UNIVERSIDADE FEDERAL DE LAVRAS

RENAN TERASSI PINTO

GENOMIC ANALYSES OF CAFFEINE METABOLISM AND PERSPECTIVES FOR GENE EDITING ON *Coffea canephora*

LAVRAS-MG
2020

RENAN TERASSI PINTO

GENOMIC ANALYSES OF CAFFEINE METABOLISM AND PERSPECTIVES FOR GENE EDITING ON *Coffea canephora*

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-graduação em Biotecnologia Vegetal, para obtenção do título de Doutor.

Luciano Vilela Paiva, PhD
Orientador

Prof. Dr. Vagner Augusto Benedito, PhD
Co-orientador

LAVRAS-MG
2020

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

Pinto, Renan Terassi.
Genomic analyses of caffeine metabolism and perspectives for gene editing on *Coffea canephora* / Renan Terassi Pinto. - 2020.
101 p. : il.

Orientador(a): Luciano Vilela Paiva.
Coorientador(a): Vagner Augusto Benedito.
Tese (doutorado) - Universidade Federal de Lavras, 2020.
Bibliografia.

1. Melhoramento genético. 2. CRISPR-Cas9. 3. Transportadores de membrana. I. Paiva, Luciano Vilela. II. Benedito, Vagner Augusto. III. Título.

RENAN TERASSI PINTO

ANALISES GENÔMICAS RELACIONADAS AO METABOLISMO DE CAFEÍNA E PERSPECTIVAS PARA A EDIÇÃO GÊNICA EM *Coffea canephora*

GENOMIC ANALYSES OF CAFFEINE METABOLISM AND PERSPECTIVES FOR GENE EDITING ON *Coffea canephora*

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-graduação em Biotecnologia Vegetal, para obtenção do título de Doutor.

APROVADA em 20 de Agosto de 2020.
Dr. Alan Carvalho Andrade - EMBRAPA
Dr. Mirian Peres Maluf - EMBRAPA
Dr. Raphael Ricon de Oliveira - EMBRAPA

Luciano Vilela Paiva, PhD
Orientador

Vagner Augusto Benedito, PhD
Co-orientador

LAVRAS – MG
2020

AKNOWLEDGEMENTS

In first place, I thank my parents, Eliana Terassi Pinto and Edilson Aparecido Pinto for being always my solid foundation and encouraging me to pursue my dreams. They were the first who taught me the value of knowledge, on the most pure way, as the only thing on life which I could keep with me during the journey. I was only able to follow my aspirations because I've always know I had their absolute love and support.

I am thankful to my wife Glicia. She is my person during this journey, the one I've chosen to stay together with for all the experiences we will live. Her love and all the companionship that comes with it turns the life into pleasant adventure and inspires me to be the best version of myself.

I want to thank my family, in general terms, for the unconditional support through life, for ever being my home on this world.

I thank to all my friends who helped me growing through the graduate years. I want to point out the companionship of Thiago, Wesley, Natália, Carlos, Ricardo, Farley, Adolfo and Fabiana, who significantly contributed to my scientific and personal growth during these years. I hope I can share many more coffee (or beer) breaks with you guys, which I truly believe is fundamental for science progress. I want to highlight the ever present and pleasant companionship of Thiago on these scientific discussions, which pushed me to always have strong arguments to beat him, rsrs.

I am thankful to Dr. Luciano Vilela Paiva for, beyond giving the opportunity to conduct my PhD under his advisement, believing in my scientific career and having patience to let me try out many ideas which I though relevant during these years. I am thankful for the many moments he pushed me and my colleagues to pursue our professional aspirations on motivational conversations.

I thank to Dr. Vagner Augusto Benedito, who embraced me as a graduate student, being essential for my professional growth during PhD. But also, I am grateful for the careful attention he had to me and Glicia during our US-living year, which made this experience possible and much more pleasant. His scientific and personal support enriched my PhD degree and has set to me an example to be followed.

I thank to Dr. Alan Carvalho Andrade for all the scientific and personal support since the Master's degree. The insightful coffee moments we had, together with Thiago, truly inspired me to be a better scientist and helped shaping the way I want to do science from now on. I hope in the future I can be a scientist of whom he will be proud of.

I am grateful to Dr. Breno Régis Santos and all the co-workers at BIOGEN, where I've started my scientific career as an undergrad student. Thank you, Dr. Breno, for believing in me and for opening the doors of the scientific world to me.

I thank CAPES for the scholarships provided, which made all this possible. Also, I am thankful to Universidade Federal de Lavras and Programa de Pós-graduação em Biotecnologia Vegetal for the opportunity of doing my Master's and PhD degrees and all the pleasant moments that happened in consequence.

I thank West Virginia University for having me as a visiting scholar and make me feeling as I belonged there. I am grateful to all the WVU staff, the professors I've met and the laboratory co-workers I've had under Dr. Vagner advisory. In special, I want to point out Estefania Tavares, Korina, Anathi Magdella, Lina Yang and Gabriel Lasmar. I want to highlight the important companionship of Gabriel during this period, the best roommate we could ever have.

I want to thank all the people not highlighted here, whom somehow participated on this journey. We cannot do anything alone, especially in science.

I am grateful, in addition to all exposed here, to the circumstances and contributions which goes beyond my comprehension and somehow could have led me towards this path.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001. I am grateful to CAPES for the grant to conduct the sandwich PhD (Programa de Doutorado Sanduíche no Exterior (PDSE), process number: 88881.190558/2018-01).

“Standing on the shoulders of giants”

(Various authors)

ABSTRACT

Humans cultivate arboreal plants for food production and industrial raw material since the early days of our society. This co-evolutionary process led to the relationship we have today with the tree crops, but the domestication course is hard to track for these species due to frequent hybridization events and the tendency of clonal propagation performed by farmers to preserve the selected phenotypes and, for modern agriculture using arboreal plants, the biggest challenge is breeding. The longer juvenile phase of these plants impairs crossing-related activities, which in turns is the basis for gene introgressions, development of segregating populations for uncovering genetic basis of traits, construction of genetic maps, among others. In face of the climate changes and resources scarcity forecast, it is important to improve the ways of tailoring tree crops and one alternative is using gene editing technology for rational molecular design. Among these arboreal plants, coffee tree is one of the most explored crops, as coffee is worldwide appreciated. Due to the upper mentioned forecast, a diploid species and generally more adapted to higher temperatures and drought like *C. canephora* could be the focus of coffee breeding programs to sustain coffee production chain. An interesting perspective to this end should consider, initially, using CRISPR-Cas-based gene editing to modify biochemical characteristics associated with coffee beverage quality and surpass self-incompatibility on *C. canephora*. Among the aspects which influences metabolite accumulation, is the membrane transporters activity and here we identified 1,847 potentially transporter-coding genes through a comprehensive genomic strategy and pointed some of them as possibly related to diterpenes, chlorogenic acids and alkaloid accumulation by gene co-expression analyses using all the public transcriptomic data for coffee. This inventory is the first of its kind constructed for coffee as well as for any tree species. One determinant substance for coffee consumption and for the plant development is caffeine but, despite its importance, little is known about the genetic bases of its accumulation on cells or the regulation of its synthesis. By using field grown *C. canephora* plants, we performed caffeine quantification on two leaf development stages as well as six points of fruit maturation. We identified putative membrane transporters and transcription factor genes co-expressed with caffeine synthesis-related ones using public RNA-seq data and analyzed its gene expression through RT-qPCR on these caffeine content-contrasting samples. Due to its potential involvement on caffeine transport and synthesis regulation, some identified genes should be further explored and a vector for CRISPR-mediated gene editing was already developed. The analyses performed here help on expanding the knowledge on genetics of caffeine accumulation in coffee and this might be part of the basis for varied rational molecular design strategies, which are essential for preparing agriculture for the challenges we expect to come.

Keywords: Breeding. CRISPR-Cas9.Caffeine.Membrane transporter. Coffee

RESUMO

Humanos cultivam plantas arbóreas para produção de alimentos e matérias-primas industriais desde os primeiros dias da nossa sociedade. Esse processo co-evolucionário culminou na atual relação com as plantas lenhosas cultivadas, porém o processo de domesticação destas espécies é difícil de mapear devido à frequencia dos eventos de hibridização e à tendência de utilização da propagação clonal para preservar fenótipos selecionados. Para a utilização destas espécies na agricultura moderna, o maior desafio é desenvolver programas melhoramento. A fase juvenil extensa dificulta a introgressão de genes, o desenvolvimento de populações segregantes para identificação de bases genéticas de caracteres, construção de mapas genéticos, dentre outros aspectos. Em decorrência das previsões de mudanças climáticas e escassez de recursos, é importante aprimorar a adaptabilidade destas plantas e, dentre as alternativas para este fim, destaca-se a utilização de edição genética. O café, produto do cultivo de espécies lenhosas do gênero *Coffea*, é mundialmente consumido e uma espécie diplóide, de forma geral, mais adptada à altas temperaturas e escasses de água como *C. canephora* pode ser o foco de programas de melhoramento genético para manutenção desta cadeia produtiva. Uma perspectiva interessante para este fim deveria considerar a utilização de edição gênica via CRISPR-Cas9 para a modificação de características bioquímicas associadas à qualidade da bebida e para superar a auto-incompatibilidade em *C. canephora*. A atividade de transportadores de membrana (TM) está entre os aspectos que influenciam o acúmulo de metabólitos e, neste trabalho, foram identificados 1.847 genes que potencialmente codificam estas proteínas por meio de uma estratégia genômica e alguns foram destacados como possilvemente relacionados ao acúmulo de diterpenos, ácidos clorogênicos e alcalóides por meio de análises de co-expressão gênica utilizando dados transcriptômicos públicos. Este é o primeiro inventório de proteínas de membrana desenvolvido para o cafeeiro, bem como para qualquer espécie arbórea. Uma das substâncias determinantes tanto para o consumo de café quanto para o desenvolvimento da planta é a cafeína porém, apesar dessa importância, há pouco conhecimento sobre as bases genéticas para seu acúmulo nas células bem como sobre a regulação de sua síntese. Utilizando plantas em campo, a concentração de cafeína foi avaliada em folhas de dois estágios de desenvolvimento e em frutos de diferentes graus de maturação. Foram identificados possíveis genes codificantes de TM e de fatores de transcrição co-expressos com genes relacionados à síntese de cafeína, utilizando dados de RNA-seq públicos, e sua expressão gênica foi avaliada via RT-qPCR nas amostras contrastantes em termos de teor de cafeína. Devido ao potencial envolvimento no transporte de cafeína e na regulação da síntese, alguns genes identificados podem ser futuramente explorados e um vetor para edição gênica via CRISPR-Cas9 já foi desenvolvido. As análises executadas neste trabalho auxiliam na expansão do conhecimento genético relacionado ao acúmulo de cafeína no cafeeiro e podem compor estratégias de design molecular racional, essenciais para aprimorar a agricultura frente aos desafios previstos.

Palavras-chave: Melhoramento genético. CRISPR-Cas9. Transportadores de membrana. Cafeeiro.

SUMMARY

PART I – GENERAL PRESENTATION

1 GENERAL INTRODUCTION

Agriculture is a fundamental activity of human society which allowed for the development of civilizations, as well as other activities such as science, art and politics, or anything that depends on a sedentary-structured human organization; and it is still the source of food and some raw materials for industry nowadays. In this way, plant and animal domestication, as the basis of this essential activity, was one the most significant cultural and evolutionary transitions on human history (LARSON *et al*., 2014).

Under this perspective, plant domestication can be defined as a co-evolutionary process, resulted from the cultivation of wild plants leading to specialization and/or the rise of new populations better fitted to human survival (PURUGGANAN *et al*., 2019). Looking from the plant side of the prism, the beginning of domestication process is an evolutionary response to anthropogenic ecosystem, involving several adaptations towards the recruitment of humans as dispersers (SPENGLER, 2020).

We can estimate that this prolonged co-evolution process from which agriculture has risen may have started about 10,000 years ago, although for different plant species and human societies this timing can vary, as multiple geographically and chronologically start points happened (FULLER *et al*., 2014). Exploring plant domestication can provide information of technological and cultural shifts in our society, but also, very important information about the evolution of many agronomically relevant traits, which is a powerful resource for plant breeding.

A significant part of the plant species that we use on modern agriculture are arboreal-shaped, having an elongated stem, woody structure and perennial growth habit (NEALE *et al*., 2017) and, unfortunately, these plants that are difficult to breed are some of the least explored in terms of domestication studies and, sometimes, even the process have slight differences from the standard seem for annual crops, as many tree crops varieties derived from hybridizations and vegetative propagation, often with unfixed introgressions (SPENGLER *et al*., 2019).

These plants will be referred here as tree crops and the challenge associated to them is the difficult on performing breeding compared to annual crops, coupled with the longer time period which these species have to endure on the field, facing increasingly instabilities both resources and climate-related as well as the rise of pest/diseases during the cultivation period. Therefore, the goals of a tree crop breeding program need to be carefully established, considering current demands, but also future forecasts and alternative ways to accelerate genetic improvements must be explored.

Despite the longer juvenile phase that enlarge crossing cycles, genetic maps for some of the main tree crops are developed to a level that allows for genome projects to be conducted, as pointed further in this work, and genomic selection to be performed (ALKIMIN *et al*., 2020). Beyond assisting conventional breeding, genetic knowledge and genomic information can be directly applied to gene editing strategies and there are already about 25 published works (gathered here) related to this approach on tree crops up to this date, using CRISPR-Cas technology (JINEK *et al*., 2012) to generate the targeted gene modifications.

Combining advancements on breeding techniques with the expanding capacity of CRISPR-based gene editing, as well as improvements on big data mining, we may be able to rationally design crops in the near future, reaching a fourth generation of crop improvement when varieties will be more precisely tailored towards society and environment demands (FERNIE and YAN, 2019). Some interesting approaches for using CRISPR-Cas technology were published recently, which will be pointed here, could greatly benefit tree crops breeding programs.

Here, we argue that *C. canephora*, the second major species used for coffee production, could be greatly improved through application of these new technologies, unleashing its potential to sustain coffee production chain. This species is one of the diploid progenitors of the allotetraploid *Coffea arabica* (SCALABRIN *et al*., 2020), which cultivation accounts for about 60% of coffee production (ICO, 2020). A simpler ploidy is an important characteristic for plant breeding and genome exploration that facilitates the association of major genes or genetic variants to target traits.

Despite some level of intraspecific variability, *C. canephora* is usually more adapted to higher temperatures and lower water supply than *C. arabica*, besides being a source of resistance to root-knot nematodes that impairs the allotetraploid species cultivation (BERTRAND *et al*., 2001; FATOBENE *et al*., 2018; SALGADO *et al*., 2019). But, there is a big concern about the acceptance of sensorial characteristics of the *C. canephora*-derived beverage as it usually differs from the standard (LEMOS *et al*., 2020; LEBOT *et al*., 2020), moreover, its gametophytic self-incompatibility (LASHERMES *et al*., 1996) hampers the development of homogenous varieties and plantations.

To give an alternative perspective for coffee tree breeding, we propose that a small set of genetic modifications discussed hereafter could be achieved by performing genomic analyses and CRISPR-based gene editing. Importantly, it is not this work’s intention to trivialize the challenge of reaching this goal, or to minimize the necessary efforts to obtain the genetic knowledge needed to perform a strategy based on multiple genetic changes on a tree crop.

From the set of efforts towards turning possible a rational molecular design of a *C. canephora* variety for coffee production, exploring its published genome (DENODEUD *et al*., 2014) and contributing to gene annotation is a significant preliminary step. To this end, a genome-wide annotation of *C. canephora* membrane proteins is provided in this work, with the intention to elucidate the membrane transporters codifying genes.

Membrane transporters are found in high numbers on plant genomes, most probable due to diversification of specialized metabolism on this kingdom (JORGENSEN *et al*., 2017) and, for a crop which the biochemical profile is the main aspect affecting consumer acceptance, the knowledge about this class of proteins is essential as they are important determinants for plant metabolite content (*i.e.* transporter-mediated vacuolar accumulation) (MARTINOIA *et al*., 2018).

We have screened the *C. canephora* genome and found about 10,000 genes that potentially codify membrane-bounded proteins. Following a comprehensive strategy, we classified 1,847 proteins as membrane transporters from 196 different protein families, based on TCDB database classification (SAIER *et al*., 2016). Along with this work, we provide an inventory with main characteristics of all these 10,000 genes.

Taking advantage of an already observed pattern in which transporters can be co-expressed (transcriptionally) with the enzymes that synthetize its substrates (DOBRITZSCH et al., 2016; PAYNE et al., 2017; DERMUTAS et al., 2019) and previous genetic knowledge about some biosynthesis routes for important substances for coffee (DENOEUD *et al*., 2014; MIZUNO *et al*., 2014; SANT'ANA *et al*., 2017; MAHESH *et al*., 2007; LALLEMAND *et al*., 2012), we performed gene co-expression analyses to speculate possible genes as potential transporters of these metabolites.

To this end, we used all the public transcriptomic data found on NCBI-SRA database for coffee (151 RNA-seq libraries) and we provide the expression profile of the 1,847 genes on this dataset (transcript per million). We found interesting associations of some of these genes to biosynthetic enzymes, which deserves further and deeper analysis to determinate their role on influencing accumulation of major coffee substances on coffee beans.

As the most studied coffee-derived substance and a determinant factor for coffee consumption, caffeine is a metabolite which deserves primary attention on an effort to manipulate biochemical attributes of this beverage. Despite of its importance, surprisingly, little is known about the transcriptional control of its biosynthesis or its accumulation process.

To contribute on expanding this knowledge, we defined genetic candidates for influencing these two aspects of caffeine accumulation route, by genomic and transcriptomic analysis using public datasets. We first performed a brief annotation of the putative transcription factor (TF) coding genes on *C. canephora* genome (DENOEUD *et al*., 2014), resulting on 1,171 candidates from 51 TF families. From this total, we found that 51 TFs could bind on the promoter region of caffeine biosynthesis genes and among them, 24 were co-expressed with the same biosynthetic genes on public available RNA-seq experiment (PRJNA339585) with different bean developmental stages and treatment with jasmonic acid.

On a parallel approach to the membrane transporter inventory mentioned above, putative membrane transporter genes were selected from phylogenetic analysis of gene families with already characterized alkaloid transporters (reviewed on SHITAN *et al*., 2014). Filtering by protein similarity and the same above mentioned co-expression analysis, six putative transporters were selected for further exploration.

From field-grown *C. canephora* plants, leaves from two developmental stages and six time points of fruit development were collected and their caffeine content evaluated. On this contrasting material, the expression level of *CcXMT*, *CcMXMT* and *CcDXMT* was analyzed, as well as of the four selected putative transporter-coding genes and four putative TF-coding. With all these analyzes which will be further detailed, we are able to point for at least on candidate gene as coding for a potential caffeine membrane transporter and one putative TF related to this alkaloid biosynthesis.

The analyses presented here help on expanding the knowledge about *C. canephora* genomics, especially related to membrane transporters annotation as well as genes involved on caffeine accumulation process, which is an important aspect for coffee breeding. Further studies are necessary to elucidate how determinant is the influence of the genes pointed here on caffeine content variation on coffee beans.

A vector to generate a CRISPR-Cas9-mediated mutation on the putative membrane transporter-coding gene, pointed here as possibly involved on caffeine accumulation, was developed and it is exposed hereafter, as part of the work in progress to analyze the function of this putative transporter and elucidate whether this protein influences caffeine content variation on *C. canephora* tissues, especially on the seeds. We hope that this work will be part of a greater effort to deepen the knowledge needed for improving coffee breeding on a broad perspective, as well as part of a stepwise strategy to tailor *C. canephora* to sustain coffee production in face of future agriculture challenges.

REFERENCES

ALKIMIM, E. R. et al. Selective efficiency of genome-wide selection in Coffeacanephora breeding.**Tree Genetics & Genomes**, v. 16, n. 3, 2020.

BERTRAND, B.; ANTHONY, F.; LASHERMES, P.Breeding for resistance to Meloidogyneexigua in Coffeaarabica by introgression of resistance genes of Coffeacanephora. **Plant pathology**, v. 50, n. 5, p.637-643, 2001.

DE LIMA SALGADO, S. M. et al. Resistance of Conilon coffee cultivar Vitoria Incaper 8142 to Meloidogyneparanaensis under field conditions. **Experimental Agriculture**, v. 56, n.1 , p.88-93, 2020.

DEMURTAS, O.C. et al.ABCC transporters mediate the vacuolar accumulation of crocins in saffron stigmas. **The Plant Cell**, v. 31, n.11, p.2789-2804, 2019.

DENOEUD, F. et al.The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. **Science**, v. 345, n. 6201, p.1181-1184, 2014.

DOBRITZSCH, M. et al.MATE transporter-dependent export of hydroxycinnamic acid amides. **The Plant Cell**, v. 28, n.2, p.583-596, 2016.

FATOBENE, B. J.et al.Coffeacanephora clones with multiple resistance to Meloidogyne incognita and M.paranaensis. **ExpAgric**, p.1-9, 2018.

FERNIE, A. R.; YAN, J. De novo domestication: an alternative route toward new crops for the future. **Molecular plant**, v. 12, n. 5, p. 615-63, 2019.

FULLER, D.Q. et al. Convergent evolution and parallelism in plant domestication revealed by an expanding archaeological record.**Proceedings of the National Academy of Sciences**, v. 111, n.17, p.6147-6152, 2014.

JINEK, M. et al.A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity.**Science**, v. 337, n. 6096, p.816-821, 2012.

LALLEMAND, L. A. et al. A structural basis for the biosynthesis of the major chlorogenic acids found in coffee.**Plant Physiology**, v. 160, n. 1, p. 249-260, 2012.

LARSON, G. et al.Current perspectives and the future of domestication studies.**Proceedings of the National Academy of Sciences**, v. 111, n. 17, p. 6139-6146, 2014.

LASHERMES, P.; COUTURON, E.; MOREAU, N.; PAILLARD, M.; LOUARN, J., Inheritance and genetic mapping of self-incompatibility in Coffea canephora Pierre**.Theoretical and Applied Genetics,** v. 93, n.3, p. 458-462, 1996.

LEBOT, V.; MELTERAS, M.; PILECKI, A.; LABOUISSE, J.P. Chemometric evaluation of cocoa (Theobroma cacao L.) and coffee (Coffea spp.) germplasm using HPTLC.**Genetic Resources and Crop Evolution**, p.1-17, 2020.

LEMOS, M. F. et al.Chemical and sensory profile of new genotypes of Brazilian Coffeacanephora.**Food chemistry**, v. 310, p. 125850, 2020.

LIU, H. J. et al. High-Throughput CRISPR/Cas9 Mutagenesis Streamlines Trait Gene Identification in Maize. **The Plant Cell**, v. 32, n. 5, p.1397-1413,2020.

MAHESH, V. et al. Functional characterization of two p-coumaroyl ester 3′-hydroxylase genes from coffee tree: evidence of a candidate for chlorogenic acid biosynthesis. **Plant molecular biology**, v. 64, n.1-2, p.145-159, 2007.

MARTINOIA, E. Vacuolar transporters–Companions on a longtime journey.**Plant physiology**, v. 176, v. 2, p.1384-1407, 2018.

MIEULET, D. et al. Unleashing meiotic crossovers in crops.**Nature plants**, v. 4, n. 12, p.1010-1016, 2018.

MIZUNO, K. et al. Conversion of nicotinic acid to trigonelline is catalyzed by N-methyltransferase belonged to motif B′ methyltransferase family in Coffeaarabica. **Biochemical and biophysical research communications**, v. 452, p. 4, p.1060-1066, 2014.

NEALE, D. B. et al .Novel insights into tree biology and genome evolution as revealed through genomics. **Annual Review of Plant Biology,** v. 68, p.457-483, 2017.

PAYNE, R. M. et al. An NPF transporter exports a central monoterpeneindole alkaloid intermediate from the vacuole. **Nature plants**, v. 3, n.2, p.1-9, 2017.

PURUGGANAN, M.D. Evolutionary insights into the nature of plant domestication.**Current Biology**, v. 29, n. 14, p.R705-R714, 2019.

SAIER JR, M. H. et al. The transporter classification database (TCDB): recent advances. **Nucleic acids research**, v. 44, n. D1, pp. D372-D379, 2016.

SANT'ANA, G. C. et al. Genome-wide association study reveals candidate genes influencing lipids and diterpenes contents in Coffeaarabica L. **Scientific reports**, v. 8, n.1, p.1-12, 2018.

SCALABRIN, S. et al. A single polyploidization event at the origin of the tetraploid genome of Coffeaarabica is responsible for the extremely low genetic variation in wild and cultivated germplasm. **Scientific reports**, v. 10, n. 1, p.1-13, 2020.

SHITAN, N.; KATO, K.; SHOJI, T.Alkaloid transporters in plants.**Plant Biotechnology**, p.14-1002, 2014.

SPENGLER III, R.N. Anthropogenic seed dispersal: rethinking the origins of plant domestication. **Trends in Plant Science**, 2020.

SPENGLER, R.N. Origins of the apple: the role of megafaunal mutualism in the domestication of Malus and rosaceous trees. **Frontiers in plant science**, v. 10, p.617, 2019.

ZSÖGÖN, A. et al. De novo domestication of wild tomato using genome editing.Nature biotechnology, v. 36, n. 12, p.1211-1216, 2018.

PART II – SCIENTIFIC PAPERS

SCIENTIFIC PAPER 1: GOING FOR THE LONG LASTING CHANGES – THE PERSPECTIVES OF GENOME EDITING FOR TREE CROPS

ABSTRACT

Hunters and gatherers at the early days of human history took advantage on fruits growing on trees to sustain themselves, as well as our primate ancestors did. Nowadays tree crops are still broadly explored for food sources and also as raw material for many industrial sectors, but breeding these plants to face current and future agriculture demands is challenging, mostly due to the longer juvenile phase, which impairs crossing and all related activities, from gene introgressions to genomics. Due to recent advances on gene editing technology, mainly because of CRISPR-Cas systems development, it is possible to predict rationally designed crops being explored on near future, with custom genetic modifications performed to achieve desired trait variants. To this end, the major determinants are genetic knowledge, advancements on genome projects and plant tissue culture, alongside with molecular biology techniques, which are developed for some arboreal species, despite not on the same level as on annual crops. The panel of CRISPR-Cas based techniques is expanding, allowing for diverse ways of modifying DNA or RNA on living organisms, as well as the delivering methods to plants are being improved to unleash broader applicability of this system. Here we point to some recent CRISPR-Cas system improvements, interesting strategies which could be harnessed for tree crops as well as an example of application on a worldwide utilized crop, the coffee tree. With this brief perspective, we seek to rise attention for the capabilities of molecular design and to the chance we have, as a society, on creating stepwise strategies to tailor tree crops towards the environment and society next decade’s demands.

Keywords: Breeding; CRISPR-Cas9; Genome; Molecular Design; *Coffea canephora*

1 INTRODUCTION

Human society organization and maintenance is essentially dependent on plant domestication and cultivation for food supply and also for obtaining some industrial raw materials. Among the crops we use for this purpose, a significant part are arboreal plants, which are cultivated to obtain diverse products, such as fruits, wood, cellulose fiber and rubber, and its cultivation requires greater land and time investment.

Beyond the inner difficulties related to raising a tree for harvesting its fruits or other plant parts, some cultivated varieties might be just recently domesticated or even have mostly unfixed traits if hybridizations and vegetative propagation methods took place on early cultivation (SPENGLER *et al*., 2019), moreover, breeding is still at its initial development compared to annual cultivated plants like maize, wheat, rice and tomato. The main reason for this is the time-costly effort to surpass the juvenile phase, which is essential for both production and crossing associated technologies, like breeding (NEALE *et al*., 2017).

The capability to adapt a cultivated plant to human society needs is fundamental for agriculture development and, since the rise of breeding on the first green revolution; we improved our ability to understand plant genetics and also to rationally modify it. With the CRISPR-based gene editing technology and advancements on genomics we are on the imminence of a fourth generation of crop improvement, marked by design breeding and the enhanced ability to tailor crops to demand, while maintaining the sustainability of agriculture system (FERNIE and YAN, 2019).

Gene editing was already performed on some arboreal plants, such as apple tree (*Malus domestica*) (CHARRIER *et al*., 2019), pomelo (*Citrus maxima*) (JIA *et al*., 2020), poplar (*Populus tomentosa*) (FAN *et al*., 2015) and coffee tree (*Coffea canephora*) (BREITLER *et al*., 2018), but the difficulty on conducting genetic studies, genome projects or even regenerating the plants *in vitro* hampers a broader application on tree crops. Moreover, in vision of the expanding capacity of CRISPR technology for genome editing, as exemplified by recent advances on base editing (GRUNEWALD *et al*., 2020) and the development of prime editing (ANZALONE *et al*., 2019; LIN *et al*., 2020), there’s still many branches to explore for tree genetic improvement.

When a tree crop is planted on the field it is supposed to endure for years or even decades, therefore a breeding strategy should consider resources limitation and climate change forecasts for a sustainable production and explore the best of technology and knowledge to increase the chances of a plantation to have good performance.

The application of CRISPR-Cas systems directly to solve breeding constraints or to identify causal genes for important traits can help on the establishment of feasible roadmaps for generating improved cultivars. Alongside of focusing on elite varieties, a diversified panel of species achieved through *de novo* domestication of wild relatives may help to secure tree cultivation systems in face of future unknowns.

2 THE CHALLENGES OF BREEDING, GENOMICS AND GENETIC MODIFICATION OF TREE CROPS

The focus of this perspective work are the tree crops, a classification of plant species distinguished by three main characteristics, an elongated stem, woody structure and perennial growth habit, which humans uses as food supply or industrial raw material (NEALE *et al*., 2017). We are attached to this group of plants since before the establishment of the first non-nomad societies, as the first hominids and also our primate ancestors, being part of the megafauna mammalians, had fruits from arboreal plants as part of their diet (SPENGLER, 2019).

Although the mutualistic relationship between humans and fruit trees exists for thousands of years, as exemplified by apple tree evolutionary studies, the process of domestication differs from what is seen for annual food crops, because hybridizations and propagation of the selected plants by cloning methods took place on many arboreal plants cultivation, leading to cultivated varieties without fixed introgressions or mutations (SPENGLER, 2019). In this way, a seed-generated F1 progeny from a cultivated variety would not exhibit the selected/improved phenotype; therefore the domestication of the species itself is compromised, analyzing through a conventional outlook.

Tree crops evolutionary studies, especially regarding forestry species, are generally less common than those for annual crops. The early domestication process on plants explored for non-food purposes, like timber harvesting, or the human influence on these species evolution, might be difficult to track due to the exploratory nature of the relationship, as humans used the whole plant without influencing the seed propagation or the reproduction success. Just more recently, breeding programs are influencing the phenotype of commercial forestry species in consistency with society demands.

In general terms, the breeding process for tree crops is challenging, mainly due to the long life cycle of the plant, but also because of the space needed to maintain the breeding populations. The breeding cycle length is determined by the duration of the juvenile phase of the plant, which in turn depends on many environmental and genetic factors, but for a simplistic example, can vary from minimum three years for *Prunus persica* (peach) to more than a decade for *Persea Americana* (avocado) (NOCKER and GARDINER, 2014).

The extended juvenile phase of arboreal plants also hampers the development of genome projects, because of the difficulty on having high resolution genetic maps of plants with long juvenile phase, which are needed for anchoring the sequenced pieces of the DNA to establish the proper positions of the genes on a chromosome context. Despite the challenge, genetic maps and genome assembly project are being elaborated for trees (BERNHARDSSON *et al*., 2019; LANGDON *et al*., 2020) and some of the most commercialized arboreal crops already have chromosome level-assembled genomes (supplementary table 1).

Among the improvements on plant biology led by genomic information availability, this knowledge allows for advances on evolution research (LIU *et al*., 2020; XIA *et al*., 2020), genome comparisons (ALONGE *et al*., 2020), gene family analyses (PINTO *et al*., 2019), and genome-wide association approaches (FERRÃO *et al*., 2020). In sum, a very powerful achievement is the comprehension of genetic factors, at DNA sequence level, that influences plant traits variation.

The union of this knowledge with recombinant DNA technologies and plant tissue culture techniques pave the way for the rising of plant genetic modification. The plant tissue culture-based regeneration processes, as well as the genetic transformation protocols are established for most of the genome-sequenced tree crops. Although, we have commercialized annual transgenic crops since 1990 and just few recent examples of regulated transgenic trees (2.6% of the total approved) could be highlighted (ISAAA, 2020).

This absence of transgenic tree crops could be related to the time-consuming *in vitro* regeneration process for these plants and also to the challenge on establishing a transformation protocol to commercial varieties, or on transferring the inserted gene from an ease-transformation variety to an elite cultivar, again, because of the long juvenile phase which make the crossing cycles a hard obstacle, but also due to the associated genetic elements that can be co-introduced into the elite cultivar from the ease-transformation variety together with the target gene, the so-called linkage drag.

We should also consider the greater time and financial investment needed to study the effect of a transgene on arboreal plants, which is a determinant variable on the cost-benefit equation. But, recently, the expansion of the genetic modification possibilities panel by the incorporation of gene editing technology (JINEK *et al*., 2012) and the associated forms of applying it to crops has become a heavy weight on the benefit side of the cost-benefit balance, making tree crops scientists and private sector leaders to think towards gene edited tree varieties being cultivated.

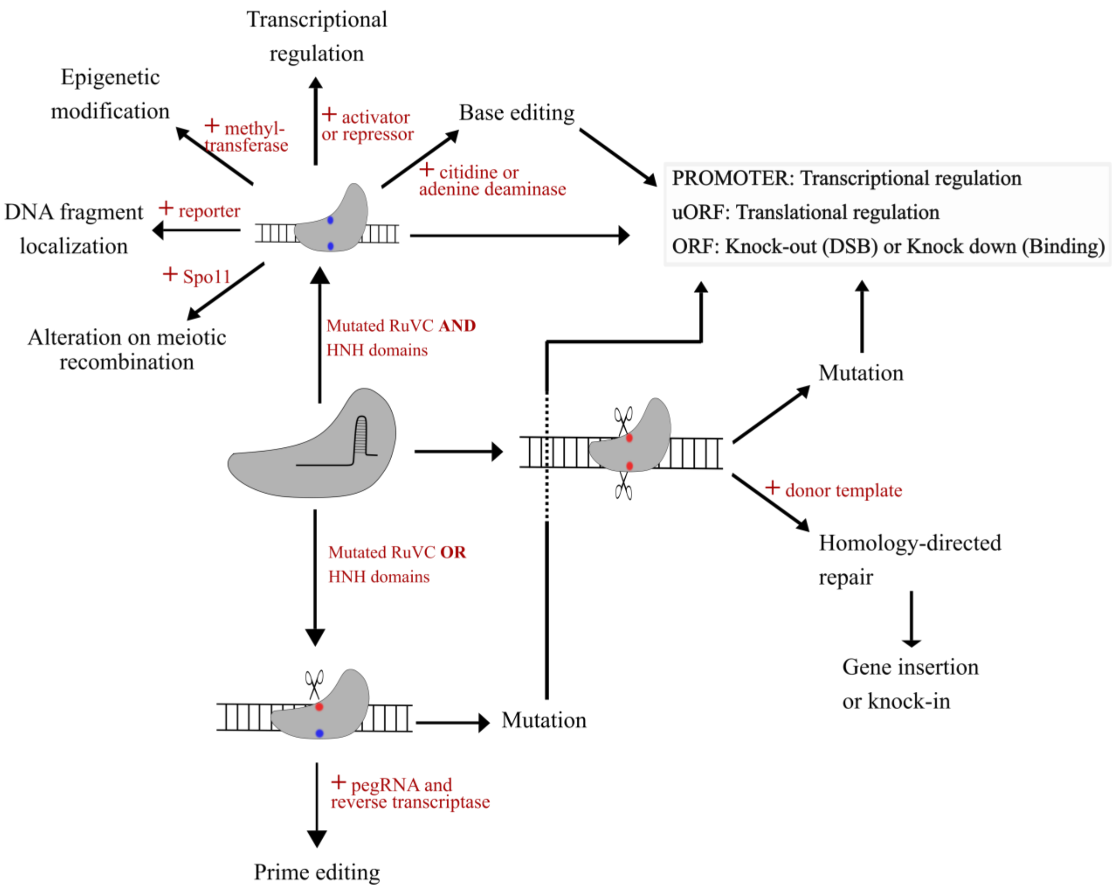
3 THE EXPANDING CAPACITY OF CRISPR-BASED GENE EDITING FOR GENETIC IMPROVEMENT

The trajectory of crop improvement could be divided into four generations, according to Fernie and Yan (2019): the first encompass breeding by phenotype-based selection, performed by independent farmers; the second incorporates mate designs, hybrid breeding, statistical analyses, use of fertilizers and pesticides and is marked by high yielding dwarf plants and the first green revolution; the third, which can be considered the second green revolution, includes transgenic and genomic breeding technologies and, finally, the fourth generation would be the current stage of crop improvement history, with the incorporation of gene editing, precision breeding and big data mining, promising to be the third green revolution.

This fourth generation of crop improvement might be marked by the efforts on increasing the sustainability of agricultural industry faced by current social and environmental demands (Fernie and Yan, 2019) and the expanding capacity of CRISPR-based gene editing technology will help to make possible the rational design of crops, which should include the arboreal plants. The current panel of the CRISPR system capacity on performing targeted genetic modifications on organisms (figure 1) allows for the rising of many models of application for tree crops that will be further discussed on the next topic of this perspective.

Before the discussion focused on tree crops, we need a brief summary of the CRISPR system diversity, that is the basics to explore the applicability for arboreal plants genetic improvement. The outbreak of this technology happened after the demonstration that a bacteria immunity system could be harnessed as a molecular tool to perform programmed genetic changes (JINEK *et al*., 2012), but for an understanding of the previous scientific breakthroughs that led to this insight, we refer to a comprehensive review (LANDER, 2016).

Figure 1 – Summary of CRISPR-Cas system versatility


Transcriptional regulation
Epigenetic modification
+ methyl-transferase
+ activator or repressor
Base editing
+ citidine or adenine deaminase
PROMOTER: Transcriptional regulation
uORF: Translational regulation
ORF: Knock-out (DSB) or Knock down (Binding)
DNA fragment localization
+ reporter
+ Spo11
Alteration on meiotic recombination
Mutated RuVC AND HNH domains
Mutation
+ donor template
Homology-directed repair
Gene insertion or knock-in
Mutated RuVC OR HNH domains
Mutation
+ pegRNA and reverse transcriptase
Prime editing

Legend: A brief summary on CRISPR-Cas system versatility. The Cas proteins complexed with sgRNAs can promote DSB, which can either generate mutations NHEJ missed repair or induce gene insetions through homology-directed repair if a donor template is provided; a dCas9 can be harnessed for multiple functions by different protein fusions, base editing, transcriptional regulation, epigenetic modification, dna fragment localization and enhance meiotic recombination; a nCas9 can also induce mutations by paired one-strand-break or harnessed for performing prime editing; the mutations can be targeted to promoter regions, to uORF and ORF, inducing transcriptional changes, translational control and knock out or knock down, respectively. Fonte: Do autor, 2020.

The main explored capability of CRISPR system and the first to be used is the targeted double strand break (DSB), induced by the nuclease CAS9 (or other type 2 CAS enzymes, like Cpf1 (ZETSCHE *et al.*, 2015; ZETSCHE *et al.*, 2017)) and guided by a programmable sgRNA designed to be complementary to a target genome region. The DSB can be repaired by the inner cellular machinery by two different pathways, the Non-homologous end joining (NHEJ) and Homology directed repair (HDR) (JIANG and DOUDNA, 2017). The most common used pathway on plant somatic cells is error-prone NHEJ (reviewed on ROZOV *et al.*, 2019) and, due to the repeatedly action of the CAS9, the targeted site can be miss repaired, causing a mutation on the desired genomic region.

The intended mutation caused by the miss repair of the DSB can be targeted to, virtually, any part of an organism's genome by designing a proper sgRNA. If it targets an open reading frame (ORF) of a gene, it has the potential to disrupt its function; in case of an UTR-ORF (uORF), it could influence the translation of the down/upstream gene (principal ORF) and if the mutation is programed to happen on a promoter region, it is possible to interfere on the gene transcription (figure 1).

Despite less common, plants can use HDR pathway in response to DSB and, on such opportunity, it is possible to incorporate a DNA fragment on a desired genome position by delivering a donor template (DNA fragment of interest) together with the CRISPR-Cas9 reagents to a cell. This approach is harder to perform, but important progress has been achieved on this gene knock in strategy on plants (LU *et al*., 2020; DONG *et al*., 2020).

Besides the outstanding capabilities of the regular CRISPR-Cas system, researchers found the way to mutate the dnase activity domains of the Cas9 protein, RuvC and HNH, making it possible to perform one strand breaks (nicking) by deactivating one of the two domains (nickase, nCas9) or turning the "scissor" Cas9 into a programmable binding protein, by shutting down all the dnase activity (dead-Cas9, dCas9) (QI *et al*., 2013; RAN *et al*., 2013) (figure 1)..

Interestingly, by using two nickases, one can make a target deletion with less chances of an off-target mutation and exploring fusions of the dCas9 with other proteins, like transcriptional activator or repressors, citidine or adenine deaminases, methyltransferase, reporter proteins and meiosis-specific endonucleases (Spo11), it is possible to influence transcription, make single base changes, epigenetic modifications, visualize/localize DNA fragments on the chromosome context and alter recombination rates during meiosis (LARSON*et* al., 2013; PAIXÃO *et* al., 2018; QIN *et al*., 2020; HILTON *et al*., 2015; DREISSIG *et al*., 2017; SARNO*et al*., 2017) (figure 1). Still, different Cas proteins can be explored for versatility of PAM sites (the recognition motif for a Cas protein, like *NGG* for Cas9) and also activities, like the RNA editing capability of Cas13, making CRISPR system a very versatile molecular tool for genetic engineering (MANGHWAR *et al*., 2019).

Some of the most recent technique breakthroughs for CRISPR-Cas system includes a protein complex (Cas9 fused to cytosine and adenine deaminases) to perform both cytosine and adenine base editing (CBE and ADE) concomitantly (GRUNEWALD *et al*., 2020), and a method to induce programmable small editions on DNA, named prime editing (ANZALONE *et al*., 2019) and already tested on crops (LIN *et al*., 2020). Based on the utilization of a nickase (nCas9) fused to a reverse transcriptase (RT), guided by a prime editing-gRNA (pegRNA), which contains the sequence to guide the complex to the target genomic region and a sequence of interest that will be incorporated on the DNA by the action of the RT after the nicking perfomed by the nCas9.

Alongside with this brief summary of CRISPR-Cas system associated techniques which are applicable on crops, we refer the reader to other recent specific reviews that can help on deepen the comprehension on the technology (WADA *et al*., 2020; McCARTY*et al*., 2020; MISHRA *et al*., 2020; MORAD and ABDULAH, 2020). Altough, it is worth to mention that CRISPR technology is being constantly upgraded; it is a field with ever-growing novelty. An example of interesting future CRISPR-Cas system capability come from studies on CRISPR associated transposases (CAST), which might be harnessed to perform long RNA-guided DNA insertions on eucaryotes efficiently, surpassing a current bottleneck on genome editing (KLOMPE *et al*., 2019; STRECKER*et al*., 2019).

It is important to highlight that the current status of CRISPR technology applicability on crops is dependent of the parallel improvement of many other areas, being a multidisciplinary effort. Alongside with plant biology techniques, sequencing technologies, genetics and bioinformatics, we need to highlight the advances on constructing vectors for the strategies involving CRISPR systems in plants (CEMARK *et al*., 2017; HAHN *et al*., 2020).

4 APPLICABILITY AND PROSPECTS OF GENETIC IMPROVEMENT USING CRISPR-CAS SYSTEMSON TREE CROPS

On the beginning of the second decade of 21st century there’s already a showcase evidencing that agriculture with gene edited tree crops is a feasible reality (supplementary table 2). The first published work with CRISPR-based gene editing on a full regenerated arboreal plant was performed on *Populus tomentosa*, a proof-of-concept *PtoPDS* knockout that resulted on albino-like poplar plantlets (FAN *et al*., 2015). The same strategy was applied to other tree crops, like citrus (*Citrus trifoliate x C. sinensis*), apple tree (*Malus x domestica*), grapevine (*Vitis vinifera*) and kiwifruit (*Actinidia spp*) (ZHANG *et al*., 2017; CHARRIER *et al*., 2019; NAKAJIMA *et al*., 2017; WANG *et al*., 2018).

Despite the potential, CRISPR-based gene editing is a very recent technology considering tree crops genetic improvement outlook, as for many of these plants the genetic transformation and full regeneration processes can take years (if it exists), not mentioning the time to identify, clone the gene and construct the vectors. As a result, not too much than a handful published reports go beyond proving CRISPR-Cas system feasibility on arboreal plants (supplementary table 2), like enhancing citrus canker resistance by editing the *CsLOB1*promoter in citrus (JIA *et al*., 2020) and increasing resistance to *Botrytis cinerea* through *VvWRKY52* knockout in grapevine (WANG *et al*., 2018).

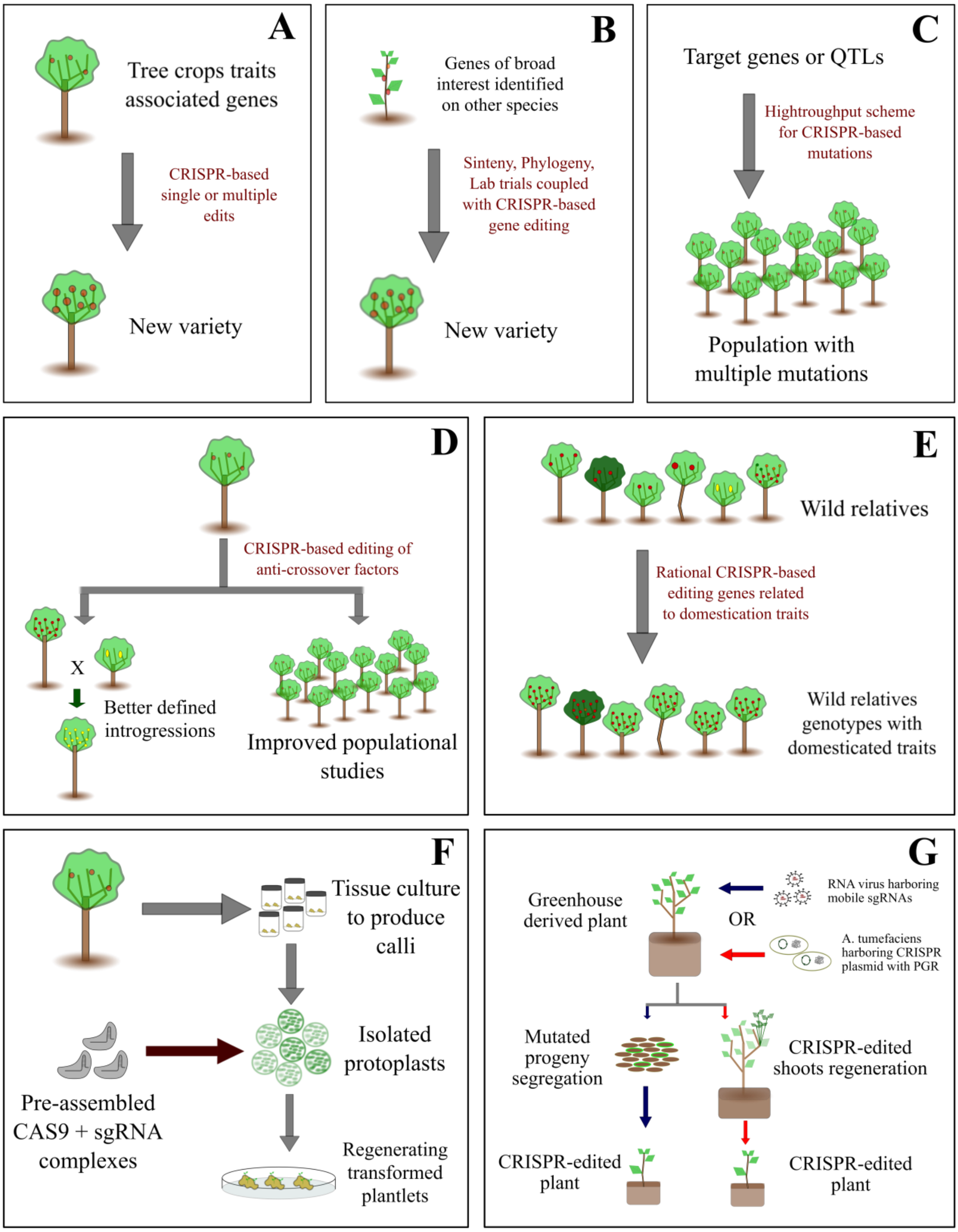
Therefore, the applications of CRISPR-Cas technology on arboreal plants are just beginning to appear, a modest state in comparison of what is being done for annual plants, due to already mentioned reasons, and just the first step on the potential of gene editing technology for agriculture. However, thinking on the recent advances on plant field, we are able to summarize some expectations for tree crops on the gene editing era (figure 2).

In first place, a straightforward approach would be looking for orthologous genes in tree crops related to others already used in genetic modification experiments in other plant species. Some interesting examples for fruit bearing trees can be appreciated from pioneering gene editing studies performed with tomato plants (*Solanum lycopersicum*), like one with a recently characterized AP2/ERF transcription factor named *EXCESSIVE NUMBER OF FLORAL ORGANS* (*ENO*), which influences floral meristem activity by directly inhibiting *SlWUS* expression and its disruption via CRISPR-Cas9 results in fruit size increase (YUSTE-LISBONA *et al*., 2020).

Other regulator of shoot apical meristem (SAM) development was already targeted in tomato plants to increase the size of its fruits, the *SlCLV3*, in which both exon and promoter-directed mutations resulted in larger fruits (ZSOGON *et al*., 2018; LI *et al*., 2018). Another interesting gene for improving fruit yields was also targeted in these two works, the flowering repressor *SELF PRUNING 5G* (*SP5G*), a member of the *CETS* family which its disruption promoted early flowering in tomato and, consequently, early yield (SOYK *et al*., 2017).

Beyond improving fruit yields, the study of *CETS* family members like the *SlSP5G*, can lead to important implications on the understanding of flowering on arboreal plants and, therefore, uncover the roadmap to manipulate the juvenile phase length to accelerate breeding or also to speed up certain fruits production. A simultaneously mutation in *SP5G*, *SP* (another *CETS* family member) and *SlER* (a leucine-rich receptor kinase), also in tomato, led to a very compact plant (KWON *et al*., 2020) and this could be applied too on fruit tree crops, anticipating that genetic differences will take place on different plant species.

Figure 2: Summary of some possible CRISPR applicability on tree crops


A
Tree crops traits associated genes
CRISPR-based single or multiple edits
New variety
B
Genes of broad interest identified on other species
Sinteny, Phylogeny, Lab trials coupled with CRISPR-based gene editing
New variety
C
Target genes or QTLs
Hightroughput scheme for CRISPR-based mutations
Population with multiple mutations
D
CRISPR-based editing of anti-crossover factors
X
Better defined introgressions
Improved populational studies
E
Wild relatives
Rational CRISPR-based editing genes related to domestication traits
Wild relatives genotypes with domesticated traits
F
Tissue culture to produce calli
Isolated protoplasts
Pre-assembled CAS9 + sgRNA complexes
Regenerating transformed plantlets
G
Greenhouse derived plant
RNA virus harboring mobile sgRNAs
OR
A. tumefaciens harboring CRISPR plasmid with PGR
Mutated progeny segregation
CRISPR-edited shoots regeneration
CRISPR-edited plant
CRISPR-edited plant

Legend: Summary of some possible CRISPR applicability on tree crops. A- Editing genes already characterized on tree crops; B- Identification of orthologous genes between tree crops and other species and modification of them by CRISPR-Cas; C- Definition of QTLs or genetic basis of important traits by high throughput CRISPR-Cas gene editing; D- Editing anticrossover factors genes to increase meiotic recombination, which improves the development of breeding population and introgressions; E- CRISPR-Cas-based gene editing to *de novo* domesticate wild-relative species; F- Pre-assembled RNP-based delivery system to transform plant cells without inserting transgenes; G- Tissue culture-free methods, by RNA virus infection or *A. tumefaciens* mediated, to perform CRISPR-Cas gene editing on tree crops. Fonte: Do autor, 2020.

Although traits related to fruit weight and plant (aerial part) morphology are important focuses for agriculture development, the 21st century agricultural systems must be prepared for the resources limitations and climatic instabilities forecast, especially the tree crops sectors, as these plants stand on the field for many years or decades. Some important traits for this scenario are related to drought and heat resistance, salt tolerance mechanisms, phosphate uptake improvement, nitrogen fixation and pest/disease resistance.

In order to support gene editing strategies focused on traits, beyond disease/pest resistance and yield, that contributes to sustain agriculture in face of the most probable future challenges, we made a list of some already studied genes in which the use of CRISPR-Cas system for a small deletion has the potential to increase fitness on an agricultural system with less resources and/or subjected to climate changes (supplementary table 3). We also refer the reader to some recent reviews (GONÇALVEZ *et al*., 2020; JANNI *et al*., 2020; HUISMAN *et al*., 2020).

Of course there's no simple solution for adapting tree crops to future agriculture demands and the application of knowledge acquired from other species, like editing orthologous genes, does not unleash the potential of molecular design. To this end, the genetics of such species must be deeply explored, which, as already pointed here, is a hard goal for arboreal species, mainly because the major part of any genetic study is based on crossings.

However, some interesting innovations using CRISPR systems can boost tree genetics also. A high throughput mutagenesis approach was recently applied on maize (*Zea mays*) using multiplexed CRISPR-Cas9 on a batch pipeline optimized with pooled transformation and low-cost barcoded deep sequencing and resulted on 118 mutated genes (412 edits in total) (LIU *et al*., 2020). The same strategy could be applied to tree crops with already sequenced genome and efficient genetic transformation protocols in order to rationally identify causal genes of important traits without time-costly crossing and random mutation experiments.

Although a rational-directed study is interesting for unraveling causal genes for known traits, it is not a substitute for segregating population analyses. Fortunately for tree crops geneticists, some initial and recent results demonstrate the capability for accelerating breeding through directed mutagenesis. Exploring classic mutagenesis assays (mostly EMS-derived mutant collections), it was already demonstrated that disruption of the helicases genes *RECQ4* and *FANCM*, which act as anticrossover factors, resulted in more than two-fold increase of recombination rates in three distantly related plant species, rice (*Oriza sativa*, monocot), tomato (*S. lycopersicum*, eudicot-asterid) and pea (*Pisumsativum*, eudicot rosid) (MIEULET *et al*., 2018).

Most of the accessions used on the study were stop-codon mutant lines (MIEULET *et al*., 2018), what points to the possibility of targeted disruptions of the anti-crossover factors by CRISPR-Cas9. The manipulation of such genes and other crossover (CO) related factors have the potential to increase CO frequency and, thus, allowing for the development of high definition genetic maps quicker. Beyond enabling achievements on genetic maps construction, controlled recombination approaches using directed nucleases are promising alternatives for breeding on a whole perspective, from improving pre-breeding population diversity to fine tuning introgressions, as obtaining favorable haplotypes, which is the foundation of breeding, depends ultimately on CO-driven recombination of genetic information (TAAGEN *et al*., 2020).

Increasing CO frequencies is also interesting for breeding strategies that involves introgressions from wild-relatives or related species to elite genotypes, as a higher recombination rate increase the chances of isolating the desired gene from nearby genetic elements, decreasing the linkage drag that could affect the genetic gain. The use of CRISPR-Cas9 to disrupt *RECQ4* was recently successfully applied on tomato and the result was a higher recombination frequency on an interspecific hybrid (*S. lycopersicum x S. pimpinelifolium*) using CRISPR-edited *recq4* plants in comparison with wild type ones (MAAGD *et al*., 2020), paving the way for applications on crop breeding, especially for arboreal plants, where successive crossing cycles to attenuate linkage drag effects is cumbersome.

Despite these imminent advances that could also be applied on transferring genetic modifications to elite crops, if the genetic basis of a trait variation is already known, it is worthier to directly modify the elite variety. However, depending on the crop, the presence of the CRISPR expression cassete as a transgene can affect its market value. Also, if the resulting CRISPR-modified variety does not contain a transgene, it can be analyzed through a product-based concept in many countries, which decreases the time and investment for commercialization (METJE-SPRINK *et al*., 2020).

Unfortunately, segregating the transgene out would require at least one extra crossing cycle, which on a tree crop can represent much more than a year, without mentioning that for many elite varieties even the backcrossing is not recommended due to the heterogenic basis of a given important phenotype and/or high ploidy level. Therefore, an interesting alternative for this process would be a transgene-free strategy to obtain gene-edited plants.

In this case, progress is been made on CRISPR machinery delivering systems, like the use of preassembled ribonucleoprotein (RNP) complexes for plant transformation. Both Cas9 and Cas12a (Cpf1) were already applied on this strategy, which requires the purified nuclease to be *ex vivo* assembled with a sgRNA prior the introduction on plant cells. The ready Cas-sgRNA complex is usually delivered to protoplasts, which are submitted to an osmotic shock (PEG-mediated) in order to incorporate the assembled CRISPR machinery and after that, regenerated into plants (WOO *et al*., 2015; KIM *et al*., 2016; BRANDT *et al*., 2020).

Some successful reports on applying this method on arboreal plants are already published for apple, grapevine and ruber tree (*Hevea brasiliensis*) (MALNOY *et al*., 2016; OSAKABE *et al*., 2018; FAN *et al*., 2020), which might reflect the great interest on this delivery system for tree crop elite varieties, as other aspects around CRISPR technology are not receiving such attention. Although not reported on arboreal species yet and more dependent on mutant screening strategies and transformation efficiency, it is possible to opt for delivering the plasmid encoding the Cas9 and sgRNAs directly to protoplasts to induce a transient transformation and select transgene-free events (LIN *et al*., 2018), avoiding the use of purified Cas proteins on preassembled complexes.

The biggest issue of these transgene-free delivering methods is the dependence on highly optimized plant tissue culture processes, as producing totipotent material, isolating protoplasts, transforming and regenerating plantlets requires great expertise and well established protocols. This methodology is not common for many plant species and sometimes can be even specific for a given genotype. The ideal scenario for CRISPR technology broad application on crops would be a transgene and tissue culture-free method of plant transformation.

Regarding this outlook, two interesting recent reports popped-out demonstrating that gene edited plants could be obtained without massive use of tissue culture (ELLISON *et al*., 2020; MAHER *et al*., 2020). The first uses a tobacco (*Nicotiana* tabacum) transgenic plant expressing the Cas9 protein, obtained by traditional processes, which is *ex vitro* infected by a RNA virus containing sgRNAs cleverly fused with plant mobile mRNAs (like from the *FT* gene). This *in planta* transformation resulted on 60 to 100% of edited progeny when one genomic site is targeted and up to 30% of progeny with three consecutive targeted edits (ELLISON *et al*., 2020).

On the second report, plant meristems are induced *de novo* by concomitant expression of developmental regulators (genes related to meristem identity maintenance, like *WUSCHEL*) and gene-editing reagents, by infecting greenhouse-growing plants. From the *de* *novo* induced meristems, edited shoots are regenerated with fixed mutations, some of them being transgene-free (MAHER *et al*., 2020), evidencing its highly potential applicability on crop genetic improvement, enabling gene editing on a straightforward way on plants which are hard to regenerate and transform *in vitro*, like most of the tree crops.

Despite the possibility of editing elite varieties, a reasonable outlook for future agriculture should consider a diverse panel of genotypes for a given crop. One perspective to achieve this scenario would be exploring wild relatives of cultivated plants. By studying domestication-related traits and uncovering its genetic basis, it is possible to induce genetic modifications to rationally convert wild plants selected characteristics (like fruit size) into cultivated varieties similar phenotypes, but preserving the wild relative background, a concept already proposed as *de novo* domestication (ZSOGON *et al*., 2017) and proved to be a feasible approach for tomato crop (ZSOGON *et al*., 2018).

Using a multiplexed CRISPR-Cas9 strategy, key genes that influences agronomical and domestication-related traits were modified in wild relatives of tomato (ZSOGON *et al*., 2018; LI *et al*., 2018) and also on an orphan crop, the groundcherry (*Physalis pruinosa*) (LEMMON *et al*., 2018), inducing a swift from a wild phenotype to a crop-similar one, but maintaining the other characteristics from the wild genetic background (like pest/disease resistance, abiotic stress adaptation traits and/or nutritional aspects), which are not common among elite varieties.

Including wild relatives into the cultivation panel, aside of elite varieties, on a fast and precise pace through *de novo* domestication and related methodologies, might provide the diversity needed to sustain the agriculture development on the beginning century and this is especially urgent for tree species.

5 A STIMULATING EXAMPLE - PERSPECTIVES FOR CROPPING A SIMPLER COFFEE TREE

Coffee production relies on two tree species cultivation, *Coffea arabica* (approximately 60% of the market) and *C. canephora*, mainly cultivated on tropical regions, with Brazil being the major exporter country (ICO, 2020). To produce one of the most appreciated beverages in the world, a coffee tree must be planted at least two years prior the bean’s harvest and processing and, after this first cycle, the plant stays on the field for yearly productions.

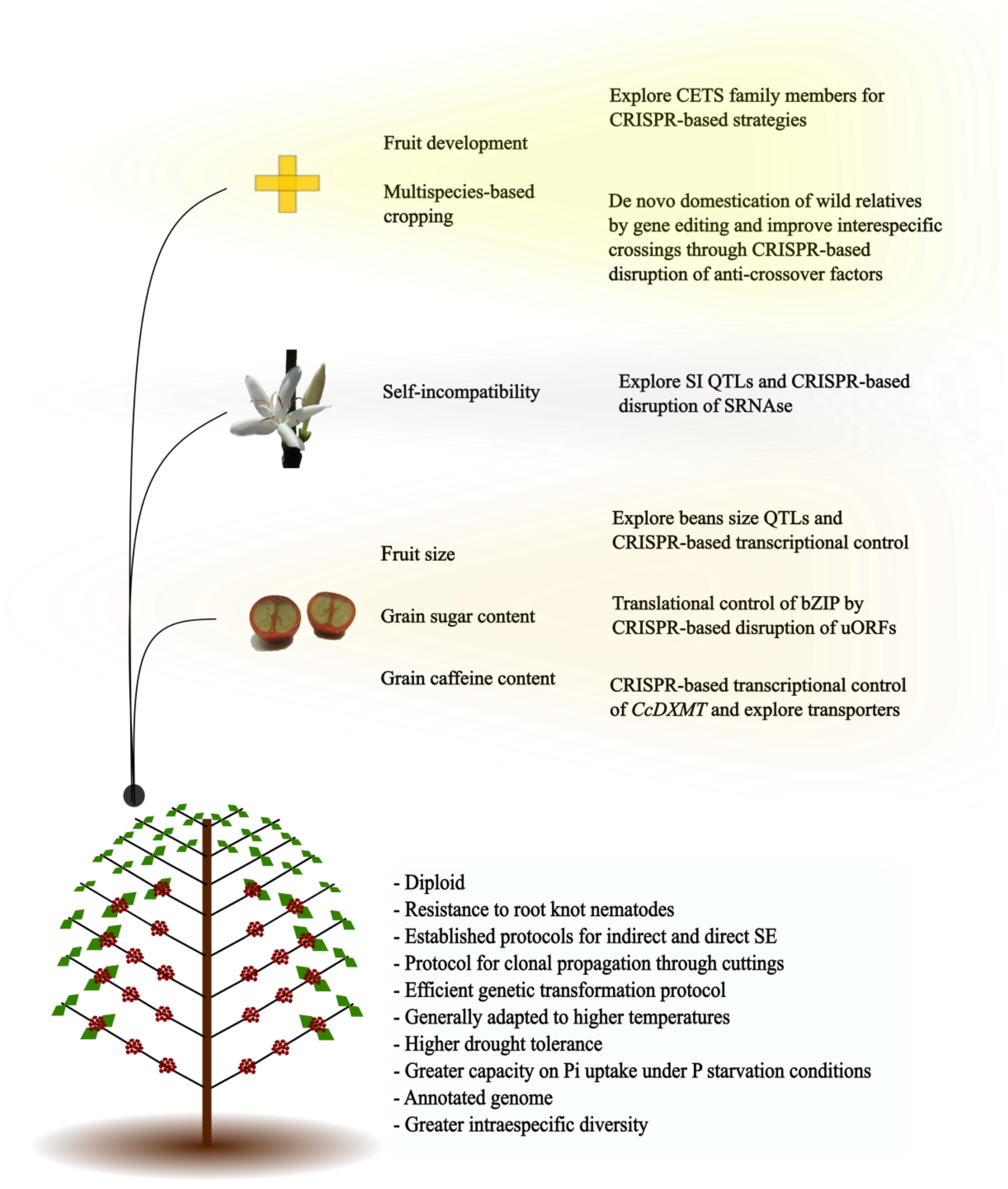
C. arabica, a natural allotetraploid hybrid (*C. canephora x C. eugenioides*) originated from a single polyploidization event, was the first species to be cultivated for coffee production and dominated the market, despite the diversity of *Coffea* genus (more than 100 species) (SCALABRIN *et al*., 2020; HAMON *et al*., 2017). Nowadays, the prevalence of *C. arabica*-derived coffee is mainly due to the sensorial characteristics of the beverage produced through its grains, usually reported as having superior quality than *C. canephora*-derived coffees, the second most common one. However, some breeding programs are recently investing on *C. canephora* genotypes for improved cup quality (LEMOS *et al*., 2020).

The registered cultivation of *C. canephora* varieties started about one hundred years ago (HERRERA and LAMBOT, 2017), a very recent process of domestication of this multi-stemmed tree that originated on central Africa. Although just recently explored, the use of this species for coffee has the potential to help sustaining its production in face challenges for this culture. Breeding programs can explore the intraspecific variation resulted from diverse regions of wild populations and the consequent trait variability.

In addition, *C. canephora* plants are generally adapted to lower water supply, higher temperatures and lower altitude cultivation conditions than *C. arabica*, despite being a source of resistance to root knot nematodes (*Meloidogyne exigua*, *M. incognita* and *M. paranaensis*) that impairs *C. arabica* cultivation in contaminated areas (BERTRAND *et al*., 2001; FATOBENE *et al*., 2018; SALGADO *et al*., 2019). Along with this, *C. canephora* is diploid; a major attribute in its favor for coffee production, as genetic studies and breeding can be performed easier, besides genomic analyses (e.g. it is the only *Coffea* species with an annotated genome (DENOEUD *et al*., 2014)) and genetic engineering approaches (BREITLER *et al*., 2018).

We argue that genetic engineering strategies should focus on tailoring *C. canephora* as the main coffee tree for massive production, which could bring the upper cited advantages to coffee cropping systems and many more derived from the exploration of other diploid *Coffea* species through interspecific hybridizations and trait introgressions. For this, we propose gene editing-based initiatives which could alter current *C. canephora* characteristics unfavorable to this perspective (figure 3).

Figure 3 – Concise strategy to genetically manipulate *C. canephora* towards the main crop for massive coffee production


Fruit development
Explore CETS family members for CRISPR-based strategies
Multispecies-based cropping
De novo domestication of wild relatives by gene editing and improve interespecific crossings through CRISPR-based disruption of anti-crossover factors
Self-incompatibility
Explore SI QTLs and CRISPR-based disruption of SRNAse
Fruit size
Explore beans size QTLs and CRISPR-based transcriptional control
Grain sugar content
Translational control of bZIP by CRISPR-based disruption of uORFs
Grain caffeine content
CRISPR-based transcriptional control of *CcDXMT* and explore transporters
- Diploid
- Resistance to root knot nematodes
- Established protocols for indirect and direct SE
- Protocol for clonal propagation through cuttings
- Efficient genetic transformation protocol
- Generally adapted to higher temperatures
- Higher drought tolerance
- Greater capacity on Pi uptake under P starvation conditions
- Annotated genome
- Greater intraespecific diversity

Legend: Concise scheme of a strategy to genetically manipulate *C. canephora* towards the main crop for massive coffee production. In addition to some intrinsic qualities of this species, some constraints regarding beverage quality in comparison to *C. arabica* standard could be addressed, like decreasing caffeine content by editing *CcDXMT* or caffeine transporters, increase sugar content by targeting uORF of *bZIP* genes and exploring QTLs to identify targets for fruit size variation; a major breakthrough for broader exploration of this species would be surpassing its self-incompatibility, which could be achieved by identifying and editing SRNAse coding genes; finally, some upgrades could be targeted, like decreasing fruit development discrepancy and explore wild-relative species, by exploring genes related to flowering control, like ones from *CETS* family and performing *de novo* domestication, respectively. Fonte: Do autor, 2020.

In order to cope with differences in cup quality, we suggest that two biochemical attributes should be targeted initially, the caffeine and sugar levels. These two compounds affect sensorial characteristics and it is known that *C. canephora* has higher quantities of the first and lower of the second, in almost two-fold scale for both (LEMOS *et al*., 2020; LEBOT *et al*., 2020). Besides biochemical characteristics, the fruit size is another important trait for coffee production, being a selective parameter during processing of the grains and, generally, *C. canephora* varieties have smaller ones.

The enzymes and respective genes that catalyze the last steps on caffeine biosynthesis (*CcXMT*, *CcDXMT* and *CcDXMT*) are known (DENODEUD *et al*., 2014), therefore, especially *CcDXMT*, could be edited to induce decrease on this alkaloid concentration on fruits. It was already demonstrated that mutations on caffeine synthase gene interfering with its expression and also probably with the coded enzyme activity may be associated with reduced caffeine synthesis in low-caffeine *C. arabica* mutants (MALUF *et al*., 2009; FAVORETTO *et al*., 2017).

In this case, a step-wise strategy would be to induce multiple and sequential CRISPR-based heritable mutations on this gene's promoter region, to obtain mutants differing on *CcDXMT* transcriptional levels and consequently, on fruit caffeine content. An alternative for the same goal would be identifying a caffeine transporter, which could be essential for its accumulation on fruit cell vacuoles, as reported in other species for diverse alkaloids and other metabolites (PAYNE et al., 2017; DEMURTAS et al., 2019; BANASIAK et al., 2020), and its CRISPR-mediated knock-out could contribute to reduce caffeine accumulation.

Besides sweetness, sugar content greatly influences the coffee aroma as a result of sugar-derived compounds formed due to roasting process, but little is explored about sugar biosynthesis on coffee tree. From studies with tomato and tobacco, *bZIP* transcription factors were identified as major regulators of sucrose biosynthesis (and some aminoacids) and, interestingly, this gene is post-transcriptionally regulated by SIRT (Sucrose Induced Repression of Translation) (THALOR *et al*., 2012; SAGOR *et al*., 2016).

These authors reported that this repression is mediated by a uORF-derived peptide, which influences the sucrose concentration threshold and demonstrated that transgenes without this regulator sequence induces greater sugar accumulation. A reasonable trial would be to identify such *bZIP* on *C. canephora* and, using CRISPR-Cas9, disrupt its uORF, which is so far considered conserved among different species, leading to possible relaxation of the sucrose concentration threshold on sugar accumulating tissues, like the grains.

A promoter-targeted CRISPR-based strategy could also favor the development of *C. canephora* varieties with a range of grain sizes, as discussed for caffeine concentration and already demonstrated for tomato fruit size variation, where the authors achieved a continuum of fruit size variation by targeting different regions of *SlCLV3* promoter via CRISPR-Cas9. To the best of our knowledge, none CRISPR-based strategy resulted on increasing grain size on dicot plants yet, but some genes that might influence this trait were already identified on rice (reviewed by FIAZ *et al*., 2019) as well as QTLs already associated with bean size variation on *C. arabica* (MONCADA *et* al., 2016) and *C. canephora* (LEROY *et al*., 2011).

Besides beverage quality and fruit morphology-related traits, a major modification needed to improve *C. canephora* cultivation and breeding is reverting its self-incompatibility (SI). The *S* locus associated with this species gametophytic SI was already mapped (LASHERMES *et al*., 1996) and studies on other species indicates that this type of SI follows the collaborative non-self recognition model, in which pistil expressed RNAses (S-RNAses) degrades the self-derived pollen, as the F-box proteins that usually inhibits these ribonucleases fails on recognize self-S-RNAses (MUNOZ-SANZ *et al*., 2020). Interestingly, a CRISPR-based knockout of S-RNAse gene was already performed and proved to be enough for inducing self-compatibility on a potato plant (*Solanum tuberosum*) (ENCISO-RODRIGUEZ *et al*., 2019), an evidence of the feasibility to solve this issue on *C. canephora* and enhance its breeding and cultivation potential.

The imminent capacity of rational molecular design using CRISPR technology goes beyond the highlighted possibilities for *C. canephora*, as was briefly explored on the previous topic. One can focus on studying *CETS* family genes (SOYK *et al*., 2017), to develop a strategy for surpassing the fruit development/flowering discrepancy, one of the most challenging issues for coffee tree cultivation. Also, in order to fully explore *Coffea* genus diversity, similarly as exposed here, other species can be strategically engineered to simulate a *de novo* domestication and integrated on breeding programs or even on cultivation systems. This diversity would help to secure the coffee supply chain in face of future challenges and also possibly enrich the panel of beverage types with distinct sensorial characteristics.

6 CONCLUDING REMARKS

Cultivating arboreal species will always be a time-consuming effort in comparison with annual crops, due to the longer juvenile phase and the time needed to harvest the products that we need from them. The breeding strategies are been improved to cope with this challenge and thanks to previous initiatives from research groups from all around the globe, there’s already substantial advancements to envision a future of rational molecular design of tree crops.

The application of CRISPR-based gene editing technologies is expanding the horizon of genetic improvement programs and, with the versatility of this system, associated with the development of cloning techniques, sequencing technology, plant transformation platforms, molecular biology and plant genetics, we might be able to upgrade agriculture on a fast pace to a point of precise tailoring plants to human society needs. But, with this power, comes the responsability on making decisions of what changes are determinant to reach true sustainable production and this is especially challenging for tree crops, for which a path taken today is intended to endure for decades.

REFERENCES

ALONGE, M. et al. Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato. **Cell,**v. 182, n.1, p. 145-161**,** 2020.

ANZALONE, A.V. et al.Search-and-replace genome editing without double-strand breaks or donor DNA.**Nature,** v. 576, n. 7785, p.149-157, 2019.

BANASIAK, J. et al. The full-size ABCG transporter of Medicagotruncatula is involved in strigolactone secretion, affecting arbuscularmycorrhiza.**Frontiers in Plant Science**, v. 11, 2020.

BERNHARDSSON, C. et al. An ultra-dense haploid genetic map for evaluating the highly fragmented genome assembly of Norway spruce (Piceaabies).**G3: Genes, Genomes, Genetics,** v. 9, n. 5, p.1623-1632, 2019.

BERTRAND, B.; ANTHONY, F.; LASHERMES, P. Breeding for resistance to Meloidogyneexigua in Coffeaarabica by introgression of resistance genes of Coffeacanephora.**Plant pathology,** v. 50, n. 5, p.637-643, 2001.

BRANDT, K. M.; GUNN, H.; MORETTI, N.; ZEMETRA, R.S.A Streamlined Protocol for Wheat (Triticumaestivum) Protoplast Isolation and Transformation With CRISPR-CasRibonucleoprotein Complexes. **Frontiers in Plant Science**, v. 11, p.769, 2020.

ČERMÁK, T. et al..A multipurpose toolkit to enable advanced genome engineering in plants.**The Plant Cell**, v. 29, n. 6, p.1196-1217, 2017.

CHARRIER, A. et al. Efficient targeted mutagenesis in apple and first time edition of pear using the CRISPR-Cas9 system. **Frontiers in plant science,** v. 10, p. 40, 2019.

DE MAAGD, R.A. et al.CRISPR/Cas inactivation of RECQ 4 increases homeologous crossovers in an interspecific tomato hybrid. **Plant biotechnology journal**, v. 18, n. 3, p.805-813, 2020.

DEMURTAS, O.C. et al**.**ABCC transporters mediate the vacuolar accumulation of crocins in saffron stigmas**. The Plant Cell**, v. 31, n. 11, p.2789-2804, 2019.

DENOEUD, F. et al.The coffeegenome provides insight into the convergent evolution of caffeine biosynthesis.**Science**, v. 345, n. 6201, p.1181-1184, 2014.

DONG, O.X. et al. Marker-free carotenoid-enriched rice generated through targeted gene insertion using CRISPR-Cas9.**Nature Communications**, v. 11, n. 1, p.1-10, 2020.

DREISSIG, S. et al. Live-cell CRISPR imaging in plants reveals dynamic telomere movements. **The Plant Journal**, v. 91, n. 4, p.565-573, 2017.

ELLISON, E.E. et al.Multiplexed heritable gene editing using RNA viruses and mobile single guide RNAs.**Nature Plants**, p.1-5, 2020.

ENCISO-RODRIGUEZ, F. et al. Overcoming self-incompatibility in diploid potato using CRISPR-Cas9.**Frontiers in plant science**, v. 10, p.376, 2019.

FAVORETTO, P. et al. Assisted-selection of naturally caffeine-free coffee cultivars—characterization of SNPs from a methyltransferase gene.**Molecular breeding**, v. 37, n. 3, p.31, 2017.

FAN, D. et al. Efficient CRISPR/Cas9-mediated targeted mutagenesis in Populus in the first generation.**Scientific reports**, v. 5, p.12217, 2015.

FAN, Y. et al.Efficient genome editing of rubber tree (heveabrasiliensis) protoplasts using CRISPR/Cas9 ribonucleoproteins.**Industrial Crops and Products**, v. 146, p.112146, 2020.

FATOBENE, B.J. et al.Coffeacanephora clones with multiple resistance to Meloidogyne incognita and M.paranaensis. **ExpAgric**, p.1-9, 2018.

FERNIE, A.R. and Yan, J. De novo domestication: an alternative route toward new crops for the future. **Molecular plant**, v. 12, n. 5, p.615-631, 2019.

FERRÃO, L.F.V. et al. Genome-wide association of volatiles reveals candidate loci for blueberry flavor. **New Phytologist**, v. 226, n.6, p.1725-1737, 2020.

FIAZ, S. et al. Applications of the CRISPR/Cas9 system for rice grain quality improvement: Perspectives and opportunities. **International journal of molecular sciences**, v. 20, n. 4, p.888, 2019.

GONÇALVES, B.X.; LIMA-MELO, Y.; DOS SANTOS MARASCHIN, F.; MARGIS-PINHEIRO, M. Phosphate starvation responses in crop roots: from well-known players to novel candidates. **Environmental and Experimental Botany**, p.104162, 2020.

GRÜNEWALD, J. et al.A dual-deaminase CRISPR base editor enables concurrent adenine and cytosine editing. **Nature Biotechnology,** v. 38, p 861-864, 2020.

HAHN, F.; KOROLEV, A.; SANJURJO LOURES, L.; NEKRASOV, V.A modular cloning toolkit for genome editing in plants.**BMC plant biology**, v. 20, p.1-10, 2020.

HAMON, P. et al. Genotyping-by-sequencing provides the first well-resolved phylogeny for coffee (Coffea) and insights into the evolution of caffeine content in its species: GBS coffee phylogeny and the evolution of caffeine content. **Molecular phylogenetics and evolution**, v. 109, p.351-361, 2017.

HERRERA, J.C.; LAMBOT, C.The Coffee Tree—Genetic Diversity and Origin.In The Craft and Science of Coffee.**Academic Press,** p. 1-16, 2017.

HILTON, I.B. et al.Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers**. Nature biotechnology**, v. 33, n. 5, p.510-517, 2015.

HUISMAN, R.; GEURTS, R.A Roadmap toward Engineered Nitrogen-Fixing Nodule Symbiosis.**Plant Communications**, v. 1, n. 1, p.100019, 2020.

INTERNATIONAL COFFEE ORGANIZATION. Coffee market report, available on http://www.ico.org, accessed on may, 2020.

JANNI, M. et al. Molecular and genetic bases of heat stress responses in crop plants and breeding for increased resilience and productivity.**Journal of Experimental Botany**, v. 71, n. 13, p. 3780-3802, 2020.

JIA, H.; WANG, N. Generation of homozygous canker-resistant citrus in the T0 generation using CRISPR-SpCas9p.**Plant Biotechnology Journal**, 2020.

JIANG, F.; DOUDNA, J.A.CRISPR–Cas9 structures and mechanisms.**Annual review of biophysics**, v. 46, p.505-529, 2017.

JINEK, M. et al.A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity.**Science**, v. 337, n.6096, p.816-821, 2012.

KIM, H. et al. CRISPR/Cpf1-mediated DNA-free plant genome editing. **Nature communications**, v. 8, n. 1, p.1-7, 2017.

KLOMPE, S.E.; VO, P.L.; HALPIN-HEALY, T.S.;STERNBERG, S.H. Transposon-encoded CRISPR–Cas systems direct RNA-guided DNA integration. **Nature**, v. 571, n. 7764, p.219-225, 2019.

KWON, C.T. et al. Rapid customization of Solanaceae fruit crops for urban agriculture. **Nature biotechnology**, v. 38, n. 2, p.182-188, 2020.

LANDER, E.S.The heroes of CRISPR.**Cell,** v. 164, n. 1-2, p.18-28, 2016.

LANGDON, K.S. Maximising recombination across macadamia populations to generate linkage maps for genome anchoring. **Scientific reports**, v. 10, n. 1, p.1-15, 2020.

LARSON, M.H. et al. CRISPR interference (CRISPRi) for sequence-specific control of gene expression.**Nature protocols**, v. 8, n. 11, p. 2180-2196, 2013.

LEMMON, Z.H. et al. Rapid improvement of domestication traits in an orphan crop by genome editing.**Nature plants**, v. 4, n. 10, p.766-770, 2018.

LEMOS, M.F. et al. Chemical and sensory profile of new genotypes of Brazilian Coffeacanephora.**Food chemistry**, v. 310, p.125850, 2020.

LEROY, T. et al. Improving the quality of African robustas: QTLs for yield-and quality-related traits in Coffeacanephora. **Tree Genetics & Genomes**, v. 7, n.4, p.781-798, 2011.

LI, T. et al. Domestication of wild tomato is accelerated by genome editing. **Nature biotechnology**, v. 36, n. 12, p.1160-1163, 2018.

LIN, C.S. et al. Application of protoplast technology to CRISPR/Cas9 mutagenesis: from single-cell mutation detection to mutant plant regeneration. **Plant biotechnology journal**, v. 16, n. 7, p.1295-1310, 2018.

LIN, Q. et al. Prime genome editing in rice and wheat.**Nature Biotechnology**, v. 38, n. 5, p.582-585, 2020.

LIU, H.J. High-Throughput CRISPR/Cas9 Mutagenesis Streamlines Trait Gene Identification in Maize. **The Plant Cell**, v. 32, n. 5, p.1397-1413, 2020.

LIU, J. et al.The chromosome-based rubber tree genome provides new insights into spurge genome evolution and rubber biosynthesis. **Molecular Plant**, v. 13, n. 2, p.336-350, 2020.

LU, Y., et al. Targeted, efficient sequence insertion and replacement in rice.**Nature Biotechnology**, p.1-6, 2020.

MAHER, M.F. et al. Plant gene editing through de novo induction of meristems.**Nature biotechnology**, v. 38, n. 1, p.84-89, 2020.

MALNOY, M. et al. DNA-free genetically edited grapevine and apple protoplast using CRISPR/Cas9 ribonucleoproteins.**Frontiers in plant science**, v. 7, p.1904, 2016.

MALUF, M.P. et al. Altered expression of the caffeine synthase gene in a naturally caffeine-free mutant of Coffeaarabica.**Genetics and Molecular Biology**, v. 32, n. 4, p.802-810, 2009.

MANGHWAR, H.; LINDSEY, K.; ZHANG, X.; JIN, S. CRISPR/Cas system: recent advances and future prospects for genome editing. **Trends in plant science**, v. 24, n.12, p.1102-1125, 2019.

MCCARTY, N.S.; GRAHAM, A.E.; STUDENÁ, L.; LEDESMA-AMARO, R.Multiplexed CRISPR technologies for gene editing and transcriptional regulation**. Nature Communications**, v. 11, n. 1, p.1-13, 2020.

METJE-SPRINK, J.; SPRINK, T.; HARTUNG, F. Genome-edited plants in the field. **Current opinion in biotechnology**, v. 61, p.1-6, 2020.

MIEULET, D. et al. Unleashing meiotic crossovers in crops.**Nature plants**, v. 4, n. 12, p.1010-1016, 2018.

MISHRA, R.; JOSHI, R.K.; ZHAO, K. Base editing in crops: current advances, limitations and future implications. **Plant Biotechnology Journal**, v. 18, n.1, p.20-31, 2020.

MONCADA, M. D. P. et al. A genetic linkage map of coffee (Coffeaarabica L.) and QTL for yield, plant height, and bean size. **Tree genetics & genomes**, v. 12, n. 1, p.5, 2016.

MORADPOUR, M.; ABDULAH, S.N.A. CRISPR/dC as9 platforms in plants: strategies and applications beyond genome editing. **Plant Biotechnology Journal**, v. 18, n. 1, p.32-44, 2020.

MUÑOZ-SANZ, J.V. et al. Self-(in) compatibility systems: target traits for crop-production, plant breeding, and biotechnology. **Frontiers in Plant Science**, v. 11, 2020.

NAKAJIMA, I. et al. CRISPR/Cas9-mediated targeted mutagenesis in grape. **PLoS One**, v. 12, n. 5, p.e0177966, 2017.

NEALE, D.B. et al. Novel insights into tree biology and genome evolution as revealed through genomics. **Annual Review of Plant Biology**, v. 68, p.457-483, 2017.

OSAKABE, Y. et al. CRISPR–Cas9-mediated genome editing in apple and grapevine.**Nature protocols**, v. 13, n. 12, p.2844-2863, 2018.

PAIXÃO, J.F.R. et al. Improved drought stress tolerance in Arabidopsis by CRISPR/dCas9 fusion with a Histone AcetylTransferase.**Scientific reports**, v. 9, n.1, p.1-9, 2019.

PAYNE, R.M. et al. An NPF transporter exports a central monoterpeneindole alkaloid intermediate from the vacuole. **Nature plants,**v. 3, n. 2, p.1-9, 2017.

PINTO, R.T. et al. Genome-wide analysis, transcription factor network approach and gene expression profile of GH3 genes over early somatic embryogenesis in Coffea spp. **BMC genomics**,v. 20, n. 1, p.1-15, 2019.

QI, L.S. et al. .Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression.**Cell**, v. 152, n. 5, p.1173-1183, 2013.

QIN, L. et al. High-efficient and precise base editing of C• G to T• A in the allotetraploid cotton (Gossypiumhirsutum) genome using a modified CRISPR/Cas9 system.**Plant biotechnology journal**, v. 18, n. 1,p.45-56, 2020.

RAN, F.A. Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity.**Cell**, v. 154, n. 6, p.1380-1389, 2013.

ROZOV, S.M.; PERMYAKOVA, N.V.; DEINEKO, E.V.The Problem of the Low Rates of CRISPR/Cas9-Mediated Knock-ins in Plants: Approaches and Solutions. **International Journal of Molecular Sciences**, v. 20, n.13, p.3371, 2019.

SAGOR, G.H.M. et al. A novel strategy to produce sweeter tomato fruits with high sugar contents by fruit-specific expression of a single bZIP transcription factor gene. **Plant biotechnology journal**, v. 14, n. 4, p.1116-1126, 2016.

SARNO, R. et al. Programming sites of meiotic crossovers using Spo11 fusion proteins. **Nucleic acids research**, v. 45, n. 19, p.e164-e164, 2017.

SCALABRIN, S. et al. A single polyploidization event at the origin of the tetraploid genome of Coffeaarabica is responsible for the extremely low genetic variation in wild and cultivated germplasm. **Scientific reports**, v. 10, n. 1, p.1-13, 2020.

SOYK, S. et al. Variation in the flowering gene SELF PRUNING 5G promotes day-neutrality and early yield in tomato. **Nature Genetics**, v. 49, n.1,pp.162-168, 2017.

SPENGLER, R.N. Origins of the apple: the role of megafaunal mutualism in the domestication of Malus and rosaceous trees. **Frontiers in plant science**, v. 10, p.617, 2019.

STRECKER, J. et al. RNA-guided DNA insertion with CRISPR-associated transposases.**Science**, v. 365, n. 6448, p.48-53, 2019.

TAAGEN, E.; BOGDANOVE, A.J.; SORRELLS, M.E. Counting on crossovers: Controlled recombination for plant breeding. **Trends in Plant Science**, v. 25, n. 5, p.455-465, 2020.

TERRA, W. et al. Resistance of Conilon coffee cultivar Vitoria Incaper 8142 to Meloidogyneparanaensis under field conditions, **Experimental Agriculture**, v. 56, n.1, p. 88-93, 2019.

THALOR, S.K. et al. Deregulation of sucrose-controlled translation of a bZIP-type transcription factor results in sucrose accumulation in leaves. **PLoS One**, v. 7, n. 3, p.e33111, 2012.

VAN NOCKER, S.; GARDINER, S.E. Breeding better cultivars, faster: applications of new technologies for the rapid deployment of superior horticultural tree crops. **Horticulture Research,** v. 1, n.1, p.1-8, 2014.

WADA, N.; UETA, R.; OSAKABE, Y.; OSAKABE, K. Precision genome editing in plants: state-of-the-art in CRISPR/Cas9-based genome engineering. **BMC Plant Biology**, v. 20, n. 1, p.1-12, 2020.

WANG, X. et al. CRISPR/Cas9-mediated efficient targeted mutagenesis in grape in the first generation.**Plant biotechnology journal**, v. 16, n. 4, p.844-855, 2018.

WANG, Z. et al. Optimized paired-sgRNA/Cas9 cloning and expression cassette triggers high-efficiency multiplex genome editing in kiwifruit. **Plant biotechnology journa**l, v.16, n. 8, p.1424-1433, 2018.

WOO, J.W. et al. DNA-free genome editing in plants with preassembled CRISPR-Cas9 ribonucleoproteins. **Nature biotechnology**, v. 33, n.11, p.1162-1164, 2015.

XU, W. et al.The genome evolution and low-phosphorus adaptation in white lupin.**Nature communications**, v. 11, n.1, p.1-13, 2020.

YUSTE-LISBONA, F.J. et al.ENO regulates tomato fruit size through the floral meristem development network. **Proceedings of the National Academy of Sciences**, v. 117, n.14,pp.8187-8195, 2020.

ZETSCHE, B. et al. Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. **Cell**, v. 163, n. 3, p.759-771, 2015.

ZETSCHE, B. et al. Multiplex gene editing by CRISPR–Cpf1 using a single crRNA array.**Nature biotechnology,** v. 35, n. 1, p.31-34, 2017.

ZHANG, F.; LEBLANC, C.; IRISH, V.F.; JACOB, Y. Rapid and efficient CRISPR/Cas9 gene editing in Citrus using the YAO promoter**. Plant cell reports**, v. 36, n. 12, p.1883-1887, 2017.

ZSÖGÖN, A. et al. De novo domestication of wild tomato using genome editing.**Nature biotechnology**, v. 36, n.12, p.1211-1216, 2018.

ZSÖGÖN, A.; CERMAK, T.; VOYTAS, D.; PERES, L. E. P.Genome editing as a tool to achieve the crop ideotype and de novo domestication of wild relatives: case study in tomato.**Plant Science**,v. 256, p.120-130, 2017.

SUPPLEMENTARY MATERIAL

Supplementary table 1: digital format (available under request)

Supplemantary table 2: digital format (available under request)

Supplamentary table 3: digital format (available under request)

SCIENTIFIC PAPER 2: A COMPREHENSIVE INVENTORY OF MEMBRANE TRANSPORTERS ON COFFEE – OPENING THE EYES FOR METABOLITE ACCUMULATION MECHANISMS

ABSTRACT

As a major crop species, the coffee plant is subjected to consumers' preferences fluctuations, besides the effects of environmental instabilities, such as climate changes and pests rise. Although, unlike many other crops, *Coffea* species have a perennial habit, with relatively long life cycle and plantations that remains on the field for many years of harvest. The breeding programs cannot rely just on time consuming crossing experiments for the identification and achievement of desired phenotypic characteristics, therefore, to cope with this challenging scenario, genomic and transcriptomic data mining is an interesting strategy to gather relevant genetic information to be applied on coffee genetic improvement. Regarding this effort, plant membrane transporters are a key target, due to its influence on plant's adaptation to biotic and abiotic stresses, relatively high amount of information available for other organisms and an interesting pattern of function conservation within homologous proteins. Therefore, we provide here a *Coffea canephora* membrane proteins inventory, with annotations for more than 10,000 genes and analyzed transcriptomic information for the putative 1,848 transporter coding genes. We also performed a comprehensive gene co-expression analysis using this dataset in order to identify potential transporters that might be determinant for the metabolites accumulation of major influence on the beverage quality and bioactivity attributes. This report points to the avenue of possibilities on genomic and transcriptomic data mining for *Coffea* genetic improvement strategies, which can lead to adapted varieties for the environmentally and commercially sustainable production of coffee.

Keywords: Caffeine, genomics, specialized metabolism, gene co-expression network.

1 INTRODUCTION

Coffee plant is a worldwide relevant crop, due to the highly appreciated beverage made from its roasted and grinded seeds, which is one of the most consumed in the world. The main producer and exporter country is Brazil, where this commodity is ranked as the 5th most relevant agriculture-derived product in terms of international commercialization (MAPA, 2019; ICO, 2020).

The beans come from plantations of two species, *Coffea arabica*, which accounts for the bigger fraction of coffee exportation market (63.01%) and *Coffea canephora* (36.99%), a plant that is usually more resistant to pests and diseases and adapted to higher temperatures, but has a different bean chemical profile that affects the common sensorial characteristics of the beverage (VAN DER VOSSEN *et al*., 2015; FATOBENE *et al*., 2019; ICO, 2020).

The two species are perennial, it takes two to three years for the first production and, after this first period, it is possible to harvest beans once a year. The relatively long life cycle for a crop make the development of breeding programs a challenge in face of consumers demand changes, climatic instabilities and the rise of new pests and diseases.

In the sense of supporting *Coffea* species genetic improvement under this perspective, it is essential to generate and explore available genomic information (DENOEUD *et al*., 2014) for better comprehension of genetic characteristics that influence the target phenotype. On a strategic approach, the gain of knowledge by mining genomic and transcriptomic data can be futher confirmed and applied to breeding programs through molecular marker-assisted crossings for desired genetic introgressions or, directly, gene editing of elite cultivars, which is proved as feasible for *Coffea ssp*. (BREITLER *et al*., 2018).

To this end, a comprehensive annotation of putative genes' characteristics and functions is a necessary effort. The membrane proteins corresponds to a significant part of any organism proteome (NAGATA *et al*., 2008) and, conveniently for a annotation strategy, their function can be more conserved among different organisms, like the conserved substrate-type affinity within some homologous transporters of diverse plant species, due to structure-dependent interactions needed for the proper transport function (UPADHYAY *et al*., 2019; TANG *et al*., 2020).

Transporter coding genes are found in higher number on plant genomes, probably due to specialized metabolism evolution in this kingdom (JORGENSEN *et al*., 2017) and are determinant for many, if not all in some manner, physiological processes. The role of transporters on plant adaptations to abiotic and biotic stresses and the applicability of this knowledge for crop species are well demonstrated (FAN *et* al., 2016; KUROMORI *et* al., 2016; ESMAEILI *et al*., 2019 STEFANNELO *et al*., 2019; UPADHYAY*et al*., 2019).

Therefore, we provide here a complete inventory of membrane protein coding genes on *C. canephora* genome and a comprehensive discrimination of the membrane transporters for this species. Annotation information is provided for more than 10,000 genes and transcriptomic data based on all currently available *Coffea ssp.* RNA-seq experiments (151 libraries) is given for the 1,848 genes classified by our approach as membrane transporter coding.

For a crop that the consumption of its product is greatly dependent on the sensory profile and bioactivity determined by the substances stored in its seeds, the understanding of metabolites transport and accumulation mechanisms is essential. The main bioactive substance that influence coffee consumption worldwide is caffeine and, surprisingly, none transporter for this compound, which may be determinant for its accumulation in the cell, was ever identified. Besides caffeine, chlorogenic acids and diterpenes are also within the important molecules for the beverage quality and bioactivity found in coffee beans (LIANG*et al*., 2016; BARBOSA *et* al., 2019; CORNELIS, 2019).

On a primary analysis with the genomic and transcriptomic dataset provided here, we constructed a gene co-expression network using information about genes involved on the biosynthesis of the already cited compounds and the transporter coding genes. Based on the hypothesis that a transporter can be co-expressed with the enzyme which metabolizes its substrate (DOBRITZSCH *et al*., 2016; PAYNE *et al*., 2017; DERMUTAS *et al*., 2019) some putative genes that may affect the accumulation of important substances on *Coffea ssp.*were identified

This is the first report of a perennial plant genome-wide transporter inventory, with information being provided that can be explored beyond the case of study presented here. Moreover, it is the first study to explore caffeine, chlorogenic acids and diterpenes putative transporters on *Coffea*, a knowledge that can serve for further applications on coffee genetic improvement programs focused on this worldwide consumed beverage quality and bioactivity.

2 RESULTS

Using the *blastp* algorithm we found 7,555 sequences in *C. canephora* proteome that match to at least one protein on the Transporter Classification Database (TCDB, http://www.tcdb.org/), which were further classified into TCDB families based on their best hit from the *blastp* result. The whole proteome was also submitted to three different transmembrane domain (TMD) prediction softwares and, by applying a filter of TMD number equal or above two, we selected 5,231, 4,003 and 3,063 sequences from HMMTOP (http://www.enzim.hu/hmmtop/), SOSUI (http://harrier.nagahama-i-bio.ac.jp/sosui/) and PHOBIUS (http://phobius.sbc.su.se/) software outputs as potential membrane transporters (figure 1A).

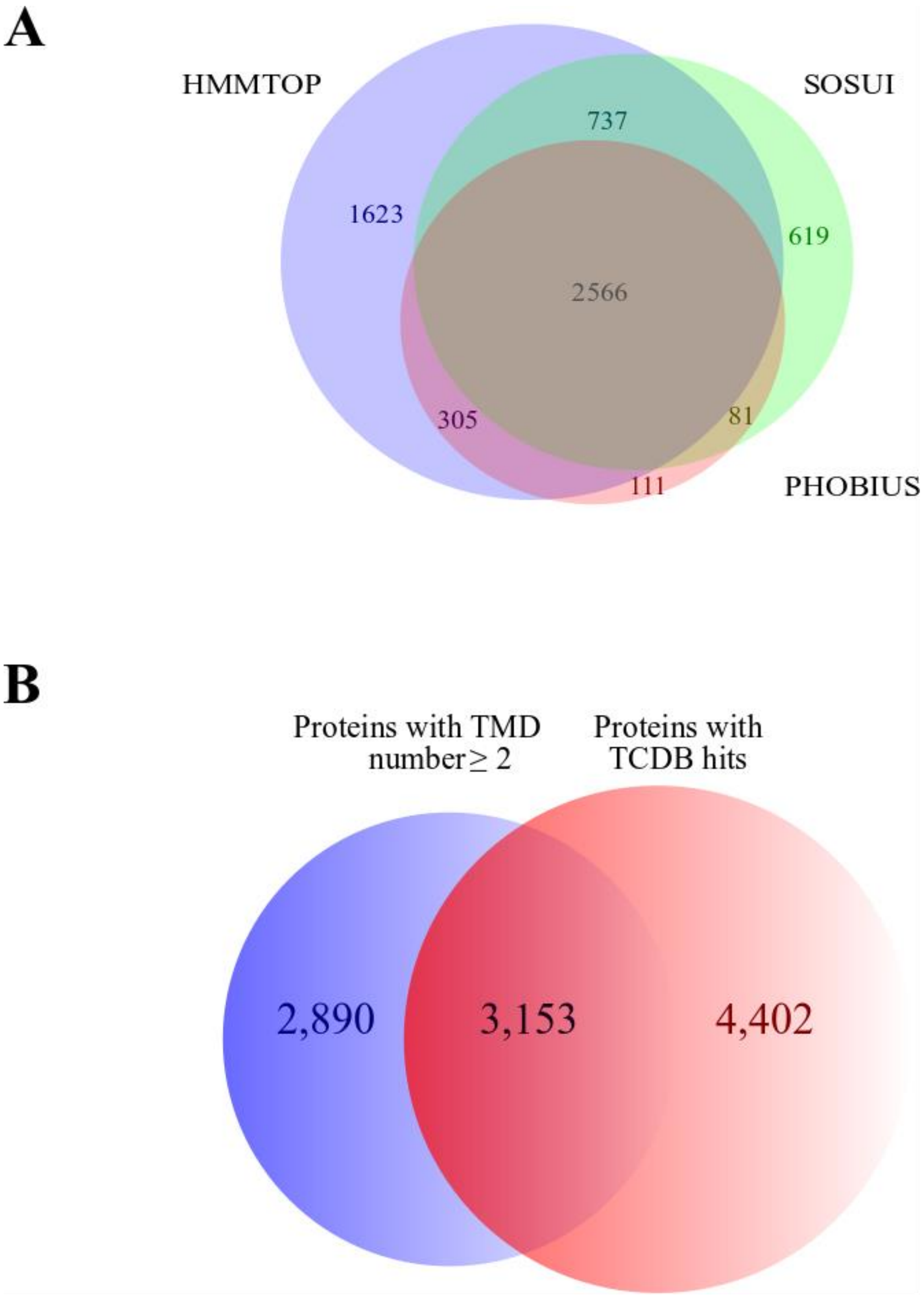
A total of 10.445 sequences were obtained from the union of these two analyses (TCDB hits and TMD number) (figure 1B) and a general annotation table was produced for them (supplementary table 1), including information about their best hit on Uniprot database (https://www.uniprot.org/), basic features (like predicted protein length and molecular weight) and subcellular localization (predicted using the DeepLoc software, http://www.cbs.dtu.dk/services/DeepLoc/).

We applied a classification score to discriminate the potential of these selected proteins on being a membrane transporter, in which the lowest potential (score "1") was given to the sequences with TMDs hits, but no TCDB hits; score "2" to sequences that have TCDB hits but none predicted TMD ; score "3" to the sequences with TCDB hits and predicted TMD number ≥ 2 in at least one software and, finally, a score "4" to the sequences with TCDB hits and TMD number ≥ 2 predicted by all the three softwares.

We found 1,847 *C. canephora* sequences with the highest classification score, which are the most probable membrane transporter proteins in this species. These were grouped by their subcellular localization to analyze the three most represented TCDB families, the average TMD number and the number of signal peptides per each cellular compartment enriched in the putative transporters group (table 1). In this set of proteins, we found 196 TCDB families and 188 transporters with 125 different signal peptides types.

In order to gain knowledge about the transcriptional regulation of the genes that codify these proteins, we decided to construct a database with all the publicly available RNA-seq datasets for coffee (*C. canephora* and *C. arabica* species) deposited on NCBI-SRA (National Center for Biotechnological Information – Sequence Read Arqchive). We found 152 transcriptome datasets, comprising diverse experimental conditions and plant tissues (supplementary table 2).

Figure 1 - Venn diagrams of C. canephora proteome-derived sequences classification



Legend: Venn diagrams of *C. canephora* proteome-derived sequences classification. 1A- the number of sequences with two or more TMDs predicted by each software and their intersections; 1-B the number of sequences with matches on TCDB and with two or more TMDs, predicted by at least one of the three prediction softwares and the intersection. Fonte: Do autor, 2020.

Table 1 - Distribution of the putative *C. canephora* membrane transporters (score “3”) among different cellular compartments, with some main features described.

Cellular compartment	Number of proteins	Overrepresented TCDB families*	TMD nº average	Signal peptides**
Cell membrane	886	2.A.1 (79); 3.A.1 (77); 2.A.17 (75)	8	101/78
End. reticulum	500	3.A.1 (34); 2.A.7 (32); 4.D.3 (21)	5	55/45
Lys-Vacuole	256	2.A.66 (36); 2.A.1 (31); 2.A.18 (11)	9	16/15
Plastid	106	2.A.7 (10); 3.A.1 (9)	6	8/8
Golgi apparatus	61	2.A.7 (9); 2.A.123 (7); 5.B.2 (5)	5	4/4
Mitochondria	26	3.A.3 (4); 2.A.1 (3); 3.A.1 (3)	5	1/1
Nucleus	2	N/A	3	0/0
Peroxisome	2	N/A	3	0/0
Total	1839	3.A.1 (131); 2.A.1 (129); 2.A.7 (121)	7	185/124

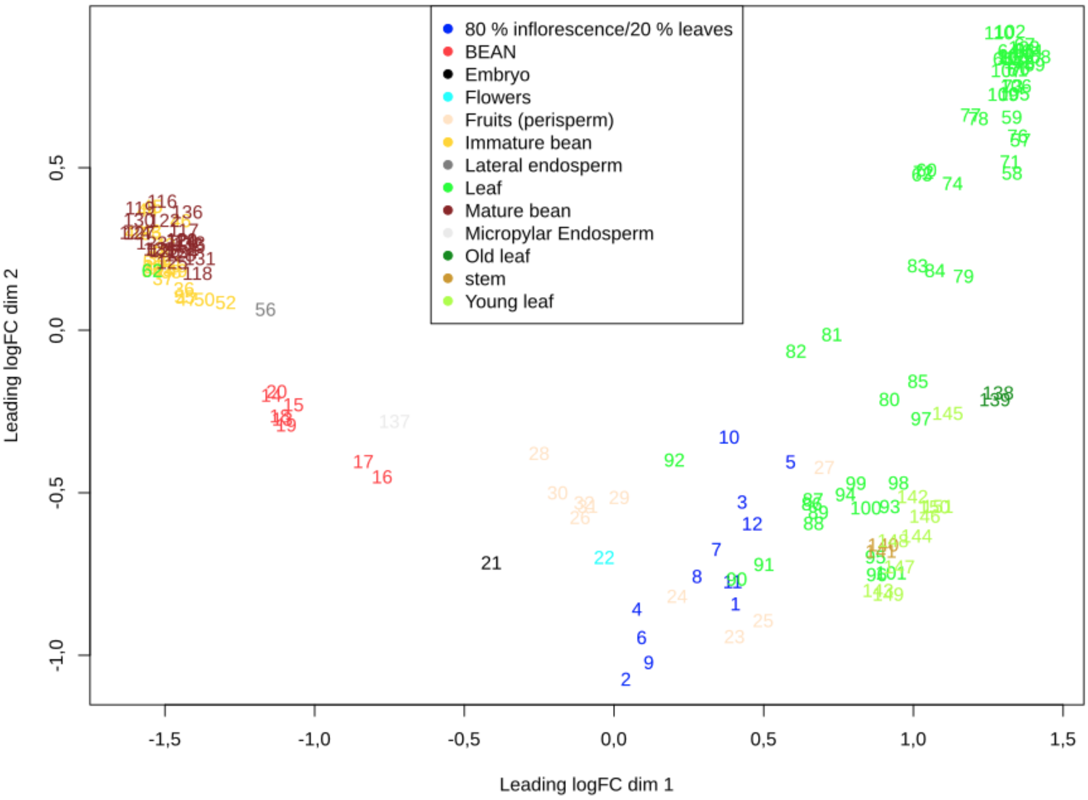
Legend: “End.reticulum” stands for endoplasmatic reticulum and “Lys-Vacuole” for lysosome or vacuole predicted proteins. Fonte: Do autor, 2020.
*Displayed as: TCDB family (number of proteins assigned to this family)
**Displayed as: number of proteins with signal peptides/number of different signal peptides

We analyzed all these RNA-seq data and, by using the Kallisto pseudoalignment (BRAY *et al*., 2016) with *C. canephora* genome as the reference, we have obtained the transcripts per million (TPM) value for each of the putative membrane transporters in all the samples (supplementary table 3). Among all the genes, nine were not expressed in any of the libraries and 20 were classified as tissue-especific for this dataset (supplementary table 4).

The 152 RNA-seq samples were distributed in 13 groups representing different plant tissues (supplementary table 5). Three of these tissue groups correspond to 69% of the samples: Leaf, with 59 samples; Imature and Mature Beans, both with 23 samples. Importantly, there were no samples for coffee roots that satisfied our criteria to be incorporated in the analysis.

The expression profile of each gene that codify the putative membrane transporters among all the RNA-seq samples analyzed was utilized to perform a biological coefficient variation (BCV) analyses, to evaluate the distribution of the datasets within the plant tissue-based classifiers (figure 2). In this analysis, the expression of all the 1839 genes in a given sample (the variable) is used to determine this sample’s position in relation to the others.

Figure 2 - Biological coefficient of variation analyses of C. canephora and C. arabica RNA-seq samples



Legend: Biological coefficient of variation analyses of the *C. canephora* and *C. arabica* RNA-seq samples utilized in this study, using the expression of the putative membrane transporters codifying genes and the classification of libraries into plant tissues. Number references can be found on supplementary table 5. Fonte: Do autor, 2020.

In the first dimension (horizontal axis) there is a clear distinction between leaf and bean-derived samples and a distinguishable difference among the other tissues, like flowers, inflorescence, embryo and perisperm. Some other specifications are also worth to highlight, like the differences within some of the bean-derived samples, in which the samples 16 and 17 that refers to green beans are clearly separated in dimension 1 from the yellow (13, 18 and 19) and red beans (14, 15 and 20) derived samples, these two groups having a slightly difference in dimension 2. It is interesting to note, however, that the membrane transporters gene expression is not suitable to distinguish the immature beans from the mature beans samples (figure 2, supplementary table 5).

In order to take advantage of this large dataset, the *C. canephora* membrane transporter annotation and transcriptome, we decided to perform an analysis to identify genes potentially involved on the transport and accumulation of some of the most important substances for the coffee beverage quality and bioactivity, the alkaloids caffeine (DENOEUD *et al*., 2014; MIZUNO *et al*., 2014) and trigonelline (MIZUNO *et al*., 2014), diterpenes like cafestol (SANT'ANA *et al*., 2017) and chlorogenic acids (MAHESH *et* al., 2007; LALLEMAND *et al*., 2012).

Genes already associated with the biosynthesis of these compounds in coffee were selected from previous studies. Their transcriptional profile on the 152 RNA-seq samples were combined with the data from the membrane transporters and a correlation analysis by Spearman's method was performed, using a threshold of $\rho \geq |0.7|$.

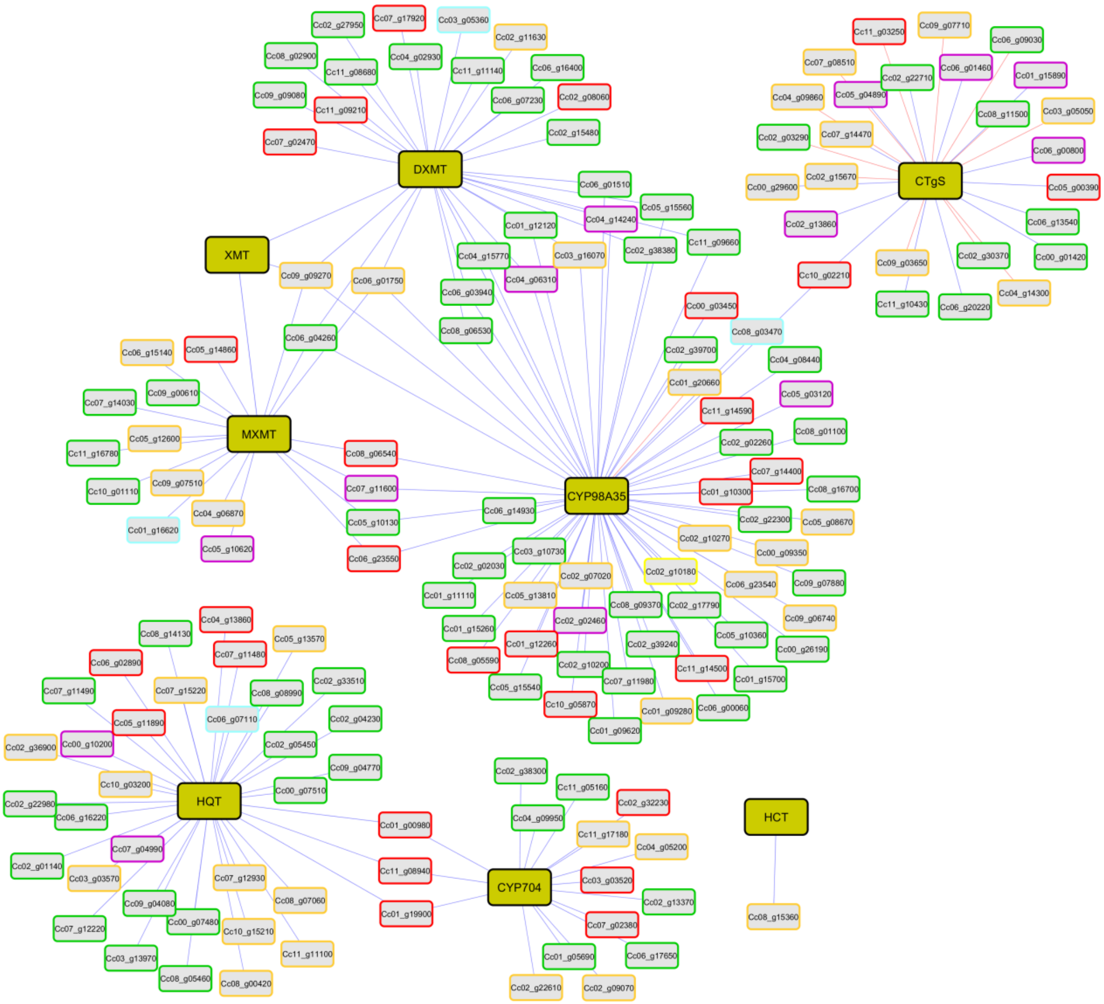
After the selection of all transporter coding genes significantly correlated (*p-value* ≤ 0.05) with the enzyme coding ones (supplementary table 6), we generated a gene co-expression network, for classes of beverage quality and bioactivity related substances. We enriched this network with information about the subcellular localization of the putative transporter, by the color code of the boxes' borders (figure 3) and all the complete annotation about each transporter is also provided (supplementary table 1). We found 100, 70 and 16 putative membrane transporters coding genes co-expressed with the chlorogenic acids, alkaloids, and diterpenes biosynthesis related genes, respectively. These connections resulted in structurally different sub-networks.

Similarly to what is observed for the putative transporters in general terms (table 1), most of the transporters genes in the co-expression network are predicted to cell membrane, while the second most common location is the endoplasmatic reticulum, followed by the lysosome/vacuole, plastid and golgi apparatus on the fifth position in this order.

Regarding alkaloid byosinthesis related genes (figure 3), it is interesting to note that two main webs for this substance class were formed, one for caffeine related genes and the other for *CTgS*, which is the gene for trigonelline biosynthesis enzyme. The genes that code for the caffeine biosynthetic ezymes are connected (*XMT* is the link between *MXMT* and *DXMT*) and there are three putative transporter genes connected to more than one of them, *Cc09_g09270* (TCDB family 9.A.43), *Cc06_g01750* (TCDB family 8.A.102) and *Cc06_g04260* (TCDB family 3.A.1). The first two genes are predicted to be located on endoplasmatic reticulum and the other on the cell membrane.

Trigonelline related subnetwork is the one formed with the highest number of negative correlations (12 out of a total of 26 connections). Among the positive correlations, five genes are predicted to be located on cell membrane, one at lysosome/vacuole and other five on plastid.

Figure 3 – Gene co-expression network of key enzyme-codifying genes and possible related transporters



Legend: Gene co-expression network between putative membrane transporters and characterized biosynthetic enzymes coding genes from alkaloid, chlorogenic acids and diterpenes routes. The correlations displayed are significant (*p-value* ≤ 0,05) and were obtained by Spearman's method ($\rho \geq |0.7|$). The blue edges refer to positive correlation and the red to negative. Boxes' border colors: green: cell membrane; red: lysosome/vacuole; orange: endoplasmatic reticulum; purple: plastid; yellow: cytoplasm; light blue: golgi apparatus. Fonte: Do autor, 2020.

Both caffeine and trigonelline related sub-networks have predicted transporter coding genes in common with *CYP98A35*, a gene involved in chlorogenic acids biosynthesis. *DXMT* shares 11 co-expressed transporter genes with *CYP98A35*, while *MXMT* shares four and *XMT* one. The three transporters that are connected to more than one caffeine biosynthesis genes are also connected to *CYP98A35* and, moreover, *DXMT* is directly connected to *CYP98A35* through a correlation of $\rho = 0.758$, slightly higher than the value found for *DXMT* and *XMT*, which is $\rho = 0.753$.

CYP98A35 forms the biggest sub-network, with 64 nodes, of which 33 corresponds to putative cell membrane localized transporters. The other two chlorogenic acids related enzyme coding genes forms unrelated sub-networks, in which *HCT* is connected with just one gene and *HQT* with 36. Three of the genes with positive correlation to *HQT* are also linked to *CYP704*, a cytochrome P450 family gene associated with diterpene biosynthesis and are predicted to vacuole subcellular localization. The diterpene biosynthesis related gene has positive correlations ($\rho \geq |0.7|$ and *p-value* ≤ 0.05) with other 13 transporter coding genes.

3 DISCUSSION

Regarding membrane transporters identification, we found that using basic structure analysis coupled with similarity search on a curated database is an interesting approach to select potential candidates. From the *C.* canephora proteome (DENOEUD *et al*., 2014), 1,848 proteins were classified as putative membrane transporters (score 4) by our selective criteria, a comparable number for what is reported for *A.* thaliana, *Oriza sativa* and *Medicago truncatula* (984, 1,200 and 1,114 proteins, respectively) (NAGATA *et al*., 2008; BENEDITO *et al*., 2010).

No other transporters inventory is available regarding tree species, which would be a better comparison to coffee. Regarding the organisms from others life kingdoms, like *E. coli* (354 transporters), *Sacharomices cerevisiae* (300), *Caernohabtidis elegans* (654) and *Homo sapiens* (754), plants have the highest number of membrane transporters proteins (NAGATA *et al*., 2008), what might be due to the specialized metabolism evolved in this kingdom (JORGENSEN *et al*., 2017).

This number (1847 putative membrane transporters out of 25,574 proteins) resulted from a primary analysis based on two filters, similarity with TCDB database proteins (SAIER *et al*., 2016) and structure composed of two or more transmembrane domains (TMDs). The last filter was applied to exclude proteins that are located on just one side of the membrane (monotopic) or span the lipid bilayer just once (bitopic), which are usually the most abundant integral membrane groups present in an organism, however, they account for the most of the receptors and proteins with enzymatic functions, besides unknown features (HUBERT *et al*., 2010; LOMIZE *et al*., 2017), what does not corresponds with our intention on generating a *C. canephora* membrane transporter inventory.

The number of proteins found with two or more TMDs was different comparing the prediction from the three softwares used (figure 1A), although we were able to select 3,153 proteins with this characteristic detected by all of them. Also, within the 6,042 sequences with TMD number ≥ 2, only 778 (12.88%) had the exact same TMD number predicted in HMMTOP (TUSNADY and SIMON, 2001), SOSUI (HIROKAWA *et al.*, 1998) and PHOBIUS (KALL *et al.*, 2007).

Among the proteins rated with score “4” in our analysis, this ratio of concordance between topology prediction softwares is increased to 27.88%. The discrepancy might be due to the differences on prediction algorithms, as HMMTOP uses a Hidden Markov model to predict the protein topology, SOSUI relies on the sequence physicochemical characteristics and PHOBIUS uses the same approach as HMMTOP, but with a signal peptide prediction algorithm coupled, to minimize overlap predictions (HIROKAWA *et al.*, 1998; TUSNADY and SIMON, 2001; KALL *et al.*, 2007). This information is described for each protein analyzed (supplementary table 1) and gives to this dataset more reliability.

We chose to not restrict the results to selected TCDB classes, to have a broad perspective on our data analysis, even though some TCDB families have members which do not directly transport molecules across the cell membranes, like the Glycan glucosyltransferase family (4.D.3), the third overrepresented family in the group of endoplasmatic reticulum localized proteins (table 1), which have members like 4.D.3.1.4 (accession O48946) that is characterized as part of a cellulose synthase complex, anchored in the membrane through 8 TMDs (PERSSON *et al.*, 2007; WATANABE *et al.*, 2015). The specific TCDB family information about each *C. canephora* protein is described in supplementary table 1.

The ATP binding cassete (ABC, 3.A.1) was the most common TCDB family found in the putative membrane transporters dataset (score “4”), with 131 members out of 1839 sequences (7.12%). Despite being ranked as the most common, the other two overrepresented families have similar proportions of representativeness, 7.01% of the proteins are assigned as members of Major Facilitator superfamily (2.A.1) and 6.58% of Drug/Metabolite transporters (2.A.7). Interestingly, the overrepresented families’ distribution varied among different cellular compartments (table1).

We decided to enrich this coffee membrane transporters dataset with transcriptomic information and, to do so, all the publicly available RNA-seq data (151 libraries) related to *C. canephora* and *C. arabica* were analysed and incorporated on the annotation. The BCV result briefly illustrates the diversity in expression patterns of coffee membrane transporter coding genes among different tissues, in accordance with the expected, as the transcriptional profile of this class of genes can be highly specific to tissue (the cell context) and environmental stimuli, as well as correlated to the substrate availability (TANG *et al*., 2020).

This transcriptional regulation of transporters coding genes is essential on the metabolism fine-tuning during the plant development and interaction with the environment, as already demonstrated in many examples, such as in response to abiotic stresses like drought and phosphate starvation (KUROMORI *et al*., 2016;CHANG *et al*., 2019), accumulation control of specialized metabolites (DEMURTAS *et al*., 2019) and developmental processes regulation (MICHNIEWICZ *et al*., 2019, GAN *et al*., 2019). However, it is important to point that some transporter genes might be ubiquitously expressed within different tissues and conditions, with no clear demonstration of regulation upon stimuli.

Within the RNA-seq dataset constructed here, there are libraries in which the overall transporters genes transcriptional profile also varies in accordance with different treatment conditions. For example, using the overall transporters genes expression pattern it is possible to differentiate the RNA-seq libraries related to somatic embryogenesis process steps (QUINTANA-ESCOBAR *et al*., 2019) and bean developmental stages (CHENG *et al*., 2018) (supplementary figures S1 and S2).

Interestingly, regarding somatic embryogenesis process experiment (QUINTANA-ESCOBAR *et al*., 2019), the samples collected before somatic embryogenesis induction treatment (14, nine and zero days before induction) are clearly horizontally distant from the ones after this point (one, two and 21 days after induction) and, within the post treatment samples, the ones collected at last time point (21 days after induction) are vertically distant from the others (supplementary figure S2). These samples from the last time point after embryogenesis induction are pointed by the authors as the representatives of the embryogenic state initiation (QUINTANA-ESCOBAR *et al*., 2019).

This observed influence of transporters is in line with the known relevance of these proteins on somatic embryogenesis, for example, mediating phytohormone and their conjugates transport, like the auxin transport through PIN family proteins (MARQUEZ-LOPES *et al*., 2018; XU *et al*., 2019), which is essential for the swifts on cell fate, besides the other molecules from primary and specialized metabolism that play a role on this process.

The same effectiveness on sample’s distribution on BCV analysis by overall transporters expression profile is noted on a bean developmental stages experiment (CHENG *et al*., 2018), where the libraries derived from green beans are clearly distant from the yellow and red beans derived ones, indicating that the transporters recruitment and, consequently, the general molecules transported in advanced fruit development stages (yellow and red beans) are closely related and diverge from that found for green beans.

This putative pattern of metabolite transport might occurred in reason of the more significant biochemical swifts on the transition from the green stage to the yellow one, which marks, for example, the end of perisperm conversion into endosperm, whereas the maturation process that begins on yellow stage persists on the red beans (DE CASTRO and MARRACCINI, 2006).

The transport of specialized metabolites between cells and within organelles is poorly explorated in *Coffea*spp., although it is of major importance, as the accumulation of substances during bean maturation determines quality attributes of the beverage. This process is dependent on substrates trafficking among organelles for biosynthesis pathways supply and compartmentalization on vacuoles for storage, all this can be performed by membrane transporters (SHITAN *et al.*, 2014a; SHITAN *et al.*, 2014b; PAYNE *et al.*, 2017; DASTMALCHI *et al*., 2019).

Even though not fully comprehended, the biosynthesis of some major substances related to coffee beverage quality and bioactivity has been explored and some genes that putatively codify enzymes involved on alkaloids, chlorogenic acids and diterpenes biosynthetic pathways are already identified (DENOEUD *et al.*, 2014; MIZUNO *et al*., 2014; LALLEMAND *et al*., 2012; MAHESH *et* al., 2007; SANT'ANA *et al.*, 2017).

Interestingly, the transcriptional profile of the genes that codify the biosynthetic enzymes and the transporter of the derived product can be correlated at some level, as well as a correlation between the substrate production and the transporter gene expression can be observed. As an example, this was already demonstrated for the accumulation of crocins in *Crocus sativus*(saffron) stigmas, in which an ABCC transporter gene is co-expressed both with one of the enzymes of the biosynthesis route and with the crocin itself and, through assays in heterologous systems (yeast and *Nicotiana tabacum*), its capability of transporting the metabolite was confirmed (DEMURTAS *et al.*, 2019).

The same hypothesis of transporter gene co-expression with biosynthetic enzymes on the potential of transporting the metabolite was tested and confirmed for coumarylagmamite transport through a MATE family protein and strictosidine through a NPF family transporter (DOBRITZSCH *et al.*, 2016; PAYNE *et al.*, 2017). Therefore, we decided to take advantage of the transporters annotation and transcriptome to identify candidate genes related to the transport of important coffee metabolites already mentioned, by constructing gene co-expression networks with biosynthetic enzyme genes found on literature and our dataset of putative membrane transporters (figure 3).

Although the enzymes XMT, MXMT and DXMT act on different steps of caffeine biosynthesis, their genes were found to be co-expressed in this dataset of 151 RNA-seq libraries and, interestingly, not co-expressed at a threshold of $\rho \geq |0.7|$ with the trigonelline synthase gene (*CTgS2*) (figure 3). Regarding the caffeine biosynthesis genes, it is already documented that their expression pattern is similar during coffee beans development stages (PERROIS *et al*., 2015) and that xanthosine conversion into caffeine is thought to occur on the same cellular environment, the cytosol (OGAWA *et al*., 2000; KUMAR *et al*., 2007), what might contribute to explain the co-regulation.

Like other alkaloids, caffeine might be stored at vacuoles and it is assumed that its biosynthesis occurs in the same organ where it accumulates (SHITAN *et al*., 2014a;ASHIHARA *et al*., 2017). On the beans, the biosynthesis starts at the perisperm, which is after converted to endosperm where the metabolite is stored, therefore, we hypothetize that at least one transporter at the cell membrane and other at the vacuole are needed for caffeine accumulation in coffee beans. Remarkably, we found one transporter gene that is co-expressed with two caffeine synthesis genes and its protein is predicted to be localized at the cell membrane, *Cc06_g04260* and, interestingly, is assigned to the ABCB subfamily (TCDB 3.A.1.201), which is one out of few membrane transporters families with members that mediate alkaloid transport in plants (figure 3) (SHITAN *et al*., 2014a).

Four putative vacuolar membrane transporters were co-expressed with *CcDXMT* (the gene that codes for the last enzyme on caffeine synthesis) and, among them, one assigned to a protein family that also have characterized alkaloid transporters members (MATE family, 2.A.66.1) (SHITAN *et* al., 2014b; UPADHYAY*et al*., 2019). Another two putative members of this family, *Cc06_g13540* and *Cc05_g00390*, respectively located on vacuole and on the cell membrane, are co-expressed with the trigonelline synthase gene and has the potential to be associated with its transport across cells and organelles (figure 3A).

Diterpenes, another important class of substances for coffee beverage quality, are part of the lipid fraction, which corresponds to a significant part of the grain weight. In our gene co-expression analysis, we found an web where 16 transporters are connected to the *C. canephora* homolog of *CaCYP704*, a cytochrome P450 gene with a co-localized SNP associated with variation in cafestol content, an exclusive *Coffea* genus diterpene (SANT'ANA *et al*., 2017).

Within this sub-network, we found one transporter from the ABC family (*Cc06_g17650*), the same family as the tobacco diterpene and sesquiterpene transporter NtPDR1 (CROUZET *et al*., 2013; PERMAN *et al*., 2017). None of the transporters in this network are located on the plastid membrane, a target location on the purpose of finding diterpenes transporters, as part of these metabolites are produced through MEP pathway, which occurs in this organelle in plants (LITCHTENTHALER, 1999; BATHE and TISSIER, 2019) and, moreover, a plastidial transporter is thought to be essential for the crosstalk between the two terpenoid metabolic routes (MVA and MEP pathways) which, intringly, is not yet identified (PICK and WEBER, 2014; HENRY *et al*., 2018).

Similarly, even though chlorogenic acids are one of the major remaining sources of antioxidants in coffee beans after the roasting, grinding and infusion process (YASHIN *et al*., 2013), there is no identified transporter protein for this class of metabolites. Therefore, we investigated possible candidate transporters through gene co-expression analysis with *HCT*, *HQT* and *CYP98A35*, all of them coding enzymes related to chlorogenic acids biosynthesis (MAHESH *et* al., 2007; LALLEMAND *et al*., 2012).

It was demonstrated that HQT uses quinic acids as substrates to CQAs production and HCT, shikimic acids for CSA. The last enzyme is also able to catalyze the production of the diester 3,5-dicaffeoylquinic acid (3,5-diCQA) (LALLEMAND *et al*., 2012). Conversely, *HQT* can perform the production of di-CQA and can be localized in two different cellular compartments, cytosol and vacuole, in tomato (MOGLIA *et al*., 2014).

There are evidences pointing to that *HQT* genes are more associated with chlorogenic acids production than *HCT*, which might be more determinant for lignin biosynthesis route (VOLPI e SILVA *et al*., 2019), what may explain their different expression profile found here and, consequently, the different number of connections with transporter genes. We found 35 putative membrane transporters genes co-expressed with *HQT* and no effective transporter gene co-expressed *HCT*. The gene *Cc08_g15360* that is co-regulated with *HCT* is assigned as a member of plant NADH oxidase family (5.B.1.1), therefore, probably not associated with the transport of chlorogenic acids.

Despite the lack of information specifically about chlorogenic acids transporters, other phenolic compounds such as flavonoids have already characterized transporters, many of them belonging to the MATE family (MARINOVA *et al*., 2007; PEREZ-DIAS *et al*., 2014; CHEN *et al*., 2018). Also relevant, one member of this same transporters family in *A. thaliana*, AtDTX18, is a hydroxycinnamic acid derivate (such as chlorogenic acids are) transporter (DOBRITZSCH *et al*., 2016).

Therefore, due to the transport of structurally similar compound, it is possible that members of MATE family are capable of mediate chlorogenic acids transport and, for this perspective, we found two genes co-expressed with *HQT* assigned as members of this family, *Cc00_g07510* and *Cc05_g11890*, predicted on the cell membrane and vacuole locations, respectively and two co-expressed with *CYP98A35*, one on the plastid, *Cc07_g11600* and the other on vacuole, *Cc07_g14400*.

The size of the sub-network related to *CYP98A35* might indicate its action plasticity in this specialized metabolism. Indeed, it was demonstrated that unlike other enzymes from the same class, it has potential to metabolize both p-coumaroylquinates and shikimates with the same efficiency (MAHESH *et* al., 2007). Interestingly, this gene is co-expressed with the last caffeine biosynthesis gene, *DXMT* and, also, is connected to some transporters in the network that are also linked to *DXMT*, *MXMT* and *CTgS*, all of them alkaloid synthesis related genes.

This observed relation between alkaloid and chlorogenic acids biosynthesis related genes and transporters, based on an expression profile composed of an extensive amount of RNA-seq libraries, is an interesting finding. It was already demonstrated that caffeine might form stable complexes with chlorogenic acids in the cell and might accumulate like this in vacuoles (WALDHASER and BAUMAN, 1996; D'AMELIO *et* al., 2015; BELAY*et al*., 2015). In line with this, within *Coffea*s pecies, the content of caffeine is correlated with chlorogenic acids (CLIFFORD *et* al., 1989; JESZKA-SKOWRON *et al*., 2016).

The gene co-expression network is in accordance with these observations and, within the identified transporters; one might perform the transport of this caffeine-chlorogenate complex, influencing the accumulation of two substances of major importance for coffee quality and bioactivity at the same time.

4 METHODS

Identification of membrane integral proteins and annotation of *Coffea canephora* membrane transporters

The *C. canephora* proteome was screened for proteins with transmembrane domains through three different softwares of membrane proteins topology prediction, HMMTOP, SOSUI and PHOBIUS. The outputs from each of the softwares were analyzed and combined to select potential membrane integral proteins, as well as the divergences between the predictions (figure 1). Venn diagrams were constructed using Meta-chart web tolls (https://www.meta-chart.com/venn).

The same proteome was used as a query on a blastp analysis against the entire TCDB dataset (http://www.tcdb.org/download.php). The output was analyzed and combined with the transmembrane domain analysis result (figure 1B). The union of these two analyses formed the group of 10,445 proteins that were used for another blastp analysis against Uniprot Uniref90 dataset and submitted to Sequence Manipulation Suite (https://sites.ualberta.ca/~stothard/javascript/) webserver for protein length and molecular weight information and to DeepLoc software (http://www.cbs.dtu.dk/services/DeepLoc/) for protein subcellular localization prediction.

A filter of transmembrane domain number equal or above two and existence of a TCDB hit result from the blastp analysis was applied to the 10,445 previously selected proteins to distinguish the sequences with higher potential of classification as membrane transporters. All the data was organized on a comprehensive table for the better use of scientific community (supplementary table 1).

Incorporation and analysis of transcriptomic information

The database Sequence Read Archive (SRA) from National Center for Biotechnological Information (NCBI) was used for searching all the RNA-seq data produced with Ilumina platform for *C. canephora* and *C. arabica*, both on single-end and paired-end configurations. 151 samples were found (supplementary table 2) and downloaded to Cyverse Discovery Environment webserver (https://de.cyverse.org/de/) through NCBI-SRA-fastq-dump-2.8.1 application. Each file was then transferred to Cyverse DNA Subway platform, where the fastqc program was executed to analyse the data. After that, through the FastX-Toolkit 0.0.14 application, the reads were trimmed, following the default parameters. Finally, Kallisto program (BRAY *et al*., 2016) was executed for each sample, using *C. canephora* genome (DENOEUD *et al*., 2014) as reference and the TPM values generated for each gene regarding each sample were downloaded and organized into a comprehensive table for the membrane transporters (supplementary table 3).

To analyse the influence of transporter coding genes expression on samples distribution, a biological coefficient variance analysis was performed and a MDS plot was produced, using the EdgeR package.

Gene co-expression network construction

For construction of the gene co-expression network, the TPM values for genes that codify enzymes related to major substances for coffee quality and bioactivity were obtained for the same cited 151 RNA-seq samples. These values were combined with those for putative membrane transporters and a Spearman's correlation analysis was performed using R software. To verify the significance of the correlation through a *p-value*, a formula was applied to the data generated: P = IDIST{ABS[r/SQRT({1-r*r}/{n-2})],[n-2],2}, using excel (USADEL *et al*., 2009).

The data, together with enrichment informations (supplementary table 1), were imported to Cytoscape 3.7 software to generate the gene co-expression network.

REFERENCES

ASHIHARA, H.; MIZUNO, K.; YOKOTA, T.; CROZIER, A.Xanthine alkaloids: occurrence, biosynthesis, and function in plants.**Progress in the Chemistry of Organic Natural Product**s, v.105, p. 1-8, 2017.

BARBOSA, M. D. S. G.; DOS SANTOS SCHOLZ, M. B.; KITZBERGER, C. S. G.; DE TOLEDO BENASSI, M. Correlation between the composition of green Arabica coffee beans and the sensory quality of coffee brews. **Food chemistry**, v. 292, p.275-280, 2019.

BATHE, U.; TISSIER, A. Cytochrome P450 enzymes: a driving force of plant diterpene diversity. **Phytochemistry,** 2019.

BELAY, A.; KIM, H.K.;HWANG, Y.H.Binding of caffeine with caffeic acid and chlorogenic acid using fluorescence quenching, UV/vis and FTIR spectroscopic techniques. **Luminescence**, v. 31, n. 2, p. 565-572, 2016.

BRAY, N.L., PIMENTEL, H., MELSTED, P. AND PACHTER, L. Near-optimal probabilistic RNA-seq quantification. **Nature biotechnology**, v. 34, n. 5, p.525-527, 2016.

BREITLER, J.C. et al. CRISPR/Cas9-mediated efficient targeted mutagenesis has the potential to accelerate the domestication of Coffeacanephora. **Plant Cell, Tissue and Organ Culture (PCTOC)**, v. 134, n. 3, p.383-39, 2018.

CHANG, M.X. et al. OsPHT1; 3 mediates uptake, translocation, and remobilization of phosphate under extremely low phosphate regimes. **Plant physiology**, v. 179, n. 2, pp.656-670, 2019.

CHEN, S.Y. et al. FaTT12-1, a multidrug and toxin extrusion (MATE) member involved in proanthocyanidin transport in strawberry fruits.**Scientiahorticulturae**, v. 231, pp.158-165, 2018.

CHENG, B.; FURTADO, A.; HENRY, R. J.The coffee bean transcriptome explains the accumulation of the major bean components through ripening. **Scientific reports**, v. 8, n. 1, p.1-11, 2018.

CHEYNIER, V. et al. Plant phenolics: recent advances on their biosynthesis, genetics, and ecophysiology. **Plant Physiology and Biochemistry**, v. 72, pp.1-20, 2013.

CLIFFORD, M. N.; WILLIAMS, T.; BRIDSON, D.Chlorogenic acids and caffeine as possible taxonomic criteria in Coffea and Psilanthus.**Phytochemistry,** v. 28, n. 3, pp.829-838, 1989.

CROUZET, J. et al. NtPDR1, a plasma membrane ABC transporter from Nicotianatabacum, is involved in diterpene transport. **Plant molecular biology**, v. 82, n.1-2, p.181-192, 2013.

D’AMELIO, N. et al. NMR reinvestigation of the caffeine–chlorogenate complex in aqueous solution and in coffee brews**. Food Biophysics**, v. 4, n. 4, p.321-330, 2009.

D'AMELIO, N. et al. NMR studies of hetero-association of caffeine with di-O-caffeoylquinic acid isomers in aqueous solution. **Food biophysics**, v. 10, n. 3, pp.235-243, 2015.

DASTMALCHI, M. et al. Purine permease-type benzylisoquinoline alkaloid transporters in opium poppy.**Plant physiology**, v. 181, n. 3, p.916-933, 2019.

DE CASTRO, R.D.; MARRACCINI, P.Cytology, biochemistry and molecular changes during coffee fruit development. **Brazilian Journal of Plant Physiology**, v. 18, n.1, p.175-199, 2006.

DEMURTAS, O.C. et al. ABCC transporters mediate the vacuolar accumulation of crocins in saffron stigmas. **The Plant Cell**, v. 31, n. 11, p.2789-2804, 2019.

DENOEUD, F. et al.The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. **Science**, v. 345, n.6201, p.1181-1184, 2014.

DOBRITZSCH, M. et al. MATE transporter-dependent export of hydroxycinnamic acid amides.**The Plant Cell**, v. 28, n. 2, p.583-596, 2016.

FAN, X.et al. Overexpression of a pH-sensitive nitrate transporter in rice increases crop yields. **Proceedings of the National Academy of Sciences**, v. 113, n. 26, p.7118-7123, 2016.

FATOBENE, B. J.; ANDRADE, V.T.; GONÇALVES, W.; GUERREIRO FILHO, O.L.I.V.E.I.R.O. Coffeacanephora clones with multiple resistance to Meloidogyne incognita and M. paranaensis. **Experimental Agriculture**, v. 55, n. 3, p.443-45, 2019.

FLOREZ, J.C. et al..High throughput transcriptome analysis of coffee reveals prehaustorial resistance in response to Hemileiavastatrix infection. **Plant molecularbiology**, v. 95, n. 6, p.607-623, 2017.

GAN, Z. et al. Downregulation of the auxin transporter gene SlPIN8 results in pollen abortion in tomato. **Plant molecular biology**, v. 99, n. 6, p.561-573, 2019.

GUERRERO, G.; SUÁREZ, M.; MORENO, G. Chlorogenic acids as a potential criterion in coffee genotype selections.**Journal of agricultural and food chemistry**, v. 49, n. 5, p.2454-2458, 2001.

HENRY, L.K. et al. Contribution of isopentenyl phosphate to plant terpenoid metabolism.**Nature plants**, v. 4, n.9, p.721-729, 2018.

HIROKAWA, T.; BOON-CHIENG, S.; MITAKU, S.SOSUI: classification and secondary structure prediction system for membrane proteins. **Bioinformatics (Oxford, England)**, v. 14, n. 4, p.378-379, 1998.

HU, G.L.; WANG, X.; ZHANG, L.; QIU, M.H.The sources and mechanisms of bioactive ingredients in coffee.**Food & function**, v. 10, n.6, p.3113-3126, 2019.

HUBERT, P. et al. Single-spanning transmembrane domains in cell growth and cell-cell interactions: More than meets the eye?.**Cell adhesion & migration**, v.4, n.2, p.313-324, 2010.

INTERNATIONAL COFFEE ORGANIZATION. Coffee market report, available on http://www.ico.org, accessed on may, 2020.

IVAMOTO, S.T.et al. Diterpenes biochemical profile and transcriptional analysis of cytochrome P450s genes in leaves, roots, flowers, and during Coffeaarabica L. fruit development.**Plant Physiology and Biochemistry**, v. 111, p.340-347, 2017.

JESZKA-SKOWRON, M.; SENTKOWSKA, A.; PYRZYŃSKA, K.; DE PEÑA, M.P. Chlorogenic acids, caffeine content and antioxidant properties of green coffee extracts: influence of green coffee bean preparation. **European Food Research and Technology**, v. 242, n.8, p.1403-1409, 2016.

JØRGENSEN, M.E. et al. Origin and evolution of transporter substrate specificity within the NPF family.**Elife,**v. 6, 2017.

KÄLL, L.; KROGH, A.; SONNHAMMER, E.L. Advantages of combined transmembrane topology and signal peptide prediction—the Phobius web server. **Nucleic acids research**, v. 35, p.W429-W432, 2007.

KUROMORI, T. et al. Overexpression of AtABCG25 enhances the abscisic acid signal in guard cells and improves plant water use efficiency**. Plant Science,**v. 251, p.75-81, 2016.

LALLEMAND, L.A. et al.A structural basis for the biosynthesis of the major chlorogenic acids found in coffee.**Plant Physiology**, v. 160, n. 1, p.249-260, 2012.

LIANG, N.; XUE, W.; KENNEPOHL, P.; KITTS, D.D. Interactions between major chlorogenic acid isomers and chemical changes in coffee brew that affect antioxidant activities. **Food chemistry**, v. 213, p.251-259, 2016.

LICHTENTHALER, H.K. The 1-deoxy-D-xylulose-5-phosphate pathway of isoprenoid biosynthesis in plants.**Annual review of plant biology**, v. 50, n. 1, p.47-65, 1999.

LOMIZE, A.L., HAGE, J.M. AND POGOZHEVA, I.D.Membranome 2.0: database for proteome-wide profiling of bitopic proteins and their dimers. **Bioinformatics,**v. 34, n. 6, p.1061-1062, 2018.

MAHESH, V. et al. Functional characterization of two p-coumaroyl ester 3′-hydroxylase genes from coffee tree: evidence of a candidate for chlorogenic acid biosynthesis. **Plant molecular biology**, v. 64, n. 1-2, p.145-159, 2007.

MARINOVA, K. et al. The Arabidopsis MATE transporter TT12 acts as a vacuolar flavonoid/H+-antiporter active in proanthocyanidin-accumulating cells of the seed coat. **The Plant Cell**, v. 19, n.6, p.2023-2038, 2007.

MÁRQUEZ-LÓPEZ, R.E. et al.Localization and transport of indole-3-acetic acid during somatic embryogenesis in Coffeacanephora.**Protoplasma**, v. 255, n. 2, p.695-708, 2018.

MICHNIEWICZ, M, et al. Transporter of IBA1 links auxin and cytokinin to influence root architecture.**Developmental cell**, v. 50, n.5, p.599-609, 2019.

MIKIHIRO, O.; YUKA, H.; NOZOMU, K. 7-Methylxanthine methyltransferase of coffee plants.**Journal of Biological** Chemistry, v. 276, n. 11, p.8213-8218, 2001.

MINISTÉRIO DA AGRICULTURA PECUÁRIA E ABASTECIMENTO. AGROSTAT – Estatística de Comércio Exterior do Agronegócio Brasileiro, available on http://indicadores.agricultura.gov.br/agrostat/index.htm, accessed on july, 2020.

MIZUNO, K. et al. Conversion of nicotinic acid to trigonelline is catalyzed by N-methyltransferase belonged to motif B′ methyltransferase family in Coffeaarabica. **Biochemical and biophysical research communications**, v. 452, n.4, p.1060-106, 2014.

MOGLIA, A. et al. Dual catalytic activity of hydroxycinnamoyl-Coenzyme A quinatetransferase from tomato allows it to moonlight in the synthesis of both mono-and dicaffeoylquinic acids. **Plant Physiology**, v. 166, n.4, p.1777-1787, 2014.

NAGATA, T., IIZUMI, S., SATOH, K. AND KIKUCHI, S. Comparative molecular biological analysis of membrane transport genes in organisms. **Plant molecular biology**, v. 66, n. 6, p.565-585, 2008.

PAYNE, R.M. et al.An NPF transport**er exports a central monoterpeneindole alkaloid intermediate from the vacuole.** Nature plants, v. 3, n. 2, p.1-9, 2017.

PÉREZ-DÍAZ, R. et al. VvMATE1 and VvMATE2 encode putative proanthocyanidin transporters expressed during berry development in Vitisvinifera L. **Plant cell reports,**v. 33, n. 7, pp.1147-1159, 2014.

PERROIS, C. et al. Differential regulation of caffeine metabolism in Coffeaarabica (Arabica) and Coffeacanephora (Robusta).**Planta**, v.241, n.1, p.179-191, 2015.

PERSSON, S. et al. Genetic evidence for three unique components in primary cell-wall cellulose synthase complexes in Arabidopsis**.Proceedings of the National Academy of Sciences,** v. 104, n. 39, p.15566-15571, 2007.

PICK, T.R.; WEBER, A.P. Unknown components of the plastidialpermeome.**Frontiers in plant science**, v. 5, p.410, 2014.

PIERMAN, B. et al. Activity of the purified plant ABC transporter NtPDR1 is stimulated by diterpenes and sesquiterpenes involved in constitutive and induced defenses. **Journal of Biological Chemistry**, v. 292, n.47, p.19491-19502, 2017.

QUINTANA-ESCOBAR, A.O. et al.Transcriptome analysis of the induction of somatic embryogenesis in Coffeacanephora and the participation of ARF and Aux/IAA genes.**PeerJ**, v. 7, p.e7752, 2019.

SAIER JR, M.H. et al. The transporter classification database (TCDB): recent advances. **Nucleic acids research**, v. 44, n.D1, p.D372-D379, 2016.

SANT'ANA, G.C. et al. Genome-wide association study reveals candidate genes influencing lipids and diterpenes contents in Coffeaarabica L. **Scientific reports**, v. 8, v.1, p.1-12, 2018.

SHITAN, N.; KATO, K.; SHOJI, T. Alkaloid transporters in plants. **Plant Biotechnology**, p.14-1002, 2014.

SHITAN, N. et al. Involvement of the leaf-specific multidrug and toxic compound extrusion (MATE) transporter Nt-JAT2 in vacuolar sequestration of nicotine in Nicotianatabacum. **PLoS One**, v. 9, n.9, 2014.

SILVA, N.; MAZZAFERA, P.; CESARINO, I.Should I stay or should I go: are chlorogenic acids mobilized towards lignin biosynthesis?. **Phytochemistry,**v. 166, p.112063, 2019.

SIMKIN, A.J.; KUNTZ, M.; MOREAU, H.; MCCARTHY, J. Carotenoid profiling and the expression of carotenoid biosynthetic genes in developing coffee grain.**Plant Physiology and Biochemistry,** v. 48, n.6, p.434-442, 2010.

STEFANELLO, N. et al. Coffee, caffeine, chlorogenic acid, and the purinergic system.**Foodand Chemical Toxicology,**v. 123, pp.298-313, 2019.

Tang, R.J. et al. Plant membrane transport research in the post-genomic era.**Plant Communications**, p.100013, 2019.

TUSNADY, G.E.; SIMON, I.The HMMTOP transmembrane topology prediction server.**Bioinformatics**, v. 17, n. 9, p.849-850, 2001.

UPADHYAY, N. et al. The multitasking abilities of MATE transporters in plants.**Journal of experimental botany**, v. 70, n.18, p.4643-4656, 2019.

USADEL, B. et al. Co-expression tools for plant biology: opportunities for hypothesis generation and caveats. **Plant, cell & environment**, v. 32, n.12, p.1633-1651, 2009.

VAN DER VOSSEN, H.; BERTRAND, B.; CHARRIER, A. Next generation variety development for sustainable production of arabica coffee (Coffeaarabica L.): a review. **Euphytica**, v. 204, n. 2, p.243-256, 2015.

WALDHAUSER, S.S.M.; BAUMANN, T.W.Compartmentation of caffeine and related purine alkaloids depends exclusively on the physical chemistry of their vacuolar complex formation with chlorogenic acids. **Phytochemistry,** v.42, n.4, p.985-996, 1996.

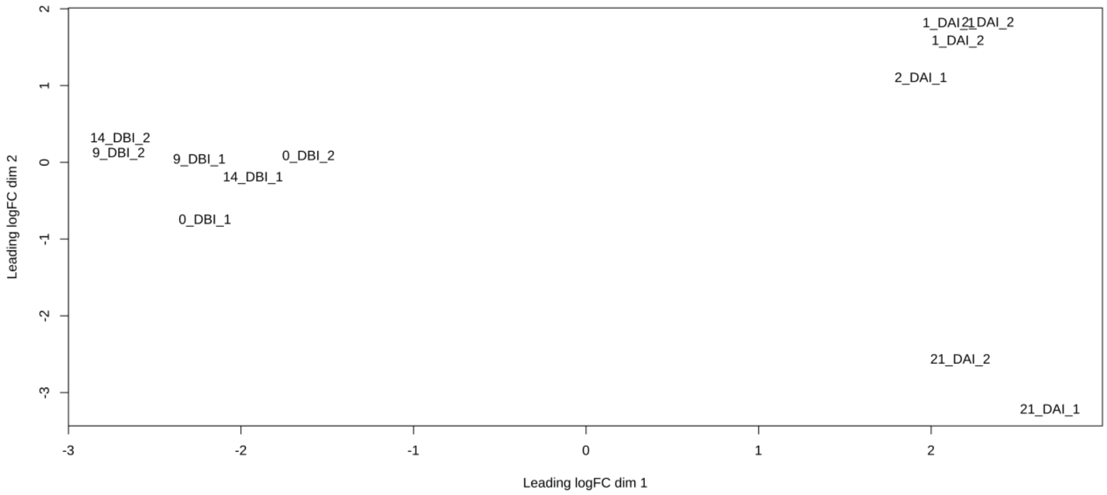
WATANABE, Y. et al. Visualization of cellulose synthases in Arabidopsis secondary cell walls. **Science,** v. 350, n.6257, p.198-203, 2015.

XU, J. et al. GhL1L1 affects cell fate specification by regulating Gh PIN 1-mediated auxin distribution. **Plant biotechnology journal**, v. 17, n.1, p.63-74, 2019.

YASHIN, A.; YASHIN, Y.; WANG, J.Y.; NEMZER, B. Antioxidant and antiradical activity of coffee.**Antioxidants**, v. 2, n.4, p.230-245, 2013.

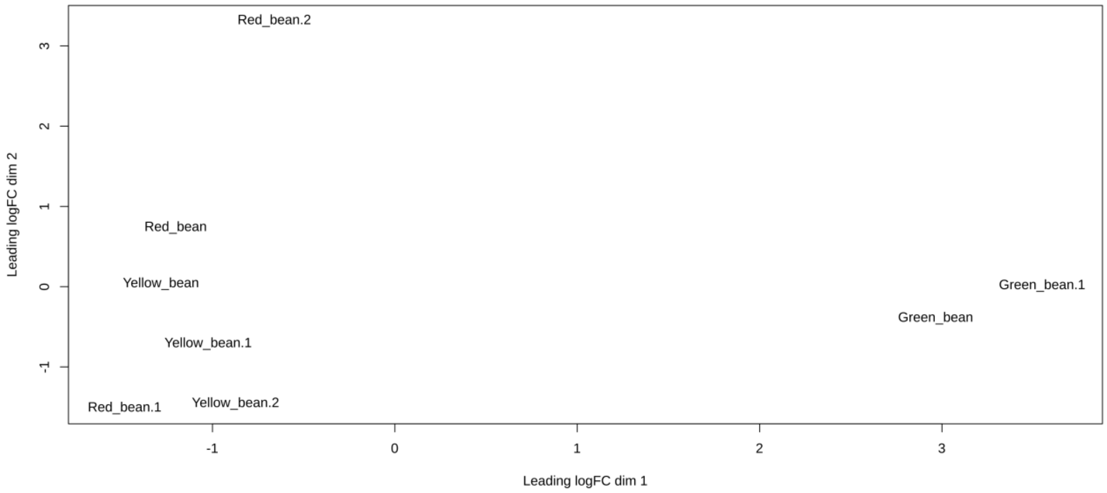
APPENDIX

Figure S1 - Biological coefficient variation using a somatic embryogenesis process RNA-seq experiment



Legend: DBI - days before induction of somatic embryogenesis; DAI - days after induction of somatic embryogenesis. Fonte: Do autor, 2020.

Figure S2 - Biological coefficient variation using coffee bean development RNA-seq experiment and gene expression profile of coffee membrane transporters.



Fonte: Do autor, 2020.

SUPPLEMENTARY MATERIAL

Supplementary table 1: digital format (available under request)

Supplemantary table 2: digital format (available under request)

Supplamentary table 3: digital format (available under request)

Supplamentary table 4: digital format (available under request)

Supplamentary table 5: digital format (available under request)

Supplamentary table 6: digital format (available under request)

SCIENTIFIC PAPER 3: A COMPREHENSIVE GENOMIC ANALYSIS TO EXPLORE MOLECULAR ASPECTS OF CAFFEINE ACCUMULATION ON COFFEE PLANTS

ABSTRACT

Coffee plant is one of the main crops which humans cultivate for making beverages. It is worldwide appreciated and consumed as part of a daily routine for billions of people and a major determinant for its consumption might be a particular substance present in this beverage, caffeine. This alkaloid affects animals' nervous system by interacting with adenosine receptors because of its structure similarity. Due to efforts on understanding the molecular basis of caffeine production on coffee plants, the three last N-methyltansferases (NMT) that participates on its biosynthesis and their respective genes are already known. However, other fundamental aspects of a metabolic pathway, like regulators, signaling receptors and transporters were never explored for caffeine metabolism. Targeting the importance of identifying transcriptional regulators of caffeine biosynthesis and caffeine transporters, because of their potential as *in planta* accumulation determinants, we performed a comprehensive screen of candidate genes using genomic and transcriptomic analyses. We have classified 1,171 genes as putative transcription factors (TF) and identified genes as possible members of ABC, MATE and PUP transporter families. Based on gene co-expression and structural analyses, we selected eight genes for further study of their gene expression profile on field derived coffee leaf and bean samples with quantified contrasting caffeine content. The gene expression profiles of NMT, TF and transporter coding genes were analyzed and we point to two genes as potential caffeine biosynthesis related TF and caffeine transporter. With this novel information, increase the perspectives of research focused on coffee genetic improvement.

Keywords: Alkaloid; ATP Binding Cassete; MATE; PUP; Transcription factor.

1 INTRODUCTION

Since the rise of agriculture and early civilizations, humans have domesticated plant species for food supply, but also for other purposes, like treating diseases and recreational uses. Throughout the history, some non-food plants became important crops for humans due to its use for producing beverages, like the coffee plant, which we use to produce a beverage linked to many relevant moments in modern human society and that is part of the daily routine of people around the globe (WITHINGTON, 2020).

Coffee and tea represent the most consumed beverages in the world (excluding water) and, interestingly, they share one metabolite that might be the main reason for its consumption, the alkaloid caffeine. The independent metabolic routes for the production of this alkaloid might have been originated from convergent evolutionary process for coffee, tea, cocoa and citrus, other caffeine producing and widely consumed crops (HUANG *et al.*, 2016).

For the coffee plant, the final steps of caffeine production pathway are three consecutive methylations, performed by three different enzymes (DENODEUD *et al.*, 2014; HUANG *et al.*, 2016). Many efforts resulted on the current knowledge about the genes that codify these enzymes on *Coffeaa arabica* and *Coffea canephora*, the major species that corresponds to 60 and 40% of coffee market, respectively. Also, we have available information about the dynamics of caffeine biosynthesis throughout coffee plant, data about expression profiles on different organs and information about possible alternative transcripts for these genes (ASHIHARA *et* al., 1996; PERROIS*et al.*, 2015; CHENG *et al.*, 2017).

Although this information is valuable and already led to scientific advances on manipulating caffeine content on coffee (ASHIHARA *et al.*, 2006; BORREL, 2012), surprisingly, little is known about any other aspect related to caffeine accumulation despite the last enzymatic steps on the biosynthesis, especially at genetic and molecular level. None transcriptional regulator, signaling receptor or membrane transporter was ever discussed, even though these are essential aspects to understand the dynamics of a metabolic pathway (GRUNEWALD *et al.*, 2020; HAYASHI *et al.*, 2020).

In this moment of increasingly available genomic and transcriptomic information, such aspects must be explored to enable data generation and curation that might benefit coffee genetic improvement programs. Moreover, the development of gene editing technologies and tissue culture protocols that allows efficient *Coffea spp.*genetic transformation, point to the possibility of precise genome manipulation for intended purposes (BREITLER *et al.*, 2018), like caffeine metabolism, which is highly determinant for both the plant-environment and product-consumer interactions.

To this end, we decided to explore the public genome of the diploid *C. canephora* to classify the transcription factor coding genes and membrane transporters related to alkaloid transport. We selected potential candidates related to caffeine metabolism using protein similarity, transcriptomic data and gene co-expression networks. Using field derived leaves and bean samples, we have tested some of these candidates regarding their expression profile concordance to caffeine content variation and caffeine-related NMT genes expression.

We found associations for two genes, one TF and one putative membrane transporter, with the caffeine variation on these tissues and the expression profile of caffeine-related NMT genes. These genes are interesting candidates for further analysis that might help to elucidate essential aspects of caffeine metabolism on coffee plants and, therefore, sustain the development of coffee genetic improvement programs focused both on this plant performance on the field and on the final beverage quality attributes.

2 RESULTS

Transcription factors and membrane transporters on *C. canephora* genome and their possible interactions with caffeine-related *NMT* genes

We performed an analysis using *C. canephora* proteome (DENODEUD *et al*., 2014) to identify all the putative transcription factors in this species. Using three different prediction softwares, PlantTFcat (DAI *et al*., 2013), iTAK (ZHENG *et al*., 2016) and PlantTFDB (JIN *et al*., 2017), we found 1,970, 1,299 and 1,256 predicted TFs, respectively (supplementary figure 1). The sequences on the intersection group within these predictions were classified as putative transcription factors. This inventory is composed by 1,171 genes, assigned to 51 different transcription factor families (supplementary table 1).

In order to prospect candidate TFs that might regulate caffeine synthesis related genes, promoter sequences (2 kb region upstream of ATG) from the genes *Cc09_g06970* (*CcXMT*), *Cc00_g24720* (*CcMXMT*) and *Cc01_g00720* (*CcDXMT*) were used as queries to identity putative binding sites related to the classified TF genes using PlantRegMap platform (TIAN *et al*., 2019). We found that 51 out of 1,171 TFs have putative binding sites on the *NMT* genes' promoters, and 91 possible interactions were identified (supplementary tables 2 and 3).

To deepen the comprehension of the caffeine accumulation process, we selected putative membrane transporters from families already associated with alkaloid transport (SHITAN *et al.*, 2014) and performed a blast analysis on *C. canephora* proteome to obtain potential candidates as caffeine transporters by protein similarity. Ten *C. canephora* sequences were selected based on this criterion (supplementary table 4), two from Multidrug and Toxic Compound Extrusion (MATE) family, two from ATP Binding Cassete (ABC) family and six from Purine Uptake Permease (PUP) family.

A RNA-seq experiment (PRJNA339585) with samples from beans at different developmental stages and leaves treated with methyl-jasmonate (MeJA) was analyzed and, using the TPM values to assess the Spearman's correlation of selected genes' expression profile, we found that 24 TFs are co-expressed with caffeine-related *NMT* genes and have binding sites on their promoters. Among them, 14 are positively correlated and 10, negatively (figure 1).

Among the caffeine transporter candidates, five were found co-expressed with at least one of the biosynthesis related genes. The whole gene co-expression network, therefore, is composed by 32 genes (three caffeine related *NMTs*, five putative transporters and 24 putative transcription factors) linked through 32 connections of 14 negative and 18 positive correlations ($\rho \geq |0.7|$) (figure 1, supplementary table 5).

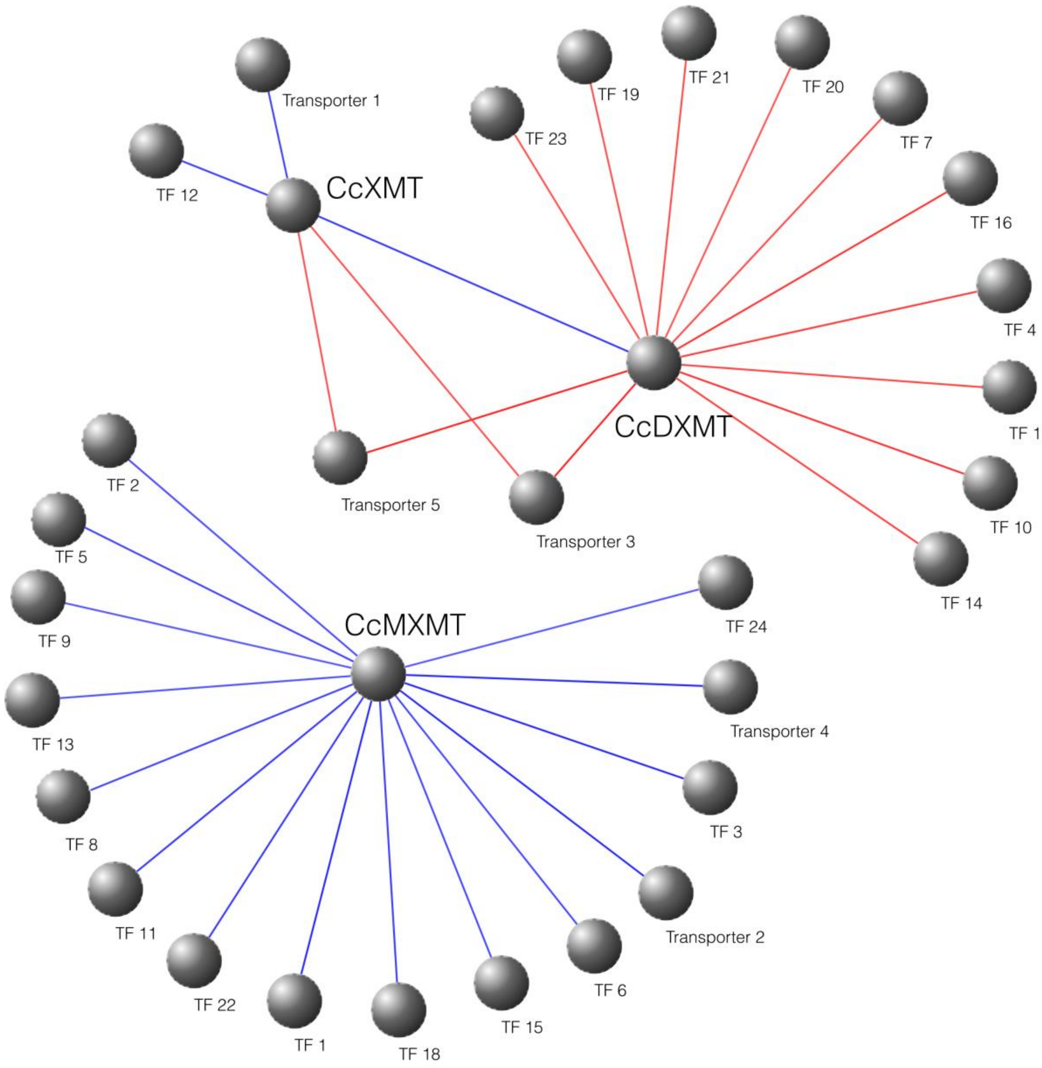
The effect of leaf age on caffeine accumulation and the expression profile of the transporters and TF genes

Eight genes from the group already analyzed here were chosen for RT-qPCR analysis on *C. canephora* tissues with potential contrast on caffeine content, four putative membrane transporter coding (*Cc01_g21810*, *Cc02_g02550*, *Cc09_g01780*, *Cc10_g06500*) and four putative TF coding ones (*Cc00_g02380*, *Cc00_g08780*, *Cc02_g30660*, *Cc08_g11060*).

On a field experiment, leaves from different coffee plants on varied positions on the plant and within the plantation were harvested. The harvest was conducted in one batch and the leaves were discriminated by their age ("Old", completely developed and "Young", on the initial development stage) (figure 2).

We found that the caffeine content was more than two-fold higher on young leaves in comparison with the old ones (figure 2). All the leaves were cut on the on the half and the same samples used for caffeine quantification were also utilized for RNA extraction to observe the expression profile of the enzyme coding and selected TF and transporter genes.

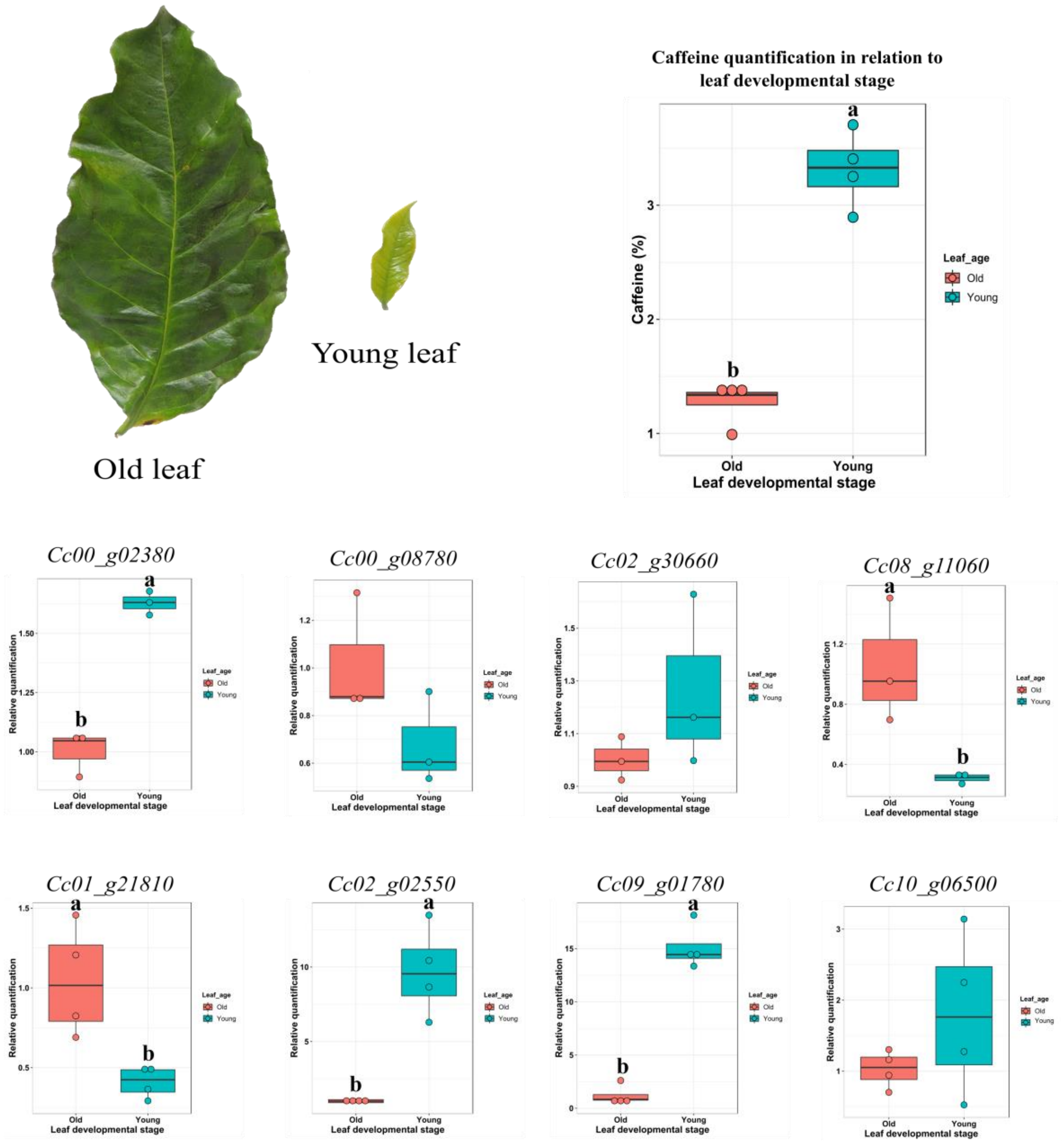
Figure 1 - Gene co-expression network with caffeine related *NMT* genes and putative membrane transporters and transcription factors selected genes.



Legend: Nodes represent each gene and edges are positive (blue) or negative (red) Spearman's correlations with ρ≥|0.7|. Gene codes for each respective node can be found on supplementary table 5. Fonte: Do autor, 2020.

There was no significant difference on expression of any biosynthesis enzyme coding gene on this two contrasting samples. One transcription factor coding gene (*Cc00_g002380*) and two membrane transporters (*Cc02_g02550* and *Cc09_g01780*) had an expression profile in concordance with caffeine content variation on coffee leaves. For three other genes (*Cc00_g008780, Cc02_g30660* and *Cc10_g06500*) there was no statistically significant difference on the expression values and the genes *Cc08_g11060* and *Cc01_g21810* (respectively TF and transporter coding) had an expression with an opposite pattern in relation to caffeine accumulation (figure 2).

Figure 2 - Caffeine content and gene expression profile (RT-qPCR) on age contrasting coffee leaves.



Legend: Transcription factor coding genes: *Cc00_g02380*, *Cc00_g08780*, *Cc02_g30660* and *Cc08_g11060*; Membrane transporter coding genes: *Cc01_g21810*, *Cc02_g02550*, *Cc09_g01780* and *Cc10_g06500*.Different letters indicate statistically significant differences using *p-value* ≤ 0.05 and Scott Knott test to compare mean values**.** Fonte: Do autor, 2020.

Caffeine content variation and expression profile of NMT, putative membrane tranporters and TFs during coffee bean development.

Fruits from *C. canephora*, *cv.* Conilon were harvested at 60, 90, 150, 210, 240 and 290 days after the flowering period (DAF) of the coffee plants. For each time point, a group of about 45 beans from nine different plants (divided into three biological replicates) was used for both caffeine and RNA extraction. We found that the caffeine concentration rises from 60 DAF until 150 DAF, which is the time point of highest caffeine content and, after that, the concentration of this alkaloid slightly decreases over the time analyzed (figure 3).

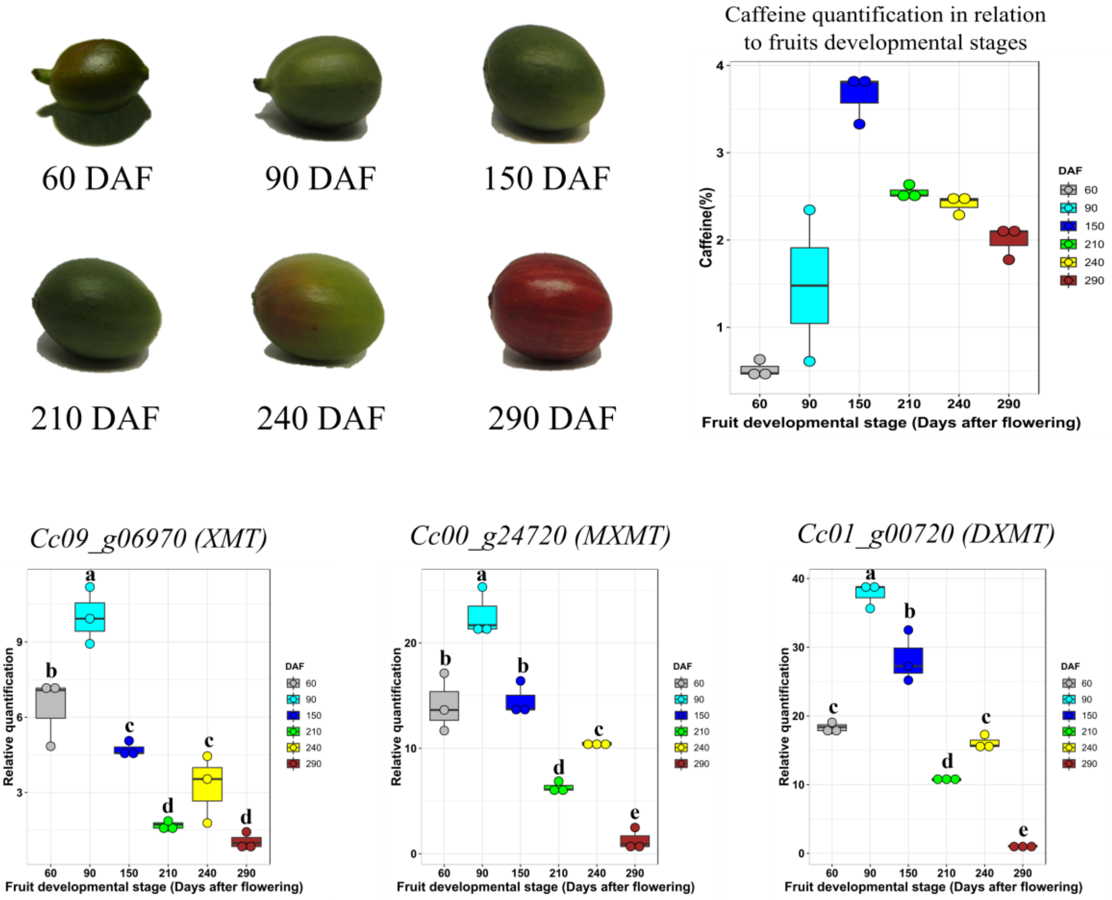


Figure 3: Caffeine content and NMT genes' expression profile (RT-qPCR) during coffee bean development. DAF - Days after flowering. Different letters indicate statistically significant differences using *p-value* ≤ 0.05 and Scott Knott test to compare mean values. Fonte: Do autor, 2020.

We observed a caffeine content variation-concordant expression profile for the *NMT* genes related to this metabolite biosynthesis (*CcXMT*, *CcMXMT*, *CcDXMT*; DENODEUD *et* *al*., 2014). All of them had an increase o gene expression untill the maximum value on 90 DAF, with a decreasing tendency after this time point (figure 3).

The same candidate TF and membrane transporter coding genes with expression profile assessed on leaves were analyzed on these samples of bean developmental stages. Among the putative membrane transporter coding genes, *Cc01_g21810* and *Cc10_g06500* had the highest expression value at 290 DAF; *Cc02_g02550* was downregulated on the intermediary bean development samples (150, 210 and 240 DAF) and upregulated on 60, 90 and 290 DAF and *Cc09_g01780* was upregulated untill the maximum point of caffeine content (150 DAF) with no drastic change in the following time points (figure 4).

For the candidate TF coding genes *Cc00_g02380*, *Cc00_g08780* and *Cc02_g30660*, we observed a similar expression profile, in which they are upregulated at the 60 and 90 DAF and on the following time points no significant change is found, execept for the switch from 240 to 290 DAF, when *Cc00_g002380* and *Cc02_g30660* are slightly down and upregulated, respectivelly. Regarding *Cc08_g11060*, a opposite tendency is observed, with upregulations from 60 to 90 DAF and from 210 to 240 DAF (figure 4).

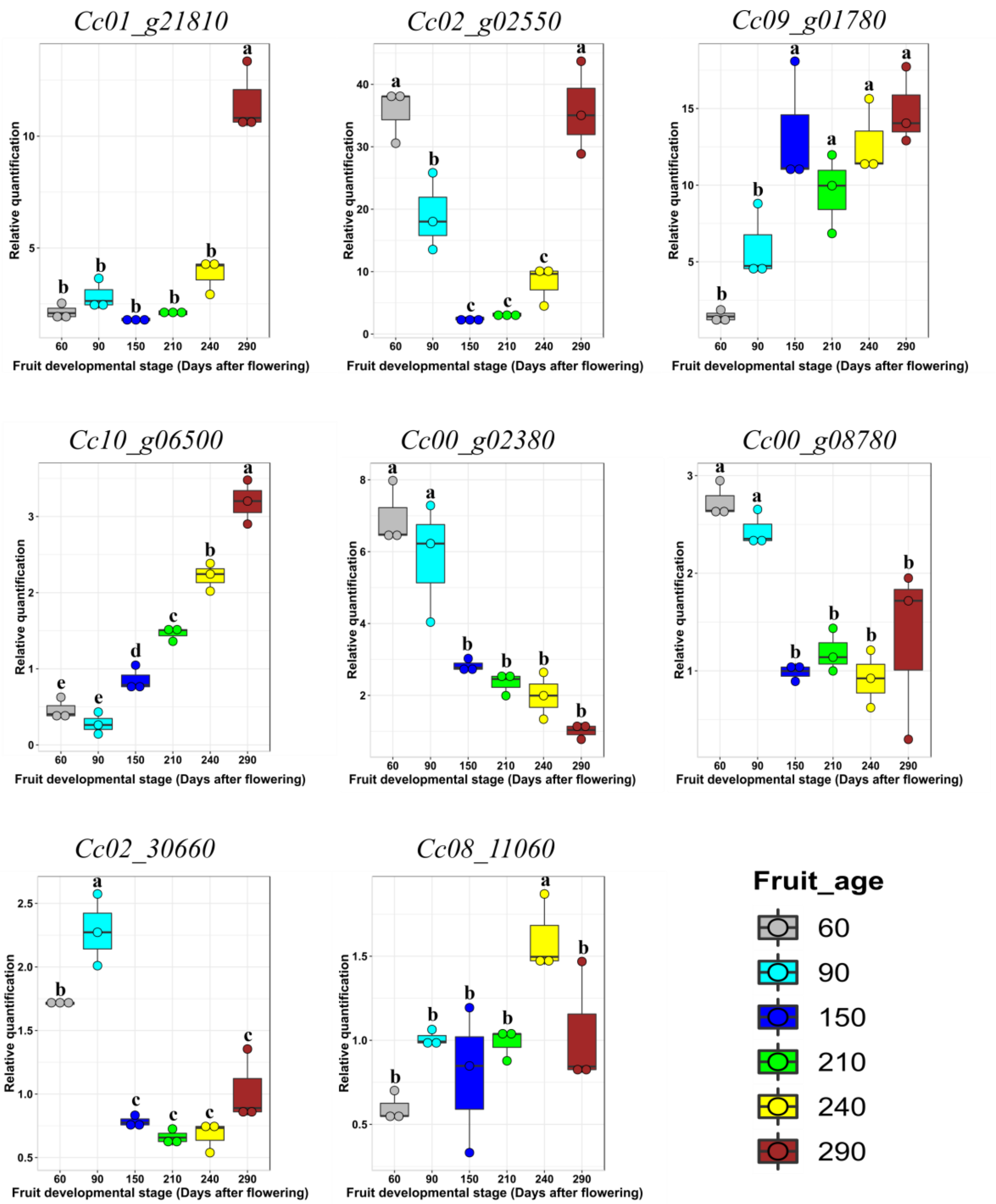
3 DISCUSSION

Comparing three different prediction softwares we classified 1,171 *C. canephora* genes (the intersection, supplementary figure 1) as potential transcription factors, number that corresponds to 4.6% of the coding sequences annotated in the genome (DENODEUD *et al*., 2014). This percentage is consistent with the mean Eudicots TF content (5.10%) found in 95 genome sequenced plants (JIN *et al*., 2017).

If the criteria of software results congruency is not considered, the union of all genes predicted as TFs by at least one prediction tool would represent 8.1%. This ratio is closer to what is observed for *Solanum lycopersicum* and *S. tuberosum* (about 8%), but, besides the close phylogenetic relationship between *C. canephora* and these two species, the formers undergone a recent whole genome triplication (WGT) event that was exclusive for *Solanaceae* (LEHTI-SHIU *et* al., 2017; QIAO*et al*., 2019).

A similar pattern is observed inside *Fabaceae* clade, in which *Medicago truncatula* has a lower percentage of TFs (about 5%) than *Glicine max* (8%), which, besides the common WGD event shared by all *Fabaceae* species, had an specific recent WGD. Considering that the ratio of TF genes number within a genome is correlated with events of polyploidization occurred during the evolution of a species , we can speculate that a TF number on the interval described here (4.6% - 8.1%), but closer to congruent result highlitghted (4.6%) is consistent with what is expected for *C. canephora* (LEHTI-SHIU *et* al., 2017; QIAO*et al.*, 2019).

Figure 4 - Expression profile (RT-qPCR) of putative membrane transporters and transcription factor coding genes during coffee bean development.



Legend: DAF - Days after flowering. Membrane transporter coding genes: *Cc01_g21810*, *Cc02_g02550*, *Cc09_g01780* and *Cc10_g06500*; TF coding genes: *Cc00_g02380*, *Cc00_g08780*, *Cc02_g30660* and *Cc08_g11060*. Different letters indicate statistically significant differences using *p-value* ≤ 0.05 and Scott Knott test to compare mean values. Fonte: Do autor, 2020.

We decided to combine structural information and transcriptional information for these genes classified as TFs in order to point candidates as caffeine biosynthesis regulators. From this dataset, 51 TFs have at least one binding site (TFBS) on the promoter region of at least one of the caffeine related NMT genes and, interestingly, 24 of these TFs are co-expressed with the *CcXMT*, *CcMXMT* and *CcDXMT* (figure 1) in the RNA-seq libraries analyzed. The concordance between the two types of information is a convenient start point for the identification of potential TFs regulating caffeine biosynthesis.

In the same way, from the members of ABC, MATE and PUP families with protein similarity to known alkaloid transporters (SHITAN *et al*., 2014), five were co-expressed with the caffeine related *NMT* genes. A gene co-expression network (figure 1) was constructed using only the TF and transporter candidates that had a potential on a structural perspective (TFBS presence or protein similarity) of participating on caffeine accumulation process. The gene co-expression of transporters and the enzymes that produce their substrates is already documented and proved to be an indicative of metabolic interaction (DOBRITZCH *et al*., 2016; PAYNE *et al*., 2017; DEMURTAS *et al*., 2019), as well as for TFs and their target genes (ZENG *et al*., 2018; ZHANG *et al*., 2020).

Caffeine is mostly found on beans (endosperm and perisperm), leaves and flowers of the coffee plant, in diverse concentrations upon different stimulli and physiological state of the organ/tissue (PERROIS *et al*., 2015; KUMAR *et al*., 2017). The role of caffeine on plant interaction with pathogens is documented (KIM *et al*., 2010; CEJA-NAVARO *et al*., 2015**)**, as well an interesting influence on pollinators attractiveness (WRIGHT *et al*., 2013), but these aspects are still poorly explored and little is known about the influence of environmental and physiological aspects on caffeine biosynthesis.

Due to the evidences that young leaves accumulates more caffeine than fully developed ones (ASHIHARA *et al*., 1996; PERROIS *et al*., 2015), we decided to use this organ to test whether some TFs and membrane transporters candidates would have a expression profile correlated to this alkaloid content variation. Out of the eight genes tested, only three were upregulated in the young leaves, where the caffeine content is higher (figure 2).

Two of these genes (*Cc09_g01780* and *Cc02_g02550*) belong to the MATE family and, conveniently, one member of this family in *Coptis japonica* that had a expression pattern correlated with an alkaloid (berberine) accumulation was confirmed as its transporter (TAKANASHI *et al*., 2017). The other gene (*Cc00_g02380*) putatively belongs to the bZIP TF family, which, to the best of our knowledge, no member has been associated to alkaloid biosynthesis activation, however, genes from this family can negatively regulate the alkaloid biosynthesis in *Catharanthus roseus* and *Camptotheca acuminata* (SIBÉRIL *et al*., 2001; CHANG *et* al., 2018).

Regarding human consumption, the most important caffeine accumulating tissue on coffee plant is the bean's endosperm, where this alkaloid is commonly stored on ratios around 1 to 2% (caffeine/bean dry mass) in the commonly comercialized coffee varieties of *C. arabica* and *C. canephora* species (BABOVA *et al*., 2016; BARBOSA*et al*., 2019). The three last reactions on the caffeine biosynthesis pathway take place during the initial to mid development of coffee beans, a period in which this organ is mostly composed by perisperm cells and the pericarp (DE CASTRO and MARRACCINI, 2006).

The accumulation of caffeine on coffee beans have its maximum peak on the moment of perisperm to endosperm conversion, which usually occurs around 120 days after flowering (DAF) for *C. canephora* (PERROIS *et* al., 2015; KUMAR*et al*., 2017). In concordance with this, the caffeine-related NMT genes are expected to be more expressed untill the bean reach this same developmental stage (PERROIS *et al*., 2015). Indeed, this was the observed pattern for *CcXMT*, *CcMXMT* and *CcDXMT* gene expression in our *C. canephora* field experiment harvested samples and also for caffeine accumulation on the beans (figure 3).

Due to this observed caffeine content variation and NMT gene expression profile, we decided to evaluate the expression of the same TF and membrane transporters candidates genes and we found diverse patterns (figure 4). Among the putative TF genes, *Cc00_g002380* had the highest positive correlation with *CcXMT* ($\rho = 0.87$), *CcMXMT* ($\rho = 0.84$) and *CcDXMT* ($\rho = 0.81$), and all the other correlation values for TFs and *NMTs* on the RT-qPCR analysis were under $\rho = 0.6$. Interestingly, this was the only TF gene that had a differential expression (with *p-value* ≤ 0.05) and was positively correlated with caffeine content variation on leaves experiment (figure 2).

Another observation worth to be highlitghted is that the gene *Cc08_g11060* has the potential to act as a repressor TF related to NMT genes, based on its expression profile, as for all the comparisons, this gene had a negative correlation both to NMT genes expression and caffeine content variation (figures 1, 2, 3 and 4).

The knowledge about transcription factor involved on caffeine biosynthesis regulation may help to design strategies for caffeine-free or pre-determined caffeine content varieties. A precise design would consider the change of caffeine content just on the endosperm, which is the used tissue on coffee preparation, without affecting any of the possible roles of caffeine on the other coffee tissue and organs, like leaves, flowers and pericarp.

To this end, a possible strategy could be focused on editing the promoter of NMT genes, on a TFBS, like the one pointed in this work as a possible cis-element necessary for the proteins coded by *Cc00_g02380* and *Cc08_g11060* binding. The editing of the TFBS would disrupt the TF-target interaction, therefore changing the spatiotemporal regulation of the *NMT* genes expression and this can be achieved by using available gene editing techniques (JINEK *et* al., 2012; ANZALONE *et al*., 2019; LIN*et al*., 2020) with already available genetic transformation protocols for the coffee plant (BREITLER *et al*., 2018).

Within the putative membrane transporter genes, an expression profile in concordance with caffeine accumulation and NMT genes regulation was observed for only one candidate, the MATE gene *Cc09_g01780*. This gene is upregulated from 60 DAF to 90 DAF, where the *NMT* genes expression hits the maximum value, and its expression profile reachs a plato at 150 DAF, the time point which we found the highest caffeine content on the beans and probably marks the perisperm to endosperm conversion.

Despite the central role of NMTs on determining the caffeine content, this alkaloid accumulation might depend on one or few transporter proteins, for cell to cell and intracellular traffic, as already demonstrated for most of the specialized metabolism-derived substances, including alkaloids (TANG *et al*., 2019). In coffee plant cells, caffeine seems to be stored on vacuoles (WALDHAUSER and BAUMANN, 1996) and the biosynthesis of this molecule occurs on the same organ where is accumulates (BAUMANN and WANNER, 1972) . These observations point to a scenario which a vacuolar transporter would be determinant for caffeine accumulation.

Modifications targeted to a putative caffeine vacuolar transporter, like the candidate pointed here (*Cc09_g01780*), could alter the amount of caffeine that is accumulated on a particular coffee cell, as the restriction on vacuolar compartimentalization of this metabolite could result on a cytoplasmatic concentration that might be toxic to the cell and pottentially can induce a decrease on its production by a feedback loop. It is already documented that genetic changes targeted to transporters can alter the metabolism of a specialized metabolite (PAYNE *et al*., 2017; DEMURTAS *et al*., 2019; BANASIAK *et al*., 2020).

This comprehensive analysis to identify genes that could affect caffeine accumlation in coffee plants is a important step towards the understanding of this metabolite biosynthesis in a wider perspective, although a deeper investigation for each candidate presented here is still needed. The possible outcomes from this study can foster the genetic improvement of coffee, regarding the influence of this alkaloid on coffee plant disease and pests resistance and coffee cup quality attributes.

4 METHODS

Transcription factors genomic inventory construction, identification of candidate membrane transporters for alkaloid transport and gene co-expression network

The proteome of *C. canephora* (DENODEUD *et al*., 2014) was used as a query for classifying which proteins have transcription factors common domains by the softwares PlantTFDB, PlantTFcat and iTAK. The data from the three softwares was compiled (supplementay table 1) and the sequences on the result intersection were classified as putative TFs. Using the webserver PlantRegMap, the promoter region (2 kb upstream of start codon) of *Cc09_g06950* (*CcXMT*), *Cc00_g24720* (*CcMXMT*) and *Cc01_g00720* (*CcDXMT*) were screened for TFBS respective to the 1,171 *C. canephora* sequences classified as TFs.

The same proteome was also utilized for a blastp analysis against members of ABC, MATE and PUP to find the most similar *C. canephora* proteins to known alkaloid transporters (SHITAN *et al*., 2014). The output *C. canephora* sequences were then aligned and a phylogenetic tree was constructed for each family, using the Neighbor joining method with *p-distance* and 1,000 booststrap repetitions, on the software MEGA. On this analysis, proteins for each of the transporter families already associated with alkaloid transport were incorporated into the pool of aligned sequences, for the identification of closely related *C. canephora* sequences.

A gene co-expression analyses beetwen 51 TFs (selected because of TFBS presence on *NMT* genes' promoter), ten putative membrane transporters and the three caffeine related *NMT* genes was performed using public RNA-seq (TPM values, supplementary table 6) data and the Spearman's correlation method, with a threshold of $\rho \geq |0.7|$. The RNA-seq data was analyzed using FastQC and the low quality reads were trimmed before the alignment/read count performed by using Kallisto (BRAY *et al*., 2016) with the *C. canephora* genome as reference, using the CyVerse webservers Discovery Environment and DNA subway (https://de.cyverse.org/de/, https://dnasubway.cyverse.org/).

Experimental design for harvesting leaves and fruits of field grown *C. canephora* plants

The plant samples (leaves and beans at diverse developmental stages) were collected from field-grown *C. canephora*, cv. Conilon at the city of Lavras, state of Minas Gerais, Brazil. For leaf harvesting, young and fully developed (Old) leaves were collected from different parts of the plants in relation to sun exposition. Each biological replicate was composed by a pool of about 18 leaves from three different plants (six per plant). The experiment was based on four biological replicates for each leaf developmental state. Each leaf, in alll of the samples, was cut on the half, in order to use the same sample pool for caffeine quantification and RNA extraction.

The beans harvesting was perfomed sequentially, at 60, 90, 150, 210, 240 and 290 days after the flowering period (DAF). Despite being collected on different days, the samples were collected on the same period of the day. For one biological replicate, six fruits were collected from three different plants (18 fruits per biological replicate) and for each treatment, we established three biological replicates. The pericarp was removed from the samples and the pool of macerated beans (perisperm and endosperm) from a biological replicate was divided into samples for RNA and caffeine extraction.

Analysis of caffeine content

Leaf and bean samples collected and selected as described above were dried at 65ºC, macerated and submited to a methanol extraction protocol (PERROIS *et al.*, 2015). The resulting extracts were analyzed on a HPLC (Shimadzu, Japan) equipped with a Ascentis Express C18 column (15cm x 2,1mm, 2,7µm), utilizing a solution of 83% of MiliQ water, 11% methanol and 6% of acetonitrile as mobile phase and a flow ratio of 2 mL.min^{-1}. The running and retention times were 7 and 3.5 minutes, respectively.

The caffeine detection was performed via an ultraviolet sensible spectrometer, on the wavelenght of 276 nm and the quantification was determined based on a analytical curve.

RNA extraction and gene expression analysis by RT-qPCR

All the samples collected were immediatelly frozen on the field on liquid nitrogen and stored on – 80 ºC. Both bean and leaf sample were macerated using liquid nitrogen , but they were submitted to different RNA extraction protocols. The RNA from leaves was extracted by using the Concert™ Plant RNA Reagent (ThermoFischer) following the procedures recommended by the manufacturer and for the beans, we used a modified CTAB-based protocol with the application of lithium chloride for RNA preciptation. All the RNAs were then purified using the Turbo DNA-free^{TM} kit (ThermoFischer), and the quality of the samples were assessed by usign spectrofotometric and electrophoresis analyses (supplementary table 7 and supplementary figure 2 and 3).

We confirmed the absence of contaminat DNA by performing a PCR with all the samples (individually) using a primer pair for a ubiquitiously expressed gene (*24S*, FREITAS*et al*., 2017), only the non-amplified samples were considered pure. The RNA samples were then converted to cDNA using the High-Capacity cDNA Reverse Transcription Kit (ThermoFischer). The qPCR was performed using SYBR Green Master Mix following the manufacturer's recommendations. The reactions efficiency were calculated using LinReg software (https://www.gene-quantification.de/download.html#linregpcr, RUIJTER *et al*., 2009) and the relative quantification values were obtained by using Pfaffl method (PFAFFL, 2001), using two endogenous genes to normalize the data (*24S* and *PP2A*, FREITAS*et al*., 2017). All the primers (supplementary table 8) were designed following optimal parameters for RT-qPCR technique. All the data was submitted to statistical analysis usign Scott Knott's test (SCOTT and KNOTT, 1974) performed on the R software (R CORE TEAM, 2013) to compare the values, considering different only the variations observed with *p-value* ≤ 0.05.

REFERENCES

ASHIHARA, H.; MONTEIRO, A.M.; GILLIES, F. M.; CROZIER, A. Biosynthesis of caffeine in leaves of coffee. **Plant Physiology**, v. 111, n. 3, p.747-75, 1996.

ASHIHARA, H. et al. Caffeine biosynthesis and adenine metabolism in transgenic Coffea canephora plants with reduced expression of N-methyltransferase genes. **Phytochemistry**, v. 67, n. 9, p.882-886, 2006.

BABOVA, O.; OCCHIPINTI, A.; MAFFEI, M.E. Chemical partitioning and antioxidant capacity of green coffee (Coffea arabica and Coffea canephora) of different geographical origin. **Phytochemistry**, v. 123, p.33-39, 2016.

BANASIAK, J. et al. The full-size ABCG transporter of Medicago truncatula is involved in strigolactone secretion, affecting arbuscular mycorrhiza.**Frontiers in Plant Science**, v. 11, 2020.

BARBOSA, M.D.S.G.; DOS SANTOS SCHOLZ, M.B.; KITZBERGER, C.S.G.; DE TOLEDO BENASSI, M. Correlation between the composition of green Arabica coffee beans and the sensory quality of coffee brews. **Food chemistry**, v. 292, p.275-280, 2019.

BAUMANN, T.W.; WANNER, H. Studies on the transport of caffeine in the coffee plant (Coffea arabica). **Planta**, v. 108, n. 1, p.11-19, 1972.

BORRELL, B. Make it a decaf: the enduring quest for a coffee bean without the buzz. **Nature,** v. 483, n. 7389, p.264-267, 2012.

BRAY, N.L., PIMENTEL, H., MELSTED, P. AND PACHTER, L. Near-optimal probabilistic RNA-seq quantification. **Nature biotechnology**, v. 34, n. 5, p.525-527, 2016.

BREITLER, J.C. et al. CRISPR/Cas9-mediated efficient targeted mutagenesis has the potential to accelerate the domestication of Coffea canephora**. Plant Cell**, Tissue and Organ Culture (PCTOC), v. 134, n. 3, p.383-394, 2018.

CEJA-NAVARRO, J.A.et a. Gut microbiota mediate caffeine detoxification in the primary insect pest of coffee. **Nature communications**, v. 6, p.7618, 2015.

CHANG, C. et al. A bZIP transcription factor, CaLMF, mediated light-regulated camptothecin biosynthesis in Camptotheca acuminata. **Tree physiology**, v. 39, n. 3, p.372-380, 2019.

CHENG, B.; FURTADO, A; HENRY, R. J. Long-read sequencing of the coffee bean transcriptome reveals the diversity of full-length transcripts. **Gigascience**, v. 6, n. 11, p. gix086, 2017.

DAI, X.; SINHAROY, S.; UDVARDI, M.; ZHAO, P. X. PlantTFcat: an online plant transcription factor and transcriptional regulator categorization and analysis tool. **BMC bioinformatics,** v. 14, n.1, p.32, 2013.

DE CASTRO, R.D.; MARRACCINI, P. Cytology, biochemistry and molecular changes during coffee fruit development. **Brazilian Journal of Plant Physiology**, v. 18, n. 1, p.175-199, 2006.

DEMURTAS, O.C. et al. ABCC transporters mediate the vacuolar accumulation of crocins in saffron stigmas. **The Plant Cell**, v. 31, n. 11, p.2789-2804, 2019.

DENOEUD, F. et al. The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. **Science**, v. 345, n.6201, p.1181-1184, 2014.

DOBRITZSCH, M. et al. MATE transporter-dependent export of hydroxycinnamic acid amides**. The Plant Cell**, v. 28, n. 2, p.583-596, 2016.

FREITAS, N.C. et al. Validation of reference genes for qPCR analysis of Coffea arabica L. somatic embryogenesis-related tissues. **Plant Cell Tissue and Organ Culture**,v. 128, p. 663–678, 2017.

GRUNEWALD, S. et al. The Tapetal Major Facilitator NPF2. 8 is Required for Accumulation of Flavonol Glycosides on the Pollen Surface in Arabidopsis thaliana. **The Plant Cell**, v. 32, n. 5, p. 1727-1748, 2020.

HAYASHI, S. et al. Genetic manipulation of transcriptional regulators alters nicotine biosynthesis in tobacco. **Plant and Cell Physiology**, v. 61, n.7, p. 1041-53, 2020.

HUANG, R.; O'DONNELL, A. J.; BARBOLINE, J. J.; BARKMAN, T. J. Convergent evolution of caffeine in plants by co-option of exapted ancestral enzymes. **Proceedings of the National Academy of Sciences**, v. 113, n. 38, p.10613-10618, 2016.

JIN, J. et al. PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. **Nucleic acids research**, p.gkw982, 2016.

KIM, Y. S.; CHOI, Y. E.; SANO, H. Plant vaccination: stimulation of defense system by caffeine production in planta. **Plant signaling & behavior**, v. 5, n. 5, p.489-493, 2010.

KUMAR, A.; NAIK, G. K.; GIRIDHAR, P. Dataset on exogenous application of salicylic acid and methyljasmonate and the accumulation of caffeine in young leaf tissues and catabolically inactive endosperms. **Data in brief**, v. 13, p.22-27, 2017.

LEHTI-SHIU, M.D. et al. Diversity, expansion, and evolutionary novelty of plant DNA-binding transcription factor families. **Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms**, v. 1860, n. 1, p.3-20, 2017.

PAYNE, R.M. et al. An NPF transporter exports a central monoterpene indole alkaloid intermediate from the vacuole. **Nature plants**, v. 3, n. 2, p.1-9, 2017.

PERROIS, C. et al. Differential regulation of caffeine metabolism in Coffeaarabica (Arabica) and Coffea canephora (Robusta). **Planta**, v. 241, n. 1, p.179-191, 2015.

PFAFFL, M.W. A new mathematical model for relative quantification in real-time RT–PCR. **Nucleic acids research**, v. 29, n. 9, p.e45-e45, 2001.

QIAO, X. et al. Gene duplication and evolution in recurring polyploidization–diploidization cycles in plants. **Genome biology**, v. 20, n. 1, p.38, 2019.

RUIJTER, J.M. et al. Amplification efficiency: linking baseline and bias in the analysis of quantitative PCR data. **Nucleic acids research**, v. 37, n. 6, p.e45-e45, 2009.

SCOTT, A.J. AND KNOTT, M. A cluster analysis method for grouping means in the analysis of variance. **Biometrics**, p.507-512, 1974

SHITAN, N.;KATO, K.; SHOJI, T. Alkaloid transporters in plants. **Plant Biotechnology**, pp.14-1002, 2014.

SIBÉRIL, Y. et al. Catharanthus roseus G-box binding factors 1 and 2 act as repressors of strictosidine synthase gene expression in cell cultures. **Plant molecular biology**, v. 45, n. 4, p.477-488, 2001.

TAKANASHI, K. et al. A multidrug and toxic compound extrusion transporter mediates berberine accumulation into vacuoles in Coptis japonica. **Phytochemistry,** v.1, n. 138, p. 76-82, 2017.

TANG, R.J.et al. Plant membrane transport research in the post-genomic era.**Plant Communications**, v.1, n. 1, p.100013, 2019.

TEAM, R.C. R: A language and environment for statistical computing, 2013.

TIAN, F. et al. PlantRegMap: charting functional regulatory maps in plants. **Nucleic acids research**, v. 48, n. D1, p.D1104-D1113, 2020.

WALDHAUSER, S. S. M.; BAUMANN, T.W. Compartmentation of caffeine and related purine alkaloids depends exclusively on the physical chemistry of their vacuolar complex formation with chlorogenic acids. P**hytochemistr**y,v. 42, n. 4 , p.985-996, 1996.

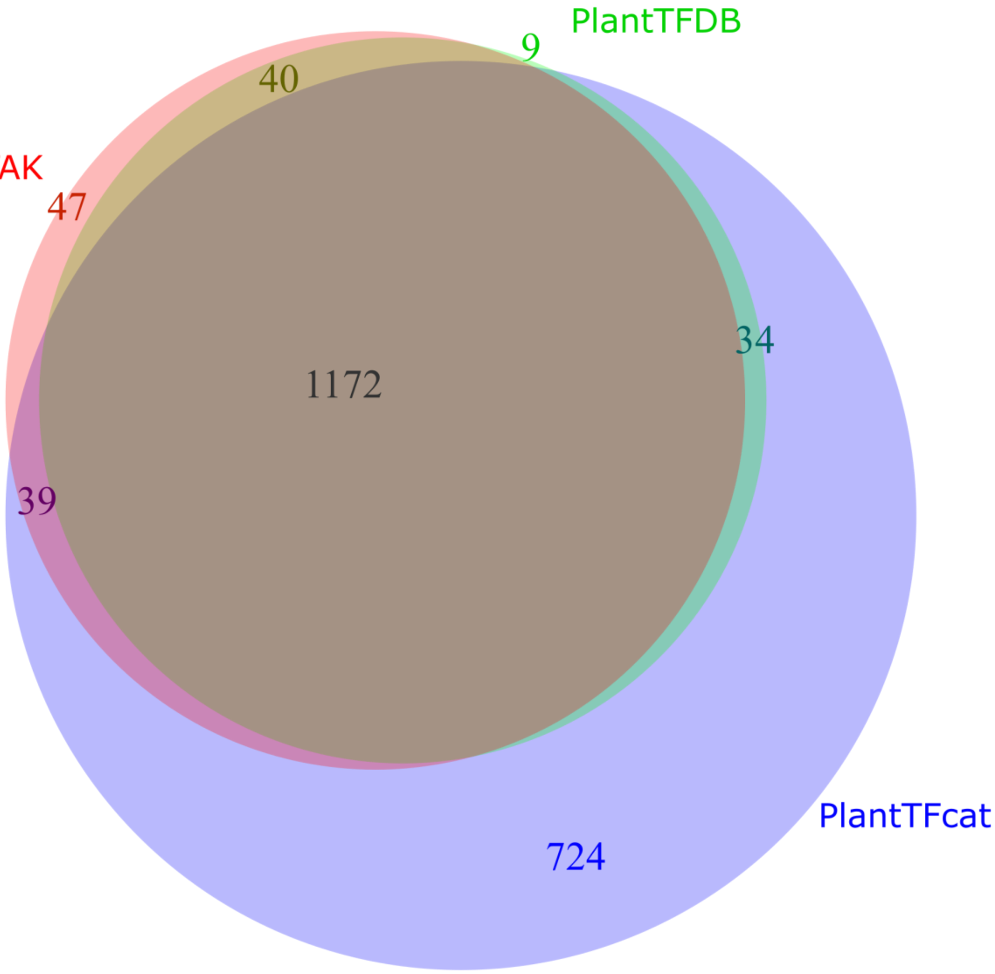
ZENG, L. et al. Identification of a G2-like transcription factor, OsPHL3, functions as a negative regulator of flowering in rice by co-expression and reverse genetic analysis. **BMC plant biology**, v. 18, n. 1, p.1-12, 2018.

ZHANG, A. et al. Transcriptome co-expression network analysis identifies key genes and regulators of ripening kiwifruit ester biosynthesis. **BMC Plant Biology**, v. 20, n. 1, p.1-12, 2020.

ZHENG, Y. et al. iTAK: a program for genome-wide prediction and classification of plant transcription factors, transcriptional regulators, and protein kinases. **Molecular plant**, v. 9, n. 12, p.1667-1670, 2016.

APPENDIX

Supplementary figure 1 - Comparison between the outputs from transcription factor's prediction softwares used on this work.



Fonte: Do autor, 2020.

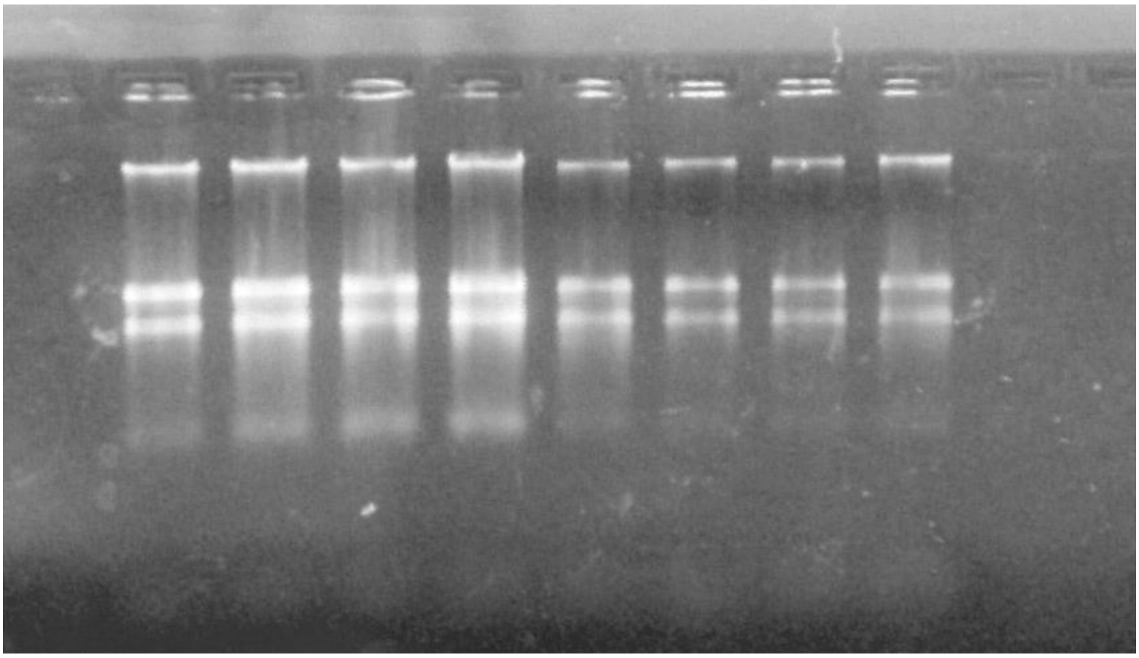
Supplementary table 7 - Description of quality parameters verified by spectrophotometry (NanoVue®) of purified RNA samples.

Sample	RNA concentration (ng/µL)	260 nm/280 nm ratio	260 nm/230 nm ratio
Young leaf(1)	146.8	1.76	2.62
Young leaf(2)	163.6	1.78	2.40
Young leaf(3)	152	1.75	2.06
Young leaf(4)	201.2	1.80	2.38
Old leaf (1)	145.2	1.76	2.28
Old leaf (2)	169.6	1.82	2.49
Old leaf (3)	153.2	1.79	2.35
Old leaf (4)	145.6	1.73	2.36

60 DAF(1)	97.07	2.04	1.94
60 DAF(2)	98.30	1.97	1.90
60 DAF(3)	101.41	1.97	1.89
90 DAF(1)	135.07	1.99	1.96
90 DAF(2)	124.26	1.96	1.95
90 DAF(3)	124.04	1.96	1.91
150 DAF(1)	103.87	1.91	1.81
150 DAF(2)	92.83	1.87	1.65
150 DAF(3)	101.53	1.97	1.82
210 DAF(1)	87.22	1.87	1.48
210 DAF(2)	203.99	1.94	1.83
210 DAF(3)	100.67	1.93	1.55
240 DAF(1)	84.45	1.87	1.68
240 DAF(2)	79.19	1.76	1.30
240 DAF(3)	128.25	1.89	1.64
290 DAF(1)	54.52	1.73	1.07
290 DAF(2)	100.08	1.97	1.84
290 DAF(3)	114.55	1.94	1.84

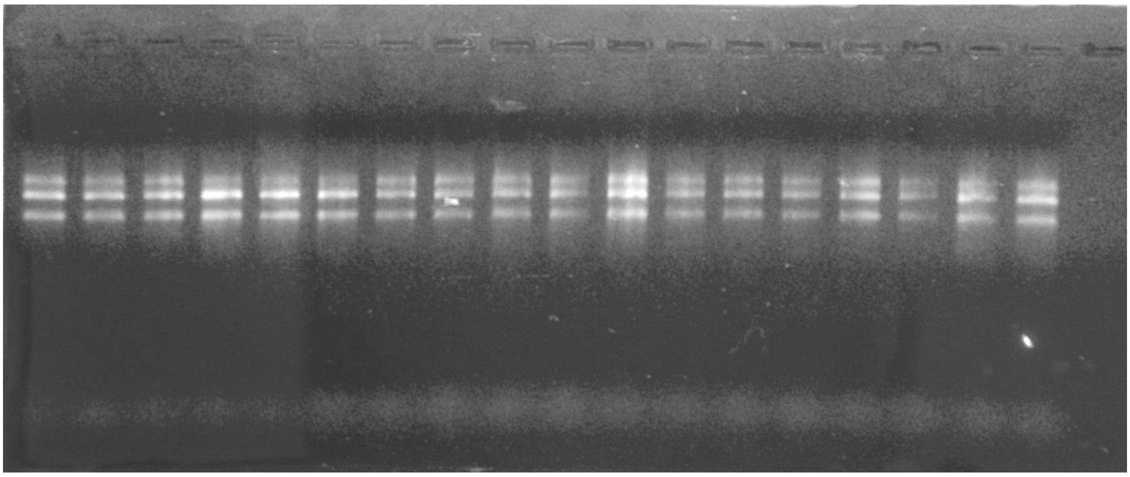
Fonte: Do autor, 2020.

Supplementary figure 2 - Electrophoresis gel visualization of purified RNA samples.



Legend: From left to right: Old Leaf 1 (OL1), OL2, OL3, OL4, Young Leaf 1 (YL1), YL2, YL3, YL4. Fonte: Do autor, 2020.

Supplementary figure 3 - Electrophoresis gel visualization of purified RNA samples



Legend: From left to right: 60DAF(1), 60DAF(2), 60DAF(3), 90DAF(1), 90DAF(2), 90DAF(3), 150DAF(1), 150DAF(2), 150DAF(3), 210DAF(1), 210DAF(2), 210DAF(3), 240DAF(1), 240DAF(2), 240DAF(3), 290DAF(1), 290DAF(2), 290DAF(3). Fonte: Do autor, 2020.

Supplementary table 8 - Description of the primers designed for RT-qPCR in this work.

Primer	**Sequence**
Cc00_g02380-F	5’ ATCCATGAGCCCAAGTCAAG 3’
Cc00_g02380-R	5’ TCCCCTAAGTGGTGCTGAAC 3’
Cc00_g08780-F	5’ GCAGCAAAAGTGGGAAGATT 3’
Cc00_g08780-R	5’ AAGTTTCTCGCATCCGCTTA 3’
Cc00_g24720-F	AACGACTTGATTGTTGAGG
Cc00_g24720-R	TGGGCCTTAAAAGTCTCC
Cc01_g00720-F	CAGCGCATGTGGCATCTG
Cc01_g00720-R	TCTTCGCAATCCTGTGGGAT
Cc01_g21810-F	5' CGGGTGATGAAGAGAGTGGT 3'
Cc01_g21810-R	5' CATTTGTTGCAGCAGCAGTAG 3'
Cc02_g02550-F	5' GCGGAAGCTATTACGATGGA 3'
Cc02_g02550-R	5' CGGAAAACGTACCACCATTC 3'
Cc02_g30660-F	5’ CTCATCGTCCGCTAACAACA 3’
Cc02_g30660-R	5’ GAGATGGCGAATGATTGGTT 3’
Cc08_g11060-F	5’ CAAAGCCCTCAACAAGAGC 3’
Cc08_g11060-R	5’ CCTCTCGTTCCCACCAATA 3’
Cc09_g01780-F	5' TGGAGGAACAATGATGCAGA 3'
Cc09_g01780-R	5' GCAGACGCTCTCTTGCTTTC 3'
Cc09_g06970-F	TATGTTGCATCTTCCGTTAGAG
Cc09_g06970-R	ACCTGTGGAATATGTCAGGT
Cc10_g06500-F	5' CAAGGAGGAGATAGAGGGTCAG 3'
Cc10_g06500-R	5' GGGGGATTAGTATCTCCGTCA 3'

Fonte: Do autor, 2020.

SUPPLEMENTARY MATERIAL

Supplementary table 1: digital format (available under request)

Supplemantary table 2: digital format (available under request)

Supplamentary table 3: digital format (available under request)

Supplamentary table 4: digital format (available under request)

Supplamentary table 5: digital format (available under request)

Supplamentary table 6: digital format (available under request)

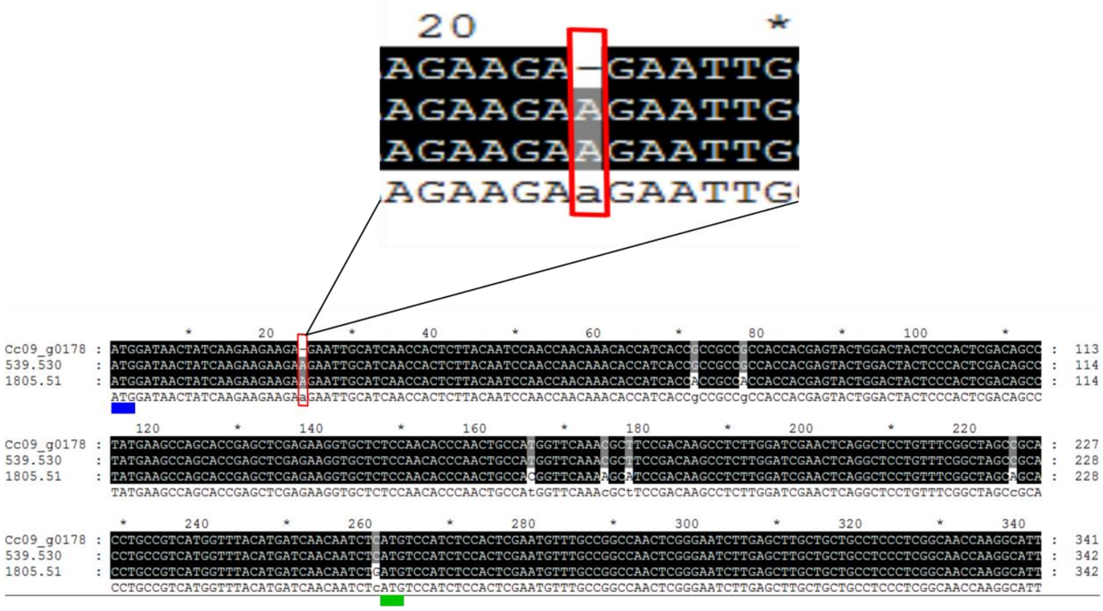
PART III – WORK ON PROGRESS AND FUTURE PROSPECTS

WORK ON PROGESS AND FUTURE PROSPECTS

In order to analyze the function of a putative membrane transporter coding gene already pointed here as a probable caffeine transporter (*Cc09_g01780*), we developed a vector for CRISPR-Cas9-based mutation on this gene. Firstly we took advantage from Coffee Genome Hub database (http://coffee-genome.org/, DENOEUD *et al*., 2014) to retrieve the nucleotide sequence of this gene and we made a comparison with the sequences deposited on Phytozome related to a yet unpublished *C. arabica* genome (https://phytozome-next.jgi.doe.gov/info/Carabica_v0_5), to rationally design sgRNAs which would have more chances to be compatible with both species, although our goal is to mutate *C. canephora.*

Using a blast search, two sequences with high similarity with *Cc09_g01780* were obtained, evm.TU.Scaffold_539.530 and evm.TU.Scaffold_1805.51, with 98.05 % and 91.42 % of identity, respectively. Interestingly, we found an important difference between the *Cc09_g01780* and the Phytozome obtained sequences, the last ones had an 261 nucleotide longer ORF, which would result on more 87 aminoacids on the translated protein. This alternative start codon is probably not annotated on *Cc09g01780* due to a missing adenine which prevents a methionine to appear on the same continuous frame (figure 1).

Figure 1 - Alignment of initial parts of *Cc09_g01780* gene with the two versions found for *C. arabica* on Phytozome


20 *
AGAAGA-GAATTG
AGAAGAAGAATTG
AGAAGAAGAATTG
AGAAGAaGAATTG
* 20 * 40 * 60 * 80 * 100 *
Cc09_g0178 : ATGGATAACTATCAAGAAGAAGA-GAATTGCATCAACCACTCTTACAATCCAACCAACAAACACCATCACCGCCGCCGCCACCACGAGTACTGGACTACTCCCACTCGACAGCC : 113
539.530 : ATGGATAACTATCAAGAAGAAGAAGAATTGCATCAACCACTCTTACAATCCAACCAACAAACACCATCACCGCCGCCGCCACCACGAGTACTGGACTACTCCCACTCGACAGCC : 114
1805.51 : ATGGATAACTATCAAGAAGAAGAAGAATTGCATCAACCACTCTTACAATCCAACCAACAAACACCATCACCACCGCCACCACCACGAGTACTGGACTACTCCCACTCGACAGCC : 114
ATGGATAACTATCAAGAAGAAGAaGAATTGCATCAACCACTCTTACAATCCAACCAACAAACACCATCACCgCCGCCgCCACCACGAGTACTGGACTACTCCCACTCGACAGCC
120 * 140 * 160 * 180 * 200 * 220
Cc09_g0178 : TATGAAGCCAGCACCGAGCTCGAGAAGGTGCTCTCCAACACCCAACTGCCATGGTTCAAACGCTTCCGACAAGCCTCTTGGATCGAACTCAGGCTCCTGTTTCGGCTAGCCGCA : 227
539.530 : TATGAAGCCAGCACCGAGCTCGAGAAGGTGCTCTCCAACACCCAACTGCCATGGTTCAAACGCTTCCGACAAGCCTCTTGGATCGAACTCAGGCTCCTGTTTCGGCTAGCCGCA : 228
1805.51 : TATGAAGCCAGCACCGAGCTCGAGAAGGTGCTCTCCAACACCCAACTGCCACGGTTCAAAAGCATCCGACAAGCCTCTTGGATCGAACTCAGGCTCCTGTTTCGGCTAGCAGCA : 228
TATGAAGCCAGCACCGAGCTCGAGAAGGTGCTCTCCAACACCCAACTGCCAtGGTTCAAAcGCtTCCGACAAGCCTCTTGGATCGAACTCAGGCTCCTGTTTCGGCTAGCcGCA
* 240 * 260 * 280 * 300 * 320 * 340
Cc09_g0178 : CCTGCCGTCATGGTTTACATGATCAACAATCTCATGTCCATCTCCACTCGAATGTTTGCCGGCCAACTCGGGAATCTTGAGCTTGCTGCTGCCTCCCTCGGCAACCAAGGCATT : 341
539.530 : CCTGCCGTCATGGTTTACATGATCAACAATCTCATGTCCATCTCCACTCGAATGTTTGCCGGCCAACTCGGGAATCTTGAGCTTGCTGCTGCCTCCCTCGGCAACCAAGGCATT : 342
1805.51 : CCTGCCGTCATGGTTTACATGATCAACAATCTGATGTCCATCTCCACTCGAATGTTTGCCGGCCAACTCGGGAATCTTGAGCTTGCTGCTGCCTCCCTCGGCAACCAAGGCATT : 342
CCTGCCGTCATGGTTTACATGATCAACAATCTcATGTCCATCTCCACTCGAATGTTTGCCGGCCAACTCGGGAATCTTGAGCTTGCTGCTGCCTCCCTCGGCAACCAAGGCATT

Legend: Alignment of initial parts of *Cc09_g01780* gene with the two versions found for *C. arabica* on Phytozome, indicating the start codon annotated on Phytozome (blue), the one for Coffee Genome Hub (green) and the polymorphism found to be responsible for this difference (a missing adenine on *Cc09_g01780*, highlighted by a red rectangle). Fonte: Do autor, 2020.

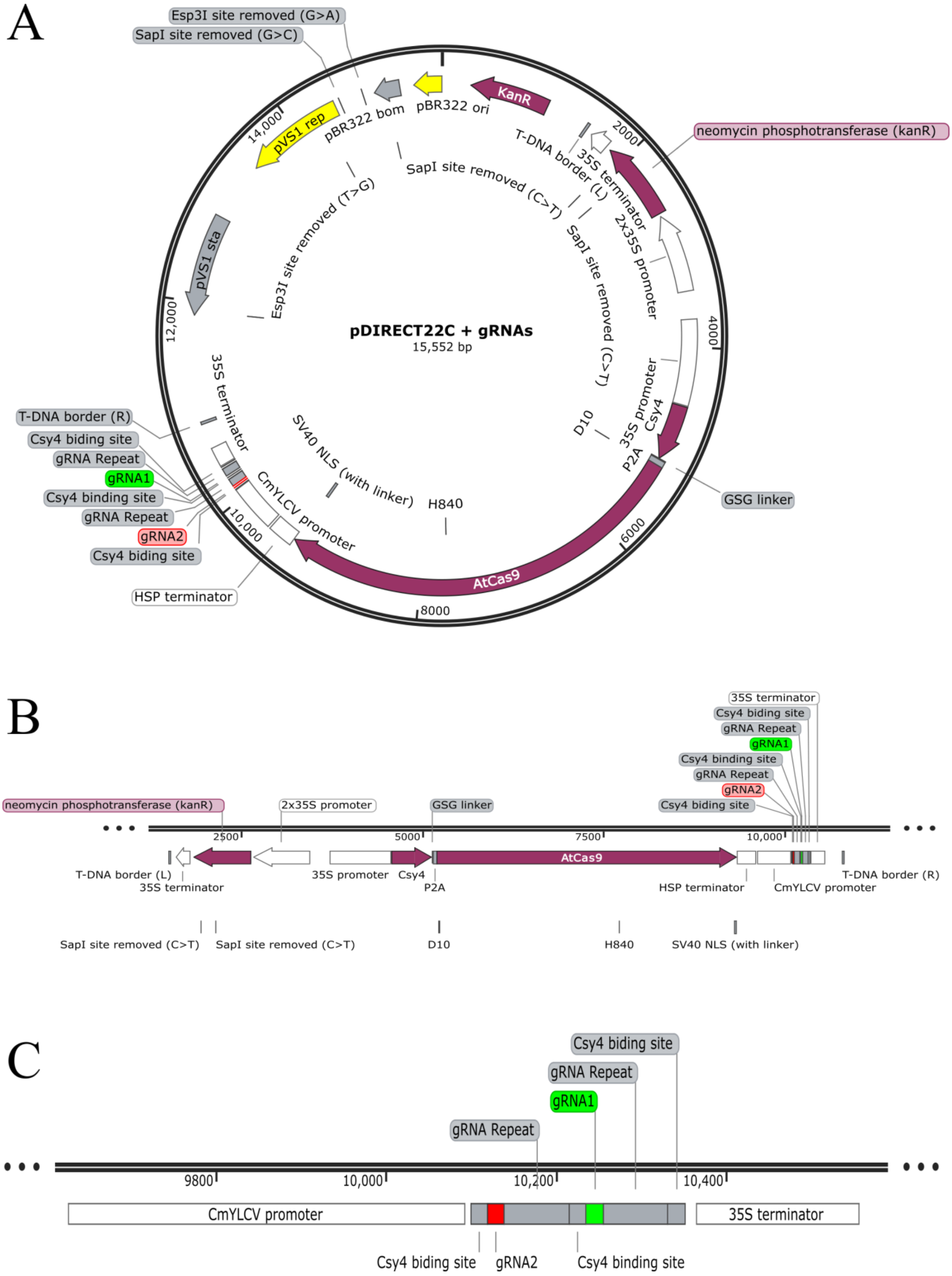
Interestingly, the smaller version of the putative protein (coded by *Cc09_g01780*) is predicted as having 11 transmembrane domains, while the longer version has 12, as analyzed using HMMTOP software (http://www.enzim.hu/hmmtop/). This last configuration seems to be more similar to the majority of plant MATE family proteins (reviewed on UPADHYAY*et al*., 2020), what points to the need of further investigations of which would be the correct form of annotation of this gene or, less probable, if this polymorphism happens between the two species resulting on such different proteins.

As a future goal, we propose that the 1,847 transporters identified and mentioned previously could be submitted to a stepwise strategy for analyzing their gene annotation. On a brief explanation, a score of compatibility to expected patterns of the transporter family could be attributed to each member, as already performed before (BENEDITO *et al*., 2010) and the information of each family would be used to evaluate the annotation of the least compatible members. To help on suggesting possible corrections to gene models, a long-read-based sequencing dataset could be explored by aligning the reads to the current annotation and verifying possible misspredictions. We argue that this effort would greatly benefit the advancement on *C. canephora* genome annotation.

We then combined information of the two annotations to design two sgRNAs to generate CRISPR-Cas9-mediated mutagenesis, with the help of free web-based designing tools (CHOP-CHOP (LABUN *et al*., 2019) and CasOFFinder (BAO *et al*., 2014). We decided to use a configuration of PAM-out, where the PAM regions of the targets are on opposite directions, expecting to generate a large deletion, using a previously described strategy for vector design (CEMARK *et al*., 2017). The aimed deletion is of 1087 bp of the gene, affecting the two first exons and deleting around 258 aminoacids, which accounts for about 59% of the protein. The vector (figure 2) is already assembled, sequenced and stored on -80 ºC on *Escherichia coli*.

This vector will be utilized for *A. tumefaciens*-mediated genetic transformation of *C. canephora* embryogenic calli and the regenerated plants will be analyzed regarding the effects of *Cc09_g01780* knockout. These analyses may help on elucidating whether this transporter is determinant for caffeine accumulation on coffee and, consequently, if its manipulation results in caffeine content variation on beans, a major biochemical characteristic associated with coffee consumption. We expect that the data presented here, along with this work on progress, will contribute to breeding programs, helping to sustainably maintain the coffee production chain.

Figure 2 - General illustration of the vector produced using Golden Gate assembly method, following the protocol provided by CEMARK and co-authors (2014).



Legend: A- view of the complete vector with main details; B- view of the T-DNA which will be inserted on the plant’s genome; C- closer view of the cassete comprising the promoter CmYLCV, which will regulate the expression of the polycistronic mRNA comprising the two sgRNAs and CSY4 binding sites for cleavage after transcription. Fonte: Do autor, 2020.

REFERENCES

BAE, S.; PARK, J.; KIM, J.S.Cas-OFFinder: a fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases. **Bioinformatics,**V. 30, n. 10, p.1473-1475, 2014.

BENEDITO, V.A. et al. Genomic inventory and transcriptional analysis of Medicagotruncatula transporters.**Plant physiology**, v. 152, n. 3, p.1716-1730, 2010.

ČERMÁK, T. et al.A multipurpose toolkit to enable advanced genome engineering in plants.**The Plant Cell**, v. 29, n.6, p.1196-1217, 2017.

DENOEUD, F. et al. The coffee genome provides insight into the convergent evolution OF CAFFEINE BIOSYNTHESIS. **SCIENCE,**V. 345, N.6201, P.1181-1184, 2014.

LABUN, K. et al. CHOPCHOP v3: expanding the CRISPR web toolbox beyond genome editing. **Nucleic acids research**, v. 47, n.W1, p.W171-W174, 2019.

TUSNADY, G.E.; SIMON, I.The HMMTOP transmembrane topology prediction server.**Bioinformatics**, v. 17, n.9, p.849-850, 2001.

UPADHYAY, N. et al. The multitasking abilities of MATE transporters in plants.**Journal of Experimental Botany**, v. 70, n. 18, p.4643-4656, 2019.