MARIANA CANSIAN SATTLER

CYTOGENOMICS IN *Coffea* L.

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Genética e Melhoramento, para obtenção do título de *Doctor Scientiae*.

Orientador: Wellington Ronildo Clarindo

VIÇOSA - MINAS GERAIS
2021

Ficha catalográfica elaborada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa

T
S253c
2021

Sattler, Mariana Cansian, 1994-
Cytogenomics in *Coffea* L. / Mariana Cansian Sattler. – Viçosa, MG, 2021.
1 tese eletrônica (82 f.): il. (algumas color.).

Orientador: Wellington Ronildo Clarindo.
Tese (doutorado) - Universidade Federal de Viçosa, Departamento de Biologia Geral, 2021.
Inclui bibliografia.
DOI: https://doi.org/10.47328/ufvbbt.2021.196
Modo de acesso: World Wide Web.

1. Café. 2. Genômica. 3. Citogenética. 4. DNA repetitivo. 5. Elementos genéticos móveis. 6. Poliploidia. I. Clarindo, Wellington Ronildo, 1979-. II. Universidade Federal de Viçosa. Departamento de Biologia Geral. Programa de Pós-Graduação em Genética e Melhoramento. III. Título.

CDD 22. ed. 572.8

Bibliotecário(a) responsável: Renata de Fátima Alves CRB6/2578

MARIANA CANSIAN SATTLER

CYTOGENOMICS IN *Coffea* L.

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Genética e Melhoramento, para obtenção do título de *Doctor Scientiae*.

APROVADA: 14 de outubro de 2021.

Assentimento:

Mariana Cansian Sattler
Autora

Wellington Ronildo Clarindo
Orientador

Aos meus pais, Sonia e Marcos

Ao meu irmão, Ramon

Dedico

AGRADECIMENTOS

À Universidade Federal de Viçosa e ao Programa de Pós-Graduação em Genética e Melhoramento, pela oportunidade de realização do curso e pela infraestrutura fornecida.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), à Fundação de Amparo à Pesquisa de Minas Gerais (FAPEMIG) e à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pelo suporte financeiro.

À Secretaria do Programa de Pós-Graduação em Genética e Melhoramento nas pessoas de Marco Túlio e Odilon, pelo auxílio e dedicação em todos os momentos.

Ao Prof. Wellington Ronildo Clarindo, pela orientação e amizade, e por fornecer conselhos e ensinamentos valiosos para o meu desenvolvimento profissional, científico e pessoal.

Às professoras Márcia, Maria Andréia e Milene, e ao professor Guilherme, pela participação na banca avaliadora e pelas colaborações fundamentais para a melhoria deste trabalho.

A todos os professores que passaram pela minha vida e contribuíram para formar as bases que fundamentam todo o conhecimento que possuo hoje, e também para a formação do meu caráter. Já dizia Paulo Freire, “O educador se eterniza em cada ser que educa.”

A todos os colegas do laboratório de Citogenética e Citometria Vegetal pelas contribuições técnicas e científicas, e pelos momentos de descontração. Em especial à Jéssica, por todo o apoio na bancada e pelas discussões que me levaram ao encantamento pelo mundo dos elementos móveis.

Aos meus pais, Sonia e Marcos, que são a base dessa conquista, pelo amor e apoio incondicionais, e por serem meus maiores exemplos de vida. Ao meu irmão, Ramon, pelo amor fraterno e amizade.

“The Cosmos is within us. We are made of star-stuff. We are a way for the Cosmos to know itself”.

(Carl Edward Sagan)

ABSTRACT

SATTLER, Mariana Cansian, D.Sc., Universidade Federal de Viçosa, October, 2021. **Cytogenomics in *Coffea* L.** Adviser: Wellington Ronildo Clarindo.

The genus *Coffea* comprises ~124 species, including *C. arabica* and *C. canephora,* which are responsible for the commercial production of coffee beans. Although most cytogenomic researches are focused on commercial species and its relatives, efforts have been made to expand the scope to wild species. However, molecular cytogenetic data are still limited to karyotype organization, especially due to low quality of the cytogenetic preparations. Considering this, the present work is focused on the cytogenomics of *Coffea* species. The first chapter provides an up to date review on the history of *Coffea* cytogenomics, initiating with the first classical cytogenetic studies, encompassing the main challenges and landmarks in cytogenetics and genomics, as well as their integration. These cytogenomic data allowed us to understand the phylogenetic relationships in *Coffea*, as well as their genomic origins, highlighting the relatively recent events of allopolyploidy. These events include the origin of the allotetraploid *C. arabica* (2n = 4x = 44, *C. canephora* x *C. eugenioides*) and the allotriploid hybrid 'Híbrido de Timor' (2n =3 x = 33, *C. arabica* x *C. canephora*). The second chapter is devoted to the study of repetitive sequences (repeatome) of the *C. eugenioides* genome, by integrating bioinformatic tools with cytogenetic mapping. We showed that repetitive sequences comprise ~47% of *C. eugenioides* genome, with mobile elements representing 45%. The Ty3/Gypsy to Ty1/Copia ratio was high (32:4), as also observed for other *Coffea* species. We mapped three Ty1/Copia (Bianca, TAR and Tork) and one Ty3/Gypsy (Athila) LTR-retrotransposon in metaphase chromosomes and interphase nuclei of *C. eugenioides*. Their distribution exhibited a clustered pattern throughout different chromosomes and regions of the chromosomes. The results obtained here are unprecedented for *Coffea*. Thus, we hope they lay the background for further investigations regarding *Coffea* cytogenomics.

Keywords: Coffee. Molecular Cytogenetics. Polyploidy. Genomics. Repeatome. Mobile Elements.

RESUMO

SATTLER, Mariana Cansian, D.Sc., Universidade Federal de Viçosa, outubro de 2021. **Citogenômica em *Coffea* L.** Orientador: Wellington Ronildo Clarindo.

O gênero Coffea compreende ~124 espécies, incluindo *C. arabica* e *C. canephora*, as quais são responsáveis pela produção comercial de grãos de café. Embora a maioria das pesquisas citogenômicas foquem nas espécies comerciais ou espécies relacionadas, esforços têm sido feitos para incluir espécies selvagens. No entanto, os dados de citogenética molecular ainda se limitam à organização do cariótipo, especialmente devido à baixa qualidade das preparações citogenéticas. Considerando o exposto, o presente trabalho tem como foco a citogenômica de espécies de *Coffea*. O primeiro capítulo apresenta uma revisão atualizada acerca da história da citogenômica em *Coffea*, iniciando com os primeiros estudos de citogenética clássica, abrangendo os principais desafios e marcos na citogenética e na genômica, assim como sua integração. Esses dados citogenômicos nos permitiram compreender as relações filogenéticas em *Coffea*, bem como suas origens genômicas, destacando os eventos relativamente recentes de alopoliploidia. Esses eventos incluem a origem do alotetraploide *C. arabica* (2n = 4x = 44, *C. canephora* x *C. eugenioides*) e do híbrido alotriploide 'Híbrido de Timor' (2n = 3 x = 33, *C. arabica* x *C. canephora*). O segundo capítulo é dedicado ao estudo de sequências repetitivas (repetoma) do genoma de *C. eugenioides*, por meio da integração de ferramentas bioinformáticas com mapeamento citogenético. Mostramos que as sequências repetitivas compreendem ~47% do genoma de *C. eugenioides*, com os elementos móveis representando 45%. A relação de Ty3/Gypsy para Ty1/Copia foi alta (32:4), assim como observado para outras espécies de *Coffea*. Nós também mapeamos três LTR-retrotransposons Ty1/Copia (Bianca, TAR e Tork) e um elemento Ty3/Gypsy (Athila) em cromossomos metafásicos e núcleos interfásicos de *C. eugenioides*. Sua distribuição exibiu um padrão agrupado, ao longo diferentes cromossomos e regiões dos cromossomos. Os resultados obtidos aqui são inéditos para *Coffea*. Portanto, esperamos que eles sirvam como base para futuras investigações a respeito da citogenômica de *Coffea*.

Palavras-chave: Café. Citogenética Molecular. Poliploidia. Genômica. Repetoma. Elementos Móveis.

SUMÁRIO

CHAPTER 1 – GENERAL INTRODUCTION

REVIEW PAPER – *Coffea* CYTOGENETICS: FROM THE FIRST KARYOTYPES TO THE MEETING WITH GENOMICS

Mariana Cansian Sattler[1]✉, Stéfanie Cristina de Oliveira[2], Maria Andréia Corrêa Mendonça[3], Wellington Ronildo Clarindo[1]

[1]Laboratório de Citogenética e Citometria, Departamento de Biologia Geral, Universidade Federal de Viçosa. ZIP 36.570-900 Viçosa – MG, Brazil.

[2]Laboratório de Citogenética e Cultura de Tecidos Vegetais, Campus de Alegre, Universidade Federal do Espírito Santo. ZIP 29.500-000 Alegre – ES, Brazil.

[3]Laboratório de Biotecnologia, Instituto Federal Goiano, Campus Rio Verde, ZIP 75.901-970 Rio Verde – MG, Brazil.

✉Corresponding author e-mail: mariana.sattler@ufv.br

Abstract

Coffea possesses ~124 species, including *C. arabica* and *C. canephora* responsible for commercial coffee production. We reviewed the *Coffea* cytogenetics, from the first chromosome counting, encompassing the karyotype characterization, chromosome DNA content, and mapping of chromosome portions and DNA sequences, until the integration with genomics. We also showed new data about *Coffea* karyotype. The 2n chromosome number evidenced the diploidy of almost all *Coffea*, and the *C. arabica* tetraploidy, as well as the polyploidy of other hybrids. Since then, other genomic similarities and divergences among the *Coffea* have been shown by karyotype morphology, nuclear and chromosomal C-value, AT and GC rich chromosome portions, and repetitive sequence and gene mapping. These cytogenomic data allowed us to know and understand the phylogenetic relations in *Coffea*, as well as their ploidy level and genomic origin, highlighting the relatively recent allopolyploidy. In addition to the euploidy, the role of the mobile elements in *Coffea* diversification is increasingly more evident, and the comparative analysis of their structure and distribution on the genome of different species is in the spotlight for future research. An integrative look at all these data is fundamental for a deeper understanding of *Coffea* karyotype evolution, including the key role of polyploidy in *C. arabica* origin. The 'Híbrido de Timor', a recent natural allotriploid, is also in the spotlight for its potential as a source of resistance genes and model for plant polyploidy research. Considering this, we also present some unprecedented results about the exciting evolutionary history of these polyploid *Coffea*.

Keywords: Classical Cytogenetics, Coffee, Cytogenomics, DNA content, Molecular Cytogenetics, Plant Breeding, Polyploidy.

1. Introduction

Coffee is one of the most important beverages worldwide, with more than 2.25 billion cups consumed daily (Denoeud et al. 2014). More than 75% of traded coffee is produced by *Coffea arabica* L. and *Coffea canephora* Pierre ex A. Froehner. *Coffea* L. (Rubiaceae) genus has a recent origin (7.87 Mya, Tosh et al. 2013) and includes more than 124 species native to Africa, Australasia, Comoros, India, Madagascar, Mascarenes and Papua New Guinea. Almost all species are diploid (2n = 2x = 22) and self-incompatible, except for the tetraploid (2n = 4x = 44) and self-compatible *C. arabica* (Hamon et al. 2017).

Cytogenetics concerns genomic and epigenomic aspects of the chromosomes, including the number, structure, organization, function, evolution, and behavior during the cell cycle and meiosis (Singh 2016; Deakin et al. 2019). *Coffea* cytogenetics was initiated in the early twentieth century with the work of von Faber (1912). As in other plant species, limitations have hampered the cytogenetic investigations in *Coffea*. The main problem has been to obtain mitotic and/or meiotic chromosomes, i. e., the biological material to conduct cytogenetic studies. Even today, one of the main challenges is to seek alternatives to obtain *Coffea* chromosomes, mainly to cytogenomics that involves in situ hybridization applications. In addition, the small size of the chromosomes and low metaphasic index from root meristems also hampered *Coffea* cytogenetics. The further advance of protocols allowed the progress in *Coffea* chromosome complement research, such as the analysis of pachytene chromosomes (Pinto-Maglio and Cruz 1987), improved methods for mitotic chromosome preparation and the replacement of root meristems by cell aggregate suspensions (CAS, Clarindo and Carvalho 2006). From this, karyotype has been expanded and refined, contributing to a better understanding of the *Coffea* genome. So, *Coffea* cytogenetics is important for taxonomic and evolutionary studies, DNA sequence mapping, functional genome annotation (Hamon et al. 2009; Yuyama et al. 2012) and coffee breeding programs (Clarindo and Carvalho 2009). In this review, we revisited all cytogenetic *Coffea* studies to report the advances and contributions in this genomic era, as well as the main challenges and perspectives for further studies. In addition, we report new data about the 2n karyotype of the diploid *C. eugenioides* and the polyploid “Híbrido de

Timor " (HT), expanding the genomic data about *Coffea* and its diversification and evolution.

2. Initial steps – chromosome number determination

The first chromosome counting in a *Coffea* from micro- and megasporogenesis evidenced that *C. arabica* shows 2n = 16 chromosomes (von Faber 1912). Twenty years later, Homeyer (1932) counted 2n = 22 chromosomes for the same species. Based on chromosome numbers of other Rubiaceae genera, such as *Sherardia* L., *Crucianella* L., *Asperula* L. and *Galium* L., Homeyer suggested the basic chromosome number of x = 11 for *Coffea*. Most striking breakthroughs in *Coffea* cytogenetics were achieved in the 1930s, mainly at the Instituto Agronômico de Campinas, Brazil, and by French and Belgian researchers. Krug (1934) reported 2n = 44 chromosomes for five *C. arabica* varieties (‘Nacional’, ‘Bourbon’, ‘Laurina’, ‘Maragogipe’ and ‘Amarelo de Botucatu’) and 2n = 22 for *C. canephora*, *Coffea liberica* Hiern and *Coffea congensis* A. Froehner, confirming the basic chromosome number of x = 11. Therefore, from the 2n chromosome number, *C. arabica* was noticed as a tetraploid species of *Coffea* (Krug 1934) and started to stand out not only for its economic relevance but also for the genomic events involved in its karyotype evolution. Despite being a tetraploid, *C. arabica* demonstrates a diploid meiotic behavior, forming only bivalents from prophase I until metaphase I. Although Krug did not carry out a detailed morphological analysis, he described the *Coffea* chromosomes as small (approximately 1 – 2 µm) and homomorphic. Due to these karyotype features, some cytogenetic applications have been applied to characterize the chromosomes and to understand the evolution of the *Coffea* genome. In the following years, the chromosome number of many other species, varieties and hybrids of *Coffea* was determined as 2n = 22 or 2n = 44 (Krug 1937; Mendes 1938; Bouharmont 1959, 1963; Sybenga 1960; Conagin and Mendes 1961).

3. Karyotype morphology and the first karyograms

Coffea chromosomes have been considered a hindrance for karyogram assembly due to their small size and high similar morphology. Mendes (1938) was the first to report a morphological characterization of *Coffea* chromosomes, describing solely those of *Coffea liberica* 'Dewevrei' (De Wild. & T. Durand) Lebrun (syn. *Coffea excelsa* Chev.). Although *C. liberica* 'Dewevrei' chromosomes were larger compared to those of other *Coffea* species, they were also considered small and morphologically similar, ranging from 3.5 (chromosome 1) to 1.5 μm (chromosome 11). Only three classes of chromosomes were distinguishable based on their total length, being designated as A, B and C. Class A was composed of three chromosome pairs with 2 – 3.5 μm, class B of four pairs with around 2 μm, and class C of four pairs with 1 – 2 μm (Mendes 1938). Bouharmont (1959, 1963) determined the number and length of the mitotic chromosomes of sixteen *Coffea* species, characterizing them as small and homomorphic and highlighting the hypothesis that the *Coffea* basic chromosome number is x = 11. Owing to the high intra- and interspecific similarity among chromosomes, the author was only able to distinguish five (1, 2, 3, 4 and 11) of the eleven chromosomes.

Pachytene chromosomes allowed to overcome the hindrance associated with the small size of mitotic metaphase chromosomes. The characterization of pachytene chromosomes with the NOR provided insights on the *C. arabica* polyploid origin by comparing the bivalent morphology with those of diploid species. One *C. arabica* NOR bivalent was similar to the *Coffea eugenioides* S. Moore and *C. liberica* 'Dewevrei' with regard to total and arm lengths. The other *C. arabica* NOR bivalent showed a simpler chromomeric pattern, most similar to the bivalents II of *Coffea salvatrix* Swynn. & Philipson and *Coffea racemosa* Lour (Pinto-Maglio and Cruz 1987).

C. arabica 22 pachytene chromosomes were after characterized, being four metacentric (1, 8, 11 and 18), fourteen submetacentric (2, 3, 4, 5, 6, 7, 9, 10, 12, 15, 16, 17, 19 and 20) and four acrocentric (13, 14, 21 and 22). Three bivalents presented the NOR (14, 20 and 21). In general, the chromomeres were located near the centromere, and similarities concerning the chromomeric pattern were shown among 54% of the 22 bivalents of *C. arabica*, evidencing some level of homoeology between

them. Based on these cytogenetic markers and by gathering information from the bibliography, the authors suggested that such similarities among *C. arabica* bivalents are consistent with a segmental allopolyploid origin (Pinto-Maglio and Cruz 1998). So, the mitotic cytogenetics showed the ploidy level of the *Coffea* species, mainly the tetraploidy of *C. arabica*, and the meiotic cytogenetics indicated the genomic origin of this species. Unfortunately, new data has not been published about the *Coffea* meiotic chromosomes.

Besides the small size and homogeneous morphology of *Coffea* chromosomes, the low seed germination rate and low metaphase index have also been pointed out as hindrances for cytogenetics (Conagin and Mendes 1961; Iacia and Pinto-Maglio 2013). To overcome this barrier, our research group (Clarindo and Carvalho 2006, 2008, 2009; Clarindo et al. 2012) replaced the root meristems by in vitro CAS as source material (Box 1, Fig. S1, Supplementary Information). Thereby, high metaphase indexes for *C. canephora*, *C. congensis*, *C. eugenioides* and *C. arabica* have been achieved, allowing to obtain chromosomes with total lengths of up to 5 μm (5.30 μm for chromosome 1 of *C. arabica*, for instance, Clarindo and Carvalho 2008). Such chromosomes were suitable for the assembly of the first karyograms at distinct levels of chromatin condensation, as well as the application of banding procedures. *C. canephora* and *C. congensis* possess 2n = 2x = 22 chromosomes, with two metacentric (4 and 9) and nine submetacentric pairs (1, 2, 3, 5, 6, 7, 8, 10 and 11). The submetacentric chromosome pair 6 of both species shows a secondary constriction (SC) in the short arm, which was confirmed for *C. canephora* by Ag-NOR and Hsc-FA. *C. arabica* has 2n = 4x = 44 chromosomes, being five pairs classified as metacentric (7, 8, 13, 14 and 20), sixteen as submetacentric (1-6, 9-12, 15-19 and 21) and one pair as acrocentric (22).

“Híbrido de Timor '' (HT), a natural hybrid between *C. canephora* and *C. arabica*, had its karyotype characterized and karyogram assembled by our adapted protocol. The semi-fertile HT ‘CIFC 4106’, a vegetatively propagated accession derived from the original HT plant, shows 2n = 3x = 33 chromosomes. Therefore, it was considered an allotriploid (Clarindo et al. 2013). We classified the chromosomes, assembled the HT ‘CIFC 4106’ karyogram and showed it for the first time now. Due to its anorthoploid condition (odd number of chromosome complements), the karyotype exhibits not only chromosome pairs, but also individual ones and groups of three or four chromosomes

(Fig. 1). Some chromosomes appear to be more similar to those of *C. canephora* or *C. arabica*, providing new cytogenetic information regarding the allopolyploid origin of HT. Such evolutionary aspects of the *Coffea* genome and interspecific hybridizations throughout the history of the genus will be discussed in the following topics.

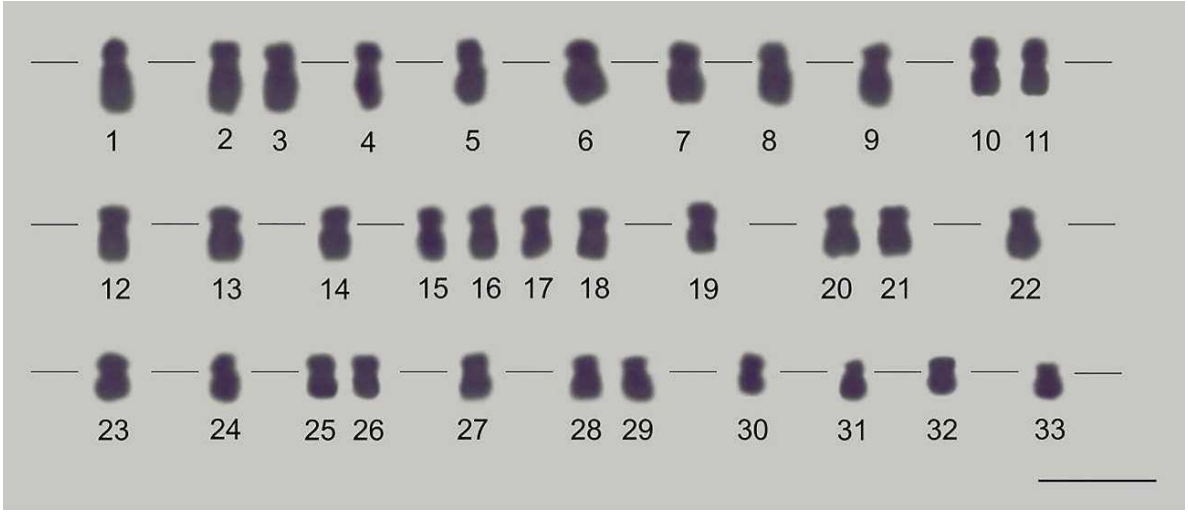

Figure 1. HT 'CIFC karyogram 4106' evidencing individual and grouped chromosomes, which were defined according to the total length and classification. 19 chromosomes (1, 4 – 9, 12 – 14, 19, 22 – 24, 27, 30 – 33) showed at least one particular cytogenetic cytogenetic feature. The paired chromosomes were 2 – 3, 10 – 11, 20 – 21, 25 – 26, and 28 – 29; and the chromosomes grouped in four were 15 – 18. Bar = 5 µm. Source: Oliveira (2017).

4. Chromosome banding and the onset of *Coffea* cytogenomics

C- and NOR-banding were initially applied in *Coffea* for mapping the constitutive heterochromatin and NOR, respectively. C-bands in *Coffea* occur preferentially at pericentromeric/centromeric regions. Corroborating to karyotype characterization, C-banding shows the karyotype asymmetry of the genus, varying from 62.3% (*C. eugenioides*) and 64.32% (*C. liberica*), with the prevalence of submetacentric chromosomes. On the other hand, the number of active NOR sites vary, with *C. kapakata*, *C. congensis*, *C. eugenioides*, *C. liberica*, *C. canephora* and *C. liberica* var. *dewevrei* displaying one pair with a positive NOR band (presumably number 3), and *C. racemosa*, *C. salvatrix* and *C. stenophylla* with two positive bands in chromosomes 1 and 3 (Pierozzi et al. 1999, 2012; Pierozzi 2013).

4'-6-diamidino-2-phenylindole ($DAPI^+$) and Chromomycin A_3 (CMA_3^+) differential staining have been carried out in *Coffea* species to map the AT- and GC-rich regions, respectively, contributing to identify specific karyotype signals in several species, even those with genetic or botanical status still under debate. $DAPI^+$ bands were only found in *C. stenophylla* among twelve analyzed *Coffea* species (Pinto-Maglio et al. 2000, 2001; Barbosa et al. 2001; Lombello and Pinto-Maglio 2004a, b, c). After DNA denaturation/renaturation in a fluorescent in situ hybridization (FISH), specific interstitial $DAPI^+$ bands in nine *Coffea* species were detected. CMA_3^+ bands were found in all *Coffea* species studied so far. In general, the strongest CMA_3^+ fluorescence signals are associated with the SC and co-localized with the ribosomal DNA (rDNA) loci detected by FISH. Additional CMA_3^+ bands may also occur adjacent to, or interspersed with, the interstitial 5S rDNA (Hamon et al. 2009).

Karyotype knowledge obtained from classical cytogenetics has been the basis for queries, which have been investigated mainly by cytogenomics, allowing a more refined *Coffea* genome characterization. Cytogenomics made possible to: (1) discover the diploid progenitors of *C. arabica* (Raina et al. 1998; Lashermes et al. 1999): (2) obtain information for application in breeding programs (Herrera et al. 2007); and (3) obtain a larger amount of details on the organization and evolution of *Coffea* genomes (Hamon et al. 2009). The identity of *C. arabica* diploid progenitors has been in the spotlight of *Coffea* research since the discovery of its tetraploid condition by Krug

(1934). Currently, wild *C. arabica* populations are found mainly in the Afromontane rainforests of southwest Ethiopia and the Boma Plateau of Sudan (Bawin et al. 2020). Although the geographic range of *C. arabica* does not overlap with that of any other *Coffea* species, the closeness with wild populations of *C. eugenioides* and of the Canephoroid species *C. canephora*, *C. congensis* and *C. brevipes* (Benth.) H.S. Irwin & Barneby, made these the most likely diploid progenitors. The first phylogenetic inferences based on chloroplast genes and rDNA confirmed the high genetic proximity among these species (Lashermes et al. 1995).

From genomic DNA probes (genomic in situ hybridization – GISH) of four diploid *Coffea* species, 22 *C. arabica* chromosomes hybridized preferentially with *C. eugenioides* probes, while the other 22 chromosomes hybridized more strongly with the *C. congensis* probes (Raina et al. 1998). Also using GISH, Lashermes et al. (1999) suggested that *C. canephora* and *C. eugenioides*, or their related ecotypes, are most likely the two diploid progenitors of the tetraploid *C. arabica*. In a phylogenomic approach using genotyping-by-sequencing, the genetic distances were estimated between *C. arabica* and 23 other species, including all those known as the most closely related to *C. arabica*, as *C. eugenioides*, *C. canephora*, *C. congensis* and *C. brevipes. C. eugenioides* and *C. canephora* were confirmed as the putative female and male progenitor species, respectively, that hybridized between 1.08 million and 543 thousand years ago (Bawin et al. 2020).

Integrating the $DAPI^+$, CMA_3^+ and rDNA sites mapping, two to five chromosome pairs were discriminated from the karyotypes of 16 *Coffea* species, including the allotetraploid *C. arabica*. However, the results were not suitable to discriminate between *C. canephora* and *C. congensis* as putative progenitors. Differences of the number of 18S and 5S rDNA locus were revealed and related to the biogeographical region of the 16 analyzed species. Most of the East African species, as *C. eugenioides*, *C. salvatrix* and *C. racemosa*, possessed two chromosome pairs with the 18S rDNA locus and one pair with the 5S rDNA, while the majority of the West and Central African species exhibited one chromosome pair with the 18S and two with the 5S rDNA (Hamon et al. 2009).

Alien chromatin in interspecific hybrids between *C. arabica* and *C. canephora*, as well as in an introgressed line derived from *C. arabica* and *C. liberica* crossing, was

identified from GISH and BAC-FISH (fluorescence in situ hybridization using bacterial artificial chromosomes). GISH results from the interspecific hybrids revealed close affinity between *C. arabica* and *C. canephora* genomes, evidencing that a low rate of structural modifications has occurred in both genomes since *C. arabica* speciation (Bawin et al. 2020). In addition, GISH/BAC-FISH in the introgressed line karyotype detected and physically located the *C. liberica*-introgressed DNA sequences carrying the S_H3 factor for resistance against *Hemileia vastatrix* Berk. & Broome (Herrera et al. 2007).

5. *Coffea* genomics and its integration with cytogenetics

During the 1990's, the availability of DNA based molecular markers allowed a rapid progress in coffee genomics. As for other crops, early genomic studies in *Coffea* were mainly focused on assessing the genetic diversity and phylogenetic relationships, constructing genetic maps and identifying quantitative trait loci (QTLs) (reviewed by Lashermes et al. 2008; de Kochko et al. 2010). Genetic diversity was investigated from several molecular markers in wild *Coffea* species and, with a larger effort, in *C. arabica* and *C. canephora*. The considerably narrow genetic basis of both wild and cultivated *C. arabica* populations was the most relevant feature noticed by these studies (Vossen 1985; Lashermes et al. 1996; Scalabrin et al. 2020). This bottleneck is mainly explained by the autogamous reproduction system, with a rate of outcrossing around 10%, and the founder effect resulted from the small number of individuals introduced to America in the first commercial plantations (Carvalho and Krug 1949; Setotaw et al. 2013). The low genetic variability is a challenge for the identification and selection of superior *C. arabica* genotypes. For this, molecular markers have played a fundamental role as to discriminate between genotypes (Sousa et al. 2017). Due to low genetic variability, breeding programs devoted efforts on the introgression of interesting genes from different species and hybrids, including *C. canephora* (Lashermes et al. 2011), *C. liberica* (Prakash et al. 2004) and HT (Setotaw et al. 2020). As the genomes of *Coffea* species exhibit considerable similarities, interspecific crossing (hybridization) is possible and often used for gene introgression (Charrier and Berthaud 1985; Anthony et al. 2011).

For *C. canephora*, the first genetic maps were constructed as soon as molecular markers became available for *Coffea* (reviewed by de Kochko et al. 2010). In 2014, the most complete genome sequence of *C. canephora* (accession number: PRJEB4211) was published along with a high-density genetic map constructed using several molecular markers, totalizing 3,230 loci distributed on 11 linkage groups, the same basic chromosome number of this species (x = 11). *C. canephora* genetic map was integrated with the sequenced genome, which covered 80% of the 710 Mbp total genome. There was a considerable variation in the physical:genetic map distances, with crossing overs occurring with higher frequencies in regions with lower density of repeats (Denoeud et al. 2014). The complexity of the *C. arabica* allotetraploid genome and its low genetic diversity were the main challenges to construct a high-density linkage map. Notwithstanding, a high-density linkage map was published for this species showing 22 linkage groups with 848 markers. This genetic map was also successfully used to identify QTLs associated with coffee yield, plant height, and bean size (Moncada et al. 2016).

As for the construction of genetic maps, the sequencing of a polyploid genome is also challenging, since the presence of homoeologous chromosomes hampers the assembly of each haploid complement. *C. arabica* draft genome (public accession number: PRJNA554647) was released in 2020, with the two component subgenomes independently assembled. In accordance with previous studies reported here, a low genetic diversity was confirmed, which was likely caused by a severe bottleneck resulting from a single event of polyploidization at the origin of the species *C. arabica* (Scalabrin et al. 2020). In addition to *C. canephora* and *C. acabica*, the *C. eugenioides* (accession number: PRJNA497891) and *Coffea humblotiana* Baill. (accession number: PRJNA665152, Raharimalala et al. 2021) genomes are also publicly available. The species *C. humblotiana* is the sole species from the *Coffea* genus endemic to the Comoros archipelago and was probably cultivated and consumed in the past. The most noteworthy feature of *C. humblotiana* is the complete absence of caffeine in seeds and leaves, which is explained by the loss of one of the *Caffeine Synthase* (*DXMT*) genes, which converts theobromine into caffeine, most likely through illegitimate recombination (Raharimalala et al. 2021). Despite the significant difference in genome size, there is a high degree of synteny between the genomes of *C. humblotiana* (2C =

0.97 pg, Razafinarivo et al. 2012) and *C. canephora* (2C = 1.43 pg, Clarindo et al. 2012) with 32.16% more nuclear DNA content.

Recent advances in next-generation sequencing technologies and bioinformatic tools provided a growing amount of available genomic data, laying the background for the integration with cytogenetics and giving rise to the cytogenomics era (Talukdar and Sinjushin 2015). The integration of the linkage and cytogenetic maps and sequencing data is fundamental to define genome regions that are not yet well-characterized (Kim et al. 2005). Because linkage map distances are based on recombination rates and not simply related to physical distances, physical mapping is also needed to confirm the locations of DNA sequences (Koo et al. 2008). Physical mapping might be obtained through genome sequencing or in situ localization using cytogenomics. The major challenge related to sequencing is the resolution of complex repetitive sequences. As the DNA must be fragmented into small contigs, usually around 100 bp, repeats create computational ambiguities during alignment and assembly, which might produce biases and errors when interpreting results (Schatz et al. 2012; Treangen and Salzberg 2012). Cytogenomics, on the other hand, is efficient to map repetitive sequences, revealing the physical localization in situ on the mitotic or meiotic chromosomes (Larracuente and Ferree 2015). Therefore, the linkage and cytogenetic maps and the sequencing are complementary, and integrating these data is fundamental for a more detailed and accurate *Coffea* genome knowdelege.

Completely *Coffea* sequenced genome released in 2014 for *C. canephora* showed that mobile elements represent more than 50% of its genome, among which ~85% belong to the LTR-retrotransposon class (Denoeud et al. 2014). Due to their high frequency (~15% to > 70%) and pivotal roles in plant genome organization, function and evolution, the analysis of mobile elements organization and distribution along the genome has been performed in several plant species (Civáň et al. 2011; Wicker et al. 2018). Mobile elements comprise DNA sequences with the ability to insert themselves (transposons) or new copies of themselves (retrotransposons) into new locations within a genome (Civáň et al. 2011). The comparative cytogenomic analysis of mobile elements distribution patterns among different species provides valuable information to understand the karyotype evolution in *Coffea*. For the sequenced genomes analyzed so far, the proportions of LTR-retrotransposons, for instance, varied from 32% for *C. humblotiana* (1C = 0.49 pg, Razafinarivo et al. 2012) to 53% for *Coffea heterocalyx*

Stoff. (1C = 0.87 pg, Noirot et al. 2003). The variation in the abundance and types of different mobile elements can reflect the divergence of botanical groups and also the evolution of species within these botanical groups (Guyot et al. 2016).

Some mobile elements have been mapped in *Coffea* species, including DNA transposons (Lopes et al. 2013), Long Terminal Repeat (LTR) retrotransposons (Yuyama et al. 2012; Herrera et al. 2013; Lopes et al. 2013) and centromeric retrotransposons (Nunes et al. 2018). Nonetheless, the low longitudinal resolution of chromosomes did not allow the precise mapping, but an overview of distribution patterns along the genomes. DNA transposons *MuDR* and *Tip100*, for instance, exhibit a preferential clustering in terminal positions of *C. canephora* and *C. eugenioides* chromosomes, while *C. arabica* showed larger numbers of interstitial signals. This distribution indicates an increased transposition activity in the allotetraploid (Lopes et al. 2013), which is consistent with the well-known hypothesis that polyploidization can induce a burst in mobile element activity (Vicient and Casacuberta 2017).

The detailed analysis of the centromeric mobile elements composition of *C. arabica*, *C. canephora* and *C. eugenioides* revealed a considerable diversity in centromeric retrotransposons of *Coffea* (CRC) from the Ty3/Gypsy superfamily, which were divided in ten groups according to the sequence and similarity of the Reverse Transcriptase domain. Generalist probes for these CRCs exhibited a variable fluorescence signal pattern among species and among chromosomes of the same species. While in *C. eugenioides* the signals were identified only to centromeric regions, in *C. canephora* and *C. arabica* the signals appeared slightly scattered along interstitial regions and less specific to centromeres. In addition, *C. arabica* presented two pairs without bright signals, which might be homologous to those chromosomes without signals from the parental diploids *C. canephora* and *C. eugenioides* (one pair each) (Nunes et al. 2018).

6. Allopolyploidy in *Coffea*

Coffea genus has a recent monophyletic origin around 5–25 Mya, and none event of whole genome duplication (WGD) occurred immediately prior to or after the irradiation of the Rubiaceae family (Orozco-Castillo et al. 1996; Wu et al. 2006; Mahé et al. 2007).

Therefore, the common ancestral shared by all *Coffea* species probably had the basic chromosome number of x = 11. In addition, the diversification and speciation of each *Coffea* species occurred by DNA sequence changes (mutations) and small chromosomal rearrangements, which have been progressively identified and characterized through classical cytogenetics and cytogenomics (Rijo 1974; Yu et al. 2011; Denoeud et al. 2014; Raharimalala et al. 2021). A micro-collinearity analysis between orthologous BACs of *C. canephora* and *C. arabica*, for instance, evidenced a high level of sequence similarity, but numerous small chromosomal rearrangements, including inversions, deletions and insertions (Yu et al. 2011).

6.1. Allotetraploidy and diversification of *C. arabica*

Krug's (1934) discovery the tetraploidy of *C. arabica* ignited a long debate on the origin (ancestry) and evolution ("omics" changes) of this species. As mentioned here, there was an agreement that *C. arabica* originated from hybridization between two diploid species with similar genomes, with the potential progenitors being *C. eugenioides* (Berthou 1983; Lopes et al. 1984; Orozco-Castillo et al. 1996; Raina et al. 1998; Lashermes et al. 1999; Ruas et al. 2003), *C. canephora* (Lashermes et al. 1997, 1999; Ruas et al. 2003; Clarindo and Carvalho 2009), *C. congensis* (Höfling and Oliveira 1981; Lashermes et al. 1997; Raina et al. 1998) and *C. brevipes* (Lashermes et al. 1997). Therefore, the focus has been the *C. arabica* ancestors. However, further studies should be accomplished to unravel the genomic and epigenomic outcomes of the *C. arabica* allopolyploid condition.

C. arabica polyploid origin from a crossing between two diploids with similar genomes is a consensus, but its genomic origin classification has been discussed. Initially, *C. arabica* was classified as a natural segmental allotetraploid that originated (Orozco-Castillo et al. 1996; Pinto-Maglio and Cruz 1998). 'Segmental allopolyploid' in this context was based on the Stebbins (1949) definition as a type of allopolyploid that contain two partially differentiated genomes, and that was originated from hybridization between species close enough to allow the partial pairing between homoeologous chromosomes. Thus, segmental allopolyploids are considered intermediaries between autopolyploids and true (or genomic) allopolyploids, as the differentiation between the

progenitor genomes is insufficient for complete allopolyploidy. Exceptionally, the allotetraploid *C. arabica* exhibits a stable diploid-like meiotic behavior (Mendelian segregation). This might be possible owing to the genetic system occurrence wherein the pairing between homoeologous chromosomes is avoided (Pinto-Maglio and Cruz 1998), such as the *Ph* gene (homoeologous pairing suppressor) found in *Triticum* L., *Avena* L., *Festuca* L., *Gossypium* L., *Nicotiana* L. and *Lolium* L. Further molecular evidence on this hypothesis suggested that the absence of homoeologous pairing in *C. arabica* is not a consequence of structural differentiation between the two parental genomes, but rather the effect of one or several pair-regulating genes, which could be similar to *Ph* (Lashermes et al. 2000).

C. arabica was also hypothesized as an amphidiploid species formed from the crossing between *C. eugenioides* as female progenitor and *C. canephora*, or its related ecotypes, as male progenitor. In addition, *C. arabica* origin is recent due to the low level of divergence between the two constitutive genomes of this species and the related parental genomes (Lashermes et al. 1999). Amphidiploid refers to segmental allotetraploids that have gone through an event of WGD after hybridization (Stebbins 1949). In the absence of homoeologous pairing, the WGD event may restore de fertility of the hybrid (homoploid), since the presence of two copies of each genome would enable the Mendelian pairing (bivalent) during meiosis I. The WGD event that gave rise to the fertile allotetraploid ancestors of *C. arabica* possibly involved either chromosome set doubling in a diploid interspecific hybrid or backcrossing of a spontaneous triploid (Lashermes et al. 1999).

The true allotetraploid nature of *C. arabica* was reinforced from classical cytogenetics and chromosomal image cytometry (Box 2) analyses (Clarindo and Carvalho 2008, 2009), corroborating with data based on molecular markers and GISH. The comparison with the karyotypes of two potential genitors, *C. canephora* and *C. congensis*, revealed the presence of identical chromosomes between *C. arabica* and both species concomitantly. In addition, *C. arabica* also exhibits a small acrocentric chromosome pair that is not present in either of these two species. Taken together these results also support the idea that only one of them participated in the origin of *C. arabica* (Clarindo and Carvalho 2009; Clarindo et al. 2012).

Cytogenetic evidence concerning the *C. arabica* origin was obtained from the SC/NOR chromosome. Three pachytene chromosomes of *C. arabica* (14, 20 and 21) have SC (Pinto-Maglio and Cruz 1998), which were NOR confirmed by 18S rDNA signals in mitotic metaphases (Hamon et al. 2009). *C. canephora* and *C. congensis* exhibit a single SC in the short arm of the chromosome 6 (Clarindo et al. 2012), also confirmed by 18S rDNA (Hamon et al. 2009). Therefore, one of the SC/NOR chromosomes in *C. arabica* karyotype was probably derived from the canephoroid genitor (most likely *C. canephora*), while the other two were possibly inherited from *C. eugenioides*.

From orthologous coding sequence divergence analysis, *C. arabica* evolutionary history was also evidenced as allotetraploid originated from a natural hybridization between *C. canephora* and *C. eugenioides*. Thus, the genome of *C. arabica* was represented by the genome formula $C^aC^aE^aE^a$ (Yu et al. 2011). The most accurate estimation for the genomic origin time of the natural hybridization is in between 0.54 and 1.08 Mya (Bawin et al. 2020). Although the current geographical range of *C. arabica* do not overlap with any other *Coffea* species, including its possible progenitors, pollen records and lake sediment cores from the Congo basin and East Africa indicate that the Afromontane rainforests regularly expanded to lower altitudes during glacial periods between 0.60 and 1.05 Mya. Corroborating, *C. canephora* and *C. eugenioides* occurred in contact zone in this area. The changing environmental conditions might have also played a role on this speciation event, weakening interspecific reproductive barriers between *C. arabica* diploid parental species. In addition, the aridification of East Africa since 0.575 Mya is one of the possible explanations for the current non-overlapped distribution areas of *C. arabica*, *C. eugenioides* and *C. canephora* (Owen et al. 2018; Bawin et al. 2020).

C. canephora and *C. arabica* 'Tall Mokka' comparison from BAC clones (~140 – 160 kb), bearing the aforementioned orthologous coding sequences, aimed to show the outcomes of the *C. arabica* allopolyploidy from a cytogenomic perspective. Despite the high degree of sequence conservation in coding regions, genomic differences were found. Major chromosomal rearrangements were observed in the intergenic regions of these BACs, including a paracentric inversion between homoeologous regions within *C. arabica* (C^a and E^a). As *C. eugenioides* was not included in this study, it was not possible to distinguish if the chromosomal inversion occurred in *C. arabica* or was

inherited from one of the two diploid progenitors. Therefore, the inversion might not be a consequence of the C. *arabica* allopolyploidy. Moreover, the specific insertion of a *Ty1-copia* retrotransposon in the C^a sub-genome of *C. arabica* was also reported, which might be related to the burst in TE activity (Yu et al. 2011).

6.2. Allotriploidy of the Timor Hybrid - the recurrent natural allopolyploidy in *Coffea*

In addition to the polyploidization event that gave rise to the well-established allotetraploid *C. arabica*, another recent allopolyploidy in the *Coffea* genus originated the hybrid named HT, or the Timor hybrid. HT arose in a plantation of *C. arabica* 'Typica', established around 1917/18 on the Timor Island, from the natural interspecific hybridization between the allotetraploid *C. arabica* and the diploid *C. canephora*. All the accessions of this hybrid originated from this single plant, or from backcrosses between this plant and its progenitor *C. arabica* (Gonçalves et al. 1978). Besides the remarkable relevance of HT as a source of resistance genes in coffee breeding, this hybrid has also gained attention for its very recent allopolyploid origin of ~100 years (Gonçalves et al. 1978; Capucho et al. 2009). HT 'CIFC 4106', which has been vegetatively propagated, is a triploid hybrid with a chromosome number 2n = 3x = 33 and a nuclear DNA content of 1C = 2.10 pg. Therefore, we considered that this accession represents the first plant of HT found in Timor island. Considering the 2n = 33 chromosomes and that the 1C value of HT 'CIFC 4106' is close to the sum of the 1C values of *C. arabica* (1C = 1.335 pg) and *C. canephora* (1C = 0.73), HT 'CIFC 4106' is an allotriploid most likely originated from the fusion of one reduced reproductive cell of *C. arabica* (n = 2x = 22) with another of *C. canephora* (n = x = 11).

In summary, the hybridization/polyploidization events involving *C. canephora*, *C. eugenioides*, HT and *C. arabica* might be explained as follows. The allotetraploid *C. arabica* originated around 0.543 and 1.08 Mya (Bawin et al. 2020) from the fusion between a reduced reproductive cell from *C. canephora* (n = x = 11) and another from *C. eugenioides* (n = x = 11, Lashermes et al. 1999). After a polyploidization event in the sterile homoploid, fertility would be restored, resulting in the fertile allotetraploid with 2x = 4x = 44 chromosomes, being 22 from *C. canephora* and 22 from *C. eugenioides* (C^a and E^a subgenomes, respectively). A natural backcross dating from

~100 years ago occurred between *C. arabica* and its progenitor *C. canephora*, involving the fusion of reduced reproductive cells from both species and resulting in the HT with 2x = 3x = 33 chromosomes. Therefore, our research group recently launched the hypothesis that the genome of HT 'CIFC 4106' is represented by the formula CC^aE^a. In order to provide more information regarding the karyotypic evolution of *C. eugenioides*, *C. canephora*, *C. arabica* and HT 'CIFC 4106', we have been combining classical and molecular cytogenetics, as well as flow cytometry for nuclear DNA content measurements (Box 2).

As shown before, based on chromosome class, the HT 'CIFC 4106' karyotype exhibits six pairs of chromosomes (2 – 3, 10 – 11, 12 – 13, 20 – 21, 25 – 26, 28 – 29). Also, one group of four chromosomes was formed (15 – 18) and the remaining chromosomes did not show homologous pairs (Fig. 1, Fig. 2). Possibly, the six chromosome pairs represent the CC^a subgenomes, being C from *C. canephora* and C^a from *C. arabica*. A look at the HT 'CIFC 4106' karyogram shows that the possible pairs 2 – 3, 10 – 11, 12 – 13, 20 – 21, 25 – 26 and 28 – 29 may represent the chromosomes 1, 4, 5, 7, 9 and 11 of *C. canephora*, respectively (Fig. 2). HT 'CIFC 4106' group 15 – 18 is similar to pair 6 and 7 of *C. canephora* (Fig. 2). Another evidence that supports the CC^aE^a genome hypothesis, specifically in relation to the E genome, is that chromosomes 1, 4 and 9 of HT 'CIFC 4106' are similar to 1, 2 and 3 of *C. eugenioides,* respectively, considering morphometry. In addition, chromosomes 7 and 11 of *C. eugenioides* are similar to 19 and 33 of HT 'CIFC 4106', respectively (Fig. 2).

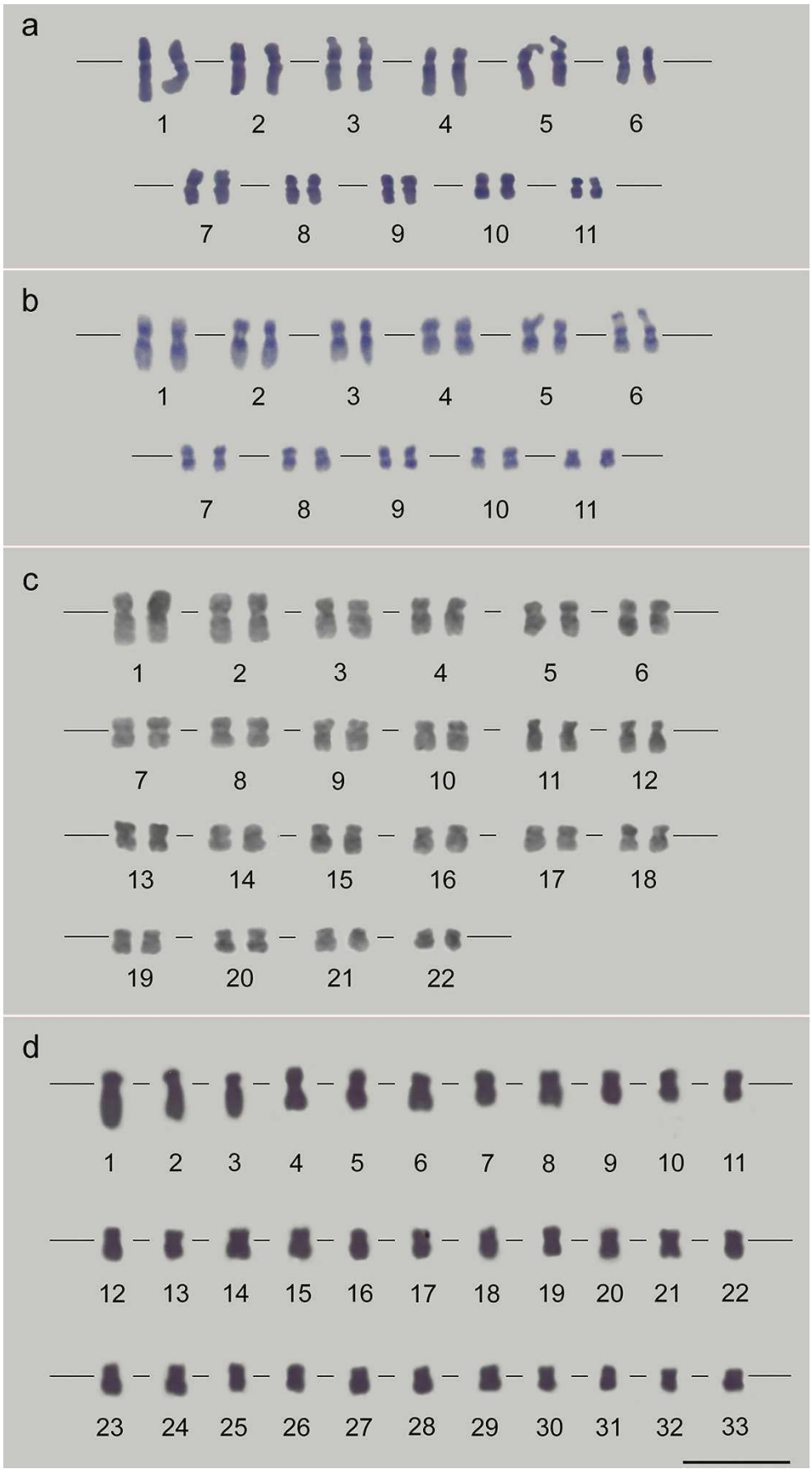
a
1 2 3 4 5 6
7 8 9 10 11
b
1 2 3 4 5 6
7 8 9 10 11
c
1 2 3 4 5 6
7 8 9 10 11 12
13 14 15 16 17 18
19 20 21 22
d
1 2 3 4 5 6 7 8 9 10 11
12 13 14 15 16 17 18 19 20 21 22
23 24 25 26 27 28 29 30 31 32 33

Figure 2. Karyograms obtained from metaphase chromosomes of: **a)** *C. eugenioides* exhibiting two metacentric chromosome pairs (7, 10), nine submetacentric (1 – 6, 8, 9 and 11) and two chromosomes pairs with SC (3 and 5); **b)** *C. canephora* with two metacentric chromosome pairs (4 and 9), nine submetacentric (1 – 3, 5 – 8, 10 and 11) and one with SC; **c)** *C. arabica* exhibits 5 metacentric chromosome pairs (7, 8, 13, 14 and 20), sixteen submetacentric (1 – 6, 9 – 12, 15 – 19 and 21) and one acrocentric pair (22); and **d)** HT 'CIFC 4106' possess six metacentric (10, 11, 19, 25, 26, 30) and twenty-seven submetacentric chromosomes (1 – 9, 12 – 18, 20 – 24, 27 – 29, 31 – 33). Bar = 5 µm. Source: Oliveira (2017).

Our new data about 5S rDNA mapping performed in *C. eugenioides* and *C. canephora* (Oliveira 2017) showed that the two diploids exhibit one 5S rDNA loci, in the interstitial region of chromosome 3 long arm for *C. eugenioides*, and in the pericentromeric portion of the long arm of chromosome 8 in *C. canephora*. The allotriploid HT 'CIFC 4106' shows two loci located interstitially in the long arm of chromosomes 9 and 13 (Fig. 3). The association of these data with chromosome class provide additional evidence for HT origin, as *C. eugenioides* chromosome 3 and HT 'CIFC 4106' chromosome 9 both exhibit one 5S rDNA loci in the same position. Therefore, the HT 'CIFC 4106' chromosome 9 was probably inherited from a reduced cell of *C. arabica* (C^aE^a), precisely from the E^a subgenome. In addition, the nuclear DNA content of HT 'CIFC 4106' (1C = 2.10 pg, Clarindo et al. 2013) is equivalent to the sum of the mean 2C nuclear genome size of *C. canephora* (2C = 1.41 pg, 1C = 0.705, Clarindo and Carvalho 2009) and the 1C nuclear value of *C. eugenioides* (2C = 1.38 pg, 1C = 0.690 pg, Sanglard et al. 2019) and HT 'CIFC 4106' chromosome number (2n = 3x = 33) corresponds to the fusion of one reproductive cell of *C. arabica* (C^aE^a; n = 2x = 22) and one of *C. canephora* (C; n = x = 11). These data also support the CC^aE^a hypothesis. A summary of the main cytogenetic features and evolutionary relations among the two allopolyploid *Coffea* and its parental diploids is depicted in Fig. 4.

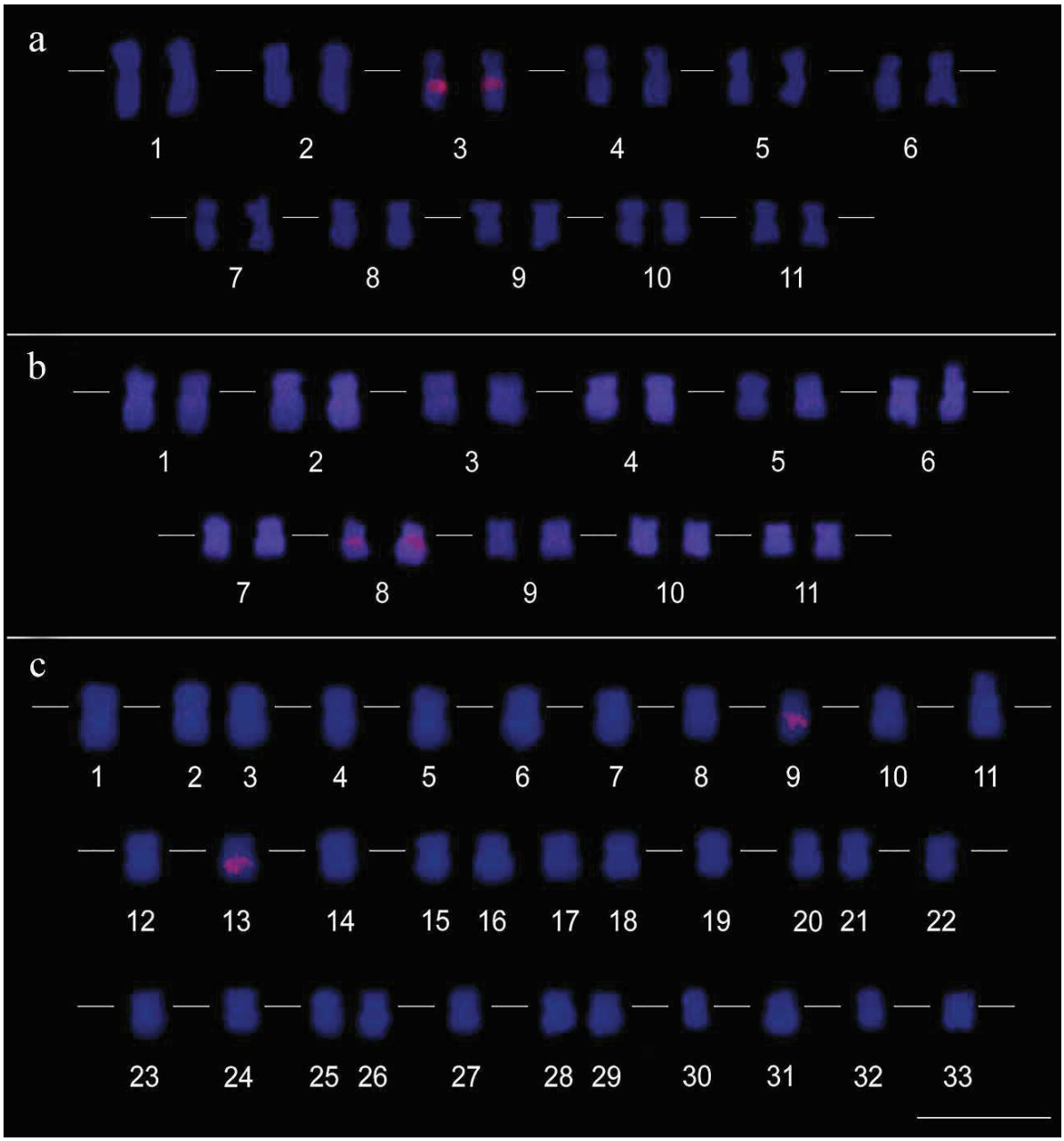
a
1
2
3
4
5
6
7
8
9
10
11
b
1
2
3
4
5
6
7
8
9
10
11
c
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33

Figure 3. FISH mapping of 5S rDNA genes (red) on *Coffea* metaphase chromosomes using probes labeled with tetramethyl-rhodamine 5-dUTP. Chromosomes were counterstained with 4',6-diamidino-2-phenylindole dihydrochloride (DAPI) (blue). **a)** *C. eugenioides* with the 5S rDNA signal at the interstitial region of chromosome 3 long arm. **b)** *C. canephora* exhibited a pericentromeric signal in the long arm of chromosome 8. **c)** HT 'CIFC 4106' has two positive 5S rDNA signals, one in chromosome 9 and the other in chromosome 13, both interstitial. Bars = 5 µm. Modified from Oliveira (2017).

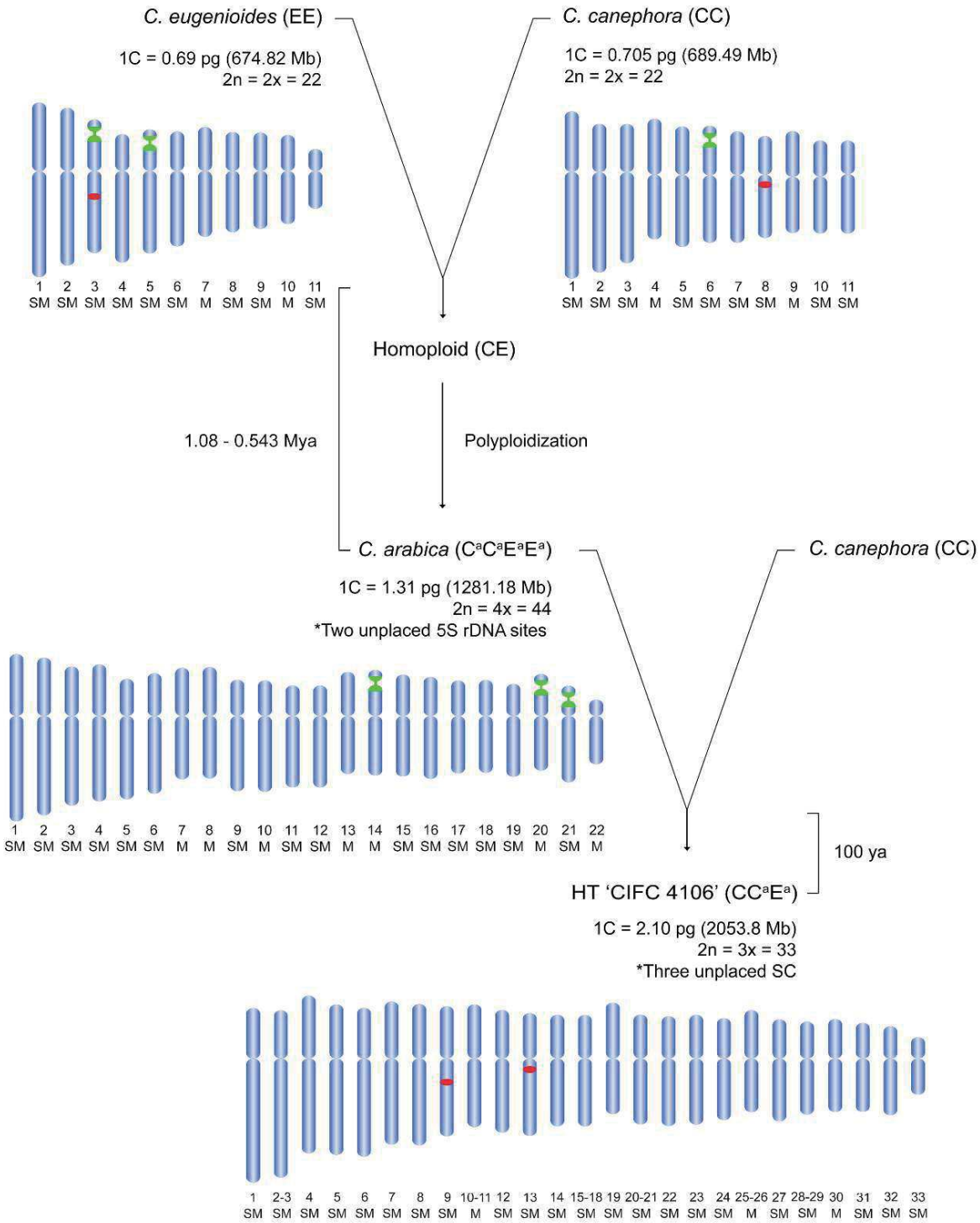
C. eugenioides (EE)
1C = 0.69 pg (674.82 Mb)
2n = 2x = 22
1 SM 2 SM 3 SM 4 SM 5 SM 6 SM 7 M 8 SM 9 SM 10 M 11 SM
C. canephora (CC)
1C = 0.705 pg (689.49 Mb)
2n = 2x = 22
1 SM 2 SM 3 SM 4 M 5 SM 6 SM 7 SM 8 SM 9 M 10 SM 11 SM
Homoploid (CE)
1.08 - 0.543 Mya
Polyploidization
C. arabica (CaCaEaEa)
1C = 1.31 pg (1281.18 Mb)
2n = 4x = 44
*Two unplaced 5S rDNA sites
1 SM 2 SM 3 SM 4 SM 5 SM 6 SM 7 M 8 M 9 SM 10 M 11 SM 12 SM 13 M 14 M 15 SM 16 SM 17 SM 18 SM 19 SM 20 M 21 SM 22 M
C. canephora (CC)
100 ya
HT 'CIFC 4106' (CCaEa)
1C = 2.10 pg (2053.8 Mb)
2n = 3x = 33
*Three unplaced SC
1 SM 2-3 SM 4 SM 5 SM 6 SM 7 SM 8 SM 9 SM 10-11 M 12 SM 13 SM 14 SM 15-18 SM 19 SM 20-21 SM 22 SM 23 SM 24 SM 25-26 M 27 SM 28-29 SM 30 M 31 SM 32 SM 33 SM

Figure 4. Summary of the main cytogenetic features with representative idiograms of two *Coffea* allopolyploids, C*. arabica* and HT 'CIFC 4106', and the diploids *C. canephora* and *C. eugenioides*, as well as their evolutionary relations. The data presented for these *Coffea* were obtained from Pinto-Maglio and Cruz (1998), Noirot et al. (2003), Clarindo and Carvalho (2006, 2009), Hamon et al. (2009), Bawin et al. (2020) and from the present work. For these *Coffea* were reported: 1C nuclear DNA content measured by flow cytometry; 2n chromosome number; ploidy level; class of each chromosome of the karyotype (M – metacentric, SM – submetacentric, A – acrocentric); number and chromosomes with secondary constrictions (SC); number of 18S (green) and 5S (red) rDNA sites.

Natural neoallopolyploids are valuable materials intensively used as evolutionary models in plant polyploidy research, such as *Spartina anglica* C. E. Hubb. (Ainouche et al. 2004), *Senecio cambrensis* Rosser, *Senecio eboracensis* Abbott & Lowe (Abbott and Lowe 2004), *Tragopogon mirus* Ownbey, *Tragopogon miscellus* Ownbey (Soltis et al. 2004), *Cardamine* × *schulzii* Urbanska-Worytkiewicz (Urbanska et al. 1997) and *Mimulus peregrinus* Vallejo-Marín (Vallejo-Marín 2012). As HT is a recent allotriploid originated only ~100 years ago, the refined study of its karyotype, in comparison to its progenitor *C. arabica*, would certainly provide substantial information on the genomic rearrangements associated with interspecific hybridization and polyploidy. Therefore, HT may also be considered a potential and valuable evolutionary model in plant polyploidy research.

7. Concluding remarks

Advances in *Coffea* cytogenetics over the years have been remarkable. Thus far, the development of improved protocols for classical and molecular cytogenetics, alongside with genomics, allowed a deeper understanding on the structure, organization and evolution of *Coffea* genomes. The relevance of these data also reaches applied research, with flow cytometry and karyotyping being used, for instance, to distinguish species, varieties and hybrids, including HT cytotypes. The

potential role of cytogenetics and flow/image cytometries on aiding *Coffea* genomes assembly is also worthy of notice (Box 2). The aim of any genome sequencing project is to achieve an assembly to the chromosome level, with each scaffold assigned and oriented onto a chromosome. This assignment of sequence contigs to chromosomes is often reached by integrating data from sequencing with those from molecular cytogenetic mapping. The basis of another relevant information, sequencing coverage, relies on reliably measuring the total genome size of the target species, for which flow cytometry has been considered the 'gold standard technique'. In addition to nuclear genome size, both flow and image cytometries also allow measuring the size of each chromosome of the karyotype, aiding to estimate the coverage at the chromosome level.

The increase of the sequencing information will also allow a deeper analysis of *Coffea* evolution, and the integration with the molecular cytogenetics is fundamental to associate the ex situ with the in situ localization of DNA sequences. Nonetheless, mitotic metaphases and pachytene chromosomes from different species of *Coffea*, especially wild species of strict geographical distribution, will still be one of the main challenges for cytogenomics. Partial genome sequence analysis already revealed interesting patterns regarding the association of mobile elements with *Coffea* diversification and the recent publication of a new genome sequence from the wild *C. humblotiana* will open new fields for comparative analysis. We also believe that the allotriploid HT is an interesting target for cytogenomics, especially to study the early genomic dynamics after an allopolyploidization event, such as the TE burst, sequence losses or gains and global effects on epigenetic regulation. Therefore, we hope that the continuous improvement in techniques and a closer integration of different research areas will lay the ground for future ground-breaking discoveries about *Coffea* genome.

8. Acknowledgments

We would like to thank the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, Brasília – DF, Brazil), Fundação de Amparo à Pesquisa do Espírito Santo (FAPES, Vitória-ES, Brazil; grants: 65942604/2014 and 82/2017), and

Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES, Brasília – DF, Brazil) for financial support.

9. References

Abbott RJ, Lowe AJ (2004) Origins, establishment and evolution of new polyploid species: *Senecio cambrensis* and *S. eboracensis* in the British Isles. Biol J Linn Soc 82:467–474. https://doi.org/10.1111/j.1095-8312.2004.00333.x

Ainouche ML, Baumel A, Salmon A (2004) *Spartina anglica* C. E. Hubbard: a natural model system for analysing early evolutionary changes that affect allopolyploid genomes. Biol J Linn Soc 82:475–484. https://doi.org/10.1111/j.1095-8312.2004.00334.x

Anthony F, Bertrand B, Etienne H, Lashermes P (2011) *Coffea* and *Psilanthus*. In: Wild crop relatives: genomic and breeding resources. Springer, Berlin, pp 41–61

Barbosa R, Pierozzi N, Silvarolla M, Pinto-Maglio C (2001) Caracterização citomolecular de três espécies de *Coffea* L.: *C. eugenioides*, *C. canephora*, e *C. dewevrei*. In: II Simpósio De Pesquisa Dos Cafés Do Brasil. Vitória, Brazil, pp 293–298

Bawin Y, Ruttink T, Staelens A, et al (2020) Phylogenomic analysis clarifies the evolutionary origin of *Coffea arabica*. J Syst Evol jse.12694. https://doi.org/10.1111/jse.12694

Berthou F (1983) Chloroplast and mitochondrial DNA variation as indicator of phylogenetic relationships in the genus *Coffea* L. Theor Appl Genet 65:77–84. https://doi.org/10.1007/BF00276268

Bouharmont J (1959) Recherches sur les affinités chromosomiques dans le genre *Coffea*. Publications de l'Institut National pour l'Etude Agronomique du Congo Belge

Bouharmont J (1963) Somatic chromosomes of *Coffea* species. Euphytica 12:254–257

Capucho AS, Caixeta ET, Zambolim EM, Zambolim L (2009) Herança da resistência do Híbrido de Timor UFV 443-03 à ferrugem-do-cafeeiro. Pesq Agropec Bras 44:276–282. https://doi.org/10.1590/S0100-204X2009000300009

Carvalho A, Krug C (1949) Agentes de polinização da flor do cafeeiro (*Coffea arabica* L.). Bragantia 9:11–24. https://doi.org/10.1590/S0006-87051949000100002

Charrier A, Berthaud J (1985) Botanical classification of coffee. In: Coffee. Springer, pp 13–47

Civáň P, Švec M, Hauptvogel P (2011) On the coevolution of transposable elements and plant genomes. J Bot 2011:1–9. https://doi.org/10.1155/2011/893546

Clarindo W, Carvalho C (2006) A high quality chromosome preparation from cell suspension aggregates culture of *Coffea canephora*. Cytologia (Tokyo) 71:243–249. https://doi.org/10.1508/cytologia.71.243

Clarindo WR, Carvalho CR (2008) First *Coffea arabica* karyogram showing that this species is a true allotetraploid. Plant Syst Evol 274:237–241. https://doi.org/10.1007/s00606-008-0050-y

Clarindo WR, Carvalho CR (2009) Comparison of the *Coffea canephora* and *C. arabica* karyotype based on chromosomal DNA content. Plant Cell Rep 28:73–81. https://doi.org/10.1007/s00299-008-0621-y

Clarindo WR, Carvalho CR, Caixeta ET, Koehler AD (2013) Following the track of "Híbrido de Timor" origin by cytogenetic and flow cytometry approaches. Genet Resour Crop Evol 60:2253–2259. https://doi.org/10.1007/s10722-013-9990-3

Clarindo WR, Carvalho CR, Mendonça MAC (2012) Cytogenetic and flow cytometry data expand knowledge of genome evolution in three *Coffea* species. Plant Syst Evol 298:835–844. https://doi.org/10.1007/s00606-012-0595-7

Conagin C, Mendes A (1961) Pesquisas citológicas e genéticas em três espécies de *Coffea*. Auto-compatibilidade em *Coffea canephora* Pierre ex Froehner. Bragantia 20:1–74

Cros J, Combes M-C, Chabrillange N, et al (1995) Nuclear DNA content in the subgenus *Coffea* (Rubiaceae): inter-and intra-specific variation in African species. Can J Bot 73:14–20. https://doi.org/10.1139/b95-002

Cros J, Gavalda M-C, Chabrillange N, et al (1994) Variations in the total nuclear DNA content in African Coffea species (Rubiaceae). Café Cacao Thé 38:3–3

De Kochko A, Akaffou S, Andrade AC, et al (2010) Advances in *Coffea* genomics. In: Advances in botanical research. Elsevier, pp 23–63

Deakin J, Potter S, O'Neill R, et al (2019) Chromosomics: bridging the gap between genomes and chromosomes. Genes 10:627. https://doi.org/10.3390/genes10080627

Denoeud F, Carretero-Paulet L, Dereeper A, et al (2014) The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. Science 345:1181–1184. https://doi.org/10.1126/science.1255274

Fehér A, Pasternak TP, Dudits D (2003) Transition of somatic plant cells to an embryogenic state. Plant Cell Tiss Organ Cult 74:201–228. https://doi.org/10.1023/A:1024033216561

Fontes BPD (2003) Citogenética, citometria de fluxo e citometria de imagem em *Coffea* spp. Universidade Federal de Viçosa, Brazil

Fowler M (1984) Large-scale cultures of cells in suspension. In: Laboratory Procedures and their Applications. Academic Press, pp 167–174

Gonçalves M, Rodrigues M, Mexia J, Daehnhardt E (1978) Melhoramento da cafeicultura em Timor face à *Hemileia vastatrix* B. & Br. Garcia de Orta-Sér Est Agr 5:3–10

Guyot R, Darré T, Dupeyron M, et al (2016) Partial sequencing reveals the transposable element composition of *Coffea* genomes and provides evidence for distinct evolutionary stories. Mol Genet Genomics 291:1979–1990. https://doi.org/10.1007/s00438-016-1235-7

Hamon P, Grover CE, Davis AP, et al (2017) Genotyping-by-sequencing provides the first well-resolved phylogeny for coffee (*Coffea*) and insights into the evolution of caffeine content in its species. Molecular Phylogenetics and Evolution 109:351–361. https://doi.org/10.1016/j.ympev.2017.02.009

Hamon P, Siljak-Yakovlev S, Srisuwan S, et al (2009) Physical mapping of rDNA and heterochromatin in chromosomes of 16 *Coffea* species: a revised view of species differentiation. Chromosome Res 17:291–304. https://doi.org/10.1007/s10577-009-9033-2

Herrera J-C, Camayo G, De-La-Torre G, et al (2013) Identification and chromosomal distribution of copia-like retrotransposon sequences in the coffee (*Coffea* L.) genome. 31:268–278

Herrera JC, D'Hont A, Lashermes P (2007) Use of fluorescence in situ hybridization as a tool for introgression analysis and chromosome identification in coffee (*Coffea arabica* L.). Genome 50:619–626. https://doi.org/10.1139/G07-034

Höfling J, Oliveira A (1981) A serological investigation of some *Coffea* species with emphasis on the origin of *C. arabica* L. Ciência e Cultura 33:66–72

Homeyer H (1932) Zur zytologie der Rubiaceen. Planta 18:640

Iacia AAS, Pinto-Maglio CAF (2013) Mapping pachytene chromosomes of coffee using a modified protocol for fluorescence in situ hybridization. AoB PLANTS 5:plt040. https://doi.org/10.1093/aobpla/plt040

Kidwell MG (2002) Transposable elements and the evolution of genome size in eukaryotes. 115:49–63. https://doi.org/10.1023/A:1016072014259

Kim J-S, Islam-Faridi M, Klein PE, et al (2005) Comprehensive molecular cytogenetic analysis of sorghum genome architecture: distribution of euchromatin, heterochromatin, genes and recombination in comparison to rice. Genetics 171:1963–1976. https://doi.org/10.1534/genetics.105.048215

Koo D-H, Jo S-H, Bang J-W, et al (2008) Integration of cytogenetic and genetic linkage maps unveils the physical architecture of tomato chromosome 2. Genetics 179:1211–1220. https://doi.org/10.1534/genetics.108.089532

Krug C (1934) Contribuição para o estudo da cytologia do gênero *Coffea*. Instituto Agronômico de Campinas

Krug C (1937) Observações citológicas em *Coffea*. III. Instituto Agronômico de Campinas 37:1–19

Larracuente AM, Ferree PM (2015) Simple method for fluorescence DNA in situ hybridization to squashed chromosomes. J Vis Exp JoVE 95:52288. https://doi.org/10.3791/52288

Lashermes P, Andrade AC, Etienne H (2008) Genomics of coffee one of the world's largest traded commodities. In: Genomics of tropical crop plants. Springer, pp 203–226

Lashermes P, Combes M-C, Ansaldi C, et al (2011) Analysis of alien introgression in coffee tree (*Coffea arabica* L.). Mol Breed 27:223–232. https://doi.org/10.1007/s11032-010-9424-2

Lashermes P, Combes M-C, Cros J, et al (1995) Origin and genetic diversity of *Coffea arabica* L. based on DNA molecular markers. ASIC 16th:10

Lashermes P, Combes M-C, Robert J, et al (1999) Molecular characterisation and origin of the *Coffea arabica* L. genome. Mol Gen Genet MGG 261:259–266. https://doi.org/10.1007/s004380050965

Lashermes P, Combes MC, Trouslot P, Charrier A (1997) Phylogenetic relationships of coffee-tree species (*Coffea* L.) as inferred from ITS sequences of nuclear ribosomal DNA. Theor Appl Genet 94:947–955. https://doi.org/10.1007/s001220050500

Lashermes P, Paczek V, Trouslot P, et al (2000) Single-locus inheritance in the allotetraploid *Coffea arabica* L. and interspecific Hybrid *C. arabica* X *C. canephora*. Journal of Heredity 91:81–85. https://doi.org/10.1093/jhered/91.1.81

Lashermes P, Trouslot P, Anthony F, et al (1996) Genetic diversity for RAPD markers between cultivated and wild accessions of *Coffea arabica*. Euphytica 87:59–64. https://doi.org/10.1007/BF00022965

Lisch D (2013) How important are transposons for plant evolution? Nat Rev Genet 14:49–61. https://doi.org/10.1038/nrg3374

Lombello RA, Pinto-Maglio CAF (2004a) Cytogenetic studies in *Coffea* L. and *Psilanthus* Hook.f. using CMA/DAPI and FISH. Cytologia 69:85–91. https://doi.org/10.1508/cytologia.69.85

Lombello RA, Pinto-Maglio CAF (2004b) Heterochromatin and rDNA sites in *Coffea* L. chromosomes revealed by FISH and CMA/DAPI II: *C canephora* cv. Apoatã, *C. salvatrix* and *C. sessiliflora*. Caryologia 57:138–143. https://doi.org/10.1080/00087114.2004.10589383

Lombello RA, Pinto-Maglio CAF (2004c) Heterochromatin and rDNA sites in *Coffea* L. chromosomes revealed by FISH and CMA/DAPI. I: *C. humilis, C. kapakata, C.* sp. Moloundou and *C. stenophylla*. Caryologia 57:11–17. https://doi.org/10.1080/00087114.2004.10589366

Lopes C, Cunha R, Blotta L (1984) Estudo taxonômico do gênero *Coffea* com emprego de métodos numéricos. Turrialba 34:421–430

Lopes FR, Jjingo D, da Silva CRM, et al (2013) Transcriptional activity, chromosomal distribution and expression effects of transposable elements in *Coffea* genomes. PLoS ONE 8:e78931. https://doi.org/10.1371/journal.pone.0078931

Mahé L, Combes M-C, Lashermes P (2007) Comparison between a coffee single copy chromosomal region and *Arabidopsis* duplicated counterparts evidenced high level synteny between the coffee genome and the ancestral *Arabidopsis* genome. Plant Mol Biol 64:699–711. https://doi.org/10.1007/s11103-007-9191-6

Maumus F (2016) Impact and insights from ancient repetitive elements in plant genomes. Curr Opin Plant Biol 30:41–46. https://doi.org/10.1016/j.pbi.2016.01.003

Mendes A (1938) Morfologia dos cromossomos de *Coffea excelsa* Chev. Instituto Agronômico de Campinas 56:1–8

Moncada MDP, Tovar E, Montoya JC, et al (2016) A genetic linkage map of coffee (*Coffea arabica* L.) and QTL for yield, plant height, and bean size. Tree Genet Genomes 12:5. https://doi.org/10.1007/s11295-015-0927-1

Noirot M, Poncet V, Barre P, et al (2003) Genome size variations in diploid African *Coffea* species. Ann Bot 92:709–714. https://doi.org/10.1093/aob/mcg183

Nunes RC, Orozco-Arias S, Crouzillat D, et al (2018) Structure and distribution of centromeric retrotransposons at diploid and allotetraploid *Coffea* centromeric and pericentromeric regions. Front Plant Sci 9:175. https://doi.org/10.3389/fpls.2018.00175

Oliveira SC de (2017) Origin of the allotriploid “Híbrido de Timor” through a karyotype comparison with its *Coffea* ancestors. Universidade Federal do Espírito Santo, Brazil

Orozco-Arias S, Isaza G, Guyot R (2019) Retrotransposons in plant genomes: structure, identification, and classification through bioinformatics and machine learning. IJMS 20:3837. https://doi.org/10.3390/ijms20153837

Orozco-Castillo C, Chalmers KJ, PoweH W, Waugh R (1996) RAPD and organelle specific PCR re-affirms taxonomic relationships within the genus *Coffea*. Plant Cell Rep 15:337–341. https://doi.org/10.1007/BF00232367

Ortega-Ortega J, Ramírez-Ortega FA, Ruiz-Medrano R, Xoconostle-Cázares B (2019) Analysis of genome size of sixteen *Coffea arabica* cultivars using flow cytometry. HortScience 54:998–1004. https://doi.org/10.21273/HORTSCI13916-19

Owen RB, Muiruri VM, Lowenstein TK, et al (2018) Progressive aridification in East Africa over the last half million years and implications for human evolution. Proc Natl Acad Sci 115:11174–11179. https://doi.org/10.1073/pnas.1801357115

Pflug JM, Holmes VR, Burrus C, et al (2020) Measuring genome sizes using read-depth, k-mers, and flow cytometry: methodological comparisons in beetles (Coleoptera). G3 10:3047–3060. https://doi.org/10.1534/g3.120.401028

Pierozzi NI (2013) Orcein, C- and NOR-banding on mitotic chromosomes of some wild diploid coffee species (*Coffea* L.). Caryologia 66:76–83. https://doi.org/10.1080/00087114.2013.787206

Pierozzi NI, Borghi TC, Silvarolla MB (2012) A karyologicals study in some species of *Coffea* L. and in the closest relative *Psilanthus travancorensis* (Wight & Arn.) J.-F. Leroy. Not Bot Horti Agrobot Cluj-Napoca 40:39–45. https://doi.org/10.15835/nbha4028195

Pierozzi NI, Pinto-Maglio CAF, Cruz ND (1999) Characterization of somatic chromosomes of two diploid species of *Coffea* L. with acetic orcein and C-band techniques. Caryologia 52:1–8. https://doi.org/10.1080/00087114.1998.10589147

Pinto-Maglio C, Barbosa R, Cuéllar T, et al (2001) Chromosome characterization in *Coffea arabica* L. using cytomolecular techniques. In: 14th International Chromosome Conference. Würzburg, Germany, p 100

Pinto-Maglio C, Cruz N (1987) Pachytene chromosome morphology in *Coffea* L. I. nucleolar chromosomes. Caryologia 40:7–23. https://doi.org/10.1080/00087114.1987.10797806

Pinto-Maglio C, Cruz N (1998) Pachytene chromosome morphology in *Coffea* L. II. *C. arabica* L. complement. Caryologia 51:19–35. https://doi.org/10.1080/00087114.1998.10589117

Pinto-Maglio C, Cuéllar T, Barbosa R (2000) Aplicação de técnicas de citogenética molecular na caracterização dos cromossomos da espécie *Coffea arabica* L. In: Anais do I Simpósio da Pesquisa dos Cafés do Brasil. Poço das Caldas, Brazil, pp 444–446

Praça-Fontes MM, Carvalho CR, Clarindo WR, Cruz CD (2011) Revisiting the DNA C-values of the genome size-standards used in plant flow cytometry to choose the “best primary standards.” Plant Cell Rep 30:1183–1191. https://doi.org/10.1007/s00299-011-1026-x

Prakash N, Marques D, Varzea V, et al (2004) Introgression molecular analysis of a leaf rust resistance gene from *Coffea liberica* into *C. arabica* L. Theor Appl Genet 109:1311–1317. https://doi.org/10.1007/s00122-004-1748-z

Raharimalala N, Rombauts S, McCarthy A, et al (2021) The absence of the caffeine synthase gene is involved in the naturally decaffeinated status of *Coffea humblotiana*, a wild species from Comoro archipelago. Sci Rep 11:8119. https://doi.org/10.1038/s41598-021-87419-0

Raina SN, Mukai Y, Yamamoto M (1998) In situ hybridization identifies the diploid progenitor species of *Coffea arabica* (Rubiaceae). Theor Appl Genet 97:1204–1209. https://doi.org/10.1007/s001220051011

Razafinarivo NJ, Rakotomalala J-J, Brown SC, et al (2012) Geographical gradients in the genome size variation of wild coffee trees (*Coffea*) native to Africa and Indian Ocean islands. Tree Genet Genomes 8:1345–1358. https://doi.org/10.1007/s11295-012-0520-9

Rijo L (1974) Observações cariológicas no cafeeiro "Híbrido de Timor." Port Acta Biol 8:157–168

Ruas PM, Ruas CF, Rampim L, et al (2003) Genetic relationship in *Coffea* species and parentage determination of interspecific hybrids using ISSR (Inter- Simple Sequence Repeat) markers. Genet Mol Biol 26:319–327. https://doi.org/10.1590/S1415-47572003000300017

Sanglard NA, Amaral-Silva PM, Sattler MC, et al (2019) Indirect somatic embryogenesis in Coffea with different ploidy levels: a revisiting and updating study. Plant Cell Tissue Organ Cult PCTOC 136:255–267. https://doi.org/10.1007/s11240-018-1511-9

Scalabrin S, Toniutti L, Di Gaspero G, et al (2020) A single polyploidization event at the origin of the tetraploid genome of *Coffea arabica* is responsible for the extremely low genetic variation in wild and cultivated germplasm. Sci Rep 10:1–13. https://doi.org/10.1038/s41598-020-61216-7

Schatz MC, Witkowski J, McCombie WR (2012) Current challenges in *de novo* plant genome sequencing and assembly. Genome Biol 13:1–7. https://doi.org/10.1186/gb-2012-13-4-243

Serrato-Capuchina A, Matute D (2018) The Role of Transposable Elements in Speciation. Genes 9:254. https://doi.org/10.3390/genes9050254

Setotaw TA, Caixeta ET, Pereira AA, et al (2013) Coefficient of parentage in *Coffea arabica* L. cultivars grown in Brazil. Crop Sci 53:1237–1247. https://doi.org/10.2135/cropsci2012.09.0541

Setotaw TA, Caixeta ET, Zambolim EM, et al (2020) Genome introgression of Híbrido de Timor and its potential to develop high cup quality *C. arabica* cultivars. J Agric Sci 12:64–76. https://doi.org/10.5539/jas.v12n4p64

Singh R (2016) Plant cytogenetics. CRC press

Soltis DE, Soltis PS, Pires JC, et al (2004) Recent and recurrent polyploidy in *Tragopogon* (Asteraceae): cytogenetic, genomic and genetic comparisons. Biol J Linn Soc 82:485–501. https://doi.org/10.1111/j.1095-8312.2004.00335.x

Sousa TV, Caixeta ET, Alkimim ER, et al (2017) Population structure and genetic diversity of coffee progenies derived from Catuaí and Híbrido de Timor revealed

by genome-wide SNP marker. Tree Genet Genomes 13:1–16. https://doi.org/10.1007/s11295-017-1208-y

Stebbins GL (1949) Variation and evolution in plants. Columbia University Press, New York, NY

Sybenga J (1960) Genetics and cytology of coffee: a literature review. 10 83–137

Talukdar D, Sinjushin A (2015) Cytogenomics and mutagenomics in plant functional biology and breeding. In: PlantOmics: the omics of plant science. Springer, New Delhi, pp 113–156

Tosh J, Dessein S, Buerki S, et al (2013) Evolutionary history of the Afro-Madagascan Ixora species (Rubiaceae): species diversification and distribution of key morphological traits inferred from dated molecular phylogenetic trees. Ann Bot 112:1723–1742. https://doi.org/10.1093/aob/mct222

Treangen TJ, Salzberg SL (2012) Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet 13:36–46. https://doi.org/10.1038/nrg3117

Urbanska KM, Hurka H, Landolt E, et al (1997) Hybridization and evolution in *Cardamine* (Brassicaceae) at Urnerboden, Central Switzerland: biosystematic and molecular evidence. Plant Syst Evol 204:233–256. https://doi.org/10.1007/BF00989208

Vallejo-Marín M (2012) *Mimulus peregrinus* (Phrymaceae): a new British allopolyploid species. PhytoKeys 14:1–14. https://doi.org/10.3897/phytokeys.14.3305

van Boxtel J, Berthouly M (1996) High frequency somatic embryogenesis from coffee leaves: factors influencing embryogenesis, and subsequent proliferation and regeneration in liquid medium. Plant Cell Tissue Organ Cult 44:7–17. https://doi.org/10.1007/BF00045907

Vicient CM, Casacuberta JM (2017) Impact of transposable elements on polyploid plant genomes. Ann Bot 120:195–207. https://doi.org/10.1093/aob/mcx078

von Faber F C (1912) Morphologisch-physiologische untersuchungen an blüten von *Coffea*-Arten. Ann Jard Buitenzorg 25:59–160

Vossen HA van der (1985) Coffee selection and breeding. In: Coffee. Springer, Boston, pp 48–96

Wicker T, Gundlach H, Spannagl M, et al (2018) Impact of transposable elements on genome structure and evolution in bread wheat. Genome Biol 19:103. https://doi.org/10.1186/s13059-018-1479-0

Wu F, Mueller LA, Crouzillat D, et al (2006) Combining bioinformatics and phylogenetics to identify large sets of single-copy orthologous genes (COSII) for comparative, evolutionary and systematic studies: a test case in the Euasterid

plant clade. Genetics 174:1407–1420. https://doi.org/10.1534/genetics.106.062455

Yotoko KSC, Dornelas MC, Togni PD, et al (2011) Does variation in genome sizes reflect adaptive or neutral processes? New clues from *Passiflora*. PLoS ONE 6:e18212. https://doi.org/10.1371/journal.pone.0018212

Yu Q, Guyot R, de Kochko A, et al (2011) Micro-collinearity and genome evolution in the vicinity of an ethylene receptor gene of cultivated diploid and allotetraploid coffee species (*Coffea*): recent speciation event of *Coffea arabica*. Plant J 67:305–317. https://doi.org/10.1111/j.1365-313X.2011.04590.x

Yuyama PM, Pereira LFP, dos Santos TB, et al (2012) FISH using a *gag* -like fragment probe reveals a common *Ty* 3- *gypsy* -like retrotransposon in genome of *Coffea* species. Genome 55:825–833. https://doi.org/10.1139/gen-2012-0081

10. Supplementary information

Box 1: A breakthrough - using plant tissue culture material as source of mitotic cells in *Coffea*

Plant cytogenetics is performed mainly from root meristems as the source of metaphase cells. However, the low germination and low metaphase index observed in *Coffea* species (Conagin and Mendes 1961) are common bottlenecks, which might be circumvented by using plant tissue culture techniques. CAS have been used as an alternative source of mitotic cells for different *Coffea* species, such as *C. canephora*, *C. congensis*, *C. arabica* (Clarindo and Carvalho 2006, 2009), *C. eugenioides* (Sanglard et al. 2019) and HT 'CIFC 4106' (Clarindo et al. 2013). CAS maintained in liquid medium display a high frequency of cell division, providing a suitable index of prometaphases and metaphases (Fowler 1984; Clarindo and Carvalho 2006). The high rate of cell division is triggered by the in vitro conditions, mainly the use of the synthetic auxin 2,4-dichlorophenoxyacetic acid (2,4-D). This growth regulator increases the level of endogenous auxins, promoting dedifferentiation, activation of cell division and proliferation (Fehér et al. 2003).

Figure S1 illustrates a workflow to establish *Coffea* CAS cultures for obtaining prometaphasic/metaphasic chromosomes for morphometric analysis, image cytometry and FISH, as well as nuclei suspensions for flow cytometry. The procedures represented in this workflow were based on van Boxtel and Berthouly (1996), Praça-Fontes et al. (2011), Clarindo and Carvalho (2008, 2009), Clarindo et al. (2012) and Sanglard et al. (2019). For induction of the indirect somatic embryogenesis pathway, leaf fragments (~1 cm^2) of a given *Coffea* species are inoculated in Petri dishes containing semisolid callus induction medium, with the abaxial surface facing upwards (Fig. S1a). This medium is usually supplemented with a combination of 2,4-D and 6-benzylaminopurine. Calli with a friable aspect and pale-yellow color are visible within ~3 months (Fig. S1b). The CAS are obtained by transferring these friable calli to Erlenmeyer flasks containing liquid callus induction medium (Fig. S1c). These cultures should remain in an orbital shaker at 100 rpm in the dark (Fig. S1c[1]). After around 2 months, the CAS can be sub cultivated and then exposed to antitubulin agents (Fig.

S1c²) and enzymatic maceration of the cell wall (Fig. S1c³) to obtain metaphase spreads.

For somatic embryos recovery, the friable embryogenic calli are transferred to conversion/maturation medium lacking the auxin 2,4-D. About 180 days later (with monthly subcultures), the calli exhibit somatic embryos in different developmental stages (Fig. S1d). The mature cotyledonary somatic embryos are transferred to germination medium for seedlings recovery (Fig. S1e). Leaf fragments from the regenerated plantlets might be used for nuclei isolation in order to produce suspensions needed for flow cytometry (Fig. S1f – g). These same plantlets also produce roots, which can be submitted to cytogenetic treatments (Fig. S1h – h^2). From the CAS (Fi. S1c – c^3) or root meristems (Fig. S1h – h^2), slides are prepared by the cellular dissociation and air-drying techniques (Fig. 1i), providing prometaphasic and/or metaphasic chromosomes suitable for chromosome number determination, morphometric characterization, image cytometry (Fig. S1j) and FISH (Fig. S1k).

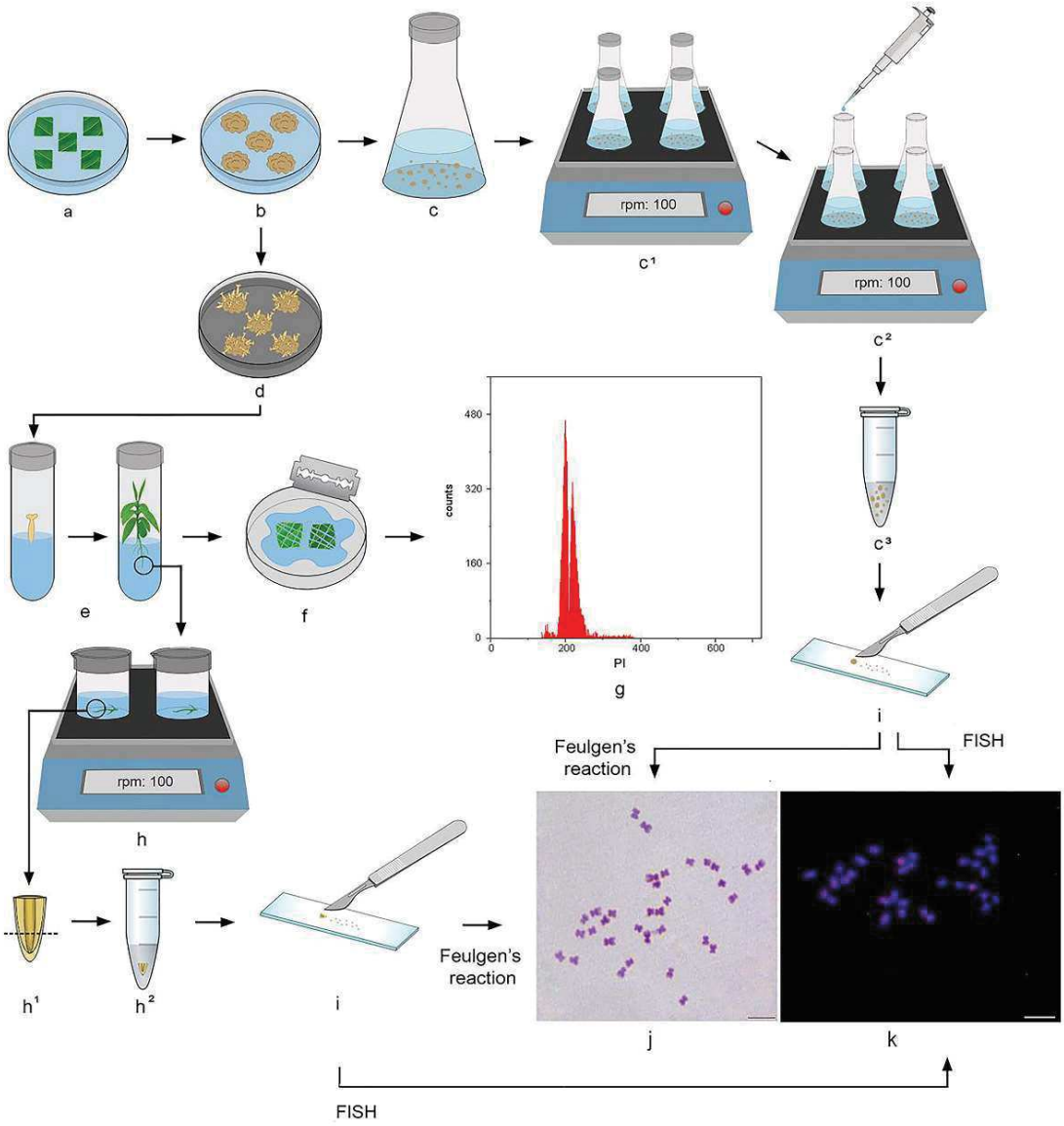

Figure S1. Schematic representation of a workflow to establish *Coffea* CAS cultures for obtaining prometaphasic/metaphasic chromosomes suitable for multiple cytogenomic analysis. More details are in the text presented in Box 1. Source: Oliveira (2017).

Box 2: Flow and image cytometries

Flow and image cytometries have been performed in *Coffea* since the early 90's for measuring the nuclear (Cros et al. 1994, 1995; Noirot et al. 2003; Clarindo and Carvalho 2009; Hamon et al. 2009; Clarindo et al. 2012; Razafinarivo et al. 2012) and chromosomal DNA content (Clarindo and Carvalho 2009). In contrast with the uniformity of chromosome number and morphology, *Coffea* species exhibit considerable genome size variation, ranging from 1C = 0.475 pg (2C = 0.95) for *C. racemosa* Lour. (Cros et al. 1995) to 1C = 0.92 pg (2C = 1.84) for *C. humilis* A. Chev. (Razafinarivo et al. 2012). Such variability in genome size has been useful for identification of *Coffea* species (Noirot et al. 2003; Clarindo and Carvalho 2009; Razafinarivo et al. 2012), varieties (Fontes 2003; Ortega-Ortega et al. 2019) and hybrids (Clarindo et al. 2013).

For genome sequencing programs, knowing the genome size is fundamental for calculating the sequencing coverage with maximum accuracy. For this, flow cytometry has been considered the 'gold standard' technique (Pflug et al. 2020). Information regarding the genome size is also relevant for evolutionary studies, as in some species, variations in nuclear DNA content is correlated to adaptive traits (Cros et al. 1995; Yotoko et al. 2011). In *Coffea*, no significant correlations were yet found between genome size and adaptive traits. Nonetheless, global trends for the genus are observed regarding the geographic distribution, genome size and leaf shape. The species with the smaller genomes and smaller/thicker leaves grow mainly in dry areas of east Africa or northern Madagascar, while those with larger genomes and wider/thinner leaves occupy the humid forests of west and central Africa (Razafinarivo et al. 2012).

Image cytometry studies in *Coffea* were performed mainly with focus on the cultivated species *C. canephora* and *C. arabica*. For the *C. canephora* cultivar 'Conilon', the estimated nuclear DNA content was 2C = 1.57 pg, while for *C. arabica*, the values were 2C = 2.62 pg in the cultivar 'Catuaí Vermelho' and 2C = 1.89 pg in 'Mundo Novo' (Fontes 2003). Through image cytometry, Clarindo and Carvalho (2009) estimated the chromosomal DNA content of the same two *Coffea* species and

identified key similarities among their karyotypes, providing relevant insights on the evolution of *C. arabica*, which are discussed in the main text.

CHAPTER 2 – RESEARCH PAPER

A CYTOGENOMIC VIEW OF THE DIPLOID *Coffea eugenioides* REPEATOME LANDSCAPE

Mariana Cansian Sattler[1]✉, Jessica Coutinho Silva[1], Stéfanie Cristina de Oliveira[2], Wellington Ronildo Clarindo[1]

[1]Laboratório de Citogenética e Citometria, Departamento de Biologia Geral, Universidade Federal de Viçosa. ZIP 36.570-900 Viçosa – MG, Brazil.

[2]Laboratório de Citogenética e Cultura de Tecidos Vegetais, Campus de Alegre, Universidade Federal do Espírito Santo. ZIP 29.500-000 Alegre – ES, Brazil.

✉Corresponding author e-mail: mariana.sattler@ufv.br

Abstract

Repetitive sequences comprise a significant amount of Angiosperm genomes and are recognized for their roles in genome organization and function. Among these sequences, mobile elements are notable due to their intrinsic mutagenic potential, which is related to the origin of adaptive novelties. Therefore, studies regarding the genomic organization and dynamics of repeated sequences are fundamental to enlighten their role in plant evolution. Here, we provide a cytogenomic characterization of the repeat content of *Coffea eugenioides*, one of the diploid progenitors of the allotetraploid *Coffea arabica* alongside with *C. canephora*. A graph-based sequence clustering algorithm was used to identify and quantify the repeatome, which comprised ~47% of the genome. Mobile elements represented 45% of the genome and, as observed for other *Coffea*, the proportion of Ty3/Gypsy elements (32%) was considerably higher compared to Ty1/Copia (4%). We also unprecedently mapped the distribution of three Ty1/Copia (Bianca, TAR and Tork) and one Ty3/Gypsy (Athila) LTR-retrotransposon in metaphase chromosomes and interphase nuclei of *C. eugenioides*. The four elements were mainly distributed in a clustered pattern throughout different portions of the chromosomes. Athila was the most abundant and co-located with the secondary constriction of chromosome 3 in some cells. The obstacles related to obtaining high-quality chromosomes from *Coffea* species have long been hampering the association of in silico analysis with the in situ mapping. Thus, we hope that the results presented here enlighten not only the composition, but also the distribution pattern of mobile elements in the genome of *C. eugenioides*, providing background for further cytogenomic investigations regarding *Coffea* repeatome.

Keywords: Coffee, Molecular Cytogenetics, Mobile Elements, LTR-Retrotransposons, RepeatExplorer2.

1. Introduction

Plant genomes contain considerably variable amounts of repetitive DNA sequences. The complete complement of repeated sequences in a genome is now acknowledged as the repeatome. In plants, the three main components of the repeatome are the rDNA genes (5S and 18S-5.8S-26S for plants), micro-, mini- and satellite DNAs, and mobile elements (Treangen and Salzberg 2012). Among them, the mobile elements are noteworthy for their intrinsic ability to insert themselves at different portions of the genome. In addition, mobile elements have a main role in eukaryotes as contributors to mutation, genome size variation, genome evolution and phenotype diversity (Kidwell 2002; Orozco-Arias et al. 2019). As the process of transposition itself is potentially mutagenic and deleterious, mobile element expression is highly controlled in plant genomes. Nevertheless, several conditions might activate their expression and result in new genetic and/or phenotypic diversity (Lisch 2013). Nowadays, a significant portion of adaptive novelties in Angiosperms are recognized as consequences of the activity of transposable elements (Serrato-Capuchina and Matute 2018). Therefore, studies regarding the genomic dynamics of transposable elements are fundamental to enlighten their role in plant genome organization, diversity and evolution (Yuyama et al. 2012; Neumann et al. 2019).

According to the mechanism of transposition, mobile elements are divided in two classes: (1) the retrotransposons (Class I) that move using an RNA intermediate (“copy-and-paste”), and (2) the DNA transposons (Class II) that transpose via DNA excision (“cut-and-paste”). Class I elements are divided based on the structure and transposition mechanism (Wicker et al. 2007). The long terminal repeats (LTR) retrotransposons are the most abundant mobile elements in plant genomes. These retrotransposons are characterized by the presence of two LTRs directly orientated in the 5’ and 3’ extremities (Wicker et al. 2007; Neumann et al. 2019). For the elements capable of self-replication (autonomous elements), there are at least two genes present between the LTRs: *gag* and *pol*. *Gag* gene encodes a capsid-like protein, and *pol* gene encodes a polyprotein including a reverse transcriptase (RT), a ribonuclease H (RH) and an integrase (INT), which provide the enzymatic machinery for reverse transcription and integration into the host DNA sequence (Feschotte et al. 2002). The

two plant LTR retrotransposons superfamilies are classified based on the order of *pol* gene domains, being INT-RT-RH for Ty1/Copia superfamily and RT-RH-INT for the Ty3/Gypsy superfamily (Wicker et al. 2007; Neumann et al. 2019). Both superfamilies are split into families considering the similarities of the coding regions, especially the RT domains (Neumann et al. 2019). As LTR-retrotransposons exhibit independent activity in chromosomes and different fates in genomes, their occurrence, amount and/or chromosome distribution are considerably variable. This variation occurs between species and also between lineages or even different tissues of the same individual (Gabriel et al. 2006; Yuyama et al. 2012; Neumann et al. 2019; Assis et al. 2020).

Coffea L. genus (Rubiaceae) is an interesting plant taxon to study repetitive sequences, since almost all species are diploids (2n = 2x = 22 chromosomes), except the allotetraploid *C. arabica* L. (2n = 4x = 44, *Coffea canephora* Pierre ex A. Froehner x *Coffea eugenioides* S. Moore), with homomorphic karyotypes (only metacentric and submetacentric chromosomes). Nevertheless, there is a substantial variation on the nuclear C-value DNA amounts. The interspecific nuclear genome size variation between diploids ranges from 1C = 0.40 (*Coffea sessiliflora* Bridson, Cros et al. 1994) to 1C = 0.88 pg (*Coffea humilis* A. Chev., Noirot et al. 2003). Therefore, the evolution in *Coffea* genomes most likely involved small mutations and rearrangements, with transposable elements being hypothesized as one of the main contributors to genome size variation and diversification (Raharimalala et al. 2021).

Studies regarding the characterization of *Coffea* repeatome substantially advanced in the last 20 years. The *Coffea* genomes analyzed so far (11 species, Denoeud et al. 2014; Guyot et al. 2016; Raharimalala et al. 2021) exhibited a variation in LTR-retrotransposon proportions ranging from 32% (*Coffea humblotiana* Baill., 1C = 0.49 pg, Razafinarivo et al. 2012) to 53% (*Coffea heterocalyx* Stoff., 1C = 0.87 pg, Noirot et al. 2003). For the allotetraploid *C. arabica*, the LTR-retrotransposons represent ~50% of the genome (Guyot et al. 2016). *C. canephora* presented a smaller proportion (42%) with a prevalence of Ty3/Gypsy (24.1%) over Ty1/Copia (6.8%) superfamily (Denoeud et al. 2014; Guyot et al. 2016).

Two Ty1/Copia elements named *Divo* and *Nana* were recently discovered and annotated in the *C. canephora* genome. The minimum estimated copy number in the

genome was 45 for *Divo* and 22 for *Nana*. Phylogenetic markers based on these two LTR-retrotransposons also revealed to be considerably useful to comprehend the genetic diversity and speciation/differentiation events of 31 species within the *Coffea* genus. While *Nana* differentiated most species, *Divo* was more informative at the intraspecific level, possibly because this LTR-retrotransposon is a more recently active family (Hamon et al. 2011). A closer analysis of *Divo* evidenced that this element belongs to the Bianca family. *Divo* is expressed in *C. canephora* plants and appears to be distributed in gene rich and gene poor regions. In addition, molecular estimations showed that the *Divo* underwent more recent activity in this species and in *C. arabica* when compared to *C. eugenioides*. The number of estimated copies was also variable among the three species: 119, 204 and 132 for *C. canephora*, *C. arabica* and *C. eugenioides*, respectively (Dupeyron et al. 2017).

The cytogenomics of mobile elements is still incipient in *Coffea*, especially due to barriers associated with obtaining good quality chromosomes for these species (Clarindo and Carvalho 2009; Yuyama et al. 2012). Until now, the fluorescent in situ hybridization (FISH) studies with mobile elements in *Coffea* provided mostly a general overview of their genomic distribution, without the assembly of karyograms and, consequently, without the mapping of these sequences into chromosomes. Nonetheless, advances have been made in the last decades to detect these sequences in *Coffea* genomes through FISH, including a Ty3/Gypsy element similar to the Tekay (*Del1*) family (Yuyama et al. 2012) and a group of Centromeric Retrotransposons (CRs) named CRC in *Coffea* (Nunes et al. 2018). Although Ty1/Copia and Ty3/Gypsy elements might be found distributed in blocks/clusters or dispersed along plant chromosomes, some families often exhibit specific patterns, such as the CRCs that are always associated with the centromeric or pericentromeric regions and have a role chromosome kinetics (Nunes et al. 2018; Assis et al. 2020).

Here, we aimed to uncover the repetitive DNA landscape of the genome of *Coffea eugenioides* S. Moore, one of the diploid progenitors of the commercial species *C. arabica* alongside with *C. canephora*, in order to address some questions, as: 1) what is the composition of *C. eugenioides* repeatome and how it compares to other *Coffea* species? 2) what is the classification of *Nana* element? 3) where are the 45S rDNA sites located in *C. eugenioides* karyotype? 4) are Bianca (*Divo*) and *Nana* detectable and mapped using FISH considering their low estimated number of copies? 5) how

different types of LTR-retrotransposons are distributed in the *C. eugenioides* chromosomes? and 6) is this distribution variable among cells or homologous chromosomes? For this, we used the RepeatExplorer2 pipeline, a graph-based sequence clustering algorithm (Novák et al. 2013), to perform the *de novo* repeat identification in *C. eugenioides* genome from publicly available next-generation sequencing data, with a closer look at the mobile elements. We also mapped the distribution of four LTR-retrotransposons using FISH in interphase nuclei and chromosomes of this diploid *Coffea*. The mapped elements included the two already identified *Coffea* lineages Bianca (*Divo*) and *Nana*, as well as two other LTR-retrotransposons identified in this work, Tork and Athila.

2. Material and Methods

2.1. Plant material

C. eugenioides plants maintained at the germplasm bank of Universidade Federal de Viçosa (Minas Gerais, Brazil) were used as biological material. The leaves were used for genomic DNA extraction, friable callus establishment by tissue culture, and nuclei isolation for cytogenetic preparations.

2.2. Repeat content analysis of C. eugenioides genome and identification of transposable element families

Paired-end Illumina reads from the *C. eugenioides* genome were obtained from the Sequence Read Archive (SRA) of the National Center for Biotechnology Information (NCBI) database under the accession number SRX7201323. Galaxy server of RepeatExplorer2 (https://repeatexplorer-elixir.cerit-sc.cz/galaxy; Novák et al. 2013) was used to filter and analyze the repeat content from *C. eugenioides* genome. RepeatExplorer2 was chosen among other bioinformatic tools in this work due to its

practicality, accessibility (available in open source), and ability to annotate repetitive elements from raw NGS data, without the need of an assembled genome. Initially, all FASTQ sequence reads were filtered by quality with default parameters (90% of bases equal to or above the quality cut-off value of 20). Then, we employed the graph-based sequence clustering tool RepeatExplorer2 pipeline on the filtered reads using default parameters, except for the queue, which was selected as long (64 GB RAM). Viridiplantae repeat protein database v3.0 was used as the reference database of transposable element protein domains (Neumann et al. 2019). The manual correction of the automated repeat annotation was performed as suggested by Novák et al. (2020).

2.3. Family classification of the *Divo* and *Nana* LTR-retrotransposons

Two *C. canephora* Ty1-*copia* LTR-retrotransposons previously annotated and named by Hamon et al. (2011) as *Divo* and *Nana* (accession numbers HM755952 and HM755953 at NCBI, respectively) were classified in the present work based on the RT domain sequences according to the system proposed by Neumann et al. (2019). For this, the RT domains were isolated from the full-length DNA sequences of both LTR-retrotransposons and used as inputs for BLASTx (Altschul et al. 1997) searches against the REDBx database provided by Neuman et al. (2019). The classification at the family level followed the 80-80-80 rule proposed by Wicker et al. (2007).

2.4. Mapping of rDNAs and LTR-retrotransposons in *C. eugenioides*

2.4.1. Probe construction

Probes for mapping the 18S rDNA and four different LR-retrotransposons in the diploid genome of *C. eugenioides* were constructed through PCR. The primers for the 18S rDNA were obtained from Unfried et al. (1989). For the *Divo* and *Nana* Ty1-*copia*

LTR-retrotransposons, primers were constructed based on the RT nucleotide sequences isolated as described in the previous section.

The RT domain sequences for constructing the primers for Tork and Athilla mapping, on the other hand, were constructed through a homology analysis using the Domain based Annotation of Transposable Elements (DANTE) tool from the RepeatExplorer2 Galaxy server, which was applied on the whole genome shotgun sequence of *C. eugenioides* (GCA_003713205.1). The RT sequences from DANTE output were filtered for a minimum alignment length of 0.8 and minimum similarity of 0.8. The nucleotide sequences from the output were extracted, manually selected and a consensus sequence was used for BLAST searches against the *C. eugenioides* genome. Primers were constructed aiming for the most conserved regions of the RT domain nucleotide sequences from the four LTR-retrotransposons. This domain is frequently used for cytogenomic analyses with LTR-retrotransposons sequences for being suggested as the most conserved (Neuman et al. 2019). The tool PrimerQuest™ from Integrated DNA Technologies was used to construct the primers (Table 1). The settings were maintained as default, except for the amplicon size interval, which was limited to 400 - 700 bp. The primers were validated *in silico* using OligoAnalyzer™ Tool from Integrated DNA Technologies, and also by performing multiple BLAST searches in the genome of *C. eugenioides.*

Table 1. Primers designed for DNA sequence amplification and probe construction for rDNA and RT domain mapping of four *C. eugenioides* LTR-retrotransposons.

Sequence name	Primers forward and reverse	Tm (°C)
18S rDNA	F: 5'- CTGCCAGTAGTCATATGC -3' R: 5'- ATGGATCCTCGTTAAGGG -3'	60
Ty1/Copia *Divo*	F: 5'- GGCTCAAGTAATCTGGACGTATG -3' R: 5'- CGGAACCGCTTTAATACCTTCT -3'	55
Ty1/Copia *Nana*	F: 5'- ACATTTGCACCAGTAACAAGAA -3' 5'- TACCATCAGCTCTCCTTATGATATT -3'	55
Ty1/Copia Tork	F: 5'- TGGATCTGGAACTTGTCCAATTA -3' R: 5'- CCCAATCTTCTATCTCTACTTATCTCC -3'	45
Ty3/Gypsy Athila	F: 5'- GTACAGGTGGTCCCGAAGAA -3' R: 5'- TATCCCGTGCTCGACCATAAA -3'	45

Probes were constructed using the PCR. For this, genomic DNA from *C. eugenioides* leaves was extracted using the GenElute Plant Genomic DNA Miniprep (Sigma). The LTR-retrotransposons and rDNA genes were amplified according to respective annealing temperature for each F and R primers (Table 1) from 50 µL of the following reaction mixture: 100 ng genomic DNA, 0.5 µM of each primer (Table 1), 200 µM of each dNTP (Promega), 2.5 U AccuTaq LA DNA Polymerase (Sigma) and 1X polymerase buffer (Sigma). PCR conditions were: initial denaturation at 95°C for 5 min; 25 cycles of denaturation at 95°C for 1 min; annealing at respective primer temperature (Table1) for 1 min; extension at 68°C for 1 min; and final extension at 68°C for 2 min. The amplification products obtained from the amplification of LTR-retrotransposon RT domains were analyzed by electrophoresis in 1.5% agarose gel stained with GelRed (Supplementary Figure 2). For labelling, the conditions were the same mentioned before, but the compounds in PCR reaction were the following: 100 ng of the respective amplified DNA, 0.5 µM of each primer, 200 µM each of dATP, dCTP and dGTP, 150 µM of dTTP, 50 µM of ChromaTide Alexa Fluor 488-5-dUTP for *Divo*, Tork and 18S rDNA or 50 µM of Tetramethylrhodamine 5-dUTP for *Nana* and Athila, 2.5 U of Platinum™ *Tfi* Exo(-) DNA Polymerase (Invitrogen) and 1X of the

enzyme reaction buffer (Invitrogen). The amplification products were also analyzed by electrophoresis in 1.5% agarose gel (adapted from Silva et al. 2020; Soares et al. 2020).

2.4.2. Metaphase chromosomes and interphase nuclei obtention from cell aggregate suspensions

Cell aggregate suspensions were obtained according to Clarindo and Carvalho (2009). The suspensions were maintained for at least four months in liquid media, with subcultures every 15 days. These suspensions were treated with 4 µM amiprophos-methyl at 30°C for 8 h under continuous agitation at 110 rpm. The cell suspension aggregates were washed in dH_2O, fixed in 3:1 methanol: acetic acid solution and stored at -20°C for at least 24 h. Posteriorly, they were washed, macerated for 2 h at 36°C in enzymatic pool (4% cellulase Sigma, 0.4% hemicellulase Sigma, 1% macerozyme Onozuka R10 Yakult, 100% pectinase Sigma) diluted in dH_2O in the proportion of 1:20 (enzyme pool: dH_2O), washed in dH_2O, fixed in 3:1 methanol: acetic acid solution and stored at -20°C. Slides were prepared by cellular dissociation and air-drying techniques on the macerated cell aggregate suspensions. The slides were selected using a contrast microscope for mapping based on the number and quality of metaphases and interphase nuclei.

2.4.3. Interphase nuclei isolation from leaves

Nuclear suspensions were obtained from *C. eugenioides* young leaf fragments (2 cm^2) by chopping (Galbraith et al. 1983) in 0.4 mL of OTTO-I (Otto 1990) supplemented with 2.0 mM dithiothreitol (Sigma®). The suspensions were filtered through a 30-µm nylon mesh (Partec®) and centrifuged at 100×*g* for 5 min. The supernatant was discarded and the pellet resuspended in 0.1 mL OTTO-I buffer for 10 min. The suspensions were fixed in methanol:acetic acid solution (3:1, Merck®) for 10 min at -20°C, centrifuged at 100×*g* for 5 min three times, and the slides prepared by the dropping technique (Abreu et al. 2014).

2.4.4. Fluorescent in situ hybridization and image analysis

Selected slides containing prometaphases/metaphases and interphase nuclei derived from cell aggregate suspensions or interphase nuclei from *C. eugenioides* leaves were aged for at least three days at 37°C. The slides were washed in 1X phosphate buffered saline (PBS) for 5 min, fixed in 4% formalin for 15 min, washed again in 1X PBS for 5 min, and submitted to an ice-cold ethanol series (70, 85 and 100%) for 5 min each. For denaturation, the slides were dried at room temperature and then submersed in a denaturation mixture containing 70% formamide and 2X saline-sodium citrate (SSC) for 2 min at 68°C. Immediately after denaturation, the slides were again submitted to the ice-cold ethanol series and dried at room temperature. The probe was denatured for 5 min at 85°C in 35 µL of a hybridization mixture containing 50% formamide, 2X SSC, 10% dextran sulfate, 1 µg competitor DNA (Human Cot-1 DNA, Invitrogen®) and 200 ng of the labeled probes. While the rDNAs were hybridized separately, the probes for the LTR-retrotransposons were hybridized in pairs, *Divo* with *Nana* and *Athila* with *Tork.* After denaturation, the mixture was immediately placed on ice for at least 5 min. For the in situ hybridization, the mixture was placed over the slides, covered with plastic coverslips (22 x 40 mm) and sealed with rubber cement. The slides were then incubated at 37°C for 24 h in Thermobrite. After hybridization, the coverslips were carefully removed and the slides submitted to the stringency washes in 2X SSC for 20 min at 45°C. Lastly, the slides were mounted in 40% glycerol in 1X PBS with 5 µM DAPI for chromosome counterstain, covered with 24 x 40 mm glass coverslips and sealed with nail polish. The images were captured with a digital video camera 12-bit CCD (Olympus) coupled to a photomicroscope Olympus BX-60 equipped with epifluorescence and immersion objective of 100X. Captured images were processed using Image ProPlus 6.1 software (Media Cybernetics) (adapted from Silva et al. 2020; Soares et al. 2020).

3. Results

3.1. The composition of the *C. eugenioides* repeatome

RepeatExplorer2 pipeline allowed us to analyze 5,026,946 reads from a total number of 275,249,010 input reads from *C. eugenioides* genome. From these 5,026,946 analyzed reads, 57% were included in top clusters (Supplementary Figure 1). The proportions of each repeat family are shown in Table 2. The satellite DNA sequences and the both rDNAs represented ~2% of the genome. LTR-retrotransposons comprised 40% of *C. eugenioides* genome. The proportions of Ty1/Copia and Ty3/Gypsy were 4% and 32%, respectively. SIRE was the most abundant (3%) among the Ty1/Copia elements, while Tekay (22%) was the most frequent Ty3/Gypsy family. Five families of Class II mobile elements were also identified and together comprised 3% of the genome (Table 2).

Divo and *Nana*, the two previously annotated Ty1/Copia LTR-retrotransposons of *C. canephora* (Hamon et al. 2011), were classified as belonging to the Bianca and TAR families, respectively, based on BLASTx homology searches using the classification system proposed by Neumann et al. (2019). The proportion of both LTR-retrotransposons on the *C. eugenioides* genome was the same, 0.20%. The *Divo* element had already been classified in a previous work as Bianca (Dupeyron et al. 2017), but the classification of *Nana* was unknown until now. The LTR-retrotransposon *Tork*, also from the Ty1/Copia superfamily, represented 0.5% of the genome, while the Ty3/Gypsy *Athila* element accounted for 5% of *C. eugenioides* genome.

Table 2. Final annotation with genome proportions of different repeat sequences in *C eugenioides*. Total repeats refer to the proportion of reads clustered by RepeatExplorer2. The percentages of different repeat classes, superfamilies and families are calculated on annotated clusters with a genome proportion higher than 0.01%.

Type of repeated sequence	Class	Superfamily	Family	Proportion (%)
Total Repeats				47.00
Satellites				1.00
rDNA 45S				0.80
rDNA 5S				0.40
Mobile elements				45.00
	Class I (LTR)			40.00
		Ty1/Copia		4.00
			SIRE	3.00
			Tork	**0.50**
			Ale	0.30
			Ikeros	0.30
			TAR (*Nana*)	**0.20**
			Bianca (*Divo*)	**0.20**
			Angela	0.10
			Ivana	0.10
		Ty3/Gypsy		32.00
			Chromovirus/Tekay	22.00
			Chromovirus/CRM	3.00
			Chromovirus/Reina	0.20
			Chromovirus/ Galadriel	0.03
			Non-chromovirus/ Athila	**5.00**
			Non-chromovirus/ Tat/Ogre	5.00
			Non-chromovirus/ Tat/Retand	0.40
	Class II			3.00
			TIR/hAT	1.00
			TIR/MuDR_Mutator	0.70
			TIR/EnSpm_CACTA	0.50
			TIR/PIF_Harbinger	0.04
			Helitron	0.09
	Unclassified			2.00

3.2. Mapping of rDNAs and LTR-retrotransposons *Divo*, *Nana*, Tork and Athila in *C. eugenioides*

The FISH procedures provided clear fluorescence signals with little or no background, allowing the mapping of target sequences in prometaphase/metaphase chromosomes and interphase nuclei of *C. eugenioides*. The 18S rDNA fluorescence signals were mapped in the terminal portions of the small arms of chromosome pairs 3 and 5 (Fig. 1). The mapping of the LTR-retrotransposons Bianca (*Divo*) and TAR (*Nana*) resulted in a few dozens of clear spotted fluorescent signals throughout the chromosomes (Fig. 2a) and interphase nuclei (Fig. 2c and Fig. 3a) of *C. eugenioides*. Some variation in the mapped sequences was observed among different prometaphases/metaphases, or even between homologous chromosomes of the same metaphase. Nonetheless, some of the fluorescent signals showed repetitiveness among the analyzed images. Three chromosome pairs (5, 9 and 11) did not show detectable fluorescence signals for both elements in any of the analyzed metaphases. For the mobile elements Tork and Athila, the clustered pattern was also observed in chromosomes (Fig. 2b) and nuclei (Fig. 2d and Fig. 3b), although Athila presented some regions with a scattered distribution. A considerable variation in the hybridization signals among the different analyzed cells was observed for all the LTR-retrotransposons, although the distribution of Bianca (*Divo*) and TAR (*Nana*) seemed more reproducible.

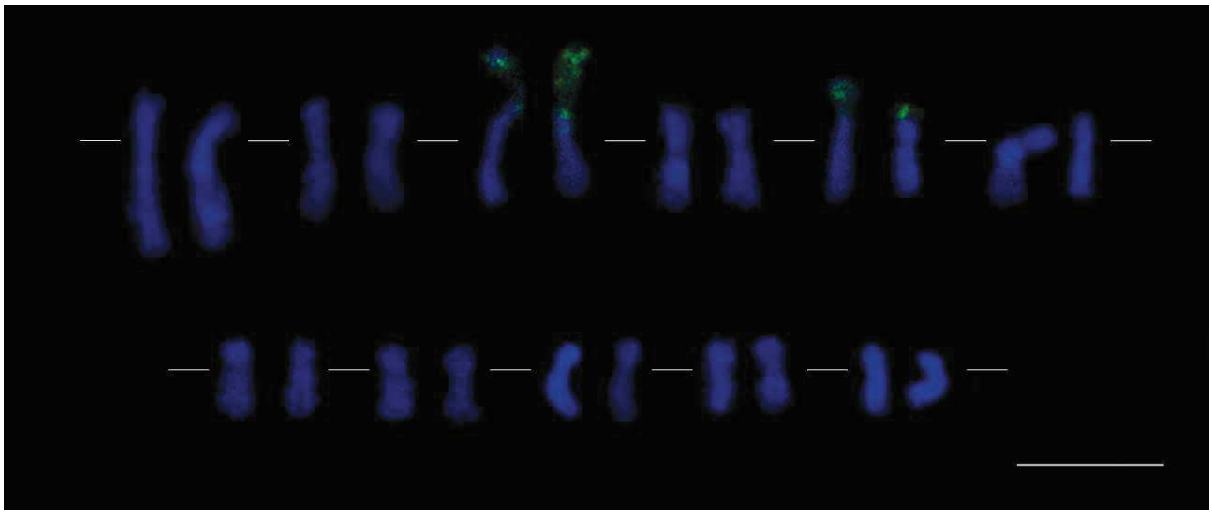

Figure 1. 18S rDNA mapping in the chromosomes of *C. eugenioides*. The 18S rDNA sites (green) co-located with the secondary constrictions, which occur in the terminal portions of the small arms of chromosome pairs 3 and 5. Bar: 5 µm.

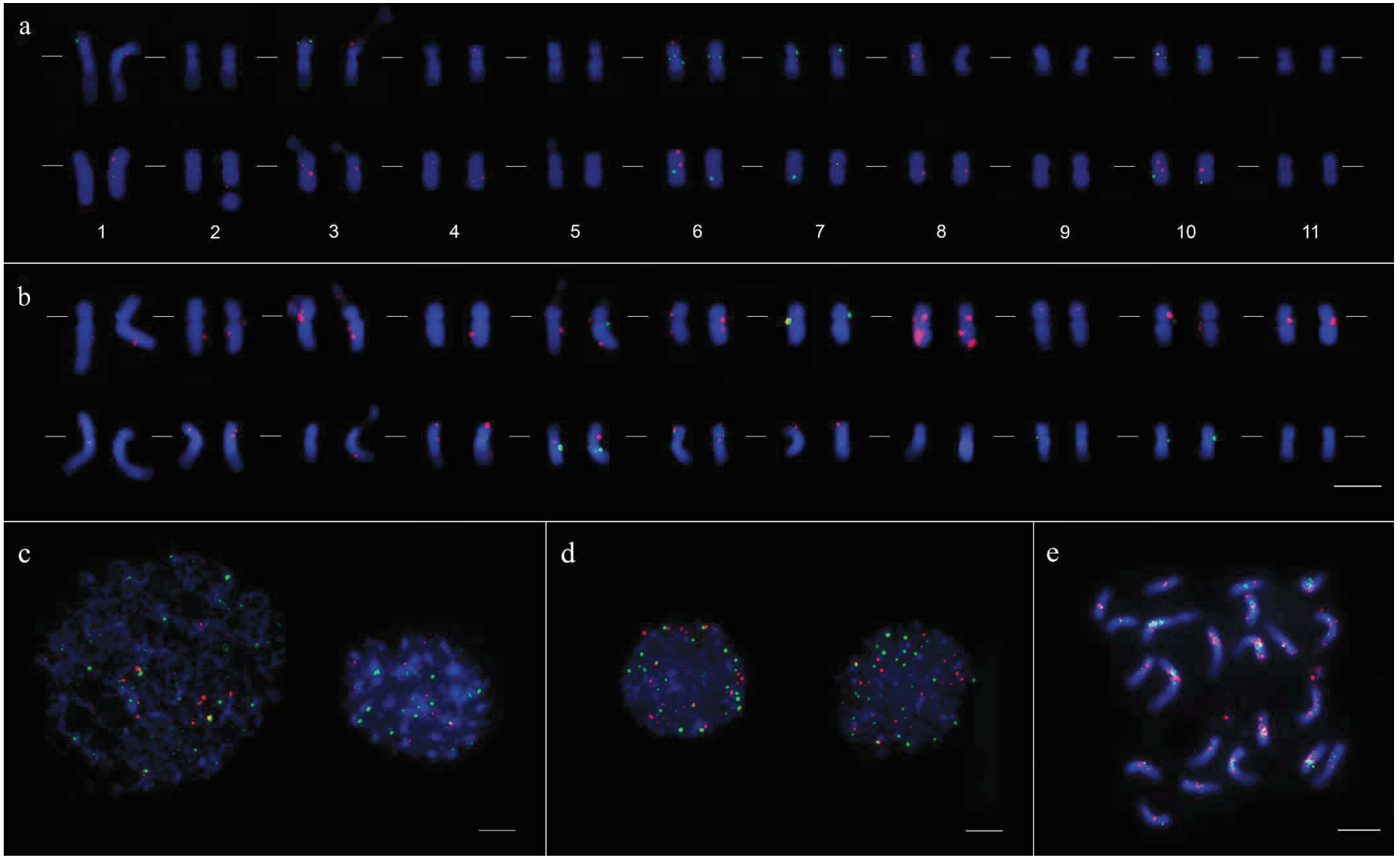
a
1
2
3
4
5
6
7
8
9
10
11
b
c
d
e

Figure 2. Mapping of the RT domains of four LTR-retrotransposons in metaphase chromosomes and interphase nuclei of *C. eugenioides* derived from cell aggregate suspensions. **a)** The Ty1/Copia mobile elements, Bianca (*Divo*, green) and TAR (*Nana*, red) occurred in dispersed clusters, located in multiple regions of the chromosomes, as terminal (pairs 1, 3 and 6), interstitial (pairs 1, 2, 4, 6, 7, and 10), and centromeric/pericentromeric (pairs 3, 6, 7, 8 and 10). The chromosome pairs 5, 9 and 11 did not show any hybridization signals. Note the heterozygous deletion oin the chromosome 2 long arm, and also a TAR (*Nana*) signal at the vicinity of the breaking point. **b)** The mapping of the Ty1/Copia Tork (green) and Ty3/Gypsy Athila (red) mobile elements also exhibited mainly a clustered pattern. The Athila element was the most abundant, occurring in all chromosomes and in diverse regions, including the secondary constriction of pair 3. Tork, on the other hand, exhibited only few clustered signals, mostly in centromeric/pericentromeric regions. For all the four mapped LTR-retrotransposons, there was a considerable variation between metaphases, although the signals from Bianca (*Divo*) and TAR (*Nana*) were more reproducible. **c)** Mapping of Bianca (*Divo*, green) and TAR (*Nana*, red) and; **d)** Tork (green) and Athila (red) in interphase nuclei from cell aggregate suspensions. **e)** A prometaphase from *C. eugenioides* cell aggregate suspensions evidencing the Tork (green) element dispersed in the centromeric portions of some chromosomes and Athila (red) spots co-localized with the secondary constrictions Bars = 5 µm.

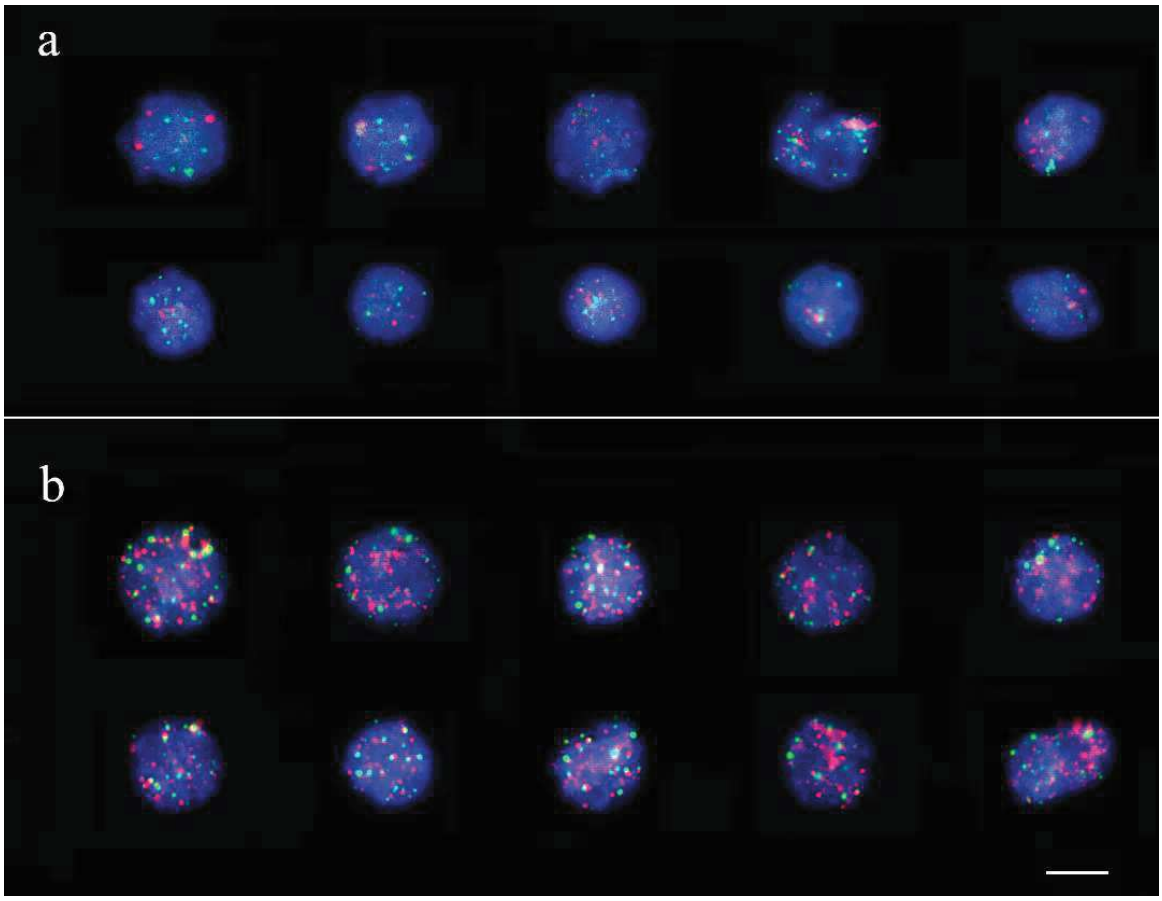

Figure 3. Mapping of the RT domains of four LTR-retrotransposons in *C. eugenioides* interphase nuclei isolated from leaves. **a)** The LTR-retrotransposons Bianca (*Divo*, green) and TAR (*Nana*, red) exhibited a small number (~10 and ~6, respectively) of spotted signals, indicating the occurrence of clusters. The number of TAR spots was smaller than Bianca for all analyzed nuclei. **b)** The LTR-retrotransposons Tork (green) and Athila (red) spots are visibly more numerous when compared to Bianca and TAR. While Tork is distributed mainly in clusters (~18 spots), Athila exhibited clusters (~28 spots) and also scattered distribution in some regions of the nuclei. In addition, Athila was the most representative LTR-retrotransposon in the genome of *C. eugenioides*, followed by Tork, and with Bianca and TAR nearly tied, which is in accordance with the estimated proportions of these elements. Bar: 5 µm.

4. Discussion

4.1. The landscape of *C. eugenioides* repeatome

C. eugenioides repeatome was analyzed and quantified in this work in a cytogenomic perspective, associating publicly available genome sequencing data and bioinformatic tools with the cytogenetic mapping. The RepeatExplorer 2 pipeline used here for the in silico analysis of *C. eugenioides* repeatome is a publicly available tool for the de novo identification of repeated sequences, making this type of analysis more accessible to different research groups and allowing to study the repetitive DNA in different plant species, such as: *Vallisneria spinulosa* S. Z. Yan (Feng et al. 2017), *Passiflora edulis* L. (Pamponét et al. 2019), *Passiflora cincinnata* Mast., *Passiflora organensis* Gardner and *Passiflora quadrangularis* L. (Sader et al. 2021), and multiple *Solanum* L. (Gaiero et al. 2019) and *Spinacia* L. species (Li et al. 2019). The graph-based clustering analysis from RepeatExplorer 2 revealed that repeated sequences accounted for 47% of *C. eugenioides* genome (Table 2), a proportion similar to the previously estimated value of 45% (Guyot et al. 2016).

Mobile elements represented 45% of the genome, with 20 identified families, 15 of Class I and 5 of Class II (Table 2). Here, we focused the further analysis on the Class I LTR-retroelements due to their high proportion on the *C. eugenioides* genome (40%). For other *Coffea* species analyzed so far, the proportion of LTR-retrotransposons on the genome varied from 32% (*C. humblotiana*) to 53% (*Coffea heterocalix* Stoff.). This proportions seems to have an association to the Chevalier botanical classification of *Coffea* and, consequently, with the biogeographical distribution. Species from the Eucoffea group (West and Central Africa), which includes *C. eugenioides*, exhibited the highest proportions, as 49 – 52% for *C. canephora* and 50 – 51% for *C. arabica*. Lower amounts were observed for the Mozambicoffea species (East and South-East Africa, *e. g. Coffea racemosa* Lour. – 39%) and Mascarocoffea species (Indian Ocean islands, *e. g. C. humblotiana* – 32%). These results evidence that these groups underwent different histories of LTR-retroelement accumulation in their genomes throughout *Coffea* evolution (Lopes et al. 2008; Guyot et al. 2016).

The ratio of Ty3/Gypsy to Ty1/Copia observed for *C. eugenioides* in this study was 8 (32%:4%), evidencing a prevalence of Ty3/Gypsy elements. This value was considerably higher than the previously observed for *C. eugenioides* (22.5%:7.5% = 3) and for other *Coffea* species, which ratios varied from 2.6 (*C. canephora*) to 4.75 (*C. humblotiana*), although in all cases the Ty3/Gypsy outnumbered Ty1/Copia elements (Denoeud et al. 2014; Guyot et al. 2016; Raharimalala et al. 2021). Such Ty3/Gypsy to Ty1/Copia ratio also seems to be related to the botanical classification and biogeographical distribution of *Coffea*, in which the species from the Mascarocoffea group exhibited higher ratios than Mozambicoffea and Eucoffea species (Guyot et al. 2016). Outside the *Coffea* genus, this ratio also is highly variable, ranging from 10 for *Gossypium arboreum* L. to 0.25 in *Elaeis guineensis* Jacq. (Li et al. 2019).

Considering Ty1/Copia super family, 8 families were identified and SIRE was the most abundant, representing 3% of the genome, a similar value compared to the previous estimate of 4.5% for *C. eugenioides* and other species from the Eucoffea group (Guyot et al. 2016). The Tork element was the second most abundant Ty1/Copia in *C. eugenioides*, with a frequency of 0.5%. Interestingly, this value is considerably smaller when compared to *C. canephora* and *C. humblotiana*, which possess a frequency of ~1.7% (Raharimalala et al. 2021). Among the 7 identified Ty3/Gypsy families, the most abundant was Tekay, with a proportion of 22%. This family is equivalent to the *Del* element presented in previous studies, in which the proportion for *C. eugenioides* was estimated as ~14% (Guyot et al. 2016). These different values observed, such as 22% and 14% for Tekay, are most likely the result of different approaches and sequence data used for repeatome analysis. For this, different bioinformatic tools must be applied because it is important to associate the data for a more accurate quantitative and qualitative analysis of repeated sequences (Maumus and Quesneville 2016). Two studies have already highlighted the potential role of SIRE and Tekay families of LTR-retrotransposons in the diversification of *Coffea* (Guyot et al. 2016; Raharimalala et al. 2021). In addition, a RT domain based phylogenetic analysis evidenced specific amplifications of Tekay (*Del*), TAT, Athila and CRM for Ty3/Gypsy and SIRE, Tork and Bianca for Ty1/Copia in the *C. canephora* genome, while little specific amplification was observed for *C. humblotiana,* a species from the Mascarocoffea group. This reinforces that differential amplification of these LTR-retrotransposon families in *C. canephora* may have occurred after the divergence of

the two species (Raharimalala et al. 2021). Due to the closeness of *C. eugenioides* to *C. canephora*, their LTR-retrotransposon evolution might have followed similar paths.

4.2. Distribution of rDNA and mobile elements in the genome of *C. eugenioides*

C. eugenioides exhibited two clusters of the 18S rDNA in the terminal regions of chromosomes 3 and 5 short arms. This was the first time that this sequence was cytogenetically mapped to the respective chromosomes of *C. eugenioides*. The 5S rDNA occurs in the long arm of pair 3 (Chapter 1, Fig. 3). The number of rDNA sites in *C. eugenioides* and other *Coffea* species has already been visualized by FISH experiments. This pattern of two chromosome pairs with a secondary constriction/18S rDNA site in the small arms, in which one of these pairs also contains the 5S rDNA site in the interstitial part of the long arm, is also observed for other East African species, as *Coffea salvatrix* Swynn. et Phil and *C. racemosa*. For Central and West African species, as *C. canephora* and *Coffea stenophylla* G. Don, only one pair possess the secondary constriction/18S rDNA site, and the 5S rDNA cluster is located in the long arm a different chromosome (Hamon et al. 2009).

The LTR-retrotransposons Bianca (*Divo*) and TAR (*Nana*) exhibited a small number of localized clusters dispersed along different chromosomes (Fig. 2a), and also in interphase nuclei from cell aggregate suspensions (Fig. 2c, d) and leaves (Fig 3a). Therefore, we confirmed by mapping the relatively low proportion of these sequences (0.22%, Table 2) and also the low estimated copy number (~ 100) in the *C. eugenioides* genome (Guyot et al. 2016). These results are in accordance with genomic results found for *C. eugenioides* and *C. canephora* (Guyot et al. 2016; Dupeyron et al. 2017). The occurrence of clear bright spots indicates that an enough number of copies occur in proximity to allow the detection of the fluorescence signal in prometaphasic/metaphasic chromosomes. In *Saccharum officinarum* L., for instance, the abundant Ale lineage of LTR-retrotransposon also exhibited widespread clusters along euchromatic regions, while the Ivana lineage, which has approximately 50 copies in the genome, was not detectable through FISH in the metaphase chromosomes (Domingues et al. 2012). The clustered pattern is a common feature of the transposable element distribution. When a considerable fraction of the genome

consists of transposable elements, new insertions, even if they occur randomly, will be more likely to occur within or near another transposable element, thus creating and expanding the clusters of mobile elements (Hua-Van et al. 2011; Domingues et al. 2012). In *Zea mays* L., for instance, ~50% of the LTR retrotransposons insert into a pre-existing copy (Stitzer et al. 2021). As these insertions within transposable elements are usually selectively neutral, the clusters are free to expand (Hua-Van et al. 2011; Domingues et al. 2012).

The Tork element (0.5%, Table 2) was another Ty1/Copia mapped in the genome *C. eugenioides* genome that exhibited a clustered distribution pattern, especially on nuclei (Fig 2b, Fig. 3b). The Ty3/Gypsy element Athila (5%, Table 2) exhibited the largest amount of fluorescence signals, with mainly clustered, but also some scattered patterns (Fig. 2b, Fig. 3b). Although some Athila clusters occurred in the centromeric or pericentromeric regions, its distribution also occurred in interstitial and terminal regions. The clustered pattern observed for Athila was different from Tekay (*Del1*), another Ty3/Gypsy mapped in *C. eugenioides* which exhibited dispersed signals (Yuyama et al. 2012). An interesting observation regarding the distribution of Athila is the localization near and/or within the secondary constriction in some metaphases/prometaphases (Fig. 2b, e), which was not the case for Tekay (*Del1*, Yuyama et al. 2012). In fish species, for instance, comparative cytogenomic studies have associated the occurrence of transposable elements within rDNA genes with the dispersion of rDNA clusters in the genomes, as well as with the occurrence of chromosomal rearrangements (Glugoski et al. 2018; Primo et al. 2018). Therefore, a closer look to the Athila LTR-retrotransposon in a comparative background might be interesting to evaluate the contribution of mobile elements to rDNA dispersion and evolution in plants.

A considerable variation in the fluorescence signals between homologous chromosomes and among different cells from *C. eugenioides* was observed and might be the result of different aspects. These aspects might be methodological, such as variation in chromatin condensation level or access of the probe due to residues of cytoplasmatic debris. Nonetheless, the methodology for slide preparation resulted in little or no cytoplasm background. The mapping of the 45S rDNA, for instance, was reproducible in all the analyzed prometaphases/metaphases. Therefore, the variation in the mobile element distribution is most likely the result of its highly polymorphic

nature (Belyayev et al. 2010). The presence or absence of a given transposable elements in a genome, as well as the distribution, might vary between different species or lineages, and also between cells and tissues of a single individual (Gabriel et al. 2006). In addition, clusters of transposable elements are hotspots for illegitimate recombination, which might lead to unequal crossing over and hence duplication of themselves, or DNA sequences at their vicinity, including genes (Pantzartzi et al. 2018).

A deletion in the long arm of one of the chromosomes of pair 2 was found. Interestingly, a TAR (*Nana*) fluorescent signal at the vicinity of the deletion breaking point was also observed (Fig. 2a). Transposable elements have been shown to be associated with chromosomal rearrangements in different species (Gabriel et al. 2006). In humans, the *Alu* element (a Class I non-LTR retrotransposon) represents more than a third of the genome. The activity of this mobile element is associated with several diseases caused by chromosomal rearrangements. Unequal crossing over events between *Alu* elements present in the same chromosome (in cis) can result in deletions or duplications of intervening sequences. On the other hand, the recombination between *Alu* elements present in different chromosomes (in trans) might lead to more complex rearrangements, as translocations (Kolomietz et al. 2002). Therefore, a more profound investigation regarding the dynamics of TAR (*Nana*) family in *Coffea* genomes would be thrilling.

Regarding the intrinsic nature of the mobile elements, stressful conditions, such as the in vitro environment, have been recognized as important triggers for transposition. Plant transposable elements are usually transcriptionally inactivated by cytosine methylation (Lisch 2009). Nonetheless, proliferative cells, as the cellular aggregate suspensions of *C. canephora*, exhibit relatively low levels of global 5-methylcytosine compared to cells in the somatic embryo regeneration step (Amaral-Silva et al. 2021), as observed for other species. In oil palm, transposable elements are also generally hypomethylated at the dedifferentiated stage and then remethylated in regenerated plants (Kubis et al. 2003). In *Arabidopsis thaliana* L., long-term suspension cultures exhibited hypomethylated transposable elements in heterochromatin regions, including the Ty3/Gypsy Athila element (Tanurdzic et al. 2008). Such epigenetic alterations might result in the transcriptional reactivation of specific families of transposable elements. Thus, the variation among transposon

classes in their epigenetic responses to in vitro culture might also be a source for the variation in genome distribution (Vining et al. 2013) and, consequently, a source of somaclonal variation (Barret et al. 2006).

5. Concluding remarks

This is the first study to perform a comprehensive description of the repeat sequences diversity and chromosomal distribution in *C. eugenioides*, by uniting the in silico analysis with the RepeatExplorer2 pipeline and the in situ mapping in metaphasic chromosomes through FISH. Mobile elements comprise 45% of *C. eugenioides* genome, with 20 families identified. Among them, 40% belong to LTR-retrotransposons. The Ty3/Gypsy to Ty1/Copia ratio (8) was considerably higher for *C. eugenioides* when compared to other *Coffea* species (~2.6 to 4.75), although further investigations using complementary bioinformatic tools and sequencing data are necessary to investigate this interspecific variation.

The in situ mapping of four LTR-retrotransposons in metaphasic chromosomes and interphase nuclei derived from cell aggregate suspensions and leaves revealed a general clustered pattern. Among these sequences were two previously described Ty1/Copia elements of *Coffea*, Bianca (*Divo*) and the *Nana* element classified here as TAR. Although only a few hundreds of copies are estimated to occur in the genome of *C. eugenioides*, they were detectable through FISH using probes for their RT domains and generated clear fluorescent spots, indicating their occurrence in clusters. The Ty1/Copia element Tork and the Ty3/Gypsy Athila were also detectable and mapped, with Athila being the most abundant among all four analyzed LTR-retrotransposons.

A considerable variation in the location of the LTR-retrotransposons was observed among different metaphases, which might be a result of the highly polymorphic nature of mobile elements. This is the first study to provide the mapping of transposable elements in metaphase chromosomes of *Coffea* species, as previous studies only reveal karyotypes, without identifying chromosomes. Therefore, we hope the results obtained here provide some background for the advance of repeatome cytogenomic investigations in *Coffea*.

6. Acknowledgments

We would like to thank the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, Brasília – DF, Brazil), Fundação de Amparo à Pesquisa do Espírito Santo (FAPES, Vitória – ES, Brazil), and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES, Brasília – DF, Brazil) for financial support.

7. References

Abreu IS, Carvalho CR, Clarindo WR (2014) Massal induction of *Carica papaya* L. Golden somatic embryos and somaclone screening by flow cytometry and cytogenetic analysis. Cytologia (Tokyo) 79:475–484. https://doi.org/10.1508/cytologia.79.475

Altschul SF, Madden TL, Schäffer AA, et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402. https://doi.org/10.1093/nar/25.17.3389

Amaral-Silva PM, Clarindo WR, Guilhen JHS, et al (2021) Global 5-methylcytosine and physiological changes are triggers of indirect somatic embryogenesis in *Coffea canephora*. Protoplasma 258:45–57. https://doi.org/10.1007/s00709-020-01551-8

Assis R, Baba VY, Cintra LA, et al (2020) Genome relationships and LTR-retrotransposon diversity in three cultivated *Capsicum* L. (Solanaceae) species. BMC Genomics 21:237. https://doi.org/10.1186/s12864-020-6618-9

Barret P, Brinkman M, Beckert M (2006) A sequence related to rice *Pong* transposable element displays transcriptional activation by in vitro culture and reveals somaclonal variations in maize. Genome 49:1399–1407. https://doi.org/10.1139/g06-109

Belyayev A, Kalendar R, Brodsky L, et al (2010) Transposable elements in a marginal plant population: temporal fluctuations provide new insights into genome evolution of wild diploid wheat. Mob DNA 1:6. https://doi.org/10.1186/1759-8753-1-6

Clarindo WR, Carvalho CR (2009) Comparison of the *Coffea canephora* and *C. arabica* karyotype based on chromosomal DNA content. Plant Cell Rep 28:73–81. https://doi.org/10.1007/s00299-008-0621-y

Cros J, Gavalda M-C, Chabrillange N, et al (1994) Variations in the total nuclear DNA content in African Coffea species (Rubiaceae). Café Cacao Thé 38:3–3

Denoeud F, Carretero-Paulet L, Dereeper A, et al (2014) The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. Science 345:1181–1184. https://doi.org/10.1126/science.1255274

Domingues DS, Cruz GM, Metcalfe CJ, et al (2012) Analysis of plant LTR-retrotransposons at the fine-scale family level reveals individual molecular patterns. BMC Genomics 13:137. https://doi.org/10.1186/1471-2164-13-137

Dupeyron M, de Souza RF, Hamon P, et al (2017) Distribution of *Divo* in *Coffea* genomes, a poorly described family of angiosperm LTR-Retrotransposons. Mol Genet Genomics 292:741–754. https://doi.org/10.1007/s00438-017-1308-2

Feng R, Wang X, Tao M, et al (2017) Genome size and identification of abundant repetitive sequences in *Vallisneria spinulosa*. PeerJ 5:e3982. https://doi.org/10.7717/peerj.3982

Feschotte C, Jiang N, Wessler SR (2002) Plant transposable elements: where genetics meets genomics. Nat Rev Genet 3:329–341. https://doi.org/10.1038/nrg793

Gabriel A, Dapprich J, Kunkel M, et al (2006) Global mapping of transposon location. PLoS Genet 2:e212. https://doi.org/10.1371/journal.pgen. 0020212

Gaiero P, Vaio M, Peters SA, et al (2019) Comparative analysis of repetitive sequences among species from the potato and the tomato clades. Ann Bot 123:521–532. https://doi.org/10.1093/aob/mcy186

Galbraith DW, Harkins KR, Maddox JM, et al (1983) Rapid flow cytometric analysis of the cell cycle in Intact plant tissues. Science 220:1049–1051. https://doi.org/10.1126/science.220.4601.1049

Glugoski L, Giuliano-Caetano L, Moreira-Filho O, et al (2018) Co-located hAT transposable element and 5S rDNA in an interstitial telomeric sequence suggest the formation of Robertsonian fusion in armored catfish. Gene 650:49–54. https://doi.org/10.1016/j.gene.2018.01.099

Guyot R, Darré T, Dupeyron M, et al (2016) Partial sequencing reveals the transposable element composition of *Coffea* genomes and provides evidence for distinct evolutionary stories. Mol Genet Genomics 291:1979–1990. https://doi.org/10.1007/s00438-016-1235-7

Hamon P, Duroy P-O, Dubreuil-Tranchant C, et al (2011) Two novel Ty1-*copia* retrotransposons isolated from coffee trees can effectively reveal evolutionary relationships in the *Coffea* genus (Rubiaceae). Mol Genet Genomics 285:447–460. https://doi.org/10.1007/s00438-011-0617-0

Hamon P, Siljak-Yakovlev S, Srisuwan S, et al (2009) Physical mapping of rDNA and heterochromatin in chromosomes of 16 *Coffea* species: a revised view of species differentiation. Chromosome Res 17:291–304. https://doi.org/10.1007/s10577-009-9033-2

Hua-Van A, Le Rouzic A, Boutin TS, et al (2011) The struggle for life of the genome's selfish architects. Biol Direct 6:19. https://doi.org/10.1186/1745-6150-6-19

Kidwell MG (2002) Transposable elements and the evolution of genome size in eukaryotes. 115:49–63. https://doi.org/10.1023/A:1016072014259

Kolomietz E, Meyn MS, Pandita A, Squire JA (2002) The role of *Alu* repeat clusters as mediators of recurrent chromosomal aberrations in tumors. Genome 35:97–112. https://doi.org/10.1002/gcc.10111

Kubis SE, Castilho AM, Vershinin AV, Heslop-Harrison JSP (2003) Retroelements, transposons and methylation status in the genome of oil palm (Elaeis guineensis)

and the relationship to somaclonal variation. Plant Mol Biol 52:69–79. https://doi.org/10.1023/a:1023942309092

Li S-F, Guo Y-J, Li J-R, et al (2019) The landscape of transposable elements and satellite DNAs in the genome of a dioecious plant spinach (*Spinacia oleracea* L.). Mob DNA 10:3. https://doi.org/10.1186/s13100-019-0147-6

Lisch D (2013) How important are transposons for plant evolution? Nat Rev Genet 14:49–61. https://doi.org/10.1038/nrg3374

Lisch D (2009) Epigenetic regulation of transposable elements in plants. Annu Rev Plant Biol 60:43–66. https://doi.org/10.1146/annurev.arplant.59.032607.092744

Lopes FR, Carazzolle MF, Pereira GAG, et al (2008) Transposable elements in Coffea (Gentianales: Rubiacea) transcripts and their role in the origin of protein diversity in flowering plants. Mol Genet Genomics 279:385–401. https://doi.org/10.1007/s00438-008-0319-4

Maumus F, Quesneville H (2016) Impact and insights from ancient repetitive elements in plant genomes. Curr Opin Plant Biol 30:41–46. https://doi.org/10.1016/j.pbi.2016.01.003

Neumann P, Novák P, Hoštáková N, Macas J (2019) Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification. Mob DNA 10:1. https://doi.org/10.1186/s13100-018-0144-1

Noirot M, Poncet V, Barre P, et al (2003) Genome size variations in diploid African *Coffea* species. Ann Bot 92:709–714. https://doi.org/10.1093/aob/mcg183

Novák P, Neumann P, Macas J (2020) Global analysis of repetitive DNA from unassembled sequence reads using RepeatExplorer2. Nat Protoc 15:3745–3776. https://doi.org/10.1038/s41596-020-0400-y

Novák P, Neumann P, Pech J, et al (2013) RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. Bioinformatics 29:792–793. https://doi.org/10.1093/bioinformatics/btt054

Nunes RC, Orozco-Arias S, Crouzillat D, et al (2018) Structure and distribution of centromeric retrotransposons at diploid and allotetraploid *Coffea* centromeric and pericentromeric regions. Front Plant Sci 9:175. https://doi.org/10.3389/fpls.2018.00175

Orozco-Arias S, Isaza G, Guyot R (2019) Retrotransposons in plant genomes: structure, identification, and classification through bioinformatics and machine learning. IJMS 20:3837. https://doi.org/10.3390/ijms20153837

Otto F (1990) DAPI staining of fixed cells for high-resolution flow cytometry of nuclear DNA. In: Methods in cell biology. Elsevier, pp 105–110

Pamponét VCC, Souza MM, Silva GS, et al (2019) Low coverage sequencing for repetitive DNA analysis in *Passiflora edulis* Sims: citogenomic characterization of transposable elements and satellite DNA. BMC Genomics 20:262. https://doi.org/10.1186/s12864-019-5576-6

Pantzartzi CN, Pergner J, Kozmik Z (2018) The role of transposable elements in functional evolution of amphioxus genome: the case of opsin gene family. Sci Rep 8:1–11. https://doi.org/10.1038/s41598-018-20683-9

Primo CC, Glugoski L, Vicari MR, Nogaroto V (2018) Chromosome mapping and molecular characterization of the Tc1/Mariner element in Rineloricaria (Siluriformes: Loricariidae). Braz Arch Biol Technol 61: e18170623. https://doi.org/10.1590/1678-4324-2018170623

Raharimalala N, Rombauts S, McCarthy A, et al (2021) The absence of the caffeine synthase gene is involved in the naturally decaffeinated status of *Coffea humblotiana*, a wild species from Comoro archipelago. Sci Rep 11:8119. https://doi.org/10.1038/s41598-021-87419-0

Razafinarivo NJ, Rakotomalala J-J, Brown SC, et al (2012) Geographical gradients in the genome size variation of wild coffee trees (*Coffea*) native to Africa and Indian Ocean islands. Tree Genet Genomes 8:1345–1358. https://doi.org/10.1007/s11295-012-0520-9

Ribeiro T, Barrela RM, Bergès H, et al (2016) Advancing *Eucalyptus* genomics: cytogenomics reveals conservation of *Eucalyptus* genomes. Front Plant Sci 7. https://doi.org/10.3389/fpls.2016.00510

Sader M, Vaio M, Cauz-Santos LA, et al (2021) Large vs small genomes in Passiflora: the influence of the mobilome and the satellitome. Planta 253:86. https://doi.org/10.1007/s00425-021-03598-0

Serrato-Capuchina A, Matute D (2018) The role of transposable elements in cpeciation. Genes 9:254. https://doi.org/10.3390/genes9050254

Silva JC, Soares FAF, Sattler MC, Clarindo WR (2020) Repetitive sequences and structural chromosome alterations promote intraspecific variations in *Zea mays* L. karyotype. Sci Rep 10:8866. https://doi.org/10.1038/s41598-020-65779-3

Soares FAF, Carvalho CR, Sattler MC, et al (2020) Plant chromosome-specific probes by microdissection of a single chromosome: Is that a reality? Front Plant Sci 11:334. https://doi.org/10.3389/fpls.2020.00334

Tanurdzic M, Vaughn MW, Jiang H, et al (2008) Epigenomic consequences of immortalized plant cell suspension culture. PLoS Biol 6:e302. https://doi.org/10.1371/journal.pbio.0060302

Treangen TJ, Salzberg SL (2012) Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet 13:36–46. https://doi.org/10.1038/nrg3117

Unfried I, Stocker U, Gruendler P (1989) Nucleotide sequence of the 18S rRNA gene from *Arabidopsis thaliana* Co10. Nucleic Acids Res 17:7513. https://doi.org/10.1093/nar/17.18.7513

Vining K, Pomraning KR, Wilhelm LJ, et al (2013) Methylome reorganization during in vitro dedifferentiation and regeneration of *Populus trichocarpa*. 13:92

Wicker T, Sabot F, Hua-Van A, et al (2007) A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8:973–982. https://doi.org/10.1038/nrg2165

Yuyama PM, Pereira LFP, dos Santos TB, et al (2012) FISH using a *gag* -like fragment probe reveals a common *Ty* 3- *gypsy* -like retrotransposon in genome of *Coffea* species. Genome 55:825–833. https://doi.org/10.1139/gen-2012-0081

8. Supplementary information

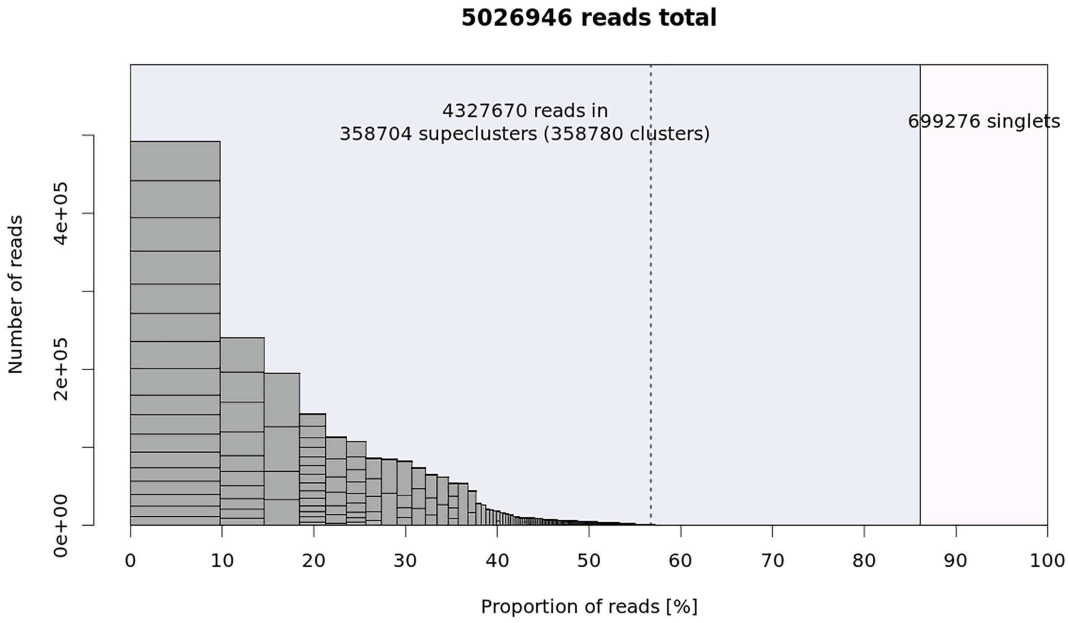

Figure S1. Graphical summary of the clustering results for *C. eugenioides* genome. Bars represent superclusters, with their heights and widths corresponding to the numbers of reads in the superclusters (y-axis) and to their proportions in all analyzed reads (x-axis), respectively. The rectangles inside each supercluster bar represent individual clusters. Top clusters, which included 57% of the total reads, are on the left of the dotted line.

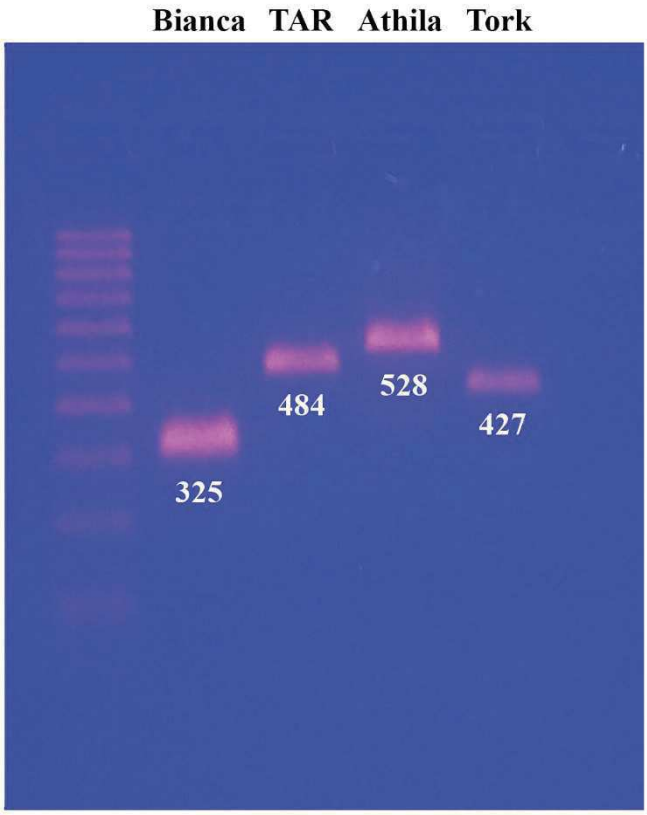

Figure S2. 1.5% agarose gel showing the amplification products generated by PCR using the primers from Table 1. These sequences were used to construct the probes for mapping the RT sequences of Bianca (*Divo*), TAR (*Nana*), Athila and Tork LTR-retrotransposons. The numbers below each band represent the size of the amplicon in base pairs. The first lane contains the 100 bp DNA Ladder.