatas no momento é molecular. Entender como as plantas organizam seus processos fisiológicos ao nível molecular pode ser o caminho para finalmente conciliar a agricultura ao desenvolvimento sustentável. O avanço da biologia molecular está tornando possível esse entendimento pela investigação de como complexas redes de elementos reguladores coordenam o funcionamento dos das plantas e outros organismos. Este trabalho de tese teve por objetivo contribuir na compreensão do metabolismo de *Coffea arabica* por processamento de dados biológicos integrados. A partir de dados de sequenciamento do genoma e transcriptomas de café foi possível identificar fenômenos evolucionários de balanceamento do número de genes, predizer e comprovar a ocorrência de metabólitos e revelar diferentes tipos de RNA envolvidos no controle do florescimento de *Coffea arabica*. Estas descobertas relacionadas à organização e possíveis tendências evolutivas do genoma podem guiar futuros trabalhos com o objetivo de manter a continuidade do cultivo de café.

Palavras-chave: Genoma. Transcriptoma. *Coffea arabica*. Biologia Molecular

ABSTRACT

Coffee plants are the source of one of the most world-wide traded commodities. From harvesting, though processing to commercializing the coffee bean moves an international market that supports the livelihood of millions. The progressing understanding of how plants function at a cellular, molecular, and physiological level has enabled successive technological breakthroughs, this, in turn, has allowed a positive balance between the supply and demand for food. These successive breakthroughs of frontiers in agricultural knowledge are taking place along centuries of civilization. At this point, one of the most relevant frontiers of biology is at the molecular level. The understanding of how plants organize their physiological processes at the molecular level may be the way to finally balance agriculture with sustainable development. Advances in molecular biology are making this understanding possible by investigating how complex networks of regulatory elements coordinate the functioning of plants and other organisms. This thesis has the objective of contributing to the effort of revealing the functional dynamics of *Coffea arabica* metabolism using integrated biological data. From genome and transcriptome sequencing data of coffee samples, I was able to identify evolutionary phenomena such as gene balance, to predict and ascertain for the presence of metabolites, and to reveal multiple types of RNAs involved in control and/or developmental processes of flowering in *Coffea arabica*. Our discoveries regarding the organization and possible evolutionary trends of this genome can guide future works with the objective of maintaining the continuity of coffee.

Keywords: Genome. Transcriptome. *Coffea arabica*. Molecular Biology

SUMMARY

0. GENERAL INTRODUCTION

The beverages extracted from coffee beans are among the most consumed in the world, approximately 2.2 billion cups a day (Denoeud et al., 2014) Supporting this market requires the collective effort of millions of workers in over 60 countries (Waller et al., 2007). That way, the continuing of the coffee culture is of special importance for the livelihood of millions, in particular, small farmers of exporter countries such as Brazil.

Approximately 70% of world coffee production comes from *Coffea arabica* L., which is the result of an interspecific cross of *Coffea eugenioides* Moore and *Coffea canephora* Pierre (Lashermes et al., 1999). As a result, *C. arabica* has two sets of diploid genomes (sub-genomes), one from each parental ancestor. It is believed that this cross was a once-in-history event and thus is the basis of the extremely low genetic variation in wild and cultivated germplasm (Scalabrin et al., 2020). In addition, the low genetic variation of *C. arabica* can also be explained by its preferential autogamous fertilization and its recent origin, probably less than a million years ago (Lashermes et al., 1999). Finally, cultivated coffee suffered from an additional genetic bottleneck effect because seeds were first moved from Ethiopia and South Sudan, which represents its primary center of diversity, to Yemen during the fourteenth century (Meyer, 1968). In the following centuries, few seeds leaked from this secondary dispersal center to constitute the basis of most modern-day varieties (Scalabrin et al., 2020).

A key feature that allowed human civilization to thrive is plant domestication (Childe et al., 1940). For thousands of years, humans have invested substantial collective effort to create specific conditions that benefit the survival and reproduction of some important food sources. Technical knowledge limitations, such as how to change a river course to better irrigate or how to use a plow to maximize energy needed to be solved. Some of these efforts to overcome agricultural problems resulted in technological breakthroughs that allowed a positive balance between the supply and demand of food. Nevertheless, low genetic diversity in plants can reduce their ability to withstand changes, making our lives a bit harder. So, a new frontier must be overcome. We need to understand how plants organize their physiological processes, even better, at the molecular level. This way, potential productivity losses due low genetic diversity can be avoided in an ever-changing world.

The adequate control of molecules within a plant cell is what ultimately allows its survival and reproduction. Most of this control is performed from genomes - long stretches of nucleotides that encode the necessary instructions for life as we know it. To better understand how *C. arabica* managed to become a successful and widespread species, amid its low genetic diversity, molecular biology techniques can be used. By reading the genome, and its immediate transcriptome product, it is possible to have a glimpse into how plants organize their physiological processes at the molecular level.

The fine controlling of metabolites is achieved by the integration of layers of biological data. For example, some transcripts - called messenger RNAs (mRNAs) - carries necessary information of how to produce a protein whereas other transcripts - called small RNAs (sRNAs) - are involved with the regulation of mRNA levels. Technologies for the sequencing of genomes and transcriptomes are revolutionary molecular tools that allow us to access those layers. Using computational power to understand this type of high-throughput data is a way of reverse engineering nature, at least in the sense of learning how it is working. In this thesis, I took advantage of sequencing data to better understand coffee physiology.

The first chapter of this thesis is an article prepared for submission to the journal *Genome Research*. In this work, I show groups of genes that are expressed in synchrony to regulate processes such photosynthesis, cell wall biogenesis, translation, transcription, catabolism and biosynthesis. Some genes of homoeologous chromosomes were found to be missing, in agreement with the gene balancing hypothesis. We also show that decision-making and execution machinery - mostly in the form of RNA processing - seems to be preferentially coordinated by the sub-genome inherited from the *C. eugenioides* ancestor. This centralization of the gene expression program in one sub-genome may have been advantageous to the establishment of *Coffea arabica.*

The second chapter is an article prepared for submission to the journal *Plant Physiology and Biochemistry*. By reconstructing the metabolic pathways that are made of transcripts in *C. arabica* leaves, we were able to predict several chemical compounds that have the potential to be natural products that are yet unknown to be occurring in coffee. From this inference we selected L-DOPA, an important metabolite that is widely used for treating Parkinson disease. Using liquid chromatography approaches we were able to show that this is a molecule naturally present in coffee. It is possible that other important compounds can be extracted from coffee leaves that are mostly regarded as a by-product of coffee culture.

The third chapter is an article prepared for submission to the journal *New Phytologist*. There we focus on the lengthy process of coffee flowering. This is an economically important trait because it is directly involved with the coffee beans production. The extended floral induction window for coffee - that in Brazil can go from February to October - can explain the asynchronous flower development that difficulties harvesting and potentially reduces the coffee bean quality (Cardon et al., 2022). Our results show many regulatory elements - mostly small RNAs - being orchestrated to control the flowering process. Among our findings is the discovery that 24-nt phasiRNAs are abundant in the months-long S3 latent stage. Shortly after the resumption of rains, the *24-PHAS* levels are drastically reduced and the floral development resumes with the development of male organs with their respective meiosis.

We hope that this thesis will fill some knowledge gaps about coffee physiology and, in the long run, help the development of novel approaches to better manage coffee and increase the portfolio of products derived from it. Additionally, long term investigations will be required to turn these finds into practical strategies to benefit millions of families that depend upon this crop.

REFERÊNCIAS

Cardon, C. H. *et al.* **Expression of coffee florigen CaFT1 reveals a sustained floral induction window associated with asynchronous flowering in tropical perennials**. *Plant Sci.* 325, 111479 (2022).

Childe, V. G. *et al.* **Man makes himself**. *Sci. Soc.* 4, (1940).

Denoeud, F. *et al.* **The coffee genome provides insight into the convergent evolution of caffeine biosynthesis.** *science* 345, 1181–1184 (2014).

Lashermes, P. *et al.* **Molecular characterisation and origin of the Coffea arabica L. genome.** ***Mol. Gen. Genet***. *MGG* 261, 259–266 (1999).

Meyer, F. G. **FAO coffee mission to Ethiopia**, 1964-1965. (1968).

Scalabrin, S. *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.* 10, 1–13 (2020).

Thipyapong, P., Hunt, M. D. & Steffens, J. C. **Antisense downregulation of polyphenol oxidase results in enhanced disease susceptibility**. *Planta* 220, 105–117 (2004).

Waller, J. M., Bigger, M. & Hillocks, R. J. **Coffee pests, diseases and their management**. (CABI, 2007).

CHAPTER 1

1. ARTICLE 1 - DIFFERENTIAL ORTHOLOG COMPOSITION OF *Coffea arabica* L. SUB-GENOMES AND ITS CONTRIBUTION TO REGULATORY NETWORKS GOVERNING ESSENTIAL BIOLOGICAL PROCESSES

If a link does not work, please copy and paste into the browser and remove eventual spaces

SUPPLEMENTAL FIGURES can be accessed using the link:

https://drive.google.com/drive/folders/16WSVHqVMv1GpnM52zw0kmDPnMZwLJgCm?usp=share_link

SUPPLEMENTAL TABLES can be accessed using the link:

https://drive.google.com/drive/folders/1wgx4F2-d0hOGn9Pbe7snsp445keq5033?usp=share_link

SCRIPTS can be accessed using the link:

https://github.com/thalescherubino/thesisChapter1

(Draft Version)

Article prepared for submission to the Genome Research Journal

Differential ortholog composition of *Coffea arabica* L. sub-genomes and its contribution to regulatory networks governing essential biological processes

Thales Henrique Cherubino Ribeiro, Raphael Ricon de Oliveira, Antonio Chalfun-Junior

Abstract

The polyploidy of *Coffea arabica* is an important trait affecting the evolution of this species. Genetic variability is scarce due to its recent origin as a interspecific hybrid from a single successful crossing event between *Coffea canephora* and *Coffea eugenioides* relatives. To further investigate the genomic composition of an allotetraploid, we coupled high-throughput methodologies of co-expression analysis and full-length protein coding genes inference. Many of the expected orthologs were found to be missing from one of the two homoeologous chromosomes. The gene expression machinery is mainly represented by single-copy essential orthologs located in the *Coffea eugenioides* sub-genome. This result suggests a preference of the transcriptional and RNA processing machinery to be regulated by one parental sub-genome. To understand the operational modules of the sub-genome's transcription, we performed co-expression analysis that revealed 23 co-regulated modules. This system-wide approach clarified how biological processes (i.e. photosynthesis, cell wall biogenesis, translation, transcription, catabolism and biosynthesis) are running in synchrony and reinforces that there is an ongoing selective pressure in *C. arabica* that constrains the number of copies of some universal orthologues. Thus, this work contributes to our understanding of genome evolution in recent polyploids and supports crop breeding programs.

1. Introduction

Coffea arabica L. (Rubiaceae) is considered the crop with the lowest level of genetic diversity reported so far [1]. It is an autogamous species that arose after a single event of interspecific hybridization, becoming the single polyploid species of the *Coffea* genus [2,3]. The few single nucleotide polymorphisms identified within *C. arabica* are not shared with any of its potential parental species, *Coffea canephora* L. and *Coffea eugenioides* Moore [1]. This finding suggested a severe genetic bottleneck effect once the variation found in *C. arabica* arose after its recent polyploidization event. In addition, no major introgressions could be verified [1]. To understand how genetic diversity impacts the coffee economy several initiatives to investigate this allopolyploidy genome - and relative species genomes - are undergoing to better understand its organization and underlying regulatory networks [4–9].

A central concept in studying allopolyploids is homoeology that is defined as "genes or chromosomes in the same species that originated by speciation and were brought back together in the same genome by allopolyploidization" [10]. Cytogenetic studies in *C. arabica* showed that large genomic duplications or deletions did not occur, confirming the low structural divergence of homoeologous chromosomes between the sub-genomes inherited from the two diploid progenitor species [11]. The only major interchange between the sub genomes seems to be a single instance of homoeologous replacement at the tip of chromosome 7 [1]. One of the tips of the inherited homoeologue from *C. canephora* was replaced with a 1.7 million bases stretch from its *C. eugenioides* counterpart in an apparent homology-directed repair of double strand break [1].

To become a functional genome, with ability to both allow a coordinated metabolism and reproductive success, *C. arabica* chromosomes needed to survive the original hybridization event and overall genome organization until a stable version was achieved [4]**.** Although the chromosomal structure is similar between the sub-genomes, bioinformatic analyses using k-mers of 51 bp suggested high levels of sequence diversity between both homoeologs [1] and differential expression of homoeologous genes was verified [12]. In the long run, these sub-genome specific sequence divergences may be the reason why polyploid genomes tend to turn into diploids [13].

The evolution of genomes after their duplication is, to some degree, governed by gene interaction networks [14]. Connected genes, such as those that form functional modules like ribosomes and transcriptional complexes, have a tendency of not being lost following tetraploidy because a co-expression imbalance may lead to reduced fitness [15]. This Gene Balance Hypothesis [16] also predicts that homoeologs that are not tightly co-regulate with other genes - they are not dosage sensitive - will eventually evolve to singletons due to purifying selection [15]. These additional copies may become pseudogenes or evolve novel functions.

The functional modules underlying gene co-expression networks may be found in contiguous regions of reduction-resistant pairs [17]. This can happen because the selective pressure to maintain the balance of dosage sensitivity loci constrains the loss of these duplicated genes. So, there is a tendency to cluster co-expressed modules on chromosomes [16]. In this work, we took advantage of a publicly available chromosome level assembly of *C. arabica* [18] to (I) predict and annotate protein coding genes (PCG), (II) distinguish the PCG content between sub-genomes and (III) identify co-expression modules to provide insights on leaf metabolism.

The *C. arabica* genome is probably in a phase between a recent whole genome duplication event and the reestablishment of a diploid state. We verified contrasting results of homoeolog gene loss and retention that may be explained within the realm of evolutionary systems biology by linking the evolution of genes to their function within networks [14]. Although both sub-genomes are working in synchrony, we believe that the kernel of this operating system - the code that controls the execution and memory allocation of other codes - is running from the *C. eugenioides* sub-genome.

1.2 Material and methods

1.2.1 RNAseq library acquisition

Illumina® RNAseq (next-generation sequencing of complementary DNA - cDNA) samples from C. arabica were retrieved from the Sequence Read Archive (SRA) using the fastq-dump tool from the SRA toolkit (v. 2.9.6-1) or retrieved directly from the sequencing company when sequenced by our group. Supplemental Table 1 shows details of each Bioproject, from which approximately 174 billion nucleic bases in 81 RNAseq libraries are publicly available though SRA. Multiple coffee tissues such as beans [19,20], leaves [21,24], seeds [22] and roots [23] were used to find exons and predict genes in this study.

1.2.2 Library quality control

RNAseq libraries were inspected for adapters using the minion tool from the kraken package [25]. Afterwards, adapters and low quality reads were processed with Trimmomatic v. 0.39 [26] using the parameters ILLUMINACLIP:3:25:6, SLIDINGWINDOW:4:20 and MINLEN:30. All the quality-controlled reads in each group from Supplemental Table 1 were used for a *de novo* transcriptome assembly and genome mapping.

1.2.3 Generating training gene structures from Short Read RNAseq data

The steps described in this section are in accordance with the *Basic Protocol 1* and *Support protocol 2* from “Predicting genes in single genomes with AUGUSTUS” [27]. All the quality-controlled reads were mapped to the *C. arabica* Caturra genome available at NCBI

(BioProject PRJNA506972) [18]. The fasta sequence of the complete genome was retrieved and used as reference to map the reads of all libraries in Supplemental Table 1 using hisat2 v. 2.1.0 with default parameters [28]. After that, the samtools package v. 1.10 [29] was used to convert sam to bam files, remove all unmapped reads, remove all multi-mappers and filter out paired-end reads with unmatching pairs. Then, picard tools [30] was used to remove read duplication. Finally, all the libraries were combined in a single bam file with the samtools merge command.

Uniquely mapped paired-end *RNAseq* fragments were then subdivided according to which sub-genome they mapped using samtools and were sorted based on their chromosome coordinates, spamming a length of 989.98 million bases. All the subsequent steps of AUGUSTUS training and prediction were run separately for each of the *C. arabica* sub-genomes. The unplaced contigs, with 104.37 million nucleotides, were not evaluated mostly because it consists of highly repetitive sequences with low PCG density. Based on the annotation Cara_1.0 it is estimated that the unplaced contigs encode approximately 13 PCG per million bases while the assembled chromosomes presented an average of 65 PCG per million bases (Supplemental Table 2).

To reduce potential noise due to coincidental alignments a filtering step was applied with “filterBam” from the AUGUSTUS package with parameters “unique”, “paired” and “pairwiseAlignment”. Next, an additional sorting step was applied. Intron information was retrieved with the program “bam2hints” with the parameter “intronsonly”. Once most of the retrieved RNAseq data was unstranded – it is not known from which strand a fragment originated – the script “filterIntronsFindStrand.pl” was applied to identify the correct strand by using genomic splice site information. In this step all reads that did not have appropriate splice site information were discarded.

Template genes for training AUGUSTUS were generated with unsupervised GeneMark-ET [31] training procedures using the script “gmes_petap.pl” with the intron information retrieved in the previous step. This *ab initio* prediction was then filtered with the script “filterGenemark.pl” to select putative genes that have support in the RNA-seq alignment. Because AUGUSTUS requires information of both the coding and non-coding sequences, the flanking region of the *ab initio* predicted genes was evaluated. The script “computeFlankingRegion.pl" was used to calculate the average length of genes and the flanking region was set to 230 for both sub-genomes (approximately half of the average mRNA size) in the script “gff2gbSmallDNA.pl”. The resulting file was processed in accordance with *protocol 2* in HOFF and STANKE (2019) [27] to remove redundant gene structures at the amino acid level. This step was performed to avoid overfitting the gene prediction model. For that reason, no two *ab initio* predictions of PCG with more than 80% of similarity in the primary structure were allowed in the training dataset. The resulting file was used during the AUGUSTUS prediction as hints for the identification of exons, introns and UTRs.

1.2.4 *De novo* transcriptome assembly and protein coding transcripts inference

All the quality-controlled reads retrieved from libraries that shared the same description (Supplemental Table 1) were submitted to Trinity v. 2.8.5 [32] to perform the *de novo* assembly of the transcriptome in each evaluated condition. The used parameters for each run were "seqType fq", "CPU 24", "full_cleanup", "max_memory 140G", "min_contig_length 50", and "no_normalize_reads". Statistics about each assembly were retrieved using the "TrinityStats.pl" script and, to remove excessive isoforms and variants for each transcripts sequences with 95% of similarity were collapsed using cd-hit-est v. 4.8.1 [33] with parameters "c 0.95", "n 10", "T 0" and "M 0".

For each of the remaining putative genes in all the assembled transcriptomes only the largest isoform was selected with the script "get_longest_isoform_seq_per_trinity_gene.pl". This was performed to avoid the effect of multiple isoforms of long genes to influence the average contig length and to speed up the further prediction steps. Additionally, all the predicted transcriptomes from *de novo* assemblies were combined in a single file and sequences with more than 80% of similarity were collapsed with cd-hit-est v. 4.8.1 [33] with parameters "c 0.80", "n 5", "T 0" and "M 0".

Next, the longest Open Reading Frame (ORF) for each transcript was inferred with Transdecoder.LongOrfs with parameter "m 16" which reflects nucleotide sequences of 48 bp. This was performed to allow a lower limit able to include short proteins that are more likely to be orphan genes. Then, the program TransDecoder.Predict v. 5.5.0 was run with the parameter "single_best_only" to select the most probable protein coding transcripts of a given genomic locus. Finally, the software Benchmarking Universal Single Copy Orthologs (BUSCO) v. 4.1.4 [34] with parameters "l eudicots_odb10", "m transcriptome" and "c 24" was used to evaluate the transcriptome-based protein coding inference [34].

1.2.5 Generating training gene structures from proteins.

The steps described in this section are in accordance with the *Alternate protocol 1* from “Predicting genes in single genomes with AUGUSTUS” [27]. The pipeline was run separately for each sub-genome. Firstly, the predicted protein sequences were mapped to the sub-genomes using GenomeThreader v 1.7.1 [35] through the encapsulating script “startAlign.pl " from the Braker software [31]. Then, the resulting alignments were converted to the Gene Transfer Format (GTF) with “gth2gtf.pl” script. Next, the flaking region length was computed with “computeFlankingRegion.pl” and was set to 978 and 1001 to *C. canephora* and *C. eugenioides* sub-genomes respectively. Those values were then used to generate a GenBank flat file with “gff2gbSmallDNA.pl" required in the subsequent training steps.

1.2.6 Generating training gene structures from mRNA.

The steps described in this section are in accordance with the *Alternate protocol 2* and *support protocol 5* from “Predicting genes in single genomes with AUGUSTUS” [27]. The pipeline was run separately for each sub-genome. The putative nucleotide sequences (CDSs) of protein coding loci identified by Transdecoder were used as a proxy for Expression Sequence Tags (EST) to help improve the prediction of exons, introns and UTR regions. To do so the Program to Assemble Spliced Alignments (PASA) v. 2.4.1 [36] was used. Firstly, the transcripts were cleaned using seqclean. Then a configuration file was created with the parameters “MIN_PERCENT_ALIGNED=0.8”, “MIN_AVG_PER_ID=0.9” and “m=50”. The PASA pipeline was called with the parameters “C”, “R”, “CPU 8” and ‘ALIGNERS blat”. The alignment was performed with the blat tool v. 35x9 [37]. Next, the ORFs were calculated with “pasa_asmbls_to_training_set.dbi” and incomplete ORFs were filtered out with custom scripts to create *a bonafide* file. Next, the flanking region length was computed with “computeFlankingRegion.pl” and was set to 1,323 and 1,343 to *C. canephora* and *C. eugenioides* sub-genomes respectively.

1.2.7 Protein coding gene prediction using extrinsic evidence

The steps described in this section are in accordance with Basic *protocol 3, alternate protocol 7* and *alternate protocol 8* from “Predicting genes in single genomes with AUGUSTUS” [27]. The pipeline was run separately for each sub-genome. Firstly, we generated hints using the paired-end *RNAseq* alignments produced in the section “Generating training

Gene Structures From Short Read RNAseq Data**”.** That information were helpful because they encoded the probable locations of introns and the coverage of transcribed exons and UTRs loci. Complementary to the intron hints were generated in *Basic protocol 1*, the exon hints where produced with “bam2wig” and them processed with the script “wig2hints.pl” with parameters “width=10”, “margin=10”, “minthresh=2”, “minscore=4”, “prune=0.1”, “src=W”, “type=ep”, “UCSC=unstranded.track”, “radius=4.5”, “pri=4” and “strand="."”.

Next, we generated hints from the protein data that can aid the prediction of CDS, introns, the correct reading frame and the position of start and stop codons. The result of the alignment with GenomeThreader (performed previously) was used as input in the script “align2hints.pl” with default parameters. Next, *Alternate Protocol 8* [27] was adapted to generate hints from *de novo* PCG predicted from Trinity and Transdecoder. Only complete sequences (with 5’ UTR, CDS and 3’ UTRs) were considered. The BLAT v. 36x9 [37] was used to align those sequences to the sub-genomes with the parameters “noHead” and “minIdentity=92”. Then, “pslCDnaFilter” was applied to filter the potentially most useful alignments with parameters “minId=0.9”, “localNearBest=0.005”, “ignoreNs” and “bestOverlap”. The hints file was then produced with “blat2hints.pl” with parameters "minintronlen=35” and “trunkSS”. Finally, all the hints from the different sources of evidence (such as proteins, cdna and RNA-seq) were combined in a single file to guide AUGUSTUS during PCG prediction.

1.2.8 Training Augustus for specie specific parameters

During all the steps of prediction of *ab initio* gene structures, the program “etraining” was run to create and/or update species-specific parameters of gene models. Each of the “etraining” runs were performed on a set of approximatively 80% of the putative gene structures and then tested in a subset containing the remaining 20%. A final training step was performed by running the script “optimize_augustus.pl”. In addition, to increase the accuracy of AUGUSTUS gene prediction and to identify potential regulatory loci in the optimization procedure of *C. arabica*, PCG were performed with the UTR training steps described in the *support protocol 5* [27].

1.2.9 Running AUGUSTUS with hints and PCG annotation

After successive stages of model training and testing for both sub-genomes the AUGUSTUS was finally run in accordance with the *Basic Protocol 4* from “Predicting genes in single genomes with AUGUSTUS” [27]. The extrinsic support from multiple sources were combined and the parameters set to “UTR=on”, “allow_hinted_splicesites=atac”, and “genemodel=complete”. Only one transcript was allowed per gene, i.e. no multiple isoforms were allowed to be reported.

The annotation of PCG was performed using blast2GO [38] by homology searches with blastp [39] against the RefSeq protein database [40]. In addition, functional analysis was also performed with InterProScan by classifying the predicted proteins into families and identification of domains and important functional sites [41]. Then, Gene Ontology (GO) terms were mapped and processed for each putative gene with the blast2GO annotation tool. That way, an annotation rule was applied to the found ontology terms for each putative gene. This rule was set to default parameters with the aim of finding the most specific annotations within a certain level of reliability. Finally, the BUSCO software v. 4.1.4 [34] was run with the predicted protein sequences from each sub-genome and parameters set to “l eudicots_odb10”, “m prot”, “--long” and “c 24” to evaluate the completeness of the protein coding inference [34].

1.2.10 Co-expression network analysis of PCG on fully expanded leaves

To infer co-expression modules (clusters) of expressed PCG in fully expanded leaves of *C. arabica* we applied procedures from the Weighted Gene Co-expression Network Analysis (WGNA) R-package [42] (Supplemental File “WGCNAsamllSeq.r”). The libraries from the bioproject ID PRJNA851465 were selected because they represent leaves under a heterogeneous set of environmental conditions which are expected to trigger multiple regulatory networks to cope with the field variability [24]. The BioSamples covered two cities (Pirapora and Varginha) during two harvest times (April and October) and with two *C. arabica* cultivars (Acauã and Catuaí). After *in house* quality processing steps, the RNAseq reads were mapped to the *C. arabica* reference genome (BioProject accession PRJNA506972) [18] using the STAR aligner v. 2.7.8 [43]. Then, fragments mapped to the gene exons of our prediction were quantified with the HTseq-count script [44] and processed with required transformations that met the requirements for the WGNA.

First, PCG with mean expression below 25 counts per library were filtered out. Next, we normalized the count data in Counts Per Million (CPM) and then we log2 transformed the matrix of CPM values to fit the assumptions of the WGCNA package. After that, we calculated an adjacency matrix of Pearson correlations between all pairs of expressed loci and raised it to a power **β** (soft threshold) of 6. The **β**=6 parameter was based on the scale free topology criterion [45]. After that, to minimize the effect of noise and spurious associations, we transformed the adjacency matrix into a Topological Overlap Matrix (TOM). Next, a dendrogram, with the co-expression modules as its branches, was inferred based on the average dissimilarity of the TOM using the Dynamic Tree Cut method. Then, we analyzed the individual co-expression modules with the R package igraph [46]**.** Finally, the GO terms for all members of each module were analyzed with the web tool agrigo2 [47] to investigate for enriched terms that provided clues about the biological processes, localization and molecular function of these co-expressed genes. The statistical analysis was performed using the singular enrichment model and a significance cutoff was set using the Benjamini and Hochberg false discovery rate [48] of 0.05.

1.3 Results

The accurate prediction of protein coding *loci* in genomes requires species-specific parameters for its underlying Hidden Markov Model (HMM) [27]. To define those parameters we gathered *ab initio* and extrinsic evidence in order to improve the accuracy and completeness of the annotation. The extrinsic evidence was based on a set of 270 billion nucleotides sequenced in 101 RNAseq libraries of multiple coffee tissues summarized in Supplemental Table 1. Both single and paired end reads were mapped to a fasta file containing the sub-genomes and excluding the unmapped contigs - regions of the genome that are repetitive and difficult to define chromosome coordinates with accuracy. In addition, these RNAseq fragments were also used for the *de novo* transcriptome assembly and transcriptome-based inference to provide full-length transcript data as an additional extrinsic evidence source for model training. This data provided hints for the coordinates of exons, introns and UTRs in the assembled *C. arabica* chromosomes.

After training a supervised machine learning model AUGUSTUS predicted a total of 69,464 full-length putative PCG being 30,162 in the *C. canephora* sub-genome and 39,302 in the *C. eugenioides* sub-genome (Supplemental Figure 1). Approximately 4.6 thousand full-length genes are putative species-specific (orphans) and 16 thousand had their expression verified in leaves (Figure 1). The putative PCG sequences, coordinates and annotation are available at https://dbi.ufla.br/lfmp/ca_annotation/

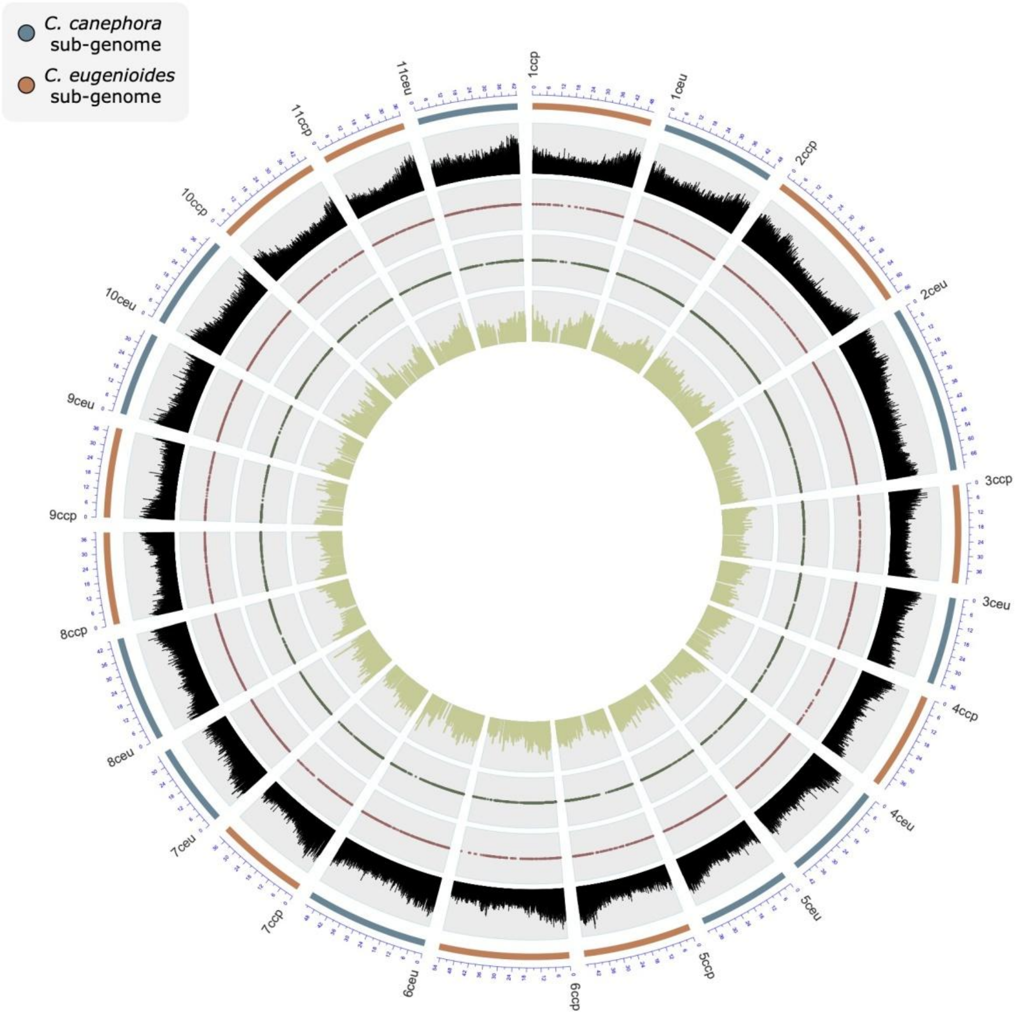

Figure 1 Genome-wide representation of *Coffea arabica* assembled homoeolog chromosomes. Concentrical cycles, left side; the outermost cycle represents chromosomes from each parental ancestor; from *C. eugenioides* (ceu) or *C. canephora* (ccp). First inner cycle; Inferred density of putative protein coding genes within a window of 0.1 million bases. Second inner cycle; red dots point out the chromosomal coordinates of potential Coffea arabica orphan genes (N = ~4.6K) . Third inner cycle; dark green dots point out the chromosomal coordinates of PCG expressed in leaves of the Experimental Group G (Table 1) (N=~16K). Innermost cycle; bar plots depicting the sum of expression of all RNAseq samples from leaves. Expression values are reported in the logarithmic base two of counts per million (CPM).

1.3.1 *De novo* transcriptome assembly and protein coding transcripts inference was an effective tool to retrieve BUSCO signatures

Because the short-read alignment-based evidence does not distinguish non-coding from coding transcripts - the *RNAseq* methodology reads transcripts with poly-A tail - we decided to *de novo* assemble the transcriptome of each of the seven groups of different sets of organs/experiments (Supplemental Table 1). Then, we filtered for the sequences with high probability of being transcribed into full-length proteins with the Transdecoder software.

The combination of the transcriptome assembly from all RNAseq libraries yielded a total of 2,518,967 transcripts. After redundancy filtering of sequences with more than 80% of similarity at the amino acid level we produced a set of 518,787 non-redundant transcripts with the average length of 491 bp, median length of 182 bp and a N50 value of 1,694. Then, the transdecoder software identified 105,598 potential complete proteins in the combined *C. arabica* transcriptome. BUSCO analyses showed that our transcriptome assembly presented 93.5% of the expected orthologs based on the expected composition of eudicots in the database eudicots_odb10 (version from 09/10/2020), which is in accordance with previous reports of gene predictions from a scaffold-level genome assembly that identified 92.4% BUSCO signatures in the “Bourbon Vermelho” variety [1]. Once we found that both the number of the full-length *de novo* predicted proteins and their BUSCO assessment were adequate, we used both the amino acid and mRNA sequences as hint sources for the model training and testing.

1.3.2 Part of the identified BUSCO signatures and the number of PCG are different between the sub-genomes

To better understand the composition of the predicted PCG in the *C. arabica* sub-genomes we performed Gene Ontology (GO), InterPro domain identification and BUSCO analysis. Because our PCG prediction was performed to maximize the quantification of RNAseq fragments mapped to exons we did not allow multiple gene isoforms that potentially shared exons. In addition, our focus was to provide a reliable reference of protein coding genes in a General Feature Format (gff) file to allow *RNAseq* expression analyses of full length PCG. With that strategy we avoided quantifying other genic loci that are not translated.

We only allowed a single isoform of each PCG. In addition, only loci containing identifiable transcription start site (tss), start codon, exon(s), stop codon and transcription termination site (tts) were reported in the gff file. This approach allowed us to increase the number of quantified *RNAseq* reads that were uniquely mapped to exons by a factor of 63% when compared to running the htseq-count script with default parameters using the NCBI *Coffea arabica* Annotation Release 100 [49].

Our procedure of inferring PCG separately for each sub-genome, instead of performing training and predicting steps in a genome-wide approach, allowed us to better understand the differences between them. Approximately 43% of the PCG were found in the *C. canephora* sub-genome with protein N50 of 987. Meanwhile, 57% of the PCG were identified in the *C. eugenioides* sub-genome with protein N50 of 777. We found these differences in the number of genes and their length intriguing, however the raw quantification of predicted PCG and their sizes cannot provide extensive insights into the actual composition and/or quality of genome wide inferences. Because of that, additional tools were applied to measure quantitatively the completeness using evolutionarily informed expectations of gene content [34].

Essential genes are significantly enriched in lineage-specific universal orthologs databases of model organisms [50]. Because of that, we applied the BUSCO tool to quantify the completeness of each sub-genome data set in terms of the expected PCG content. The BUSCO result for the PCG in the *C. canephora* sub-genome reported 73.9 % of Complete (C) signatures (being 68.6 % Single (S) copies and 5.3 % Duplicated (D)), 5% Fragmented (F) and 21.1% Missing (M). In addition, the BUSCO result for the PCG in the *C. eugenioides* sub-genome reported 77.6% C signatures (being 72.6% S and 5% D), 4.3% F and 18.1% M. The high proportion of missing universal orthologs within sub-genomes made us wonder if their PCG compositions are different. We found that missing terms of one sub-genome are present in the other.

The union of complete BUSCO signatures of predicted PCG in both sub-genomes sums up to 2,132 (approximately 92% of the signatures in the eudicots_odb10 reference database - version from 09/10/2020). In addition, the union of fragmented signatures from the sub-genomes summed up to 184. After filtering out overlaps between Fragmented and Complete signatures across the sub-genomes, the proportion of non-missing terms increased to 2,211 (95% of eudicots_odb10 database). Those fragmented signatures are from sequences matches with lengths below two standard deviations from the BUSCO group mean sequence length[34]. In our annotation the fragmented signatures may have arisen because of our choice of only reporting a single isoform per gene. We configured AUGUSTUS to select the most likely isoform when multiple were presented. However, it is not guaranteed that the reported transcript of a given gene is the longest or even the main isoform. In addition, our choice of reporting only full-length PCG may have influenced the total number of genes evaluated during the BUSCO analyses.

Surprisingly, 327 (15.3%) complete BUSCO signatures were exclusively derived from the *C. canephora* sub-genome. Meanwhile, 413 (19.4%) BUSCO signatures were exclusively derived from the *C. eugenioides* sub-genome. When we accessed the PCG with those complete BUSCO matches that are exclusive from the *C. canephora sub-genome* (365 genes; Supplemental Supplemental Table 3) and the ones that are exclusive from the *C. eugenioides* sub-genome (454 genes; Supplemental Supplemental Table 3). We verified that the total number of signatures is lower compared to the total number of PCG with signatures. This difference in the number of signatures and the universal orthologous proteins is explained by the finding that some PCG was assigned to more than one BUSCO signatures and also because of the potential presence of multiple copies of some universal orthologs in a sub-genome.

Those sub-genome-specific BUSCO signatures - that are not shared with the other set of homoeologs chromosomes - can occur in a single copy (S) or multiple copies (duplicated - D) configuration. However, in some lineages, complete BUSCO signatures are more likely to be found in singletons because they are evolving under single copy control [51]. We propose that there is an ongoing selective pressure in *C. arabica* that constrains the number of copies of universal orthologues. Similar stronger selection constraints on the evolution of essential genes were verified in other lineages such as fungi, arthropods and vertebrates [51].

1.3.3 A high proportion of universal orthologs are sub-genome-specific singletons

The recent-polyploid nature of *C. arabica* gives an interesting perspective of a recently formed genome because there is a tendency of duplicated genes being silenced or removed shortly after tetraploidy formation [52]. We found that about 60% of the identified universal ortholog signatures were found to be shared by both sub genomes - they were homoeologous pairs. Homoeologous are genes in the same species that originated by speciation and were later brought back together in the same genome by allopolyploidization [10]. We initially hypothesized that each sub-genome would have roughly the same number of exclusive BUSCO signatures because this gene loss would be a random process. But we verified an imbalance towards the *C. eugenioides* sub-genome keeping more universal orthologs than its *C. canephora* counterpart.

Fast and system-wide gene loss is a strategy to escape selective pressure in recently formed allopolyploids [17,53]. After polyploidization, they must survive extensive genomic reprogramming to get rid of transcriptional imbalances that cause reduced fitness [15]. It is possible that the verified lack of about 40% of universal orthologs - that are potentially missing homoeologs - was caused by the purifying process that benefits the retention of a single copy of some types of genes [16].

The tendency for the loss of duplicated genes also drives genomes towards a simpler and more stable configuration, both structurally and in code base content [13,16]. We presume that if the sub-genomes are segregated from each other, they would not be able to survive due the lack of fundamental molecular codebase that are better kept in single copy. It is possible that the evolutionary tendency for the following *Coffea arabica* generations is to revert to a diploid state.

1.3.4 Few GO terms of universal orthologs are exclusively enriched in the *C. canephora* sub-genome

To further elucidate the functions of those sub-genome specific universal orthologues, we analyzed their enriched GO terms. There are fourteen enriched GO terms for the *C. canephora* sub-genome exclusive BUSCO signatures (Supplemental Table 4). They are involved with biological processes (BP) such as cellular response to stress (GO:0033554, pval = $5.7E^{-7}$), DNA repair (GO:0006281, pval = $2.4E^{-5}$) and DNA recombination (GO:0006310, pval = $1.60E^{-6}$). Their molecular function is primarily helicase activity (GO:0004386, pval = $2.7E^{-4}$) and the cellular component term chromosome (GO:0005694, pval = $1.6E^{-5}$). We compared these fourteen enriched GO terms and we found out that ten are shared between both sub-genomes specific BUSCO signatures.

Among the enriched GO terms of sub-genome specific BUSCO signatures, there are few exclusively found in the *C. canephora* sub-genome. They are DNA repair (GO:0006281, pval = $2.4E^{-5}$), double-strand break repair (GO:0006302, pval = $2.5E^{-4}$), serine-type endopeptidase activity (GO:0004252, pval = $4.4E^{-4}$) and helicase activity (GO:0004386, pval = $2.7E^{-4}$). It is important to note that despite being exclusively enriched for BUSCO signatures within a sub-genome these BP terms are not necessarily lacking in one of the homoeolog chromosomes because thousands of genes are not universal orthologs. So, although the DNA repair and response to stress capabilities are shared between sub-genomes, we suppose that some important component of the helicase activity is coordinated by the *C. canephora* sub-genome. This sub-genome enrichment of a specific part of the universal orthologs may be a way of keeping some relevant influence over the transcriptional and genome duplication machinery. It is also possible to be a form of protection from double-strand breaks promoted by the *C. eugenioides* counterpart.

1.3.5 The *C. eugenioides* sub-genome seems to be the main coordinator of gene expression in *C. arabica*

Following the trend of the total number of PCG, that is higher in the *C. eugenioides* sub-genome, the number of enriched GO terms from its exclusive BUSCO signatures is also higher compared to the *C. canephora* homoeolog. There are 167 GO terms enriched for the BUSCO signatures found exclusively in the *C. eugenioides* sub-genome (Supplemental Table 4). Ten of these 167 enriched terms are shared with *C. canephora* suggesting that *C. eugenioides* is the sub-genome with more ability of transcribing universal orthologs.

The most enriched PB term is RNA processing (GO:0006396, pval = 6.3E^{-13}) followed by nuclear transport (GO:0051169, pval = 1.9E^{-11}). This transport is mainly characterized as RNA export from nucleus (GO:0006405, pval = 6.0E^{-7}). In addition, we also found other highly significant terms such as ncRNA processing (GO:0034470, pval = 5.9E^{-6}) with its child terms tRNA processing (GO:0008033, pval = 4.9E^{-5}) and rRNA processing (GO:0006364, pval = 2.4E^{-3}). Also among the BP we could distinguish the terms mRNA processing (GO:0006397, pval = 1.0E^{-5}), organelle organization (GO:0006996, pval = 1.0E^{-9}) with its child terms chromosome organization (GO:0051276, pval = 2.1E^{-5}) and chloroplast organization (GO:0009658,pval = 1.8E^{-3}). Finally, we found the molecular function terms of RNA methyltransferase activity (GO:0008173, pval = 2.7E^{-9}) and protein binding (GO:0005515, pval = 5.9E^{-9}) being highly enriched as well as the cellular component terms of membrane-bounded organelle (GO:0043227, pval = 1.5E^{-25}) with its child terms nucleus (GO:0005634, pval = 4.8E^{-10}) and chloroplast (GO:0009507, pval = 5.7E^{-23}). Taken together, these results show that an important proportion of *C. arabica*'s gene expression machinery (GO:0010467, pval = 1.70E^{-5}) is controlled by the *C. eugenioides* sub-genome.

1.3.6 PCG in the *C. canephora* sub-genome tend to be longer and enzymes are preferentially found in the *C. eugenioides* sub-genome

Besides the finding that there are fewer PCG encoded in the *C. canephora* sub-genome we also found that they tend to be longer than PCG encoded in the *C. eugenioides* counterpart (Supplemental Figure 2). In addition, we found that the contrasting number of PCG between the sub-genomes is even more pronounced in enzymes characterized by the InterPro scan analyses. The number of identified enzymes is always higher for the *C. eugenioides* sub-genome for every evaluated enzymatic class (Supplemental Figure 3). For example, there are 116 oxidoreductases encoded in the *C. canephora* sub-genome while 1,449 in the *C eugenioides* sub-genome. The abundance of this class of enzymes in the *C. eugenioides* sub-genome is 12.5 times higher than *C. canephora* sub-genome. A similar pattern is verified for transferases (7 times), Hidrolases (8.9 times), Lyases (8 times), isomerases (7.57 times), ligases (3 times) and Translocases (8.84 times).

It is possible that by controlling the RNA processing and transporting machinery and, in practical terms, the gene expression (GO:0010467, pval = 1.70E^{-5}), the *C. eugenioides* sub-genome are effectively causing the erosion of the gene content in *C. canephora* homoeolog chromosomes. It is also possible that a more direct - and faster - interference may be archived by TE insertions leading to pseudogenization [54]. This way, the *C. eugenioides* sub-genome - with its advantage control of transcription and chromosome organization - may recruit Transposable Elements (TE) to be inserted into *C. canephora* loci. Over the past few millennia this process would render affected genes unrecognizable and explaining why there are fewer enzymes in the *C. canephora* sub-genome. In addition, this can also explain why many of the remaining genes in the *C. canephora* sub-genome are longer than their *C. eugenioides* counterpart.

1.3.7 Most of the GO terms of BUSCO signatures in *C. canephora* sub-genome are shared with its *C. eugenioides* counterpart, but the reciproque is not true

The difference in the number of PCG, their length, BUSCO signatures and enzymatic composition between the *C. arabica* sub-genomes made us wonder if the overall distribution of GO terms is also different between the PCG of each sub-genome, and not only a specific feature of the universal orthologs. To address this, we carefully examined the sub-genome specific BLAST2GO functional annotation results.

We found that the number of GO terms for BP is 86% higher in *C. eugenioides* sub-genome than for *C. canephora* (Figure 2A). From the 2,874 BP terms identified in *C. eugenioides* sub-genome PCG 1,414 (49%) are exclusive. Meanwhile a small proportion, only 82 (5%) of the 1,542 BP terms, are exclusively found in the *C. canephora* sub-genome (Figure 2A). Interestingly, among the exclusive BP terms from the *C. eugenioides* sub-genome we could distinguish terms such as “chromatin assembly”, “chromatin maintenance”, “chromatin organization involved in negative regulation of transcription” and “production of siRNA involved in chromatin silencing by small RNA”. Taken together, these *C. eugenioides* sub-genome specific BP terms reinforce the thesis that this sub-genome is controlling the gene expression while also controlling the chromatin organization.

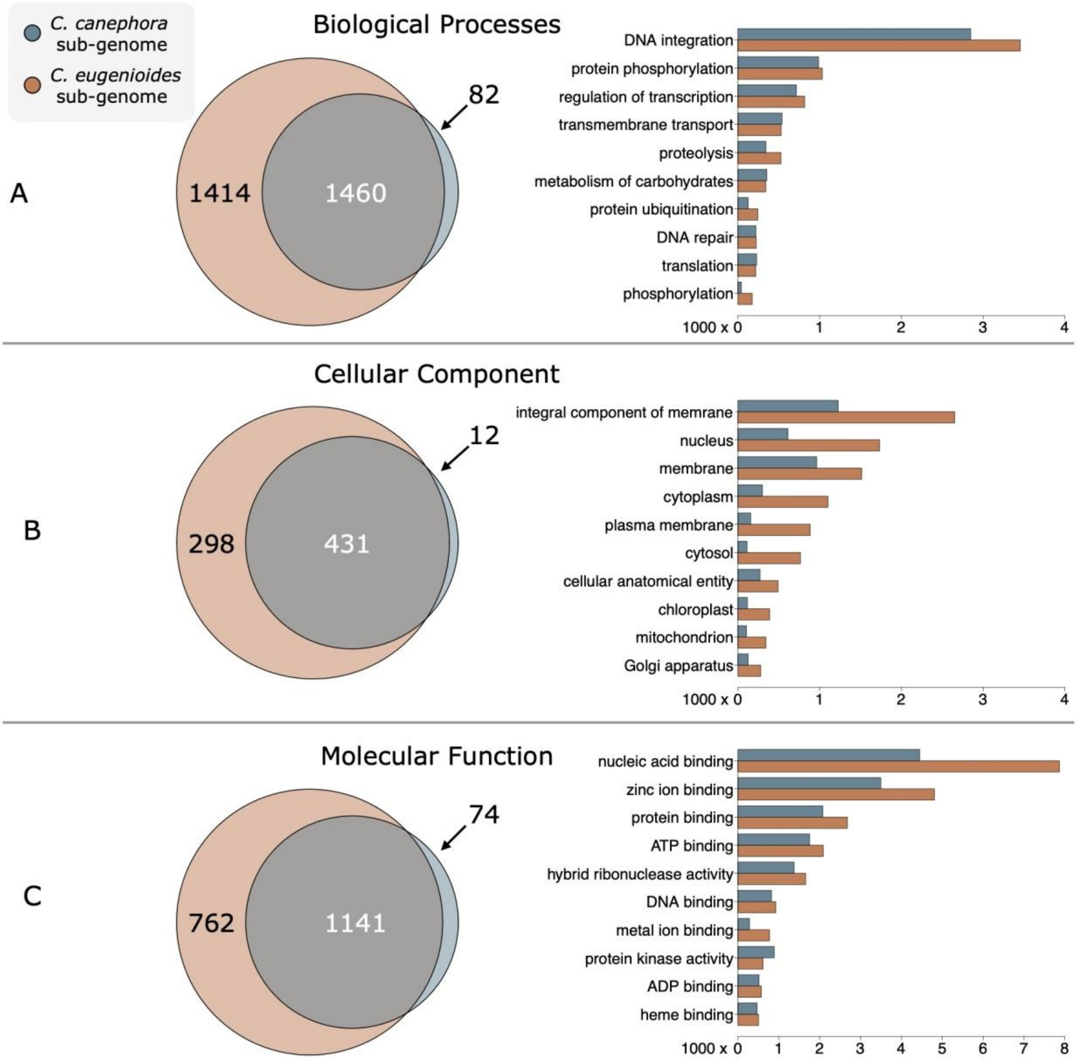

Figure 2 Venn diagrams of identified GO terms for sub-genome specific genes with BUSCO signatures for the three main categories of Biological Processes (A), Cellular Component (B) and Molecular Function (C). Most of the GO terms were identified in the *C. eugenioides* sub-genome while a high proportion of them are not shared with its homoeologous counterpart. On the right, barplots showing a comparison of the content of PCG in each of the most abundant terms by sub-genomes. The most abundant Biological Process is DNA integration while the Molecular Function is particularly enriched for nucleic acid binding. Following the trend of total number of PCG, the *C. eugenioides* is the main contributor for all the Cellular Components, in particular the nucleus and membranes.

Regarding Cellular Component (CC) terms we found a similar pattern in which the number of individual terms is 65% higher in the *C. eugenioides* sub-genome (Figure 2B). From the 729 individual CC terms from the *C. eugenioides* sub-genome, 298 (41%) are exclusive while only twelve of 443 (3%) terms are exclusively derived from the *C. canephora* sub-genome (Figure 2B). Finally, regarding the Molecular Function (MF) terms the number of individual terms is 57% higher in the *C. eugenioides* sub-genome (Figure 2C). From the 1,903 individual MF terms from the *C. eugenioides* sub-genome, 762 (40%) are exclusive while seventy four of 1,215 (6.1%) are exclusively occurring in the *C. canephora* sub-genome (Figure 2B). We suggest that the *C. arabica* genome is still evolving towards balanced gene composition, but the *C. canephora* sub-genome suffers most of this gene loss.

1.3.8 DNA integration and nucleic acid binding GO terms are dominated by Transposable Elements

In addition to the composition of GO terms being different between the sub-genomes, the number of PCG in each of the GO categories is also different. Not surprisingly, the *C. eugenioides* sub-genome is often encoding more genes within each GO category, in particular the top 10 larger categories in number of genes (bar plots at the right of 2). This outlying proportion of genes with the terms DNA integration, nucleus and nucleic acid binding led us to investigate the PCG composition of the top terms in each sub-genome (Figure 3).

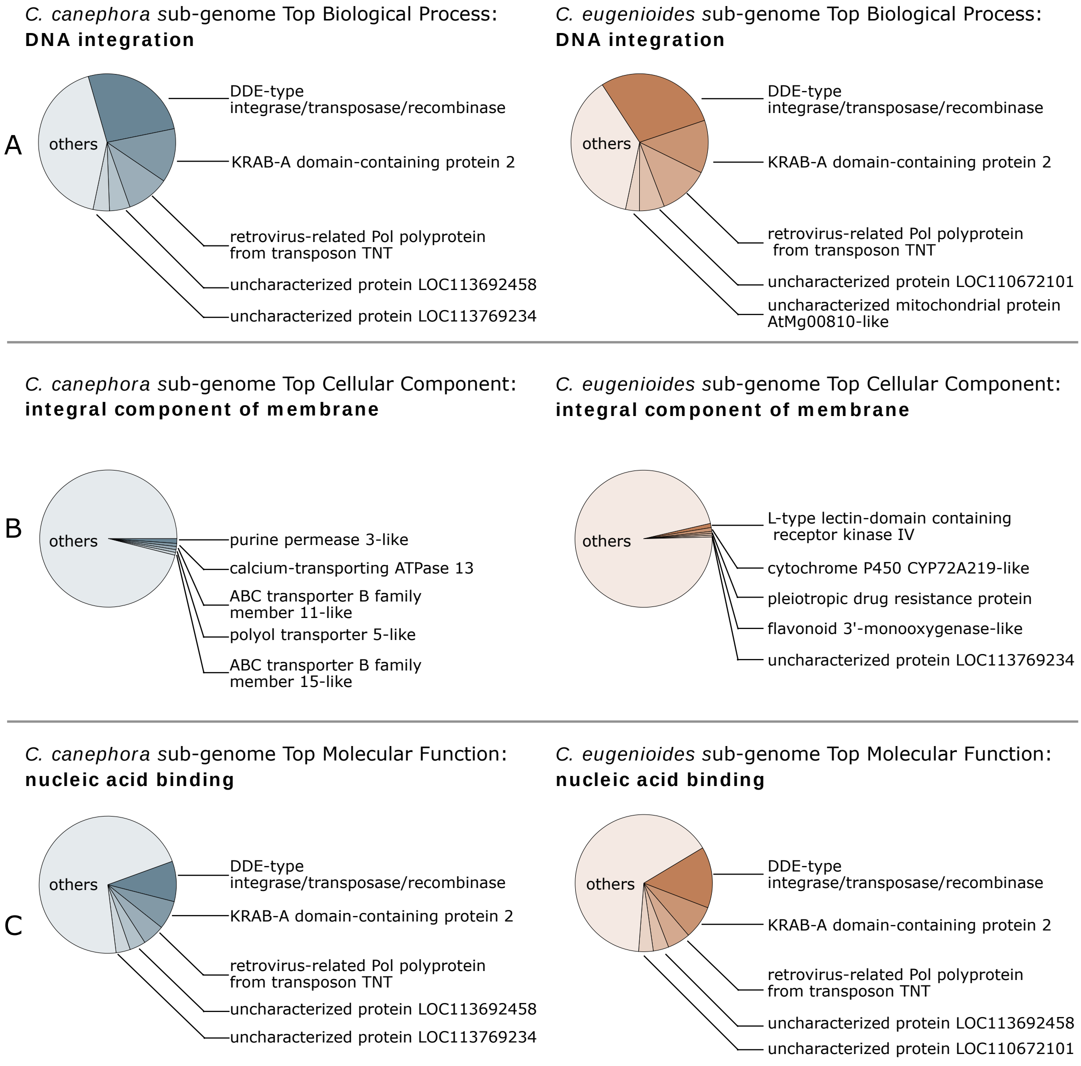

Figure 3 Proportion of the most abundant terms in each GO category of Biological Processes (A), Cellular Component (B) and Molecular Function (C). Known and apparent novel Transposable Elements are specially enriched in the top BP and top MF categories while the top CC category has a diverse composition. The BP term with higher number of members in both sub-genomes is “DNA integration” with 3,454 genes in the *C. eugenioides* sub-genome and 2,847 in the *C. canephora* sub-genome. The top CC “integral component of membrane” has 2,653 genes in the *C. eugenioides* sub-genome and 1,227 genes in the *C. canephora* sub-genome. The top MF “nucleic acid binding” has 7,867 genes in the *C. eugenioides* sub-genome and 4,447 in the *C. canephora* sub-genome.

This combination of abundant terms related to DNA binding made us wonder if those transcripts would be TE. In the BP category, the term “DNA integration” is mainly composed by genes identified as *DDE-TYPE INTEGRASE/TRANSPOSASE/RECOMBINASE* that accounts for approximately 28% of the PCG in this category (Figure 3A) followed by *KRAB-A DOMAIN-CONTAINING PROTEIN 2* (~13%) and *RETROVIRUS-RELATED POL POLYPROTEIN FROM TRANSPOSON TNT* (~11%). The other 52% of PCG in the “DNA integration” category are mostly identified as uncharacterized proteins. Many of these genes are also reported under the most enriched MF of “nucleic acid binding” where the *DDE-type* genes account for about 13%, *KRAB-A* 7% and the transposon *TNT* 5% (Figure 3C). Once Kruppel associated box (KRAB-ZFPs) are domains only present in tetrapod vertebrates [55] and homology-based searches to the nr database in NCBI returned TE related matches, we believe that these reported KARB-A domain containing proteins are, in fact, miss annotated TE.

Interestingly, the BP term “DNA integration” is more than 50% composed by PCG that corresponds for only five genic families and the MF term “nucleic acid binding” is following a similar pattern, with the five more abundant families accounting for about 33% of its members. On the other hand, the CC most enriched term “integral component of membrane” has a much more diverse composition (Figure 3C). The top five gene families account for less than 5% of all genes belonging to the “integral component of membrane” category while the other 789 families or individual genes are summing up the remaining 95% of membrane components. Nevertheless, the high proportion of TE in the genome composition can be verified in the second larger CC term “nucleus”. So, there is a wide-spread dissemination of TEs that seems to be a strong force shaping the evolution of *C. arabica.*

1.3.9 Part of the core diurnal leaf operating system was captured while transcribing execution orders

After exploring the *C. arabica* sub-genome organization, we performed a regulatory network analysis to understand its transcriptional activation and regulatory modules related to leaf metabolism. To do so, we used 24 RNAseq libraries previously published from fully expanded leaves collected during the morning and under a heterogeneous set of environmental conditions [24].

Our RNAseq based regulatory network inference was performed using 88,620,764 of uniquely mapped paired-end reads that aligned to approximately 51,000 loci that potentially encode PCG. These 88 million RNAseq reads were set apart from a group of 27 millions of multi-mappers that could not provide locus-specific mapping resolution and potentially bias the analysis by not meeting WGCNA microarray-based assumptions. An additional filtering procedure of low expressed loci was important to exclude 34,000 loci that accounted for only 3.7% of the transcripts. Low expressed genes have the potential of biasing the analysis by interfering with distribution assumptions and potentially incurring false-positive results due to their relatively low contribution to the transcriptome. Finally, a total of 16,610 constitutively expressed and accurately mapped putative PCG loci were evaluated.

This counting data from RNA fragments was processed under the WGCNA guidelines [42] to infer twenty three modules of co-expressed genes. Module names were assigned to random colors. We then searched for enriched GO terms in each module and overall results are reported in Supplemental Table 5. After that, the modules were grouped into eight clusters - named from A to H - based on the correlation of their eigengenes (Figure 4, Supplemental Dataset 1), that are individual transcripts that summarizes the overall trend in expression of any given module [45].

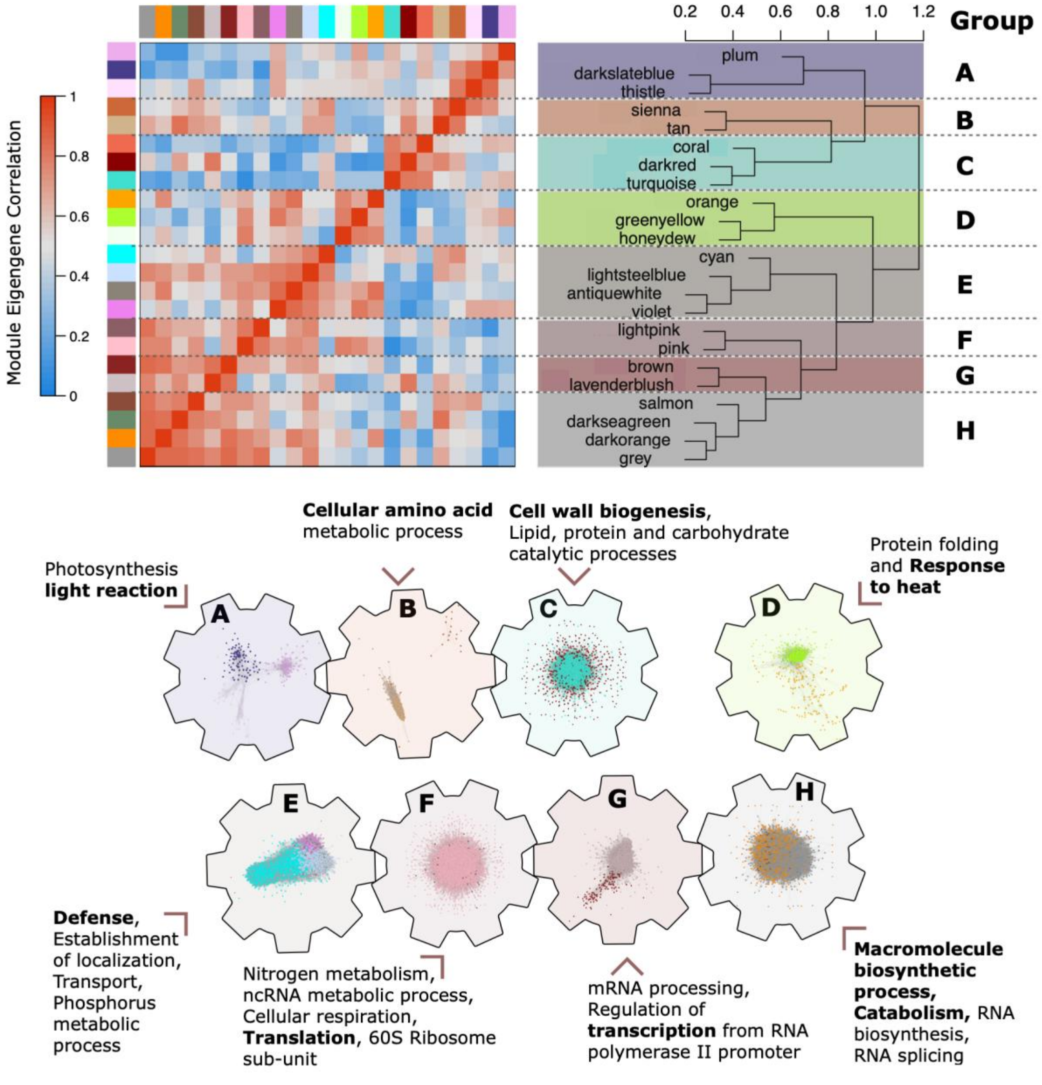

Figure 4 Co-expression inference revealed 23 functional modules involved with important processes of diurnal leaf metabolism. The heatmap on the left shows that the eigengenes of some functional modules are also co-expressed forming functionally larger groups. The hierarchical clustering on the top left provided delimitation for co-expression groups based on

a height cut-off of 0.75. On the bottom, the schematic representation of the interconnectedness of 16,610 expressed genes. Groups of co-expressed genes, represented as engaged or disengaged gears, shows that processes such the light-reaction of photosynthesis and the amino acid and cell wall metabolism are co-regulated. Meanwhile, defense, translation, transcription and other biosynthetic processes are forming a yet larger custer of functional modules. Some of the modules are formed by macro molecules such ribosome subunits, RNA polymerase II, spliceosomes and chromosome components.

Statistical singular enrichment analyses of GO terms associated with each module and groups of modules allowed us to unravel important biological processes coordinated by those inferred regulatory networks. This hybrid group and module-wise approach of analyses allowed us to reveal biological processes that are underlying the diurnal metabolism of *C. arabica* leaves. For example, we found that the 266 genes in group A are primarily involved in the light reactions of photosynthesis (GO:0019684, $\text{pval}_{\text{Darkslateblue}}$ = 7.30E^{-7}) and carbohydrate biosynthetic processes (GO:0016051, $\text{pval}_{\text{Plum}}$ = 3.4E^{-4}). The proteins encoded by those genes are located in the chloroplasts (GO:0009507, $\text{pval}_{\text{Plum}}$ = 9.5E^{-4}). Interestingly, a *TREHALOSE-PHOSPHATE SYNTHASE* is a key component of this group because it is the most central node, in terms of total number of connections (degree), in the Plum module.

The trehalose-6-phosphate is an important signaling metabolite that regulates carbon assimilation and sugar status in plants by balancing the synthesis and breakdown of starch [56]. Meanwhile, the closely related group B, with 448 PCG divided in two modules, Sienna and Tan, is primarily involved with the cellular amino acid metabolic process (GO:0006520, pval_{Tan} = 2.70E^{-5}). In addition, the term chloroplast (GO:0009507, pval_{Tan} = 9.7E^{-4}) and others chloroplast related terms are shared between groups A and B suggesting a co-regulation between the PCG involved in light reactions of photosynthesis and cellular amino acid metabolic process in *C. arabica* leaves.

Next, 3,583 putative PCG from group C are enriched for lipid (GO:0006629, $\text{pval}_{\text{Darkred}}$ = 7.40E^{-11}), protein (GO:0019538, $\text{pval}_{\text{Darkred}}$ = 6.70E^{-8}) and carbohydrate (GO:0044723, $\text{pval}_{\text{Darkred}}$ = 1.40E^{-5}) metabolic processes - more precisely catabolic processes (GO:0009056, $\text{pval}_{\text{Darkred}}$ = 2.00E^{-7}) - in the cytoplasm (GO:0005737, $\text{pval}_{\text{Darkred}}$ = 7.90E^{-14}). The largest member of group C, the Turquoise module with 2,021 PCG components, was mostly enriched for cell wall biogenesis (GO:0042546, $\text{pval}_{\text{Turquoise}}$ = 8.50E^{-21}). In addition, the small, but significantly enriched for photosynthesis (GO:0015979,$\text{pval}_{\text{coral}}$ = 5.10E^{-12}) and thylakoid (GO:0009579, $\text{pval}_{\text{Coral}}$ = 2.20E^{-13}), Coral module was found to be directly involved with the generation of precursor metabolites and energy (GO:0006091, $\text{pval}_{\text{Coral}}$ = 1.80E^{-5}). Cell walls are dynamic structures that are constantly being maintained for enhanced structural and defense capabilities[57]. This group C is showing that part of the photosynthetic machinery, specifically the photosystem I (GO:0009522, $\text{pval}_{\text{Coral}} = 6.80\text{E}^{-11}$) are in transcriptional coordination with cell wall biogenesis and catabolic processes.

Group D is the most distant in terms of their eigengenes correlation with other modules. It is composed of three modules named Orange, Greenyellow and Honeydew summing up to 695 putative PCG. Its largest module, Greenyellow, is specially enriched for protein folding (GO:0006457, $\text{pval}_{\text{Greenyellow}} = 7.30\text{E}^{-55}$), response to heat (GO:0009408, $\text{pval}_{\text{Greenyellow}} = 1.00\text{E}^{-19}$) and response to reactive oxygen species (GO:0000302, $\text{pval}_{\text{Greenyellow}} = 8.20\text{E}^{-15}$). The most central gene in terms of eigengene centrality is a (chloroplastic) *SMALL HEAT SHOCK PROTEIN*. These findings allowed us to propose that this group provides response mechanisms to perceived thermic variations in the environment. This mechanism seems to be coordinated across membranes such as the ones that envelops the endoplasmic reticulum (GO:0005783, $\text{pval}_{\text{Greenyellow}} = 1.40\text{E}^{-7}$). Changes in the membrane fluidity are used as sensors of the temperature variation [58]. These signals can be transmitted to the cytoplasm from the endomembrane system (GO:0012505, $\text{pval}_{\text{Greenyellow}} = 5.10\text{E}^{-6}$) by proteic transmembrane transporters (GO:0022857, $\text{pval}_{\text{Orange}} = 4.2\text{E}^{-4}$). There, they can interact with protein folding mechanisms to prevent heat related damage [59]. It is possible that this group is intrinsically more reliant on the environment - more specifically the heat - and this may be the reason why these modules behave as an outgroup in the eigengene correlation cladogram (Figure 4).

The next group was assigned to the letter E and aggregates a total of 1,660 putative PCG that compose four modules; Cyan, Lightsteelblue, Antiquewhite and Violet. Here, we found a correlation between modules involved with establishment of localization, transport, defense and protein phosphorylation. Protein phosphorylation is an important post-translational modification that regulates protein functions and controls cellular processes during the diurnal cycle in different Arabidopsis organs and seedlings [60].

The enriched GO term establishment of localization (GO:0051234, $\text{pval}_{\text{Cyan}} = 1.80\text{E}^{-11}$, $\text{pval}_{\text{Lightsteelblue}} = 3.90\text{E}^{-9}$, $\text{pval}_{\text{Antiquewhite}} = 7.2\text{E}^{-4}$) is shared between three module members of group E. Permeating the other enriched BP of this group is the term phosphorus metabolic process (GO:0006793, $\text{pval}_{\text{Cyan}} = 2.80\text{E}^{-21}$, $\text{pval}_{\text{Lightsteelblue}} = 6.3\text{E}^{-4}$). Another important phosphorus-related term is cellular response to phosphate starvation (GO:0016036, $\text{pval}_{\text{Violet}} = 4.40\text{E}^{-15}$) that is particularly enriched in the Violet module. These phosphate starvation responses are a collective array of morphological and physiological adaptive changes that are critical for plant survival when phosphorus is limited [61]. The physiological responses for phosphate starvation also play roles in the regulatory mechanism of phosphate usage [61]. In addition, several inorganic phosphate (P_i) transporters have been implicated to compose defense response mechanisms [62].

Supporting the importance of phosphoric compounds to plant defense strategy is the finding that a T-DNA insertions in a loci encoding for a member of *PHOSPHATE TRANSPORTER (PHT)* family increases *Arabidopsis thaliana* susceptibility to virulent *Pseudomonas* strains [63] . Here we found two members of the *PHT* family being highly connected members of the Cyan module and co-expressed with *SALICYLIC ACID-BINDING PROTEIN,* known to be involved with salicylic acid-dependent immune responses [64].

The two larger modules of group E, Cyan and Lightsteelblue with 890 and 579 members respectively, are especially enriched for membrane-related CC terms (GO:0016020, $pval_{Cyan}$ = 1.20E^{-13}, $pval_{Lightsteelblue}$ = 4.20E^{-7}). In addition, the Cyan module is particularly enriched for the BP response to salicylic acid (GO:0023051, $pval_{Cyan}$ = 3.30E^{-5}), exocytosis (GO:0006887, $pval_{Cyan}$ = 7.70E^{-7}) and golgi vesicle transport (GO:0048193, $pval_{Cyan}$ = 1.20E^{-6}). Meanwhile, Lightsteelblue module is enriched for terpenoid biosynthetic process (GO:0006721, $pval_{Lightsteelblue}$ = 4.50E^{-6}) and defense response (GO:0006952, $pval_{Lightsteelblue}$ = 1.10E^{-5}). These findings reinforce the importance of phosphorus for the immune system of plants. The members of this group E, in particular the *Cyan* module, can be further investigated to better elucidate how plants defend themselves from other organisms.

The next group was assigned to the letter F and is composed of two modules, one is the largest of all modules in this analysis and is called *Pink* with 4,847 putative PCG. The other component of this group has only sixty three PCG is called *Lightpink.* Not surprisingly, Pink is also the module with more enriched GO terms, 899 of them. Meanwhile, all enriched 43 GO terms from *Lightpink* are shared with the larger module. The *Pink* module is particularly enriched for the term organo-nitrogen compound metabolic process (GO:1901564, $pval_{Pink}$ = 1.5E^{-138}). In addition, it is also significantly enriched for many other important BP terms such ncRNA metabolic process (GO:0034660, $pval_{Pink}$ = 3.60E^{-29}), rRNA processing (GO:0006364, $pval_{Pink}$ = 3.30E^{-22}), RNA splicing, via transesterification reactions (GO:0000375, $pval_{Pink}$ = 1.20E^{-14}), ligase activity, forming aminoacyl-tRNA and related compounds (GO:0016876, $pval_{Pink}$ = 7.40E-07) and cellular respiration (GO:0045333, $pval_{Pink}$ = 8.50E^{-16}).

The co-regulation of group F is so organized that their proteins are transported to multiple intracellular components (GO:0005622, $pval_{Pink}$ = 1.75E^{-166}). Members of this groups are a relevant part of the protein composition of the cytosol (GO:0005829, $pval_{Pink}$ = 1.9E^{-60}), mitochondrial electron transport, NADH to ubiquinone (GO:0006120, $pval_{Pink}$ = 5.80E^{-5}), the full ubiquinol to cytochrome c (2.70E-09, $pval_{Pink}$ = 2.70E^{-9}) and mitochondrial respiratory chain complex III (GO:0005750, $pval_{Pink}$ = 3.20E^{-8}). They are also enriched for chloroplast thylakoid (GO:0009534, $pval_{Pink}$ = 1.10E^{-19}), nucleus (GO:0005634, $pval_{Pink}$ = 1.10E^{-23}) and apparently compose 44% of the ribosomal protein complex (GO:0005840, $pval_{Pink}$ = 3.30E^{-113}). Their components are so wide-spread within the cell that the few remaining CC terms not enriched in the pink module are the ones related to the extracellular part.

It is possible that the key for understanding this group - that spans most of the CC volume - relies on genes closely related to the Pink module eigengene, a 60S ribosomal protein L24-like. In fact 277 putative PCG of the Pink module were identified as ribosomal proteins, mostly 60S ribosomal proteins. So, it is possible that the whole translational machinery (GO:0006412, $pval_{Pink}$ = 1.20E^{-134}) correlates directly to the rhythm of ribosome assembly (GO:0042255, $pval_{Pink}$ = 9.80E^{-14}) and biogenesis (GO:0042254, $pval_{Pink}$ = 9.90E^{-36}). That way, others important processes that are directly involved with translation, such as the transcriptional steps of the gene expression (GO:0006352, $pval_{Pink}$ = 2.40E^{-7}), RNA splicing (GO:0008380, $pval_{Pink}$ = 8.20E^{-16}) and organo-nitrogen compound metabolic process, such amino acids (GO:0008652, $pval_{Pink}$ = 1.8E^{-4}) and nucleotides (GO:0009117, $pval_{Pink}$ 1.10E^{-21}), are intrinsically interlinked.

The following cluster of modules is called group G and is composed of two modules summing up a total of 945 putative PCG. The main component of this group, the Lavenderblush module, is particularly enriched for GO terms relative to nucleus (GO:0005634, $pval_{Lavenderblush}$ = 1.60E^{-17}) and gene expression (GO:0010467, $pval_{Lavenderblush}$ = 3.10E^{-11}, $pval_{Brown}$ = 2.30E^{-5}). Although the “gene expression” term is also enriched in the previous group F here the descending terms differ. In this group, the influence on gene expression (GO:0010468, $pval_{lavenderblush}$ = 8.10E^{-9}) is mainly performed by the regulation of transcription from RNA polymerase II promoter (GO:0006357, $pval_{lavenderblush}$ = 2.70E^{-11}) and chromosome organization (GO:0051276, $pval_{lavenderblush}$ = 3.70E^{-6}). This group shares with group F important aspects of the gene expression process but they constitute different parts of the machinery. The importance of group G seems to be related to the transcription while group F is mostly related to the translation.

Supporting the finding that group G is an fundamental part of transcription are the enriched terms RNA helicase activity (GO:0003724, $pval_{Lavenderblush} = 2.00E^{-5}$), histone modifications (GO:0016570, $pval_{Lavenderblush} = 3.70E^{-06}$) such as histone methylation (GO:0016571, $pval_{Lavenderblush} = 2.90E^{-05}$) and histone acetyltransferase complex (GO:0000123, $pval_{Lavenderblush} = 5.50E^{-5}$) that are related to two constranting epigenetic processes that modulates the access of RNA polymerases to PCG loci. That way, it is not surprising that the eigengene of the Lavenderblush module is a ATP-dependent helicase, from the *C. canephora* sub-genome, that seems to be encompassing the transcriptional regulation of the whole group.

The final cluster of modules is group H. This group shows a coordination of expression of 4,103 genes divided into four modules. This group is mainly involved with macromolecule biosynthetic processes and is also enriched for hundreds of GO terms, 496 of them. Two modules account for about 97% of this group PCG content, the Darkorange with 1,040 and Grey with 2,930. Traditionally in the WGCNA methodology, the grey module corresponds to the set of genes which have not been clustered in any other module. Nevertheless, this is not the case in our analyses and this cluster is a bonafide module. Both Darkorange and Grey shares 121 enriched GO terms, including regulation of metabolic process (GO:0019222, $pval_{Darkorange} = 3.40E^{-8}$, $pval_{Grey} = 1.10E^{-9}$) and macromolecular complex (GO:0032991, $pval_{Darkorange} = 4.99E^{-6}$, $pval_{Grey} = 1.20E^{-7}$). However, even sharing a significant proportion of enriched GO terms, they are also individually enriched for module-specific terms.

The Darkorange module is enriched for photosynthesis, light reaction (GO:0019684, $pval_{Darkorage} = 1.10E^{-6}$), chlorophyll metabolic process (GO:0015994, $pval_{Darkorange} = 2.00E^{-5}$), generation of precursor metabolites and energy (GO:0006091, $pval_{Darkorange} = 3.10E^{-6}$). Other photosynthesis related terms are also significantly enriched. Meanwhile, Grey is particularly enriched for terms related with mRNA processing (GO:0006397, $pval_{Darkorange} = 2.90E^{-34}$) such as mRNA splicing, via spliceosome (GO:0000398, $pval_{Grey} = 7.40E^{-26}$) or covalent chromatin modification (GO:0016569, $pval_{Grey} = 1.90E^{-12}$).

We noted that even sharing the enriched GO term “macromolecular complex” there are different complexes in each of the two larger modules of group H. The child terms of “macromolecular complex” in the Darkorage module are divided into photosystems I

(GO:0009522, $\text{pval}_{\text{Darkorage}}$ = 2.00E^{-5}), photosystems II (GO:0009523, $\text{pval}_{\text{Darkorage}}$ = 6.0E^{-4}) and ubiquitin ligase complex (GO:0000151, $\text{pval}_{\text{Darkorage}}$ = 4.80E^{-6}). Meanwhile, "macromolecular complex" has diverse child terms in the Grey module, including the top enriched "spliceosomal complex" (GO:0005681, $\text{pval}_{\text{grey}}$ = 6.00E^{-16}) followed by others ribonucleoproteins such pre-ribosome (GO:0030684, $\text{pval}_{\text{grey}}$ = 2.00E^{-9}), small nucleolar ribonucleoprotein complex (GO:0005732, $\text{pval}_{\text{grey}}$ = 5.6E^{-4}) and Mre11 complex (GO:0030870, $\text{pval}_{\text{grey}}$ = 1.80E^{-6}). Another complexes that are formed by PCG in the Grey module include the transferase complex (GO:0061695, pval = 8.9E^{-4}) that binds phosphorus-containing groups in the holoenzyme DNA-directed RNA polymerase II (GO:0016591, $\text{pval}_{\text{Grey}}$ = 1.6E^{-3}).

Members of the Grey module are include components of chromosomes in the form of cohesin (GO:0008278, $\text{pval}_{\text{Grey}}$ = 2.9E^{-4}) and H4 histone acetyltransferase complex (GO:1902562, pval = 8.20E^{-5}). So, we conclude that the Grey module is majoritarily involved in the transcription, partially because of its role in regulation of RNA biosynthetic process (GO:2001141, $\text{pval}_{\text{Grey}}$ = 1.10E^{-5}). Meanwhile, the Darkorange module is mainly involved with photosynthesis, more precisely the chlorophyll biosynthetic process machinery (GO:0015995, $\text{pval}_{\text{Darkorange}}$ = 1.6E^{-3}). Underlying the whole group H, biosynthetic and catabolic processes are enriched and engaged.

Finally, we also evaluated the proportion of PCG from each sub-genome in each module, however, no clear pattern emerged. Some modules were preferentially composed by PCG from a specific sub-genome but this data do not clearly correlate with total number of genes in a module nor the interconnectedness measured as the module density. Nevertheless, modules enriched for GO terms related to photosynthesis, transport and transcription are preferentially composed by PCG expressed from the *C. eugenioides* sub-genome. Similar results from EST of homoeolog data suggested that *C. arabica* may have specific physiological contributions derived from specific ancestors [12]. This finding supports our hypothesis that the *C. eugenioides* sub-genome is the main controller of gene expression whereas the *C. canephora* sub-genome may function as an important source of alleles to cope with the environment.

1.4 Discussion

This system-wide approach to investigate the transcriptome leaves of the allotetraploid *C. arabica*, a tropical perennial plant, revealed that processes such photosynthesis, cell wall biogenesis, translation, transcription, catabolism and biosynthesis are running in synchrony. Seemly contrasting rhythms and some clear noise are also perceived. To fully appreciate this complex orchestra, one would need to visit this concert hall throughout the day and throughout the year. Nevertheless, we had the opportunity to visit this spectacle while being simultaneously performed by two assemblies of players (PCG in sub-genomes). It is yet not clear if one sub-genome is louder than the other. Nevertheless, we suggest that the conductor is on the *C. eugenioides* side.

Our evaluation of reads uniquely mapped to the sub-genomes showed that homoeologous chromosomes are, in the majority of the cases, equally used for transcription (Supplemental Figure 4). Nevertheless, we verified that in chromosomes 2 and 10 the mapped reads derived from the *C. eugenioides* seem to be the majority when compared to the reads from its *C. canephora* counterpart. On a lesser level of intensity and significance, it appears that the Chromosome 6 from *C. canephora* ancestor is preferentially transcribed in comparison to its homoeologous (Supplemental Figure 4).

Other authors, using EST analyses, found that 48% of the *C. arabica* was transcribed from the *C. canephora* sub-genome and 52% were transcribed from the *C. eugenioides* sub-genome [12]. They also found that in 29% of 2,646 contigs had a higher contribution of one sub-genome in comparison to the other: 13% of the contigs had more ESTs from *C. eugenioides* sub-genome and 16% of contigs had more ESTs from *C. canephora* sub-genome [12]. Our results support the finding of differential sub-genome transcription. Here we show that chromosome organization and RNA processing PCG are preferentially encoded in the *C. eugenioides* sub-genome. Many of their homoeologs may have suffered processes of pseudogenization or gene deletion.

We hypothesize that by regulating the transcription and RNA transport, the *C. eugenioides* sub-genome has an edge in the decision making regulatory networks. However, the differential contribution of homoeologous genes to the transcriptome does not necessarily correlate with genome-wide transcription levels [65]. That way, one specific allele from a sub-genome may be preferentially expressed in a given condition, but it is unlikely, in the allotetraploid state of the *C. arabica* genome, that one sub-genome can monopolize the whole transcriptome.

Many universal orthologs - that represent the expected eudicot essential genes - are missing from the sub-genomes - especially the one inherited from the *C. canephora* ancestral parent (Figure 5). We assume that this lack of sub-genome-specific essential genes, reflected in the missing BUSCO signatures, is due to selective pressures that benefit gene expression under single copy control [51]. It is possible that after the origin of *C. arabica* processes of genome reorganization occurred during its evolution, ultimately causing the loss of homoeotic genes in the *C. canephora* sub-genome [66]. It is common that after episodes of whole genome duplications multiple genes are lost due to processes of epigenetic silencing, pseudogenization and chromosome level deletions of segments containing genes [17].

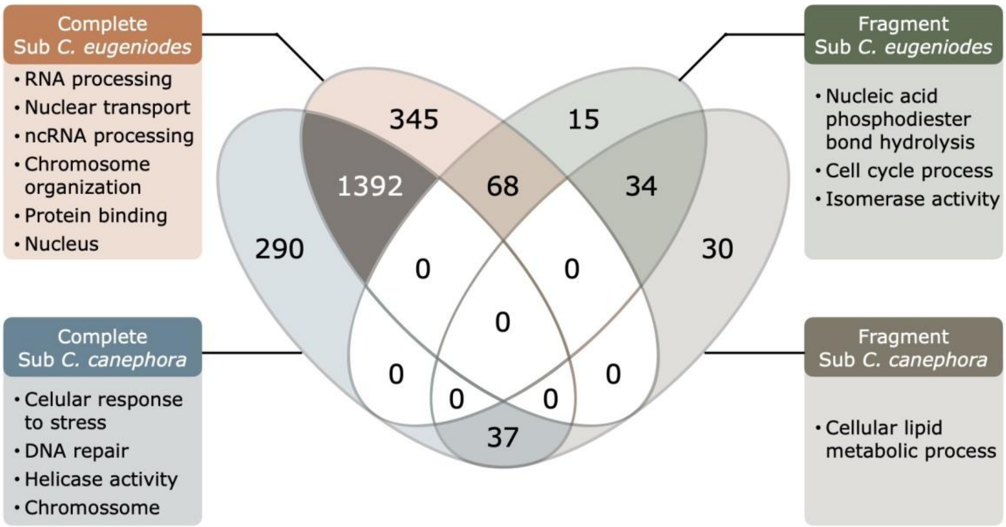

Figure 5 Venn diagram comparing the content of complete and fragmented BUSCO signatures of each *C. arabica* sub-genomes. About 63% (1,393) of the found BUSCO signatures are shared between both sub-genomes. A large proportion of expected universal orthologues (~13%) were not found in the *C. canephora* sub-genome. The *C. eugenioides* sub-genome is particularly enriched for complete signatures related to RNA processing, chromosome organization and nucleus while the fewer complete signatures in the *C. canephora* sub-genome are related to responses to stresses and DNA repair.

It is yet not clear if the interspecific hybridization that created *C. arabica* had occurred in the scale of millions [66] or tens of thousands of years [1,8], but it is agreed that it is a recent event in evolutionary time scale [8,12]. Recently formed polyploid species have higher extinction rates than their diploid relatives, rarely surviving over the long term and, in most cases, are evolutionary dead-ends [67]. Nevertheless, the few thriving individuals leave a substantial legacy in plant genomes [67]. Allopolyploidy can induce rapid genome evolution that may have contributed to the successful establishment of newly formed species [68,67].

We address that the key for allowing the survival and reproduction of *C. arabica* is the homoeologous gene loss through sequence deletion. Evidence for this hypothesis is that sequence elimination is a common and rapid phenomenon that can be verified as early as the generation F1 of newly formed hybrids [69,70]. It is also possible that the missing genes in the *C. eugenioides* sub-genome turned to pseudo-genes because of the finding that gene lengths in this sub-genome are usually longer than its homoeolog counterpart. Alternatively, it is possible that the abundant PCG with GO terms related to "DNA integration"- such hundreds of *DDE-TYPE INTEGRASE/TRANSPOSASE/RECOMBINASE* in both sub genomes - are promoting ongoing insertions to *C. eugenioides* sub-genome that may be causing this gene enlargement and pseudogenization.

A previous report on a set of 9,047 genes in a *C. arabica* scaffold level genome found homoeologous copies being retained in 98% of the analyzed dataset and suggested that homoeolog losses are not a random event [4]. These authors stated that the majority of the 2% of homoeologous losses were probably due to sequence homogenization instead of sequence deletion because they verified a lack of homoeologous single nuclear polymorphisms [4]. Our results are apparently contradicting these findings, mostly because of methodological differences. We focused our analysis on the expected homoeologous with BUSCO signatures in the *C. arabica* genome and not on a subset of genes selected by DNAseq depth - mapped to the reference genome of the relative species *C. canephora* [9] - and polymorphisms [4].

Genetic and epigenetic changes are common consequences of polyploidization in many major agricultural crop plants [71], including wheat [68,69], cotton[72], rapeseed [53] and sugarcane [73]. Among the genetic changes verified in recent polyploids is gene loss [52]. A study with resynthesized *Brassica napus* lines - independently derived by hybridizing double haploids of *Brassica oleracea* and *Brassica rapa* - revealed the occurrence of a 'genomic shock' leading to gene loss in early generations following the polyploid formation [53]. The maintained copies tend to concentrate in one of the homoeologous chromosomes [17]. A similar rapid gene loss event may have happened early in *C. arabica* evolution and may be the reason why essential universal orthologs are missing from the sub-genomes (Figure 5).

1.4.1 The gene loss in *Coffea arabica* is pronounced for homoeologous essential genes controlling gene expression.

An important process in shaping genomes is purifying selection that tends to remove duplicated genes after whole genome duplications[74]. However, this same purifying selection causes a tendency of preservation of balanced gene activity [75]. The evolution towards balance is favored by natural selection as evidenced by the finding that angiosperms are paleopolyploid and many of them are now in a diploid state as the tetraploid genome tends to merge [75].

The latest whole genome duplication event (the α duplication) in *Arabidopsis thaliana* happened between 83 to 86 million years ago [75]. During this geologic time, only 30% of *Arabidopsis genes* have retained syntenic copies, suggesting a large impact of gene loss on angiosperm evolution [75]. Some duplicated genes have a tendency of escaping this selective pressure whereas others are easy prey as predicted by the Gene Balance Hypothesis [16]. This resistance to genic duplication loss is a common evolutionary phenomenon if the duplication affects dosage dependent genes [16]. That way, genes which the number of its protein products are stoichiometrically calibrated to other related proteins - as the components of ribosomes - have an evolutionary tendency of keeping its duplicated copies once imbalance can cause reduced fitness [16].

This balancing in the number of copies is often reflected in co-expression modules [17,76]. The group of genes that are connected due to serialized molecular function - such enzymes in a pathway or transcriptional cascades - will be more resilient to gene loss while other groups will be more advantageous existing in single copies in the haplotype. The identified co-expression modules of *C. arabica* leaf transcriptome tended to have transcriptome components preferentially derived from one sub-genome, instead of having roughly the same number of components from each sub genome (Supplemental Table 5). It is possible that this differential module composition is a consequence of a balancing process that calibrates homoeologous copy numbers.

The regulation of transcription in *C. arabica* seems to be preferentially encoded by the *C. eugenioides* sub-genome. Module Brown, which is significantly enriched for the GO term “gene expression”, has a strong tendency of being composed by genes in the *C. eugenioides* sub-genome (65%). We initially assumed that the RNA processing machinery - the hardware component of gene expression - would mostly work as the product of dosage dependent genes. Nevertheless, the lack of essential loci related to gene expression in *C. canephora* sub-genome shows that pressures of single copy control were an important force shaping coffee evolution.

The concentration of the transcriptional decision making machinery in the *C. eugenioides* codebase may have also provided advantageous phenotypic homeostasis in response to the environment [77,78]. The maintenance of this genomic rearrangements may have been facilitated by the preferential autogamy of *Coffea arabica*. Once it lost its selfincompatibility machinery [66], it was also possible that other pieces of its chromosomes were rapidly lost due to inbreeding depression.

C. arabica is a young species [1] that seems to be evolving towards a simpler genome configuration. The stabilization of its genome left marked differences in homoeolog gene composition which was made evident by the loss of universal orthologues. In the million years to come, we expect a natural tendency of gene loss and chromosome merging. Finally, this system-wide approach clarified how biological processes (i.e. photosynthesis, cell wall biogenesis, translation, transcription, catabolism and biosynthesis) are running in synchrony. Thus, this work contributes to comprehending genome evolution of recent polyploids and supports crop breeding programs through future functional studies considering the eigengenes, unknown and/or species-specific genes found in coffee.

REFERENCES

1. Scalabrin, S. *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.* **10**, 1–13 (2020).

2. Charrier, A. & Berthaud, J. Botanical Classification of Coffee. in *Coffee: Botany, Biochemistry and Production of Beans and Beverage* (eds. Clifford, M. N. & Willson, K. C.) 13–47 (Springer US, 1985). doi:10.1007/978-1-4615-6657-1_2.

3. Davis, A. P., Govaerts, R., Bridson, D. M. & Stoffelen, P. An annotated taxonomic conspectus of the genus Coffea (Rubiaceae). *Bot. J. Linn. Soc.* **152**, 465–512 (2006).

4. Lashermes, P., Hueber, Y., Combes, M.-C., Severac, D. & Dereeper, A. Inter-genomic DNA Exchanges and Homeologous Gene Silencing Shaped the Nascent Allopolyploid Coffee Genome (Coffea arabica L.). *G3 GenesGenomesGenetics* **6**, 2937–2948 (2016).

5. Tran, H. *et al.* SNP in the Coffea arabica genome associated with coffee quality. *Tree Genet. Genomes* **14**, 1–15 (2018).

6. Tran, H. T. M., Ramaraj, T., Furtado, A., Lee, L. S. & Henry, R. J. Use of a draft genome of coffee (Coffea arabica) to identify SNPs associated with caffeine content. *Plant Biotechnol. J.* **16**, 1756–1766 (2018).

7. Mekbib, Y. *et al.* Whole-genome resequencing of Coffea arabica L. (Rubiaceae) genotypes identify SNP and unravels distinct groups showing a strong geographical pattern. *BMC Plant Biol.* **22**, 69 (2022).

8. Cenci, A., Combes, M.-C. & Lashermes, P. Genome evolution in diploid and tetraploid Coffea species as revealed by comparative analysis of orthologous genome segments. *Plant Mol. Biol.* **78**, 135–145 (2012).

9. Denoeud, F. *et al.* The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. *science* **345**, 1181–1184 (2014).

10. Glover, N. M., Redestig, H. & Dessimoz, C. Homoeologs: What Are They and How Do We Infer Them? *Trends Plant Sci.* **21**, 609–621 (2016).

11. Pinto-Maglio, C. A. F. Cytogenetics of coffee. *Braz. J. Plant Physiol.* **18**, 37–44 (2006).

12. Vidal, R. O. *et al.* 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 (2010).

13. Wolfe, K. H. Yesterday’s polyploids and the mystery of diploidization. *Nat. Rev. Genet.* **2**, 333–341 (2001).

14. Koonin, E. V. & Wolf, Y. I. Evolutionary systems biology: links between gene evolution

and function. *Curr. Opin. Biotechnol.* **17**, 481–487 (2006).

15. Papp, B., Pál, C. & Hurst, L. D. Dosage sensitivity and the evolution of gene families in yeast. *Nature* **424**, 194–197 (2003).

16. Freeling, M. & Thomas, B. C. Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome Res.* **16**, 805–814 (2006).

17. Sankoff, D., Zheng, C. & Zhu, Q. The collapse of gene complement following whole genome duplication. *BMC Genomics* **11**, 313 (2010).

18. Johns Hopkins University. Coffea arabica V. Caturra Genome. (2018).

19. Cheng, B., Smyth, H. E., Furtado, A. & Henry, R. J. Slower development of lower canopy beans produces better coffee. *J. Exp. Bot.* **71**, 4201–4214 (2020).

20. Cheng, B., Furtado, A. & Henry, R. J. The coffee bean transcriptome explains the accumulation of the major bean components through ripening. *Sci. Rep.* **8**, 1–11 (2018).

21. de Oliveira, R. R. *et al.* Elevated temperatures impose transcriptional constraints on coffee genotypes and elicit intraspecific differences in thermoregulation. *Front. Plant Sci.* **11**, 2020.03.07.981340 (2020).

22. Stavrinides, A. K. *et al.* Seed comparative genomics in three coffee species identify desiccation tolerance mechanisms in intermediate seeds. *J. Exp. Bot.* **71**, 1418–1433 (2020).

23. dos Santos, T. B. *et al.* An integrated analysis of mRNA and sRNA transcriptional profiles in Coffea arabica L. roots: insights on nitrogen starvation responses. *Funct. Integr. Genomics* **19**, 151–169 (2019).

24. Cardon, C. H. *et al.* Expression of coffee florigen CaFT1 reveals a sustained floral induction window associated with asynchronous flowering in tropical perennials. *Plant Sci.* **325**, 111479 (2022).

25. Davis, M. P. A., van Dongen, S., Abreu-Goodger, C., Bartonicek, N. & Enright, A. J. Kraken: A set of tools for quality control and analysis of high-throughput sequence data. *Methods* **63**, 41–49 (2013).

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

27. Hoff, K. J. & Stanke, M. Predicting Genes in Single Genomes with AUGUSTUS. *Curr. Protoc. Bioinforma.* **65**, e57 (2019).

28. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).

29. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

30. Picard toolkit. *Broad Institute, GitHub repository* (2019).

31. Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M. & Stanke, M. BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics* **32**, 767–769 (2016).

32. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).

33. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinforma. Oxf. Engl.* **22**, 1658–1659 (2006).

34. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).

35. Gremme, G., Brendel, V., Sparks, M. E. & Kurtz, S. Engineering a software tool for gene structure prediction in higher organisms. *Inf. Softw. Technol.* **47**, 965–978 (2005).

36. Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, 1–22 (2008).

37. Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).

38. Götz, S. *et al.* High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res.* **36**, 3420–3435 (2008).

39. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl Acids Res* **25**, (1997).

40. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733-745 (2016).

41. Blum, M. *et al.* The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.* **49**, D344–D354 (2021).

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

43. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

44. Anders, S., Pyl, P. T. & Huber, W. HTSeq - A Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).

45. Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* **4**, (2005).

46. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal Complex Syst.* **1695**, 1–9 (2006).

47. Yan, H. *et al.* agriGO v2.0: a GO analysis toolkit for the agricultural community, 2017 update. *Nucleic Acids Res.* **45**, W122–W129 (2017).

48. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B Methodol* **57**, (1995).

49. Coffea arabica Annotation Report. https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Coffea_arabica/100/ (2018).

50. Waterhouse, R. M. *et al.* BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol. Biol. Evol.* **35**, 543–548 (2018).

51. Waterhouse, R. M., Zdobnov, E. M. & Kriventseva, E. V. Correlating Traits of Gene Retention, Sequence Divergence, Duplicability and Essentiality in Vertebrates, Arthropods, and Fungi. *Genome Biol. Evol.* **3**, 75–86 (2011).

52. Buggs, R. J. A. *et al.* Rapid, Repeated, and Clustered Loss of Duplicate Genes in Allopolyploid Plant Populations of Independent Origin. *Curr. Biol.* **22**, 248–252 (2012).

53. Gaeta, R. T., Pires, J. C., Iniguez-Luy, F., Leon, E. & Osborn, T. C. Genomic Changes in Resynthesized Brassica napus and Their Effect on Gene Expression and Phenotype. *Plant Cell* **19**, 3403–3417 (2007).

54. Yang, L., Takuno, S., Waters, E. R. & Gaut, B. S. Lowly Expressed Genes in Arabidopsis thaliana Bear the Signature of Possible Pseudogenization by Promoter Degradation. *Mol. Biol. Evol.* **28**, 1193–1203 (2011).

55. Urrutia, R. KRAB-containing zinc-finger repressor proteins. *Genome Biol.* **4**, 1–8 (2003).
56. Ponnu, J., Wahl, V. & Schmid, M. Trehalose-6-Phosphate: Connecting Plant Metabolism and Development. *Front. Plant Sci.* **2**, (2011).

57. Vaahtera, L., Schulz, J. & Hamann, T. Cell wall integrity maintenance during plant development and interaction with the environment. *Nat. Plants* **5**, 924–932 (2019).

58. Murata, N. & Los, D. A. Membrane Fluidity and Temperature Perception. *Plant Physiol.* **115**, 875–879 (1997).

59. Mittler, R., Finka, A. & Goloubinoff, P. How do plants feel the heat? *Trends Biochem. Sci.* **37**, 118–125 (2012).

60. Uhrig, R. G., Schläpfer, P., Roschitzki, B., Hirsch-Hoffmann, M. & Gruissem, W. Diurnal changes in concerted plant protein phosphorylation and acetylation in Arabidopsis organs and seedlings. *Plant J.* **99**, 176–194 (2019).

61. Chiou, T.-J. & Lin, S.-I. Signaling Network in Sensing Phosphate Availability in Plants. *Annu. Rev. Plant Biol.* **62**, 185–206 (2011).

62. eWang, G., eZhang, C., Battle, S. L. & eLu, H. *The Phosphate Transporter PHT4;1 is a Salicylic Acid Regulator Likely Controlled By the Circadian Clock Protein CCA1*. https://doaj.org/article/93006ec94bb04334be9e39037925415c (2014).

63. Wang, G.-Y. *et al.* Circadian Clock-Regulated Phosphate Transporter PHT4;1 Plays an Important Role in Arabidopsis Defense. *Mol. Plant* **4**, 516–526 (2011).

64. Chan, C., Liao, Y.-Y. & Chiou, T.-J. The Impact of Phosphorus on Plant Immunity. *Plant Cell Physiol.* **62**, 582–589 (2021).

65. Mochida, K., Yamazaki, Y. & Ogihara, Y. Discrimination of homoeologous gene expression in hexaploid wheat by SNP analysis of contigs grouped from a large number of expressed sequence tags. *Mol. Genet. Genomics* **270**, 371–377 (2004).

66. Lashermes, P. *et al.* Molecular characterisation and origin of the Coffea arabica L. genome. *Mol. Gen. Genet. MGG* **261**, 259–266 (1999).

67. Arrigo, N. & Barker, M. S. Rarely successful polyploids and their legacy in plant genomes. *Curr. Opin. Plant Biol.* **15**, 140–146 (2012).

68. Ozkan, H., Levy, A. A. & Feldman, M. Allopolyploidy-induced rapid genome evolution in the wheat (Aegilops-Triticum) group. *Plant Cell* **13**, 1735–1747 (2001).

69. Shaked, H., Kashkush, K., Ozkan, H., Feldman, M. & Levy, A. A. Sequence Elimination and Cytosine Methylation Are Rapid and Reproducible Responses of the Genome to Wide Hybridization and Allopolyploidy in Wheat. *Plant Cell* **13**, 1749–1759 (2001).

70. Ma, X.-F. & Gustafson, J. P. Timing and rate of genome variation in triticale following allopolyploidization. *Genome* **49**, 950–958 (2006).

71. Madlung, A. & Wendel, J. F. Genetic and Epigenetic Aspects of Polyploid Evolution in Plants. *Cytogenet. Genome Res.* **140**, 270–285 (2013).

72. Wendel, J. F., Schnabel, A. & Seelanan, T. Bidirectional interlocus concerted evolution following allopolyploid speciation in cotton (Gossypium). *Proc. Natl. Acad. Sci.* **92**, 280–284 (1995).

73. Zhang, J. *et al.* Sugarcane genetics and genomics. *Sugarcane Physiol. Biochem. Funct. Biol.* 623–643 (2013).

74. Kondrashov, F. A., Rogozin, I. B., Wolf, Y. I. & Koonin, E. V. Selection in the evolution of gene duplications. *Genome Biol.* **3**, research0008.1 (2002).

75. Bowers, J. E., Chapman, B. A., Rong, J. & Paterson, A. H. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* **422**, 433–438 (2003).

76. Coate, J. E., Schlueter, J. A., Whaley, A. M. & Doyle, J. J. Comparative Evolution of Photosynthetic Genes in Response to Polyploid and Nonpolyploid Duplication. *Plant*

Physiol. **155**, 2081–2095 (2011).

77. Bertrand, B. *et al.* The greater phenotypic homeostasis of the allopolyploid Coffea arabica improved the transcriptional homeostasis over that of both diploid parents. *Plant Cell Physiol.* **56**, 2035–2051 (2015).

78. Marques, I. *et al.* A transcriptomic approach to understanding the combined impacts of supra-optimal temperatures and co2 revealed different responses in the polyploid coffea arabica and its diploid progenitor c. Canephora. *Int. J. Mol. Sci.* **22**, 1–21 (2021).

CHAPTER 2

2. ARTICLE 2 - METABOLIC PATHWAY RECONSTRUCTION INDICATES THE PRESENCE OF IMPORTANT MEDICAL COMPOUNDS IN *Coffea* SUCH AS L-DOPA.

If a link does not work, please copy and paste into the browser and remove eventual spaces

SUPPLEMENTAL FIGURES can be accessed using the link:

https://drive.google.com/drive/folders/1ETTiPulbt4ldin4zP7wWX8Rik9Awz8vW?usp=share_link

SUPPLEMENTAL TABLES can be accessed using the link:

https://drive.google.com/drive/folders/1GFsjnjj6OvvcgfwJ8UxIfg4GbYZ99AoB?usp=share_link

SUPPLEMENTAL DATASETS can be accessed using the link:

https://drive.google.com/drive/folders/1HSUIDf2e0LnrzpVdXrWqu74JObnpIBUH?usp=share_link

SCRIPTS can be accessed using the link:

https://github.com/thalescherubino/thesisChapter2

(Draft Version)

Article prepared for submission to the Plant Physiology and Biochemistry Journal

Metabolic pathway reconstruction indicates the presence of important medical compounds in *Coffea* such as L-DOPA.

Thales Henrique Cherubino Ribeiro, Raphael Ricon de Oliveira, Taís Teixeira das Neves, Wilder Douglas Santiago, Bethania Leite Mansur, Adelir Aparecida Saczk, Mario Lucio Vilela de Resende, Antonio Chalfun-Junior

Abstract

The use of transcriptomic data to make inferences about plant metabolomes is a useful tool to help the discovery of important compounds in the available biodiversity. In this work we applied *in silico* techniques to reveal possible metabolites in the leaves of *Coffee arabica* L. By mapping RNAseq reads sequenced from fully expanded leaves to protein coding genes we were able to access metabolic pathways responsible for the production of several compounds of economic importance. L-DOPA, the precursor of dopamine, is a common product of *POLYPHENOL OXIDASES* (PPOs). This compound is widely used as a treatment of the human neurodegenerative condition called Parkinson's disease. Here, we applied *in vitro* studies to validate *in silico* results showing that L-DOPA is a naturally occurring metabolite in coffee.

2.1 Introduction

Coffee (Rubiaceae) is an important crop in which beans are harvested and roasted before being traded as a commodity [1]. It is produced mostly in tropical countries and is an important source of livelihood for millions of smallholder farmers and workers involved in the various steps of the coffee bean processing and trade [2]. *Coffea arabica* L. is the main source of the coffee beans. This species is an interspecific hybrid of *Coffea canephora* Pierre and *Coffea eugenioides* Moore ancestors [3]. Of those parental species, only *C. canephora* is also cultivated for economic purposes. The polyploidy of *C. arabica* (2n = 44) may provide physiological advantages to cope with abiotic stresses, improve phenotypic homeostasis [4] and allow a broader diversification of metabolite compounds - when compared to progenitor species - through the differential expression of homoeologous genes [5].

Apart from its beans, the coffee leaves have the potential to also become a source of metabolites of economic importance [6–10]. The tea produced from coffee leaves is rich in natural polyphenolic compounds such as chlorogenic acids and xanthones which are important dietary antioxidants that significantly benefit human health [7]. To further extend the portfolio of known bioactive compounds in coffee leaves, we applied bioinformatic methodologies with the aim of investigating metabolic pathways that may be constitutively expressed in *Coffea canephora* and *Coffea arabica* leaves. Our *in silico* analyses showed that *POLYPHENOL OXIDASES* (PPOs) and *DOPA DESCARBOXILASES* (DDCs) are expressed in leaves of both the economically important coffee species.

PPOs are type-III copper containing metalloenzymes divided into three types; tyrosinases (TYRs, EC 1.14.18.1 and EC 1.10.3.1), catechol oxidases (COs, EC 1.10.3.1) and aurone synthases (AUSs). AUS is a type of PPOs that are responsible for the synthesis of yellow pigments in petals of various Asteraceae species [11]. PPOs is one of the oldest enzymes known [12] and are wide-spread across all life kingdoms [13–19] with biological roles varying [16,20]. PPOs catalyzes the oxidation of catechol to o-quinone in the presence of oxygen. The main difference between TYRs and COs is that the latter can only catalyzes the oxidation of catechol (i.e., o-diphenol) to the corresponding o-quinone whereas the former can catalyze both the monooxygenation of monophenols and the oxidation of catechols [21].

In animals and many microorganisms PPOs are directly involved in the production of Melanin pigments by oxidizing L-tyrosine (TYR) to L-DOPA (L-3,4-dihydroxyphenylalanine; levodopa), and others metabolites, ultimately producing dark color pigments [20,22]. Similarly, plant PPO can produce dark-brownish compounds [23]. This browning process is the result of the oxidation of phenolics to quinones that are highly reactive intermediates involved in senescence, wounding and response to pathogens [23]. After fruit harvest, the accumulation of those metabolites becomes evident. In many plant-derived foodstuffs those reactions can reduce their nutritional quality and perceived commercial value [12,24,25].

It has been estimated that half of the world's fruits and vegetable crops are lost due to postharvest deteriorative reactions [26]. This browning might be a side effect of fundamental defense responses because when PPOs are transcriptionally repressed in tomatoes (*Lycopersicon esculentum* L.) their susceptibility to the *P. syringae* is increased [23] whereas the overexpression of PPOs reduces the susceptibility to the same pathogen [27].

Apart from tyrosine, PPOs can accept diverse types of both monophenols and *o*-diphenols as substrates in different species and tissues. The monophenol substrates include, but are not limited to, 4-methylphenol [28], 4-propylphenol [29], 4-tert-butylphenol [30], 4-

aminophenol [31] and p-tyrosol [29]. Similarly, *o*-diphenol substrates include 4-methylcatechol [28], L-3,4-dihydroxyphenylalanine (L-DOPA) [32], 4-tert-butylcatechol [32], 3,4-dihydroxyphenethylamine (dopamine) [29], caffeic acid [29], and 5-caffeoylquinic acid (5CQA; Chlorogenic Acid) [33]. In *Coffea arabica* leaves and endosperm, the most efficient PPO substrates was found to be 5CQA followed by 4-methilcatecol, caffeic acid and catechol [33]. There is evidence that *C. arabica* cultivars in which leaves have higher concentration of 5CQA are more resistant to *Hemileia vastatrix,* a pathogen fungus to coffee plants and the causal agent of coffee leaf rust [6]. In addition, the PPOs activity is higher in young coffee leaves and decreases with increasing leaf length and age [33,34].

In humans and other animals L-DOPA is an important precursor of the neurotransmitter dopamine [35,36]. The depletion of dopamine in the human brain causes the neurodegenerative condition called parkinson's disease [37]. In humans, dopamine cannot cross the morphological barrier at the blood-brain interface while L-DOPA can [38]. Once L-DOPA enters the central nervous system it is converted into dopamine by the enzyme DOPA Descarboxilase (DDC; EC 4.1.1.28). Because of the ability of the L-DOPA compound to cross the blood-brain barrier and then be metabolized by DDC into dopamine in the brain, it is used as the standard treatment for Parkinson's disease [35].

Plant DDCs are common enzymes that mediate numerous secondary reactions [39]. Nevertheless, the full extension of their biological function in plant growth and development remains unknown. DDC over-expression in apple trees increased both the dopamine levels and salt tolerance [39]. This effect may be due to enhanced maintenance of ion homeostasis that is verified after dopamine treatment [40]. It has been reported that dopamine can promote nutrient uptake, transport and distribution as well as promoting the down-regulation of senescence related genes [41]. In cucumber, it has been shown to meditate photosynthesis, carbon and nitrogen metabolism and reduce damage under nitrate stress [42]. Because dopamine can mediate important biological processes in plants, its production by DDC may be of relevance in coffee.

The full phenolic composition of *C. arabica* remains unknown [43,44]. To investigate potential metabolites we evaluate the transcriptome profile of multiple *Coffea arabica* RNAseq samples that are publicly available at the Sequence Read Archive (SRA) from the National Center for Biotechnology Information (NCBI). Then, we compared the expressed genes with metabolic pathways available at the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [45].

We focused our attention on genes coding for the enzymes PPOs and DDC that are present in multiple copies in the genomes of *C. arabica* and *C. canephora.* We used High Performance Liquid Chromatography with tandem Mass Spectrometry (HPLC-MS/MS) techniques to show that L-DOPA is a phenolic metabolite that naturally occurs in *Coffea* leaves and fruits. To our knowledge, L-DOPA has never been reported as a naturally occurring phenolic compound in *Coffea*, though there is *in vitro* evidence that a PPO extracted from coffee endosperm can accept L-DOPA as a substrate, however with an activity 25.6 times lower than its preferred substrate, 5CQA [33]. The co-expression of PPO and DDC suggests that dopamine is also present in leaves of those species.

Those results show that *in silico* analysis coupled with analytical chemistry techniques is a powerful combination of toolsets to allow the identification of compounds of economic and pharmacological importance in plants. In addition, these findings may provide additional base to the use of coffee leaves as a source of phenolic metabolites with medicinal, phototherapeutic and economic value.

2.2 Material and Methods

2.2.1 Identification of enzyme coding genes in *Coffee* genome and inference of metabolic pathways

A total of 25,605 predicted protein sequences in *Coffea canephora* were downloaded from the Coffee genome hub [46]. Then, they were analyzed with the blast2GO [47] suite to search potential enzymes and their respective Enzyme Codes (EC). We used the resulting list of 1,141 non redundant EC in the online KEGG mapper tool [45] to find metabolic pathways that were possibly active in *Coffea.* We included in our search the following reference KEGG pathway map databases: Carbohydrate metabolism (1.1), Energy metabolism (1.2), Lipid metabolism (1.3), Amino acid metabolism (1.5), Metabolism of cofactors and vitamins (1.8), Metabolism of terpenoids and polyketides (1.9) and Biosynthesis of secondary metabolites (1.10).

2.2.2 Characterization of PPOs and DDCs coding genes in *Coffea*

The genome sequence for the *Coffea arabica* Caturra-red cultivar was retrieved from the NCBI under BioProject accession PRJNA506972 [48]. Then we predicted protein coding genes using AUGUSTUS v. 3.3.3 [49]. The annotation of protein coding genes was performed using blast2GO [47]. Sequences with the enzymatic code for PPOs (EC 1.14.18.1 and EC 1.10.3.1) and DDC (EC 4.1.1.28) where selected from both *Coffea arabica* and *Coffea canephora* and conserved domain analyses were performed using hidden Markov models by aligning the selected protein sequences against the Pfam domain database v35 [50] with the HMMER software v3.3.2 [50]. Finally, the respective coding sequences for each putative PPO and DDC were aligned to the NCBI non-redundant (nr) protein database using blastx v 2.12.0+ [51].

We considered Coffee PPOs those protein sequences that (1) presented significant hits to the three typical plant PPO domains in the following amino-carboxyl order: Tyrosinase (PF00264), Polyphenol oxidase middle domain (PPO1_DWL; PF12142) and PPO1_KFDV (PF12143); (2) sequences with at least 70% of coverage and 50% of identity to the 3D-chistolograph verified PPO structure from *Ipomoea batatas* (UniProtKB/Swiss-Prot: Q9MB14.2) [52]; (3) sequences with the five blastx top-hits of known plant PPOs. In addition, we considered the coffee DDC homologs those protein sequences that (1) presented significant hits to the Pyridoxal-dependent decarboxylase conserved domain (PF00282.22); (2) sequences with at least 70% of coverage and 50% of identity to the curated DDC from *Papaver somniferum* (UniProtKB/Swiss-Prot: P54768); (3) sequences with the five blastx top-hits of known plants DDC - and not the highly similar proteins L-tryptophan decarboxylase (TDC2-like). Then, their physicochemical properties (length of amino acid sequence, molecular weight, and isoelectric point) were determined with the ExPASy Proteomics tool (**https://web.expasy.org/protparam/**).

2.2.3 Phylogenetic Analysis

Representative protein sequences for PPOs and DDC were retrieved from NCBI's nr database for the following taxa: Amborellales, Arecales, Asparagales, Asterales, Brassicales, Cannabaceae, Cucurbitales, Fabales, Gentianales, Ginkgoales, Lamiales, Liliales, Lycophytes, Malpighiales, Malvaceae, Poales, Ranunculales, Rosales, Solanales, Vitales and Zingiberales. A Tyrosinase from *Homo sapiens* (GenBank accession AAA61244.1) was used as an out-group for plant PPOs and a human DDC (NCBI accession: NP_000781.2) was used as an out-group for plant DDC. The selected plant species represents diverse phylogenetic groups of higher plants. To this subset of PPOs or DDCs we added the respective homologs from *Coffea arabica* and *Coffea canephora.*

The multiple protein sequence alignment was performed with MAFFT v. 7.505 using the iterative refinement method incorporating global pairwise alignment information (G-INS-i) [53,54]. For inferring phylogenetic relationships, coffee sequences with more than 98% of identity at the protein level were collapsed into a single representative sequence. Phylogenetic trees were inferred with PHYLIP [55] v.3.696 with 1000 bootstrap replicates, using the Jones-Taylor-Thornton substitution model [56] and neighbour-joining clustering method [57]. The consensus tree was chosen by the majority rule and drawn using the Interactive Tree of Life (iTOL v. 6.5.8) webtool [58]. Transfer signal peptides were inferred with the online tool LOCALIZER v. 1.0.4 [59].

2.2.4 Expression evaluation of *PPOs* and *DDCs*

To identify expressed PPOs and DDCs in *Coffea arabica* leaves, we downloaded paired-end RNAseq libraries available at the Sequence Read Archive (SRA) of the NCBI under bioproject ID PRJNA851465 [60]. In summary, the experiment was conducted in Brazilian farms of two cities (Pirapora and Varginha) during two harvest times (April and October) and with two *C. arabica* cultivars (Acauã and Catuaí). After *inhouse* quality processing steps, the RNAseq reads were mapped to the *Coffea arabica* reference genome (BioProject accession PRJNA506972 [48]) using the STAR aligner v. 2.7.8 [61]. Then, fragments mapped to gene exons were quantified with the HTseq-count script [62], analyzed with the edgeR [63] and expression-based heatmap produced with the heatmap.2 function from gplot package [64].

2.2.5 Extraction of L-DOPA from *Coffea arabica* leaves

The extraction procedure was based on a sustainable, simple and robust method for L-DOPA extraction recently developed for *Vicia faba* [65]. We collected *Coffea arabica* leaves and immediately macerated them with liquid nitrogen until a fine and homogeneous powder was produced. Samples of 200 mg were collected in 15 ml tubes with 5 ml of acetic acid 0.1%. Then, we homogenized samples for 20 minutes with a magnetic shaker and subsequently centrifuged at 13,000 rpm for 10 minutes at environment temperature (± 25 °C). Next, we collected the supernatants and a second extraction step was performed with the remaining biomass. Lastly, we mixed and filtered both supernatants in a membrane and immediately submitted to chromatographic analysis.

2.2.6 Liquid Chromatographic analysis and validation parameters

The analyses were performed at the Brazilian National Institute of Coffee Science and Technology (Instituto Nacional de Ciência e Tecnologia do Café; INCT-Café) at the Federal University of Lavras (Universidade Federal de Lavras; UFLA). The liquid chromatographic runs were performed with a Shimadzu HPLC equipment composed of a high pressure quaternary pump model LC-20AT, a degasser DGU-20A5, an interface CBM-20A, an automatic injector SIL-20A-HT and an UV-Vis detector SPD-20A. The used column was a Zorbax Eclipse XDB-C18 (4.6 x 250 mm, 5 µm) connected to an XDB-C18 pre-column (4.6 x 12.5 mm, 5 µm).

The L-DOPA analysis was performed with the methodology proposed by Elbarbry *et al.* (2019) [66] with modifications. L-DOPA standard was purchased from Sigma-Aldrich (St. Louis, MO, USA). Mobile phase chemicals were all of HPLC analytical grade; metanol (Merck), glacial acetic acid (J.T.Baker) and type I water from a Milli-Q system.

We used the external standardization method to apply quantification procedures. For the analytical curves, we diluted a stock solution with the L-DOPA standard in perchloric acid (1,000 µg mL^{-1}). From that stock solution we prepared the analytical curve by varying the concentration from 0.1 to 200 µg mL^{-1}. The selected mobile phase for the compound elution was acetic acid 1% in water (Solvent A) and methanol (Solvent B) at the 95:5 (v/v) and flow rate of 1,0 mL min^{-1}. We eluted samples and standards in isocratic mode at 30 ºC in the column oven. The used light wavelength was 282 nm and the injection volume was 20 µL.

We filtered the biological samples and standard solutions in 0.45 µm polyethylene membrane (Millipore) and injected directly into the chromatographic system. The injections of the standards and biological samples were performed in triplicate, with the analyte identity confirmed by the retention time and the peak profile of the sample compared to that of the standard solution.

To ensure the analytical quality of the results, we evaluated multiple parameters such as selectivity, linearity, detection limit (DL), quantification limit (QL), precision (in terms of coefficient of variation; CV) and accuracy (recovery). All the procedures required to evaluate those parameters were performed to guarantee the standardization of the method [67–69]. Firstly, we evaluated the selectivity by adding to a pool of samples in different quantities of the L-DOPA standard. Then, we evaluated the linearity by inferring the linear regression equation and its respective correlation coefficient (R^2). A R^2 greater than 0.99 was considered as evidence of an ideal fit of the data to the model.

To verify the ascertainment of detection (DL) and quantification (QL) limits we considered the parameters related to the selectivity linear regression curve. To this end, we applied the following equations: DL = 3 x (s/S) and QL =10 x (s/S), where *s* is the standard deviation estimate of the linear regression model and *S* is its slope.

The precision was calculated by using the intermediate precision method. To do so, we repeated the HPLC analysis for 5 days by evaluating the readings of standard solutions with three known concentrations (1.0, 50.0 and 100.0 µg mL^{-1}). At the end, the Coefficient of Variation (CV), expressed as a percentage, was calculated with the function CV = (s/DMC) x 100, where *s* is the estimated standard deviation and *DMC* is the Determined Mean Concentration.

Finally, we evaluated the accuracy by running recovery assays using three random samples fortified with standard solutions at three concentration levels (1.0, 50.0 and 100.0 µg mL^{-1}). The recovery, expressed as a percentage of L-DOPA, was determined using the equation: recovery = [(measured concentration)/(expected concentration)] x 100.

2.2.7 LC-MS/MS for qualitative analyses

To verify the occurrence of L-DOPA in *Coffea,* sample extracts in triplicate from *Coffea arabica* and *Coffea canephora* leaves, flowers and fruits were analyzed by LC-MS/MS. Those analyses were performed in an Agilent Technologies system consisting of a binary pump, a degassing unit, a G4226A autosampler, a column oven and a triple quadrupole mass spectrometer (QqQ G6420A). The system was controlled by MassHunter Workstation Software (Version B.08.00). The separation was carried out on Zorbax Eclipse XDB-C18, 4.6 x 250 mm x 5 µm, thermostated at 30 ºC, using a mobile phase composed of acetic acid 1% in water (Solvent A) and methanol (Solvent B) at the 95:5 (v/v), with a flow rate of 1.0 mL min^{-1}. Full scan spectra were acquired from m/z 10 to 500. Identification of L-DOPA was performed in Multiple Reaction Monitoring (MRM) mode detecting the transitions m/z 198 → m/z 152, m/z 198 → m/z 107 and m/z 198 → m/z 135 [70–72].

2.3 Results

2.3.1 Exploratory analyses of metabolic pathways shows that PPOs and DDCs are present in the genome of *Coffea arabica* and *Coffea canephora*

Our exploratory analyses of metabolic pathways in *Coffea* revealed important enzymes central for the survival of plants, such as the ones involved in the carbon fixation in photosynthetic organs (Supplemental Figure 1). Among the metabolites predicted to occur along the investigated pathways (Supplemental Figure 2) we focused our attention on L-DOPA that is a substrate of DDC and a product of PPO (Figure 1).

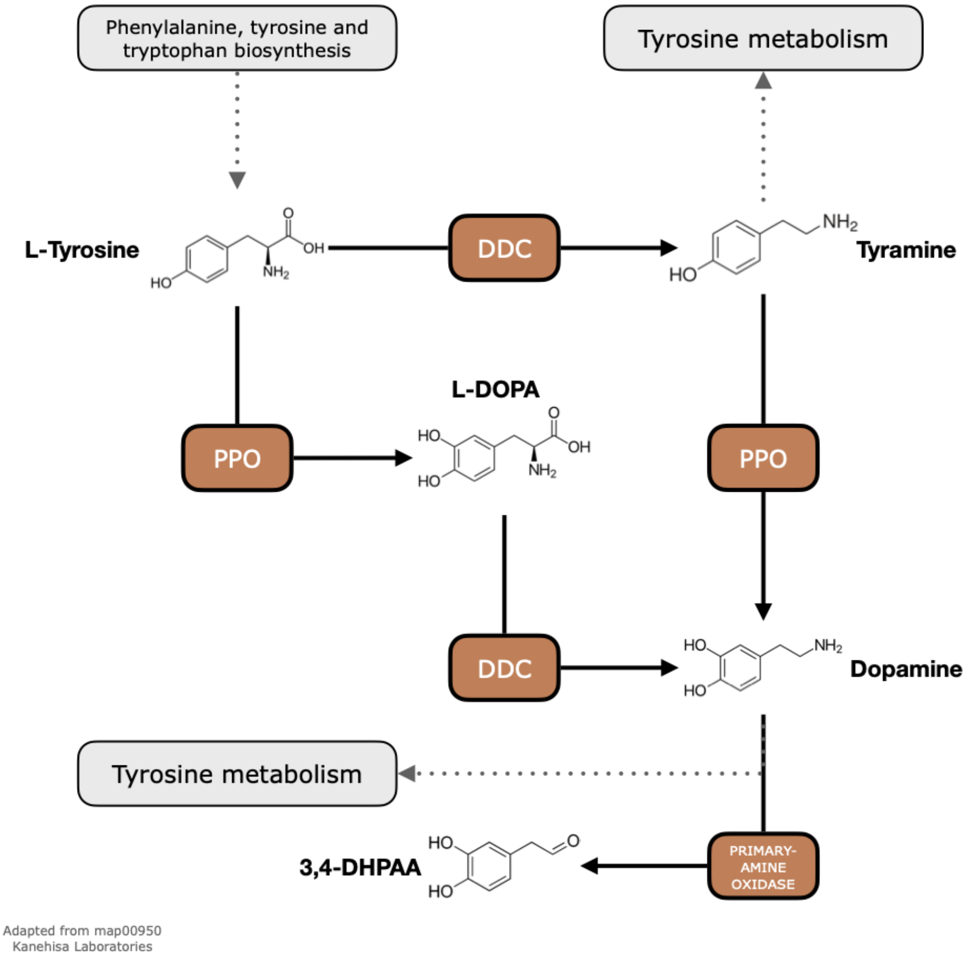

Figure 1 Representation of part of the isoquinoline alkaloid biosynthesis pathway according to the KEGG map00950. Here we show only the enzymes which gene could be identified in the Coffee genome. *POLYPHENOL OXIDASE* (PPOs; EC:1.14.18.1 and EC:1.10.3.1) and *DOPA DESCARBOXILASE* (DDC; EC:4.1.1.28) can both accept the amino acid L-tyrosine as a substrate and produce L-DOPA (L-3,4-dihydroxyphenylalanine) and Tyramine, respectively.

> Then, L-DOPA becomes an intermediate metabolite that can be used as a substrate to DDC to produce Dopamine. Alternatively, Dopamine can also be produced from a Tyramine substrate in a PPO enzyme. Among other factors, the final concentration of L-DOPA and Dopamine will depend on the downstream pathways such as the Tyrosine metabolism or the synthesis of 3,4-DHPAA 3,4-Dihydroxyphenylacetaldehyde) by a Primary-amine oxidase (EC:1.4.3.21).

We found eight *PPOs* in the *Coffea arabica* genome of which seven are encoded by the *C. canephora* sub-genome and one by the *C. eugenioides* sub-genome (Supplemental Table 1). In addition, we found three *PPOs* in the genome of *Coffea canephora* (Supplemental Table 1). All the PPOs identified in the *Coffea arabica* genome were named PPO.CAR followed by a number from 1 to 8 whereas *Coffea canephora* PPOs were named PPO.CCA followed by a number from 1 to 3. *Coffea*'s PPOs mean length is 566 amino acids ranging from 423 (PPO.CCA1) to 584 (PPO.CAR6). Their mean molecular weight is 63,336 g mol^{-1} ranging from 48,056 g mol^{-1} to 65,183 g mol^{-1} and mean isoelectric point (pI) of 6.28 ranging from 5.55 (PPO.CAR6) to 6.85 (PPO.CAR8).

Regarding DDCs, we found six genomic loci in the *Coffea arabica* genome being four encoded by the *C. eugenioides* sub-genome and two by the *C. canephora* sub-genome. In addition, we found three DDCs in the genome of *Coffea canephora* (Supplemental Table 1). All the DDCs identified in the *Coffea arabica* were named DDC.CAR followed by a number from 1 to 6 whereas *Coffea canephora* DDCs were called DDC.CCA followed by a number from 1 to 3. *Coffea*'s DDCs mean length is 508 amino acids ranging from 480 (DDC.CAR6 and DDC.CCA3) to 537 (DDC.CAR2 and DDC.CCA2). Their mean molecular weight is 56,463 g mol^{-1} ranging from 53,376 g mol^{-1} to 59,696 g mol^{-1} and mean isoelectric point (pI) of 6.08 ranging from 5.77 (DDC.CAR6 and DDC.CA3) to 6.28 (DDC.CAR4).

The phylogeny inference for PPOs recapitulates the evolutionary pattern of angiosperms reported in the Angiosperm Phylogeny Website [73] as shown in Figure 2A. The three main clusters are representing the groups Asterids (containing the orders Solanales, Lamiales and Gentianales), Rosids (containing the orders Rosales, Fabales, Fagales, Brassicales and Malvales) and Monocots (containing the orders Arecales, Asparagales, Poales and Zingiberales). Mostly all *Coffea* PPOs were clustered within the Asterid group while a single PPO from *Coffea canephora* (PPO.CCA3) and two PPOs from *Coffea arabica* (PPO.CAR7 and PPO.CR8) were placed together with the rosales *Trema orientale* and *Morus notabilis* (Figure 2A).

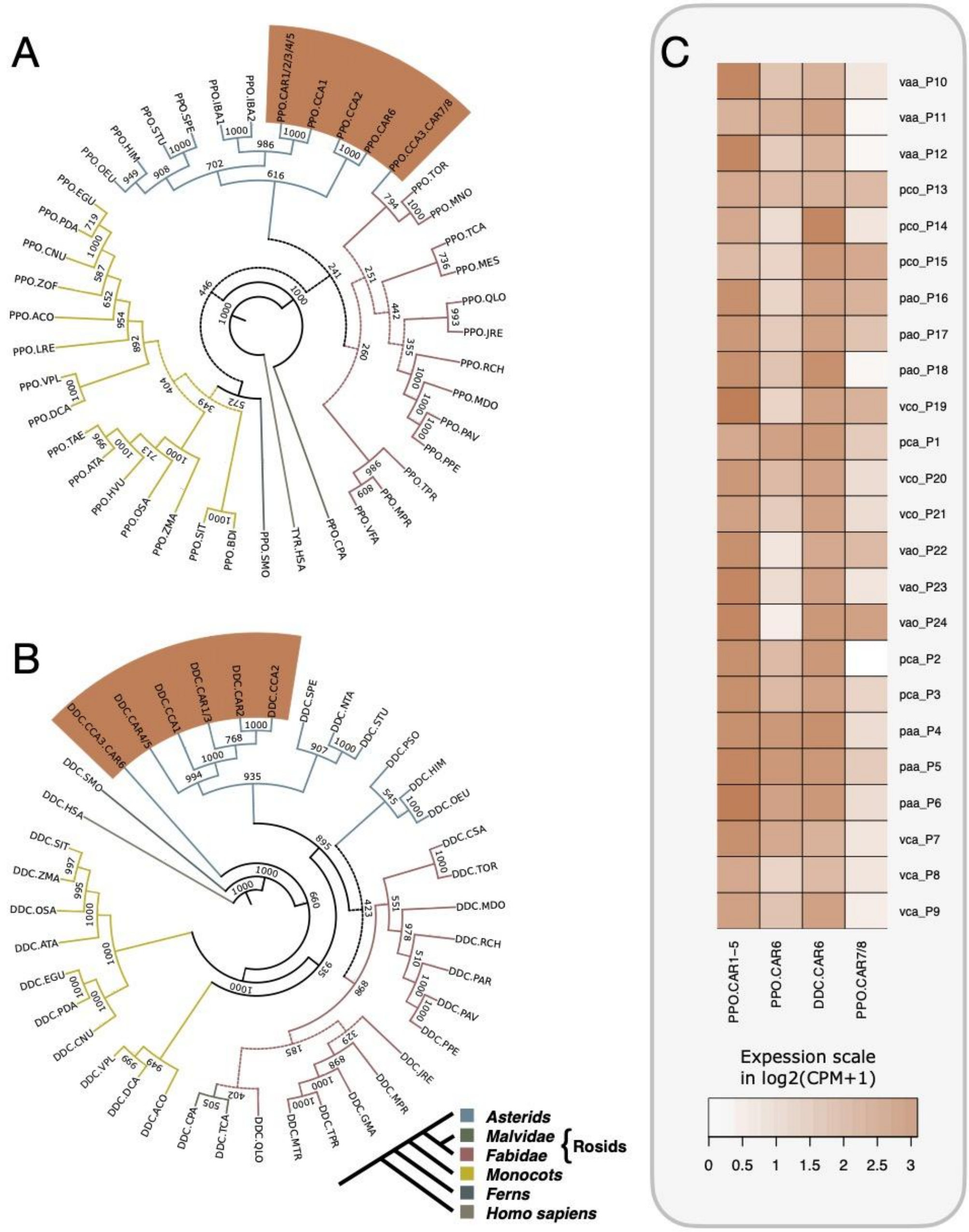

Figure 2 Consensus tree using the Neighbor-Joining clustering method depicting the evolutionary relationships of *Coffea arabica* (CAR) and *Coffea canephora* (CCA) polyphenol oxidases (PPOs) and DOPA decarboxylases (DDCs) as well as the Expression profile of their respective genes in *C. arabica* leaves. **(A)** phylogenetic tree of PPOs, five PPOs in *C. arabica* (PPO.CAR 1 to 5) with identity above 98% were collapsed during phylogeny inference. PPO.CAR1/2/3/4/5 were clustered together with other PPOs from *C. arabica, C. canephora* and other members of the Asterid group. One PPO from *C. canephora* (PPO.CCA3) and two PPOs from *C. arabica* (PPO.CAR7 and PPO.CAR8) were clustered into the Rosid group suggesting that they are under functional diversification. **(B)** phylogenetic tree of DDCs, five DDCs from *C. arabica* (DDC.CAR 1 to 5) clustered with two from *C. canephora* (DDC.CCA1 and 2) as well as other members of Asterids. However, two highly similar DDCs, being one from *C. canephora* (DDC.CCA3) and other from *C. arabica* (DDC.CAR6), were clustered outside of any other flowering plant group suggesting a diversification of this gene in *Coffea. (***C)** Heatmap representation of the expressed *PPOs* and *DDCs* using RNAseq

data from *Coffea arabica* fully expanded leaves in a field experiment with two cultivars (Acauã or Catuaí), two harvest times (April or October) grown in farms of two Brazilian cities (Pirapora or Varginha). Expression values are normalized in Counts Per Million (CPM) and represented in $\log_2$(CPM+1) scale. Each line in the heatmap represents a sequenced library from eight biological samples; Varginha, Catuaí, October (vco); Varginha, Catuaí, April (vca); Varginha, Acauã, October (vao); Varginha, Acauã, April (vaa); Pirapora, Catuaí, October (pco); Pirapora, Catuaí, April (pca); Pirapora, Acauã, October (pao) and Pirapora, Acua, April (paa). Each BioSample consists of three biological replicates. The *PPO.CAR1-5* (representing five highly similar loci with identity above 98%) is constitutively expressed in all analyzed samples while PPO.CAR6 and the divergent PPO.CAR7/8 are less expressed. Regarding *DDCs*, only *DDC.CAR6* was found to be expressed in *C. arabica* leaves. In both phylogenetic trees (**A** and **B**) node numbers correspond to the sum of occurrences of pairs of groups or individual sequences that clustered together in a total of 1,000 bootstraps; dashed lines represent nodes in which group pairs are clustered together in less than 500 (50%) of the bootstraps. The selected species and their respective codes are *Ananas comosus* (ACO), *Aegilops tauschii* (ATA), *Brachypodium distachyon* (BDI), *Coffea arabica* (CAR), *Coffea canephora* (CCA), *Cocos nucifera* (CNU), *Carica papaya* (CPA), *Cannabis sativa* (CSA), *Dendrobium catenatum* (DCA), *Elaeis guineensis* (EGU), *Glycine max* (GMA), *Handroanthus impetiginosus* (HIM), *Homo sapiens* (HSA), *Hordeum vulgare* (HVU), *Ipomoea batatas* (IBA), *Juglans regia* (JRE), *Lilium regale* (LRE), *Malus domestica* (MDO), *Manihot esculenta* (MES), *Morus notabilis* (MNO), *Mucuna pruriens* (MPR), *Medicago truncatula* (MTR), *Nicotiana tabacum* (NTA), *Olea europaea* (OEU), *Oryza sativa* (OSA), *Prunus armeniaca* (PAR), *Prunus avium* (PAV), *Phoenix dactylifera* (PDA), *Prunus persica* (PPE), *Papaver somniferum* (PSO), *Quercus lobata* (QLO), *Rosa chinensis* (RCH), *Setaria italica* (SIT), *Selaginella moellendorffii* (SMO), *Solanum pennellii* (SPE), *Solanum tuberosum* (STU), *Triticum aestivum* (TAE), *Theobroma cacao* (TCA), *Trema orientale* (TOR), *Trifolium pratense* (TPR), *Vicia faba* (VFA), *Vanilla planifolia* (VPL), *Zea mays* (ZMA) and *Zingiber officinale* (ZOF). GenBank or UniProtKB/Swiss-Prot IDs are available in Supplemental Table 2 (PPOs) and Supplemental Table 3 (DDCs). Small phylogenetic tree at the bottom is based on the Angiosperm Phylogeny Website [53] and colors represents specific phylogenetic groups, homologue *Homo sapiens* proteins were add as the out-group.

Seven DDC sequences were clustered within the Asterid group being five from *Coffea arabica* and two from *Coffea canephora* (Figure 2B). A cluster of containing two DDCs, one from *Coffea canephora* (DDC.CCA3) and other from *Coffea arabica- subgenome C. canephora* - (DDC.CAR6) with more than 99% of identity were placed outside of any of the main phylogenetic clusters suggesting that these sequences are under evolutionary divergence after a gene duplication event that happened prior to the origin of the *Coffea arabica*.

Our expression analysis based on the identified *PPOs* and *DDCs* in the *Coffea arabica* genome suggests that all *PPOs* described here are expressed in fully expanded leaves of adult plants (Figure 2C). The PPO.CAR1/2/3/4/5 (represented as PPO.CAR1-5 in the heatmap) are a group of highly similar PPOs with more than 98% of sequence identity at the protein level. In addition, these PPO.CAR1-5 display higher expression levels in *Coffea arabica* leaves when compared to the more distant CAR.PPO6 and the two PPOs clustered within the rosid group CAR.PPO7/8. Regarding *DDCs*, only *DDC.CAR6* was found to be expressed in *C. arabica* leaves.

2.3.2 Chromatographic analyses shows that L-DOPA is a naturally occurring metabolite in *Coffea* leaves

The mean retention time for the L-DOPA standard solution was 3.51 ± 0.41 minutes (Supplemental Figure 3A). The selectivity was accessed by adding the standard solution to samples without L-DOPA. The fortified solutions were prepared by adding L-DOPA in two concentrations (50 and 100 µg mL^{-1}). Then the fortified samples were compared to control samples without L-DOPA in the HPLC runs. Doing so we were able to observe that the separation had no interference from the matrix in the identification of L-DOPA (Supplemental Figure 3B). The correlation coefficient (R^2), detection limit (DL), quantification limit (QL), precision (CV) and accuracy (recovery) are presented in Table 1.

Parameter	**L-DOPA**
B	1694.5
A	177.4
R^2	0.99998
DL (µg mL^{-1})	0.81
QL (µg mL^{-1})	2.73
Recovery (%)	81 to 104
CV (%)	0.38 to 1.11

Table 1 - Analytical parameter for the method standardization

The measured R-squared (R^2) for L-DOPA in *Coffee* leaves was 0.99998 using the HPLC. This is evidence for a strong linear correlation between L-DOPA concentration and the peak area once R^2 values above 0.99 are widely accepted as indicators of linear relationships between chemical compound concentration and its HPLC signal [67,69]. This can be observed in the calibration curve for quantification of L-DOPA in human plasma using a similar approach which presents an $R^2 > 0.99$ [66].

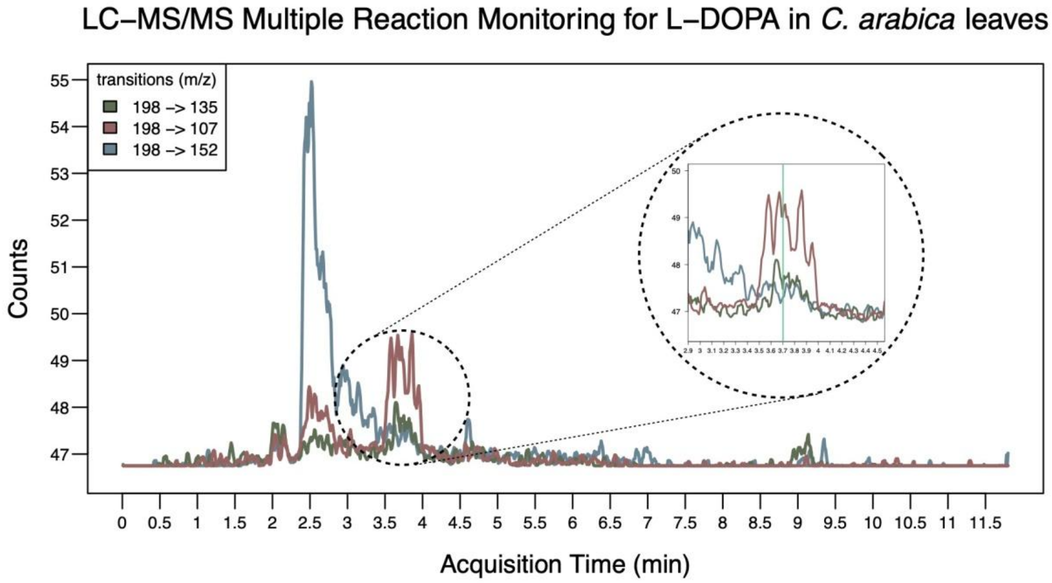

Figure 3 Sequential Mass Spectrometry with Multiple Reaction Monitoring (MRM) profile. In the expected retention time (tR ≈ 3.7 min, highlighted circular section) the presence of the 3 specified transitions is verified, indicating that the L-DOPA molecule in the *Coffea arabica* leaves sample. In addition, the presence of higher intensity interference in the transition 198 > 152 (Blue) is notable, this transition may suggest the decarboxylation of L-DOPA in the carbon C9 resulting in the production of dopamine.

Using LC-MS/MS with Multiple Reaction Monitoring (MRM) we were able to ascertain that the peak around the 3.7 min of acquisition time is indeed L-DOPA that is naturally being synthesized by *Coffea arabica* fully expanded leaves and fruits (Table 2). This is because at this specific time we could identify all the three reported transitions that are characteristic of L-DOPA[70–72]. In addition, L-DOPA was also found in *Coffea canephora* fully expanded leaves (Figure 3 and Supplemental Dataset 1). In both species L-DOPA could not be verified in flowers because the LC-MS/MS signal was indistinguishable from the background noise. For this same reason we could not certify for the occurrence of L-DOPA in *Coffea canephora* fruits.

Species	Tissue	SampleID	Counts	Signal Above Noise Level

Coffea arabica	Leaves	1	48	Yes
		2	48	Yes
		3	48	Yes
	Flowers	4	47	No
		5	47	No
		6	47	No
	Fruits	7	48.5	Yes
		8	49	Yes
		9	49.5	Yes
Coffea canephora	Leaves	10	48.5	Yes
		11	48	Yes
		12	49	Yes
	Flowers	13	48	No
		14	48	No
		15	48	No
	Fruits	16	47	No
		17	48	No
		18	48	No

Table 2 - LC-MS/MS analysis results from samples of leaves, flowers and fruits extracted from *Coffea arabica* and *Coffea canephora*. Samples with readings above background noise have the presence of L-DOPA certified.

2.4 Discussion

Coffee leaves are by-products of the coffee culture that, in comparison to coffee beans, are mostly disregarded in studies concerning their chemical constituents [6,74,75]. However, the studies available show that they are rich in echinoids, flavonoids, xanthones, caffeine and its tea has the potential of being an excellent functional beverage [10] traditionally consumed for over 200 years by locals in the countries where coffee plants are grown [74]. To further investigate the potential compounds in fully expanded coffee leaves, we applied a series of *in silico* analysis to infer metabolic pathways that may be active. Then, we select a portion of the isoquinoline alkaloid biosynthesis pathway according to the KEGG [45] representation (map00950) which involves the enzymes PPO (EC 1.14.18.1, EC 1.10.3.1) and DDCs (EC 4.1.1.28 - Figure 1). Finally, we investigated the occurrence of the L-DOPA compound using HPLC and LC-MS/MS techniques.

2.4.1 Multiple *PPO* copies are present in *Coffea arabica* and *Coffea canephora* genomes

Multiple sequence alignment of PPOs showed that all coffee PPOs identified in this work are similar regarding their primary structure and presented the expected conserved domains (PF00264, PF12142 and PF12143) in the same order and positions in comparison to other functional plant PPOs. In addition, 10 out 11 coffee PPOs possessed a chloroplast transport signal peptide between amino acids 1 to 40 (Supplemental Figure 4A). It is well documented that many plant PPOs have this N-terminal domain containing a thylakoid transfer signal peptide to allow translocation through the chloroplast [76].

Only a PPO from *Coffea canephora*, named PPO.CCA1, did not present this signal peptide. It is not clear for us if this region is indeed absent in this *locus* - meaning that this specific PPO occurs outside the chloroplast - or if the predicted transcription start site is, in fact, upstream of the actual reported coordinates. Functional non chloroplastic plant PPOs were verified for poplar [77], snapdragon [78] and are predicted to occur in monocots, such as rice and maize, and eudicots such as columbine [17]. The discovery of non-plastidic PPOs can help in the discovery of additional roles for PPOs in plants [17].

The PPO C-terminal domain PF12143 (which PFAM description refers to as “unknown function domain; DUF_B2219”) is well conserved in all coffee PPOs. This domain is sometimes regarded to be a blocking-device for the active site via a placeholder residue. But this gatekeeping system is not functional in many plants, being *Malus domestica* and *S.*

lycopersicum[29] exceptions. On the other hand, the C-terminal domain is lacking in other species such as *Vitis vinifera*[79] in which it is believed that the C-terminal domain is a functional copper transporter system that works prior to a proteolytic cleavage [80].

The number of *PPO* in plant genomes vary due to lineage-specific duplications, expansion or loss [17]. *Arabidopsis,* a Malvidae, does not code *PPOs* whereas *Carica papaya* - also a Malvidae - encodes four *PPOs* [17]. Meanwhile, *Glycine max* and *Populus trichocarpa* genomes encode eleven *PPO*. We found three *PPOs* in the *Coffea canephora* genome that may have originally arisen due to the ancestral whole genome triplication event of eudicots [46,81]. Interestingly, we found eight *PPOs* in the *Coffea arabica* genome being seven of them located in the sub genome derived from the ancestral parent *Coffea canephora* and one from the ancestral parent *Coffea eugenioides.*

It is not clear why there is this difference in the number of PPO in the *Coffea arabica* sub-genomes and also in the genome of now living *Coffea canephora* plants. It may be due to events that occurred before the origin of *C. arabica* such as the ones that may have occurred in the respective lineages of its parent progenitors. This way it is possible that *C. canephora* ancestors retained duplicated PPOs loci due to dosage sensitivity gene balancing [82]. Later on, this trait was kept in *Coffea arabica.*

2.4.2 *PPO* copies are expressed in *Coffea arabica* leaves

RNAseq expression analysis showed that fully expanded *Coffea arabica* leaves are constitutively transcribing *PPOs* (Figure 2C). However, due to the high similarity of PPOs at the coding sequence level, the determination of which *PPO* loci are producing transcripts is not possible with current similarity-based bioinformatic approaches. For example, PPO.CAR1 and PPO.CAR2 are 100% identical throughout their length of 1,734 nucleotides. In addition, PPO.CAR3, 4 and 5, all having the same length of 1,734 nucleotides, are more than 98% similar to PPO.CAR1/2. For that reason, when transcript fragments from NGS are mapped to them, the algorithms cannot discern if all loci are being expressed or it is just a subset of them.

These five highly similar *PPOs* are independent loci in the chromosome 5 originated from the parent ancestral *C. canephora* (Supplemental Dataset 2). They are all single exon genes occurring near to each other suggesting that they emerged recently from a local gene duplication event. The *PPO.CAR6* is also located in the chromosome 5 originated from the parent ancestral *C. canephora,* its protein has a similarity of 96.75 % to a PPO in chromosome 5 of the of now living *C. canephora* plants and is also expressed in fully expanded leaves (Figure 2C). Finally, CAR.PPO7 and CAR.PPO8 are 98.79 % similar to each other but only ~50% similar to any other PPO in *C. arabica.* That may be the reason why they are clustered outside the rosid group in figure 2A. The *PPO.CAR7* locus is in the chromosome 2 originated from the parent ancestral *C. canephora* whereas *PPO.CAR8* is the only loci that is encoded in the sub-genome originated from the ancestral parent *C. eugenioides* and it is also in the chromosome 2 of its sub-genome.

2.4.3 Multiple *DDC* copies are present in *Coffea arabica* and *Coffea canephora* genomes but only one copy is expressed in *Coffea arabica* fully expanded leaves

Multiple sequence alignment of DDCs showed that all coffee DDCs are similar regarding their primary structure and presented the expected Pyridoxal-dependent decarboxylase conserved domain (Pyridoxal_deC - PF00282) in the middle section of these sequences (Supplemental Figure 4B). This domain occupies approximately 60% to 70% of the length of all DDCs evaluated and it's the only conserved domain characteristic of DDCs.

In *Coffea arabica* we found four DDCs encoded by the *C. eugenioides* sub-genome (two in chromosome 11e, one in chromosome 1e and another in chromosome 9e) and two encoded by the *C. canephora* sub-genome (both close to each other in chromosome 9c). Similarly, the genome of now living *C. canephora* preset three DDCs being one in chromosome 1 and two organized *in tandem* in chromosome 9. Interestingly, all but two coffee DDCs, one from *C.arabica* DDC.CAR6 and other from *C. canephora* DDC.CCA3, are single exon genes. These multi exon DDCs are located in chromosome 1 of their respective genomes and were the most divergent coffee DDCs in our phylogenetic analysis (Figure 2B and Supplemental Dataset 2).

The identity of the multi-exonic DCCs, DDC.CCA3 and DDC.CAR6, are 99% throughout their protein length of 480 amino acids. The DDC.CAR6 from *Coffea arabica* is in the sense (+) strand on its genome whereas the DDC.CCA3 from *Coffea canephora* genome is in the antisense (-) strand. In addition, these multi-exon DDCs are shorter and present only ~55% of identity to other DDCs in coffee. The expression of *DDC* in *Coffea arabica* leaves was only verified for *DDC.CAR6* (Figure 2C). It is not clear if this coffee DDC is preferentially catalyzing the conversion of L-tyrosine to Tyramine or L-DOPA to Dopamine and it is also possible that these enzymes are catalyzing different reactions. Some DDCs seem to be sequences exclusively found in coffee because our phylogenetic inferences clustered the DDC CAR6 and CCA3 outside of any higher plant group evaluated in this study (Figure 2B). Finally, the apparently silenced *DDC 1-5* in leaves may be active in other leaves under a different set of environmental conditions.

2.4.4 Chromatography analyses confirmed the presence of L-DOPA in *Coffea* leaves and fruits.

The technical DL and QL for our chromatography analysis were 0.81 and 2.73 µg mL^{-1}, respectively for L-DOPA in coffee extracts. Both values were above the reports for human plasma, DL of 0.025 and QL of 0.1 µg/mL [66]. On the other hand, Pavón-Pérez *et al*. (2019) [83] reported a DL of 0.01 mg L^{-1} and a QL of 0.05 mg L^{-1} using LC-MS/MS on *Vicia faba* extracts. Meanwhile, the reported QL for L-DOPA in rat plasma is 25.0 ng/mL [71] which is higher than our findings. The differences found in these parameters may arise from differences in the studied matrices and in the chromatographic conditions, such as in the equipment and/or methodologies used for the detection and quantification.

Our recovery assays to determine the accuracy of the technique returned average values ranging from 81 to 104% (with CV ranging from 0.38 to 1.11 %). Those values are between the analytical acceptance range of 70 to 120% with ± 20% of precision (CV) [68]. Thus, by the results found in this study regarding the recovery values for the L-DOPA compound, we propose that the applied method showed a satisfactory good recovery. Our findings are in accordance with other works that reported recovery yields of 94 to 117% (relative standard deviation of $\leq$ 5.66) [83], 98 to 106% (CV $\leq$ 15%) [66]. In addition, our recovery was higher than the values reported for the quantitation of L-DOPA in rat plasma by HPLC–UV-Vis in which values ranged from 46.5 to 50.1% (CV $\leq$ 10.3%) [71]. Our methodology presented a CV ranging from 0.38 to 1.11%. In this work, the CV value is below 5%, which is the precision lower limit for compounds found in low concentrations in biological extracts [67–69].

Although the samples of fully expanded leaves from *C. arabica* and *C. canephora* and fruits from *Coffea arabica* presented analytical signals below 50 ua it is possible to confirm the L-DOPA presence in those organs (Figure 3). In addition, the presence of higher intensity interference in the MRM transition 198 > 152 is notable, since the choice of mass transitions in MRM mode allows analysis in qualitative (confirmatory) mode. Finally, this specific transition may suggest the decarboxylation of L-DOPA in C9 resulting in the production of dopamine.

2.5 Conclusion

Our *in silico* analysis showed that *PPOs* and *DDCs* are present in multiple copies in the *Coffea arabica* genome and that some of these genes are expressed in fully expanded leaves. L-DOPA, one of the products of the PPO enzyme, was found to be present in both *C. arabica* and *C. canephora* leaves. This presence of L-DOPA in fully expanded leaves may suggest that this metabolite is a component of a pathway to promote defensive mechanisms against pathogens, which may involve its conversion to dopamine by *DDC*s. If dopamine is also a naturally occurring metabolite in coffee leaves it may be part of a mechanism to alleviate nutrient deficiency-induced stresses [42].

Future works will be required to fully reveal the importance of L-DOPA, PPOs and DCCs in coffee. Younger leaves may present enhanced L-DOPA concentrations once PPO activity was verified to be higher there [33,34]. Additionally, different coffee varieties or wild *Coffea* species may be an enhanced source of L-DOPA. This work advances towards the purpose of using coffee leaves as a source of compounds with nutraceutical importance. Finally, we demonstrate that *in silico* analysis is an effective tool to predict metabolic pathways which intermediate compounds can be verified using *in vitro* approaches such HPLC and related techniques. Using the same *in silico* approach we found other potential pathways that may also guide *in vitro* studies to reveal important metabolites in coffee.

REFERENCES

1. International Coffee Organization - Trade Statistics Tables. https://www.ico.org/trade_statistics.asp.

2. Pham, Y., Reardon-Smith, K., Mushtaq, S. & Cockfield, G. The impact of climate change and variability on coffee production: a systematic review. *Clim. Change* **156**, 609–630 (2019).

3. Lashermes, P. *et al.* Molecular characterisation and origin of the Coffea arabica L. genome. *Mol. Gen. Genet. MGG* **261**, 259–266 (1999).

4. Bertrand, B. *et al.* The greater phenotypic homeostasis of the allopolyploid Coffea arabica improved the transcriptional homeostasis over that of both diploid parents. *Plant Cell Physiol.* **56**, 2035–2051 (2015).

5. Vidal, R. O. *et al.* 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 (2010).

6. Silva, F. L. F. *et al.* The concentration of polyphenolic compounds and trace elements in the Coffea arabica leaves: Potential chemometric pattern recognition of coffee leaf rust resistance. *Food Res. Int.* **134**, 109221 (2020).

7. Monteiro, Â. *et al.* Dietary Antioxidants in Coffee Leaves: Impact of Botanical Origin and Maturity on Chlorogenic Acids and Xanthones. *Antioxidants* **9**, 6 (2019).

8. Ashihara, H., Monteiro, A. M., Gillies, F. M. & Crozier, A. Biosynthesis of Caffeine in Leaves of Coffee. *Plant Physiol.* **111**, 747–753 (1996).

9. Campa, C. *et al.* A survey of mangiferin and hydroxycinnamic acid ester accumulation in coffee (Coffea) leaves: biological implications and uses. *Ann. Bot.* **110**, 595–613 (2012).

10. de Almeida, R. F. *et al.* Nutraceutical compounds: Echinoids, flavonoids, xanthones and caffeine identified and quantitated in the leaves of Coffea arabica trees from three regions of Brazil. *Food Res. Int.* **115**, 493–503 (2019).

11. Molitor, C. *et al.* Latent and active aurone synthase from petals of C. grandiflora: a polyphenol oxidase with unique characteristics. *Planta* **242**, 519–537 (2015).

12. Mayer, A. M. & Harel, E. Polyphenol oxidases in plants. *Phytochemistry* **18**, 193–215 (1979).

13. Goldfeder, M., Kanteev, M., Isaschar-Ovdat, S., Adir, N. & Fishman, A. Determination of tyrosinase substrate-binding modes reveals mechanistic differences between type-3 copper proteins. *Nat. Commun.* **5**, 4505 (2014).

14. Mauracher, S. G. *et al.* High level protein-purification allows the unambiguous

polypeptide determination of latent isoform PPO4 of mushroom tyrosinase. *Phytochemistry* **99**, 14–25 (2014).

15. Kim, H. *et al.* A cold-adapted tyrosinase with an abnormally high monophenolase/diphenolase activity ratio originating from the marine archaeon Candidatus Nitrosopumilus koreensis. *Biotechnol. Lett.* **38**, 1535–1542 (2016).

16. Mayer, A. M. Polyphenol oxidases in plants and fungi: Going places? A review. *Phytochemistry* **67**, 2318–2331 (2006).

17. Tran, L. T., Taylor, J. S. & Constabel, C. P. The polyphenol oxidase gene family in land plants: Lineage-specific duplication and expansion. *BMC Genomics* **13**, 395 (2012).

18. Lai, X., Soler-Lopez, M., Wichers, H. J. & Dijkstra, B. W. *PLOS ONE* **11**, e0161697 (2016).

19. FERNÁNDEZ, E., SANCHEZ-AMAT, A. & SOLANO, F. Location and Catalytic Characteristics of a Multipotent Bacterial Polyphenol Oxidase. *Pigment Cell Res.* **12**, 331–339 (1999).

20. Solano, F. Melanins: Skin Pigments and Much More—Types, Structural Models, Biological Functions, and Formation Routes. *New J. Sci.* **2014**, e498276 (2014).

21. Sánchez-Ferrer, Á., Neptuno Rodríguez-López, J., García-Cánovas, F. & García-Carmona, F. Tyrosinase: a comprehensive review of its mechanism. *Biochim. Biophys. Acta BBA - Protein Struct. Mol. Enzymol.* **1247**, 1–11 (1995).

22. Körner, A. & Pawelek, J. Mammalian tyrosinase catalyzes three reactions in the biosynthesis of melanin. *Science* **217**, 1163–1165 (1982).

23. Thipyapong, P., Hunt, M. D. & Steffens, J. C. Antisense downregulation of polyphenol oxidase results in enhanced disease susceptibility. *Planta* **220**, 105–117 (2004).

24. Queiroz, C., Mendes Lopes, M. L., Fialho, E. & Valente-Mesquita, V. L. Polyphenol Oxidase: Characteristics and Mechanisms of Browning Control. *Food Rev. Int.* **24**, 361–375 (2008).

25. Mayer, A. & Harel, E. Phenoloxidases and their significance in fruit and vegetables. *Food Enzymol.* **1**, 373–398 (1991).

26. Martinez, M. V. & Whitaker, J. R. The biochemistry and control of enzymatic browning. *Trends Food Sci. Technol.* **6**, 195–200 (1995).

27. Li, L. & Steffens, J. C. Overexpression of polyphenol oxidase in transgenic tomato plants results in enhanced bacterial disease resistance. *Planta* **215**, 239–247 (2002).

28. Li, Y., Zafar, A., Kilmartin, P. A., Reynisson, J. & Leung, I. K. H. Development and Application of an NMR-Based Assay for Polyphenol Oxidases. *ChemistrySelect* **2**, 10435–10441 (2017).

29. Kampatsikas, I., Bijelic, A. & Rompel, A. Biochemical and structural characterization of tomato polyphenol oxidases provide novel insights into their substrate specificity. *Sci. Rep.* **9**, 4022 (2019).

30. Gasparetti, C. *et al.* Discovery of a new tyrosinase-like enzyme family lacking a C-terminally processed domain: production and characterization of an Aspergillus oryzae catechol oxidase. *Appl. Microbiol. Biotechnol.* **86**, 213–226 (2010).

31. Hakulinen, N., Gasparetti, C., Kaljunen, H., Kruus, K. & Rouvinen, J. The crystal structure of an extracellular catechol oxidase from the ascomycete fungus Aspergillus oryzae. *JBIC J. Biol. Inorg. Chem.* **18**, 917–929 (2013).

32. McLarin, M.-A. & Leung, I. K. H. Substrate specificity of polyphenol oxidase. *Crit. Rev. Biochem. Mol. Biol.* **55**, 274–308 (2020).

33. Mazzafera, P. & Robinson, S. P. Characterization of polyphenol oxidase in coffee. 12 (2000).

34. MONDOLOT, L. *et al.* Evolution in Caffeoylquinic Acid Content and Histolocalization During Coffea canephora Leaf Development. *Ann. Bot.* **98**, 33–40 (2006).

35. Ovallath, S. & Sulthana, B. Levodopa: History and Therapeutic Applications. *Ann. Indian Acad. Neurol.* **20**, 185–189 (2017).

36. Guigoni, C. *et al.* Involvement of Sensorimotor, Limbic, and Associative Basal Ganglia Domains in L-3,4-Dihydroxyphenylalanine-Induced Dyskinesia. *J. Neurosci.* **25**, 2102–2107 (2005).

37. Höglinger, G. U. *et al.* Dopamine depletion impairs precursor cell proliferation in Parkinson disease. *Nat. Neurosci.* **7**, 726–735 (2004).

38. Hardebo, J. E. & Owman, C. Barrier mechanisms for neurotransmitter monoamines and their precursors at the blood-brain interface. *Ann. Neurol.* **8**, 1–11 (1980).

39. Wang, Y. *et al.* Overexpression of the tyrosine decarboxylase gene MdTyDC confers salt tolerance in apple. *Environ. Exp. Bot.* **180**, 104244 (2020).

40. Li, C. *et al.* Dopamine alleviates salt-induced stress in Malus hupehensis. *Physiol. Plant.* **153**, 584–602 (2015).

41. Liang, B. *et al.* Dopamine alleviates nutrient deficiency-induced stress in Malus hupehensis. *Plant Physiol. Biochem.* **119**, 346–359 (2017).

42. Lan, G., Jiao, C., Wang, G., Sun, Y. & Sun, Y. Effects of dopamine on growth, carbon metabolism, and nitrogen metabolism in cucumber under nitrate stress. *Sci. Hortic.* **260**, 108790 (2020).

43. Monente, C., Ludwig, I. A., Irigoyen, A., De Peña, M.-P. & Cid, C. Assessment of Total (Free and Bound) Phenolic Compounds in Spent Coffee Extracts. *J. Agric. Food Chem.* **63**, 4327–4334 (2015).

44. Farah, A. & Donangelo, C. M. Phenolic compounds in coffee. *Braz. J. Plant Physiol.* **18**, 23–36 (2006).

45. Aoki, K. F. & Kanehisa, M. Using the KEGG database resource. *Curr. Protoc. Bioinforma.* **11**, 1–12 (2005).

46. Dereeper, A. *et al.* The coffee genome hub: a resource for coffee genomes. *Nucleic Acids Res.* **43**, D1028–D1035 (2015).

47. Götz, S. *et al.* High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res.* **36**, 3420–3435 (2008).

48. Johns Hopkins University. Coffea arabica V. Caturra Genome. (2018).

49. Hoff, K. J. & Stanke, M. Predicting Genes in Single Genomes with AUGUSTUS. *Curr. Protoc. Bioinforma.* **65**, e57 (2019).

50. Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).

51. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl Acids Res* **25**, (1997).

52. Klabunde, T., Eicken, C., Sacchettini, J. C. & Krebs, B. Crystal structure of a plant catechol oxidase containing a dicopper center. *Nat. Struct. Biol.* **5**, 1084–1090 (1998).

53. Katoh, K. & Toh, H. Recent developments in the MAFFT multiple sequence alignment program. *Brief. Bioinform.* **9**, 286–298 (2008).

54. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).

55. Felsenstein, J. *PHYLIP (phylogeny inference package), version 3.5 c*. (Joseph Felsenstein., 1993).

56. Jones, D. T., Taylor, W. R. & Thornton, J. M. The rapid generation of mutation data matrices from protein sequences. *Bioinformatics* **8**, 275–282 (1992).

57. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425 (1987).

58. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296 (2021).

59. Sperschneider, J. *et al.* LOCALIZER: subcellular localization prediction of both plant and effector proteins in the plant cell. *Sci. Rep.* **7**, 44598 (2017).

60. Cardon, C. H. *et al.* Expression of coffee florigen CaFT1 reveals a sustained floral

induction window associated with asynchronous flowering in tropical perennials. *Plant Sci.* **325**, 111479 (2022).

61. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

62. Anders, S., Pyl, P. T. & Huber, W. HTSeq - A Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).

63. Chen, Y. *et al.* edgeR: differential analysis of sequence read count data User's Guide. *R Package* 1–121 (2020).

64. Warnes, M. G. R., Bolker, B., Bonebakker, L., Gentleman, R. & Huber, W. Package 'gplots'. *Var. R Program. Tools Plotting Data* (2016).

65. Polanowska, K., Łukasik, R., Kuligowski, M. & Nowak, J. Development of a Sustainable, Simple, and Robust Method for Efficient l-DOPA Extraction. *Molecules* **24**, 2325 (2019).

66. Elbarbry, F., Nguyen, V., Mirka, A., Zwickey, H. & Rosenbaum, R. A new validated HPLC method for the determination of levodopa: Application to study the impact of ketogenic diet on the pharmacokinetics of levodopa in Parkinson's participants. *Biomed. Chromatogr.* **33**, e4382 (2019).

67. Snyder, L. R., Kirkland, J. J. & Dolan, J. W. *Introduction to modern liquid chromatography*. (John Wiley & Sons, 2011).

68. Ribani, M., Bottoli, C. B. G., Collins, C. H., Jardim, I. C. S. F. & Melo, L. F. C. Validação em métodos cromatográficos e eletroforéticos. *Quím. Nova* **27**, 771–780 (2004).

69. Harris, D. C. Análise Química Quantitativa. 7a edição. *Rio Jan. LTC* (2008).

70. César, I. C. *et al.* Development and validation of a high-performance liquid chromatography–electrospray ionization–MS/MS method for the simultaneous quantitation of levodopa and carbidopa in human plasma. *J. Mass Spectrom.* **46**, 943–948 (2011).

71. Chi, J., Ling, Y., Jenkins, R. & Li, F. Quantitation of levodopa and carbidopa in rat plasma by LC–MS/MS: The key role of ion-pairing reversed-phase chromatography. *J. Chromatogr. B* **1054**, 1–9 (2017).

72. Yang, G. *et al.* Development and validation of an LC-MS/MS method for simultaneous quantification of levodopa and MD01 in rat plasma and its application to a pharmacokinetic study of mucuna pruriens extract. *Biomed. Chromatogr.* **30**, 1506–1514 (2016).

73. Stevens, P. F. Angiosperm Phylogeny Website. Version 13. *Angiosperm Phylogeny Website Version 13* (2016).

74. Patay, É. B., Bencsik, T. & Papp, N. Phytochemical overview and medicinal importance of Coffea species from the past until now. *Asian Pac. J. Trop. Med.* **9**, 1127–

1135 (2016).

75. Chen, X.-M., Ma, Z. & Kitts, D. D. Effects of processing method and age of leaves on phytochemical profiles and bioactivity of coffee leaves. *Food Chem.* **249**, 143–153 (2018).

76. Kampatsikas, I., Bijelic, A., Pretzler, M. & Rompel, A. A Peptide-Induced Self-Cleavage Reaction Initiates the Activation of Tyrosinase. *Angew. Chem. Int. Ed.* **58**, 7475–7479 (2019).

77. Tran, L. T. & Constabel, C. P. The polyphenol oxidase gene family in poplar: phylogeny, differential expression and identification of a novel, vacuolar isoform. *Planta* **234**, 799–813 (2011).

78. Ono, E. *et al.* Localization of a flavonoid biosynthetic polyphenol oxidase in vacuoles. *Plant J.* **45**, 133–143 (2006).

79. Virador, V. M. *et al.* Cloning, sequencing, purification, and crystal structure of Grenache (Vitis vinifera) polyphenol oxidase. *J. Agric. Food Chem.* **58**, 1189–1201 (2010).

80. Kanteev, M., Goldfeder, M. & Fishman, A. Structure-function correlations in tyrosinases. *Protein Sci. Publ. Protein Soc.* **24**, 1360–1369 (2015).

81. Jiao, Y. *et al.* A genome triplication associated with early diversification of the core eudicots. *Genome Biol.* **13**, 1–14 (2012).

82. Freeling, M. & Thomas, B. C. Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome Res.* **16**, 805–814 (2006).

83. Pavón-Pérez, J., Oviedo, C. A., Elso-Freudenberg, M., Henríquez-Aedo, K. & Aranda, M. LC-MS/MS METHOD FOR L-DOPA QUANTIFICATION IN DIFFERENT TISSUES OF VICIA FABA. *J. Chil. Chem. Soc.* **64**, 4651–4653 (2019)

CHAPTER 3

3. ARTICLE 3 - THE FLORAL DEVELOPMENT OF THE ALLOTETRAPLOID *Coffea arabica* L. IS CORRELATED WITH A DYNAMIC REPROGRAMMING OF SMALL RNAS

If a link does not work, please copy and paste into the browser and remove eventual spaces

SUPPLEMENTAL FIGURES can be accessed using the link:

https://drive.google.com/drive/folders/1x5mWnN44nS_YnGMPe_zsSQOdRojqDIGA?usp=share_link

SUPPLEMENTAL TABLES can be accessed using the link:

https://drive.google.com/drive/folders/1V-jcbyoHTbPScXzHMGrjelOdtFyUp5JQ?usp=share_link

SUPPLEMENTAL DATASETS can be accessed using the link:

https://drive.google.com/drive/folders/1njWvI2m27afqDx4Yn0svifwt1Q3JTx6d?usp=share_link

SCRIPTS can be accessed using the link:

https://github.com/thalescherubino/thesisChapter3

(Draft Version)

Article prepared for submission to the New Phytologist Journal

The floral development of the allotetraploid *Coffea arabica L* is correlated with a dynamic reprogramming of small RNAs

Thales Henrique Cherubino Ribeiro, Patricia Baldrich, Raphael Ricon de Oliveira, Christiane Noronha Fernandes-Brum, Sandra Mathioni, Thaís Cunha de Sousa Cardoso, Matheus de Souza Gomes, Laurence Rodrigues do Amaral, Kellen Kauanne Pimenta de Oliveira, Gabriel Lasmar dos Reis, Blake C. Meyers, Antonio Chalfun-Junior

Abstract

Non-coding and coding RNAs are key regulators of plant growth, development, and stress responses. To investigate the types of transcripts accumulated during the vegetative to reproductive transition in the *Coffea arabica* L., we sequenced small RNA libraries from eight developmental stages up to anthesis. This data was combined with messenger RNA and degradome sequencing of two important development stages that marks the transition of an apparent latent to a rapid growth stage. In addition, we took advantage of multiple *in silico* tools to characterize genomic loci producing small RNAs such as phasiRNAs, miRNAs and tRFs. Our differential and co-expression analysis showed that some types of small RNAs such as tRNAs, snoRNAs, snRNAs and phasiRNAs preferentially accumulate in a stage-specific manner whereas miRNAs accumulate in a family-stage specific manner, related to modulated hormonal responses and transcription factor expression. The accumulation of 24-nt phasiRNAs in a latent stage suggests a stabilizing functionality to synchronize flowering and/or an epigenetic imprint mechanism.

3.1 Introduction

Ribonucleic acids (RNA) are involved in all stages of a plant life cycle. They can be compared to computer-code subroutines loaded to the Random Access Memory (RAM) in every smart-phone or similar device that runs on an software-based operating system. Our analogy of RNAs and computer codes can only go thus far because the plant transcriptome is not only software, but it is also hardware. Although it is human nature to try to categorize things in different classes, RNAs often blur definition borders. While a computer code is immaterial, the RNAs are code, matter and (enzymatic)action. For example, a transfer RNA (tRNA) can fulfill its classical role as an amino acid carrier [1], can work as an interference signal to quantitatively counteract other molecular program [2], regulate genome stability by targeting Transposable Elements (TE) transcripts [3] and be transported to other life kingdom to negotiate the establishment of a symbiotic relationship [4].

Biological codes expressed in the form of RNAs are not like human-made computer codes. There is also an additional level of complexity. While computer codes often come from a hard-disc (even when they are running from the "cloud") a plant RNAs script/hardware can rise from multiple versions of a genome co-existing inside a cell. For example, *Coffea arabica* L.*,* the source of most of the coffee beverages consumed world-wide, has two copies of a diploid genome [5]. Each version of its genome comes from one ancestral progenitor from an intraspecific cross between two ancient specimens, *Coffea eugenioides* Moore and *Coffea canephora* Pierre [6]. The *C. arabica* allopolyploidy (2 n = 4 x = 44) can benefit an organism by improving transcriptional homeostasis over its diploid parents [7]. Nevertheless, allopolyploidy can also add more complexity during important processes such as the meiotic cell division [8].

Like any other plants, *Coffea arabica* has a complex operational system that runs in its cells. This system is based upon chemical reactions between diverse molecules such as chromosomes (DNA), enzymes (proteins or RNAs), chemical gradients (osmotic potentials separated by membranes), hormones and many others. This interaction among diverse endogenous players underlies complex regulatory networks that will respond to a chaotic environment producing a seemingly predictable phenotype [9]. For example, healthy *Coffea arabica* plants will mostly unquestionably produce flowers and then seeds that will have a chance to start a new cycle in the near future. However, more detailed outcomes such as the number of flowers, fruits and the ratio of green and red fruits is yet impossible to predict in any given moment [10].

This non-uniformity in *Coffea arabica* flowering and ripening process is a serious economic problem that affects its production, even more so in the actual climate change progression [11,12]. This lack of uniformity is in part due to the long phenological cycle of coffee that is biennial [13]. In other words, it takes two years from the transition of vegetative buds to flowers, fruits, and the latter senescence of fruitful branches [14]. In addition, as a perennial plant, it must keep the shoot apical meristem in a vegetative state to allow simultaneous - but spatially separated - growth and reproductive development [13]. Although some endogenous and environmental factors are known to be involved in the vegetative to reproductive transition of coffee [13], the role of small RNAs - in special non-coding RNAs (ncRNAs) - is a topic often overlooked. To address this lack of knowledge we performed a broad overview of what RNA programs (transcripts) are being loaded across stages of the flower differentiation process in *Coffea arabica*.

Here we evaluated messenger RNAs (mRNAs), micro interfering RNAs (miRNAs), phased small interfering RNAs (phasiRNAs), small nuclear RNAs (snRNAs), small nucleolar RNAs (snoRNAs) and transfer RNAs (tRNAs). Doing so, we found that different types of small RNAs are preferentially accumulated in a stage-specific manner. In addition, our analysis also rendered important insights on how these RNAs interact with other pieces of the molecular machinery and regulatory networks such as hormone crosstalk and defense response systems.

Finally, we identified changes in the accumulation of sRNAs during two contrasting development stages that are usually classified into a single stage by anatomical characteristics [15]. Buds between 3 mm and 6 mm are typically referred to as G4 and considered dormant or latent. Although this stage may, anatomically, look latent for months, our transcription data shows that 21 and 24 nt phasiRNAs are being accumulated. After environmental stimuli, such as a long period of drought followed by rain, the levels of phasiRNAs are sharply reduced whereas other sRNAs such snoRNAs and tRNAs are drastically, but briefly, increased. This finding leaded us to propose novel transcriptomic-based classification that adapts the Morais *et al.* [15] phenology classification to discriminate buds > 3 mm and < 6 mm into two new stages: an early stage now called S3 -transcriptionally characterized by the accumulation of 24-nt phasiRNAs - and a late stage called S4 (characterized by the fast accumulation of tRNAs, snoRNAs and the resuming of developmental programs.

3.2 Material and methods

3.2.1 Plant material

To obtain data for RNAseq, small RNAseq (sRNAseq) and Parallel Analysis of RNA Ends (PARE), we selected 5-year-old *Coffea arabica* plants of two cultivars; "Siriema VC4" and "Catuaí Vermelho IAC 144". Each biological replicate had material pooled from two individuals. These plants were grown at the experimental field in the Federal University of Lavras (UFLA), Brazil (21°13' S, 44°58' W) and were maintained with standard cultivation practices. After harvesting, all samples were immediately frozen in liquid nitrogen and stored at -80 ºC until total RNA extraction. Sample descriptions are available in the Supplemental

Table 1 and the sampled contrasting developmental stages were collected and categorized based on an adaptation of the phenological characterization proposed by Morais *et al.* [15].

3.2.2 Library preparation and Illumina sequencing

Total RNA was isolated with PureLink® Plant RNA Reagent (Invitrogen). Then, for the sRNAseq libraries, we performed the size selection using denaturing Urea-PAGE gels and library construction using the TruSeq Small RNA Library Preparation kit following the protocol described by Mathioni *et al.*, 2017 [16].

For the RNAseq of two development stages (S4 and S5) the libraries were prepared with Illumina TruSeq stranded RNAseq preparation kit. The sampling for the RNAseq libraries was the following; 2 biological replicates x 2 development stages (S4 and S5) x 2 cultivars, rendering 8 strand-specific single-end sequenced libraries with a total of about 376 million reads. Finally, PARE libraries for S4, S5, pre-meiotic, meiotic and pos-meiotic samples were constructed using the protocol described by German *et al.*, 2009 [17]. Sequencing was performed at the University of Delaware Sequencing and Genotyping Center in the Delaware Biotechnology Institute using an Illumina HiSeq 2500 sequencer.

3.2.3 Genome-wide identification of miRNAs

We identified conserved mature and precursor miRNAs in the *Coffea arabica* genome using the procedure described by de Souza Gomes *et al* [18]. This pipeline includes steps such as filtering for GC content from 20 to 65%, Minimum Free Energy below -20 kcal/mol and selection of candidate mature sequence with identity greater than 85% to mature sequences of plant miRNAs registered in the miRbase Release 22.1 [19]. Then, we enriched this set of conserved miRNAs with additional novel, and conserved, miRNAs predicted using the miRador tool [20]. miRador utilizes the most up-to-date criteria to accurately identify plant miRNAs taking advantage of sRNA-seq data [21]. Finally, we developed custom Python scripts to merge miRNAs with identical mature sequences predicted by both methodologies to reduce mapping biases due to multi-mapped reads in different genome loci.

3.2.4 Genome-wide identification of *PHAS* loci

We mapped ~1.2 billion quality-controlled sRNAseq reads from 8 stages ranging from nodes containing undetermined vegetative cells to flowers in addition to pre-meiotic, meiotic and post-meiotic anthers to the *Coffea arabica* L. (var Caturra) genome [22] with Bowtie [23] version 1.3.0. No gaps or mismatches were allowed. Next, we sorted the alignments files and merged them with samtools [24]. The resulting bam file was processed with ShortStack [25] to predict pashing locus. The parameters were --mmap u, --nohp and --mismatches 0. Subsequently, we manually evaluated the predicted loci based on their score, length, overlapping with proteins or TEs, what proteins and TEs, the ratio of alignments between strands, overall phasing pattern and complexity (measured as the number of distinct aligned reads divided by the total abundance of a given locus). Lower complexity values - close to 0 - indicated loci covered by a few dominant small RNAs while higher values - close to 1 - indicate loci with a diverse set of small RNAs . The alignments were manually inspected with the Integrated Genome Viewer (IGV) [26] version 2.12.2. Doing so, we classified the 803 candidate *PHAS* loci as True *PHAS* Locus, Long 24-*PHAS*-like locus (L24P-like) or, if the characteristics of the locus was ambiguous, unknown type (UNK).

3.2.5 Differential accumulation analysis of phasiRNA, tRNAs, snoRNA, and snRNAs

To evaluate the accumulation profile of different types of small RNAs we manually curated a reference fasta file containing the sequences of non redundant mature miRNAs, all 803 *PHAS/PHAS-like* predicted loci and 307 representative sequences of non-redundant tRNAs, 157 snRNAs and 1,203 snoRNAs. Next, we mapped the ~1.2 billion quality controlled sRNAseq reads from nodes containing undetermined vegetative cells (S0) to flower using Bowtie [23] 1.3.0 with parameters -k 1 and -v 0 (no mismatches allowed). We carried out differential analysis with edgeR [27] and a given small RNA producing locus was deemed differentially accumulated between two contrasting development stages if its false discovery rate (FDR) was below 0.05 and fold change ≥ 2.

3.2.6 Co-expression network analysis of small RNAs

To infer co-expression modules (clusters) of sRNAs, we applied procedures from the Weighted Gene Co-expression Network Analysis (WGCNA) R-package [28]. The input for the correlation analysis was the same count data we used in the differential accumulation analysis with some required transformations. First, we normalized the count data in Counts Per

Million (CPM) and then we log2 transformed the data to better fit the assumptions of the WGCNA package. Next, we calculated an adjacency matrix of Pearson correlations between all pairs of transcribed small RNAs and raised it to a power **β** (soft threshold) of 6. The **β** = 6 parameter was based on the scale free topology criterion [29]. After that, to minimize the effect of noise and spurious associations, we transformed the adjacency matrix into a Topological Overlap Matrix (TOM). Next, a dendrogram, with the co-expression modules as its branches, was inferred based on the average dissimilarity of the TOM using the Dynamic Tree Cut method. Finally, we analyzed the individual co-expression modules with the R package igraph [30].

3.2.7 Genome-wide target prediction

To validate the siRNA targets profiled by degradome sequencing (PARE) we used sPARTA [31] with parameters --map2DD, --validate, -minTagLen 18 and -tarScore N. These degradome analyses were performed to identify genic and intergenic cleavage sites of mature miRNAs and genic cleavage sites of tRNAs-Derived Fragments (tRFs). The target cutoff score was set to ≤5 for miRNA targets and, more restrictively, ≤3 for tRFs. In the miRNA target prediction we considered the non-redundant mature sequences described previously.

The tRFs used for this degradome analysis were processed as follows; First, all quality controlled smallSEQ reads were mapped to the tRNA reference available at the NCBI annotation accession GCF_003713225.1 [22]. All fragments mapped without mismatches were then selected based on their length (minimum of 18 and maximum of 24 nt) and expression in CPM (minimum of 1CPM). Doing so, we selected 7,458 unique tRFs that were used by sPARTA to identify putative cleavage sites. We set the sPARTA minTagLen to 18 because about 16% of tRFs had this length and most of them are sub-sequences from longer tRFs.

3.2.8 Differential Expression analysis of protein coding genes

Approximately 376 million single-end RNAseq reads from the stages S4 and S5 were sequenced in eight libraries: four biological replicates for each stage of "Siriema VC4" and "Catuaí Vermelho IAC 144" cultivars. After quality control with trimmomatic [32] v.0.33 (parameters ILLUMINACLIP:./adapters:3:25:6 SLIDINGWINDOW:4:28 MINLEN:30) 272 million reads were mapped to the genome with STAR [33] aligner v.2.7.1. Then, we quantified the reads uniquely mapped to exons in the genome using the htseq-count script [34]. We carried out differential expression analysis with edgeR [27] and a given protein coding gene was deemed differentially expressed (DE) between S4 and S5 stages if its false discovery rate (FDR) was less than 0.05 and fold change ≥ 2. Then, we searched for enriched gene ontology (GO) terms using the online tool agrigo v.2 [35].

3.3. Results

3.3.1 The higher number of miRNAs precursor loci in *Coffea arabica* compared to *Arabidopsis* may reflect a 100 million years of divergence

A total of 557 candidate miRNA precursor loci from 296 miRNA families were found in the *C. arabica* genome (Figure 1 - First inner cycle). Those precursors produce 447 nonredundant mature miRNAs. We discovered that 45 of those precursors are generating mature miRNAs representing putative novel family members (Figure 1 - First inner cycle, violet points). A total of 205 miRNA precursors were found to be encoded in the *C. eugenioides* sub-genome and 263 from the *C. canephora* sub-genome (Figure 1 - First inner cycle, black dots). Additionally, 89 miRNA precursors were found in the unplaced contigs.

Among the conserved miRNA families in the Asterids, *miR160* was only identified in the *Coffea canephora* sub-genome whereas the *miR828* family is only found in the *Coffea eugenioides* sub-genome. Both families were also not identified in the unplaced contigs. The *miR827* family, that is conserved in angiosperms [36], seems to be absent in *Coffea arabica*. In addition, the *miR173* family, which in *Arabidopsis thaliana* is known for triggering the biogenesis of both trans-acting small interfering RNAs (tasiRNAs) TAS1 and TAS2 [37], is not present in the *Coffea arabica* genome. The families having the greatest numbers of loci were *miR482* (20 loci), *miR395* (19), *miR169* (16), *miR167* (14) and *miR171* (13) (Table S2) whereas the most accumulated families were, in decreasing order, *miR166, miR396, miR482, miR319 and miR8155* (Supplemental Figure 1).

The genome of extant *Coffea canephora* plants, possibly from the same species of one of the ancestral progenitors of *Coffea arabica* [6], is also enriched with multiple copies of some *MIR* families; *miR482* (7 loci), *miR395* (9), *miR169* (10), *miR167* (7) and *miR171* (16) [38]. *Arabidopsis thaliana* has no *miR482* [39], which is the most abundant in terms of the number of loci in *Coffea arabica*. However, *Arabidopsis thaliana* still encodes other abundant loci in *Coffea arabica* such as *miR395* (5 loci), *miR169* (14), *miR167* (4) and *miR171* (3) [40]. *Arabidopsis lyrata*, a Rosid believed to have diverged from *Arabidopsis thaliana* about ten million years ago [41], has the same number of loci for families *miR169*, *miR167* and *miR171*, but contains two additional members of *miR395* family (7 loci) when compared to *A. thalina* [40].

The expansion of the number of miRNA loci in *Coffea arabica* can be explained by the *MIR* net rate of flux (birth-death). In the *Arabidopsis* lineage, this rate was estimated to be a net gain of 1.2 to 3.3 *MIR* genes per million years [40]. According to miRbase v. 22 [19] there are 326 miRNA precursors in *Arabidopsis thaliana* genome [19]. It is estimated that *Coffea canephora* diverged from *Vitis vinifera*, a basal rosid, about 114 to 125 million years ago [42,43]. This allows us to infer that *Coffea arabica* and *Arabidopsis* shared a common ancestor at roughly the same time that *Coffea canephora* and *Vitis vinifera* did. Therefore, in more than 100 million years of evolution and the emergence of the allopolyploid *Coffea arabica*, it is possible that hundreds of microRNA genes were born and lost causing this difference of 231 *MIR* precursors between *Coffea arabica* and *Arabidopsis thaliana.*

3.3.2 The majority of 21-*PHAS* loci correspond to disease resistance proteins triggered by *miR482* superfamily

Plant phasiRNAs can be processed from long noncoding RNAs (lncRNAs) or protein coding transcripts that, after a precise cleavage guided by a miRNA, generate secondary siRNAs by the activity of DICER-like enzymes. We used small RNA libraries from vegetative and reproductive organs to identify loci producing phasiRNAs in *Coffea arabica*. We identified 173 21-*PHAS* loci, most of them from disease resistance (R) genes (Figure 1 - second internal cycle, green dots; Supplemental Table 2). We found miRNA triggering at least 51 *21-PHAS*, including 23 triggered by *miR482* family members. In addition, 15 *21-PHAS* loci were found to be triggered by a putative novel *MIR* family, the candidate *miR245889*. Similarity based analysis suggests that this candidate diverged from the *miR482* superfamily (Supplemental Figure 2).

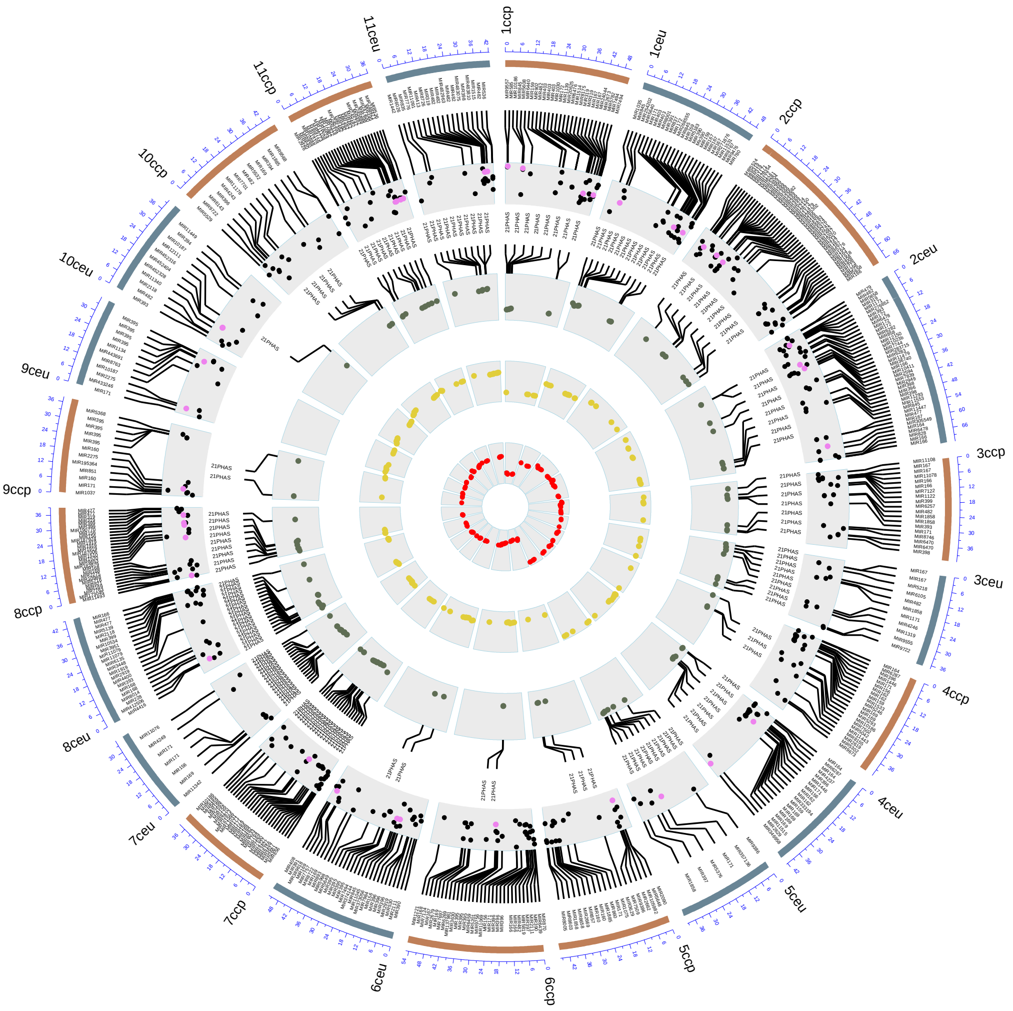

Figure 1 Genome-wide distribution of *MIRs*, *PHAS* and *PHAS*-like loci in the allopolyploid *Coffea arabica.* The outermost cycle represents chromosomes from each parental ancestor; from *C. eugenioides* (ceu) or *C. canephora* (ccp). First inner cycle; Chromosomal coordinates of miRNA precursors, black dots are conserved families while violet dots are putative novel. Second inner cycle; green dots point out the chromosomal coordinates of 21-*PHAS* loci. Third inner cycle; point out the chromosomal coordinates of 24-*PHAS* loci. Innermost cycle; red dots point out the chromosomal coordinates of *24-PHAS*-like loci

Our prediction of loci with phasing patterns allowed us to identify two candidate tasiRNA loci. Their miRNA target sites were similar to *miR390* and *miR828*. Further manual analysis of those loci revealed that they are the respective orthologs of *TAS3* and *TAS4* (Supplemental figure 3). We did not find any orthologs for *TAS1* or *TAS2* in accordance with the lack of its conserved trigger, *miR173*. We also identified two *DICER-like 2* (*DCL2*), one in chromosome 6e and another in 9c, as loci generating phasiRNAs. This phasing of *DCL2* has been described before in Fabaceae [44], and Solanaceae [45]. We were not able to identify a specific trigger of *DCL2* phasing. It is possible that another small RNA is performing this task.

3.3.4 24-*PHAS* loci are not triggered by any expressed miRNA

The presence of 24-nt phasiRNAs was recently reported in many eudicots, and in those reports, these small RNAsare highly enriched in reproductive organs, specifically anthers [46–48]. Because several phased siRNA annotation methods can frequently mistake heterochromatic siRNAs with *24-PHAS* loci [49] we manually evaluated all putative *PHAS* loci with the IGV genome browser. Doing so, we were able to identify 189 24-*PHAS* (Figure 1 - third internal cycle with yellow dots; Supplemental Table 3). Of those loci, 56 overlap with annotated protein coding genes (PCG) that were mostly identified as "uncharacterized proteins" by the similarity based local alignment tool BLASTP [50]. That way, these putative protein-coding loci could be lncRNAs that were misannotated. Meanwhile, other loci seem to be *bonafide* PCGs that, due to their proximity to the 24-*PHAS* loci*,* may be transcribed and processed as such. In addition, 58 of those phasiRNAs overlapped with transposable elements (TEs) such as Gypsy (34%) and hAT (22%).

Unexpectedly, we were unable to find any apparent miRNA triggers of the loci with the 24-nt phasing pattern. *miR2275*, broadly present in eudicots, is the usual trigger of 24-nt phasiRNA biogenesis [47]. However, it is known to be absent in some lineages such Brassicales, Caryophyllales, Cucurbitaceas and Lamiales. In *Coffea arabica,* we found four putative *miR2275* precursor loci in the genome, but without evidence of expression in the analyzed organs. Something similar was previously reported for the Solanales tomato and petunia in which 24-nt reproductive phasiRNAs were found to be abundant in meiotic anthers, nevertheless, no apparent miRNA target site was present [47]. MEME analysis of the 189 *Cofea arabica* 24-*PHAS* loci showed that 181 of those loci contain 24-nt rich-A motifs, with statistical significance [51] (exemplified in the Supplemental Figure 4). However, no expressed miRNA (or any of the evaluated small RNA) has complementarity to those putative target sites once miRNAs tend to be poor in A-U stretches. It is also possible that due to the relatively weak pairing of adenine and thymine, with two hydrogen bonds instead of three from the guanine and cytosine, the double strand of DNA can be easily separated allowing transcription and further processed into double-stranded RNA. Then, further degradation in the precise 24-nt pattern by a yet unknown mechanism can take place.

3.3.5 Hundreds of loci present a 24 nt phasing pattern but could not be included in the 24-*PHAS* category

There were 175 *Long 24-PHAS-like* loci (L24P-like; Figure 1 - Fourth internal cycle with red dots; Supplemental Table 4). We gave these loci this new name for several reasons. Those loci passed our filtering criteria for a phasiRNA locus, and they were mostly derived from non-repetitive regions of the genome. Similar to the 24-*PHAS* loci, they also did not seem to have any obvious triggers. However, during the manual examination at the IGV genome browser, they appeared too long to be considered typical phasiRNA generating loci. Their median length of L24P-like is 10.3 kb whereas the median length of precursors of *Coffea arabica* 24-*PHAS* loci is around 1.8 kb.

Those L24P-like resemble sirenRNAs (small-interfering RNA in endosperm) in terms of their length and stage specificity (Figure 2A, Cluster V in Figure 4C and Supplemental Figure 5). sirenRNAs are 24-nt siRNAs firstly identified and abundant in endosperm [52,53]. They map predominantly to genic and intergenic regions rather than transposable elements - except for edges of *hAT* and *Helitron* TE families [54]. They are derived from approximately 200 loci in *Brassica rapa* but are also present in diverse other angiosperms [52,53]. sirenRNAs are transcribed in maternal organs and are believed to trigger DNA methylation in filial tissues, establishing epigenetic marks in the next generation [53]. The siren loci are usually larger than other types of siRNA producing loci and their derived siRNAs are more likely to map uniquely in the genome when compared to other categories of siRNAs [53]. In addition, they tend to represent more than 90% of the accumulated siRNAs in developing seeds [53]. In accordance with the siRNAs tendency, the number of mapped reads of *Coffea arabica* L24P-like are substantially higher than the 24-*PHAS* (two-tailed t-test unequal variance $p = 2.5E^{-09}$) and are predominantly multi-mapers (Figure 2B).

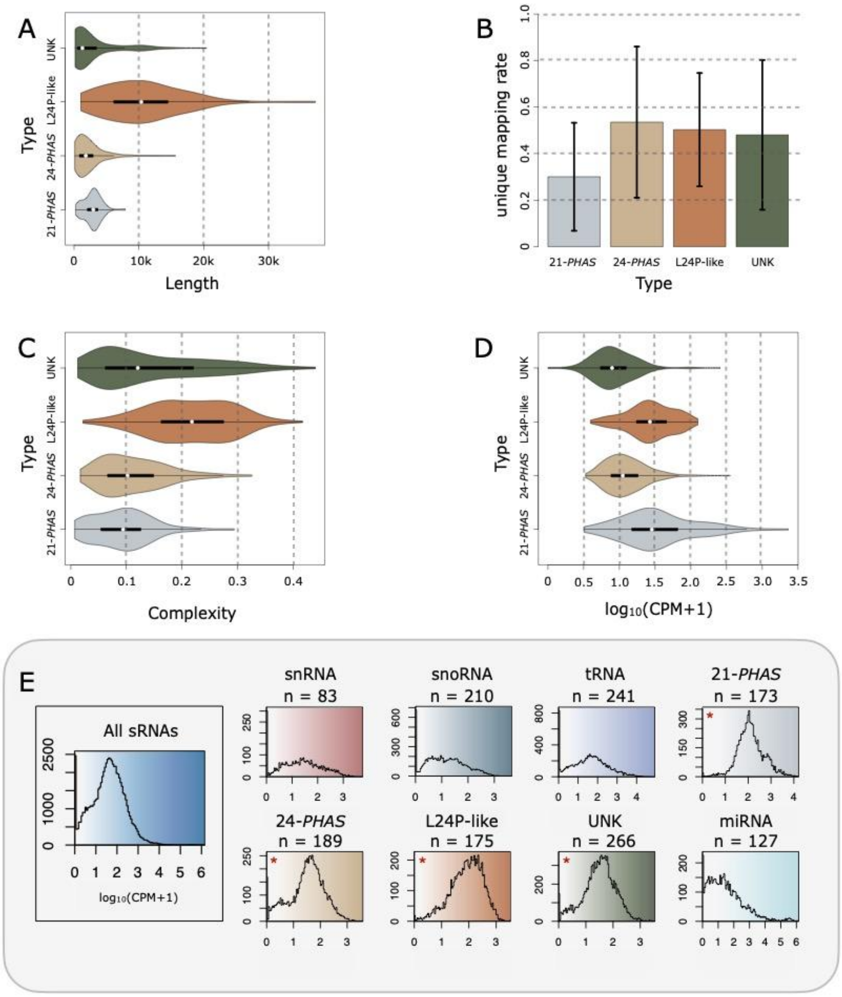

Figure 2 Overall features of phasing loci types in *Coffea arabica* during S0 (node) to flower development. **(A)** Length of different types of phasing loci displayed in kilobases. The *L24P-like* type contains the larger loci producing sRNAs. **(B)** The fraction of uniquely mapping reads aggregated across phased small RNA loci types. The *21-PHAS* sRNAs tend to, proportionally, map less uniquely than the other types. **(C)** Complexity of different types of phasing loci. The complexity parameter is calculated between 0 to 1 where 0 represents a less diverse set of sRNA fragments and 1 a more diverse set of fragments mapped to a loci. The *L*24P-like type has enhanced complexity compared to other types meaning that this type produces more diverse sRNAs. **(D)** Aggregated abundance across different types of phasing loci displayed as logarithmic base 10 of Counts Per Million (CPM) plus one. The *21-PHAS* is the type with most sRNAs fragments mapped to them while the UNK is the less abundant. **(E)** Abundance histogram (with 100 bins) of the different types of expressed sRNA producing loci displayed as logarithmic base 10 of Counts Per Million (CPM) plus one. Leftmost histogram shows the combined histogram of all expressed sRNAs loci (n = 1462). Histograms marked with an asterix (*) shows loci with phasing patterns. (A,B,D) Boxes in the violin plots represent the interquartile range (Q1 to Q3) while the white circles represent the median and whiskers are set to 1.5 times the interquartile range. Maximum and minimum values are delimited at the extremities of the kernel density plots

A total of 266 loci identified as phasiRNA producers by the ShortStack software did not pass our manual evaluation step (Supplemental Table 5). Many reasons can be given for not considering these loci as true phasiRNAs, including a low phasing score (less than 15), their presence in long repetitive regions, strand biases, the number of produced siRNA or a combination of factors. Those sequences were classified in the unknown (*UNK*) category although they demonstrate phasing patterns (Supplemental Figure 6).

Finally, to standardize the L24P-like designation, we proposed that long 24-*PHAS*-like loci must have a phasing score above 15 (calculated using the predictive algorithm described by Chen *et al.*[56]), more than 1,800 nt in length and be phasing at a 24-nt interval. We found that the L24P-like tended to be more complex than the other phasing loci (Figure 2C) meaning they produce a more diverse set of sRNA fragments and their phasing pattern is less clear. Although longer than the other types of phasing loci, they accumulated to a lower level than 21-*PHAS* loci, but accumulated to a higher level than 24-*PHAS* and *UNK* (Figure 2D and E). In addition, 132 of these loci overlap with TE domains, mainly Gypsy (22%), Copia (14%), Helitrons (8%) and hAT (8%).

3.3.6 The flower transition is accompanied by an extensive reprogramming of small RNAs

The transition from vegetative to reproductive stage of coffee meristems - and the following branch senescence - is a biannual process [14]. During these two years, multiple regulatory networks must interpret endogenous and environmental cues to control the change of meristems in the node from a vegetative stage to become flowers and fruits. This program allows multiple floral meristems - induced at different times - to reach anthesis at roughly the same moment [56]. Thus, to better understand the roles of sRNAs in this molecular orchestra, we sequenced and analyzed a total of 250 million sRNAseq reads extracted from eight developmental stages from node with buds undetermined vegetative growth (S0) until anthesis (described in Supplemental Table 1). All the evaluated reads were successfully mapped to a manually curated reference genome and transcriptome.

The class of sRNAs with most mapped reads was the mature miRNAs - accounting for 78% of mapped reads - followed by tRNAs (7.8%) and *21-PHAS* loci (6%) (Supplemental Figure 7). Both Spearman correlation and Multidimensional Scaling analysis showed that libraries grouped into four main clusters, mostly formed by the same or sequential developmental stages (Supplemental Figure 8). Moreover, libraries at the same stage clustered together independently of the coffee genotypes (Siriema or Catuaí) suggesting few differences between different cultivars. Thus, the libraries of the same development stage from both genotypes were analyzed without distinction.

When contrasting sRNA abundance in S0 (nodes with buds under undetermined vegetative growth) with the S1 stage, only 18 small RNA loci were found to be Differentially Accumulated (DA). This finding suggests that S1 (node containing swollen buds) is still in a vegetative stage. However, when we compared sRNAs in S0 samples to more advanced stages (S2 to Flower; reproductive) some clear patterns emerged (Figure 3, Supplemental Dataset 1).

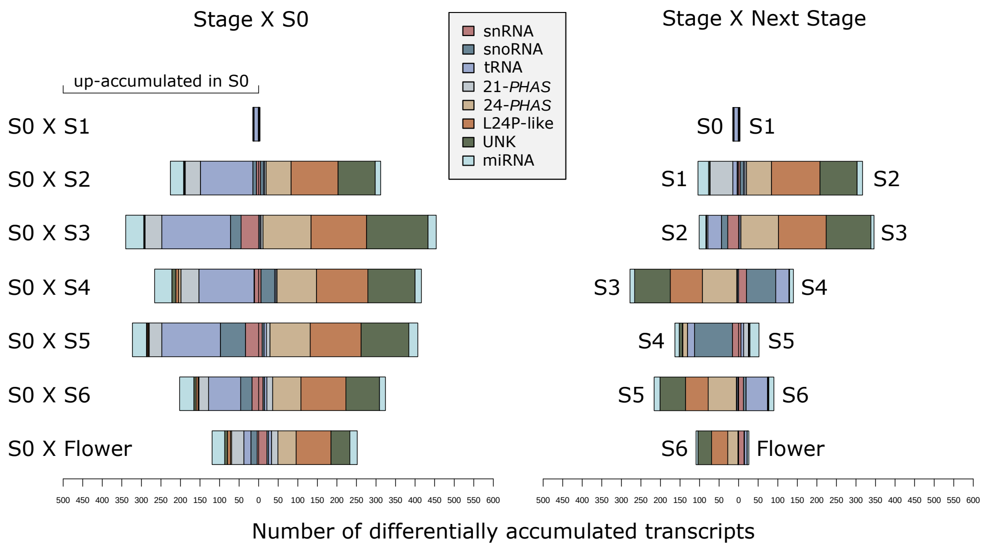

Figure 3 Number of differentially accumulated transcripts based on the small RNA sequencing quantification. Colors represent RNA types, bar represents the number of up accumulated transcripts in a given pairwise contrast between development stages. Left side summarizes the contrast of all stages against S0. Right side summarizes the contrast of a given stage against its subsequent stage. S0; node (with buds) undetermined vegetative growth. S1; Node containing swollen buds. S2; buds <3 mm. S3; Buds >3 mm and <6 mm - early stage. S4; Buds >3 mm and <6 mm - late stage. S5; buds from 6 to 10 mm, light green color. S6; buds >10 mm, white color. Flower; flowers after anthesis

Firstly, we found that tRNAs are preferentially accumulated in the vegetative S0 stage, and the same pattern is followed by the *21-PHAS* loci (Figure 3, left side). On the other hand, 24-*PHAS*, L24P-like, and UNK are preferentially accumulated in the reproductive organs (S2, S3, S4, S5, S6 and Flower). As described above, some of the *PHAS-like* loci (L24P-like and

UNK) may be siren RNAs [52,53] or paternal epigenetically activated small interfering RNAs (easiRNAs) [57] because of their accumulation in reproductive tissues.

Although always more accumulated in reproductive samples, those 24-*PHAS* and 24-*PHAS-like* loci show a progressive reduction in accumulation after their peak in the S3 (Figure 3, right side; supplemental figure 9). Following a contrasting trend, during the transition from S3 to S4 we observed a tendency of increased accumulation of snRNAs, snoRNAs and tRNAs (Supplemental Figure 10). During S4 there is a peak of accumulation of these small RNAs (Figure 3, right side). Subsequently, in S5, they are sharply reduced. After that stage, snRNA and snoRNA levels were constant. tRNAs, on the other hand, were preferentially accumulated in S0 and, by S3, they are strongly reduced. Subsequently, some of the tRNA levels increased after S4. Taken together, our analysis shows that there is a contrasting tendency in the accumulation of tRNAs, snRNAs and snoRNAs compared to the 24-*PHAS*, L24P-like, and UNK. A pontual turning information is the transition of S3 to S4, stages that are traditionally regarded as a single stage called G4 [15].

3.3.7 The balance of *miR156* and *miR172* suggests juvenility restoration in S4

Identical mature paralogs and homoeologs *MIR* sequences were collapsed in a dataset of 447 nonredundant mature miRNAs (Supplemental Dataset 2). Two distinctive groups of miRNAs were found to be preferentially accumulated in vegetative or reproductive stages. The first group was more abundant in the non-reproductive S0 and S1, and are composed by 18 candidate miRNA families, including *miR164, miR169, miR171, miR172, miR319, miR394, miR396, miR399*, in addition to two candidate novel *MIR* families (Supplemental Figure 11). The second group was composed of four candidate miRNA families — *miR156*, *miR171*, and two novel *miR* families - their accumulation was higher in latter stages such S5 or S6 (Supplemental Figure 12).

We noted that the *miR156* family abundance was reduced in the stages S0, S1, and S2, while some *miR156* family members were preferentially accumulated from S3 to flower, with a peak of accumulation verified in S4. The members of the *miR172* family followed a contrasting pattern, more abundant in S0, S1 and S2 (Supplemental Figure 13).

In gametophytes of *Arabidopsis thaliana,* it was verified that members of *miR156* family are *de novo* reactivated to restore the juvenile phase in each generation [58]. The activation of members of *miR156* family in S3, and its subsequent peak at S4 suggest that during these stages the processes of pollen or ovule sporogenesis and gametogenesis are taking place. The decreased accumulation of *miR172* family members in S3 and S4 is in agreement with the finding that the floral induction takes place early in individual buds [56] - probably short before S2 stage. Once an inflorescence meristem is established in a bud and gets at a S3, it will stay latent while other buds are also being formed. However, the development will coordinately continue until anthesis [56].

3.3.8 Co-expression analysis shows that apart from miRNAs, small RNA loci of the same type are preferentially co-regulated

To investigate how the different types of sRNAs accumulate during S0 to anthesis, the a weighted gene co-expression network analysis was performed with the WGCNA package. This analysis yielded five co-expression modules, each assigned a Roman numeral from I to V (Figure 4A, Supplemental Table 6). Each module is predominantly composed of a specific type of sRNA: tRNAs, snoRNA, 21-*PHAS* or 24-*PHAS*/L24*P*-like/UNK. We identified, in each module, their eigengene (also known as eigenvector), that is a single transcribed element (edge) that summarizes the Euclidean mean of a module's regulatory trend [59]. In addition, a filtering step selected the most co-regulated members in each module by setting an adjacency threshold above the 95th quantile (Supplemental Dataset 3).

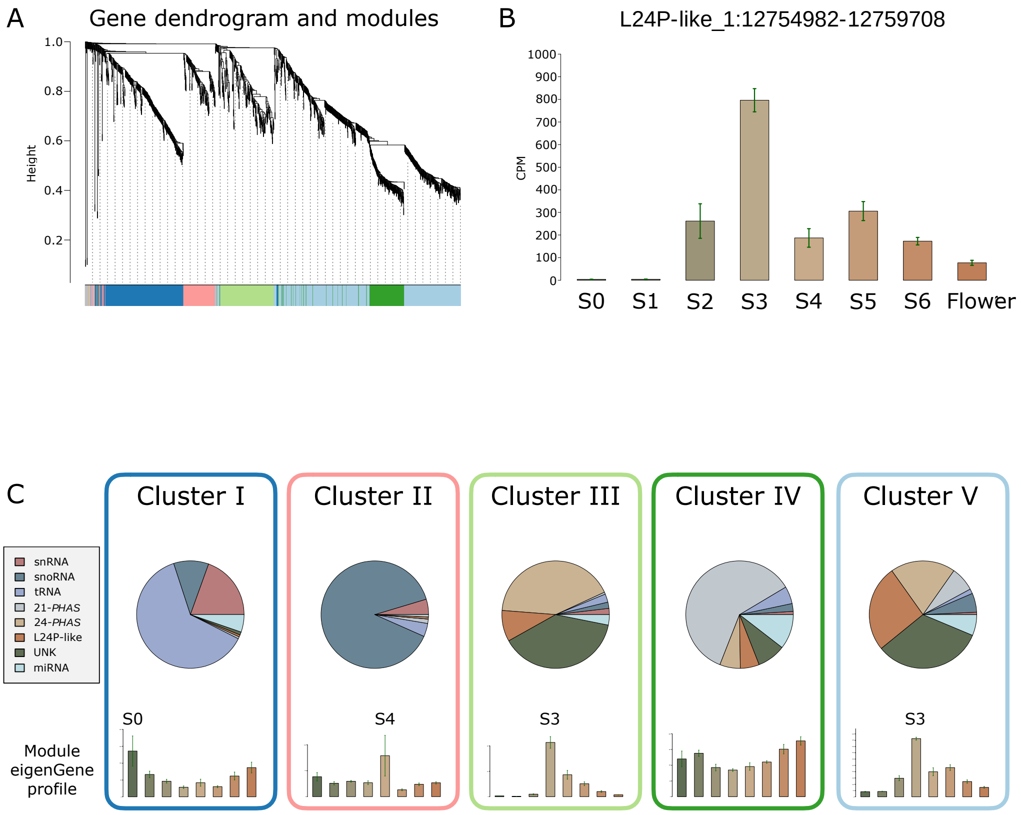

Figure 4 Co-expression analysis shows different types of sRNA producing loci are accumulated in a stage-specific way. **(A)** Gene dendrogram obtained by average linkage hierarchical clustering. The color row underneath the dendrogram shows the module assignment determined by the Dynamic Tree Cut. Different types of small RNA producing loci are clustered based on their accumulation profiles. **(B)** Example of a *L24P-like* locus among the top connected nodes of Cluster V module showing its characteristic low accumulation level in S0 and S1, an increase in S2 and its peak in the S3 followed by a reduction onward. **(C)** Co-expression module composition (pie charts) and barplot representation of its eigengene in the stages, from left to right, S0, S1, S2, S3, S4, S5, S6 and Flower. Cluster I; composed mostly by tRNAs with expression peak in S0. Cluster II; composed mostly by snoRNAs with expression peak in S4. Cluster III; composed preferentially by *24-PHAS* and a significant proportion of *L24P-like* loci and UNK. This module is characterized by an expression peak in S3. Cluster IV; composed mostly by *21-PHAS* loci with the majority of members preferentially accumulated in S0 and S1, although many members are also constitutively present in all evaluated stages and some are accumulated in S6 and Flower. Cluster V; Similar but larger than Cluster III. This module is composed mostly by *24-PHAS*, *L24P-like* and UNK loci with expression peak in S3. However, contrary to Cluster III, this module displays an enhanced proportion of *L24P-like* loci

The largest co-expressed module is designated Cluster V, containing 578 members. The major types of sRNA-producing loci are *24-PHAS* and 24-nt phasing loci, in particular the unknown type (UNK). Its eigengene is a L24P.like with peak accumulation in the S3 stage. To report the most co-accumulated vertices we applied the adjacency cutoff, expressed in the form of a Topological Overlap Matrix - TOM). Doing so, we found highly co-regulated nodes representing members of the miRNA families *miR171, miR396, miR156, miR319, miR399* and three candidate novel *MIR* families. The majority of elements of this module are weakly accumulated in S0, S1 and S2, but in S3 they present a sharp increase (Figure 4C).

The Cluster III follows a similar trend to Cluster V, its main components are also 24-*PHAS* and 24-nt phasing loci but it has fewer members: 165. Its eigengene is from the UNK type with peak accumulation in S3. It had a negligible contribution in S0, S1 and S2. The top connected loci are all from 24-*PHAS and* 24-nt phasing like loci (Figure 4C).

The Cluster IV is primarily composed of 21-*PHAS* loci (~60% of its 216 members). After filtering for the top connected nodes, the proportion of *21-PHAS* loci increased to 82%. Additionally, miRNAs from families *miR156* and *miR482* were among the top connected vertices of this regulatory module. We identified four valine tRNAs (with anticodons CAC and AAC) among those highly connected nodes. No clear accumulation peak was observed as members of this module are stable along the stages, although a reduction in the S3 is perceived (Figure 4C).

Regarding the Cluster I, the majority of its members are tRNAs (60% of 336 vertices) but it also has a significant proportion of snRNAs and snoRNAs (19% of and 10% respectively). Nevertheless, after the filtering for the top connected vertices, the proportion of tRNAs increases to 92%. This module eigengene is a tryptophan tRNA (anti-codon CCA) that is predominantly accumulated in S0 and Flower (Figure 4C). The overall accumulation profile of members of the Cluster I seems to follow a similar trend of Cluster IV and, likewise, following a contrary trend to Cluster V and Cluster III.

The Cluster II is predominantly composed of snoRNAs (89% of its 154 members). This proportion rises to 93% when we only consider the top connected vertices. Members of this module, specially the snoRNAs, are preferentially accumulated at the S4 (Figure 4C, Supplemental Figure 10). This pattern is contrasting to the other modules and - because of its exclusive peak in S4 - we hypothesize that RNA metabolic processes and ribosome synthesis are active pathways in the resumption of development in response to water after its deficit.

3.3.9 Cluster IV is enriched for 21-*PHAS* loci that are composed of sRNAs produced from resistance genes

Cluster IV was found to be the module with higher adjacency between its members, showing that their accumulation profile is fine tuned. It is possible that the verified co-regulation of Cluster IV is relatively constant from S0 to Flower (Figure 4C). We supposed this long term regulatory control is achieved by the large proportion of disease resistance gene transcripts that are being processed into 21-nt phasiRNAs by the targeting of the most abundant miRNA family *miR482* in terms of total number of loci. The transcriptional investment on defense systems seems to be of constant importance in the plant budget and a tightly-controlled decision making system must be always available to cut costs and/or call for a cease fire.

3.3.10 miRNAs are central members of all modules, except the snoRNA rich cluster II

No module was predominantly composed of miRNAs (Figure 4C). Instead, miRNAs are distributed across all modules and are among the top connected nodes in Clusters I, III, IV and V. Those modules displayed contrasting compositions, i.e Cluster IV is dominated by *21-PHAS* loci and has members of miRNA candidate families *miR482, miR403, miR162 and miR167* among its top connected vertices. Cluster V, a module enriched by *24-PHAS* and *24-nt phasing-like* loci, has among its top nodes members of miRNA candidate families *miR171, miR319, miR396, miR399*. Both clusters IV and V present different members from family *miR156* (Supplemental Table 6). The cluster I has the miR candidate families *miR8155 miR5139* being co-regulated with t-RNAs. Cluster II, enriched for snoRNAs, is the exception once none of top connected vertices are miRNAs candidates. Its accumulation profile is the most distinct being characterized by a peak accumulation of snoRNAs in S4, a brief stage that is followed by fast morphological differentiation.

3.3.11 S4 to S5 marks an important transition accompanied by changes in the accumulation levels of miRNAs and their target genes

The transition of S4 to S5 is a key step of *Coffea arabica* flowering development because it marks the transition from the seemingly latent buds S4 (Buds > 3 mm and < 6 mm) to an active and fast stage of development in S5 (buds from 6 to 10mm with light green color) [13,60]. This transition occurs in response to environmental stimuli, such as rains or irrigation while plants are still at S3. It is possible that S4 marks the preparation of the molecular machinery to allow the resuming of the reproductive development and the triggering of anthesis [13,60]. The differential accumulation analysis of small RNAs showed that during S4 the *24-PHAS* and similar 24-nt phasing loci accumulation levels are abruptly reduced (Figure 3, Supplemental Figure 9). Meanwhile, snRNAs and snoRNAs levels are rapidly increased (Figure 3, Supplemental Figure 10).

To better investigate the transition, we produced RNAseq libraries from S4 and S5 stages as well as PARE libraries of samples from S4, S5 and anthers in pre-meiotic, meiotic and post-meiotic stages. The miRNA target prediction rendered a total of 3,213 miRNA-genic target pairs (Supplemental Table 7). We also identified 2,003 inter-genic miR-target pairs in the *Coffea arabica* whole genome (Supplemental Table 8). Figure 5 summarizes the main findings regarding the accumulation levels of miRNAs and their target genes.

aaa

3.3.12 The miR families *miR396, miR156* and *miR172* are master modulators of PCG in S4 and S5

The most up accumulated microRNA in S4 compared to S5 is a member of family *miR396* (*ccp-miR396e-3p*) which we found to be targeting ten *GROWTH-REGULATING FACTOR* (*GRF*) loci (Supplemental Table 9). In grapevine, mutations in the *miR396* binding site of *GRF4* caused changes in the inflorescence bunch architecture [61]. Our RNAseq differential expression analysis of PCG found that all of these *miR396* targeted *GRF* are significantly more abundant in S4 than S5 (Supplemental table 10). As shown in *Vitis vinifera,* the modulation of *GRF* by *miR396* can be engineered to promote elongated pedicels due to an enhanced cell division rate or an extended window of cell proliferation [61]. In *Coffea arabica* this modulation of *GFR* by *miR396* might loosen the architecture of fruit bunches, that is usually compact. This loosening would be achieved by an elongation of the peduncle and that could be beneficial to facilitate the harvesting without causing too much damage to the plant, once large-scale mechanical harvesting techniques are known for its detrimental effects to branches and leaves.

Although the most accumulated *MIR* gene in S4 is a member of *miR396* family it is important to note that there are also some family members that are up-accumulated in S5.

Whereas the up-S4 *ccp-miR396e-3p* targets *GFRs,* the group of loci encoding for three mature miRNAs up-accumulated in S5 are targeting genes such as *TIC110-chloroplastic*, *THREONINE-TRNA LIGASE,* and *AGAMOUS-LIKE MADS-BOX AGL62*. Of these target genes, only *AGL62* was found to be Differentially Expressed (DE), being up-regulated in S5. In *Arabidopsis thaliana* the *AGL62* is only expressed in seed endosperm, regulating cellularization and acting as an upstream activator of *InvINH1*, an endosperm-specific invertase inhibitor [62,63]. The *InvINH1* in *Arabidopsis* ultimately regulates embryo growth rate during the early stage of seed development [62,63]. In *Coffea arabica* we identified the gene SubC.e_6694 - characterized as a putative invertase inhibitor - beeing up-regulated in S5. It is possible that this regulatory network of *miR396-AGL62-InvINH1* is important to control the growth rate of coffee floral buds by regulating the supplying energy (sucrose) when needed.

Five members of family *miR156* were more accumulated in S4 than S5 and we found them to be targeting seventeen *SQUAMOSA PROMOTER-BINDING-like* (*SPL*), preferentially at the pre meiotic stage. Only one *SPL* was up-regulated in the stages S4 compared to S5 (Supplemental Tables 9 and 10). The occurrence of both miRNAs and their targets expressed in the same stage without strong evidence of reduction in the target transcript levels is in agreement with the function that miRNAs can work as a fine tuning regulatory mechanism [64], instead of an on/off switch (Supplemental Figure 14).

In the leaves of *Arabidopsis thaliana*, the interaction between the *miR156*-*SPL3* module and *FLOWERING LOCUS T (FT)* is part of the regulatory mechanism controlling flowering time in response to ambient temperature [65]. The overexpression of *miR156* in leaves at low temperature (16 ºC) was associated with down-regulation of *FT* whereas overexpression of *miR156*-resistant *SPL3* was associated with higher levels of *FT* and an early flowering phenotype [65]. Here, a *FT-like* gene was found to be up-regulated in S5 compared to S4. Meanwhile, most *miR156* family members are less accumulated in S5 and a *SPL* was also down in S5.

We identified a *floral homeotic AGAMOUS-like* being targeted by the group of up-regulated *miR156* family members in S4. The targeting evidence was verified during the post-meiotic stage. We were unable to find evidence of differential expression of this specific *AGAMOUS-like* in the S4 to S5 transition. However, four other *AGAMOUS-like AGL80*, *AGL62* and two copies of *AGL15* are upregulated in S5. In *Arabidopsis thaliana*, the floral repressor *miR156* is positively regulated by the *AGL15* [66] which overexpression causes the delaying of the floral transition [67]. It is possible that this up-regulation of *AGL15* is activating some *miR156* in S5 to counteract the *SPL* inducing *FT* effect in an attempt to precisely balance flower development.

The *miR156-SPL* interaction is a conserved endogenous cue for the transition from vegetative to reproductive phase [65]. In *Arabidopsis thaliana,* this transition is controlled by a regulatory network that processes long-day cues through the interaction of gibberellins (GA), DELLA, SPL and the balance of *miR156/miR172* [68]. The DELLA proteins repress flowering. On the other hand, GA promotes the degradation of DELLA and, in turn, promotes flowering. DELLA directly binds to *miR156*-targeted *SPL* transcription factors, which promote flowering by activating *miR172* and some MADS-box genes [68].

No GO terms were enriched for GA biosynthesis or response during S4 to S5 transition. However, we found five gene encoding GIBBERELLIN-BETA-DIOXYGENASES and one GIBBERELLIN-REGULATED proteins being up-regulated in S4. Meanwhile, three GIBBERELLIN-BETA-DIOXYGENASES, two GIBBERELLIN 20 OXIDASES (GA20OX) and two GIBBERELLIN-REGULATED genes were up-regulated in S5 (Supplemental Table 10). In *Arabidopsis thaliana*, *GA20OX* contributes to the induction of flowering by long days [69] . In addition, *DELLA* transcripts were also found to be up-regulated in both S4 and S5. The roles of *DELLA*, *MADS-box* and GA biosynthesis and response genes in coffee flower development are still unclear and require further investigation.

Among the mature miRNAs more abundant in S5 than S4, we found a single member of *miR156* family. Like the group of mature *miR156* that accumulate in S4, this S5-enriched copy also targets *SPL* transcripts (Supplemental Figure 15, Supplemental Table 11). This targeting of *SPLs* might function to reduce the activity of *FT* during S5 to slow floral development The up-regulation of *AGL15* in S5 may explain why this *miR156* member is more accumulated in libraries of later developmental stages such as S5, S6 and Flower. In addition, it shows a regulatory motif of *AGLs* and *miR156* family members where some *AGL* are targeted by *miR156* and, in turn, *AGL15* can promote the expression of some *miR156* [70] in a negative feedback loop. This co-regulatory network motif may be an interesting mechanism by which coffee plants synchronize the flowering process.

3.3.14 Ethylene responsive genes, such *AP2*, are DE during the transition of S4 to S5 and are target by *miR172* family members

Degradome analyses showed two mature *miR172* family members, identified as *miR172.2.ab,* targeting four *ETHYLENE-RESPONSIVE TRANSCRIPTION FACTOR RELATED TO APETALA2-7-like* (*ERF-RAP2-like*) mainly in S4 but, to a lesser extent, also in S5 (Supplemental Figure 16, Supplemental Table 7). It is important to note that *miR172.2.ab* was not found to be differentially accumulated between S4 and S5 but we found one of its targets, the *APETALA2/ETHYLENE RESPONSIVE FACTOR* (*AP2/ERF*), and another 41 genes identified as *ERF* being more expressed in S4 than S5 (Supplemental Table 10). Although the *AP2/ERF* superfamily name refers to ethylene, the responsiveness to this growth regulator is not a universal feature of this superfamily [71]. The *AP2/ERF* superfamily are known to be involved in transcriptional regulation of a variety of biological processes related to growth and development [72], as well as various responses to environmental stimuli [73] such as responses to biotic and abiotic stresses [74].

This mature *miR172.2.ab* also targets two genes identified as "*floral homeotic protein APETALA 2"* (*AP2*) of which one was found to be upregulated in S4. The *Arabidopsis thaliana* AP2 may promote early flowering identity [75]. It also has a subsequent function on the transition of the inflorescence meristem to a floral meristem [76] and plays a central role in the specification of sepals and petals [77]. *AP2* has dual molecular roles working as both a transcriptional activator and repressor by modulating gene expression in logical, reinforcing circuits [70]. For example, *AP2* can repress *miR172* and, simultaneously, be repressed by it [70]. *AP2* can also induce the expression of *AGL15*, a floral repressor, and directly down-regulate the transcription of floral activators like *SUPPRESSOR OF OVEREXPRESSION OF CONSTANS1* (*SOC1*) [70] and *FT* [78]. Here we found a contrasting expression profile of the *MADS-box SOC1* that is up-regulated in S4 while *FT* is up-regulated in S5 (Figure 5).

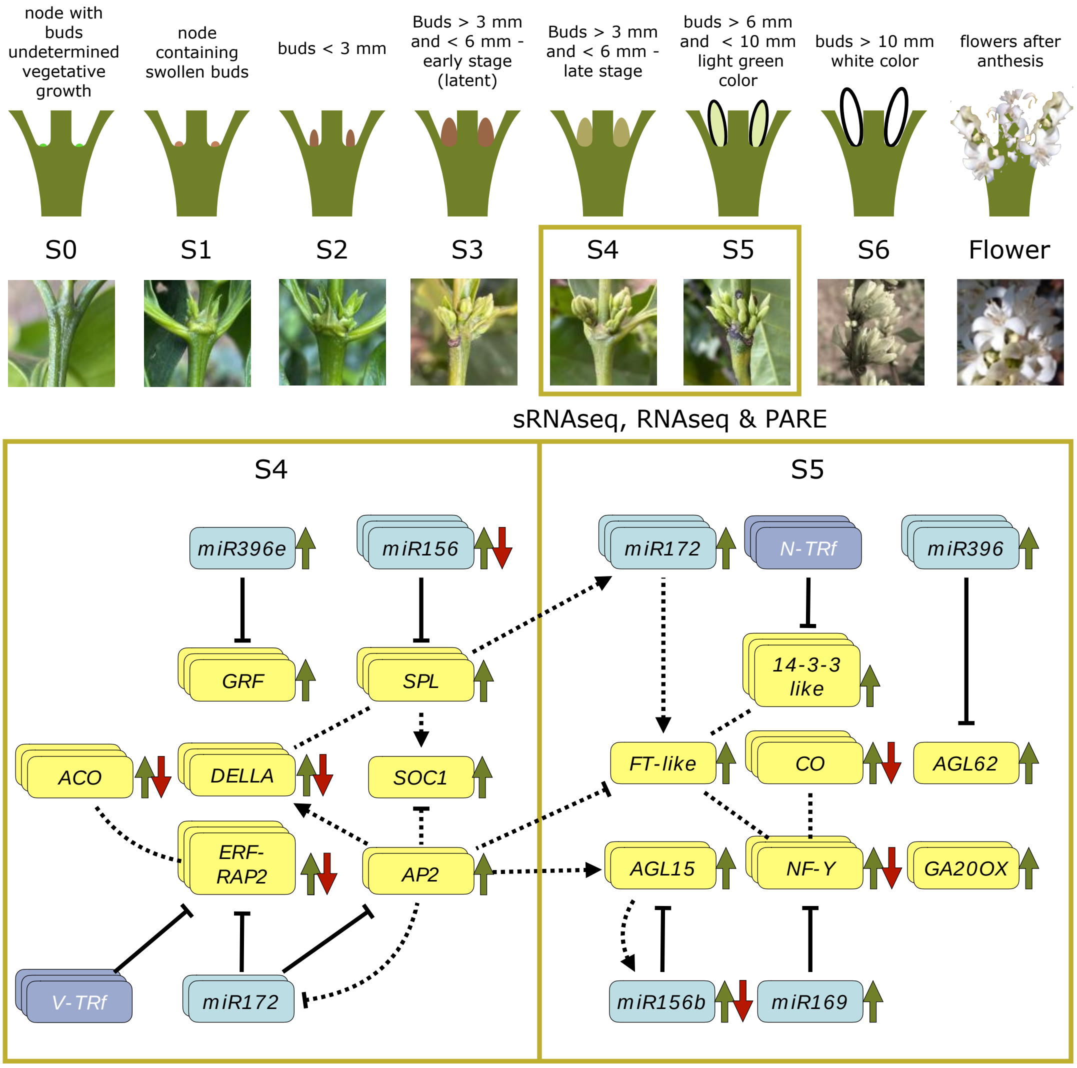

Figure 5 Schematic overview of the S0 to flower development and the core regulatory network of miRNA/t-RNAs and Protein Coding Genes governing the S4 to S5 transition. The combination of sRNAseq, RNAseq and PARE analysis reveals conserved motifs of gene regulation. The divided rectangle depicts the contrasting transcription profile between S4 and S5. Blue boxes represent *MIR* genes while Darker blue represents tRNA fragments (tRFs), yellow boxes represent Protein Coding Genes (PCGs). Single boxes show that a specific gene is being transcribed while two stacked boxes show that two homologs of that gene are transcribed and three stacked boxes show that three, or more, homologs of are transcribed. Green arrows pointed upwards represent genes being statistically more expressed in the S4 or S5 stage - according to the side which a gene is. Green and red arrows in opposite directions show that different homologs - or family members - of a gene present a contrasting transcription trend. Continuous black arrows with blunt ends between a interferingRNA and a PCG shows that the targeting was verified using PARE data. Dotted lines without arrows show that there is literature support of protein-protein interaction. Dotted lines with arrows between two genes shows that there is literature support that one gene is a direct activator of the other gene in other plant species. Refer to main text for references and complete gene names

3.3.15 The enzyme responsible for the final step in ethylene biosynthesis is present in both S4 and S5 but the precursor step is missing

Among a myriad of biological processes, ethylene is also an important anthesis regulator in *Coffea arabica* [13]. Ethylene biosynthesis involves two dedicated steps which a S-adenosyl-L-methionine (SAM) precursor is converted to 1-aminocyclopropane-1-carboxylic acid (ACC) by ACC-synthase (ACS) and then it is transformed into ethylene *by* 1-AMINOCYCLOPROPANE-1-CARBOXYLATE OXIDASE (ACO) [79,80]. Nevertheless, the regulation of its biosynthesis is far more complex and occurs at multiple regulatory layers in specific tissues [81].

Here, in addition to the 42 *ERF* genes that are more expressed in S4 than S5, we found 12 *ACO* genes up-regulated in S4. The second and third most DE genes in S4 are *ACOs*. In contrast, we found ten other *ERF* and eight *ACO* being up-regulated in S5, but to a significantly lower level of fold change (Supplemental Table 10). In addition, no *ACS* were found to be DE between S4 and S5. These results are in agreement with the thesis that the step of the conversion of SAM to ACC occurs in other tissues and are transported to the inflorescence meristem instead of being synthesized there [82].

In this work, we show similar components of ethylene biosynthesis and response being up or down-regulated across contrasting developmental stages. Similar results were previously reported by Almeida-Lima *et al.* [82]. We show that the ethylene biosynthesis from its ACC precursor in buds is regulated by a complex multilevel control circuit, of which we could reveal a small fraction.

3.3.16 Transcription factors related to the Circadian regulation of FT are targets of *miR169* in S4

Other relevant genic miRNA-target pairs of abundant miRNAs in S5 are *miR169* targeting eleven *NUCLEAR TRANSCRIPTION FACTOR Y* (*NF-Y*) subunits, one of which is up-regulated in S4. Other seven *NF-Y* genes that were not found to be targets in the PARE analysis, were found to be DE. Three were down-regulated in S4 whereas four up-regulated in S5. Some members of *NF-Y* play important roles in the control of photoperiod-dependent flowering time as protein-protein integrators to allow DNA target specificity of transcriptional regulators. *NF-Y* are known to allow the binding of CONSTANS (CO) to the CCAAT-box containing region of *FT* [83–85]. Here we found nine *CO-like* transcripts being up-regulated in S4 whereas four were up-regulated in S5. This finding of DE CO-like genes in S4 and S5 may suggest that the circadian clock within buds is an important feature for the decision making regulatory circuit of floral development in *Coffea* and that this circuit is linked with the regulatory mechanism based on *miR169*.

3.3.17 tRNAs are among the most up accumulated sRNA genes in S5 and are putatively targeting protein coding genes and TE

miRNAs are important elements involved in the transition from S4 to S5 and the most abundant type of sRNA in the samples from nodes containing undetermined vegetative cells (S0) to flower. Nevertheless, we noticed that the most differentially accumulated sRNAs, in terms of fold change, in S5 are four tRNAs. They are, at least, one hundred times more abundant in S5 than S4 (Supplemental Figure 17). The most abundant sRNAs S5 were two tRNA lysine (anticodon CUU) followed by a tRNA asparagine (anticodon AUU) and tRNA threonine (anticodon UGU). The asparagine and threonine tRNAs have relatively low abundance but the differences in their fold change are remarkable. Because of this sudden increase, we examined if they had some regulatory role during flower development or were just a by-product of the enhanced transcriptomic activity that underlies the development. We found 1,373 tRNFs potentially targeting 1,131 PCGs (Supplemental Table 12).

Three of the four most differentially accumulated tRNAs in S5 compared to S4 represented tRFs with protein coding gene targets. These tRNAs are a transfer-RNA asparagine (anticodon AUU) and two tRNA lysine (anticodon CUU). Degradome analysis shows a asparagine tRF with 24 nt in length “UCCUCAGUAGCUCAGUGGUAGAGC” potentially targeting a gene coding for a carbohydrate esterase enzyme during the pre-meiotic stage. Meanwhile the two most differentially accumulated tRNA, both of them lysine (anticodon CUU), were found producing a 24 nt tRFs ”CUCAGUCGGUAGAGCGCAAGGCUC” potentially targeting two genes encoding for “intracellular protein transport protein USO1-like”. Finally, one of these highly differentially accumulated lysine tRNAs was found to be producing a 22 nt tRF “CGAACCCACGACCACAAGGUUA” potentially targeting a gene encoding for a chloroplastic protein identified as “PEP-RELATED DEVELOPMENT ARRESTED 1” in post-meiotic anthers.

One of the most abundant tRFs, specially in S0, has a length of 21 nt “AUCAGAGUGGCGCAGCGGAAG" and is a fragment derived from two methionine tRNA (anticodon CAU) loci in chromosomes 4 and 8, both inherited from the *C. eugenioides* parent ancestor. This tRNF is the most abundant with validated genic targets from PARE (Supplemental Table 13, Supplemental Figure 18). In post-meiotic anther stages, it was shown by PARE analysis to match to 24 cleaved genic loci including KRAB-A domain containing proteins and three DDE-type integrase/transposase/recombinase transcripts (Supplemental Data 12). Elsewhere, we reported that loci identified as *DDE* and *KRAB-A* are widespread across the *C. arabica* genome. Because KRAB are domains only present in tetrapod vertebrates [86] and homology based searches to the nr database in NCBI returned TE related matches; therefore, we believe that these reported KRAB-A domains containing proteins are, in fact, mis annotated TEs.

Interestingly, we found six ethylene-responsive transcription factors predicted to be targeted by tRFs. Two of them were target by the tRF 'GUUGCUGUGGUGUAGUGG' that are 18 nt reverse complement sequences derived from the 3` end of valine tRNA loci (anticodon UAC) with PARE signals in S4, S5 and post-meiotic samples. In addition, two 14-3-3-like protein-A transcripts were found to be targeted by the tRF 'CCACAAGGUCGAAGGUUC' from 5 asparagine tRNA loci (anticodon GUU) in meiotic and post-meiotic PARE libraries. Three other *14-3-3-like* genes were found to be up-regulated in S5 and are known to be intracellular receptors for the FT florigen in rice shoot apical cells [87].

Whether these, and about a thousand other tRF-target pairs are true is a discussion still open to debate (or at least validated with additional experiments). Although we used stringent cutoffs both for selecting the accumulated tRFs and the sPARTA-identified targets, about 60% of the tRF-target pairs were of length below 20 nt and the shortest small RNA with a verified ability of being loaded into an ARGONAUTE protein is 20 nt. On the other hand, 173 (15%) of the predicted tRFs genic targets were found to be DE between S4 and S5 (Supplemental Figure 19). Additionally, it is also possible that many smaller than 20 nt tRFs are sub-sequences of longer tRFs that can be loaded in ARGONAUTE or that the tertiary structure of the tRNA can guide some endonuclease to a target transcript.

3.3.18 The transcriptome of S4 is responsive to hormones and involved in biosynthesis of RNAs whereas S5's is focused on the production of cell walls

Small nucleolar RNAs (snoRNAs) and their associated proteins are ancient devices that mediate maturation of ribosomal RNA (rRNA) [88]. The S4 and S5 stages mark an evident change in the accumulation of snoRNAs that are abundant in S4. In S5, there is a relative reduction in the levels of not only snoRNAs but also other sRNA types (Supplemental Table 14, Figure 3). In addition, a total of 39 ribosomal-related PCG were found to be DE between S4 and S5. The up-regulated ribosomal-related genes in S4 are predominantly coding for 60S ribosomal proteins and ribosome biogenesis proteins, whereas the up-regulated in S5 are 50S and 30S ribosomal protein components. Different components of ribosome subunits are transcribed in distinct phases.

Because ribosomes are central to the protein translation pathways, we performed an overall analysis of GO terms of the 7,283 PCGs that were DE between S4 and S5 stages (Supplemental Table 10). We classified those DE genes into 684 GO terms in the categories Biological Process (BP) Molecular Function (MF) and Cellular Component (CC) - Supplemental Tables 15 and 16.

In S4, 3,502 PCG were up-regulated in comparison to S5. The most enriched GO term was RNA biosynthetic process (GO:0032774, pval = 2.5E^{-36}) with its child term rRNA metabolic process (GO:0016072, pval = 4.4E^{-2}) in addition to response to hormone (GO:0009725, pval = 2.25E^{14}), response to auxin (GO:0009733, pval= 7.42E^{-10}), response to ethylene (GO:0009723, pval = 1.4E^{-2}), meiosis1 (GO:0007127, pval = 8.14E^{-6}) and stamen development (GO:0048433, 4.2E^{-2}). The most enriched CC term was nucleus (GO:0005634, pval = 3.8E^{-33}), followed by the child terms DNA repair complex (GO:1990391, pval = 5.9E^{-11}) and nuclear replication fork (GO:0043596, pval = 2E^{-7}).

In S5, 3,781 PCG were up-regulated in comparison to S4. GO terms were particularly enriched for BP such as carbohydrate metabolic process (GO:0005975, pval = 5.3E^{-49}), transmembrane transport (GO:0055085,pval = 2.4E^{-29}), localization (GO:0051179, pval = 9.8E^{-20}) and plant-type cell wall organization or biogenesis (GO:0071669, pval = 6.20E^{-6}) with its child terms xyloglucan metabolic process (GO:0010411, pval = 1.6E^{-2}) and xylan metabolic process (GO:0045492, pval = 2.42E^{-6}). In addition, the enriched CC terms were mostly related to cell wall (GO:0005618, pval = 6.54E^{-11}), photosystem (GO:0009521, pval = 3.98E^{-8}) and plasmodesma (GO:0009506, pval = 4.2E^{-4}).

The S4 is marked by important processes. The development of male organs with their respective meiosis processes, integration of hormonal signals, and intense metabolism of RNAs, in particular snoRNAs, are characteristic events of buds in S4. Next, during S5, the expressed genes are mainly involved in metabolic processes to develop cell wall structures. The metabolism of carbohydrates for building compounds, such as xyloglucan and xylan, is anatomically evident once in a matter of days the buds can grow from 6 mm to 10 mm, an increase of more than 60% in their length.

3.3.19 Resistance genes related to *21-PHAS* are differentially expressed between S4 and S5

Finally, *21-PHAS* loci, which were processed mainly from R genes, were constitutively present during all the evaluated stages. Their relative stability in accumulation during S4 and S5 led us to inquire if R genes were DE between those stages. We found 209 R genes DE between both conditions, with 152 up-regulated in S4 and 57 up-regulated in S5. About a third of these DE R genes were identified as “late blight resistance”. In addition, we found 79 DE genes identified as “Leucine Rich Repeaters repector like” (*LRR*) that are known for recognizing effectors in a direct or indirect fashion and promote R-gene-mediated immunity [89]. It is not clear why different R genes are DE between both stages while the proportion of late blight resistance genes is constant. It is possible that different pathogenic fungi are trying to infest the buds during different stages of bud development [90]. It is also possible that the regulatory networks governing defense strategy promote the transcription of different combinations of pathogen recognition and R genes at different stages in a trial and error fashion to find optimum responses to pathogens.

3.4 Discussion

We conducted a genome-wide annotation of miRNAs and phasiRNA in the allotetraploid *Coffea arabica.* This data was then integrated with publicly available annotations of tRNAs, snRNAs and snoRNAs [22], allowing us to investigate the abundance profiles of different RNAs during reproductive organ development. Finally, we coupled these datasets with RNAseq differential expression data from two key stages (S4 and S5) and degradome analysis.

We show that tRNAs are preferentially accumulated in S0 whereas 24-*PHAS* and 24-PHAS-like loci are preferentially accumulated in S3 (Figure 3 - left side, Figure 4 - Clusters I, III, IV and V). The 24-*PHAS* accumulation coincides with an extended period of anatomical latency in which buds are apparently dormant - S3. This stage-specific accumulation also resembles the spatio-temporal dependent expression of many 24-nt phasiRNAs in species such as barley [91], wheat [91], maize [92], tomato [47], petunia [47] and soybean [93]. It is possible that some of those sRNAs producing loci that are phasing with a 24-nt signature are responsible for guiding the machinery of epigenetic imprint, some of them may be siren or easi RNAs [49,52,53,57]. Meanwhile, snRNAs and snoRNAs are preferentially accumulated in S4, in accordance with the most enriched GO term RNA biosynthetic process that are characteristic of its DE genes (Supplemental Figure 10). It is possible that the re-activation of the transcription is an important feature for resuming flower development.

3.4.1 The *miR482* family are expanding and evolving rapidly in *Coffea arabica* possibly to allow proper response control to pathogens

The miRNAs, in accordance with their fine tuning regulatory role [64] , were found to have families preferentially accumulated in vegetative or reproductive organs (Figure 3 - left side; Supplemental Figures 11 and 12). *miR482* is the family with more precursor loci - at least 20 - which is a similar number to spruce with at least 24 loci [94]. In addition, *miR482* is the third most accumulated family in terms of total number of fragments mapped to their mature sequences. This higher number of *miR482* loci, and the emergence of a putative novel mature *miR482* in *Coffea* (Novel-245889.2.ab), suggests a fast evolutionary rate of these family members. We hypothesize that this reflects an army race between evolving components of plant resistance systems and evolutionary strategies of potential and known pathogens [95].

The *miR482* family is mainly targeting disease resistance genes that were found to be phased in a 21-nt pattern. *miR482* members are central nodes of the co-expression cluster IV, dominated by 21-*PHAS* (Figure 4). During the S4 stage, we identified 26 R genes up-regulated that were also found to be 21-*PHAS* loci. The ability of mRNAs from R genes to be processed and used for post-transcriptional silencing can strengthen the regulatory network by targeting the original and other similar LRRs [94,96]. In fact, the Cluster IV was verified to have the highest mean co-expression (adjacency) value of all modules (Supplemental Dataset 3).

The accumulation of members of Cluster IV tends to be relatively constant along S0 to anthesis (Figure 3C) suggesting that the *21-PHAS* loci, which corresponds to the majority of its members, are not directly involved in reproductive tasks. They are fine-tuned defense systems to provide proper control of responses to pathogens. Misregulation of R genes can lead to severe fitness costs to plants [95]. If the translation of a given R is too high, it can lead to deleterious auto-immune responses [97]. On the other hand, too slow responses due to lack of R gene transcription will allow pathogens to take over the plant cell [98]. An inappropriate response timing and strength may ultimately lead to programmed cell death [90,95].

3.4.2 The S3 to S4 and S4 to S5 transitions marks changes in sRNA abundance

The transition from S3 to S4 is marked by an evident change in the accumulation of not only *24-PHAS*/24P-like/UNK loci but also snoRNAs, snRNAs and tRNAs (Figure 3). This transition occurs in response to water after a long period of drought and is followed by rapid development. In S4, there is a clear decrease in the accumulation of 24-nt phasiRNAs while structural RNA levels such as snRNAs, tRNAs and especially snoRNAs are sharply increased. This accumulation of the snoRNAs in S4 is what defines the Cluster II (Figure 3 - right side, Figure 4 and Supplemental Data 3). This accumulation of snoRNAs was correlated with 60S ribosomal protein components which may control expression of PCG via enhanced translation efficiency [99]. It is also possible that some snoRNAs fragments can be loaded into ARGONAUTE to promote post-transcriptional gene silencing in a similar way to miRNAs[100].

The next transition, S4 to S5, is marked by a slow decrease in the accumulation of *24-PHAS* that will continue until anthesis (Figure 3). Nevertheless, miRNAs levels increase in S5 suggesting tight regulation of developmental processes. Eighteen mature miRNAs from the families *miR156, miR162, miR164, miR169, miR172, miR1919, miR396*, *miR398* and 5 novel families candidates were up-regulated in S5 when compared to S4 (Supplemental Table 14). Our degradome analysis showed that many of these miRNAs are targeting important genes related to flower development and response to ethylene (Figure 5). Our results reinforce the thesis that in *Coffea arabica* the ethylene is a key hormone governing the flower development [13,82].

3.5 Conclusion

Although extremely conserved in angiosperms, the core regulatory circuit of floral induction and development varies in terms of what cues it will follow to promote flowering in a given species. This directly correlates with the widespread evolutionary strategy of maximizing the reproductive success of a set of genes in an organism. *Coffea arabica* plants are tropical perennial plants with a sustained floral induction window that, in Brazil, occurs from February to October [56]. This long induction window coupled with environmental variability and different rates of branch growth leads to asynchronous flowering.

Nevertheless, we suggest that,as far as the transcriptome is concerned, the floral development is an event controlled at multiple levels that evolved to promote a synchronized anthesis in the face of endogenous and exogenous variability. This is possible because of complex interactions of PCG and diverse types of RNAs programs trying to decide the best timing for anthesis.

Because of the long floral induction window, different branch nodes are producing floral buds at different months. There must be a delaying program to stall the development of early formed buds. This “wait until the signal comes” takes place in S3 together with the accumulation of 24-*PHAS* and similar sRNAs. We suggest that the accumulation of these phasiRNAs may be part of some important stabilizing mechanism to keep cells waiting. It is also possible that some of these 24-*PHAS* loci may comprise a system for long-term epigenetic imprinting in response to the environment.

REFERENCES

1. Keam, S. P. & Hutvagner, G. tRNA-Derived Fragments (tRFs): Emerging New Roles for an Ancient RNA in the Regulation of Gene Expression. *Life* **5**, 1638–1651 (2015).

2. Grigoriev, A. Transfer RNA and Origins of RNA Interference. *Front. Mol. Biosci.* **8**, 708984 (2021).

3. Martinez, G., Choudury, S. G. & Slotkin, R. K. tRNA-derived small RNAs target transposable element transcripts. *Nucleic Acids Res.* **45**, 5142–5152 (2017).

4. Rhizobial tRNA-derived small RNAs are signal molecules regulating plant nodulation. https://www.science.org/doi/full/10.1126/science.aav8907.

5. Scalabrin, S. *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.* **10**, 1–13 (2020).

6. Lashermes, P. *et al.* Molecular characterisation and origin of the Coffea arabica L. genome. *Mol. Gen. Genet. MGG* **261**, 259–266 (1999).

7. Bertrand, B. *et al.* The greater phenotypic homeostasis of the allopolyploid Coffea arabica improved the transcriptional homeostasis over that of both diploid parents. *Plant Cell Physiol.* **56**, 2035–2051 (2015).

8. Lloyd, A. & Bomblies, K. Meiosis in autopolyploid and allopolyploid Arabidopsis. *Curr. Opin. Plant Biol.* **30**, 116–122 (2016).

9. Rego-Costa, A., Débarre, F. & Chevin, L.-M. Chaos and the (un)predictability of evolution in a changing environment. *Evolution* **72**, 375–385 (2018).

10. de Oliveira, R. R., Cesarino, I., Mazzafera, P. & Dornelas, M. C. Flower development in Coffea arabica L.: New insights into MADS-box genes. *Plant Reprod.* **27**, 79–94 (2014).

11. Davis, A. P. *et al.* High extinction risk for wild coffee species and implications for coffee sector sustainability. *Sci. Adv.* **5**, eaav3473 (2019).

12. de Oliveira, R. R. *et al.* Elevated temperatures impose transcriptional constraints on coffee genotypes and elicit intraspecific differences in thermoregulation. *Front. Plant Sci.* **11**, 2020.03.07.981340 (2020).

13. López, M. E. *et al.* An overview of the endogenous and environmental factors related to the *Coffea arabica* flowering process. *Beverage Plant Res.* **1**, 1–16 (2021).

14. De Camargo, Â. P. & De Camargo, M. B. P. Definição e Esquematização das Fases Fenológicas do Cafeeiro Arábica nas Condições Tropicais do Brasil. *Bragantia* **60**, 65–68 (2001).

15. Morais, H., Caramori, P. H., Koguishi, M. S. & De Arruda Ribeiro, A. M. Escala

fenológica detalhada da fase reprodutiva de coffea arabica. *Bragantia* **67**, 257–260 (2008).

16. Mathioni, S. M., Kakrana, A. & Meyers, B. C. Characterization of Plant Small RNAs by Next Generation Sequencing. *Curr. Protoc. Plant Biol.* **2**, 39–63 (2017).

17. German, M. A., Luo, S., Schroth, G., Meyers, B. C. & Green, P. J. Construction of Parallel Analysis of RNA Ends (PARE) libraries for the study of cleaved miRNA targets and the RNA degradome. **4**, 356–362 (2009).

18. de Souza Gomes, M., Muniyappa, M. K., Carvalho, S. G., Guerra-Sá, R. & Spillane, C. Genome-wide identification of novel microRNAs and their target genes in the human parasite Schistosoma mansoni. *Genomics* **98**, 96–111 (2011).

19. Kozomara, A., Birgaoanu, M. & Griffiths-Jones, S. miRBase: from microRNA sequences to function. *Nucleic Acids Res.* **47**, D155–D162 (2019).

20. Hammond, R. K., Gupta, P., Patel, P. & Meyers, B. C. miRador: a fast and precise tool for the prediction of plant miRNAs. *bioRxiv* 2021.03.24.436803 (2021) doi:10.1101/2021.03.24.436803.

21. Axtell, M. J. & Meyers, B. C. Revisiting criteria for plant microRNA annotation in the era of big data. *Plant Cell* **30**, 272–284 (2018).

22. Johns Hopkins University. Coffea arabica V. Caturra Genome. (2018).

23. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).

24. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

25. Axtell, M. J. ShortStack: Comprehensive annotation and quantification of small RNA genes. *Rna* **19**, 740–751 (2013).

26. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration | Briefings in Bioinformatics | Oxford Academic. https://academic.oup.com/bib/article/14/2/178/208453?login=true.

27. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).

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

29. Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* **4**, (2005).

30. Csardi, G. & Nepusz, T. The igraph software package for complex network research.

InterJournal Complex Syst. **1695**, 1–9 (2006).

31. Kakrana, A., Hammond, R., Patel, P., Nakano, M. & Meyers, B. C. SPARTA: A parallelized pipeline for integrated analysis of plant miRNA and cleaved mRNA data sets, including new miRNA target-identification software. *Nucleic Acids Res.* **42**, 1–13 (2014).

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

33. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

34. Anders, S., Pyl, P. T. & Huber, W. HTSeq - A Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).

35. Yan, H. *et al.* agriGO v2.0: a GO analysis toolkit for the agricultural community, 2017 update. *Nucleic Acids Res.* **45**, W122–W129 (2017).

36. Cuperus, J. T., Fahlgren, N. & Carrington, J. C. Evolution and functional diversification of MIRNA genes. *Plant Cell* **23**, 431–442 (2011).

37. Allen, E., Xie, Z., Gustafson, A. M. & Carrington, J. C. microRNA-Directed Phasing during Trans-Acting siRNA Biogenesis in Plants. *Cell* **121**, 207–221 (2005).

38. Fernandes-Brum, C. N. *et al.* A genome-wide analysis of the RNA-guided silencing pathway in coffee reveals insights into its regulatory mechanisms. *PLOS ONE* **12**, e0176333 (2017).

39. Zhu, Q.-H. *et al.* miR482 Regulation of NBS-LRR Defense Genes during Fungal Pathogen Infection in Cotton. *PLoS ONE* **8**, e84390 (2013).

40. Fahlgren, N. *et al.* MicroRNA Gene Evolution in Arabidopsis lyrata and Arabidopsis thaliana. *Plant Cell* **22**, 1074–1089 (2010).

41. Koch, M. A., Haubold, B. & Mitchell-Olds, T. Comparative Evolutionary Analysis of Chalcone Synthase and Alcohol Dehydrogenase Loci in Arabidopsis, Arabis, and Related Genera (Brassicaceae). *Mol. Biol. Evol.* **17**, 1483–1498 (2000).

42. Wikström, N., Savolainen, V. & Chase, M. W. Evolution of the angiosperms: calibrating the family tree. *Proc. R. Soc. B Biol. Sci.* **268**, 2211–2220 (2001).

43. Guyot, R. *et al.* Ancestral synteny shared between distantly-related plant species from the asterid (Coffea canephora and Solanum Sp.) and rosid (Vitis vinifera) clades. *BMC Genomics* **13**, 103 (2012).

44. Zhai, J. *et al.* MicroRNAs as master regulators of the plant NB-LRR defense gene family via the production of phased, trans-acting siRNAs. *Genes Dev.* **25**, 2540–2553 (2011).

45. Baldrich, P. *et al.* The Evolutionary History of Small RNAs in the Solanaceae. 2021.05.26.445884 Preprint at https://doi.org/10.1101/2021.05.26.445884 (2022).

46. Pokhrel, S. & Meyers, B. C. Heat-responsive microRNAs and phased small interfering RNAs in reproductive development of flax. *Plant Direct* **6**, (2022).

47. Xia, R. *et al.* 24-nt reproductive phasiRNAs are broadly present in angiosperms. *Nat. Commun.* **10**, 627 (2019).

48. Pokhrel, S., Huang, K. & Meyers, B. C. Conserved and non-conserved triggers of 24-nucleotide reproductive phasiRNAs in eudicots. *Plant J.* **107**, 1332–1345 (2021).

49. Polydore, S., Lunardon, A. & Axtell, M. J. Several phased siRNA annotation methods can frequently misidentify 24 nucleotide siRNA-dominated *PHAS* loci. *Plant Direct* **2**, (2018).

50. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl Acids Res* **25**, (1997).

51. Bailey, T. L. & Elkan, C. Fitting a Mixture Model By Expectation Maximization To Discover Motifs In Biopolymer. 9.

52. Rodrigues, J. A. *et al.* Imprinted expression of genes and small RNA is associated with localized hypomethylation of the maternal genome in rice endosperm. *Proc. Natl. Acad. Sci.* **110**, 7934–7939 (2013).

53. Grover, J. W. *et al.* Abundant expression of maternal siRNAs is a conserved feature of seed development. *Proc. Natl. Acad. Sci.* **117**, 15305–15315 (2020).

54. Burgess, D., Chow, H. T., Grover, J. W., Freeling, M. & Mosher, R. A. Ovule siRNAs methylate protein-coding genes in trans. *BioRxiv* 2021–06 (2022).

55. Chen, H.-M., Li, Y.-H. & Wu, S.-H. Bioinformatic prediction and experimental validation of a microRNA-directed tandem trans-acting siRNA cascade in Arabidopsis. *Proc. Natl. Acad. Sci.* **104**, 3318–3323 (2007).

56. Cardon, C. H. *et al.* Expression of coffee florigen CaFT1 reveals a sustained floral induction window associated with asynchronous flowering in tropical perennials. *Plant Sci.* **325**, 111479 (2022).

57. Martinez, G. *et al.* Paternal easiRNAs regulate parental genome dosage in Arabidopsis. *Nat. Genet.* **50**, 193–198 (2018).

58. Gao, J. *et al.* A robust mechanism for resetting juvenility during each generation in Arabidopsis. *Nat. Plants* **8**, 257–268 (2022).

59. Newman, M. E. J. Modularity and community structure in networks. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 8577–8582 (2006).

60. Majerowicz, N. & Söndahl, M. R. Induction and differentiation of reproductive buds in Coffea arabica L. *Braz. J. Plant Physiol.* **17**, 247–254 (2005).

61. Rossmann, S. *et al.* Mutations in the miR396 binding site of the growth-regulating factor gene VvGRF4 modulate inflorescence architecture in grapevine. *Plant J.* **101**, 1234–1248 (2020).

62. Kang, I.-H., Steffen, J. G., Portereiko, M. F., Lloyd, A. & Drews, G. N. The AGL62 MADS Domain Protein Regulates Cellularization during Endosperm Development in Arabidopsis. *Plant Cell* **20**, 635–647 (2008).

63. Hoffmann, T. *et al.* The identification of type I MADS box genes as the upstream activators of an endosperm-specific invertase inhibitor in Arabidopsis. *BMC Plant Biol.* **22**, 18 (2022).

64. Debernardi, J. M., Woods, D. P., Li, K., Li, C. & Dubcovsky, J. MiR172-APETALA2-like genes integrate vernalization and plant age to control flowering time in wheat. *PLOS Genet.* **18**, e1010157 (2022).

65. Kim, J. J. *et al.* The microRNA156-SQUAMOSA PROMOTER BINDING PROTEIN-LIKE3 Module Regulates Ambient Temperature-Responsive Flowering via FLOWERING LOCUS T in Arabidopsis. *Plant Physiol.* **159**, 461–478 (2012).

66. Serivichyaswat, P. *et al.* Expression of the Floral Repressor miRNA156 is Positively Regulated by the AGAMOUS-like Proteins AGL15 and AGL18. *Mol. Cells* **38**, 259–266 (2015).

67. Adamczyk, B. J., Lehti-Shiu, M. D. & Fernandez, D. E. The MADS domain factors AGL15 and AGL18 act redundantly as repressors of the floral transition in Arabidopsis. *Plant J.* **50**, 1007–1019 (2007).

68. Yu, S. *et al.* Gibberellin Regulates the Arabidopsis Floral Transition through miR156-Targeted SQUAMOSA PROMOTER BINDING–LIKE Transcription Factors. *Plant Cell* **24**, 3320–3332 (2012).

69. Xu, Y. L., Gage, D. A. & Zeevaart, Jad. Gibberellins and Stem Growth in Arabidopsis thaliana (Effects of Photoperiod on Expression of the GA4 and GA5 Loci). *Plant Physiol.* **114**, 1471–1476 (1997).

70. Yant, L. *et al.* Orchestration of the Floral Transition and Floral Development in Arabidopsis by the Bifunctional Transcription Factor APETALA2. *Plant Cell* **22**, 2156–2170 (2010).

71. APETALA2/Ethylene Responsive Factor (AP2/ERF) transcription factors: mediators of stress responses and developmental programs - Licausi - 2013 - New Phytologist - Wiley Online Library. https://nph.onlinelibrary.wiley.com/doi/full/10.1111/nph.12291.

72. Nakano, T., Suzuki, K., Fujimura, T. & Shinshi, H. Genome-Wide Analysis of the ERF Gene Family in Arabidopsis and Rice. *Plant Physiol.* **140**, 411–432 (2006).

73. Xie, Z., Nolan, T. M., Jiang, H. & Yin, Y. AP2/ERF Transcription Factor Regulatory Networks in Hormone and Abiotic Stress Responses in Arabidopsis. *Front. Plant Sci.* **10**, (2019).

74. Abiri, R. *et al.* Role of ethylene and the APETALA 2/ethylene response factor superfamily in rice under various abiotic and biotic stress conditions. *Environ. Exp. Bot.* **134**, 33–44 (2017).

75. Jofuku, K. D., Boer, B. G. W. den, Van Montagu, M. & Okamuro, J. K. Control of Arabidopsis Flower and Seed Development by the Homeotic Gene APETALA2. *Plant Cell* **6**, 1211–1225 (1994).

76. Drews, G. N., Bowman, J. L. & Meyerowitz, E. M. Negative regulation of the Arabidopsis homeotic gene AGAMOUS by the APETALA2 product. *Cell* **65**, 991–1002 (1991).

77. Krogan, N. T., Hogan, K. & Long, J. A. APETALA2 negatively regulates multiple floral organ identity genes in Arabidopsis by recruiting the co-repressor TOPLESS and the histone deacetylase HDA19. *Dev. Camb. Engl.* **139**, 4180–4190 (2012).

78. Zhu, Q.-H. & Helliwell, C. A. Regulation of flowering time and floral patterning by miR172. *J. Exp. Bot.* **62**, 487–495 (2011).

79. Houben, M. & Van de Poel, B. 1-aminocyclopropane-1-carboxylic acid oxidase (ACO): The enzyme that makes the plant hormone ethylene. *Frontiers in Plant Science* vol. 10 (2019).

80. Zhang, Z., Schofield, C. J., Baldwin, J. E., Thomas, P. & John, P. Expression, purification and characterization of 1-aminocyclopropane-1-carboxylate oxidase from tomato in Escherichia coli. *Biochem. J.* **307**, 77–85 (1995).

81. Pattyn, J., Vaughan-Hirsch, J. & Poel, B. V. de. The regulation of ethylene biosynthesis: a complex multilevel control circuitry. *New Phytol.* **229**, 770–782 (2021).

82. Lima, A. A. *et al.* Drought and re-watering modify ethylene production and sensitivity, and are associated with coffee anthesis. *Environ. Exp. Bot.* **181**, 104289 (2021).

83. Petroni, K. *et al.* The Promiscuous Life of Plant NUCLEAR FACTOR Y Transcription Factors. *Plant Cell* **24**, 4777–4792 (2012).

84. Ben-Naim, O. *et al.* The CCAAT binding factor can mediate interactions between CONSTANS-like proteins and DNA. *Plant J.* **46**, 462–476 (2006).

85. Kumimoto, R. W. *et al.* The Nuclear Factor Y subunits NF-YB2 and NF-YB3 play additive roles in the promotion of flowering by inductive long-day photoperiods in Arabidopsis. *Planta* **228**, 709–723 (2008).

86. Urrutia, R. KRAB-containing zinc-finger repressor proteins. *Genome Biol.* **4**, 1–8 (2003).

87. Taoka, K. *et al.* 14-3-3 proteins act as intracellular receptors for rice Hd3a florigen. *Nature* **476**, 332–335 (2011).

88. Bertrand, E. & Fournier, M. J. *The snoRNPs and Related Machines: Ancient Devices That*

Mediate Maturation of rRNA and Other RNAs. *Madame Curie Bioscience Database [Internet]* (Landes Bioscience, 2013).

89. Meyers, B. C., Kaushik, S. & Nandety, R. S. Evolving disease resistance genes. *Curr. Opin. Plant Biol.* **8**, 129–134 (2005).

90. Ribeiro, T. H. C. *et al.* Transcriptome analyses suggest that changes in fungal endophyte lifestyle could be involved in grapevine bud necrosis. *Sci. Rep.* **10**, 9514 (2020).

91. Bélanger, S., Pokhrel, S., Czymmek, K. & Meyers, B. C. Premeiotic, 24-nucleotide reproductive phasiRNAs are abundant in anthers of wheat and barley but not rice and maize. *Plant Physiol.* **184**, 1407–1423 (2020).

92. Zhai, J. *et al.* Spatiotemporally dynamic, cell-type-dependent premeiotic and meiotic phasiRNAs in maize anthers. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 3146–3151 (2015).

93. Arikit, S. *et al.* An atlas of soybean small RNAs identifies phased siRNAs from hundreds of coding genes. *Plant Cell* **26**, 4584–4601 (2014).

94. Xia, R., Xu, J., Arikit, S. & Meyers, B. C. Extensive Families of miRNAs and PHAS Loci in Norway Spruce Demonstrate the Origins of Complex phasiRNA Networks in Seed Plants. *Mol. Biol. Evol.* **32**, 2905–2918 (2015).

95. de Vries, S., Kloesges, T. & Rose, L. E. Evolutionarily Dynamic, but Robust, Targeting of Resistance Genes by the miR482/2118 Gene Family in the Solanaceae. *Genome Biol. Evol.* **7**, 3307–3321 (2015).

96. MicroRNA Superfamily Regulates Nucleotide Binding Site–Leucine-Rich Repeats and Other mRNAs | The Plant Cell | Oxford Academic. https://academic.oup.com/plcell/article/24/3/859/6097249.

97. Chakraborty, J., Ghosh, P. & Das, S. Autoimmunity in plants. *Planta* **248**, 751–767 (2018).

98. Ralstonia solanacearum Extracellular Polysaccharide Is a Specific Elicitor of Defense Responses in Wilt-Resistant Tomato Plants | PLOS ONE. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0015853.

99. Kufel, J. & Grzechnik, P. Small Nucleolar RNAs Tell a Different Tale. *Trends Genet.* **35**, 104–117 (2019).

100. Falaleeva, M. & Stamm, S. Processing of snoRNAs as a new source of regulatory non-coding RNAs. *BioEssays* **35**, 46–54 (2013).

4. General Conclusions

The coffee genome and its multiple transcriptome facets can be abstracted as a complex maze of interactions. Even then, we know that this perception is an oversimplification. During the writing of bioinformatic scripts and this thesis manuscript we could not dispel the urge of comparing the genome, and its transcriptome counterpart, to a computer operating system calling programs along its runtime.

In chapter one, I describe a story about an ancestral merging of two genomes. This was a rare event, probably a once in history incident. It is known that many crops are allopolyploids and hybrid polyploid species of economic importance can be artificially produced. Nevertheless, the low genetic diversity of the *C. arabica* genome suggests that there are complications for allopolyploid establishment in this genus. One of the complications we suppose had to be overcome was the duplication of the gene expression machinery.

Most computers require a kernel as a fundamental part of its operating system. This kernel is a system to allocate Random Access Memory (RAM) and execute code in an organized way. In other words, the kernel is a code to organize the running of multiple other codes simultaneously. In our plant/computer analogy the kernel is analog to transcription and its regulation. Well, when we found out that most of the RNA processing machinery codebase was missing from *C. eugenioides* sub genome - in the form of missing universal orthologs - we could not help but to think that the plant operating system's kernel had to be deduplicated to avoid corrupted memory faulting. So, we believe that one of the initial steps that allowed the establishment of *C. arabica* was the simplification of the kernel - one of the inherited copies had to go away as it was the *C. canephora* version. It seems that the choice of the name "*eugenioides" was* somewhat prophetic.

In the long run, polyploids tend to simplify back to diploidy. If *C. arabica* will follow this tendency it is yet unknown. Nevertheless, along hundreds of million years of evolution all flowering plants were, at some point, a polyploid. In chapters two and three, I showed that some genes were lost or duplicated in the sub-genomes. POLYPHENOL OXIDASES (PPOs) are important enzymes involved in defense responses and other processes (Thipyapong et al., 2004). In *C. arabica,* we found copies of this enzyme being encoded by five loci in chromosome 5 that was inherited from *C. canephora* and two additional copies in the respective chromosome 2 of both sub-genomes. The expansion of this gene may be advantageous. Meanwhile, members of miRNA families are mostly equally distributed between the sub-genomes suggesting that they are dosage-sensitive genes.

If we can’t precisely understand what is going on, the future is yet more opaque. So, why not try to speed up change? What would happen if we try to “debug” the *Coffea arabica* genomic code-base? Simplify to the point where it becomes a diploid? Try to recreate the allopolyploid, but this time giving the RNA processing machinery to *C. canephora*? If we know that genetic diversity is scarce, the creation of novelty may be our way forward.

REFERÊNCIAS

Cardon, C. H. *et al.* Expression of coffee florigen CaFT1 reveals a sustained floral induction window associated with asynchronous flowering in tropical perennials. *Plant Sci.* **325**, 111479 (2022).

Childe, V. G. *et al.* Man makes himself. *Sci. Soc.* **4**, (1940).

Denoeud, F. *et al.* The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. *science* **345**, 1181–1184 (2014).

Lashermes, P. *et al.* Molecular characterisation and origin of the Coffea arabica L. genome. *Mol. Gen. Genet. MGG* **261**, 259–266 (1999).

Meyer, F. G. FAO coffee mission to Ethiopia, 1964-1965. (1968).

Scalabrin, S. *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.* **10**, 1–13 (2020).

Thipyapong, P., Hunt, M. D. & Steffens, J. C. Antisense downregulation of polyphenol oxidase results in enhanced disease susceptibility. *Planta* **220**, 105–117 (2004).

Waller, J. M., Bigger, M. & Hillocks, R. J. *Coffee pests, diseases and their management.* (CABI, 2007).