RENATA DE CASTRO NUNES

ESTRUTURA, DIVERSIDADE E DISTRIBUIÇÃO CROMOSSÔMICA DE LTR-RETROTRANSPOSONS EM *Coffea arabica* E SEUS GENOMAS PARENTAIS

Londrina
2018

RENATA DE CASTRO NUNES

ESTRUTURA, DIVERSIDADE E DISTRIBUIÇÃO CROMOSSÔMICA DE LTR-RETROTRANSPOSONS EM *Coffea arabica* E SEUS GENOMAS PARENTAIS

Londrina

2018

RENATA DE CASTRO NUNES

ESTRUTURA, DIVERSIDADE E DISTRIBUIÇÃO CROMOSSÔMICA DE LTR-RETROTRANSPOSONS EM *Coffea arabica* E SEUS GENOMAS PARENTAIS

Tese apresentada ao Programa de Pós-Graduação em Genética e Biologia Molecular, da Universidade Estadual de Londrina, como requisito parcial para a obtenção do título de Doutor.

Orientador: Dr. André Luís Laforga Vanzela

Coorientador: Dr. Romain Guyot

Londrina
2018

Ficha de identificação da obra elaborada pelo autor, através do Programa de Geração Automática do Sistema de Bibliotecas da UEL

Nunes, Renata de Castro.
ESTRUTURA, DIVERSIDADE E DISTRIBUIÇÃO CROMOSSÔMICA DE LTR-RETROTRANSPOSONS EM Coffea arabica E SEUS GENOMAS PARENTAIS / Renata de Castro Nunes. - Londrina, 2018.
112 f. : il.

Orientador: Dr. André Luís Laforga Vanzela.
Coorientador: Dr. Romain Guyot.
Tese (Doutorado em Genética e Biologia Molecular) - Universidade Estadual de Londrina, Centro de Ciências Biológicas, Programa de Pós-Graduação em Genética e Biologia Molecular, 2018.
Inclui bibliografia.

1. Coffea - Tese. 2. LTR RETROTRANSPOSONS - Tese. 3. CRM - Tese. 4. centromeres - Tese. I. Vanzela, Dr. André Luís Laforga . II. Guyot, Dr. Romain. III. Universidade Estadual de Londrina. Centro de Ciências Biológicas. Programa de Pós-Graduação em Genética e Biologia Molecular. IV. Título.

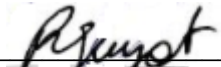
RENATA DE CASTRO NUNES

ESTRUTURA, DIVERSIDADE E DISTRIBUIÇÃO CROMOSSÔMICA DE LTR-RETROTRANSPOSONS EM *Coffea arabica* E SEUS GENOMAS PARENTAIS

Tese apresentada ao Programa de Pós-Graduação em Genética e Biologia Molecular, da Universidade Estadual de Londrina, como requisito parcial para a obtenção do título de Doutor.

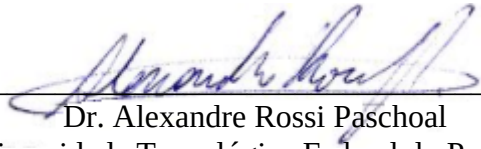
BANCA EXAMINADORA

Orientador: Dr. André Luís Laforga Vanzela
Universidade Estadual de Londrina - UEL



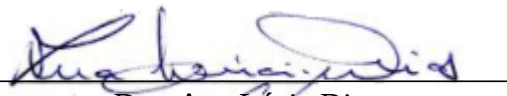
Dr. Romain Guyot
Institut de recherche pour le développement - IRD

Dr. Douglas Silva Domingues
Universidade Estadual Paulista - UNESP



Dr. Alexandre Rossi Paschoal
Universidade Tecnológica Federal do Paraná - UTFPR

Dr. Luiz Filipe Protasio Pereira
Empresa Brasileira de Pesquisa Agropecuária - EMBRAPA



Dra. Ana Lúcia Dias
Universidade Estadual de Londrina - UEL

Londrina, 23 de fevereiro de 2018.

DADOS CURRICULARES DA AUTORA

RENATA DE CASTRO NUNES - nascida em 01 de novembro de 1986, na cidade de Ponte Nova localizada no Estado de Minas Gerais, filha de Joana de Deus Castro Nunes e João Lopes Nunes. Em março de 2008 ingressou no curso de Ciências Biológicas pela Universidade Federal de Lavras, onde em dezembro de 2011 graduou-se com o título de Bacharel em Ciências Biológicas. Durante a graduação foi bolsista de Iniciação Científica pela Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) realizando também outras atividades como monitorias e estágios. Em março de 2012 ingressou no curso de mestrado em Agronomia - Genética e Melhoramento de Plantas pela Universidade Estadual Paulista "Júlio de Mesquita Filho", Campus de Jaboticabal-SP, onde desenvolveu o projeto de dissertação na área de Olericultura/Fitopatologia. Durante o mestrado, foi bolsista inicialmente do CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) e em seguida passou a ser bolsista da FAPESP (Fundação de Amparo a Pesquisa do Estado de São Paulo). Em março de 2014 ingressou no curso de doutorado em Genética e Biologia Molecular pela Universidade Estadual de Londrina (UEL) com projeto de tese em Genética molecular do gênero *Coffea*. Em 2015, participou do programa de doutorado sanduíche no Institut de Recherche pour le Développement (IRD, França) onde adquiriu experiência em bioinformática. Durante o doutorado, contou com o apoio financeiro da Capes (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior).

“Jamais se poderá expressar em frias letras a ternura de um filho ao compreender os sacrifícios de seus pais”

Raumsol

A DEUS por tudo que tens feito em minha vida. Pela alegria de viver, pela minha família, pelos meus amigos, pelo ar que respiro, e pelas oportunidades que possibilitam que eu cresça a cada dia.

AGRADEÇO

Aos meus pais, João e Deuzinha, pelos ensinamentos, pelo apoio, paciência e confiança.

DEDICO

Aos meus irmãos, Marcelo e Tamara, pela amizade de sempre. À cunhada Lu por todo carinho, aos meus amados sobrinhos por fazerem os meus dias mais felizes e ao meu noivo Thiago por todo amor e paciência

OFEREÇO

"Ama sempre, fazendo pelos outros o melhor que possas realizar.

Age auxiliando.

Serve sem apego.

E assim vencerás"

Chico Xavier

AGRADECIMENTOS

A Deus pela presença constante em minha vida.

Aos meus pais, João e Deuzinha, pelos ensinamentos. Vocês são exemplos de dignidade, humildade e amor.

Aos meus irmãos, Marcelo e Tamara, pela amizade e companheirismo.

Aos meus sobrinhos, João Marcelo e Pedro, amores da minha vida.

À cunhada Lú pelo apoio de sempre.

Ao Thiago por todo carinho, amor e paciência.

Aos demais familiares que sempre torceram por mim.

À minha eterna avó Santa que deixou esta vida para ser meu anjo protetor.

À Universidade Estadual de Londrina e ao curso de Pós-Graduação em Genética e Biologia Molecular pela valiosa contribuição em minha formação profissional.

À Capes (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) pela concessão de auxílio financeiro e bolsa de doutorado.

Agradeço ao Prof. Dr. André Luís Laforga Vanzela por todo conhecimento compartilhado, pela confiança depositada em meu trabalho, pelas conversas e conselhos aplicados a minha carreira acadêmica.

Ao Dr. Romain Guyot agradeço a paciência e toda atenção dispensada durante minha estadia na França. Como coorientador e pessoa notável, terá sempre meu respeito e gratidão.

Aos membros da banca examinadora pela disponibilidade e sugestões de correção.

Aos professores do programa de Pós-graduação em Genética e Biologia Molecular pelos conhecimentos compartilhados.

Aos funcionários da Universidade Estadual de Londrina (UEL) por todo auxílio.

Aos pesquisadores do Institut de Recherche pour le Développement (IRD, França) por toda contribuição.

A equipe LCDV pelo companheirismo e convívio harmonioso dentro e fora do laboratório. Agradeço pela contribuição científica, pela amizade, pelas risadas, conselhos, e que o olhar pelo próximo seja sempre recíproco.

Aos amigos Maíra, Isa, Jamine, Luana, Elias e Guilherme pelo acolhimento em Montpellier/France.

À minha grande amiga Joana pela sincera amizade e por tornar os meus dias em Londrina mais felizes e animados.

Aos meus amigos de Ponte Nova e Lavras que sempre me deram força para continuar em frente.

Agradeço também a todos, que de alguma forma, torceram por minha vitória, seja pela ajuda constante ou por uma palavra de amizade!

NUNES, Renata de Castro. **Estrutura, diversidade e distribuição cromossômica de LTR-retrotransposons em *Coffea arabica* e seus genomas parentais.** 2018. 112 p. Tese (Doutorado em Genética e Biologia Molecular) – Universidade Estadual de Londrina, Londrina. 2018.

RESUMO

LTR-retrotransposons (LTR-RT) são abundantes nos genomas das plantas, sendo *Gypsy* e *Copia* os mais representativos. A região centromérica acumula diferentes arranjos de sequências repetitivas, como DNA satélite e retrotransposons. Neste estudo foram triados LTR-RT da família *CRM*, baseado em domínios conservados da gag-POL nos genomas de *Coffea canephora*, *C. eugenioides* e *C. arabica*. Essas sequências foram anotadas e comparadas para verificar a diversidade desses elementos nos três genomas, e entender a dinâmica dos *CRM* na formação do híbrido. Adicionalmente, sondas foram preparadas para a localização física de *CRM* por FISH. *Gypsy* representou a maioria dos retrotransposons (>50%), sendo que dessa fração, *CRM* representou ~20%. A comparação dos domínios conservados da transcriptase reversa de *CRM* permitiu caracterizar dez grupos de retrotransposons centroméricos de *Coffea* (CRC), com >99% de identidade. Esses foram chamados de A, B, C, D, E, F, G, H, X e Y. A comparação da posição dos domínios conservados, usando a análise gráfica do Mauve, confirmou esses grupos. Exceto pelo grupo A, os demais CRC não apresentaram cromodomínio, mas exibiram a cadeia poli-A e o CR motif. Apenas o grupo D em *C. canephora* e Y em *C. arabica* foram espécie-específicos. A FISH com a sonda da transcriptase reversa de *CRM* mostrou sinais acumulados preferencialmente nas regiões proximais nos três genomas. Contudo, *C. eugenioides* mostrou menos sinais intersticiais em relação às outras duas espécies. Os sinais de hibridização foram heterogêneos nas três espécies, com diferenças na intensidade dos sinais centroméricos, distribuição dispersa, além de ausência de sinais em alguns cromossomos. Em geral, os sinais de FISH foram colocalizados com a heterocromatina DAPI $^{+}$, comum nas regiões proximais de *C. canephora* e *C. arabica*. Nossos resultados mostram que a família *CRM* é diversa nos genomas de *Coffea*, seja do ponto de vista da estrutura do LTR-RT de cada CRC, como também da ocorrência e distribuição cariotípica, com sinais fora das regiões proximais, além de diferenças nos tamanhos. Apesar dessa diversidade, os cromossomos com sinais proximais mais intensos observados em *C. canephora* e *C. eugenioides*, também foram vistos em *C. arabica*, o que fortalece a hipótese da origem desse híbrido.

Palavras-chaves: Centrômeros. Cromovírus. LTR-RT. Transcriptase reversa.

NUNES, Renata de Castro. **Structure, diversity and chromosome distribution of centromeric LTR-retrotransposons in *Coffea arabica* and its parental genomes.** 2018. 112 p. Thesis (Doctorate in Genetics and Molecular Biology) – Universidade Estadual de Londrina, Londrina. 2018.

ABSTRACT

LTR-retrotransposons (LTR-RT) are abundant in plant genomes; *Gypsy* and *Copia* are the most representative. The centromeric region accumulates different arrangements of repetitive sequences, such as satellite DNA and retrotransposons. In this study, LTR-RT of the *CRM* family were screened based on the gag-POL conserved domains in *Coffea canephora*, *C. eugenioides* and *C. arabica* genomes. These sequences were annotated and compared to assess the diversity of these elements in the three genomes, and understand *CRM* dynamics in the formation of the hybrid. Additionally, probes were prepared for physically locating *CRM* using FISH. *Gypsy* represented the majority of retrotransposons (>50%), from which *CRM* represented ~20%. The comparison between the conserved domains of *CRM* reverse transcriptase allowed the characterization of ten groups of centromeric retrotranspons of *Coffea* (CRC) with >99% identity. These groups were called A, B, C, D, E, F, G, H, X and Y. Comparison between the conserved domain position, using graphic analysis in Mauve, confirmed these groups. Except for group A, the remainder of the CRC did not possess chromodomain, but exhibited poli-A chain and CR motif. Only group D in *C. canephora* and Y in *C. arabica* were species-specific. FISH with *CRM* reverse transcriptase showed preferentially accumulated signals in the proximal regions in all three genomes. However, *C. eugenioides* showed less interstitial signals in relation to the other two species. Hybridization signals were heterogeneous in the three species, with differences in the intensity of centromeric signals, dispersed distribution and signal absence in some chromosomes. In general, FISH signals were colocalized with $DAPI^{+}$ heterochromatin, commonly observed in the proximal regions of *C. canephora* and *C. arabica*. These results show the *CRM* family is diverse in *Coffea* genomes whether regarding structure of each CRC LTR-RT or karyotypic occurence and distribution, with signals outside proximal regions as well as different in size. Despite this diversity, chromosomes with more intense proximal signals, observed in *C. canephora* and *C. eugenioides*, were also seen in *C. arabica*, which supports the hypothesis of this hybrid's origin.

Keywords: Centromeres. Chromovirus. LTR-RT. Reverse transcriptase.

LISTA DE ILUSTRAÇÕES

LISTA DE DADOS SUPLEMENTARES

LISTA DE TABELAS

LISTA DE ABREVIAÇÕES

cDNA	DNA complementar
CenH3	*Centromeric histone 3*
cpDNA	DNA dos cloroplastos
DIRS	sequências intermediárias repetidas de *Dictyostelium* (*Dictyostelium Intermediate Repeat Sequence)*
DNA	ácido desoxirribonucleico (*Deoxyribonucleic acid*)
DNAr	DNA ribossômico
DNAsat	DNA satélite
ERVs	*endogenous retroviruses*
Ets	elementos transponíveis
FISH	hibridização in situ (*fluorescent in situ hybridization*)
INT	integrasse
ISSR	*inter simple sequence repeat*
ITS	*internal transcribed spacer*
LINE	sequências longas interespaçadas (*long interspersed nuclear element*)
LTR	sequências terminais repetidas (*long terminal repeat*)
Mpb	mega pares de base
ORF	*open reading frames*
pb	pares de base
PCR	reação em cadeia da polimerase (*polymerase chain reaction*)
PLE	*Penelope-like element*
POL	cadeia poligênica (*Poligenic chain*)
PR	protease

Rep	*replication initiator*
RFLP	*restriction fragment length polymorphism*
RH	RNAseH
RNA	ácido ribonucleico (*ribonucleic acid*)
RT	transcriptase reversa (*reverse transcriptase*)
SINE	sequências curtas interespaçadas (*short interspersed nuclear element*)
TIR	repetições terminais invertidas (*terminal inverted repeat*)

SUMÁRIO

1. FUNDAMENTAÇÃO TEÓRICA

1.1. Sequências de DNA repetitivo

A maior parte dos genomas vegetais é composta por DNA repetitivos, que se diferenciam quanto à composição das sequências, a distribuição física e o papel nos genomas, bem como pela homologia e distribuição nas espécies (Schmidt e Heslop-Harrison, 1998). As sequências de DNA repetitivo podem estar organizadas em blocos consecutivos (*tandem*), que podem ser de DNA codificante, como por exemplo, as sequências de DNA ribossômico, e também de DNA não codificante, que englobam, por exemplo, os microsatélites (1-10 pb), os minisatélites (10-100 pb) e os DNA satélites (>100 pb) (Bennetzen, 2000; Heslop-Harrison e Schmidt, 2007). Esses últimos, podem alcançar milhões de cópias, podendo se movimentar nos cromossomos por dispersão equilocal e equidistante (Hemleben et al., 2007; Zakrzewski et al., 2010; Bardella et al., 2014). Algumas sequências repetidas podem ainda se acumular de forma dispersa ou agrupadas, como os elementos transponíveis (Jurka et al., 2007) **(Figura I).**

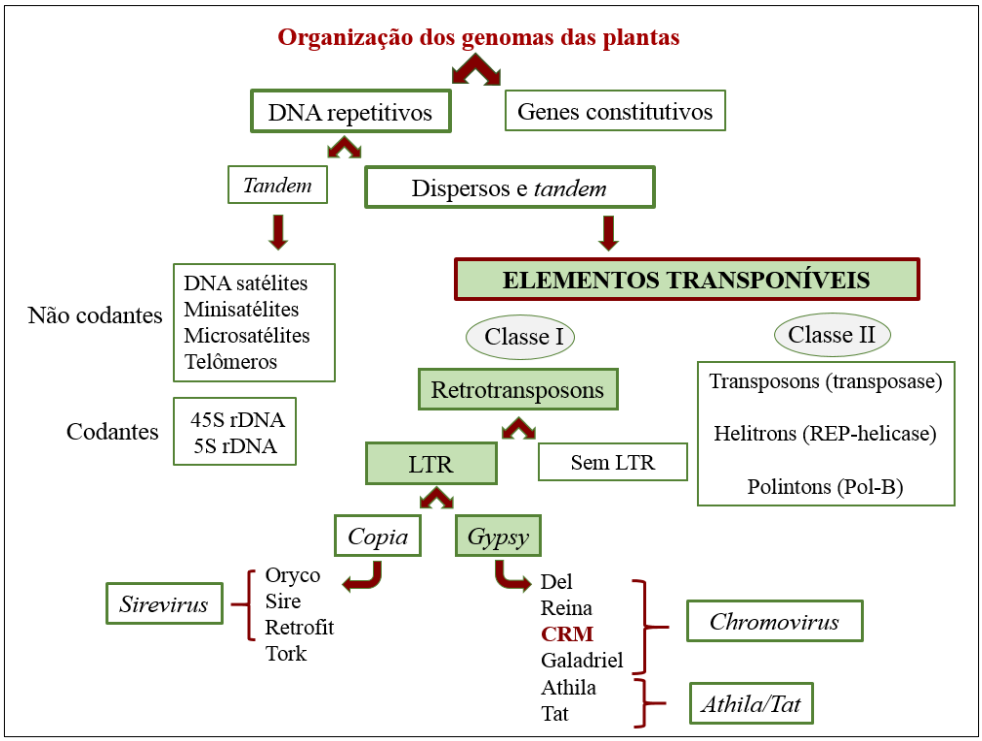


Figura I. Organização geral dos genomas das plantas.

1.2. Elementos transponíveis

A caracterização genética dos elementos transponíveis (ETs) iniciou-se com estudos em milho desenvolvidos por Bárbara McClintock (1951), nos quais foi demonstrada a existência de fatores capazes de se moverem e influenciarem a expressão de genes. Depois de sua descoberta original, os elementos transponíveis foram encontrados em todos os organismos (Bennetzen, 2000). Esses elementos podem compreender até 80% do DNA total em plantas, como em trigo (Brenchley et al., 2012) e milho (Schnable et al., 2009). No genoma das orquídeas e no tomate, os ETs correspondem a 60% das sequências (Cai et al., 2015; Mehra et al., 2015). Em *Coffea canephora,* eles representam 50% do genoma (Denoeud et al., 2014), em videira 40% (The French–Italian Public Consortium for Grapevine Genome Characterization, 2007) e em arroz representam um pouco menos, em torno de 35% (International Rice Genome Sequencing Project, 2005). A repetibilidade desses ETs nos genomas pode levar à recombinação desigual, o que pode causar uma reorganização cromossômica de pequenas ou de grandes proporções, ou causar outros efeitos, como a inativação, a criação e a regulação de genes (Bennetzen e Wang, 2014; Chaparro et al., 2015). Os elementos de transposição podem também ser classificados como autônomos, quando possuem todas as enzimas necessárias para a transposição de modo independente, enquanto que os não autônomos utilizam os componentes codificados pelos autônomos para realizar a transposição (Jurka et al., 2007).

Diante da abundância e da evidente diversidade dos ETs, e mediante a dificuldade de nomeá-los, Wicker et al. (2007) propuseram um sistema unificado de nomenclatura, na intenção de integrar conceitos anteriormente adotados, fornecendo um consenso entre as várias classificações conflitantes e sistemas de nomeação existentes. Esta classificação

divide os elementos transponíveis hierarquicamente em classe, subclasse, ordem, superfamília, família e subfamília.

1.2.1. Elementos transponíveis de classe 1

Elementos transponíveis que realizam a transposição por meio de um RNA intermediário pertencem a Classe 1, e são chamados de retrotransposons. Neste processo, uma cópia de cDNA gerada pela transcriptase reversa é inserida em algum ponto do genoma alvo (Wicker et al., 2007). Cada ciclo completo de replicação produz, portanto, uma nova cópia do elemento, fazendo dos retrotransposons os principais contribuintes para a fração repetitiva dos grandes genomas (Kumar e Bennetzen, 1999).

Os retrotransposons podem ser divididos em cinco ordens: LTR retrotransposons, DIRS, PLE, LINEs e SINEs (Wicker et al., 2007). Os retrotransposons com LTR, assim chamados por possuírem longas repetições terminais (LTRs - *Long Terminal Repeats*), são os mais abundantes devido à sua mobilidade (Grandbastien, 2015), e contribuem para a variação do tamanho e estrutura dos genomas observados nas plantas (Piegu et al., 2006; Heslop-Harrison e Schwarzacher, 2011; Tenaillon et al., 2011). Estes retrotransposons são compostos pela ORF *gag*, responsável por codificar uma poliproteína estrutural para partículas semelhantes às de vírus, e pela cadeia poligênica (*pol*) que codifica transcriptase reversa (RT), RNAse (RH), integrase (INT) e proteinase aspártica (PR) (Wicker et al., 2007). Segundo esses autores, a ordem de retrotransposons LTR é ainda subdividida em cinco superfamílias: *Copia*, *Gypsy*, *Bel-Pao*, *Retrovirus* e *ERVs* (do inglês: *endogenous retroviruses*). As superfamílias *Copia* e *Gypsy* são as mais importantes dos genomas eucarióticos devido à sua frequência e diversificação (Flavell et al., 1992; Wicker et al., 2007) e são amplamente distribuídas no reino vegetal (Marco e Marín, 2005). Essas superfamílias se diferem pela ordem da proteína integrase na *pol*,

sendo em *Gypsy* (PR-RT-RH-INT) e em *Copia* (PR-INT-RT-RH) (Llorens et al., 2009). As relações entre essas superfamílias são usualmente estabelecidas com base em análises de sequências de RT, e são divididas em famílias, como por exemplo, *Sire*, *Oryco*, *Retrofit* (*Sirevirus*) e *Tork* em *Copia*, e *CRM*, *Galadriel*, *Reina* (*Chromovirus*) e *Athila/Tat* em *Gypsy* (Llorens et al., 2009) **(Figura I).**

1.2.2. Elementos transponíveis de classe 2

Segundo Wicker et al. (2007), os elementos dessa classe são separados em duas subclasses: 1 e 2. A primeira compreende os elementos transponíveis do tipo "corta-e-cola" da ordem TIR, caracterizados por suas repetições terminais invertidas de comprimento variável. Esta ordem é subdividida em nove superfamílias (Tc1–Mariner, hAT, Mutator, Merlin, Transib, P, PiggyBac, PIF- Harbinger e CACTA) que apresentam uma ORF para transposase. Essas superfamílias, distinguem-se pelo tamanho das repetições terminais invertidas e das duplicações de sítio alvo (Wicker et al., 2007). Durante a transposição dos elementos desta classe, o transposon é excisado de um local e inserido em outro. Ambas as reações são catalisadas pela transposase, que reconhece as repetições terminais invertidas e corta ambas as fitas de DNA em cada extremidade, ligando-as posteriormente ao seu sítio alvo (Jurka et al., 2007). Os elementos da segunda subclasse (Helitrons e Maverick ou Polintons), se movimentam em um processo de transposição, replicação sem a quebra da dupla fita de DNA, sendo por isso acentuadamente diferentes dos elementos inclusos na subclasse 1 (Wicker et al., 2007). A transposição dos elementos da ordem Helitron, descrita por Kapitonov e Jurka (2006), começa a partir de uma quebra num sítio iniciador de replicação específico chamado *Rep* (do inglês - replication initiator) codificado no transposon. Em seguida a extremidade livre 3-OH, recém produzida pela quebra, funciona como um *primer* para a síntese da fita

líder, facilitada pela helicase (a qual é codificada em uma ORF no transposon) e por algumas proteínas de replicação do hospedeiro, incluindo a DNA polimerase. Ao final da replicação do transposon, o sítio iniciador de replicação catalisa uma reação de transferência de cadeia, que resulta em um DNA de fita dupla, composto por uma fita "antiga" e por aquela recém-sintetizada, e um DNA de cadeia simples, proveniente de uma das fitas "antigas", que é liberado e que reintegra-se ao genoma hospedeiro. Os elementos transponíveis da ordem Maverick, também conhecidos como Polintrons, são cercados por TIRs e podem codificar até onze proteínas (Kapitonov e Jurka, 2006). A maioria deles codifica a DNA polimerase B e uma integrase; no entanto, eles não contêm codificação para transcriptase reversa, sugerindo que eles passem por transposição replicativa sem intermediários de RNA (Kapitonov e Jurka, 2006). Segundo esses autores, a transposição pode ocorrer pela excisão do transposon de uma única fita de DNA, seguida por replicação extracromossômica e integração em um novo sítio.

1.2.3. DNA centromérico

O centrômero é a região responsável pela coesão de cromátides irmãs e pela segregação cromossômica regular durante a mitose e meiose, sendo essencial para o desenvolvimento e proliferação celular em todos os organismos (Lermontova et al., 2015; Marques et al., 2015; Han et al., 2016). Apesar de sua função ser conservada nos eucariotos, os centrômeros das plantas apresentam uma diversidade considerável em tamanho, estrutura e composição (Comai et al., 2017). Além disso, as sequências de DNA centroméricas apresentam poucas semelhanças entre as espécies estreitamente relacionadas (Plohl et al., 2014; Feng et al., 2015). A organização e a complexidade dos centrômeros variam consideravelmente em diferentes organismos (Jiang et al., 2003; Henikoff e Dalal, 2005). Em *Saccharomyces cerevisiae*, os centrômeros consistem apenas

em 125 pb de sequência única (Clarke 1998, 1990). Em contraste, os centrômeros da maioria dos eucariotos são muito mais complexos, como em arroz (Zhang et al., 2004) e milho (Jin et al., 2004). Acredita-se que a identidade e a herança do centrômero estão associadas à organização da cromatina, que difere dos outros domínios cromossômicos pela presença de uma variante da histona H3 (CenH3), que localiza-se exclusivamente no centrômero de todos os eucariotos estudados e constitui um marcador de funcionalidade dessa região (Jiang et al., 2003; Wang et al., 2009). Os centrômeros podem conter longos arranjos de DNA satélite, que podem estar associados com retrotransposons da família *CRM* (Houben et al., 2007; Marques et al., 2015; Santos et al., 2015). A família *CRM* (Centromeric Retrotransposon of Maize), chamada assim pois foi inicialmente reconhecida em milho, pertence ao clado dos cromovírus e está localizada preferencialmente em regiões cromossômicas proximais (Du et al., 2010; Sharma e Presting, 2014).

A inserção do *CRM* no genoma pode ter um papel importante na função e evolução dos centrômeros das plantas (Zhong et al., 2002; Neumann et al., 2011; Gao et al., 2015). Os membros desta família possuem um domínio *zinc finger* chamado de cromodomínio (CHRomatin Organization MOdifer domain), na posição C-terminal da integrase, capaz de reconhecer as proteínas CENH3 associada às regiões do centrômero, ou podem ainda possuir outro domínio chamado CR *motif* também localizado na posição C-terminal da integrase, estendendo-se para o LTR 3 ' (Houben et al., 2007; Neumann et al., 2011). Retrotransposons centroméricos (CR) foram encontrados em todos os centrômeros das gramíneas já estudadas, como milho (*CRM*; Nagaki et al., 2003), arroz (CRR; Bao et al., 2006) cana-de-açúcar (CRS; Nagaki e Minoru, 2005) e trigo (CRW, Liu et al., 2008). Em uma pesquisa abrangente, elementos CR de 12 famílias de plantas monocotiledôneas e dicotiledôneas foram classificados em três grupos, de acordo com suas propriedades

estruturais e distribuição cromossômica: o Grupo A que possui um CR *motif* e aparece localizado em regiões centroméricas, o Grupo B sem qualquer domínio específico e o Grupo C contendo um cromodomínio, localizados dispersos ao longo dos cromossomos (Neumann et al., 2011). No entanto, a estrutura, diversidade e a contribuição dos elementos CR para a organização da região do centrômero das plantas ainda não está clara.

1.3. O gênero *Coffea* e os aspectos citogenéticos

O gênero *Coffea* pertence à família Rubiaceae e compreende 125 espécies (Hamon et al., 2017). *Coffea arabica* L. e *C. canephora* Pierre ex A. Froehner são as duas espécies de maior importância econômica do gênero, sendo o café uma das *commodities* mais valorizadas nos mercados internacionais (ICO, International Coffee Organization, http://www.ico.org, 2018). Muitos estudos são direcionados para o entendimento da organização genômica de *C. arabica*, enquanto as espécies selvagens são utilizadas para subsidiar os programas de melhoramento, como fonte de características relacionadas à resistência a diferentes doenças e adversidades ambientais (Herrera et al., 2002; Prakash et al., 2002). Evidências botânicas indicam que *C. arabica*, espécie alotetraploide ($2n = 4x = 44$), foi originada no platô da Etiópia Central onde ainda pode ser encontrada em estado selvagem. Essa espécie foi originada por um cruzamento relativamente recente, há menos de 1 milhão de anos atrás, entre *C. canephora* e *C. eugenioides*, ambas diploides com $2n = 22$ (Lashermes et al., 1999; Yu et al., 2011; Cenci et al., 2012). Essas duas últimas espécies divergiram há cerca de 4,2 milhões de anos (Yu et al., 2011), e estudos de segmentos do genoma dos cloroplastos (cpDNA) sugeriram que um ancestral de *C. eugenioides* ou uma espécie próxima seria o ancestral materno de *C. arabica* (Tesfaye et al., 2007).

Estudos para determinar a morfologia dos cromossomos somáticos de café foram realizados inicialmente em *C. excelsa* Chev., reconhecida atualmente como *C. liberica* var. dewevrei (Mendes, 1938). Contudo, este autor encontrou dificuldades técnicas, uma vez que os cromossomos são muito pequenos (1 a 3 µm) e morfologicamente similares. Clarindo e Carvalho (2008, 2009) desenvolveram os primeiros cariogramas de alta definição para *C. arabica* e *C. canephora*. De acordo com esses autores, *C. arabica* é composta por cinco pares metacêntricos, 16 pares submetacêntricos e um par acrocêntrico, originados de genitores com cariótipos muito semelhantes, enquanto que *C. canephora* é composta por dois pares metacêntricos e nove submetacêntricos, sendo que os pares 1, 10 e 11 de *C. canephora* são muito similares aos pares 1, 19 e 21 de *C. arabica*. Essas informações básicas reforçam a ideia de *C. canephora* como genitor de *C. arabica*, como também observado em estudos de ITS (*internal transcribed spacer*) (Lashermes et al., 1997), RFLP (*restriction fragment length polymorphism*) e hibridação genômica *in situ* (Lashermes et al., 1999) e ISSR (*inter-simple sequence repeat*) (Ruas et al., 2003).

Estudos com bandamento C e fluorocromos cromomicina A3 (CMA3) e 4,6-diamino-2-fenilindol (DAPI) foram realizados para determinar a ocorrência e a localização da heterocromatina em espécies de *Coffea* (Lombello e Pinto-Maglio, 2004a; 2004b; Hamon et al., 2009). Estudos de bandamento CMA/DAPI, em sete espécies de *Coffea*, mostraram variações no número e na localização das bandas CMA, enquanto que as bandas $DAPI^+$ foram encontradas somente em uma das espécies estudadas (Lombello e Pinto-Maglio, 2004a; 2004b). Em um estudo realizado por Hamon et al. (2009) em 16 espécies de *Coffea,* cinco espécies também apresentaram variações com relação ao número de bandas CMA, enquanto que as bandas $DAPI^+$ foram encontradas nas regiões intersticiais na maioria das espécies e sem definição do par.

Em relação à localização física do DNA ribossômico (DNAr), Lombello e Pinto-Maglio (2004a; 2004b) mostraram dois sítios terminais de DNAr 45S em seis espécies de *Coffea* e dois sítios intersticiais de DNAr 5S em cada espécie. Contudo, Raina et al. (1998) revelaram quatro sítios de DNAr 45S e seis de 5S em *C. arabica*, com dois sinais de 5S localizados em um cromossomo de cada homólogo (sem identificação do par). Segundo Hamon et al. (2009), o número de segmentos de DNAr em *Coffea* está correlacionado com a possível região de origem das espécies na África. Para Mishima et al. (2002) sítios de 5S geralmente são eliminados depois do processo de poliploidia, enquanto os sítios de 45S são mais conservados, por ocorrência de diferentes eventos, como *crossing-over* desigual, conversão de genes e transposição (Leitch e Heslop-Harrison, 1993; Taketa et al., 1999; Hasterok et al., 2001). Com relação aos ETs, o processo de hibridização parece ter favorecido o acúmulo desses elementos no híbrido, já que duas famílias de transposons (MuDR e Tip100), bem como uma família de retrotransposons com LTR (Del) apresentaram sinais de hibridização agrupados, preferencialmente em posições terminais cromossômicas, nos parentais *C. eugenioides* e *C. canephora*, enquanto que os sinais intersticiais e /ou proximais foram observados em maior número em *C. arabica* var. tipica, sugerindo um aumento na atividade de transposição no híbrido em comparação com as espécies parentais (Lopes et al., 2013).

1.4. Citogenômica

A citogenética molecular possibilita a realização de estudos sobre a estrutura cromossômica de maneira mais detalhada, usando uma associação de técnicas convencionais, moleculares e de bioinformática. Esse conjunto de ferramentas possibilita caracterizar e localizar de maneira rápida e eficiente, genes, sequências repetitivas,

cromossomos isolados, assim como genomas inteiros (Kato et al., 2005; Danilova e Birchler, 2008; Marques et al., 2015; Santos et al., 2015; Ribeiro et al., 2016).

A citogenômica é uma prática na qual sequências podem ser baixadas de bancos de dados, ou montadas a partir de *reads* curtos e empregadas em uma abordagem citológica, visando o estudo da biologia cromossômica e nuclear focando, por exemplo, na diversidade de famílias de DNA repetitivo de diferentes naturezas. Há vários exemplos onde a citogenômica foi empregada com sucesso no estudo da evolução genômica e cromossômica em plantas.

Zakrzewski et al. (2010), por exemplo, usaram análises de bioinformática e FISH para verificar a amplificação e distribuição de minisatélites em *Beta vulgaris*. Esses autores observaram sequências de moderada a altamente repetitivas, com sinais dispersos em regiões heterocromáticas nos cromossomos. Heckmann et al. (2013) estudaram a distribuição de sequências repetitivas nos cromossomos holocêntricos de *Luzula elegans,* e observaram que os retrotransposons da superfamília *Copia* são os mais representativos no genoma, com sequências distribuídas de maneira dispersa nos cromossomos. No genoma de *Erianthus arundinaceus,* a superfamília *Copia* também apresentou grande representatividade, porém, os sinais de hibridização estavam localizados nas extremidades da maioria dos cromossomos (Huang et al., 2017). Santos et al. (2015), buscaram retrotransposons ativos de *Brachiaria decumbens* em um transcriptoma de raiz, focando em regiões proteicas conservadas da cadeia poligênica. Esses autores conseguiram identificar e localizar *in situ* famílias de *Gypsy* entre espécies diploides e poliploides, mostrando um acúmulo diferencial de cada família ao longo dos cromossomos, tanto em blocos quanto dispersos. Em cana-de-açúcar, famílias de LTR-RT apresentaram comportamentos distintos, podendo impactar o genoma de diversas

formas, por exemplo, provocando mecanismos de silenciamento de genes (Domingues et al., 2012).

Recentemente, o genoma sequenciado de *C. canephora* revelou uma contribuição importante de elementos transponíveis (~50%), sendo a maioria deles LTR-retrotransposons (Denoeud et al., 2014). Estudos dessas sequências repetitivas no transcriptoma de três espécies de (*C. arabica*, *C. canephora* e *C. racemosa*) sugerem que esses elementos estão ativos no genoma (Lopes et al., 2008; 2013). Estudos utilizando FISH em *Coffea*, mostraram que sondas de *Copia* apareceram predominantemente dispersas ao longo de cromossomos de *C. arabica*, *C. canephora* e *C. eugenioides*, mas com diferenças óbvias nos sinais acumulação de acordo com cromossomos parentais (Herrera et al., 2013). Ao contrário, sondas de *Gypsy* apareceram em blocos nos cromossomos de *C. canephora* e dispersos em *C. eugenioides*, e ambos os padrões de distribuição foram observados em *C. arabica* (Yuyama et al., 2012). A origem e a função desta variação de perfis de distribuição entre espécies diploides e poliploides permanecem desconhecidas até o momento, no entanto, acredita-se que as diferenças encontradas na distribuição dos retroelementos podem ser em função da sua atividade variável durante a evolução dos genomas das espécies ou de rearranjos cromossômicos (Belyayev et al., 2001; Santini et al., 2012).

2. OBJETIVOS

2.1. Geral

O objetivo deste estudo foi identificar, classificar e estudar a diversidade e distribuição de retrotransposons centroméricos (elementos CR) nos cromossomos de *C. arabica* e seus parentais, buscando sinais de hibridização, relações genômicas e cariotípicas.

2.2. Específicos

a. Identificar, agrupar e analisar as sequências da família *CRM* para verificar a diversidade desses elementos nos três genomas e, assim, entender a dinâmica dos *CRM* nos genomas parentais e no híbrido;

b. Estudar detalhadamente os elementos CR das sequências completas (> 90% de identidade) dos três genomas e eleger as sequências mais representativas de cada agrupamento;

c. Fazer uma análise comparativa dos perfis de hibridização dos retrotransposons centroméricos nos três genomas para compreender a distribuição dos retrotransposons centroméricos nos cromossomos dos parentais e do híbrido;

d. Definir o padrão de bandas CMA/DAPI nas três espécies de *Coffea* para relacionar com os perfis de hibridização e fazer associações entre os CR e as regiões heterocromáticas.

REFERÊNCIAS BIBLIOGRÁFICAS

Bao, W., Zhang, W., Yang, Q., Zhang, Y., Han, B., Gu, M., et al. (2006). Diversity of centromeric repeats in two closely related wild rice species, *Oryza officinalis* and *Oryza rhizomatis*. *Mol. Genet. Genomics* 275(5), 421-430. doi: 10.1007/s00438-006- 0103-2

Bardella, V. B., Da Rosa, J. A. and Vanzela, A. L. L. (2014). Origin and distribution of AT-rich repetitive DNA families in *Triatoma infestans* (Heteroptera). *Infect Genet. Evol.* 23, 106-114. doi: 10.1016/j.meegid.2014.01.035

Belyayev, A., Raskina, O. and Nevo, E. (2001) Chromosomal distribution of reverse transcriptase- containing retroelements in two Triticeae species. *Chromosome Res*. 9, 129-136. doi: 10.1023/A:1009231019833

Bennetzen, J. L. (2000) Transposable element contributions to plant gene and genome evolution. *Plant Mol. Biol.* 42 (1), 251-269. doi: 10.1023/A:1006344508454

Bennetzen, J. L. and Wang, H. (2014) The contributions of transposable elements to the structure, function, and evolution of plant genomes. *Annu. Rev. Plant Biol*. 65, 505-30. doi: 10.1146/annurev-arplant-050213-035811

Brenchley, R., Spannagl, M., Pfeifer, M., Barker, G.L.A., D´ Amore, R., Allen, A. M., et al. (2012) Analysis of the bread wheat genome using whole genome shot- gun sequencing. *Nature* 491, 705-710 doi:10.1038/nature11650

Cai, J., Liu, X., Vanneste, K., Proost, S., Tsai, W.C., Liu, K.W., et al. (2015) The genome sequence of the orchid *Phalaenopsis equestris*. *Nat. Genet.* 47, 65-76. doi: 10.1038/ng.3149

Cenci, A., Combes, M. C. and Lashermes, P. (2012) Genome evolution in diploid and tetraploid *Coffea* species as revealed by comparative analysis of orthologous genome segments. *Plant Mol. Bio.* 78, 135-45. doi: 10.1007/s11103-011-9852-3

Chaparro, C., Gayraud, T., de Souza, R. F., Domingues, D. S., Akaffou, S., Vanzela, A. L. L., et al. (2015). Terminal-repeat retrotransposons with GAG domain in plant

genomes: a new testimony on the complex world of transposable elements. *Genome Biol. Evol.* 7(2), 493-504. doi: 10.1093/gbe/evv001

Clarindo, W. R. and Carvalho, C. R. (2008) First *Coffea arabica* karyogram showing that this species is a true allotetraploid. *Plant Syst. Evol.* 274, 237-241. doi:10.1007/s00606-008-0050-y

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

Clarke, L. (1990) Centromeres of budding and fission yeasts. *Trends Genet.* 6, 150-154. doi: 10.1016/0168-9525(90)90149-Z

Clarke, L. (1998) Centromeres: proteins, protein complexes, and repeated domains at centromeres of simple eukaryotes. *Curr. Opin. Genet. Dev*. 8, 212-218. doi:10.1016/S0959-437X(98)80143-3

Comai, L., Maheshwari, S. and Marimuthu, P. A. (2017) Plant centromeres. *Curr. Opin. Plant. Biol.* 36, 158-167. doi: 10.1016/j.pbi.2017.03.003

Danilova, T.V. and Birchler, J.A. (2008) Integrated cytogenetic map of mitotic metaphase chromosome 9 of maize: resolution, sensitivity, and banding paint development. *Chromosoma*, 117 (4), 345-356. doi: 10.1007/s00412-008-0151-y

Denoeud, F., Carretero-Paulet, L., Dereeper, A., Droc, G., Guyot, R., Pietrella, M., et al. (2014). The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. *Science* 345, 1180-1184. doi: 10.1126/science.1255274

Domingues, D.S., Cruz, G.M.Q., Metcalfe, C.J., Nogueira, F.T.S., Vicentini, R., Alves, C. S., et al. (2012) Analysis of plant LTR-retrotransposons at the fine-scale family level reveals individual molecular patterns. *BMC Genomics* 13, 137. doi: 10.1186/1471-2164-13-137

Du, J., Tian, Z., Hans, C. S., Laten, H. M., Cannon, S. B., Jackson, S. A., et al. (2010). Evolutionary conservation, diversity and specificity of LTR-retrotransposons in

flowering plants: insights from genome-wide analysis and multi-specific comparison. *Plant J.* 63(4), 584-598. doi: 10.1111/j.1365-313X.2010.04263.x

Feng, C., Liu, Y., Su, H., Wang, H., Birchler, J. and Han, F. (2015). Recent advances in plant centromere biology. *Sci. China Life Sci.* 58, 240-245. doi: 10.1007/s11427-015-4818-3

Flavell, A. J., Dunbar, E., Anderson, R., Pearce, S.R., Hartley, R. and Kumar, A. (1992) Ty1-copia group retrotransposons are ubiquitous and heterogeneous in higher plants. *Nucleic Acids Res.*, 20, 3639-3644. doi: 10.1093/nar/20.14.3639

Gao, D., Jiang, N., Wing, R. A., Jiang, J., and Jackson, S. A. (2015). Transposons play an important role in the evolution and diversification of centromeres among closely related species. *Front. Plant Sci.* 6, 216. doi: 10.3389/fpls.2015.00216

Grandbastien, M. A. (2015). LTR-retrotransposons, handy hitchhikers of plant regulation and stress response. *Biochim. Biophys. Acta.* 849(4), 403-16. doi: 10.1016/j.bbagrm.2014.07.017

Hamon, P., Siljak-Yakovlev, S., Srisuwan, S., Robin, O., Poncet, V., Hamon, 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, 291-394. doi:10.1007/s10577-009-9033-2

Hamon, P., Grover C. E., Davis, A. P., Rakotomalala, J. J., Raharimalala, N. E., Albert, V. A. 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: GBS coffee phylogeny and the evolution of caffeine content. *Mol. Phylogenet. Evol*. 109, 351-361. doi: 10.1016/j.ympev.2017.02.009

Han, J., Masonbrink, R.E., Shan, W., Song, F., Zhang, J., Yu, W., et al. (2016). Rapid proliferation and nucleolar organizer targeting centromeric retrotransposons in cotton. *Plant J.* 88, 992-1005. doi: 10.1111/tpj.13309.

Hasterok, R., Wolny, E.I., Hosiawa, M., Kowalczyk, M., Kulak-Ksiazczyk, S., Ksiazczyk, T., et al. (2006) Ribosomal DNA is an effective marker of *Brassica* chromosomes. *Ann Bot.* 97, 205-216. doi:10.1093/aob/mcj031

Heckmann, S., Macas, J., Kumke, K., Fuchs, J., Schubert, V., Ma, L., et al. (2013) The holocentric species *Luzula elegans* shows interplay between centromere and large-scale genome organization. *Plant J.* 73, 555-565. doi: 10.1111/tpj.12054

Hemleben, V., Kovarik, A., Torres-Ruiz, R.A., Volkov, R.A. and Beridze, T. (2007) Plant highly repeated satellite DNA : molecular evolution, distribution and use for identification of hybrids. *Syst. Biodivers.* 5(3), 277-289. doi: 10.1017/S147720000700240X

Henikoff, S. and Dalal, Y. (2005). Centromeric chromatin: what makes it unique? *Curr. Opin. Genet. Dev.* 15, 177-184. doi: 10.1016/j.gde.2005.01.004

Herrera, J. C., Camayo, G., De-La-Torre G., Galeano, N., Salcedo, E., et al. (2013). Identification and chromosomal distribution of Copia-like retrotransposon sequences in the coffee (*Coffea* L.) genome. *Agron. Colomb.* 31(3), 269-278.

Herrera, J.C., Combes, M.C., Anthony, F., Charrier, A. and Lashermes, P. (2002) Introgression into the allotetraploid coffee (*Coffea arabica L.*): segregation and recombination of the *C. canephora* genome in the tetraploid interspecific hybrid (*C. arabica* × *C. canephora*). *Theor. Appl. Genet.* 104, 661-668. doi: 10.1007/s001220100747

Heslop-harrison, J. S. and Schmidt, T. (2007) Plant Nuclear Genome Composition. *E. L. S.* 1-8. doi: 10.1002/9780470015902.a0002014

Heslop-Harrison, J. S. and Schwarzacher, T. (2011). Organisation of the plant genome in chromosomes. *Plant J.* 66, 18-33. doi: 10.1111/j.1365-313X.2011.04544.x

Houben, A., Schroeder-Reiter, E., Nagaki, K., Nasuda, S., Wanner, G., Murata, M., et al. (2007). CENH3 interacts with the centromeric retrotransposon cereba and GC-rich satellites and locates to centromeric substructures in barley. *Chromosoma* 116(3), 275-283. doi: 10.1007/s00412-007- 0102-z

Huang, Y., Luo, L., Hu, X., Yu, F., Yang, Y., Deng, Z., Wu, J., Chen, R. and Zhang, M. (2017) Characterization, genomic organization, sbundance, and chromosomal distribution of Ty1-copia retrotransposons in *Erianthus arundinaceus*. *Front. Plant Sci.* 8, 924. doi: 10.3389/fpls.2017.00924

ICO, International *Coffee* Organization. Disponível em <http://www.ico.org>. 01/01/2018

International Rice Genome Sequencing Project (2005) The map- based sequence of the rice genome. *Nature* 436, 793-800. doi: 10.1038/nature03895

Jiang, J., Birchler, J. A., Parrott, W. P. and Dawe, R. K. (2003) A molecular view of plant centromeres. *Trends Plant Sci.* 8(12), 570-575. doi:10.1016/j.tplants.2003.10.011

Jin, W., Melo, J. R., Nagaki, K., Talbert, P. B., Henikoff, S., Dawe, R. K., et al. (2004). Maize centromeres: organization and functional adaptation in the genetic background of oat. *Plant Cell* 16(3), 571-581. doi: 10.1105/tpc.018937

Jurka, J., Kapitonov, V. V., Kohany, O. and Jurka, M. V. (2007) Repetitive sequences in complex genomes: structure and evolution. *Annu. Rev .Genomics Hum. Genet*. 8, 241-259. doi:10.1146/annurev.genom.8.080706.092416

Kapitonov, V. V. and Jurka, J. (2006) Self-synthesizing DNA transposons in eukaryotes. *Proc. Natl Acad. Sci.* 103, 4540-4545. doi:10.1073/pnas.0600833103

Kato, A., Vega, J. M., Han, F., Lamb, J. C. and Birchler, J. A. (2005) Advances in plant chromosome identification and cytogenetic techniques. *Curr. Opin. Plant Biol*. 8 (2), 148-154. doi:10.1016/j.pbi.2005.01.014

Kumar, A. and Bennetzen, J. (1999) Plant retrotransposons. *Annu. Rev. Genet*. 33, 479-532. doi: 10.1146/annurev.genet.33.1.479

Lashermes, P., Combes, M. C., Trouslot, P. and 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. doi: 10.1007/s001220050500

Lashermes, P., Combes, M.C., Robert, J., Trouslot, P. D'Hont, A. and Anthony F., et al. (1999) Molecular characterization and origin of the *Coffea arabica L*. genome. *Mol. Gen. Genet.* 261, 259-266. doi: 10.1007/s004380050965

Leitch, I. J. and Heslop-Harrison, J. S. (1993) Physical mapping of 4 sites of 5S rDNA sequences and one site of the alpha-amylase-2 gene in barley (*Hordeum vulgare*). *Genome* 36, 517-523. doi: 10.1139/g93-071

Lermontova, I., Sandmann, M., Mascher, M., Schmit, A.C. and Chaboute, M.E. (2015) Centromeric chromatin and its dynamics in plants. *Plant J*. 83, 4-17. doi:10.1111/tpj.12875

Liu, Z., Yue, W., Li, D., Wang, R. R. C., Kong, X., Lu, K., et al. (2008). Structure and dynamics of retrotransposons at wheat centromeres and pericentromeres. *Chromosoma*. 117(5), 445-456. doi: 10.1007/s00412-008-0161-9

Llorens, C., Muñoz-Pomer, A., Bernad, L., Botella, H., and Moya, A. (2009). Network of eukaryotic LTR retroelements beyond phylogenetic trees. *Biol. Direct.* 4, 41. doi: 10.1186/1745-6150- 4-41dynamics

Lombello, R. A. and Pinto-Maglio, C. A. F. (2004a) 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 (1), 11-17. doi: 10.1080/00087114.2004.10589366

Lombello, R. A. and Pinto-Maglio, C. A. F. (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 (2), 138-143. doi: 10.1080/00087114.2004.10589383

Lopes, F. R., Jjingo, D., Da Silva, C. R., Andrade, A. C., Marraccini, P., Teixeira, J. B., et al. (2013). Transcriptional activity, chromosomal distribution and expression effects of transposable elements in *Coffea* genomes. *PloS One*. 8(11), e78931. doi: 10.1371/journal.pone.0078931

Lopes, F., Carazzolle, M. F., Pereira, G. A. G., Colombo, C. A., and Carareto, C. M. A. (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.doi: 10.1007/s00438-008-0319-4

Marco, A. and Marín, I. (2005) Retrovirus-like elements in plants. *Recent Res. Devel. Plant Sci*. 3, 1-10.

Marques, A., Ribeiro, T., Neumann, P., Macas, J., Novak, P., Schubert, V., et al. (2015). Holocentromeres in *Rhynchospora* are associated with genome-wide centromere-specific repeat arrays interspersed amongst euchromatin. *PNAS*. 112, 13633. doi: 10.1073/pnas.1512255112

McClintock, B. (1951) Chromosome organization and genic expression. *Cold Spring Harb. Symp. Quant. Biol*. 16, 13-47 doi: 10.1101/SQB.1951.016.01.004

Mehra, M., Gangwar, I. and Shankar, R. (2015) A deluge of complex repeats: the *Solanum* genome. *PLoS One* 10, 1-38 doi: 10.1371/journal.pone.0133962

Mendes, A. J. T. (1938) Morfologia dos cromossomos de *Coffea* excelsa Chev. *Boletim Técnico*, 56, 1-8.

Mishima, M., Ohmido, N., Fukui, K. and Yahara, T. (2002) Trends in site number change of rDNA loci during polyploid evolution in Sanguisorba (Rosaceae). *Chromosoma* 110, 550-558. doi:10.1007/s00412-001-0175-z

Nagaki, K., Song, J., Stupar, R. M., Parokonny, A. S., Yuan, Q., Ouyang, S., et al. (2003). Molecular and cytological analyses of large tracks of centromeric DNA reveal the structure and evolutionary dynamics of maize centromeres. *Genetics*. 163, 759-770.

Nagaki, K., and Minoru, M. (2005). Characterization of CENH3 and centromere associated DNA sequences in sugarcane. *Chromosome Res.* 13(2), 195-203. doi: 10.1007/s10577-005-0847- 2

Neumann, P., Navrátilová, A., Koblížková, A., Kejnovský, E., Hřibová, E., Hobza, R., et al. (2011). Plant centromeric retrotransposons: a structural and cytogenetic perspective. *Mobile DNA* 2(1), 4. doi: 10.1186/1759-8753- 2-

Piegu, B., Guyot, R., Picault, N., Roulin, A., Saniyal, A., Kim, H., et al. (2006). Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. *Genome Res.* 16(10), 1262-1269. doi: 10.1101/gr.5290206

Plohl, M., Meštrovic, N. and Mravinac, B. (2014) Centromere identity from the DNA point of view. *Chromosoma* 123(4), 313-325. doi:10.1007/s00412-014-0462-0

Prakash, N.S., Combes M. C., Somanna, N. and Lashermes P. (2002) AFLP analysis of introgression in coffee cultivars (*Coffea arabica L.)* derived from a natural interspecific hybrid. *Euphytica* 124, 265-271. doi:10.1023/A:1015736220358

Raina, S. N., Mukai, Y. and Yamamoto, M. (1998) *In situ* hybridization identifies the diploid progenitor species of *Coffea arabica* (*Rubiaceae*). *Theor. Appl. Genet* . 9, 1024-1029. doi: 10.1007/s001220051011

Ribeiro, T., Marques, A., Novák, P., Schubert V., Vanzela, A. L. L., Macas, J., et al.(2016) Centromeric and non-centromeric satellite DNA organisation differs in holocentric *Rhynchospora* species. *Chromosoma* 126, 325-335. doi: 10.1007/s00412-016-0616-3

Ruas, P. M., Ruas, C. F., Rampim, L., Carvalho, V. P., Ruas, E. A. and Sera, T. (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. doi: 10.1590/S1415-47572003000300017

Santini, S., Cavallini, A., Natali, L., Minelli, S., Maggini, F., Cionini, P.G. (2002) Ty1/copia-and Ty3/gypsy-like DNA sequences in *Helianthus* species. *Chromosoma* 111, 192-200. doi:10.1007/s00412-002-0196-2

Santos, F. C., Guyot, R., Do Valle, C. B., Chiari, L., Techio, V. H., Heslop-Harrison, P., et al. (2015). Chromosomal distribution and evolution of abundant retrotransposons in plants: Gypsy elements in diploid and polyploid *Brachiaria* forage grasses. *Chromosome Res.* 23(3), 571-582. doi: 10.1007/s10577-015-9492-6

Schmidt, T. and Heslop Harrison, J.S. (1998) Genomes, genes and junk: the large scale organization of plant chromosomes. *Trends Plant. Sci.* 3, 195-199. doi:10.1016/S1360-1385(98)01223-0

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternk, S., et al. (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science* 326, 1112-1116. doi: 10.1126/science.1178534

Sharma, A. and Presting, G. G. (2014). Evolution of centromeric retrotransposons in grasses. *Genome Biol. Evol.* 6(6), 1335-1352. doi: 10.1093/gbe/evu096

Taketa, S., Harrison, G.E. and Heslop-Harrison, J. S. (1999) Comparative physical mapping of the 5S and 18S-25S rDNA in nine wild *Hordeum* species and cytotypes. *Theor. Appl .Genet.* 98, 1-9. doi: 10.1007/s001220051033

Tenaillon, M. I., Hufford, M. B., Gaut, B. S. and Ross-Ibarra, J. (2011). Genome size and transposable element content as determined by high-throughput sequencing in maize and *Zea luxurians*. *Genome Biol. Evol*. 3, 219-229. doi: 10.1093/gbe/evr008

Tesfaye, K., Borsch, T., Kim Govers, K. and Bekele E. (2007) Characterization of *Coffea* chloroplast microsatellites and evidence for the recent divergence of *C. arabica* and *C. eugenioides* chloroplast genomes. *Genome* 50 (12), 1112-29. doi: 10.1139/G07-088

The French–Italian Public Consortium for Grapevine Genome Characterization (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*. doi:10.1038/nature6148

Wang, G., Zhang, X. and Jin, W. (2009) An overview of plant centromeres. *JGG.* 36, 529-537. doi: 10.1016/S1673-8527(08)60144-7

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., et al. (2007). A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* 8, 973-982. doi: 10.1038/nrg2165

Yu, Q., Guyot, R., de Kochko, A., Byers, A., Navajas-Pérez, R., Langston, B. J., et al. (2011). Micro-collinearity and genome evolution in the vicinity of an ethylene receptor gene of cultivated diploid and allotetraploid coffee species (*Coffea). Plant J*. 67(2), 305-317. doi: 10.1111/j.1365-313X.2011.04590.x

Yuyama, P. M., Pereira, L. F. P., Santos, T. B., Sera, T., Vilas-Boas, L. A., Lopes, F. R., et al. (2012). FISH using a gag-like fragment probe reveals a common Ty3-Gypsy-like retrotransposon in genome of *Coffea* species. *Genome.* 55(12), 825-833. doi: 10.1139/gen-2012- 0081

Zakrzewski, F., Wenke, T., Holtgräwe, D., Weisshaar B. and Schmidt, T. (2010) Analysis of a c0t-1 library enables the targeted identification of minisatellite and satellite families in *Beta vulgaris*. *BMC Plant Biol.*10, 8. doi: 10.1186/1471-2229-10-8

Zhang, Y., Huang, Y., Zhang, L., Li, Y., Lu, T., Lu, Y., et al. (2004). Structural features of the rice chromosome 4 centromere. *Nucleic Acids Res.* 32(6), 2023-2030. doi: 10.1093/nar/gkh521

Zhong, C. X., Marshall, J. B., Topp, C., Mroczek, R., Kato, A., Nagaki, K., et al. (2002). Centromeric retroelements and satellites interact with maize kinetochore protein CENH3. *Plant Cell* 14(11), 2825-2836. doi: 10.1105/tpc.006106

Manuscrito submetido para a revista Frontiers in Plant Science.

Structure and distribution of centromeric retrotransposons at diploid and allotetraploid *Coffea* centromeric and pericentromeric regions

Abstract

Centromeric regions of plants are generally composed of large array of satellites from a specific lineage of *Gypsy* LTR-retrotransposons, called Centromeric Retrotransposons (CR). Repeated sequences interact with a specific H3 histone, playing a crucial function on the kinetochore formation. To study the structure and composition of centromeric regions in the genus *Coffea*, we annotated and classified into ten distinct families Centromeric Retrotransposons sequences (called hereafter CRC) from *C. arabica* genome and its two diploid ancestors: *Coffea canephora* and *C. eugenioides*. Ten distinct CRC (Centromeric Retrotransposons in *Coffea*) families were found. The sequence mapping and FISH experiments of CRC Reverse Transcriptase domains in *C. canephora*, *C. eugenioides* and *C. arabica* clearly indicate a strong and specific targeting mainly onto proximal chromosome regions, which can be associated also with heterochromatin. PacBio genome sequence analyses of putative centromeric regions on *C. arabica* and *C. canephora* chromosomes showed an exceptional density of one family of CRC elements, and the complete absence of satellite arrays, contrasting with usual structure of plant centromeres. Altogether, our data suggest a specific centromere organization in *Coffea*, contrasting with other plant genomes.

Key words: Coffee, *CRM* lineages, FISH, *Gypsy*, pseudochromosomes, proximal chromosome regions, centromeres.

Introduction

LTR-retrotransposons pertain to the Class I of Transposable Elements (TEs), and they move via the synthesis of an intermediate RNA using 'copy and paste' mechanisms (Wicker et al., 2007). Due to their mobility, LTR-retrotransposons are the most abundant TEs (Grandbastien, 2015). They contribute to the variation of genome size and structure observed in plants (Piegu et al., 2006; Heslop-Harrison and Schwarzacher, 2011; Tenaillon et al., 2011).

LTR-retrotransposons are classified into *Copia* and *Gypsy* superfamilies according to their coding domain internal organization (Schnable et al., 2009; Gao et al., 2012; Bennetzen and Wang, 2014). Each *Copia* and *Gypsy* superfamily is sub-classified into lineages and families (Wicker et al., 2007), according to coding region similarities and overall structures (Llorens et al., 2009). For plant genomes, *Copia* is sub-classified into *Tork, Retrofit, Oryco, SIRE* and *Bianca*, while *Gypsy* is sub-classified into *TAT*, *Athila*, *Galadriel, Reina, Del* and *CRM* (Llorens et al., 2009; Llorens et al., 2011), based on Reverse-Transcriptase (RT) domain phylogenetic analyses. *Gypsy* lineages are also grouped into different branches according to the presence of a chromodomain; grouping together *Galadriel, Reina, Del* and *CRM* lineages into the Chromovirus branch.

Copia and *Gypsy* superfamilies can be found distributed in blocks or dispersed along plant chromosomes (Lopes et al., 2013, Santos et al., 2015; Zhang et al., 2017). One notable exception is the Centromeric Retrotransposon lineage of Chromovirus (*CRM* or Centromeric Retrotransposon of Maize), which appears located preferentially into proximal chromosome regions or "centromeric regions" (Nagaki et al., 2005; Bao et al., 2006; Liu et al., 2008; Du et al., 2010; Sharma and Presting, 2014). *CRMs* carry heterogeneous domains at the C-terminus of the integrase that may be linked to their chromosomal distribution. A chromodomain (CHRomatin Organization MOdifer

domain) or a targeting domain called CR motif were identified (Houben et al., 2007; Neumann et al., 2011). These domains are probably able to interact with the CENH3 protein, suggesting that Centromeric Retrotransposons (CR) participate in centromere function. Plant centromeric regions can be composed of large arrays of CR elements inserted into specific satellite DNA (Cheng et al., 2002; Houben et al., 2007; Santos et al., 2015; Marques et al., 2015). Although relatively few centromeric regions have been studied in plants, especially due to difficulties to sequence and assemble of regions with a high content of repetitive sequences, Neumann et al. (2011) separated CR elements into three groups according to their properties and chromosomal distribution: Group A carrying a CR motif and Group B lacking any targeting domain, both localized in centromeric regions; and Group C containing a chromodomain and dispersed along chromosomes.

The *Coffea* (Rubieaceae) comprises 125 species (Hamon et al., 2017). All species are diploids, except *Coffea arabica* ($2n = 4x = 44$), that arose from a recent hybridization between *C. canephora* and *C. eugenioides* (Lashermes et al., 1999; Yu et al., 2011). The recent sequencing of *C. canephora* genome revealed an important contribution of transposable elements (>50%). Most of them fell into the LTR-retrotransposons order (Denoeud et al., 2014). Several international sequencing initiatives are targeting the *C. arabica* genome using Pacific Biosciences (PacBio) single molecule sequencing (Mueller et al., 2015). This technique, allowing the sequencing of complex regions with a high content of repeated sequences, offers the opportunity to study the composition and organization of centromeric regions. In this study, we identified and compared 10 families of Centromeric Retrotransposons in the forthcoming PacBio genomes of *C. canephora*, *C. eugenioides* and *C. arabica*. In situ hybridization using conserved RT probes showed CRCs located in proximal and interstitial chromosome regions. Finally, annotation and

comparison of centromeric region rich in CRC elements revealed dynamic changes targeting LTR retrotransposons, but also the complete absence of tandem repeats usually associated with CRC elements.

Material and methods

Genome sequencing

Genomic DNA was extracted from leaves using DNeasy Plant Maxi Qiagen Kit. For long read sequencing, 20 Kb libraries were prepared following Pacific Biosciences (PacBio) protocol and Blupippin size selection. Sequencing was performed on the PacBio RSII platform, and specifications are described in **Supplemental data 1**. For short read sequencing, libraries were prepared with the KAPA HyperPlus kits, following manufacturer recommendation and sequenced on Illumina HiSeq2500 using PE flow cells and V4 chemistry. Genomes were assembled using Falcon and Falcon unzip from Pacific Bioscience (https://github.com/PacificBiosciences/FALCON).

Organism	**sequencing chemistry**	**collection protocol [min]**	**SMRT cell nb**	**output [Gb]**	**Coverage**	**PF read nb**	**PF Polymerase Read N50 [bp]**
Coffea arabica	P4/C2 and P5/C3	180	174	72	56x	7,246,499	10,610
Coffea canephora	P6/C4	240	65	55	77x	3,721,573	16,988
Coffea eugenoides	P6/C4	240	60	39	58x	3,219,296	15,481

Supplemental data 1. Description of genome sequencing specification.

In silico analyses

Genomes of *C. canephora* (DH200-94-V.2), *C. eugenioides* (BU-A) and *C. arabica* (accession Et39), were kindly provided by the Arabica Coffee Genome Consortium (ACGC) with the single molecule real-time (SMRT, Pacific Biosciences—PacBio). The three genomes were sequenced using the long-read Pacific Bioscience technology (Mueller et al., 2015). *Coffea canephora* genome assembly was finished using both Bionano genome mapping and Dovetail Hi-C scaffolding technologies (ACGC, unpublished results).

Transposable element annotations and analyses

Sequenced genomes served as source for searching and comparing LTR-retrotransposons using the LTR_STRUC (McCarthy and McDonald, 2003). Putative retrotransposons sequences were classified into *Gypsy* and *Copia* superfamilies according to their similarity against the Gypsy Database protein domains (http://www.gydb.org/index.php/Main_Page) as implemented in the *Impactor* program (Orozco et al. unpublished. Available upon request). Putative reverse transcriptase (RT) domain from the *Gypsy* superfamilies were identified using BLASTX (Altschul et al., 1997) and extracted and translated into amino acids using Genewise (Birney et al., 2004) with a minimum length of 150 residues as in Guyot et al. (2016). For each coffee genome, RT domains from *Gypsy* LTR-RTs were aligned using MUSCLE (Edgar, 2004) with RT reference domains from the Gypsy Database. Aligned sequences were used to construct a bootstrapped neighbor joining phylogenetic tree (1,000 bootstrap) with ClustalW (Thompson et al., 1994), edited using FigTree (http://tree.bio.ed.ac.uk/software/figtree/).

The coffee sequences from the *CRM* lineage and called hereafter CRC (Centromeric Retrotransposons of *Coffea*) sequences were identified from the NJ tree. These sequences were sub-classified into groups according to tree conformation and bootstrap values.

Groups were validated by alignments using dotter (Sonnhammer and Durbin, 1995), stretcher (EMBOSS) and plotcon (EMBOSS). LTR sequences with 99% identity based on LTR_STRUC were annotated using Artemis (Rutherford et al., 2000). Complete (i.e. a LTR-retrotransposon containing both LTR domains) and putative autonomous (i.e. a LTR-retrotransposon containing all coding domain involved in its mobility) elements were compared and grouped with the Mauve tool (http://darlinglab.org/mauve/mauve.html). Non-autonomous elements were classified into TRIM, LARD and TR-GAG according to their length and domains as in Chaparro et al. (2015) and implemented in the *Parallan* program (Orozco et al. unpublished). A representative element of each group was submitted to GenBank under the following accession: A MG242426; B MG242427; C MG242428; D MG242429; E MG242430; F MG242431; G MG242432; H MG242433; Y MG242434; X MG242435.

In silico estimation of CRC elements copy number and distribution

Assessment of the CRC elements copy number in *C. canephora*, *C. arabica* and *C. eugenioides* PacBio sequences was done as in Dupeyron et al. (2017). Briefly, each representative copy of CRC groups was used for similarity searches against genomes using Censor (http://www.girinst.org/downloads/software/censor/). Copies are sorted according to their completeness and percentage of similarity when compared to the representative copy. Insertion times of selected LTR-RT were estimated as proposed by SanMiguel et al. (1998) and Guyot et al. (2016), with a substitution rate of 1.3×10^{-8}, established by Ma and Bennetzen (2004). The distribution of RT domains was carried out using RepeatMasker (-div 20 option) while the distribution of complete elements, LTR and non-autonomous elements was performed using Censor with a minimum of 80% of nucleotides identity and 80% of sequence coverage.

The centromeric region annotation was performed using RepeatMasker (-div 20 option)

and edited with Artemis, and transposable elements density along genomic sequences was carried out using DensityMap (Guizard et al., 2016).

Plant Materials, DNA extraction and probes production

Seedlings of *C. arabica*, *C. canephora* and *C. eugenioides* were obtained from the Agronomic Institute of Paraná (IAPAR), Londrina, Paraná, Brazil, cultivated in pots in the green house of the Laboratory of Cytogenetics and Plant Diversity, State University of Londrina, Brazil. DNA extraction was performed as described by Romano et al. (1999). Quickly, young leaves were collected, macerated in liquid nitrogen and treated with CTAB extraction buffer. DNA was purified with phenol:chloroform (1:1, v:v) and chloroform:isoamyl alcohol (24:1, v:v) and precipitated in absolute ethanol. DNA concentration was estimated using a NanoDrop 2000 Spectrophotometer (Thermo Scientific). Primers were designed using OligoPerfect™Designer (http://tools.lifetechnologies.com). A conserved region located in the predicted Reverse Transcriptase (RT) coding region of each CRC group was amplified by PCR using a pair of RT primers (Forward: 5'ACTGTCGGGCTGTAAATGCT; Reverse: 5'CTGCGAACTCACGACATAGC). Reactions were done using *C. arabica*, *C. canephora* and *C. eugenioides* genomic DNA as template, in a mix composed by 0.5 μL Taq Polymerase (5 U/μL), 2.5 μL 10× buffer, 2.5 μL $MgCl_2$ (50 mM), 1 μL of dNTP (10 mM), 1 μL of each primer at 10 mM and H_2O, in a final volume of 25 μL. Reactions were checked with 1% agarose gel electrophoresis. Probes were obtained by PCR, using the product of a first PCR as template, in a new reaction containing dGTP (25%), dCTP (25%), dTTP (25%), dATP (17.5%) and Cy3-dUTP (7.5%).

Cytogenetic analyses

Mitotic chromosomes were obtained from root tips treated with a saturated solution of paradichlorobenzene (PDB) for 1h at room temperature plus 23 h at 14 °C. Samples were fixed in a fresh solution of methanol: acetic acid (3:1, v:v) for 24 h, and stored at -20 °C, or used immediately. Root-tips were softened in 2% cellulase plus 20% pectinase (v:v), both Sigma, at 37 °C for 5 h, and squashed in a drop of 60% acetic acid. The cover slips were removed after freezing in liquid nitrogen, slides were air dried and used in FISH or C-CMA/DAPI banding procedures.

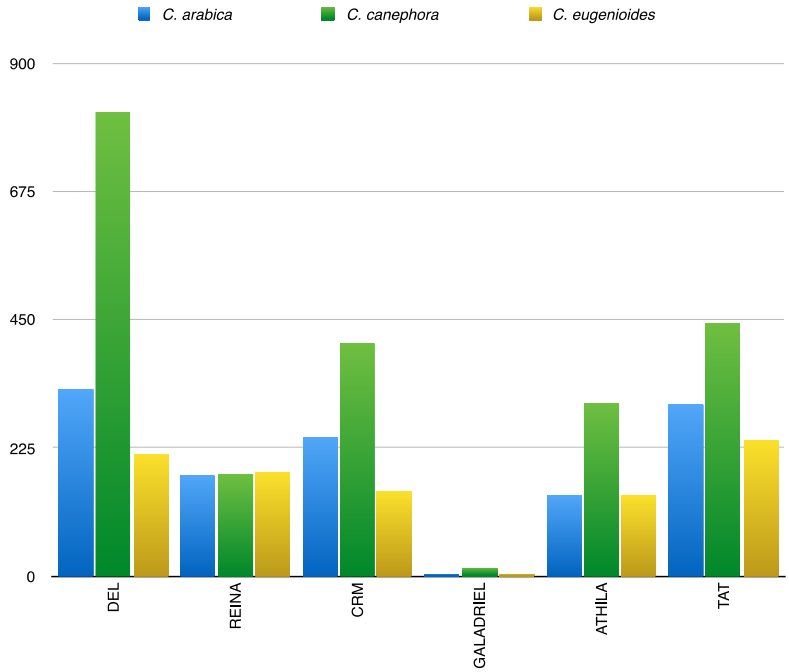
For FISH, a mixture of 30 µL containing 100% formamide (15 µL), 50% polyethylene glycol (6 µL), 20× SSC (3 µL), 100 ng calf thymus DNA (1 µL), 10% SDS (1 µL) and 100 ng probes (4 µL), was treated at 70 °C for 10 min, placed on ice and immediately applied to the samples. Denaturation/hybridization was performed at 95 °C, 50 °C and 38 °C, ten minutes each, followed by 37 °C overnight in a humidified chamber. Post-hybridization washes were carried out in SSC buffer with about 70% stringency, mounted in 23 µL antifade solution (90% glycerol, 2.3% DABCO, 2% 20 mM Tris–HCl, pH 8.0, plus 1 µL of 2 µg/mL DAPI and 1 µL of 2.5 mM $MgCl_2$).

Chromosome banding was done using three days aged slides incubated in a solution of 45% acetic acid, 5% barium hydroxide and 2× SSC, pH 7.0 (Schwarzacher et al., 1980, with modifications). Samples were stained with 0.5 mg/mL CMA_3 for 1.5 h and 2 mg/mL DAPI for 30 min, and finally stained with a medium composed of glycerol/McIlvaine buffer (pH 7.0) 1:1 plus 2.5 mM $MgCl_2$. FISH and C-CMA/DAPI chromosome images were acquired in gray-scale mode using a Leica DM4500B microscope, equipped with a Leica DFC300FX camera, and overlapped with blue for DAPI, greenish-yellow for CMA and red for Cy3, and processed using the Leica LAS software. Images were optimized for contrast and brightness using the GIMP 2.8 Image Editor.

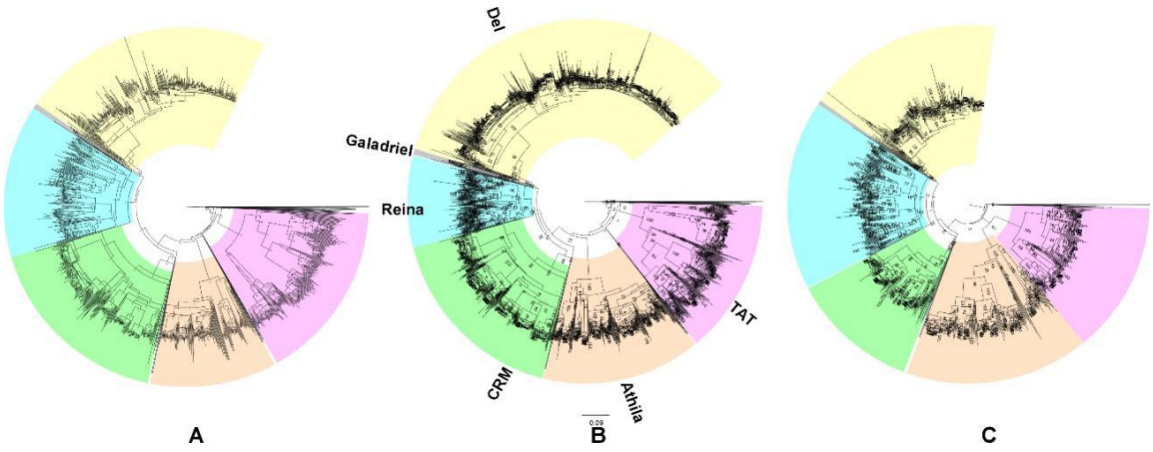
Results

T*he Gypsy superfamily and the CRM lineage in coffee genomes*

The search for complete LTR-retrotransposons sequences in *C. canephora*, *C. eugenioides* and *C. arabica* allowed to recognize 7,195, 3,590, and 3,877 elements, respectively. These were predicted and classified into 1,021 *Copia* and 2,222 *Gypsy* (*C. canephora*), 668 *Copia* and 950 *Gypsy* (*C. eugenioides*) and 743 *Copia* and 1226 *Gypsy* (*C. arabica*). The remaining predicted elements were identified into non-autonomous LTR-retrotransposons or into unclassified autonomous elements according to similarities to GAG-POL regions available at the Gypsy Database. For the *Gypsy* superfamily, the LTR-retrotransposon lineages (*Del, Galadriel, Reina, CRM, Athila* and *TAT*), were found in the three *Coffea* genomes, using a BLAST based analysis and a RT based phylogenetic analysis (**Supplemental data 2 and 3**), and the *CRM* lineage was particularly analyzed. *CRM* lineage represented 499, 223 and 262 of complete annotated elements in *C. canephora, C. eugenioides and C. arabica* genomes, respectively. Manual inspection revealed that 367 (73.55%), 124 (55.61%) and 113 (43.13%) elements were found complete for *C. canephora, C. eugenioides and C. arabica*, respectively, since no large deletion affected these sequences. The RT amino acid sequences of the *CRM* lineage from the three coffee species were grouped together, aligned and displayed with a N.J. phylogenetic tree (**Supplemental data 4**). Ten phylogenetic groups were defined according to the structure and similarity of these domains (**Table 1 and Supplemental data 5**). The Centromeric Retrotransposons of *Coffea* were grouped and named here as follow A, B, C, E, D, F, G, H, X and Y.

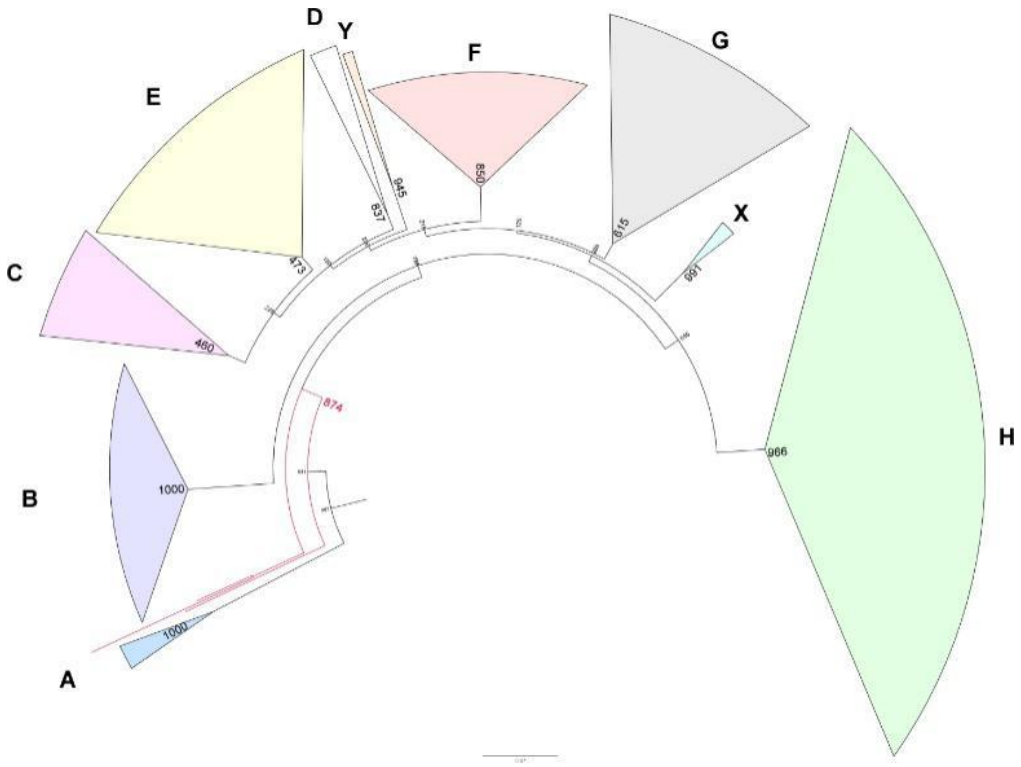


Supplemental data 2. Number of LTR-retrotransposons elements in the different *Gypsy* lineages in *Coffea arabica*, *C. canephora* and *C. eugenioides* as detected by LTR_STRUC.



Supplemental data 3. RT-based phylogenetic analysis of *Gypsy* LTR-retrotransposons predicted in *Coffea arabica* (A), *C. canephora* (B) and *C. eugenioides* (C) identified by LTR_STRUC. The names of *Gypsy* lineages are indicated. Phylogenetic trees were based

on protein alignments of Reverse Transcriptase domains. 1,226, 2,222 and 950 recovered domains were used for *C. arabica*, *C. canephora* and *C. eugenioides*.



Supplemental data 4. NJ Phylogenetic tree of RT domains from 604 autonomous CRC elements from *Coffea canephora*, *C. eugenioides* and *C. arabica*.

Colors represent the 10 CRC groups as follow: A, B, C, D, E, F, G, H, Y and X. In red are represented the branch of representative *CRM* RT domains from the Gypsy db (*CRM*, Beetle1 and Cereba). Bootstraps were indicated.

Supplemental data 5. Distribution of CRC groups on the *Coffea eugenioides*, *C. canephora* and *C. arabica* predicted complete elements by LTR_STRUC.

Genomes	Absolute numbers and CRC groups									
	A	**B**	**Y**	**C**	**E**	**D**	**F**	**G**	**X**	**H**
CRcc	5	68	0	16	53	6	43	53	0	91
CRce	2	20	0	11	10	0	7	0	2	31
CRca	0	1	2	2	15	0	21	8	3	49
Genomes	**Relative numbers (%) and CRC groups**									
	A	**B**	**Y**	**C**	**E**	**D**	**F**	**G**	**X**	**H**
CRcc	1.4	20.3	0	4.7	15.8	1.7	12.8	15.8	0	27.1
CRce	2.4	24.1	0	13.2	12.0	0	8.4	0	2.41	37.3
CRca	0	0.9	1.9	1.9	14.8	0	20.7	7.9	2.9	48.5

The following letters: A. B. C. D. E. F. G. H. X and Y correspond to CRC groups, as for instance: CRcc_group_A. CRcc = centromeric retrotransposons of *C. canephora*; CRce = centromeric retrotransposons of *C. eugenioides*; CRca = centromeric retrotransposons of *C. arabica*.

Table 1. Matrix of RT domain identity between CRC groups in *Coffea eugenioides*, *C. canephora* and *C. arabica.*

	CR Groups								
	C. canephora	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**
C. eugenioides	**X**	48 %	57 %	61 %	57 %	60 %	61 %	64 %	56 %
	A	**88 %**	49 %	48 %	44 %	48 %	47 %	47 %	45 %
	B	48 %	**91 %**	58 %	54 %	57 %	58 %	58 %	55 %
	C	45 %	56 %	**82 %**	57 %	59 %	58 %	59 %	57 %
	E	49 %	57 %	60 %	59 %	**92 %**	61 %	61 %	57 %
	F	47 %	57 %	60 %	59 %	61 %	**93 %**	61 %	59 %
	H	46 %	55 %	58 %	57 %	58 %	60 %	59 %	**80 %**
	C. arabica	**X**	**Y**	**B**	**C**	**E**	**F**	**G**	**H**
C. eugenioides	**X**	**87 %**	59 %	58 %	60 %	60 %	60 %	64 %	57 %
	A	46 %	47 %	48 %	47 %	47 %	47 %	47 %	44 %
	B	57 %	57 %	**97 %**	57 %	57 %	57 %	58 %	54 %
	C	59 %	59 %	56 %	**84 %**	60 %	57 %	60 %	57 %
	E	59 %	60 %	57 %	60 %	**93 %**	61 %	61 %	57 %
	F	60 %	59 %	58 %	60 %	61 %	**90 %**	61 %	59 %
	H	58 %	56 %	55 %	57 %	58 %	59 %	59 %	**80 %**
	C. arabica	**Y**	**X**	**B**	**C**	**E**	**F**	**G**	**H**
C. canephora	**A**	47 %	47 %	49 %	48 %	48 %	48 %	47 %	45 %
	B	56 %	57 %	**91 %**	58 %	58 %	57 %	58 %	54 %
	C	61 %	59 %	58 %	**93 %**	60 %	60 %	60 %	56 %
	D	57 %	58 %	54 %	56 %	60 %	57 %	58 %	57 %
	E	60 %	60 %	58 %	60 %	**97 %**	61 %	61 %	57 %
	F	59 %	60 %	58 %	59 %	61 %	**94 %**	61 %	59 %
	G	60 %	63 %	58 %	60 %	61 %	60 %	**98 %**	58 %
	H	56 %	57 %	55 %	56 %	57 %	58 %	58 %	**92 %**

The letters A, B, C, D, E, F, G, H, X and Y correspond to CRC groups, as defined by the phylogenetic analysis. Values highlighted in grey represent the highest percentage of identity observed between groups.

About 60 CRC sequences per genome, from the different groups, and showing >99% of nucleotide identity between both LTR of the same element were carefully annotated and compared (**Figure 1**). Only elements from the A group presented a chromodomain, with zinc finger/HHCC motif at their C-terminus downstream the INT region, while elements from other groups exhibited a CR motif (**Figure 1B**) at their C-terminal regions, and a poly-A motif upstream the GAG region (data not shown). Autonomous elements of each group showed a variable length from 5,971 bp (Group_A) to 8,088 bp (Group_D), and a LTR size from 661 bp (Group_F) to 781 bp (Group_Y).

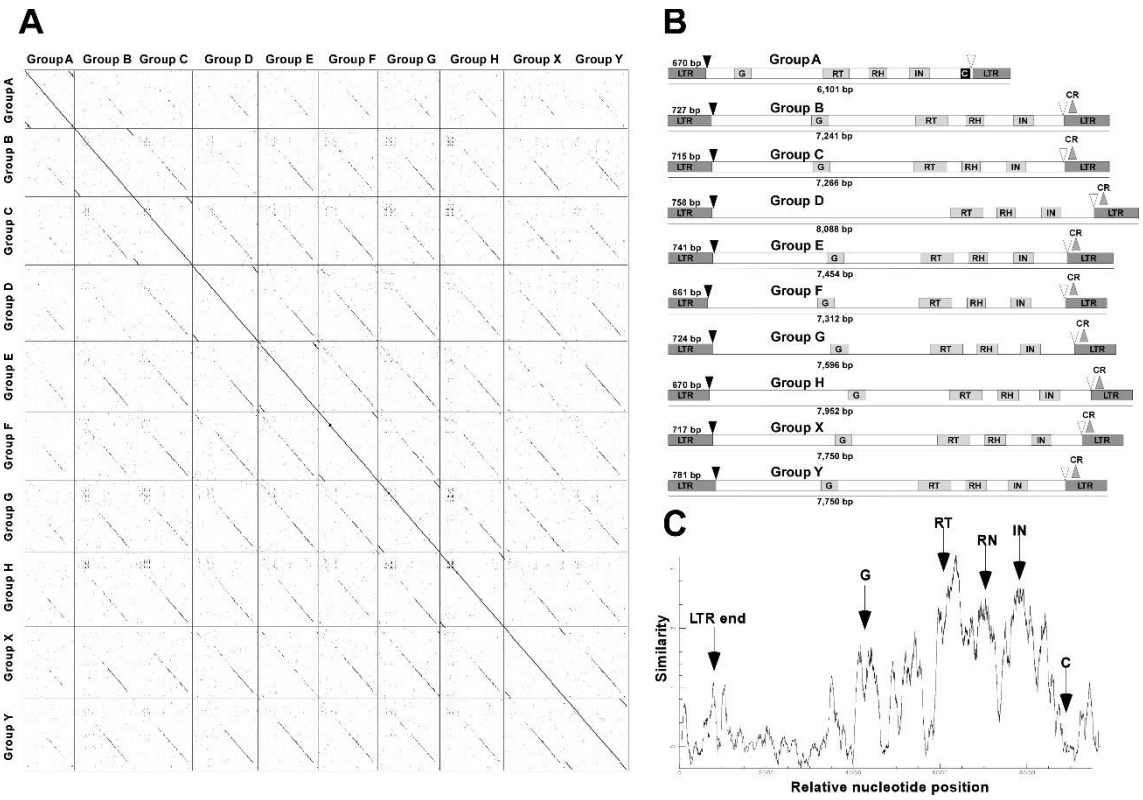


Figure 1. Structure and conservation of the *Gypsy* CRC LTR-retrotransposons in *Coffea arabica*, *C. canephora* and *C. eugenioides*.

A. Dotter alignments between the 10 groups of CRC found by LTR_STRUC against themselves.

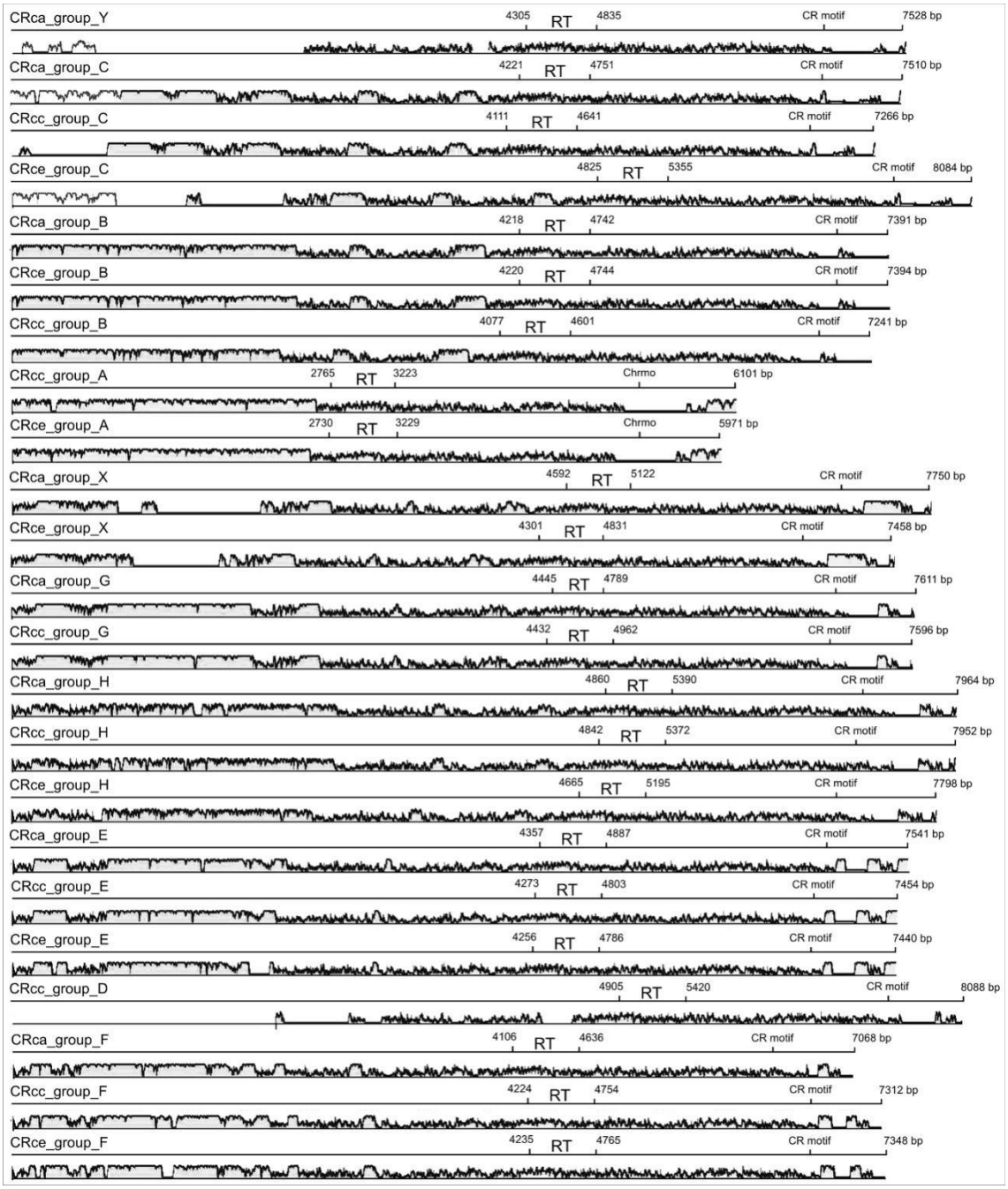
B. Structural features of CRC groups. LTR: Long Terminal Repeats; G: GAG domain, RT: Reverse Transcriptase; RH: RNAse H; IN; Integrase; C: Chromodomain, CR: CR motif. The dark arrows indicate the PBS sites while the white arrows indicate the PPT sites.

C. Nucleotide similarity plot with the 10 groups of CRC. The positions of the different domains are indicated.

The alignment of complete elements into a matrix of nucleotides comparison showed discontinuous lines between groups, suggesting interrupted conservation along the different CRCs (**Figure 1A**). This discontinuous similarity was also confirmed with a

nucleotide similarity plot of the full-length sequences of the 10 CRC groups (**Figure 1C**). The RT domain comparison at the nucleotide level showed a high conservation among elements within each group, independent of the species they are issued (from 80 to 98%), and a distant conservation between elements of different groups, i.e. from 45 to 64% (**Table 1**). These results suggest that CRCs are distributed among different families in the *Coffea* genus.

The additional comparison of elements using the Mauve software revealed different conservation status along and among sequences (**Supplemental data 6**), and confirmed the existence of these ten CRC groups. These findings showed that phylogenetically close groups (**Supplemental data 4**) presented also a similar general structure (**Figure 1**), such as those that can be observed in the groups F and G (**Tabela 1, Figures 1, Supplemental data 4).** The grouping obtained after the Mauve analysis was consistent with the grouping using the reverse transcriptase sequences and support that this similarity seems to be higher within each group than between the different CRC groups (**Supplemental data 6**). It is important to mention also that only *C. arabica* and *C. canephora* exhibited species-specific CRC. The group D appeared only in *C. canephora*, and group Y was observed only in *C. arabica*. The group A was found into parental genomes, and the group G occurred in *C. canephora* and *C. arabica.* The group X, differently, appeared in *C. eugenioides* and *C. arabica*, and members of the groups B, C, E, F and H were common in these three genomes (**Supplemental data 6**).

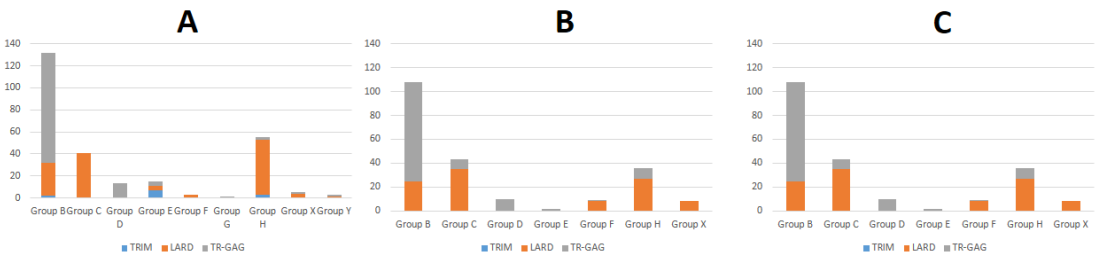


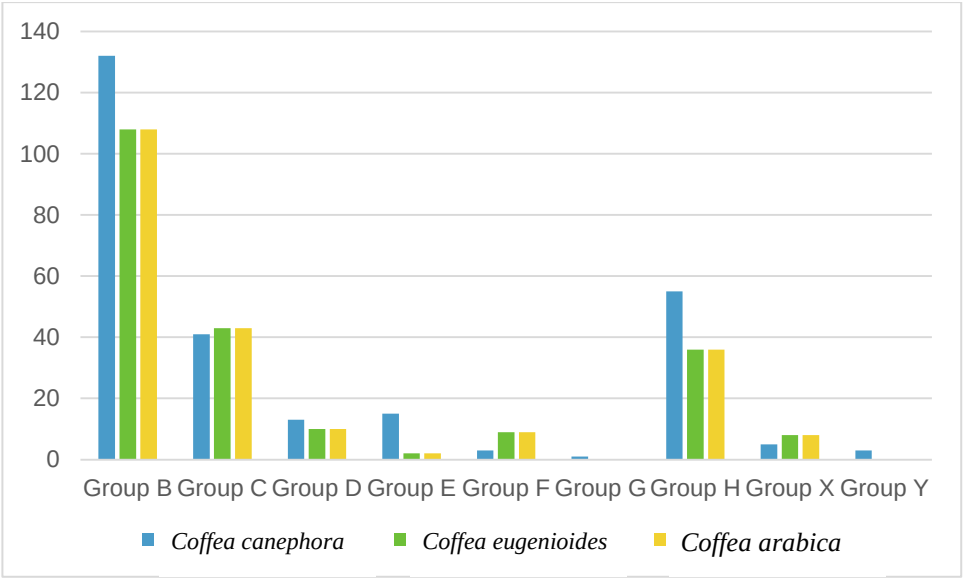
Supplemental data 6. Structural analysis of CRC elements. The A group showed a chromodomain (Chmo) positioned downstream of terminal INT regions. In members of groups B, C, D, E, F, G, H, X and Y the CR motif was positioned downstream of terminal INT regions.

Non-autonomous CRC elements in Coffea

Non-autonomous CRC elements, lacking any coding regions as seen in TRIMs (Terminal

Repeat in Miniature) or LARDs (Large Retrotransposon Derivative, or lacking the POL polyprotein region as in TR-GAGs, were also identified (Chaparro et al., 2015). CRC group alignments (80% identity cutoff) against the putative non-autonomous elements exhibited different structures, such as TRIMs (only in *C. canephora*), LARDs and TR-GAGs. The counting showed 268, 216 and 216 putative non-autonomous CRC for *C. canephora*, *C. eugenioides* and *C. arabica*, respectively (**Supplemental data 7**). Among them the group B (mainly TR-GAG elements), the H (mainly LARD elements) and the group C showed the highest number of copies, whatever the genome analyzed. Only the chromodomain of group A did not show similarity to any non-autonomous element.





Supplemental data 7. Copy number of putative non-autonomous CRC elements (TRIM, LARD and TR-GAG) in *Coffea arabica*, *C. canephora and C. eugenioides*.

Up: Copy number of non-autonomous CRC elements for each group.

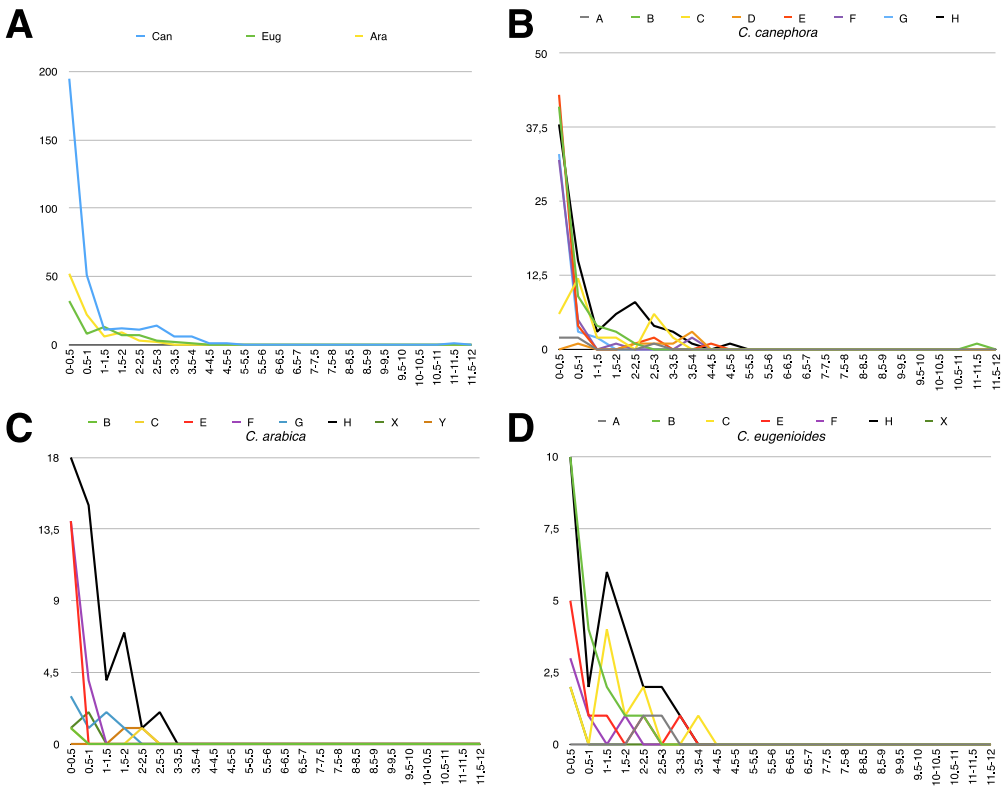
Down: Copy number of each type of non-autonomous elements for each CRC groups A) *C. canephora* B) *C. eugenoides* and C) *C. arabica.*

***In silico* copy number estimation and insertion time of 10 CRC families**

A total copy number of 359, 278 and 473 CRC elements (with >80% of both coverage and identity) were found in *C. canephora*, *C. eugenioides* and *C. arabica*, respectively. Beside conserved copies, fragmented copies (with >10% of coverage and >80% of identity) represented 2,055, 2,064 and 3,478 CRC elements in *C. canephora*, *C. eugenioides* and *C. arabica*, respectively (**Table 2**). For the three species, elements from the groups H and B outnumber the other groups for complete (80-80) and fragmented copies (80-10). The allotetraploid genome of *C. arabica* contains, as expected, the highest copy number when compared to the diploid genomes of *C. canephora* and *C. eugenioides*. The nucleotide divergence and relative insertion time of complete CRC copies suggest a relatively recent insertion, or a high conservation of the whole sequences with a similar pattern in *C. arabica*, *C. eugenioides* and *C. canephora* (**Supplemental data 8A**). For each CRC group, three peaks of copy number accumulation were observed for the H group in *C. canephora, C. eugenioides* and *C. arabica*, while for the C group four and two peaks were noted for *C. eugenioides*, and for *C. canephora* and *C. arabica* (**Supplemental data 8B, 8C and 8D**). Other and successive small peaks of copy number accumulation were observed for the E group, for example. This result suggested that the insertions of CRC are relatively recent, but that ancient activities may be detected, particularly for the group H.

Table 2. Estimation of the copy numbers (autonomous and non-autonomous) of CRC elements in the *Coffea canephora*, *C. eugenioides* and *C. arabica* genome sequences.

	C. canephora Copies (80-80)	*C. canephora* Partial copies (80-10)	*C. eugenioides* Copies (80-80)	*C. eugenioides* Partial copies (80-10)	*C. arabica* Copies (80-80)	*C. arabica* Partial copies (80-10)
Group A	8	85	7	103	13	156
Group B	81	841	66	705	121	1149
Group C	18	86	16	202	28	303
Group D	6	63	19	96	18	164
Group E	49	259	39	188	50	476
Group F	60	144	29	148	63	265
Group G	47	90	3	55	20	153
Group H	84	412	88	460	142	674
Group X	1	13	4	0	7	17
Group Y	5	62	7	107	11	121
Total	**359**	**2055**	**278**	**2064**	**473**	**3478**



Supplemental data 8. Estimation of insertion times of CRC groups in *Coffea arabica C. canephora and C. eugenioides*.

A. Insertion times of all CRC groups. B. Insertion times of each group in *C. canephora*. C. Insertion times of each group in *C. arabica.* D. Insertion times of each group in *C. eugenioides.* Insertion times were estimated using a substitution rate of $1.3x10^{-8}$ (Ma and Bennetzen 2004).

The distribution of CRC RT sequences along the *C. canephora* pseudochromosomes (**Figure 2**) showed that for some of them there is a clear accumulation of RT sequences in the central regions (pseudochromosomes 1, 2, 4, 5, 6, 8, 9 and 10). For the others, RT sequences were less concentrated, exhibiting a dispersed pattern, such as in the pseudochromosomes 3, 7 and 11. When we compare the distribution these sequences of each CRC group along pseudochromosomes, it is possible to note that only the groups E and H showed a clear accumulation into median regions **(Supplemental data 9**).

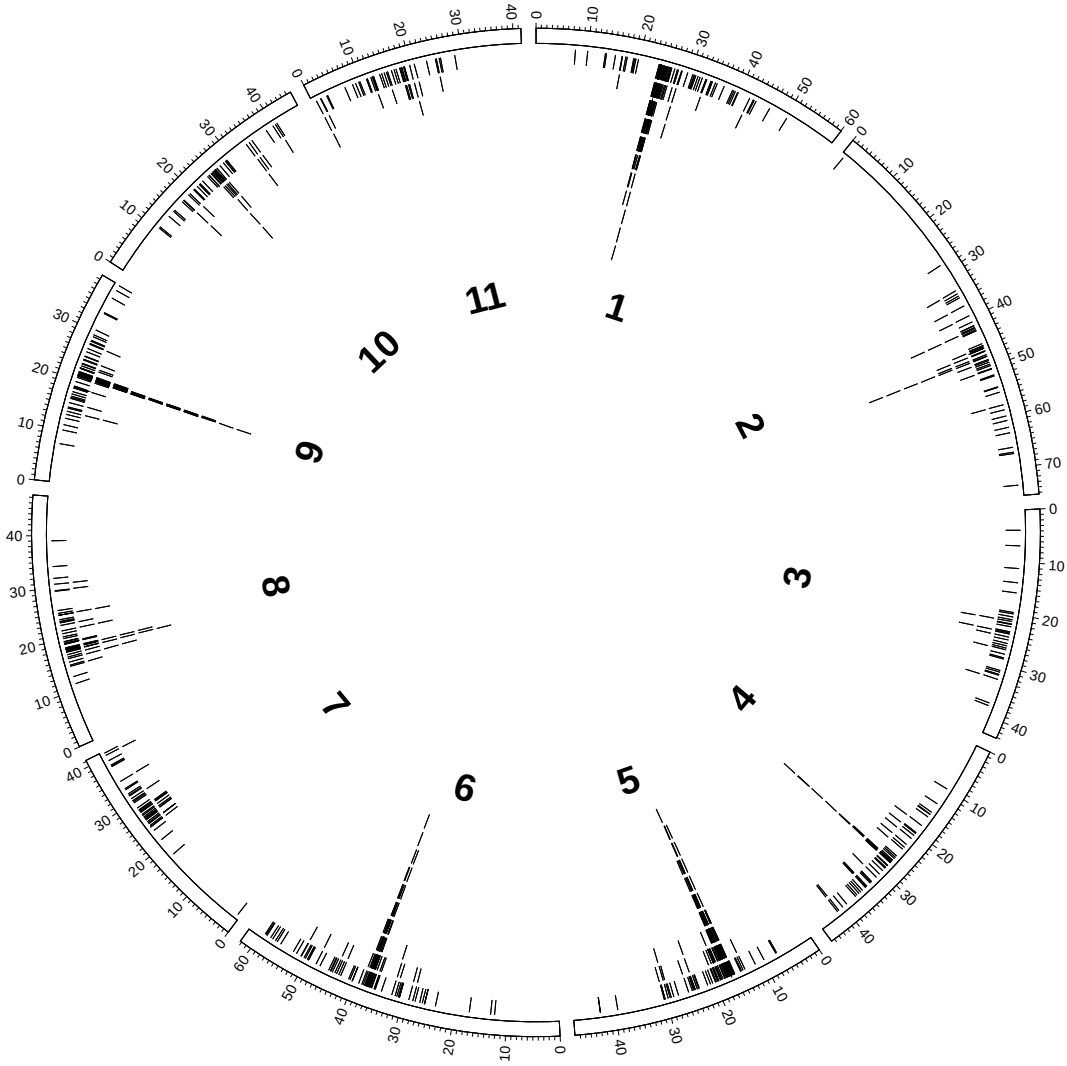
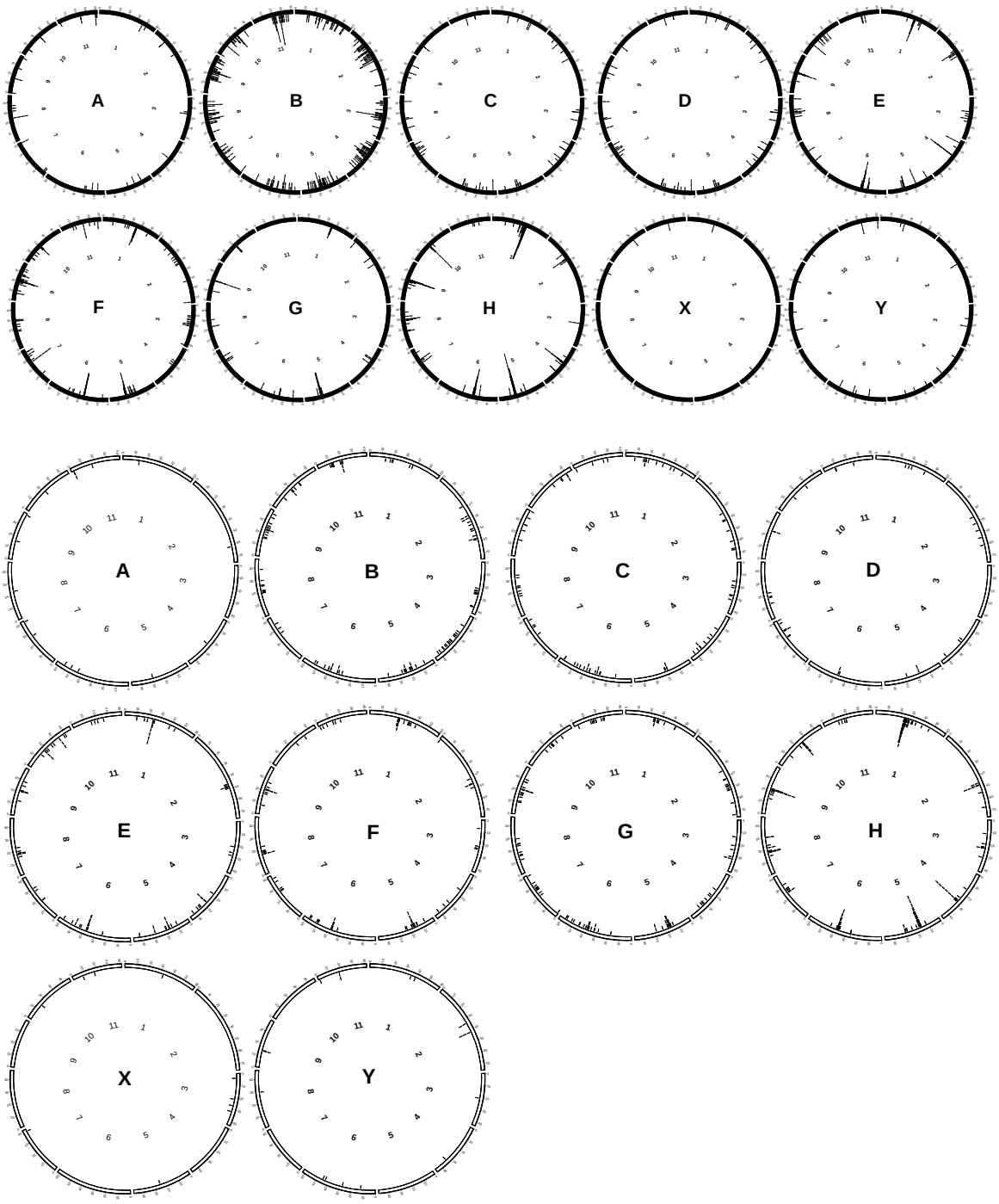


Figure 2. Distribution of RT domain from CRC groups along assembled pseudochromosomes of *Coffea canephora*. Black lines represent the position of RT domains as found by RepeatMasker with a minimum of 400 aligned bases.



Supplemental data 9.

Up: In silico distribution of RT domain from the different CRC groups along assembled pseudochromosomes of *Coffea canephora*. Each circle represents the distribution of one CRC group. Black lines represent the position of RT domains as found by RepeatMasker (-div 20).

Down: In silico distribution of LTR regions from the different CRC groups along assembled pseudochromosomes of *C. canephora*. Each circle represents the distribution of one CRC group. Black lines represent the position of RT domains as found by Censor with a minimum of 80 % of identity over 80 % of the length of the reference sequence.

Cytogenetic analysis

FISH using a probes for RT conserved region, common for all CRC groups (**Supplemental data 10**), showed signals with differences in sizes and brightness on *C. arabica, C. canephora* and *C. eugenioides* nuclei. Signals were distributed in all regions of differentiated cell nuclei (**Figures 3A and F, and Figure 4A-C**), and in a Rabl-like organization in undifferentiated cells (**Figure 3E**). Brighter signals appeared located in the proximal chromosome regions (**see Table 3**), but with variations within and between karyotypes of diploid species *C. canephora* with two signals (**Figures 3B-D**) and *C. eugenioides* with four signals (**Figures 3G-I**), and with six signals in the allotetraploid *C. arabica* (**Figures 4D-I**)**.** In addition to the predominant signals into proximal regions, and few chromosomes displayed scattered signals in proximal/interstitial dots (except for *C. eugenioides*). This is probably due to a smaller copy numbers of CRC RT sequences in these chromosomes. Chromosomes with few or undetectable FISH signals were also observed in *C. canephora* (one pair), *C. eugenioides* (one pair) and *C. arabica* (two chromosome pairs).

Primer position

```
Forward
ACTGTCGGGCTGTAAATGCT.........|.........|.........|.........|.........|.........|.........|.........100
CRcc_group_A
ATAGCCATGCAATCAACAAGATAACAATCAAGTACAGGTTTCCCATACCCAGACTTGATGATATGATTGATATGATGGCTGGGTCAACTATCTACACTAA
CRce_group_A
ATAGCCGTGCAATCAATAAGATAACAATCAAGTATAGGTTTCCCATACCTAGATTTGATAATATGATTGATATGATGACTGGTACTATTGTCTACACTAA
CRcc_group_B
ATTGTCGTGCAATTAATAAAATCACTGTTAAGTATCGTCATCCCATTCCTAGGTTAGACGATATGTTAGAAGAATTGCATGGGGCAATTATTTTCACTAA
```

CRce_group_B
ATTGTCGTGCAATTAATAAAATTACTGTCAAGTATCGTCATCCTATTCCTAGGTTAGATGATATGTTAGAAGAATTGCATGGGGCAATCATTTTCACTAA
CRca_group_B
ATTGTCGTGCAATTAATAAAATTACTGTCAAGTATCGTCATCCTATTCCTAGGTTAGATGATATGTTAGAAGAATTGCATGGGGCAATCATTTTCACTAA
CRca_group_Y
ACTGCCGAGCTACCAATACCATAACGGTAAAGTATCGTCATCTCATCCCTCATCTAGATGACATGTTAGATGAATTACACGGGGCAATTATATTCACAAA
CRcc_group_C
ATTGTAGAGCCGTAAATGCCATCACGGTAAAGTATCGCCACCCTATACCTCGCTTAGATGACATGCTTGATGAGTTACATGGGGCTGTGTTCTTTACCAA
CRce_group_C
ATTGTAGGGCCGTAAATGCCATCACGGTAAAGTATCGTCACCCTATACCTCGCTTAGATGACATGCTTGATGAGTTACATGGGGCTGTGTTCTTTACCAA
CRca_group_C
ATTGCAGAGCCGTAAATGCCATCACGGTAAAGTATCGCCACCCTATACCTCGCTTAGATGACATGCTTGATGAGTTACATGGGGCTGTGTTCTTTACCAA
CRcc_group_E
ACTGTAGGGCTGTTAATGCAATAACGGTAAAATATCGTCACCCTATTCCTAGACTTGATGATATGTTAGATGAGCTATATGGTGCTGTGATTTTCACTAA
CRce_group_E
ACTGTAGGGCCGTTAATGCAATAACGGTAAAATATCGTCACCCTATTCCTAGACTTGATGATATGTTAGATGAGCTATATGGTGCTGTGATTTTCACTAA
CRca_group_E
ACTGTAGGGCTGTTAATGCAATAACGGTAAAATATCGTCACCCTATTCCTAGACTTGATGATATGTTAGATGAGCTATATGGTGCTGTGATTTTCACTAA
CRcc_group_D
ACTGCCGAGCTGTCAACGCCATCACAGTTAAATATCGTCACCCTATATCTCGTTTAGATGATATACTTGACGAATTATATGGTGCAGTCATATTCACTAA
CRcc_group_F
ACTGTCGAGCAGTCAATGCAATCACGGTAAAATATCGTCACCCTATCCCTAGGTTAGATGATATGTTAGATGAATTGCATGGTGCCATTATATTTACTAA
CRce_group_F
ACTGTCGAGCAGTCAATGCAATCACGGTAAAATATCGTCACCCTATCCCTAGGTTAGATGATATGTTAGATGAATTGCATGGTGCCATTATATTCACTAA
CRca_group_F
ACTGTCGAGCAGTCAATGCAATCACGGTAAAATATCGTCACCCTATCCCTAGGTTAGATGATATGTTAGATGAATTGCATGGTGCCATTATATTTACTAA
CRcc_group_G
ATTGTAGAGCCATAAATGCCATAACGGTAAAGTATCGCCATCCCATACCTCGATTAGATGACATGCTTGATGAGTTACATGGTGCTATTATATTTACTAA
CRca_group_G
ATTGTAGAGCCATAAATGCCATAACGGTAAAGTATCGCCATCCCATACCTCGATTAGATGACATGCTTGATGAGTTACATGGTGCTATTATATTTACTAA
CRce_group_X
ATTGTCGTGCCATTAACGCCATCACGATAAAGTATCGCCATCCTATACCACGTTTAGATGACATGCTTGATGAACTTCATGGTGCCGTTATTTTCACTAA
CRca_group_X
ATTGTCGTGCCACTAATGCCATCACGGTAAAGTATCACCATCCTATACCACGTTTAGATGACATGCTTGATGAACTTCATGGTGCCGTTATTTTCACTAA
CRcc_group_H
ACTGTCGGGCTGTGAATGCCATCACTGTCAAATATCGGCATCCCATCCCTCGACTTGATGATATGCTGGACGAACTCGACGGTGCCATTATTTTCACCAA
CRce_group_H
ACTGTCGGGCTGTGAATGCTATCACTGTTAAATATCGTCATCCCATTCCTAGACTTGATGATATGCTTGATGAACTCGATGGGGCCATTATTTTCACCAA
CRca_group_H
ACTGTCGGGCTGTGAATGCTATCACTGTCAAATATCGGCATCCCATCCCTCGACTTGATGATATGCTGGACGAACTCGACGGTGCCATTATTTTCACCAA
CRcc_group_A
AATTGACTTGCGAAAGGGGTACTATCAAATCAGAATTCGTCCAGGTGATGAGTGGAAAACAGCCTTTGAGACGAAGGATGGTCTCTATGAGTGGCTAGTG
CRce_group_A
AATTGACTTGCGAAAGGGGTACTGTTAGATCAGAATTCGTTCAGGTGATGAGTGGAAAACAGCCTTCAAGGCGAAGGATGGTCTCTACGAGTGGCTAGTG
CRcc_group_B
AATTGATTTAAGATCTGGGTATCATCAATTAAGGATAAAGGAAGGTGACGAATGGAAAACTGCTTTCAAAACAAAGTATGGTCTTTATGAGTGGCTTGTG
CRce_group_B
AATTGACTTAAGATCTGGTTATCATCAAATAAGGATAAAGGAAGGAGACGAATGGAAAACTGCTTTCAAAACAAAGTATGGTCTTTATGAGTGGCTTGTG
CRca_group_B
AATTGACTTAAGATCTGGTTATCATCAAATAAGGATAAAGGAAGGAGACGAATGGAAAACTGCTTTCAAAACAAAGTATGGTCTTTATGAGTGGCTTGTG
CRca_group_Y
AATTGATCTGAAATCTGGATATCATCAAATTAGAATTAAAGAGGGGGACGAATGGAAGACTGCATTTAAGACTAAGTACGGGTTATATGAGTAGTTAGTC
CRcc_group_C
AATTGATCTCAAAAGTGGATACCATCAAATTAGGATTAAGGAGGGGGATGAATGGAAAACTGCCTTCAAAACTAAGTATGGATTGTATGAGTGGTTAGTG
CRce_group_C
AATTGATCTCAAAAGTGGGTACCATCAAATTAGGATTAAGGAGGGGGATGAATGGAAAACTGCCTTTAAAACTAAGTATGGATTGTATGAGTGGTTAGTG
CRca_group_C
AATTGATCTCAAAAGTGGGTACCATCAAATTAGGATTCAGGAGGGGGATGAATGGAAAACTGCCTTCAAAACTAAGTATGGCTTGTATGAGTGGTTAGTG

```
CRcc_group_E
AATTGATCTAAAAAGCGGTTATCATCAAATTAGGATGAAAGAGGGGGATGAATGGAAAACGGCCTTTAAAACCAAACATGGCTTGTACGAGTGGTTAGTT
CRce_group_E
AATTGATCTAAAAAGCGGTTATCATCAAATTAGGATGAAAGAGGGGGATGAATGGAAAACGGCCTTTAAAACCAAACATGGCTTGTACGAGTGGTTAGTT
CRca_group_E
AATTGATCTAAAAAGCGGTTATCATCAAATTAGGATGAAAGAGGGGGATGAATGGAAAACGGCCTTTAAAACCAAACATGGCTTGTACGAGTGGTTAGTT
CRcc_group_D
AATAGATCTTAAGAGTGGCTACCATCAAATTCGAATGAAAGAAGGTGATGAGTAGAAGACAACCTTCAAGACTAAGTATGGTCTGTATGAGTGGCTAGTT
CRcc_group_F
AATTGATTTGAAAAGTGGTTACCACCAAATAAGAATGAAAGAGGGAGATGAGTGGAAAACCGCTTTCAAAACTAAACATGGTCTATACGAGTGGTTAGTA
CRce_group_F
AATTGATTTGAAAAGTGGTTACCACCAAATAAGAATGAAAGAGGGAGATGAGTGGAAAACCGCTTTCAAAACTAAACATGGTCTATATGAGTGGTTAGTA
CRca_group_F
AATTGATTTGAAAAGTGGTTACCACCAAATAAGAATGAAAGAGGGAGATGAGTGGAAAACCGCTTTCAAAACTAAACATGGTCTATACGAGTGGTTAGTA
CRcc_group_G
GATTGACTTGAAAAGTGGTTATCACCAAATACGCATGAAAGAAGGGGATGAGTGGAAAACAGCATTTAAAACAAAACATGGTCTGTATGAGTGGTTGGTC
CRca_group_G
GATTGACTTGAAGAGTGGTTATCACCAAATACGCATGAAAGAAGGGGATGAGTGGAAAACAGCATTTAAAACAAAACATGGTCTGTATGAGTGGTTGGTC
CRce_group_X
AATTGATCTTAAAAGTGGTTATCATCAAATTAGGATGAAAGAAGGAGATGAATGGAAAACTGCATTTAAGACTAAACATGGTTTGTATGAGTGGTTAGTC
CRca_group_X
AATTGATCTTAAAAGTGGTTATCATCAAATTAGGATAAAAGAAGGAGATGAATGGAAAACTGTATTTAAGACTAAACATGATTTGTATGAGTGGTTAGTC
CRcc_group_H
AATTGACTTGAGGAGTGGCTATCATCAAATCCGAATGAAGGAAGGCGATGAGTGGAAGACGGCCTTCAAAACCAAACATGGTCTCTATGAGTGGCTGGTC
CRce_group_H
AATTGACTTAAGGAGTGGCTATCATCAAATTAGGATGAAAGAAGGTGATGAGTGGAAAACAGCCTTCAAAACCAAACATGGTCTCTATGAGTGGCTAGTC
CRca_group_H
AATTGACTTGAGGAGTGGCTATCATCAAATCCGAATGAAGGAAGGCGATGAGTGGAAGACGGCCTTCAAAACCAAACATGGTCTCTATGAGTGGCTAGTC

.........|.........|.........|.........|.........|.........|.........|.........|.........|.........200

CRcc_group_A
ATGCCATTTGGCCTCTCTAATGCTCCCAGCACCTTTATGAGATTCATGACGCGGGTATTTCAGCCCTTTATTGGCCATTTTTTGGTTGTCTATTTCGATG
CRce_group_A
ATGCCGTTTGGCCTCTCTAATGCTCCTAGCACCTTTATGAGATTCATGACACAGGTGTTTCAGCCCTTTATTGGCCTTTTTTTGGTTGTCTATTTCGATG
CRcc_group_B
ATGCCTTTTGGCTTAACAAACGCTCCAAGTACTTTCATGAGATTAATGAACCATGTTTTAAGGCATTTTCTTGGGAAGTTCGTTGTTGTTTATTTTGATG
CRce_group_B
ATGCCTTTCGGCTTAACAAACGCTCCAAGTACTTTCATGAGGTTAATGAACCATGTTTTAAGGCATTTTCTTGGGAAGTTTGTTGTTGTTTATTTTGATG
CRca_group_B
ATGCCTTTCGGCTTAACAAACGCTCCAAGTACTTTCATGAGGTTAATGAACCATGTTTTAAGGCATTTTCTTGGGAAGTTTGTTGTTGTTTATTTTGATG
CRca_group_Y
ATACCTTTTGGCCTTACTAATGCACCAAGCACCTTTATGCGACTAATGAACCATGTCTTGCGTGAGTACTTAGGCAAATTTGTAGTAGTATATTTTGATG
CRcc_group_C
ATGCCTTTTGGGCTGACTAATGCACCTAGTACCTTCATGAGATTAATGAACCATGTTCTACGTCCATTCTTAGGGAAATTTGTGGTAGTTTATTTTGACG
CRce_group_C
ATGCCTTTTGGGCTCACTAATGCACCTAGCACCTTCATGAGGTTAATGAACCATGTTCTGCGTCCATTCTTAGGAAAATTCGTGGTAGTTTACTTTGATG
CRca_group_C
ATGCCTTTTGGGCTAACTAATGCACCTAGTACCTTCATGAGGTTAATGAATCATGTTCTGCGTCCATTCTTAGGAAAATTTGTGGTAGTTTACTTTGACG
CRcc_group_E
ATGCCATTTGGACTAACTAATGCACCTAGTACGTTCATGAGGTTGATGAACCATGTACTTCGTCCATTCCTTGGAAAATTTGTAGTTGTATATTTTGACG
CRce_group_E
ATGCCATTTGGACTAACTAATGCACCTAGTACGTTCATGAGATTGATGAACCATGTACTTCGTCCATTCCTTGGAAAATTTGTAGTTGTATATTTTGACG
CRca_group_E
ATGCCATTTGGACTAACTAATGCACCTAGTACATTCATGAGGTTGATGAACCATGTACTTCGTCCATTCCTTGGAAAATTTGTAGTTGTATATTTTGACG
CRcc_group_D
ATACCTTTTGGTTTGACTAATGTTCCTAGTACATTTATGCACTTTATGAACCATGTGCTTCGCCCTTTCTTAAGTAAATTTGTTGTTGTTTATTTTGACG
CRcc_group_F
ATGCCATTTGGTTTAACTAATGCACCTAGTACCTTCATGCGTTTAATGAACCATGTTTTGAGACCTTTTCTTGGGAAATTTGTAGTGGTTTACTTTGATG
```

```
CRce_group_F
ATGCCATTTGGCTTAACTAATGCACCTAGTACCTTCATGCGTTTGATGAACCATGTTTTGAGACCTTTTCTTGGGAAATTTGTAGTGGTTTACTTTGATG
CRca_group_F
ATGCCATTTGGTTTAACTAATGCACCTAGTACCTTCATGCGTTTAATGAACCATGTTTTGAGACCTTTTCTTGGGAAATTTGTAGTGGTTTACTTTGATG
CRcc_group_G
ATGCCTTTTGGCTTAACTAATGCACCTAGTACCTTCATGAGATTAATGAACCATGTTTTGCGTTCTTTTATTGGTAAATTTGTGGTAGTCTACTTTGATG
CRca_group_G
ATGCCTTTTGGCTTAACTAATGCACCTAGTACCTTCATGAGATTAATGAACCATGTTTTGCGTTCTTTTATTGGTAAATTTGTGGTAGTCTACTTTGATG
CRce_group_X
ATGCCATTTGGCCTGACTAATGCCCCTAGTACTTTCATGCGTTTAATGAACCATGTCTTACGTGCGTTCTTGGACAAATTTGTGGTAGTCTACTTTGATG
CRca_group_X
ATGCCATTTGGCCTGACTAATACCCCTAGTACCTTCATGCGTTTGATGAACCATGTCTTACGTGCTTTCTTGGACAAATTTGTGGTAGTCTACTTCGATG
CRcc_group_H
ATGCCTTTTGGCTTGACCAATGCCCCTAGCACCTTTATGCGCCTCATGAACCATGTCTTGAGACACTTCATTGGTAAATTCGTCATTGTTTACTTTGATG
CRce_group_H
ATGCCTTTTGGCTTGACTAATGCTCCTAGCACTTTCATGCGACTAATGAACCATGTTTTGAGGCACTTTATTGGCAAGTTTGTCATTGTTTACTTTGATG
CRca_group_H
ATGCCTTTTGGCTTGACCAATGCCCCTAGCACCTTTATGCGCCTCATGAACCATGTCTTGAGACACTTCATTGGTAAATTCGTCATTGTTTACTTTGATG

.........|.........|.........|.........|.........|.........|.........|.........|.........|.........300

CRcc_group_A
ATATCTTAATTTACAGCAAGAACAAAGAGGAGCACATTAACCATCTTCAACAAGTAATGAGAGTTCTTCGTCAAAGTTCGCAGGGCATCAA---------
CRce_group_A
ATATTTTAATTTATAGCAAGAACAAAGAGGAGCACATTAACCATCTTCAGCAAGTAATGCGAGTTCTTCGTCAAGAAAAGCTCTATATCAACTTGAAGAA
CRcc_group_B
ATATACTGATCTTTAGCAAGTCTTTAGAAGAGCATGTTGAGCATTTGCGACTTGTTTTAAGTGCCTTGCGTGAAAATAGGCTATTTGCTAACATGGAAAA
CRce_group_B
ATATATTGATTTTTAGCCAGTCTTTAGATGAACATGTTGAGCATTTGCGACTTGTTTTAAGTGCCTTGCGTGAAAATAGGTTGTTTGCTAACATGGAAAA
CRca_group_B
ATATATTGATTTTTAGCCAGTCTTTAGATGAACATGTTGAGCATTTGCGACTTGTTTTAAGTGCCTTGCGTGAAAATAGGTTGTTTGCTAACATGGAAAA
CRca_group_Y
ATATTCTCATTTATAGCACAAGTCTAGAGGAACATTTACAGCATGTTAAACTTGTGTTAGAAATACTTCGAAAGGAACGCCCATATGCTAATCTAAAGAA
CRcc_group_C
ATATCCTAATTTATAGCAAATCTTATGATGAACACCTAGAACATATTAGGGCTGTTATGGATGTACTTCGAAGAGAAAAGCTCTATGCCAATCTCAAGAA
CRce_group_C
ATATCCTAATTTATAGCAAATCTTATGATGAACACCTCGAGCATATTAGGGCTGTTATGGATGTACTTCGACGAGAGCAGCTCTATGCCAATCTCAAGAA
CRca_group_C
ATATCCTAATTTATAGCAAATCTTATGATGAGCACCTAGAGCATATTAGGGCTGTTATGGATGTACTTCGAAGAGAGAAGCTCTATGCCAATCTCAAGAA
CRcc_group_E
ATATCCTGATTTATAGTAGGAGCTTCGATGAGCATGTTGAACATGTGAAGCTTGTTCTTGATGTACTTCGAAGGGAAAAGCTCTATGCTAACCTTAAGAA
CRce_group_E
ATATCCTGATTTATAGTAGGAGCTTAGATGAGCATGTTGAACATGTGAAGCTTGTTCTTGATGTACTTCGAAGGGAAAAGCTCTATGCTAACCTTAAGAA
CRca_group_E
ATATCCTGATTTATAGTAGGAGCTTCGATGAGCATGTTGAACATGTGAAGCTTGTTCTTGATGTACTTCGAAGGGAAAAGCTCTATGCTAACCTTAAGAA
CRcc_group_D
GTATTCTCATATACAGCAAGAGTCCTGAGGAACATGTAGCACATGTACGAACTCTTCTTGATGTTTTGCGTAGGGAGAGGTTATTTGCTAACCTTGGCAA
CRcc_group_F
ATATTCTGATTTATAGCAGAAGCTTAGAGGAACACCTTGAGCACCTCAAAGCCATCTTTGAAGTACTTCGAAGGGAAAGGCTATATGCCAACCTTAAGAA
CRce_group_F
ATATTCTGATTTATAGTAGAAGCTTAGAGGAACACTTTGAGCACCTCAAAGCCATCTTTGCAGTACTTCGAAGGGAAAGGCTATATGCCAACCTTAAGAA
CRca_group_F
ATATTCTGATTTATAGCAGAAGTTTAGAAGAACACCTTGAGCACCTCAAAGCCATCTTTGAAGTACTTCGAAGGGAAAGGCTATATGCCAACCTTAAGAA
CRcc_group_G
ACATATTGATCTATAGTAAGAGTACAGAAGAGCATGTTGTGCATGTACGAATGGTCTTAGATGCACTTCGAAAGGCGAGCCTCTATGCTAACCTTAAGAA
CRca_group_G
ACATATTGATCTATAGTAAGAGTACAGAAGAGCATGTTGTGCATGTACGAATGGTCTTAGATGCACATCGAAAGGCGAGCCTCTATGCTAACCTTAAGAA
CRce_group_X
ATATCCTCATTTATAGTAAAAGTTTAGATGAGCATGTTGATCATGTTAAAGCTGTTTTAGAGGTTCTTCGAAGGGAACACTTGTATGCTAATCTTCAAAA
```

```
CRca_group_X
ATATCCTCATTTATAGTAAAAGTTTAGATGAGCATGTTGAGCATGTTAAAGCTATTTTAGAGGTTCTTCGAAGGGAACACTTGTATGCTAATCTTCAAAA
CRcc_group_H
ACATCTTGATATATAGTCGTAGTGAGCAAGAGCACTTGGAGCATGTACGACTAGTTCTTGAGACGCTTCGCCAAGCACAACTCTACGCTAACCTCAAGAA
CRce_group_H
ACATCCTGATCTATAGTCGTAGTGAGCATGAGCACTTAGAGCATGTGAGATTAGTTCTTGAGACACTTCGCCAGGCAAGTCTATACGCCAACCTCAAGAA
CRca_group_H
ACATCTTGATATATAGTCGTAGTGAGCAAGAGCACTTGGAGCATGTACGACTAGTTCTTGAGACGCTTCGCCAAGCACAACTCTACGCTAACCTCAAGAA
.........|.........|.........|.........|.........|.........|.........|.........|.........|........400

CRcc_group_A  ----------------------------------------------------------
CRce_group_A  GTGTACTTTCATGGCTCCTAGTATTGTATTTTTGGG----------AGTTCGCAG
CRcc_group_B  ATGTGTCTTTTGCACTTCTGAGGTTAACTTTCTTGGTTATATTGTTAGTTCGCAG
CRce_group_B  ATGTGTCTTCTGCACTCCTGAAGTTAATTTCCTTGGATATATTGTTAGTTCGCAG
CRca_group_B  ATGTGTCTTCTGCACTCCTGAAGTTAATTTCCTTGGATATATTGTTAGTTCGCAG
CRca_group_Y  ATGTACTTTTTGCACTGACCAACTAGCGTTCCTAGGCTATGTTGTGAGTTCGCAG
CRcc_group_C  GTGCAATTTTTGCACTAATGAGCTTGTGTTTCTAGGGTTTGTTATAAGTTCGCAG
CRce_group_C  GTGTAATTTTTGCACTAATGAGCTTGTGTTTCTAGGGTTTGTTATAAGTTCGCAG
CRca_group_C  GTGCAATTTTTGCACTAACGAACTTGTGTTCCTAGGGTTTGTTATAAGTTCGCAG
CRcc_group_E  GTGCTCATTTTGTACTGATCAACTTGTCTTCCTAGGCTTTGTTGTGAGTTCGCAG
CRce_group_E  GTGCTCCTTTTGTACTGATCAACTTGTCTTCCTAGGCTTTGTTGTGAGTTCGCAG
CRca_group_E  GTGCTCCTTTTGTACTGATCAACTTGTCTTCCTAGGCTTTGTTGTGAGTTCGCAG
CRcc_group_D  ATGTATTTTCTGCACTAATGAGCTTGTTTTCCTTGGTTATAAGGTTAGTTCGCAG
CRcc_group_F  GTGCACATTTTGCACTGATCGTGTTGTGTTTCTAGGATATGTTGTAAGTTCGCAG
CRce_group_F  GTGCACATTTTGCACTGATCGTGTTGTGTTTCTAGGATATGTTGTAAGTTCGCAG
CRca_group_F  GTGCACATTTTGCACTGATCGTGTTGTGTTTCTAGGATATGTTGTAAGTTCGCAG
CRcc_group_G  GTGTTCTTTTTGCACTAATCAACTTGTCTTCCTAGGTTATGTTGTGAGTTCGCAG
CRca_group_G  GTGTTCTTTTTGCACTAATCAACTTGTCTTCCTAGGTTATGTTGTGAGTTCGCAG
CRce_group_X  GTGTACCTTTTGCACTAGCGAGATTGTGTTCCTAGGATATGTTGTAAGTTCGCAG
CRca_group_X  GTGTATCTTTTGCACTAATGAGGTTGTGTTCCTAGGATATGTTGTAAGTTCGCAG
CRcc_group_H  ATGTACTTTTTGTACTAATGAACTTGTGTTTCTTGGCTATGTGGTAAGTTCGCAG
CRce_group_H  GTGCACCTTTTGTACTAACGAGCTTGTGTTTTTAGGCTATGTGGTAAGTTCGCAG
CRca_group_H  ATGTACTTTTTGTACTAATGAACTTGTGTTTCTTGGCTATGTGGTAAGTTCGCAG
              .........|.........|.........|.........|.........|....455bp
Reverse       ................................GCTATGTCGTGAGTTCGCAG
```

Primer position

Supplemental data 10. Nucleotide alignment of RT domain sequences of all CRC groups and selection of PCR primers.

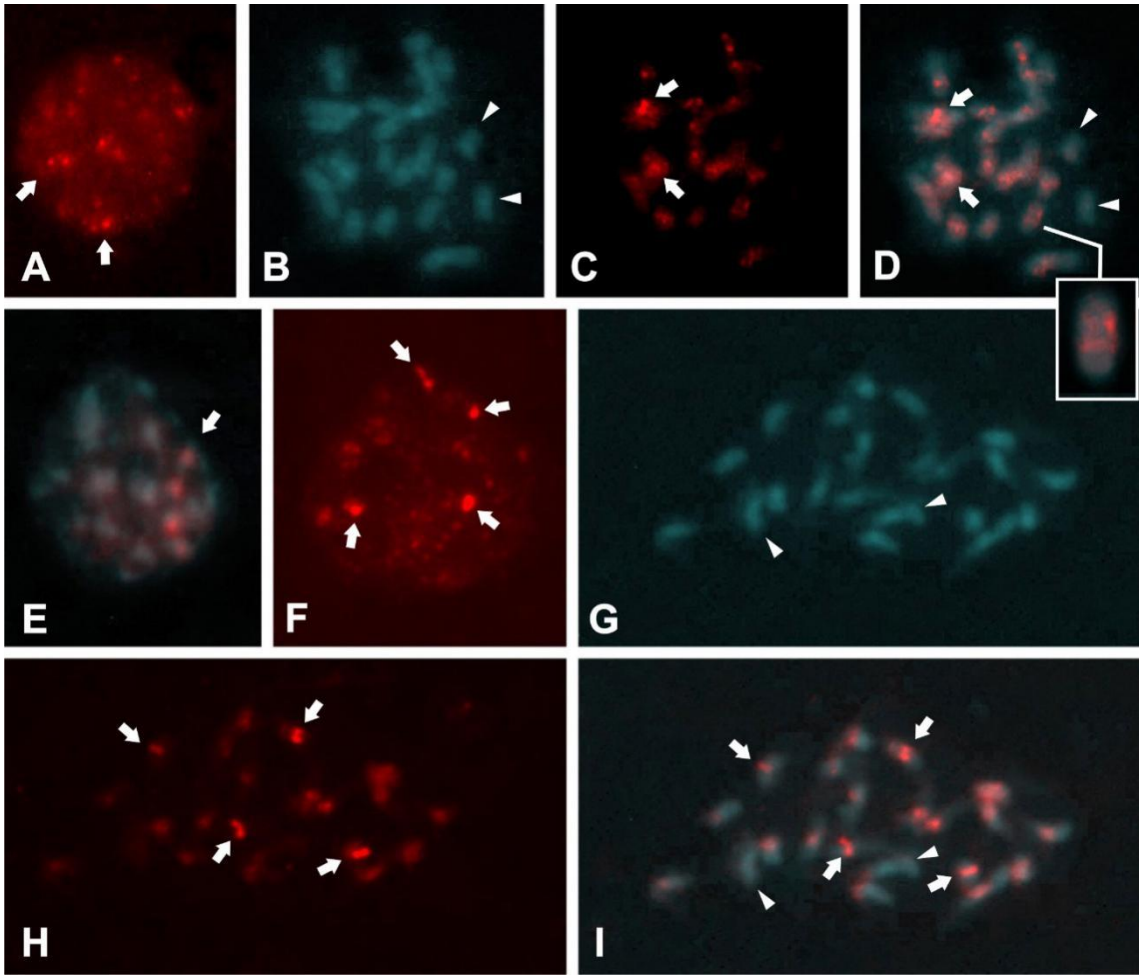


Figure 3. Fluorescence *in situ* hybridization (FISH) in nucleus and metaphases stained with DAPI (blue) and RT-CRC probe hybridized with Cy3-dUTP (red) in *Coffea canephora* (A-D) and *C. eugenioides* (E-I). (A) Nucleus with scattered signals and two brighter signals (arrows). Metaphase stained with DAPI (B), showing RT-CRC FISH signals (C-D) in the centromeres, proximal regions, including few chromosomes with scattered signals and proximal/interstitial dots (box), in red acquired and merged images. E) Undifferentiated nucleus of *C. eugenioides* (Cy3/DAPI merged), showing scattered signals and four brighter signals Rabl-like organized, that are typical of centromeric location. Scattered and four large signals can also be observed in the red stained unpolarized nucleus (F). Arrows point out the large FISH signals. (G-I) Prometaphase stained with DAPI and hybridized with RT-CRC probe. FISH indicates a predominance

of centromeric-pericentromeric signals, including the four large signals detected in the nuclei (arrows). Arrowheads in B, D, G and I indicate chromosomes without hybridization signals. Bar = 10 µm.

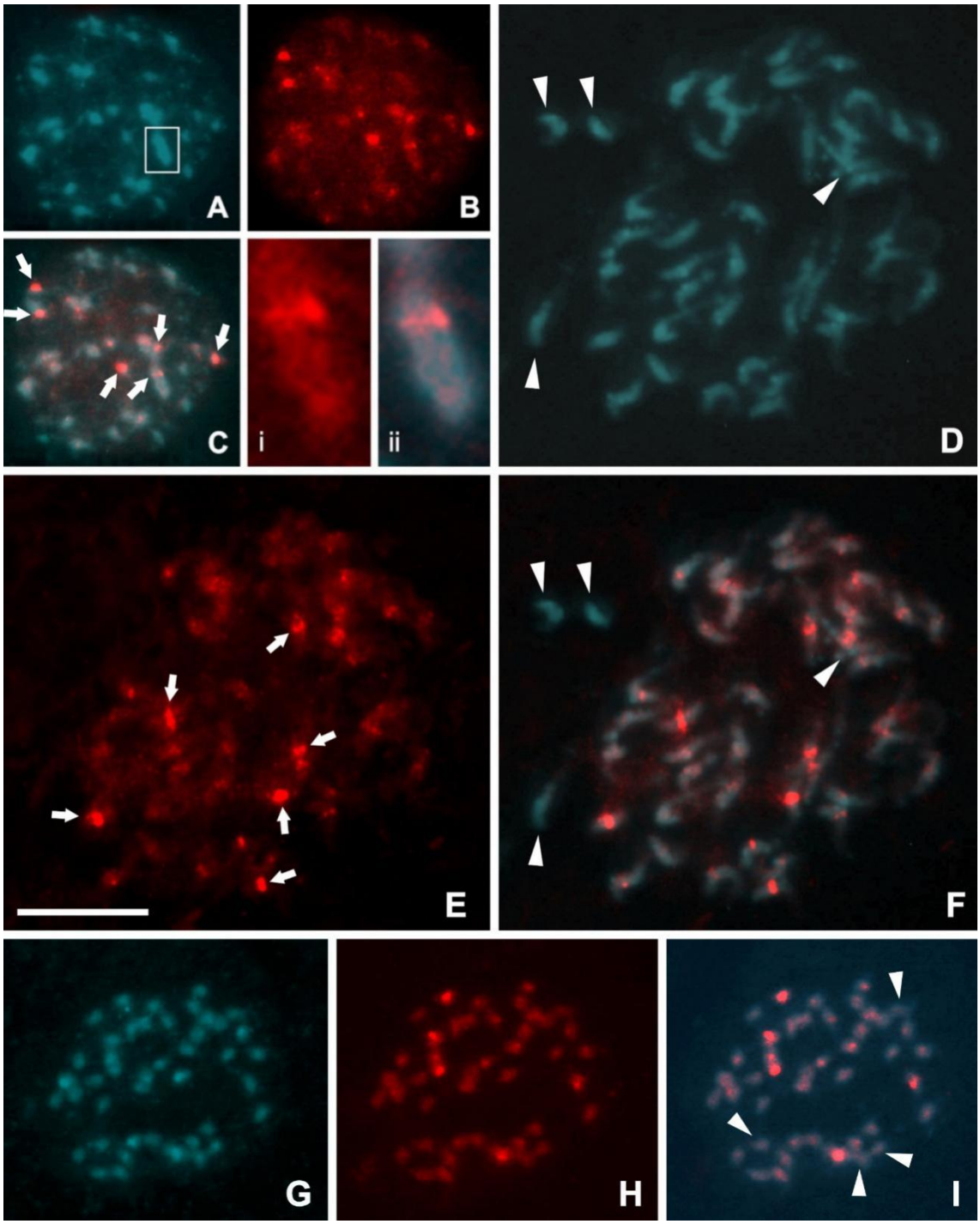


Figure 4. Fluorescence *in situ* hybridization (FISH) in nucleus, prometaphases and metaphases of *Coffea arabica.* Samples stained with DAPI appear in A and D, and with RT-CR FISH signals (red) are in the others. Nucleus showing scattered signals and with six brighter signals (B), that are better observed in the merged image in C (arrows). Boxes i and ii (merged) show a well-defined RT-CRC signal into regions with more condensed chromatin. Prometaphases and metaphases hybridized with the RT-CRC probe (E-I) showing scattered signals, but with predominance of concentrated signals in the centromeric-pericentromeric regions (arrows in E). Arrowheads in D, F and I indicate chromosomes without hybridization signals. Bar = 10 µm.

Table 3. Cytogenetic distribution of CRC RT domains in *Coffea canephora*, *C. eugenioides* and *C. arabica.*

Chromosome Location	Chromosomal pairs with FISH signals		
	C. canephora	*C. eugenioides*	*C. arabica*
Centromeric	7	7	9
Proximal & dispersed	1	2	7
Proximal & interstitial dots	1	0	3
Interstitial & dispersed	1	1	1
No signals	1	1	2
Total	11	11	22

The C-CMA/DAPI banding indicated that C-CMA^{+}/DAPI^{-} were associated to NOR bearing chromosomes in these three species. In *C. canephora* and *C. arabica,* C-CMA^{+}/DAPI^{+} bands were accumulated in proximal regions (**Figures 5A-B and E-F**), while these bands were absent or few accumulated in *C. eugenioides* (**Figures 5C-D**). In

this last species, C-CMA^{+}/C-DAPI^{-} bands seem to be inconspicuous in the proximal regions of some chromosomes and absent in most of them (**Figures 5C-D**). These results showed also that C-CMA^{+} and C-DAPI^{+} heterochromatin can be co-localized with RT CRC hybridization signals for *C. canephora* and *C. arabica* chromosomes, but not for *C. eugenioides*.

Besides the predominance of CRC associated with heterochromatin in the proximal chromosome regions in *C. canephora and C. arabica*, detected here using FISH versus chromosome banding, a comparative map between *C. canephora* chromosomes and RT CRC domain mapping along pseudochromosomes have been drawn **(Figure 6)**. This map showed a good visual correlation between regions with higher CRC concentration into pseudochromosomes and the proximal regions detected using FISH.

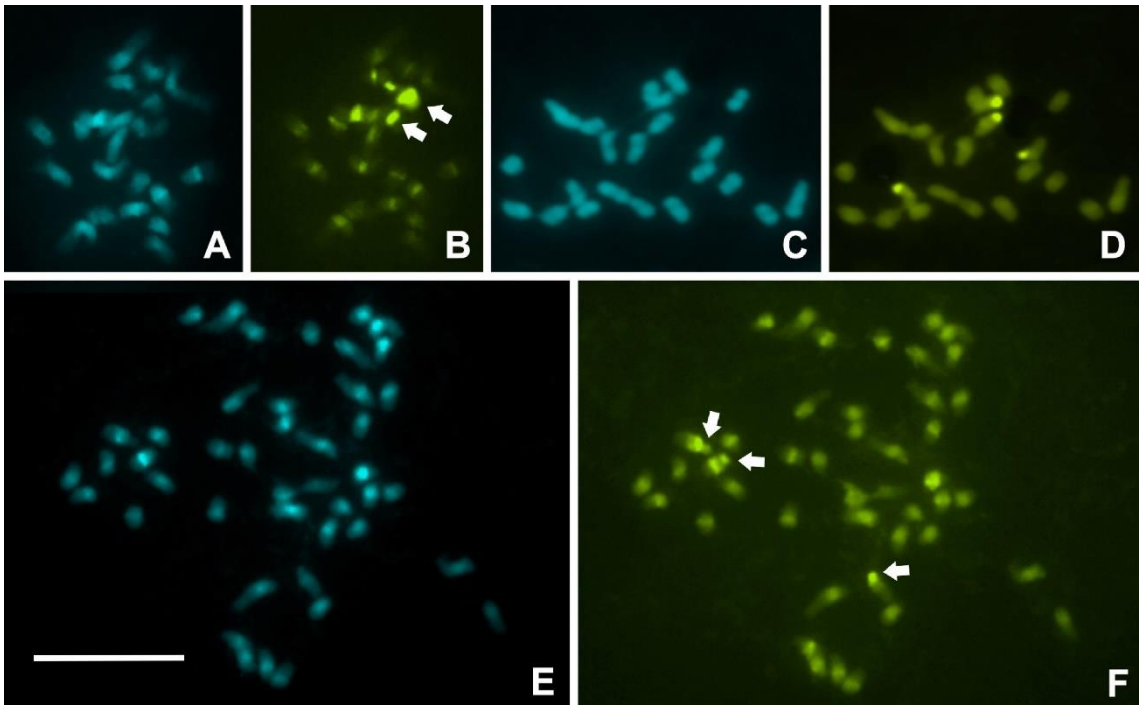


Figure 5. C-CMA/DAPI banding in *Coffea canephora* (A-B), *C. eugenioides* (C-D), and *C. arabica* (E-F), showing an accumulation of C-CMA^{+}/DAPI^{+} bands in the proximal

regions of *C. canephora* and *C. arabica*, and absence of these bands in *C. eugenioides*. However, *C. eugenioides* seems to be inconspicuous C-CMA^{+}/DAPI^{-} bands (thin bands of difficult visualization indicated as arrowhead), in the proximal regions of some chromosomes that are not present in the other two species (C-D). Arrows indicate C-CMA^{+}/DAPI^{-} bands accumulated in the terminal regions that are associated to nucleolar organizing regions (data not shown). Bar = 10 µm.

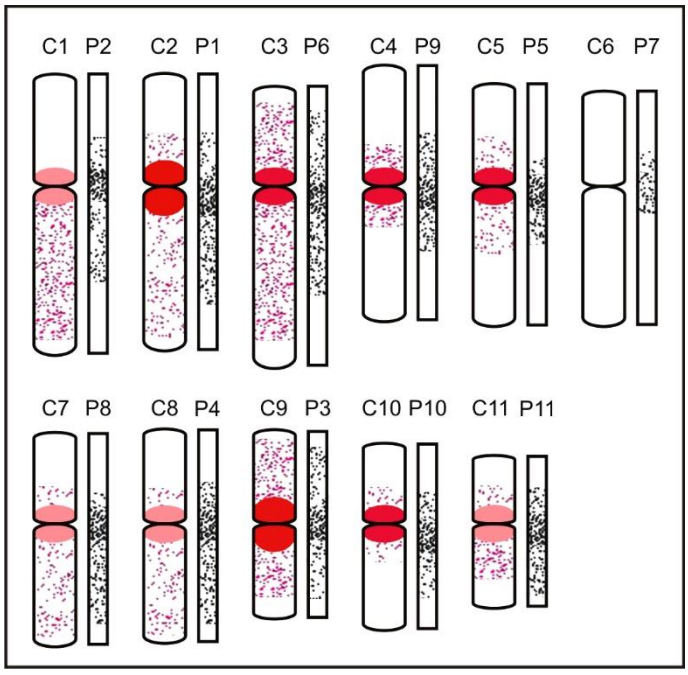


Figure 6. Comparative map between *Coffea canephora* cytological FISH observation with a CRC RT domain probe. C in left correspond to chromosomes and P in right correspond to CRC RT domain mapping along *C. canephora* pseudochromosomes.

The C. canephora and C. arabica chromosome 5 putative centromeric regions are enriched of CRC elements

Based on the FISH data and localization of RT CRC on *C. canephora* genome sequences, the pseudochromosome 5 has been selected for further analysis. The density of transposable elements (light green, annotated on *C. canephora;* Denoeud et al., 2014) and

full-length CRC elements (dark green) were displayed along the pseudochromosome 5 from *C. canephora* (**Figure 7 A**) and along the pseudochromosome 5 sub-genome *C. canephora* from *C. arabica* (**Figure 7 B**). Data showed a high density of CRC elements in the median part for both orthologous pseudochromosomes. A dot-plot of 4 Mb length around these regions in *C. canephora* and *C. arabica* (**Figure 7 C**), suggest a conservation where CRC elements density (dark green) is the highest. Annotations of highest density regions containing CRC elements of *C. canephora* and *C. arabica*, with 1.2 Mb and 800 kb length, respectively (**Figure 7 D**), revealed that 94.1% and 91,7% of these regions consisted of transposable elements. LTR retrotransposons and non-autonomous derivatives represent 84,4% and 79,7% and CRC elements represent 33,7% and 35% in *C. canephora* and *C. arabica*, respectively, whereas transposons account for 0% and 0.7%. Interestingly, the CRC family H, represents alone 17,84% and 25,28% of the analyzed regions in *C. canephora* and *C. arabica*, suggesting of a local enrichment. Beside CRC, the Del lineage is the most redundant with 15,9% and 9,6%. A detailed annotation was performed for the centromeric region of *C. arabica* pseudochromosome 5. Ninety-one complete or partial CRC elements were annotated for which 76 fell into the H family. Twenty-three complete and 13 putative non-autonomous CR elements carrying both intact LTR ends were recovered and their insertion times were estimated. Seventeen of them have a very recent insertion time (> 1 Mya), similarly to estimation at the genome scale (**Supplemental data 8**). In these regions rich in CRC elements, no tandem repeats were observed in *C. canephora* and in *C. arabica* assembled sequences.

Insertion of CRC elements into tandem arrays were directly searched in raw *C. canephora* PacBio reads, before their assembly, using BLAST and dot-plot. Here again no tandem repeats associated with CRC elements of the H family were found. The density of transposable elements (light green, annotated on *C. canephora*; Denoeud et al., 2014) and

full-length CRC elements (dark green) were also displayed along all pseudochromosome from *C. canephora*, *C. arabica* and *C. eugenioides* (**Supplemental data 11, 12 and 13**). Most of the pseudochromosomes showed a clear peak of accumulation.

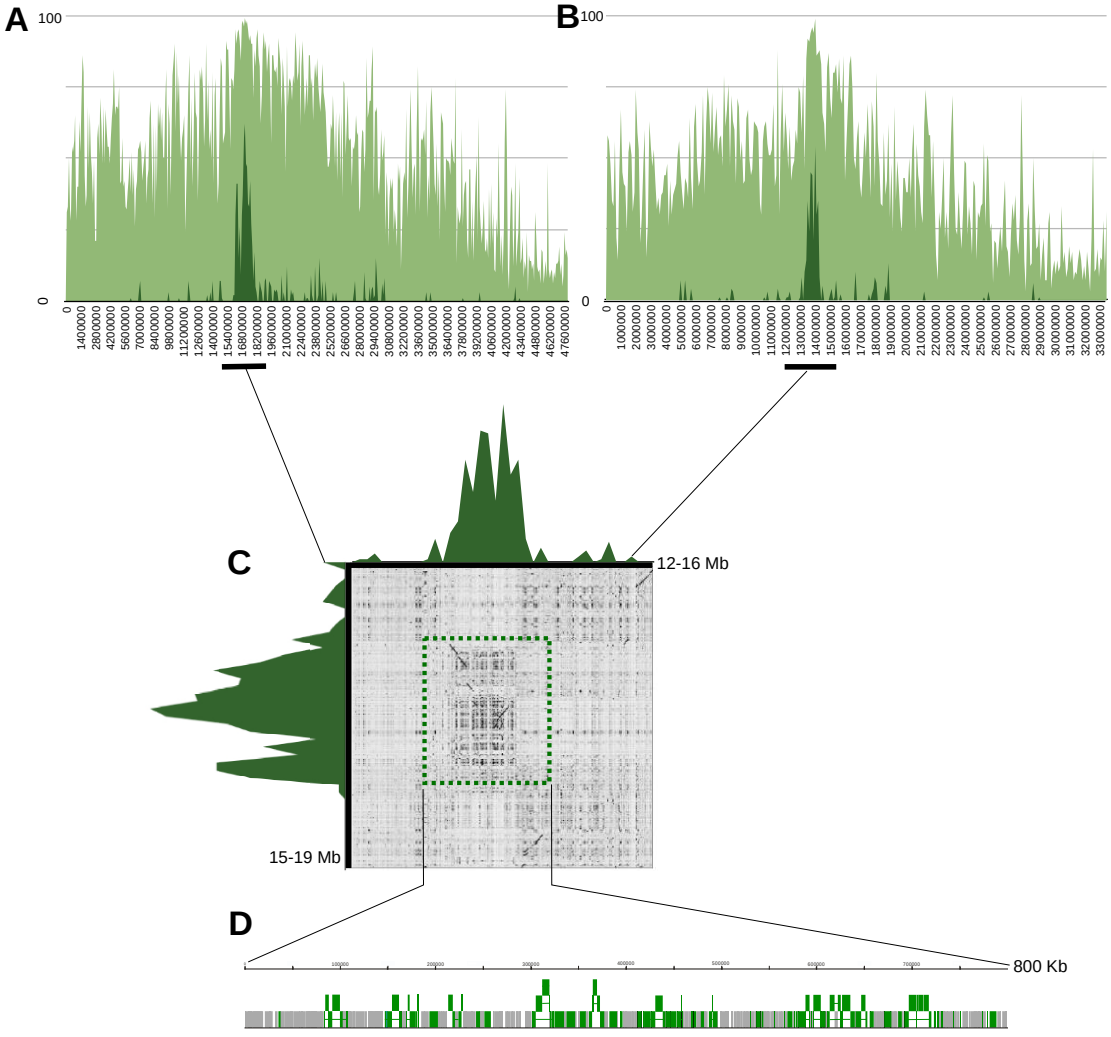


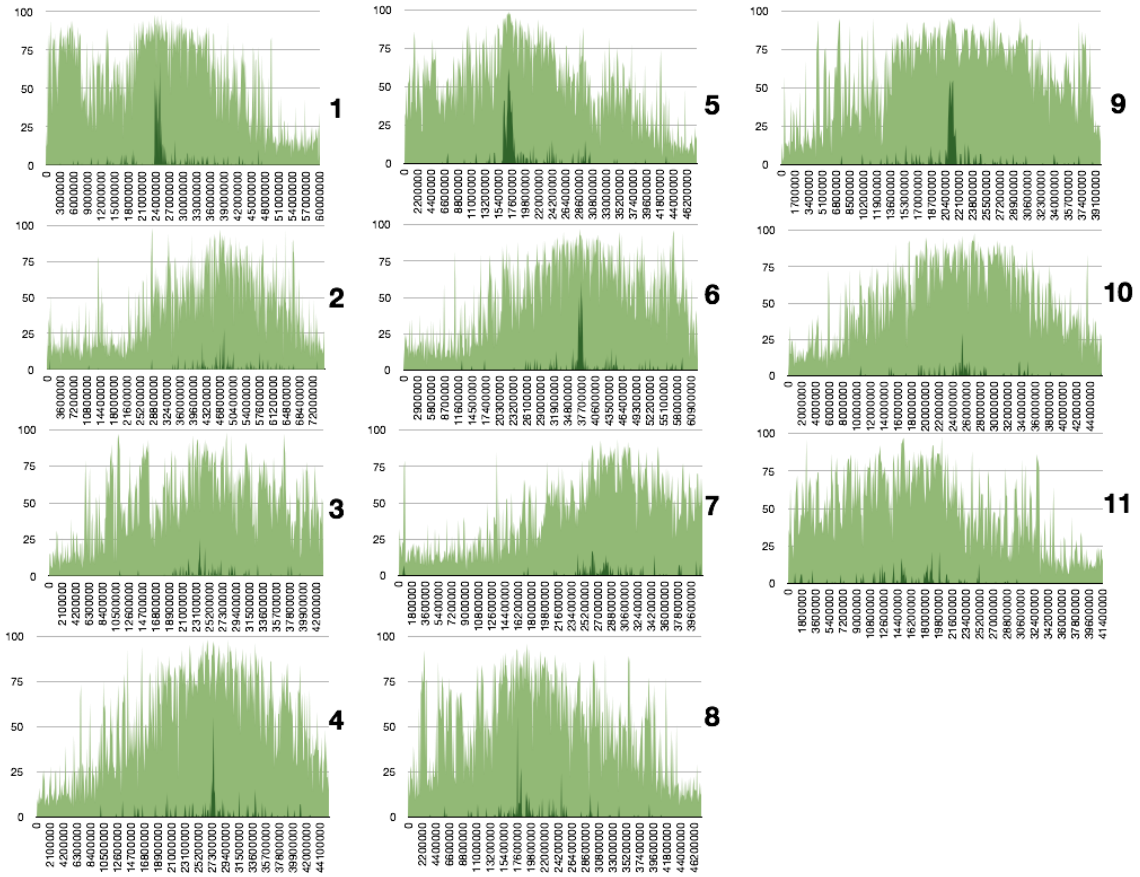
Figure 7. Structure and annotation of pseudo-chromosomes 5 from *C. canephora* and *C. arabica.*

A. Density of transposable elements (light green) and CR elements (dark green) of pseudo-chromosomes 5 from *C. canephora.* X-axis represents the density of elements en percentage and Y-axis the coordinates of the pseudochromosomes.

B. Density of transposable elements (light green) and CR elements (dark green) of pseudo-chromosomes 5 from the *C. canephora* sub-genome in *C. arabica*

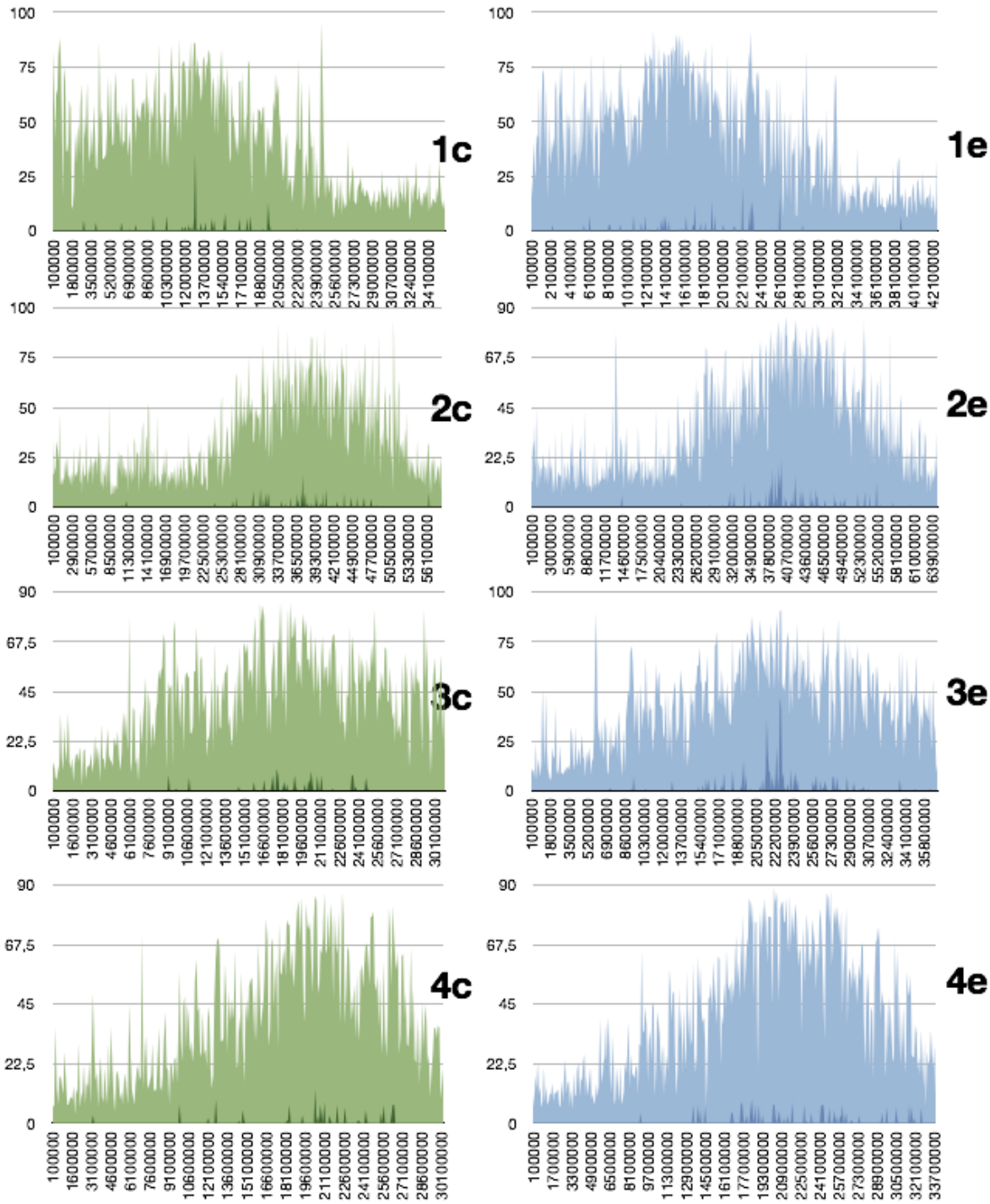
C. Dot-plot graphical view of sequence comparison of 4 Mb *in C. arabica* (horizontal) and *C. canephora* (vertical). Dark green peaks represent the density of CR elements in these regions.

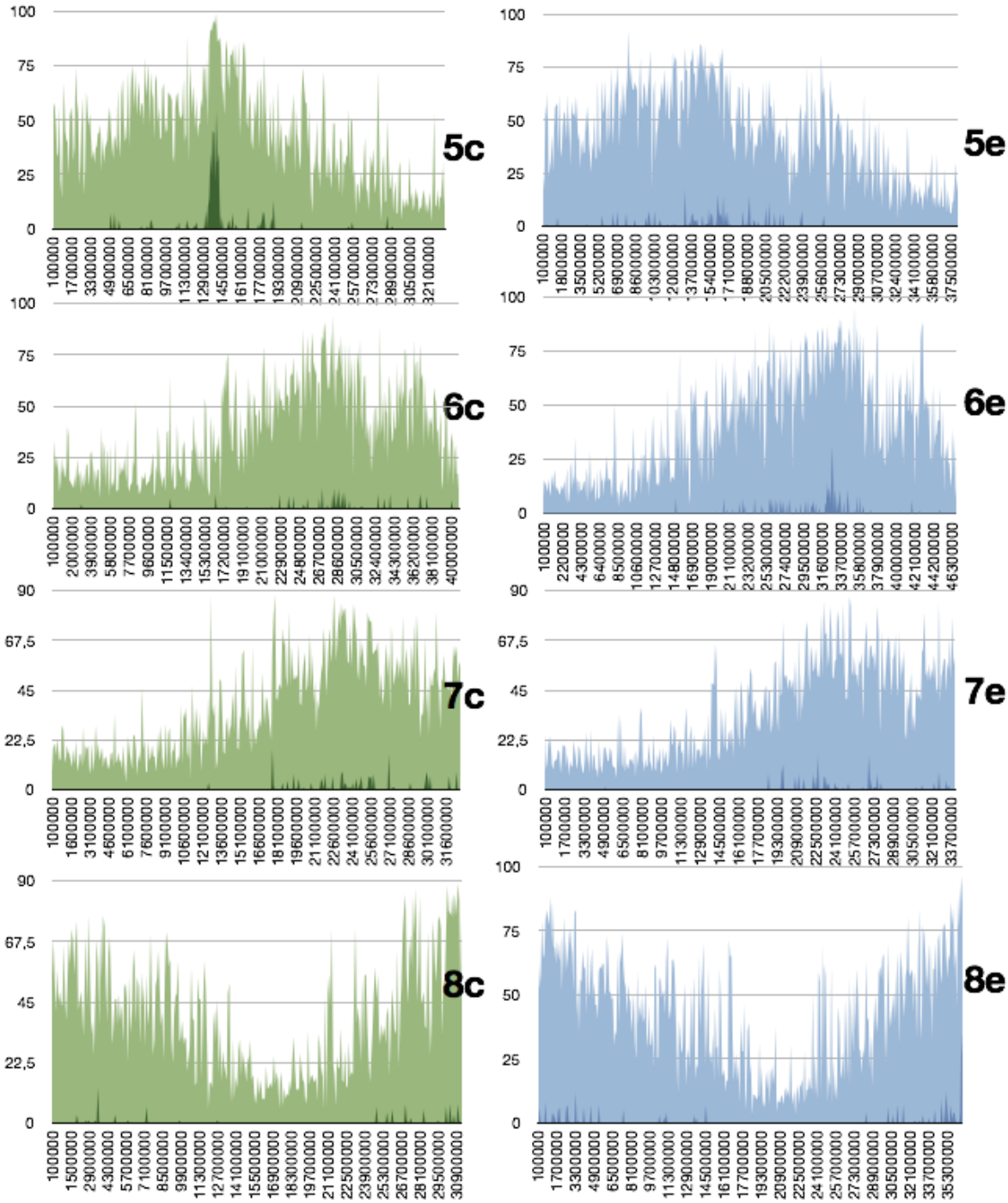
D. Sequence organization of the 800 kb centromeric region in *C. arabica*. Grey blocks represent transposable elements and green blocks are CR elements.

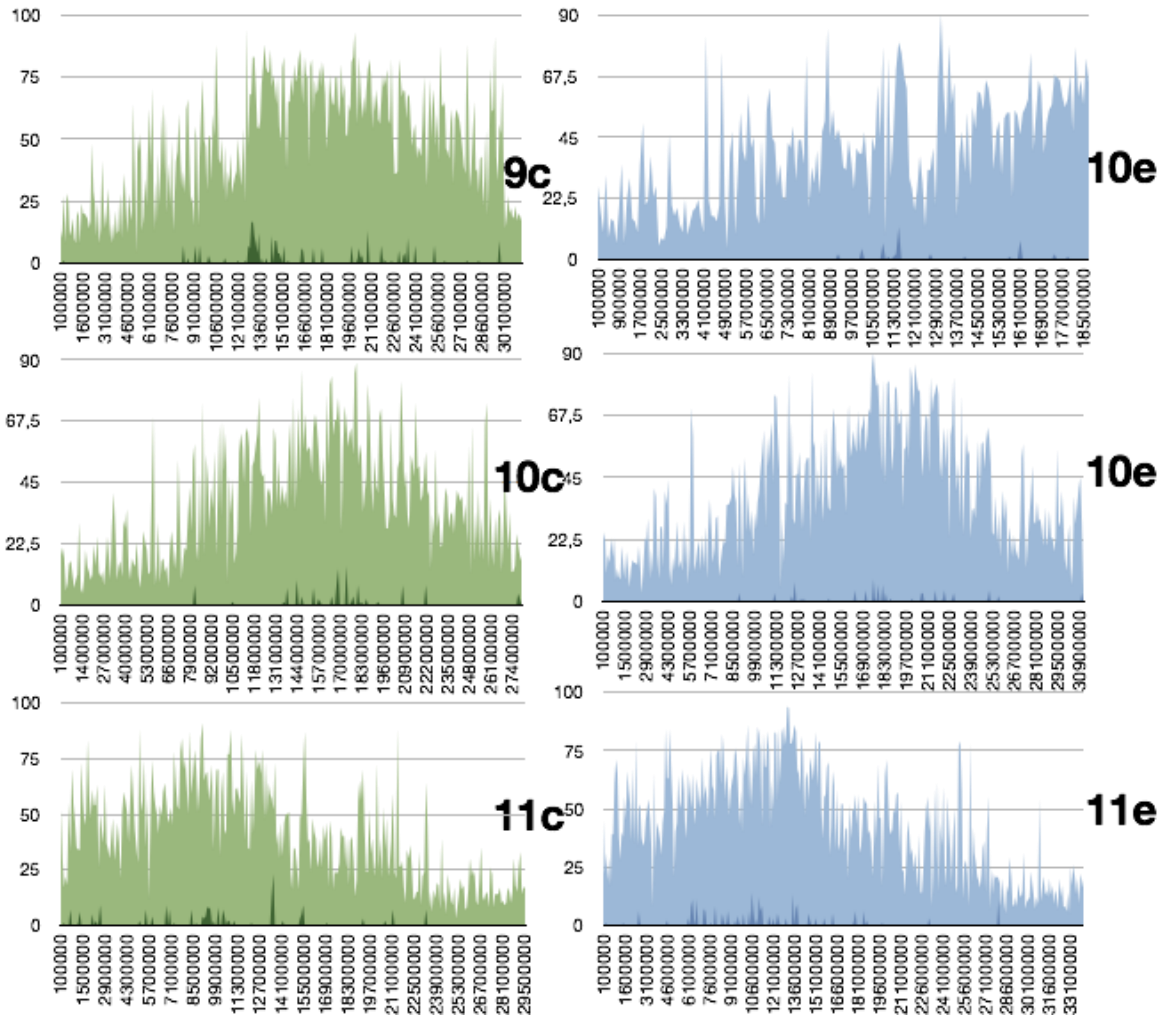


Supplemental data 11. The density of transposable elements (light green, annotated on *C. canephora;* Denoeud et al., 2014) and full-length CRC elements (dark green) along all *C. canephora* PacBio pseudochromosomes. Y-axis represents the density of transposable elements (A percentage calculated as the length (bp) of transposable elements over a tilling window of 100,000 bp) and X-axis the bin coordinates (every 100,000 bp) along

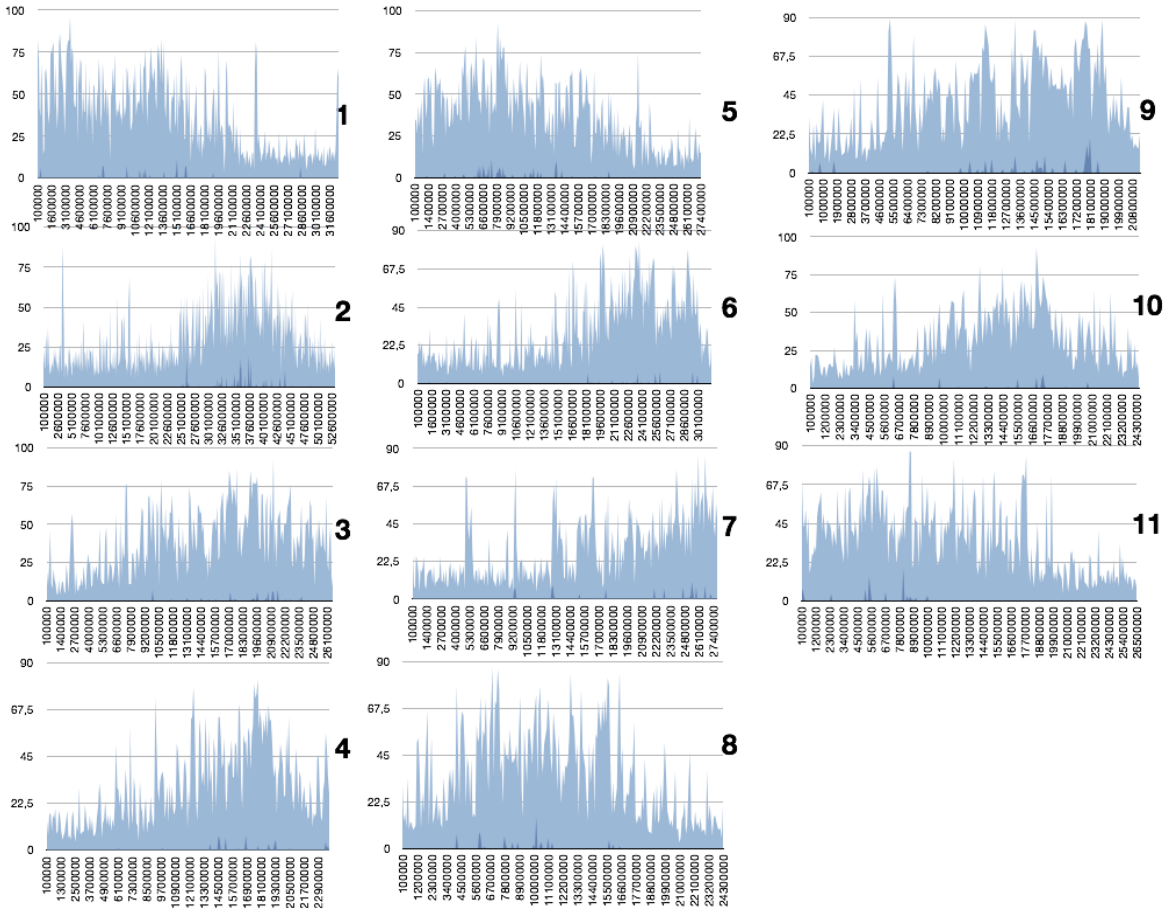
each pseudochromosomes. Densities were calculated by DensityMap (Guizard et al., 2016).




5c
5e
6c
6e
7c
7e
8c
8e



Supplemental data 12. The density of transposable elements (light green or light blue, annotated on *C. canephora;* Denoeud et al., 2014) and full-length CRC elements (dark green or dark blue) along all *C. arabica* PacBio pseudochromosomes. Y-axis represents the density of transposable elements (A percentage calculated as the length (bp) of transposable elements over a tilling window of 100,000 bp) and X-axis the bin coordinates (every 100,000 bp) along each pseudochromosomes. Densities were calculated by DensityMap (Guizard et al., 2016). 1c to 11c represent *C. canephora* subgenome and 1e to 11e represent *C. eugenioides* subgenomes.



Supplemental data 13. The density of transposable elements (light blue, annotated on *C. canephora;* Denoeud et al., 2014) and full-length CRC elements (dark blue) along all *C. eugenioides* PacBio pseudochromosomes. Y-axis represents the density of transposable elements (A percentage calculated as the length (bp) of transposable elements over a tilling window of 100,000 bp) and X-axis the bin coordinates (every 100,000 bp) along each pseudochromosomes. Densities were calculated by DensityMap (Guizard et al., 2016).

Discussion

Characterization of CRC elements in Coffea yields ten distinct groups

Despite numerous centromeric retrotransposons elements identified in monocot and dicot species (Neumann et al., 2011), their diversity and classification into types, as well as their respective contribution to the structure of centromeric regions is poorly known for most higher plant groups. In this study, we identified ten groups of Centromeric Retrotransposons of *Coffea* (CRC) in the genomes of *Coffea arabica*, an allotetraploid species and its two diploids parents, *C. canephora* and *C. eugenioides*. This work was based on high coverage of PacBio reads used for *Coffea arabica*, *C. canephora* and *C. eugenioides* genomes produced by the Arabica Coffee Genome Consortium (ACGC; Mueller et al., 2015). Centromeric retrotransposons in plants were initially organized into three groups, based on the presence of a CR domain extending into the 3' LTR and a chromodomain at the C terminus of the POL polyprotein (Neumann et al., 2011). In *Coffea* the ten identified groups fall into two of these groups: those possessing a CR motif (most of them, group "A" from Neumann et al., (2011), corresponding to our B, C, D, E, F, G, H, X and Y groups) and those carrying a terminal chromodomain-like (group "C" from Neumann et al., (2011), corresponding to our A group). These data indicate that centromeric retrotransposons could be more diverse in plants than previously proposed by Neumann et al. (2011).

Chromodomain might target integration of chromovirus LTR retrotransposon into heterochromatic chromosome regions (Novikova, 2009), and these specificities could allow the *CRM* accumulation into proximal chromosome regions, such as in *Coffea*, or may be still associated with epigenetic changes mechanisms (Houben et al., 2007, Neumann et al., 2011). However, most of CRC groups (B, Y, C, E, D, F, G, X and H) that are similar to the "C" group of Neumann et al. (2011), did not have any

chromodomain nor zinc finger domains, but carried a CR motif. This motif appears particularly important for centromeric retrotransposons to target the heterochromatin (Gao et al., 2008), but they are probably not associated with epigenetic changes in H3 histones (Neumann et al., 2011). The "B" group of centromeric retrotransposons, as defined by Neumann et al. (2011), without CR motif nor chromodomain, was not identified in the autonomous elements set in *C. arabica*, *C. canephora* and *C. eugenioides* genomes. This group has been probably lost or degenerated during the evolution of the *Coffea* genus, since group "B" was identified in another dicotyledonous, such as *Vitis*, *Arabidopsis*, *Medicago* and *Populus* (Neumann et al., 2011). Another possibility is the group B of Neumann has been lost or degenerated earlier during the evolution of the Rubiaceae family or the Asterids branch of dicots, because the genera previously mentioned belong to the Rosids branch.

Non-autonomous centromeric retrotransposons identified in *Coffea* belong to different families: TRIM (Terminal Repeat in Miniature, Witte et al., 2001), LARD (large retrotransposon derivative, Kalendar et al., 2004) or lacking the POL polyprotein region such as TR-GAG (Chaparro et al 2015). This last family was also found in rice (Nagaki et al., 2005). Non-autonomous CRC shared similarities with the nine autonomous CRC groups containing CR motif, suggesting a direct relationship between autonomous and non-autonomous elements, as well as they could indicate that non-autonomous CRC may use the enzymatic machinery of complete elements for their own mobility (Wicker et al., 2007).

In silico copy numbers and insertion time of CRC families

The *C. arabica* genome contains a higher number of complete CRC copies than the related diploid *C. canephora* or *C. eugenioides* genomes, and it is in accordance to relationships between the polyploidization and copy number variation observed for other

retrotransposons in allopolyploid genomes (Parisod et al., 2010). However, the cumulative number of CRC copies is higher for the two diploid than for the allotetraploid species, suggesting that changes occurred either during the hybridization steps leading to *C. arabica* or very recently, after the hybridization, independently in the different genomes. CRC groups may have been amplified very recently in these three genomes, but with higher amplitude in *C. canephora* during the last million years. However, it remains unclear if the CRC copy number variation is only due to differential rates of amplification or if this variation is due to an efficient process of elimination via unequal or illegitimate recombination (Bennetzen, 2007). Two groups with the highest copy number (B and H) in the three species also showed recent peaks of insertion time, suggesting they were amplified recently in the *Coffea* genomes. The only exception is the B group of *Coffea*, which seems to have an ancient origin in *C. arabica*. The number of *C. arabica* CRC observed in present days compared to its progenitors should be carefully interpreted, because the present germplasms of *C. canephora* and *C. eugenioides* studied recently can be accumulated some differences in relation to those which gave rise the amphidiploidy in *C. arabica*. In addition, we have also to consider that all the worldwide *C. arabica* collection had been originated from a few Ethiopian individuals (Carvalho, 1946), and they have been extensively submitted to agronomic breeding selection.

The E and H CRC groups target putative centromeric regions in Coffea

Along plant chromosomes, *Copia* and *Gypsy* superfamilies can be found distributed in blocks and scattered (Lopes et al., 2013; Santos et al., 2015; Zhang et al., 2017). One notable exception is the *Gypsy* Centromeric Retrotransposon lineage, located preferentially into centromeric and proximal regions (Du et al., 2010; Sharma and Presting, 2014). In Coffea species, the distribution of CRC families showed two contrasting situations. One family, the group B with E and H. While the B group appears

scattered along *C. canephora* pseudochromosomes, whereas the H and in a lesser extent the E group, appeared clustered into proximal chromosome regions, as expected for Centromeric Retrotransposons (Sanseverino et al., 2015).

Although it was possible to separate ten CRC groups using complete sequences, the high identity (>90%) of RT regions made difficult the design of specific primers for each group. While specific FISH for each CRC family was impossible with RT-domains, other and more divergent regions such as LTR or gag was relatively inaccurate results.

Results of FISH using a generic RT-*CRM* probe is in agreement with a targeting of chromodomain and CR motif into centromeric regions associated to CENH3 (Houben et al., 2007; Neumann et al., 2011; Li et al., 2013), suggesting an interaction between these elements and centromeric proteins.

Our cytological observations suggested that the hybridization profile is variable among species and chromosomes in *Coffea*. In *C. eugenioides*, FISH signals were strictly associated to centromeric regions, whereas in *C. canephora* and *C. arabica* signals appear less specific to centromeres, and scattered along interstitial regions. This could be the result of a small CRC RT copy numbers hybridized. We hypothesize the two pairs without bright signals in *C. arabica* could be homologous chromosomes to those without FISH signals from the parental genomes (one pair each). Scattered FISH signals using CR probe were also reported in *Saccharum spontaneum* (Zhang et al., 2017). Surprisingly one chromosomes pair in *C. canephora* and *C. eugenioides* and two in *C. arabica* did not exhibit evident centromeric signals. All these variable hybridization patterns could be associated also with differential occurrence of proximal C-CMA^{+}/DAPI^{+} bands, that were observed in *C. canephora* and *C. arabica*, and absent or difficult to distinguish in *C. eugenioides*. The heterochromatin accumulation may be associated with increase and expansion of CRC elements beyond the centromere towards the interstitial regions

observed in *C. canephora* and *C. arabica*. However, additional tests are necessary to confirm this assumption, especially in relation to equilocal dispersion (Schweizer and Loidl , 1987) of repetitive DNA families into proximal regions of *Coffea* chromosomes. In addition, it is possible that, CR elements containing the 3' terminal CR motif, and that represent a fraction of the all CR families, would be more likely inserted into the putative centromeric regions, while the other CRCs (lacking the CR motif) could be less specific and occupy other chromosomal regions. CRC elements carrying a CR motif may also present diverse pattern of insertion, i.e. they can be specific to putative centromeric regions (E and H groups) and/or to interstitial regions (B group). The presence of the CR motif may be not the *sine qua non* condition for a putative centromeric targeting and that other mechanisms may intervene for chromosomal regions targeting by chromoviruses in plants. Chromatin immunoprecipitation followed by sequencing (ChIP-Seq) using antibodies against the centromere-specific histone H3 of Coffee are now required to validate putative centromeric regions as active centromeres.

The putative centromeric region of chromosome 5 is mainly composed of the H family

Repetitive DNA families, such as centromeric retrotransposons and tandem repeats, participate in the complex organization of centromeric regions, especially of the kinetochore formation (Neumann et al., 2011). In *Coffea*, 23 CRCs were predicted as elements that have some role in the centromeric regions, as observed in other plant groups (Han et al., 2010; Sanei et al., 2011). However, it has not been yet clarified what CRC types (complete, truncated, partial or non-autonomous) may participate in kinetochore formation. The presence of partial and truncated elements on proximal chromosome regions suggests that unequal and illegitimate recombination mechanisms may also act on centromeric regions in a neutral manner (Bennetzen, 2007). CR elements were frequently associated with satellite DNA repeats in centromeric regions of other plant

species (Cheng et al., 2002; Lim et al., 2007), except for the wheat chromosome 3B, only composed of CRW retrotransposons families (Li et al., 201). This observation may suggest that CR elements alone might be sufficient to ensure the kinetochore function. But more detailed annotations of centromeric regions of Coffee trees are necessary to understand the composition and the evolution of such critical chromosomal regions.

The diversity in types and chromosomal insertions of CRCs gave a more complex view of the structure and evolution of centromeric regions in *Coffea,* especially in relation to LTR-RTs along hybridization process. *Coffea arabica* showed an accumulation of proximal heterochromatin associated with more dispersed CRC profile on the chromosomes, suggesting that the roles and effects of centromeric retrotransposons can extend beyond the proximal domains. In the near future, the characterization of centromere sequences in diploid and allotetraploid *Coffea* genomes will bring more insights into the evolution of these chromosomal regions that play a crucial role in the cell life cycle.

Acknowledgments

The authors thank the Brazilian agencies Fundação Araucária, CNPq and CAPES-Agropolis for financial support and the Agronomic Institute of Paraná (IAPAR), Londrina, Paraná, Brazil for Coffee seedlings. R.G. was supported by a Special Visiting Scientist grant from the Ciência sem Fronteiras program under the reference ID 84/2013 (CNPq/CAPES) and the Arabica Coffee Genome Consortium (ACGC) for providing unpublished data. The authors also thank the Centro de Bioinformática y Biología Computacional (BIOS), fort he kind use of the cluster service.

References

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 25, 3389-0402. doi: 10.1093/nar/25.17.3389

Bao, W., Zhang, W., Yang, Q., Zhang, Y., Han, B., Gu, M., et al. (2006). Diversity of centromeric repeats in two closely related wild rice species, *Oryza officinalis* and *Oryza rhizomatis*. *Mol. Genet. Genomics*. 275(5), 421-430. doi: 10.1007/s00438-006-0103-2

Bennetzen, J. L. (2007). Patterns in grass genome evolution. *Curr. Opin. Plant Biol.* 10, 176-81. doi: 10.1016/j.pbi.2007.01.010

Bennetzen, J. L., and Wang, H. (2014). The contributions of transposable elements to the structure, function, and evolution of plant genomes. *Annu. Rev. Plant. Biol.* 65, 505-530. doi: 10.1146/annurev-arplant-050213-035811

Birney, E., Clamp, M., and Durbin, R. (2004). GeneWise and Genomewise. *Genome Res.* 14, 988-95. doi: 10.1101/gr.1865504

Carvalho, A. (1946). Distribuição geográfica e classificação botânica do gênero *Coffea* com referência especial à espécie Arabica. V. Origem e classificação botânica do *C. arabica* L. *Separata dos Boletins da Superintendência dos Serviços do Café.* 21, 174-180.

Chaparro, C., Gayraud, T., de Souza, R. F., Domingues, D. S., Akaffou, S., Vanzela, A. L. L., et al. (2015). Terminal-repeat retrotransposons with GAG domain in plant genomes: a new testimony on the complex world of transposable elements. *Genome Biol. Evol.* 7(2), 493-504. doi: 10.1093/gbe/evv001

Cheng, Z., Dong, F., Langdon, T., Ouyang, S., Buell, C. R., Gu, M., et al. (2002). Functional rice centromeres are marked by a satellite repeat and a centromere-specific retrotransposon. *The Plant Cell.* 14(8), 1691-1704. doi: 10.1105/tpc.003079

Denoeud, F., Carretero-Paulet, L., Dereeper, A., Droc, G., Guyot, R., Pietrella, M., et al. (2014). The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. *Science*. 345, 1180-1184. doi: 10.1126/science.1255274

Du, J., Tian, Z., Hans, C. S., Laten, H. M., Cannon, S. B., Jackson, S. A., et al. (2010). Evolutionary conservation, diversity and specificity of LTR−retrotransposons in flowering plants: insights from genome−wide analysis and multi−specific comparison. *Plant J.* 63(4), 584-598. doi: 10.1111/j.1365-313X.2010.04263.x

Dupeyron, M., de Souza, R. F., Hamon, P., Kochko, A., Crouzillat, D., Couturon, E., et al. (2017). Distribution of Divo in *Coffea* genomes, a poorly described family of angiosperm LTR-Retrotransposons. *Mol. Genet. Genomics.* 1-14. doi: 10.1007/s00438-017-1308-2

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32(5), 1792-1797. doi: 10.1093/nar/gkh340 reveals the dynamics of TE-driven genome evolution. Genome Biol. Evol. 5(5):954-965.

Gao, X., Hou, Y., Ebina, H., Levin, H. L., and Voytas, D. F. (2008). Chromodomains direct integration of retrotransposons to heterochromatin. *Genome Res.* 18, 359-369. doi: 10.1101/gr.7146408

Gao, D., Chen, J., Chen, M., Meyers, B. C., and Jackson, S. (2012). A highly conserved, small LTR retrotransposon that preferentially targets genes in grass genomes. *PloS One*. 7(2), e32010. doi: 10.1371/journal.pone.0032010

Grandbastien, M. A. (2015). LTR-retrotransposons, handy hitchhikers of plant regulation and stress response. *Biochim. Biophys. Acta.* 849(4), 403-16. doi: 10.1016/j.bbagrm.2014.07.017

Guizard, S., Piégu, B., and Bigot, Y. (2016). DensityMap: a genome viewer for illustrating the densities of features. *BMC Bioinformatics.* 7(1), 204. doi: 10.1186/s12859-016-1055-0

Guyot, R., Darré, T., Dupeyron, M., de Kochko, A., Hamon, S., Couturon, E., 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. doi: 10.1007/s00438-016-1235-7

Hamon, P., Grover, C. E., Davis, A. P., Rakotomalala, J. J., Raharimalala, N. E., Albert, V. A., 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: GBS coffee phylogeny and the evolution of caffeine content. *Mol. Phylogenet. Evol.* 109, 351-361. doi: 10.1016/j.ympev.2017.02.009

Han, Y., Wang, G., Liu, Z., Liu, J., Yue, W., Song, R., and Jin, W. (2010). Divergence in centromere structure distinguishes related genomes in *Coix lacryma-jobi* and its wild relative. *Chromosoma*. 119(1), 89-98. doi: 10.1007/s00412-009-0239-z

Heslop-Harrison, J. S., and Schwarzacher, T. (2011). Organisation of the plant genome in chromosomes. *Plant J.* 66, 18-33. doi: 10.1111/j.1365-313X.2011.04544.x

Houben, A., Schroeder-Reiter, E., Nagaki, K., Nasuda, S., Wanner, G., Murata, M., et al. (2007). CENH3 interacts with the centromeric retrotransposon cereba and GC-rich satellites and locates to centromeric substructures in barley. *Chromosoma.* 116(3), 275-283. doi: 10.1007/s00412-007-0102-z

Kalendar, R., Vicient, C. M., Peleg, O., Anamthawat-Jonsson, K., Bolshoy, A., and Schulman, A. H. (2004). Large retrotransposon derivatives: abundant, conserved but nonautonomous retroelements of barley and related genomes. *Genetics.* 166(3), 1437-1450. doi: 10.1534/genetics.166.3.1437

Lashermes, P., Combes, M. C., Robert, J., Trouslot, P., D'Hont, A., Anthony, F., et al. (1999). Molecular characterisation and origin of the *Coffea arabica* L. genome. *Mol. Gen. Genet.* 261(2), 259-266. 10,1007 / s004380050965

Li, B., Choulet, F., Heng, Y., Hao, W., Paux, E., Liu, Z., et al. (2013). Wheat centromeric retrotransposons: the new ones take a major role in centromeric structure. *Plant J. 73*(6), 952-965. doi: 10.1111/tpj.12086

Lim, K. B., Yang, T. J., Hwang, Y. J., Kim, J. S., Park, J. Y., Kwon, S. J., and Kim, H. I. (2007). Characterization of the centromere and peri-centromere retrotransposons in *Brassica rapa* and their distribution in related Brassica species. *Plant J.* 49(2), 173-183. doi: 10.1111/j.1365-313X.2006.02952.x

Liu, Z., Yue, W., Li, D., Wang, R. R. C., Kong, X., Lu, K., et al. (2008). Structure and dynamics of retrotransposons at wheat centromeres and pericentromeres. *Chromosoma.* 117(5), 445-456. doi: 10.1007/s00412-008-0161-9

Llorens, C., Muñoz-Pomer, A., Bernad, L., Botella, H., and Moya, A. (2009). Network dynamics of eukaryotic LTR retroelements beyond phylogenetic trees. *Biol. Direct.* 4, 41. doi: 10.1186/1745-6150-4-41

Llorens, C., Futami, R., Covelli, L., Domínguez-Escribá, L., Viu, J. M., Tamarit, D., et al. (2011). The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. *Nucleic Acids Res.* 39, D70-D74. doi: 10.1093/nar/gkq1061

Lopes, F. R., Jjingo, D., Da Silva, C. R., Andrade, A. C., Marraccini, P., Teixeira, J. B., et al. (2013). Transcriptional activity, chromosomal distribution and expression

effects of transposable elements in *Coffea* genomes. *PloS One.* 8(11), e78931. doi: 10.1371/journal.pone.0078931

Ma, J., and Bennetzen, J. L. (2004). Rapid recent growth and divergence of rice nuclear genomes. *PNAS.* 101(34), 12404-12410. doi: 10.1073/pnas.0403715101

Marques, A., Ribeiro, T., Neumann, P., Macas, J., Novak, P., Schubert, V., et al. (2015). Holocentromeres in *Rhynchospora* are associated with genome-wide centromere-specific repeat arrays interspersed amongst euchromatin. *PNAS*. 112, 13633. doi: 10.1073/pnas.1512255112

McCarthy, E. M., and McDonald, J. F. (2003). LTR_STRUC: a novel search and identification program for LTR-retrotransposons. *Bioinformatics.* 19, 362-367. doi: 10.1093/bioinformatics/btf878

Mueller, L., Strickler, S. R., Domingues, D. S., Pereira, L. F. P., Andrade, A. A., Marraccini, P. et al. (2015). Towards a Better Understanding of the *Coffea* Arabica Genome Structure. *In : Proceedings of the 25th International Conference on Coffee Science*. ASIC. 42-45.

Nagaki, K., Neumann, P., Zhang, D., Ouyang, S., Buell, C. R., Cheng, Z., et al. (2005). Structure, divergence, and distribution of the CRR centromeric retrotransposon family in rice. *Mol. Biol. Evol.* 22(4), 845-855. doi: 10.1093/molbev/msi069

Neumann, P., Navrátilová, A., Koblížková, A., Kejnovský, E., Hřibová, E., Hobza, R., et al. (2011). Plant centromeric retrotransposons: a structural and cytogenetic perspective. *Mobile DNA.* 2(1), 4. doi: 10.1186/1759-8753-2-4

Novikova O. 2009. Chromodomains and LTR-retrotransposons in plants. Comm. & Integr. Biol. 2:158-162.

Parisod, C., Alix, K., Just, J., Petit, M., Sarilar, V., Mhiri, C., et al. (2010). Impact of transposable elements on the organization and function of allopolyploid genomes. *New Phytol*. 186, 37-45. doi: 10.1111/j.1469-8137.2009.03096.x

Piegu, B., Guyot, R., Picault, N., Roulin, A., Saniyal, A., Kim, H., et al. (2006). Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. *Genome Res.* 16(10), 1262-1269. doi: 10.1101/gr.5290206

Romano, E., and Brasileiro, A. C. M. Extração de DNA de plantas: Soluções para problemas comumente encontrados. (1999). *Biotecnologia, Ciência e Desenvolvimento*. 9, 40-43.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M. A., et al. (2000). Artemis: sequence visualization and annotation. *Bioinformatics.* 16, 944-945. doi: 10.1093/bioinformatics/16.10.944

Sanei, M., Pickering, R., Kumke, K., Nasuda, S., and Houben, A. (2011). Loss of centromeric histone H3 (CENH3) from centromeres precedes uniparental chromosome elimination in interspecific barley hybrids. *PNAS*, 108(33), E498-E505. doi: 10.1073/pnas.1103190108

SanMiguel, P., Gaut, B. S., Tikhonov, A., Nakajima, Y., and Bennetzen, J. L. (1998). The paleontology of intergene retrotransposons of maize. *Nat. Genet.* 20, 43-45. doi: 10.1038/1695

Sanseverino, W., Hénaff, E., Vives, C., Pinosio, S., Burgos-Paz, W., Morgante, M., et al. (2015). Transposon Insertions, Structural Variations, and SNPs Contribute to the Evolution of the Melon Genome. *Mol. Biol. Evol.* msv152. doi: 10.1093/molbev/msv152

Santos, F. C., Guyot, R., Do Valle, C. B., Chiari, L., Techio, V. H., Heslop-Harrison, P., et al. (2015). Chromosomal distribution and evolution of abundant retrotransposons in plants: *Gypsy* elements in diploid and polyploid *Brachiaria* forage grasses. *Chromosome Res.* 23(3), 571-582. doi: 10.1007/s10577-015-9492-6

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternk, S., et al. (2009). The B73 maize genome: complexity, diversity, and dynamics. *Science.* 326, 1112-1116. doi: 10.1126/science.1178534

Schwarzacher, T., Ambros, P., and Schweizer, D. (1980). Application of Giemsa banding to orchid karyotype analysis. *Plant Syst. Evol.* 134, 293-297. doi: 10.1007/BF00986805

Schweizer, D., and Loidl, J. (1987). A model for heterochromatin dispersion and the evolution of C band patterns. *Chrom. Today*. 9, 61-74. doi: 10.1007/978-94-010-9166-4_7

Sharma, A., and Presting, G. G. (2014). Evolution of centromeric retrotransposons in grasses. *Genome Biol. Evol.* 6(6), 1335-1352. doi: 10.1093/gbe/evu096

Sonnhammer, E. L., and Durbin, R. (1995). A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene.* 167(1-2), GC1-10. doi: 10.1016/0378-1119(95)00714-8

Tenaillon, M. I., Hufford, M. B., Gaut, B. S., and Ross-Ibarra, J. (2011). Genome Size and Transposable Element Content as Determined by High-Throughput Sequencing in Maize and *Zea luxurians. Genome Biol. Evol.* 3, 219-229. doi: 10.1093/gbe/evr008

The Arabica Coffee Genome Consortium. 2014. Towards a Better Understanding of the Coffea Arabica Genome Structure. In: Association for Science and Information on Coffee (ed) International Conference on Coffee Science. Cogito, Armenia. 42-45.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22(22), 4673-4680. doi: 10.1093/nar/22.22.4673

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., et al. (2007). A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* 8, 973-982. doi: 10.1038/nrg2165

Witte, C. P., Le, Q. H., Bureau, T., and Kumar, A. (2001). Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes. *PNAS*. 98(24), 13778-13783. doi: 10.1073/pnas.241341898

Yu, Q., Guyot, R., de Kochko, A., Byers, A., Navajas−Pérez, R., Langston, B. J., et al. (2011). Micro−collinearity and genome evolution in the vicinity of an ethylene receptor gene of cultivated diploid and allotetraploid coffee species (*Coffea*). *Plant J.* 67(2), 305-317. doi: 10.1111/j.1365-313X.2011.04590.x

Zhang, W., Zuo, S., Li, Z., Meng, Z., Han, J., Song, J. et al. (2017). Isolation and characterization of centromeric repetitive DNA sequences in *Saccharum spontaneum*. *Sci. Rep*. 7. doi: 10.1038/srep41659

3. CONCLUSÕES

a. Foi possível reconhecer 10 grupos CRC. O grupo A exibiu um cromodomínio, enquanto o restante exibiu uma cadeia poli-A e um CR *motif*;

b. Existem mais semelhanças dentro dos grupos do que entre eles;

c. O grupo D é específico para *C. canephora* e o grupo Y para *C. arabica*. Alguns grupos foram encontrados nos parentais e em *C. arabica*, o que reforça a origem híbrida;

d. O grupo H é o mais representativo nos três genomas;

e. Nossa observação citológica sugere que o perfil de hibridização é variável entre espécies e cromossomos em *Coffea*. Os sinais de FISH predominam nas regiões proximais, mas com intensidades variáveis entre os cromossomos. Contudo, os sinais de FISH também foram observados em regiões intersticiais;

f. Este perfil da FISH parece ser equivalente à localização do CRC nos pseudocromossomos de *C. canephora* e também à ocorrência diferencial de bandas proximais C-CMA $^{+}$ / DAPI $^{+}$;

g. Em *Coffea*, os CRM não estão localizados exclusivamente nas regiões proximais. Nós sugerimos que aqueles elementos que contêm o cromodomínio parecem ser menos específicos da região centromérica, enquanto que aqueles elementos que possuem o CR *motif* parecem ser mais específicos desta região. No entanto, devido ao perfil disperso de hibridização identificado em grupos com CR *motif*, nós acreditamos que a presença do CR *motif* pode não ser a única condição associada à marcação da região centromérica, e devem existir outros mecanismos responsáveis pela marcação desta região.

4. ARTIGO PUBLICADO

Data da publicação: 15/02/2018

frontiers in Plant Science

ORIGINAL RESEARCH
published: 15 February 2018
doi: 10.3389/fpls.2018.00175

Structure and Distribution of Centromeric Retrotransposons at Diploid and Allotetraploid *Coffea* Centromeric and Pericentromeric Regions

Renata de Castro Nunes[1], Simon Orozco-Arias[2], Dominique Crouzillat[3], Lukas A. Mueller[4], Suzy R. Strickler[4], Patrick Descombes[5], Coralie Fournier[5], Deborah Moine[5], Alexandre de Kochko[6], Priscila M. Yuyama[1], André L. L. Vanzela[1] and Romain Guyot[2,7*†]*

[1] Laboratory of Cytogenetics and Plant Diversity, Department of General Biology, Center for Biological Sciences, State University of Londrina, Londrina, Brazil, [2] Department of Electronics and Automatization, Universidad Autónoma de Manizales, Colombia, [3] Nestlé R&D Tours, Notre-Dame d'Oé, Tours, France, [4] Boyce Thompson Institute, Cornell University, Ithaca, NY, United States, [5] Nestlé Institute of Health Sciences, Lausanne, Switzerland, [6] Institut de Recherche pour le Développement, UMR DIADE, EvoGec, Montpellier, France, [7] Institut de Recherche pour le Développement, CIRAD, Univ. Montpellier, UMR IPME, Montpellier, France

OPEN ACCESS

Edited by:
Tian Tang,
Sun Yat-sen University, China

Reviewed by:
Zeljka Pezer,
Rudjer Boskovic Institute, Croatia
Jinfeng Chen,
University of California, Riverside, United States

***Correspondence:**
André L. L. Vanzela
andrevanzela@uel.br
Romain Guyot
romain.guyot@ird.fr

†Present Address:
Romain Guyot,
Centro Nacional de Investigaciones de Café—Cenicafé, Chinchiná-Manizales, Colombia

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Plant Science

Received: *20 October 2017*
Accepted: *30 January 2018*
Published: *15 February 2018*

Citation:
de Castro Nunes R, Orozco-Arias S, Crouzillat D, Mueller LA, Strickler SR, Descombes P, Fournier C, Moine D, de Kochko A, Yuyama PM, Vanzela ALL and Guyot R (2018) Structure and Distribution of Centromeric Retrotransposons at Diploid and Allotetraploid Coffea Centromeric and Pericentromeric Regions. Front. Plant Sci. 9:175. doi: 10.3389/fpls.2018.00175

Centromeric regions of plants are generally composed of large array of satellites from a specific lineage of *Gypsy* LTR-retrotransposons, called Centromeric Retrotransposons. Repeated sequences interact with a specific H3 histone, playing a crucial function on kinetochore formation. To study the structure and composition of centromeric regions in the genus *Coffea*, we annotated and classified Centromeric Retrotransposons sequences from the allotetraploid *C. arabica* genome and its two diploid ancestors: *Coffea canephora* and *C. eugenioides.* Ten distinct CRC (Centromeric Retrotransposons in *Coffea*) families were found. The sequence mapping and FISH experiments of CRC Reverse Transcriptase domains in *C. canephora, C. eugenioides,* and *C. arabica* clearly indicate a strong and specific targeting mainly onto proximal chromosome regions, which can be associated also with heterochromatin. PacBio genome sequence analyses of putative centromeric regions on *C. arabica* and *C. canephora* chromosomes showed an exceptional density of one family of CRC elements, and the complete absence of satellite arrays, contrasting with usual structure of plant centromeres. Altogether, our data suggest a specific centromere organization in *Coffea*, contrasting with other plant genomes.

Keywords: coffee, CRM lineages, FISH, *Gypsy*, pseudochromosomes, proximal chromosome regions, centromeres

INTRODUCTION

LTR-retrotransposons pertain to the Class I of Transposable Elements (TEs), they move via the synthesis of an intermediate RNA using "copy and paste" mechanisms (Wicker et al., 2007). Due to their mobility, LTR-retrotransposons are the most abundant TEs (Grandbastien, 2015). They contribute to the variation of genome size and structure observed in plants (Piegu et al., 2006; Heslop-Harrison and Schwarzacher, 2011; Tenaillon et al., 2011).

LTR-retrotransposons are classified into *Copia* and *Gypsy* superfamilies according to their coding domain internal organization (Schnable et al., 2009; Gao et al., 2012; Bennetzen and Wang, 2014). Each *Copia* and *Gypsy* superfamily is sub-classified into lineages and families (Wicker et al., 2007), according to coding region similarities and overall structures (Llorens et al., 2009). For plant genomes, *Copia* is sub-classified into *Tork, Retrofit, Oryco, SIRE,* and *Bianca*, while *Gypsy* is sub-classified into *TAT, Athila, Galadriel, Reina, Del,* and *CRM* (Llorens et al., 2009, 2011), based on Reverse-Transcriptase (RT) domain phylogenetic analyses. *Gypsy* lineages are also grouped into different branches according to the presence of a chromodomain; grouping together *Galadriel, Reina, Del,* and *CRM* lineages into the Chromovirus branch.

Copia and *Gypsy* superfamilies can be found distributed in blocks or dispersed along plant chromosomes (Lopes et al., 2013; Santos et al., 2015; Zhang et al., 2017). One notable exception is the Centromeric Retrotransposon lineage of Chromovirus (*CRM* or Centromeric Retrotransposon of Maize), which appears located preferentially into proximal chromosome regions or "centromeric regions" (Nagaki et al., 2005; Bao et al., 2006; Liu et al., 2008; Du et al., 2010; Sharma and Presting, 2014). CRMs carry heterogeneous domains at the C-terminus of the integrase that may be linked to their chromosomal distribution. A chromodomain (CHRomatin Organization MOdifer domain) or a targeting domain called CR motif were identified (Houben et al., 2007; Neumann et al., 2011). These domains are probably able to interact with the CENH3 protein, suggesting that Centromeric Retrotransposons (CR) participate in centromere function. Plant centromeric regions can be composed of large arrays of CR elements inserted into specific satellite DNA (Cheng et al., 2002; Houben et al., 2007; Marques et al., 2015; Santos et al., 2015). Although relatively few centromeric regions have been studied in plants, especially due to difficulties to sequence and assemble regions with a high content of repetitive sequences, Neumann et al. (2011) separated CR elements into three groups according to their properties and chromosomal distribution: Group A carrying a CR motif and Group B lacking any targeting domain, both localized in centromeric regions; and Group C containing a chromodomain and dispersed along chromosomes.

The *Coffea* genus (Rubieaceae) comprises 125 species (Hamon et al., 2017). All species are diploids, except *Coffea arabica* ($2n = 4x = 44$), that arose from a recent hybridization between *C. canephora* and *C. eugenioides* (Lashermes et al., 1999; Yu et al., 2011). The recent sequencing of *C. canephora* genome revealed an important contribution of transposable elements (>50%). Most of them fell into the LTR-retrotransposons order (Denoeud et al., 2014). Several international sequencing initiatives are targeting the *C. arabica* genome using Pacific Biosciences (PacBio) single molecule sequencing (Mueller et al., 2015). This technique, allowing the sequencing of complex regions with a high content of repeated sequences, offers the opportunity to study the composition and organization of centromeric regions. In this study, we identified and compared 10 families of Centromeric Retrotransposons in the forthcoming PacBio genomes of *C. canephora, C. eugenioides,* and *C. arabica*. In situ hybridization using conserved RT probes showed CRCs located in proximal and interstitial chromosome regions. Finally, annotation and comparison of centromeric region rich in CRC elements revealed dynamic changes targeting LTR retrotransposons, but also the complete absence of tandem repeats usually associated with CRC elements.

MATERIALS AND METHODS

Genome Sequencing

Genomic DNA was extracted from leaves using DNeasy Plant Maxi Qiagen Kit. For long read sequencing, 20 Kb libraries were prepared following Pacific Biosciences (PacBio) protocol and Blupippin size selection. Sequencing was performed on the PacBio RSII platform, and specifications are described in Supplemental data 1. For short read sequencing, libraries were prepared with the KAPA HyperPlus kits, following manufacturer recommendation and sequenced on Illumina HiSeq2500 using PE flow cells and V4 chemistry. Genomes were assembled using Falcon and Falcon unzip from Pacific Bioscience (https://github.com/PacificBiosciences/FALCON).

In Silico Analyzes

Genomes of *C. canephora* (DH200-94-V.2), *C. eugenioides* (BU-A) and *C. arabica* (accession Et39), were kindly provided by the ACGC (2014) with the single molecule real-time (SMRT, Pacific Biosciences—PacBio). The three genomes were sequenced using the long-read Pacific Bioscience technology (Mueller et al., 2015). *C. canephora* genome assembly was finished using both Bionano genome mapping and Dovetail Hi-C scaffolding technologies (ACGC, unpublished results).

Transposable Element Annotations and Analyses

Sequenced genomes served as source for searching and comparing LTR-retrotransposons using the LTR_STRUC (McCarthy and McDonald, 2003). Putative retrotransposons sequences were classified into *Gypsy* and *Copia* superfamilies according to their similarity against the Gypsy Database protein domains (http://www.gydb.org/index.php/Main_Page) as implemented in the *Impactor* program (Orozco et al. unpublished. Available upon request). Putative reverse transcriptase (RT) domain from the *Gypsy* superfamilies were identified using BLASTX (Altschul et al., 1997) and extracted and translated into amino acids using Genewise (Birney et al., 2004) with a minimum length of 150 residues as in Guyot et al. (2016). For each coffee genome, RT domains from *Gypsy* LTR-RTs were aligned using MUSCLE (Edgar, 2004) with RT reference domains from the Gypsy Database. Aligned sequences were used to construct a bootstrapped neighbor joining phylogenetic tree (1,000 bootstrap) with ClustalW (Thompson et al., 1994), edited using FigTree (http://tree.bio.ed.ac.uk/software/figtree/).

The coffee sequences from the *CRM* lineage and called hereafter CRC (Centromeric Retrotransposons of *Coffea*) sequences were identified from the NJ tree. These sequences were sub-classified into groups according to tree conformation and bootstrap values. Groups were validated by alignments using dotter (Sonnhammer and Durbin, 1995), stretcher

(EMBOSS) and plotcon (EMBOSS). LTR sequences with 99% identity based on LTR_STRUC were annotated using Artemis (Rutherford et al., 2000). Complete (i.e., a LTR-retrotransposon containing both LTR domains) and putative autonomous (i.e., a LTR-retrotransposon containing all coding domain involved in its mobility) elements were compared and grouped with the Mauve tool (http://darlinglab.org/mauve/mauve.html). Non-autonomous elements were classified into TRIM, LARD, and TR-GAG according to their length and domains as in Chaparro et al. (2015) and implemented in the *Impactor* program (Orozco et al., unpublished). A representative element of each group was submitted to GenBank under the following accession: A MG242426; B MG242427; C MG242428; D MG242429; E MG242430; F MG242431; G MG242432; H MG242433; Y MG242434; X MG242435.

In Silico Estimation of CRC Elements Copy Number and Distribution

Assessment of the CRC elements copy number in *C. canephora*, *C. arabica,* and *C. eugenioides* PacBio sequences was done as in Dupeyron et al. (2017). Briefly, each representative copy of CRC groups was used for similarity searches against genomes using Censor (http://www.girinst.org/downloads/software/censor/). Copies are sorted according to their completeness and percentage of similarity when compared to the representative copy. Insertion times of selected LTR-RT were estimated as proposed by SanMiguel et al. (1998) and Guyot et al. (2016), with a substitution rate of 1.3×10^{-8}, established by Ma and Bennetzen (2004). The distribution of RT domains was carried out using RepeatMasker (–div 20 option) while the distribution of complete elements, LTR and non-autonomous elements was performed using Censor with a minimum of 80% of nucleotides identity and 80% of sequence coverage.

The centromeric regions annotation was performed using RepeatMasker (–div 20 option) and edited with Artemis, and transposable elements density along genomic sequences was carried out using DensityMap (Guizard et al., 2016).

Plant Materials, DNA Extraction, and Probes Production

Seedlings of *C. arabica*, *C. canephora,* and *C. eugenioides* were obtained from the Agronomic Institute of Paraná (IAPAR), Londrina, Paraná, Brazil, cultivated in pots in the green house of the Laboratory of Cytogenetics and Plant Diversity, State University of Londrina, Brazil. DNA extraction was performed as described by Romano and Brasileiro (1999). Quickly, young leaves were collected, macerated in liquid nitrogen and treated with CTAB extraction buffer. DNA was purified with phenol:chloroform (1:1, v:v) and chloroform:isoamyl alcohol (24:1, v:v) and precipitated in absolute ethanol. DNA concentration was estimated using a NanoDrop 2000 Spectrophotometer (Thermo Scientific). Primers were designed using OligoPerfect^{TM}Designer (http://tools.lifetechnologies.com). A conserved region located in the predicted Reverse Transcriptase (RT) coding region of each CRC group was amplified by PCR using a pair of RT primers (Forward: 5′ACTGTCGGGCTGTAAATGCT; Reverse: 5′CTGCGAACTCACGACATAGC). Reactions were done using *C. arabica, C. canephora,* and *C. eugenioides* genomic DNA as template, in a mix composed by 0.5 μL Taq Polymerase (5 U/μL), 2.5 μL 10 × buffer, 2.5 μL $MgCl_2$ (50 mM), 1 μL of dNTP (10 mM), 1 μL of each primer at 10 mM and H_2O, in a final volume of 25 μL. Reactions were checked with 1% agarose gel electrophoresis. Probes were obtained by PCR, using the product of a first PCR as template, in a new reaction containing dGTP (25%), dCTP (25%), dTTP (25%), dATP (17.5%), and Cy3-dUTP (7.5%).

Cytogenetic Analyses

Mitotic chromosomes were obtained from root tips treated with a saturated solution of paradichlorobenzene (PDB) for 1 h at room temperature plus 23 h at 14°C. Samples were fixed in a fresh solution of methanol: acetic acid (3:1, v:v) for 24 h, and stored at −20°C, or used immediately. Root-tips were softened in 2% cellulase plus 20% pectinase (v:v), both Sigma, at 37°C for 5 h, and squashed in a drop of 60% acetic acid. The cover slips were removed after freezing in liquid nitrogen, slides were air dried and used in FISH or C-CMA/DAPI banding procedures.

For FISH, a mixture of 30 μL containing 100% formamide (15 μL), 50% polyethylene glycol (6 μL), 20× SSC (3 μL), 100 ng calf thymus DNA (1 μL), 10% SDS (1 μL), and 100 ng probes (4 μL), was treated at 70°C for 10 min, placed on ice and immediately applied to the samples. Denaturation/hybridization was performed at 95, 50, and 38°C, 10 min each, followed by 37°C overnight in a humidified chamber. Post-hybridization washes were carried out in SSC buffer with about 70% stringency, mounted in 23 μL antifade solution (90% glycerol, 2.3% DABCO, 2% 20 mM Tris–HCl, pH 8.0, plus 1 μL of 2 μg/mL DAPI, and 1 μL of 2.5 mM $MgCl_2$).

Chromosome banding was done using 3 days aged slides incubated in a solution of 45% acetic acid, 5% barium hydroxide, and 2× SSC, pH 7.0 (Schwarzacher et al., 1980, with modifications). Samples were stained with 0.5 mg/mL CMA_3 for 1.5 h and 2 mg/mL DAPI for 30 min, and finally stained with a medium composed of glycerol/McIlvaine buffer (pH 7.0) 1:1 plus 2.5 mM $MgCl_2$. FISH and C-CMA/DAPI chromosome images were acquired in gray-scale mode using a Leica DM4500B microscope, equipped with a Leica DFC300FX camera, and overlapped with blue for DAPI, greenish-yellow for CMA and red for Cy3, and processed using the Leica LAS software. Images were optimized for contrast and brightness using the GIMP 2.8 Image Editor.

RESULTS

The *Gypsy* Superfamily and the *CRM* Lineage in Coffee Genomes

The search for complete LTR-retrotransposons sequences in *C. canephora*, *C. eugenioides* and *C. arabica* allowed to recognize 7,195, 3,590, and 3,877 elements, respectively. These were predicted and classified into 1,021 *Copia* and 2,222 *Gypsy* (*C. canephora*), 668 *Copia* and 950 *Gypsy* (*C. eugenioides*)

and 743 *Copia* and 1226 *Gypsy* (*C. arabica*). The remaining predicted elements were identified into non-autonomous LTR-retrotransposons or into unclassified autonomous elements according to similarities to GAG-POL regions available at the Gypsy Database. For the *Gypsy* superfamily, the LTR-retrotransposon lineages (*Del, Galadriel, Reina, CRM, Athila,* and *TAT*), were found in the three *Coffea* genomes, using a BLAST based analysis and a RT based phylogenetic analysis (Supplemental datas 2, 3), and the *CRM* lineage was particularly analyzed.

The *CRM* lineage represented 499, 223, and 262 of complete annotated elements in *C. canephora, C. eugenioides, and C. arabica* genomes, respectively. Manual inspection revealed that 367 (73.55%), 124 (55.61%), and 113 (43.13%) elements were found complete for *C. canephora, C. eugenioides, and C. arabica*, respectively, since no large deletion affected these sequences. The RT amino acid sequences of the *CRM* lineage from the three coffee species were grouped together, aligned and displayed with a N.J. phylogenetic tree (Supplemental data 4). Ten phylogenetic groups were defined according to the structure and similarity of these domains (**Table 1** and Supplemental data 5). The Centromeric Retrotransposons of *Coffea* were grouped and named here as follow A, B, C, E, D, F, G, H, X, and Y.

About 60 CRC sequences per genome, from the different groups, showing >99% of nucleotide identity between both LTR of the same element were carefully annotated and compared (**Figure 1**). Only elements from the A group presented a chromodomain, with zinc finger/HHCC motif at their C-terminus downstream the INT region, while elements from other groups exhibited a CR motif (**Figure 1B**) at their C-terminal regions, and a poly-A motif upstream the GAG region (data not shown). Autonomous elements of each group showed a variable length from 5,971 bp (Group_A) to 8,088 bp (Group_D), and a LTR size from 661 bp (Group_F) to 781 bp (Group_Y).

The alignment of complete elements into a matrix of nucleotides comparison showed discontinuous lines between groups, suggesting interrupted conservation along the different CRCs (**Figure 1A**). This discontinuous similarity was also confirmed with a nucleotide similarity plot of the full-length sequences of the 10 CRC groups (**Figure 1C**). The RT domain comparison at the nucleotide level showed a high conservation among elements within each group, independent of the species they are issued (from 80 to 98%), and a distant conservation between elements of different groups, i.e., from 45 to 64% (**Table 1**). These results suggest that CRCs are distributed among different families in the *Coffea* genus.

Non-autonomous CRC Elements in *Coffea*

Non-autonomous CRC elements, lacking any coding regions as seen in Terminal Repeat in Miniature (TRIMs) or Large Retrotransposon Derivative (LARDs), or lacking the POL polyprotein region as in TR-GAGs, were also identified (Chaparro et al., 2015). CRC group alignments (80% identity cutoff) against the putative non-autonomous elements exhibited different structures, such as TRIMs (only in *C. canephora*), LARDs and TR-GAGs. The counting showed 268, 216 and 216 putative non-autonomous CRC for *C. canephora*, *C. eugenioides,* and *C. arabica*, respectively (Supplemental data 6). Among them, the group B (mainly TR-GAG elements), the H (mainly LARD elements) and the group C, showed the highest number of copies, whatever the genome analyzed. Only the chromodomain of group A did not show similarity to any non-autonomous element.

In Silico Copy Number Estimation and Insertion Time of 10 CRC Families

A total copy number of 359, 278, and 473 CRC elements (with >80% of both coverage and identity) were found in *C. canephora*, *C. eugenioides,* and *C. arabica*, respectively. Besides conserved copies, fragmented copies (with >10% of coverage and >80% of identity) represented 2,055, 2,064, and 3,478 CRC elements in *C. canephora*, *C. eugenioides,* and *C. arabica*, respectively (**Table 2**). For the three species, elements from the groups H and B outnumbered the other groups for complete (80-80) and fragmented copies (80-10). The allotetraploid genome of *C. arabica* contains, as expected, the highest copy number when compared to the diploid genomes of *C. canephora* and *C. eugenioides*.

The nucleotide divergence and relative insertion time of complete CRC copies suggest a relatively recent insertion, or a high conservation of the whole sequences with a similar pattern in *C. arabica, C. eugenioides,* and *C. canephora* (Supplemental data 7A). For each CRC group, three peaks of copy number accumulation were observed for the H group in *C. canephora, C. eugenioides* and *C. arabica*, while for the C group four and two peaks of copy number accumulation were noted for *C. eugenioides*, and for *C. canephora* and *C. arabica* (Supplemental datas 7B–D). Other and successive small peaks of copy number accumulation were observed for the E group, for example. This result suggested that the insertions of CRC are relatively recent, but that ancient activities may be detected, particularly for the group H.

The distribution of CRC RT sequences along the *C. canephora* pseudochromosomes (**Figure 2**) showed that for some of them there is a clear accumulation of RT sequences in the central regions (pseudochromosomes 1, 2, 4, 5, 6, 8, 9, and 10). For the others, RT sequences were less concentrated, exhibiting a dispersed pattern, such as in the pseudochromosomes 3, 7 and 11. When we compare the distribution of these sequences of each CRC group along pseudochromosomes, it is possible to note that only the groups E and H showed a clear accumulation into median regions (**Figure 2**).

Cytogenetic Analysis

FISH using a probe for RT conserved region, common for all CRC groups (Supplemental data 8), showed signals with differences in sizes and brightness on *C. arabica, C. canephora,* and *C. eugenioides* nuclei. Signals were distributed in all regions of differentiated cell nuclei (**Figures 3A,F**, **4A–C**), and in a Rabl-like organization in undifferentiated cells (**Figure 3E**). Brighter signals appeared located in the proximal chromosome regions (see **Table 3**), but with variations within and between karyotypes of diploid species *C. canephora* with two signals (**Figures 3B–D**) and *C. eugenioides* with four signals (**Figures 3G–I**), and with

TABLE 1 | Matrix of RT domain identity between CRC groups in *Coffea eugenioides*, *C. canephora,* and *C. arabica*.

		CR Groups							
	C. canephora	**A (%)**	**B (%)**	**C (%)**	**D (%)**	**E (%)**	**F (%)**	**G (%)**	**H (%)**
C.eugenioides	X	48	57	61	57	60	61	64	56
	A	88	49	48	44	48	47	47	45
	B	48	91	58	54	57	58	58	55
	C	45	56	82	57	59	58	59	57
	E	49	57	60	59	92	61	61	57
	F	47	57	60	59	61	93	61	59
	H	46	55	58	57	58	60	59	80
	C. arabica	**X (%)**	**Y (%)**	**B (%)**	**C (%)**	**E (%)**	**F (%)**	**G (%)**	**H (%)**
C.eugenioides	X	87	59	58	60	60	60	64	57
	A	46	47	48	47	47	47	47	44
	B	57	57	97	57	57	57	58	54
	C	59	59	56	84	60	57	60	57
	E	59	60	57	60	93	61	61	57
	F	60	59	58	60	61	90	61	59
	H	58	56	55	57	58	59	59	80
	C. arabica	**Y (%)**	**X (%)**	**B (%)**	**C (%)**	**E (%)**	**F (%)**	**G (%)**	**H (%)**
C. canephora	A	47	47	49	48	48	48	47	45
	B	56	57	91	58	58	57	58	54
	C	61	59	58	93	60	60	60	56
	D	57	58	54	56	60	57	58	57
	E	60	60	58	60	97	61	61	57
	F	59	60	58	59	61	94	61	59
	G	60	63	58	60	61	60	98	58
	H	56	57	55	56	57	58	58	92

The letters A, B, C, D, E, F, G, H, X, and Y correspond to CRC groups, as defined by the phylogenetic analysis. Values highlighted in gray represent the highest percentage of identity observed between groups.

six signals in the allotetraploid *C. arabica* (**Figures 4D–I**). In addition to the predominant signals into proximal regions, few chromosomes displayed scattered signals in proximal/interstitial dots (except for *C. eugenioides*). This is probably due to a smaller copy numbers of CRC RT sequences in these chromosomes. Chromosomes with few or undetectable FISH signals were also observed in *C. canephora* (one pair), *C. eugenioides* (one pair), and *C. arabica* (two chromosome pairs).

The C-CMA/DAPI banding indicated that C-CMA$^+$/DAPI$^-$ were associated to NOR bearing chromosomes in these three species. In *C. canephora* and *C. arabica,* C-CMA$^+$/DAPI$^+$ bands were accumulated in proximal regions (**Figures 5A,B,E,F**), while these bands were absent or few accumulated in *C. eugenioides* (**Figures 5C,D**). In this last species, C-CMA$^+$/C-DAPI$^-$ bands seem to be inconspicuous in the proximal regions of some chromosomes and absent in most of them (**Figures 5C,D**). These results showed also that C-CMA$^+$ and C-DAPI$^+$ heterochromatin can be co-localized with RT CRC hybridization signals for *C. canephora* and *C. arabica* chromosomes, but not for *C. eugenioides*.

The *C. canephora* and *C. arabica* Chromosome 5 Putative Centromeric Regions Are Enriched of CRC Elements

Based on the FISH data and localization of RT CRC on *C. canephora* genome sequences, the pseudochromosome 5 has been selected for further analysis. The density of transposable elements (light green, annotated on *C. canephora;* Denoeud et al., 2014) and full-length CRC elements (dark green) were displayed along the pseudochromosome 5 from *C. canephora* (**Figure 6A**) and along the pseudochromosome 5 sub-genome *C. canephora* from *C. arabica* (**Figure 6B**). Data showed a high density of CRC elements in the median part for both orthologous pseudochromosomes. A dot-plot of 4 Mb length around these regions in *C. canephora* and *C. arabica* (**Figure 6C**), suggest a conservation where CRC elements density (dark green) is the highest. Annotations of highest density regions containing CRC elements of *C. canephora* and *C. arabica*, with 1.2 Mb and 800 kb length, respectively (**Figure 6D**), revealed that 94.1 and 91.7% of these regions consisted of transposable elements. LTR retrotransposons and non-autonomous derivatives represent

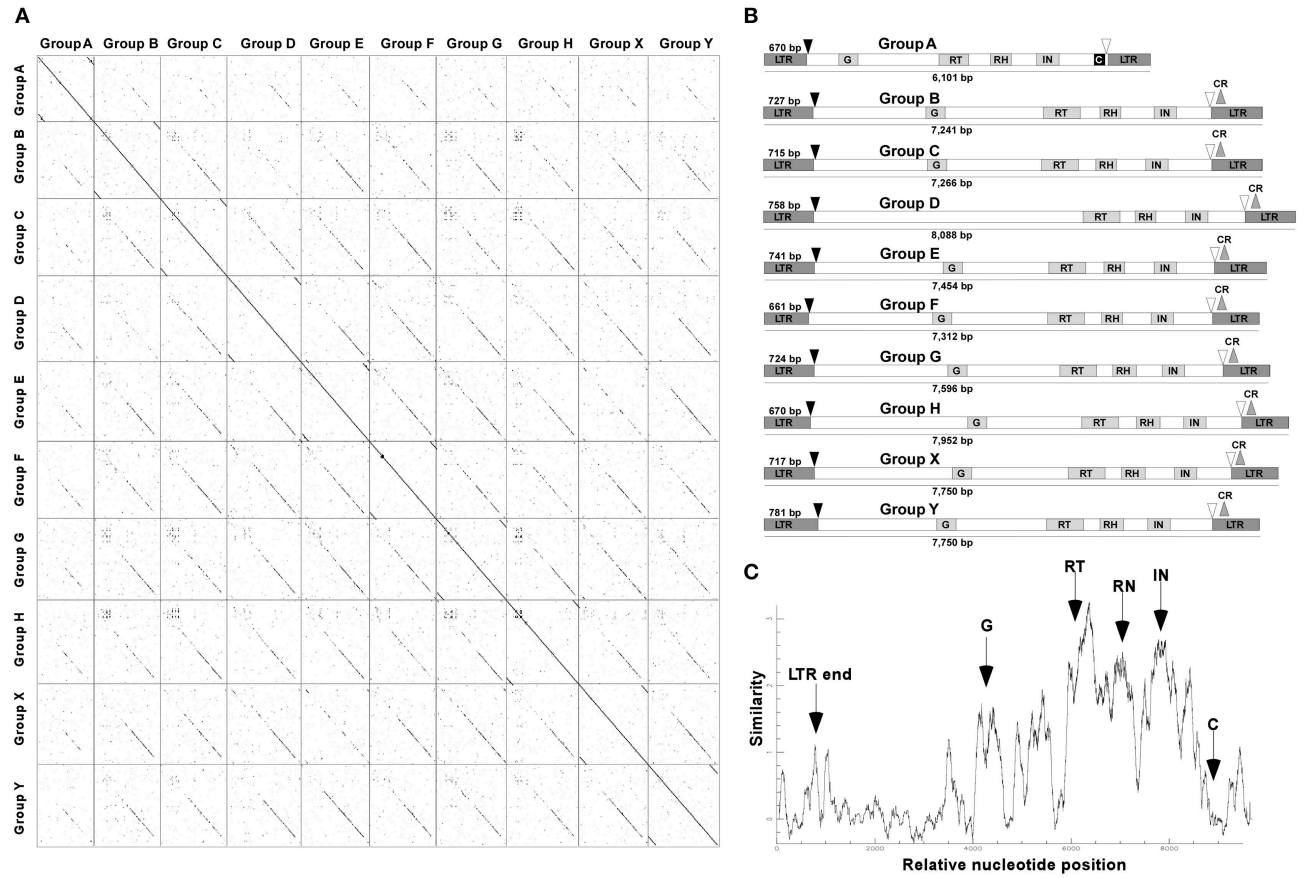


FIGURE 1 | Structure and conservation of the *Gypsy* CRC LTR-retrotransposons in *Coffea arabica*, *C. canephora* and *C. eugenioides*. **(A)** Dotter alignments between the 10 groups of CRC found by LTR_STRUC against themselves. **(B)** Structural features of CRC groups. LTR, Long Terminal Repeats; G, GAG domain; RT, Reverse Transcriptase; RH, RNAse H; IN, Integrase; C, Chromodomain; CR, CR motif. The dark arrows indicate the PBS sites while the white arrows indicate the PPT sites. **(C)** Nucleotide similarity plot with the 10 groups of CRC. The positions of the different domains are indicated.

84.4 and 79.7% and CRC elements represent 33.7 and 35% in *C. canephora* and *C. arabica*, respectively, whereas transposons account for 0 and 0.7%. Interestingly, the CRC family H, represents alone 17.84 and 25.28 of the analyzed regions in *C. canephora* and *C. arabica,* suggesting a local enrichment. Beside CRC, the Del lineage is the most redundant with 15.9 and 9.6%. A detailed annotation was performed for the centromeric region of *C. arabica* pseudochromosome 5. Ninety-one complete or partial CRC elements were annotated for which 76 fell into the H family. Twenty-three complete and 13 putative non-autonomous CR elements carrying both intact LTR ends were recovered and their insertion times were estimated. Seventeen of them have a very recent insertion time (>1 Mya), similarly to estimation at the genome scale (Supplemental data 7). In these regions rich in CRC elements, no tandem repeats were observed in *C. canephora* and in *C. arabica* assembled sequences. Insertion of CRC elements into tandem arrays were directly searched in raw *C. canephora* PacBio reads, before their assembly, using BLAST and dot-plot. Here again no tandem repeats associated with CRC elements of the H family were found. The density of transposable elements (light green, annotated on *C. canephora;* Denoeud et al., 2014) and full-length CRC elements (dark green) were also displayed along all pseudochromosome from *C. canephora, C. arabica and C. eugenioides* (Supplemental datas 9–11). Most of the pseudochromosomes showed a clear peak of accumulation.

DISCUSSION

Characterization of CRC Elements in *Coffea* Yields 10 Distinct Groups

Despite numerous centromeric retrotransposons elements identified in monocot and dicot species (Neumann et al., 2011), their diversity and classification into types, as well as their respective contribution to the structure of centromeric regions is poorly known for most higher plant groups. In this study, we identified 10 groups of Centromeric Retrotransposons of *Coffea* (CRC) in the genomes of *C. arabica*, an allotetraploid species and its two diploid parents, *C. canephora* and *C. eugenioides*. This work was based on high coverage of PacBio reads used for *C. arabica, C. canephora,* and *C. eugenioides* genomes produced by the ACGC (Mueller et al., 2015). Centromeric

TABLE 2 | Estimation of the copy numbers of CRC elements in the *Coffea canephora*, *C. eugenioides,* and *C. arabica* genome sequences.

	C.canephora **copies (80–80)**	***C.canephora*** **partial copies (80–10)**	***C.eugenioides*** **copies (80–80)**	***C.eugenioides*** **partial copies (80–10)**	***C.arabica*** **copies (80–80)**	***C.arabica*** **partial copies (80–10)**
Group A	8	85	7	103	13	156
Group B	81	841	66	705	121	1,149
Group C	18	86	16	202	28	303
Group D	6	63	19	96	18	164
Group E	49	259	39	188	50	476
Group F	60	144	29	148	63	265
Group G	47	90	3	55	20	153
Group H	84	412	88	460	142	674
Group X	1	13	4	0	7	17
Group Y	5	62	7	107	11	121
Total	359	2055	278	2,064	473	3,478

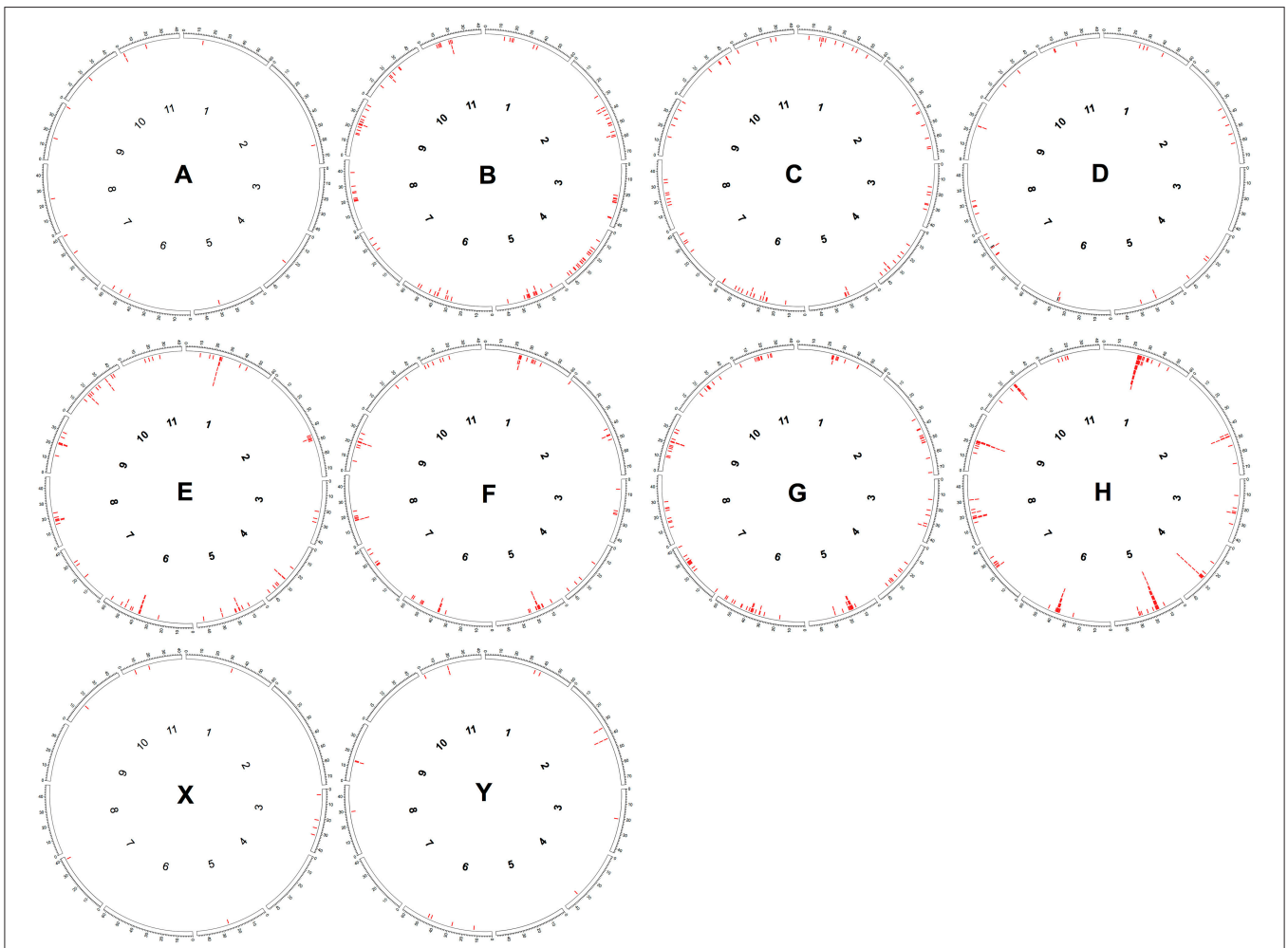


FIGURE 2 | *In silico* distribution of RT domains from CRC groups along PacBio assembled pseudochromosomes of *Coffea canephora.* Each circle represents the distribution of one CRC group. Red lines represent the position of RT domains as found by RepeatMasker.

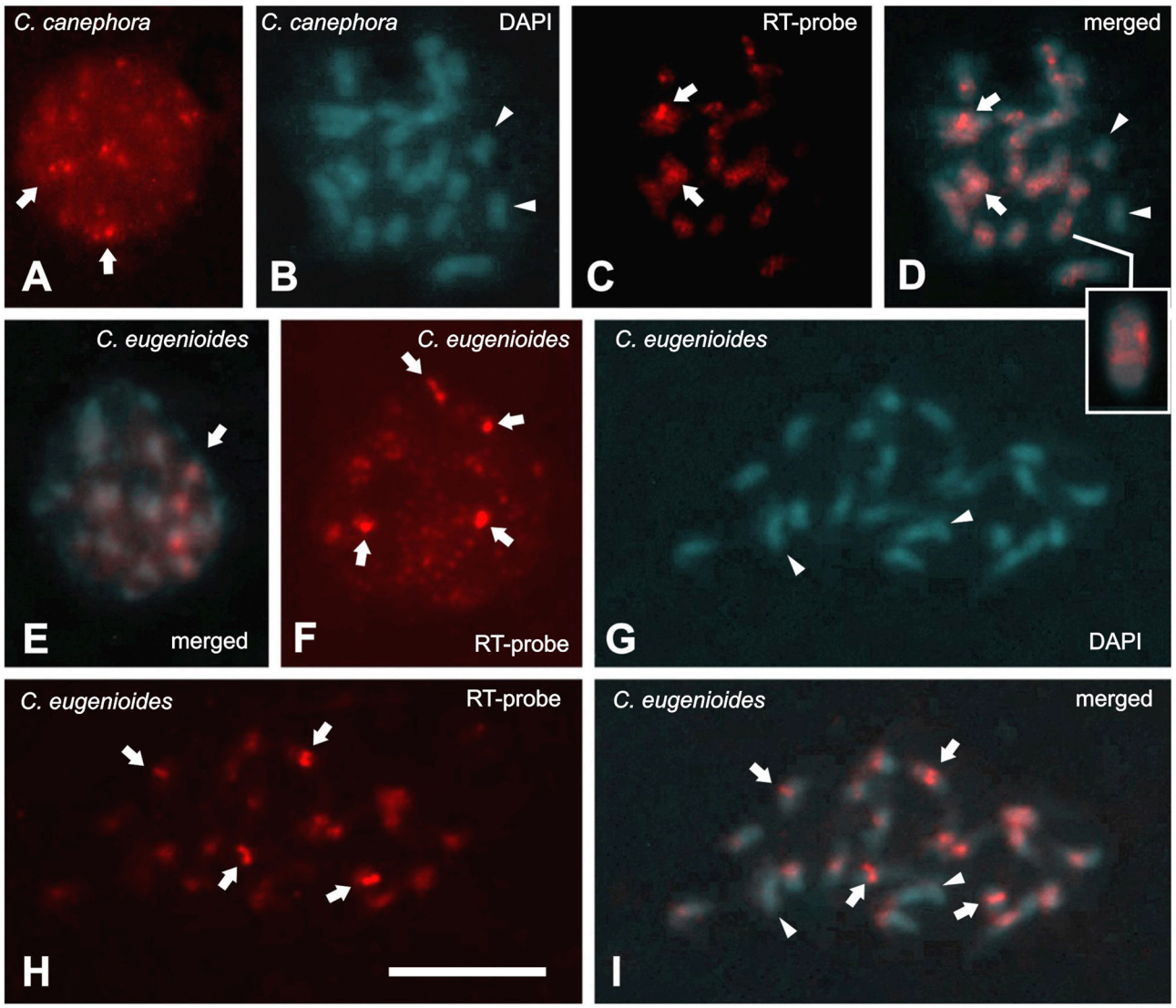


FIGURE 3 | Fluorescence *in situ* hybridization (FISH) in nucleus and metaphases stained with DAPI (blue) and RT-CRC probe hybridized with Cy3-dUTP (red) in *Coffea canephora* **(A–D)** and *C. eugenioides* **(E–I)**. **(A)** Nucleus with scattered signals and two brighter signals (arrows). Metaphase stained with DAPI **(B)**, showing RT-CRC FISH signals **(C,D)** in the centromeres, proximal regions, including few chromosomes with scattered signals and proximal/interstitial dots (box), in red acquired and merged images. **(E)** Undifferentiated nucleus of *C. eugenioides* (Cy3/DAPI merged), showing scattered signals and four brighter signals Rabl-like organized, that are typical of centromeric location. Scattered and four large signals can also be observed in the red stained unpolarized nucleus **(F)**. Arrows point out the large FISH signals. **(G–I)** Prometaphase stained with DAPI and hybridized with RT-CRC probe. FISH indicates a predominance of centromeric-pericentromeric signals, including the four large signals detected in the nuclei (arrows). Arrowheads in **B, D, G,** and **I** indicate chromosomes without hybridization signals. Bar = 10 μm.

retrotransposons in plants were initially organized into three groups, based on the presence of a CR domain extending into the 3′ LTR and a chromodomain at the C terminus of the POL polyprotein (Neumann et al., 2011). In *Coffea* the 10 identified groups fall into two of these groups: those possessing a CR motif (most of them, group "A" from Neumann et al. (2011), corresponding to our B, C, D, E, F, G, H, X, and Y groups) and those carrying a terminal chromodomain-like (group "C" from Neumann et al. (2011), corresponding to our A group). These data indicate that centromeric retrotransposons could be more diverse in plants than previously proposed by Neumann et al. (2011).

Chromodomain might target integration of chromovirus LTR retrotransposons into heterochromatic chromosome regions (Novikova, 2009), and these specificities could allow the *CRM* accumulation into proximal chromosome regions, such as in *Coffea*, or may be still associated with epigenetic mechanisms (Houben et al., 2007; Neumann et al., 2011). However, most of CRC groups (B, Y, C, E, D, F, G, X, and H) that are similar to the "C" group of Neumann et al. (2011), did not have any chromodomain nor zinc finger domains, but carried a CR motif. This motif appears particularly important for centromeric retrotransposons to target the heterochromatin (Gao et al., 2008), but they are probably not associated with epigenetic changes

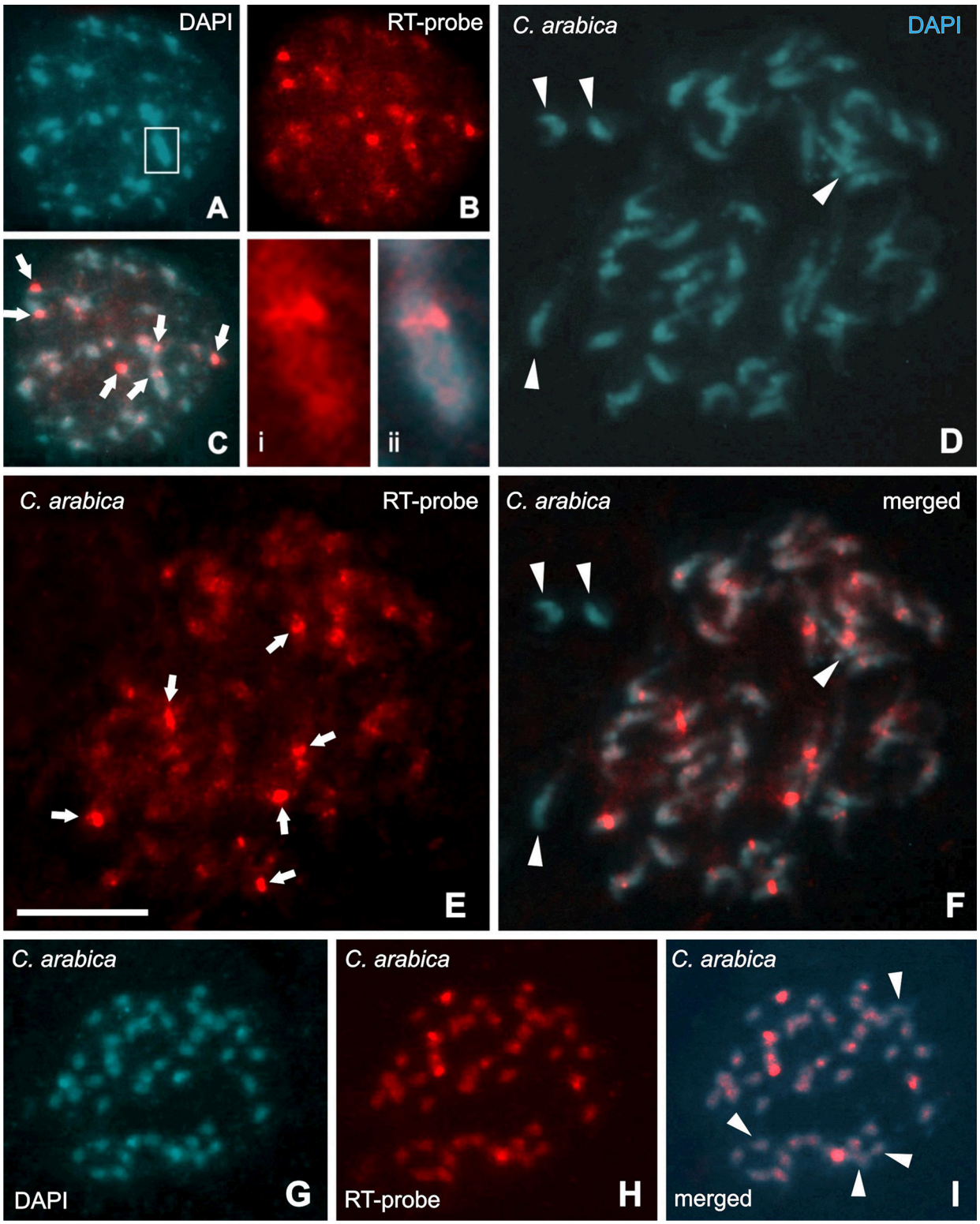


FIGURE 4 | Fluorescence *in situ* hybridization (FISH) in nucleus, prometaphases and metaphases of *Coffea arabica.* Samples stained with DAPI appear in **(A,D)**, and with RT-CR FISH signals (red) are in the others. Nucleus showing scattered signals and with six brighter signals **(B)**, that are better observed in the merged image in **(C)** (arrows). Boxes **i** and **ii** (merged) show a well-defined RT-CRC signal into regions with more condensed chromatin. Prometaphases and metaphases hybridized with the RT-CRC probe **(E–I)** showing scattered signals, but with predominance of concentrated signals in the centromeric-pericentromeric regions (arrows in **E**) Arrowheads in **(D,F,I)** indicate chromosomes without hybridization signals. Bar = 10 μm.

TABLE 3 | Cytogenetic distribution of CRC RT domains in *Coffea canephora*, *C. eugenioides,* and *C. arabica*.

Chromosome Location	Chromosomal pairs with FISH signals		
	C. canephora	*C. eugenioides*	*C. arabica*
Centromeric	7	7	9
Proximal & dispersed	1	2	7
Proximal & interstitial dots	1	0	3
Interstitial & dispersed	1	1	1
No signals	1	1	2
Total	11	11	22

in H3 histones (Neumann et al., 2011). The "B" group of centromeric retrotransposons, as defined by Neumann et al. (2011), without CR motif nor chromodomain, was not identified in the autonomous elements set in *C. arabica, C. canephora,* and *C. eugenioides* genomes. This group has been probably lost or degenerated during the evolution of the *Coffea* genus, since group "B" was identified in other dicotyledonous, such as *Vitis*, *Arabidopsis*, *Medicago,* and *Populus* (Neumann et al., 2011). Another possibility is that the group B of Neumann has been lost or degenerated earlier during the evolution of the Rubiaceae family or the Asterids branch of dicots, because the genera previously mentioned belong to the Rosids branch.

Non-autonomous centromeric retrotransposons identified in *Coffea* belong to different families: (TRIM, Witte et al., 2001), (LARD, Kalendar et al., 2004), or lacking the POL polyprotein region such as TR-GAG (Chaparro et al., 2015). This last family was also found in rice (Nagaki et al., 2005). Non-autonomous CRC shared similarities with the nine autonomous CRC groups containing CR motif, suggesting a direct relationship between autonomous and non-autonomous elements, as well as they could indicate that non-autonomous CRC may use the enzymatic machinery of complete elements for their own mobility (Wicker et al., 2007).

In Silico Copy Numbers and Insertion Time of CRC Families

The *C. arabica* genome contains a higher number of complete CRC copies than the related diploid *C. canephora* or *C. eugenioides* genomes, and it is in accordance to relationships between the polyploidization and copy number variation observed for other retrotransposons in allopolyploid genomes (Parisod et al., 2010). However, the cumulative number of CRC copies is higher for the two diploid than for the allotetraploid species, suggesting that changes occurred either during the hybridization steps leading to *C. arabica* or very recently, after the hybridization. CRC groups may have been amplified very recently in these three genomes, but with higher amplitude in *C. canephora* during the last million years. However, it remains unclear if the CRC copy number variation is only due to differential rates of amplification or if this variation is due to an efficient process of elimination via unequal or illegitimate recombination (Bennetzen, 2007). Two groups with the highest copy number (B and H) in the three species also showed recent peaks of insertion time, suggesting they were amplified recently in the *Coffea* genomes. The only exception is the B group of *Coffea*, which seems to have an ancient origin in *C. arabica*. The number of *C. arabica* CRC observed in present days compared to its progenitors should be carefully interpreted, because the present germplasms of *C. canephora* and *C. eugenioides* studied recently can have accumulated some differences in relation to those which gave rise to the amphidiploidy in *C. arabica*. In addition, we have also to consider that the worldwide *C. arabica* collection had been originated from a few Ethiopian individuals (Carvalho, 1946), and they have been extensively submitted to agronomic breeding selection.

The E and H CRC Groups Target Putative Centromeric Regions in *Coffea*

Along plant chromosomes, *Copia* and *Gypsy* superfamilies can be found distributed in blocks and scattered (Lopes et al., 2013; Santos et al., 2015; Zhang et al., 2017). One notable exception is the *Gypsy* Centromeric Retrotransposon lineage, located preferentially into centromeric and proximal regions (Du et al., 2010; Sharma and Presting, 2014). In *Coffea* species, the distribution of CRC families showed two contrasting situations. One family, the B group, appears scattered along *C. canephora* pseudochromosomes, whereas the H and, in a lesser extent, the E group, appeared clustered into proximal chromosome regions, as expected for Centromeric Retrotransposons (Sanseverino et al., 2015).

Although it was possible to separate 10 CRC groups using complete sequences, the high identity (>90%) of RT regions made difficult the design of specific primers for each group. While specific FISH for each CRC family was impossible with RT-domains, other and more divergent regions such as LTR or GAG gave inaccurate results.

Results of FISH using a generic RT-CRM probe is in agreement with a targeting of chromodomain and CR motif into centromeric regions associated to CENH3 (Houben et al., 2007; Neumann et al., 2011; Li et al., 2013), suggesting an interaction between these elements and centromeric proteins.

Our cytological observations suggested that the hybridization profile is variable among species and chromosomes in *Coffea*. In *C. eugenioides,* FISH signals were strictly associated to centromeric regions, whereas in *C. canephora* and *C. arabica* signals appear less specific to centromeres, and scattered along interstitial regions. This could be the result of a small CRC RT copy numbers hybridized. We hypothesize the two pairs without bright signals in *C. arabica* could be homologous chromosomes to those without FISH signals from the parental genomes (one pair each). Scattered FISH signals using CR probe were also reported in *Saccharum spontaneum* (Zhang et al., 2017). Surprisingly one chromosome pair in *C. canephora* and *C. eugenioides* and two in *C. arabica* did not exhibit evident centromeric signals. All these variable hybridization patterns could be associated also with differential occurrence of proximal C-CMA$^+$/DAPI$^+$ bands, that were observed in

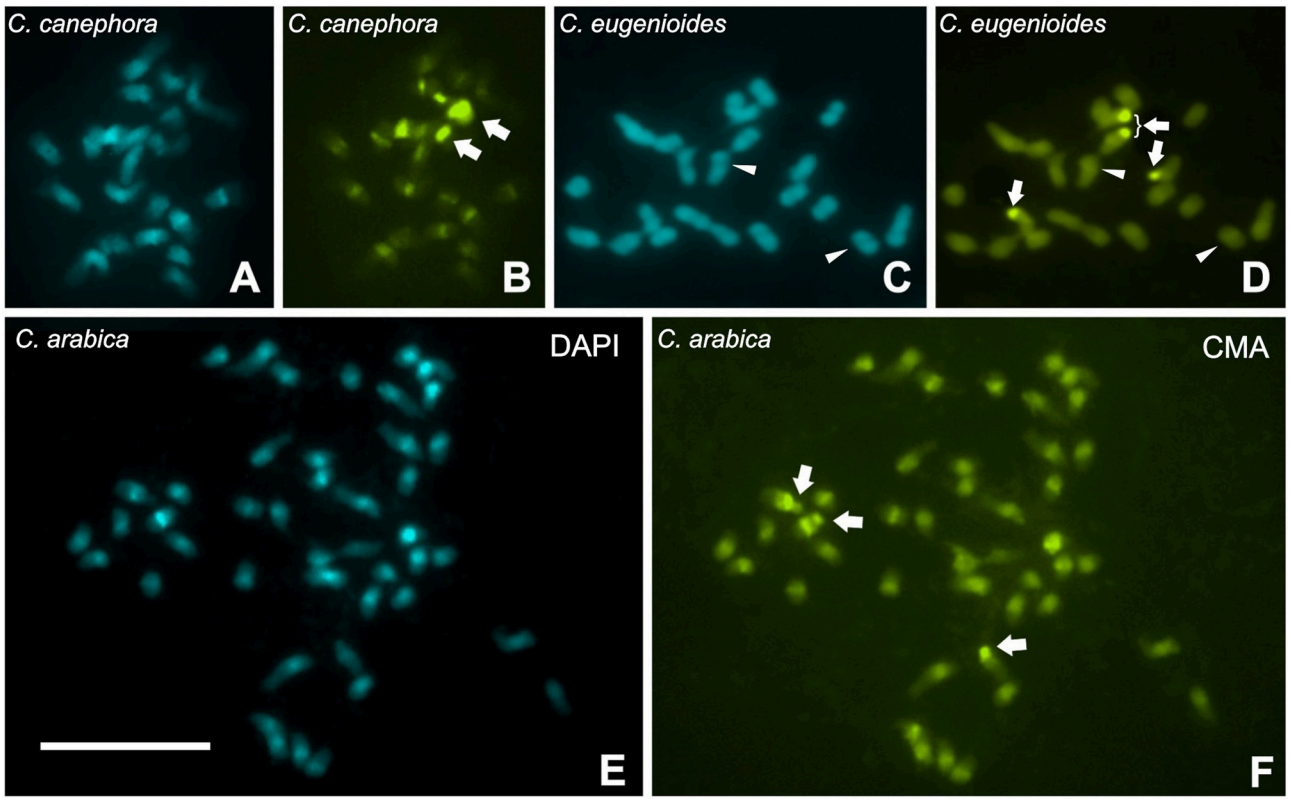


FIGURE 5 | C-CMA/DAPI banding in *Coffea canephora* **(A,B)**, *C. eugenioides* **(C,D)**, and *C. arabica* **(E,F)**, showing an accumulation of C-CMA$^+$/DAPI$^+$ bands in the proximal regions of *C. canephora* and *C. arabica*, and absence of these bands in *C. eugenioides*. However, *C. eugenioides* seems to be inconspicuous C-CMA$^+$/DAPI$^-$ bands (thin bands of difficult visualization indicated as arrowhead), in the proximal regions of some chromosomes that are not present in the other two species **(C,D)**. Arrows indicate C-CMA$^+$/DAPI$^-$ bands accumulated in the terminal regions that are associated to nucleolar organizing regions (data not shown). Bar = 10 μm.

C. canephora and *C. arabica*, and absent or difficult to distinguish in *C. eugenioides*. The heterochromatin accumulation may be associated with increase and expansion of CRC elements beyond the centromere toward the interstitial regions observed in *C. canephora* and *C. arabica*. However, additional tests are necessary to confirm this assumption, especially in relation to equilocal dispersion (Schweizer and Loidl, 1987) of repetitive DNA families into proximal regions of *Coffea* chromosomes. In addition, it is possible that, CR elements containing the 3′ terminal CR motif, and that represent a fraction of the all CR families, would be more likely inserted into the putative centromeric regions, while the other CRCs (lacking the CR motif) could be less specific and occupy other chromosomal regions.

CRC elements carrying a CR motif may also present diverse pattern of insertion, i.e., they can be specific to putative centromeric regions (E and H groups) and/or to interstitial regions (B group). The presence of the CR motif may be not the *sine qua non* condition for a putative centromeric targeting and that other mechanisms may intervene for chromosomal regions targeting by chromoviruses in plants. Chromatin immunoprecipitation followed by sequencing (ChIP-Seq) using antibodies against the centromere-specific histone H3 of Coffee are now required to validate putative centromeric regions as active centromeres.

The Putative Centromeric Region of Chromosome 5 Is Mainly Composed of the H Family

Repetitive DNA families, such as centromeric retrotransposons and tandem repeats, participate in the complex organization of centromeric regions, especially of the kinetochore formation (Neumann et al., 2011). In *Coffea*, 23 CRCs were predicted as elements that have some role in the centromeric regions, as observed in other plant groups (Han et al., 2010; Sanei et al., 2011). However, it has not been yet clarified what CRC types (complete, truncated, partial, or non-autonomous) may participate in kinetochore formation. The presence of partial and truncated elements on proximal chromosome regions suggests that unequal and illegitimate recombination mechanisms may also act on centromeric regions in a neutral manner (Bennetzen, 2007). CR elements were frequently associated with satellite DNA repeats in centromeric regions of other plant species (Cheng et al., 2002; Lim et al., 2007), except for the wheat chromosome 3B, only composed of CRW retrotransposons families (Li et al.,

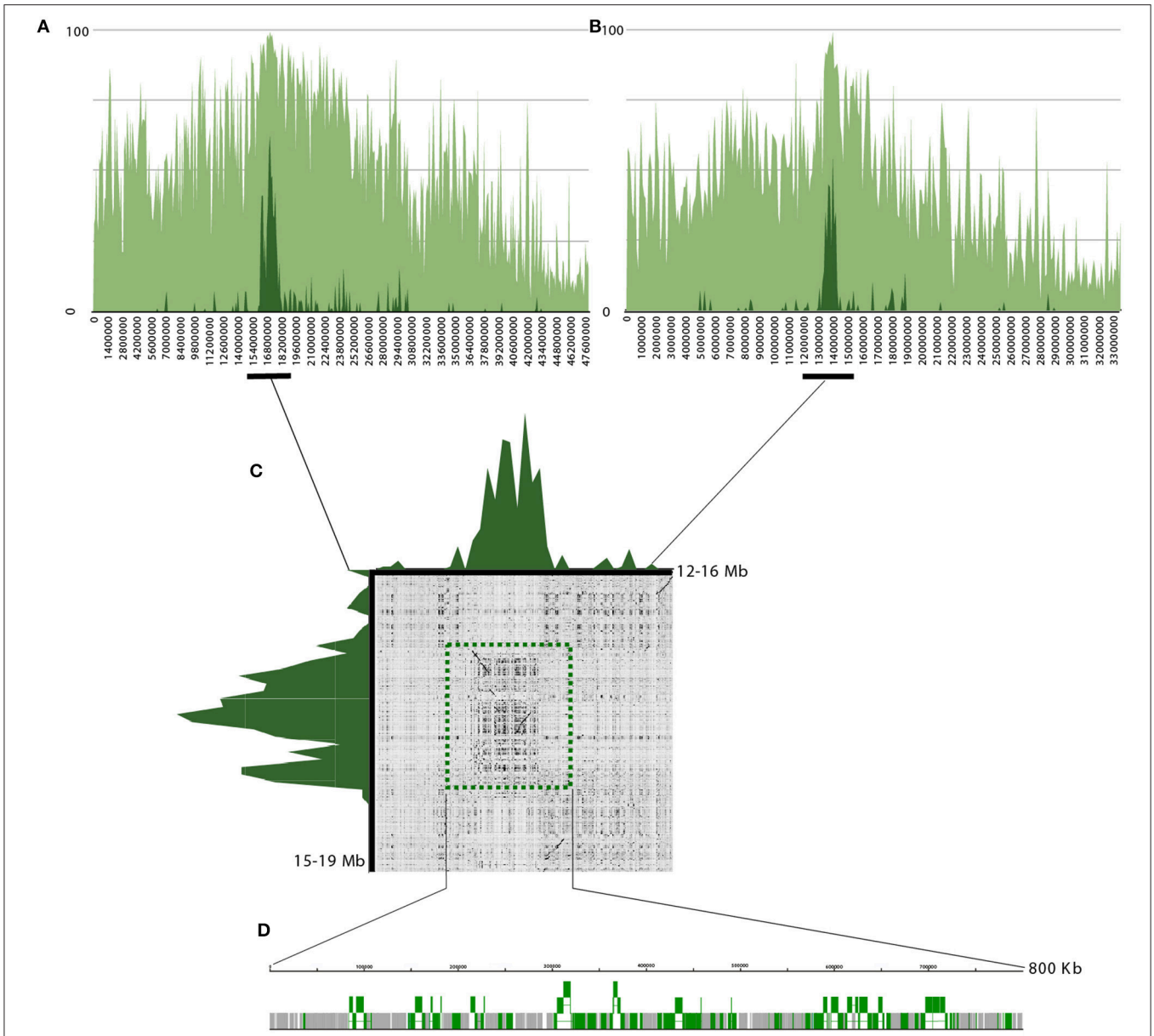


FIGURE 6 | Structure and annotation of pseudo-chromosomes 5 from *C. canephora* and *C. arabica*. **(A)** Density of transposable elements (light green) and CR elements (dark green) of pseudo-chromosomes 5 from *C. canephora*. X-axis represents the density of elements en percentage and Y-axis the coordinates of the pseudochromosomes. **(B)** Density of transposable elements (light green) and CR elements (dark green) of pseudo-chromosomes 5 from the *C. canephora* sub-genome in *C. arabica* **(C)**. Dot-plot graphical view of sequence comparison of 4 Mb *in C. arabica* (horizontal) and *C. canephora* (vertical). Dark green peaks represent the density of CR elements in these regions. **(D)** Sequence organization of the 800 kb centromeric region in *C. arabica*. Gray blocks represent transposable elements and green blocks are CRC elements.

2013). This observation may suggest that CR elements alone might be sufficient to ensure the kinetochore function. But more detailed annotations and validation of centromeric regions of Coffee trees are necessary to understand the composition and the evolution of such critical chromosomal regions.

The diversity in types and chromosomal insertions of CRCs gave a more complex view of the structure and evolution of centromeric regions in *Coffea*, especially in relation to LTR-RTs along hybridization process. *C. arabica* showed an accumulation of proximal heterochromatin associated with more dispersed CRC profile on the chromosomes, suggesting that the roles and effects of centromeric retrotransposons can extend beyond the proximal domains. In the near future, the characterization of centromere sequences in diploid and allotetraploid *Coffea* genomes will bring more insights into the evolution of these chromosomal regions that play a crucial role in the cell life cycle.

AUTHOR CONTRIBUTIONS

AV and RG: directed researches; RdCN and PY: performed FISH and bioinformatics and SO-A performed bioinformatics; PD, CF, and DM: performed sequencing and LM and SS performed genome assembly; AV, RG, AdK, and DC: wrote the manuscript.

ACKNOWLEDGMENTS

The authors thank the Brazilian agencies Fundação Araucária, CNPq and CAPES-Agropolis for financial support and the Agronomic Institute of Paraná (IAPAR), Londrina, Paraná, Brazil for Coffee seedlings. RG was supported by a Special Visiting Scientist grant from the Ciência sem Fronteiras program under the reference ID 84/2013 (CNPq/CAPES) and the ACGC for providing unpublished data. The authors also thank the Centro de Bioinformática y Biología Computacional (BIOS), fort he kind use of the cluster service.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2018.00175/full#supplementary-material

REFERENCES

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–0402. doi: 10.1093/nar/25.17.3389

Bao, W., Zhang, W., Yang, Q., Zhang, Y., Han, B., Gu, M., et al. (2006). Diversity of centromeric repeats in two closely related wild rice species, *Oryza officinalis* and *Oryza rhizomatis*. *Mol. Genet. Genomics*. 275, 421–430. doi: 10.1007/s00438-006-0103-2

Bennetzen, J. L. (2007). Patterns in grass genome evolution. *Curr. Opin. Plant Biol.* 10, 176–181. doi: 10.1016/j.pbi.2007.01.010

Bennetzen, J. L., and Wang, H. (2014). The contributions of transposable elements to the structure, function, and evolution of plant genomes. *Annu. Rev. Plant. Biol.* 65, 505–530. doi: 10.1146/annurev-arplant-050213-035811

Birney, E., Clamp, M., and Durbin, R. (2004). Genewise and genomewise. *Genome Res.* 14, 988–995. doi: 10.1101/gr.1865504

Carvalho, A. (1946). Distribuição geográfica e classificação botânica do gênero Coffea com referência especial à espécie Arabica. V. Origem e classificação botânica do *C. arabica* L. *Separata dos Boletins da Superintendência dos Serviços do Café*. 21, 174–180.

Chaparro, C., Gayraud, T., de Souza, R. F., Domingues, D. S., Akaffou, S., Vanzela, A. L. L., et al. (2015). Terminal-repeat retrotransposons with GAG domain in plant genomes: a new testimony on the complex world of transposable elements. *Genome Biol. Evol.* 7, 493–504. doi: 10.1093/gbe/evv001

Cheng, Z., Dong, F., Langdon, T., Ouyang, S., Buell, C. R., Gu, M., et al. (2002). Functional rice centromeres are marked by a satellite repeat and a centromere-specific retrotransposon. *Plant Cell.* 14, 1691–1704. doi: 10.1105/tpc.003079

Denoeud, F., Carretero-Paulet, L., Dereeper, A., Droc, G., Guyot, R., Pietrella, M., et al. (2014). The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. *Science* 345, 1180–1184. doi: 10.1126/science.1255274

Du, J., Tian, Z., Hans, C. S., Laten, H. M., Cannon, S. B., Jackson, S. A., et al. (2010). Evolutionary conservation, diversity and specificity of LTR-retrotransposons in flowering plants: insights from genome-wide analysis and multi-specific comparison. *Plant J.* 63, 584–598. doi: 10.1111/j.1365-313X.2010.04263.x

Dupeyron, M., de Souza, R. F., Hamon, P., Kochko, A., Crouzillat, D., Couturon, E., et al. (2017). Distribution of Divo in *Coffea* genomes, a poorly described family of angiosperm LTR-retrotransposons. *Mol. Genet. Genomics* 292, 741–754. doi: 10.1007/s00438-017-1308-2

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. doi: 10.1093/nar/gkh340

Gao, D., Chen, J., Chen, M., Meyers, B. C., and Jackson, S. (2012). A highly conserved, small LTR retrotransposon that preferentially targets genes in grass genomes. *PLoS ONE* 7:e32010. doi: 10.1371/journal.pone.0032010

Gao, X., Hou, Y., Ebina, H., Levin, H. L., and Voytas, D. F. (2008). Chromodomains direct integration of retrotransposons to heterochromatin. *Genome Res.* 18, 359–369. doi: 10.1101/gr.7146408

Grandbastien, M. A. (2015). LTR-retrotransposons, handy hitchhikers of plant regulation and stress response. *Biochim. Biophys. Acta* 849, 403–416. doi: 10.1016/j.bbagrm.2014.07.017

Guizard, S., Piégu, B., and Bigot, Y. (2016). DensityMap: a genome viewer for illustrating the densities of features. *BMC Bioinformatics* 7:204. doi: 10.1186/s12859-016-1055-0

Guyot, R., Darré, T., Dupeyron, M., de Kochko, A., Hamon, S., Couturon, E., 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. doi: 10.1007/s00438-016-1235-7

Hamon, P., Grover, C. E., Davis, A. P., Rakotomalala, J. J., Raharimalala, N. E., Albert, V. A., 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: GBS coffee phylogeny and the evolution of caffeine content. *Mol. Phylogenet. Evol.* 109, 351–361. doi: 10.1016/j.ympev.2017.02.009

Han, Y., Wang, G., Liu, Z., Liu, J., Yue, W., Song, R., et al. (2010). Divergence in centromere structure distinguishes related genomes in *Coix lacryma-jobi* and its wild relative. *Chromosoma* 119, 89–98. doi: 10.1007/s00412-009-0239-z

Heslop-Harrison, J. S., and Schwarzacher, T. (2011). Organisation of the plant genome in chromosomes. *Plant J.* 66, 18–33. doi: 10.1111/j.1365-313X.2011.04544.x

Houben, A., Schroeder-Reiter, E., Nagaki, K., Nasuda, S., Wanner, G., Murata, M., et al. (2007). CENH3 interacts with the centromeric retrotransposon cereba and GC-rich satellites and locates to centromeric substructures in barley. *Chromosoma* 116, 275–283. doi: 10.1007/s00412-007-0102-z

Kalendar, R., Vicient, C. M., Peleg, O., Anamthawat-Jonsson, K., Bolshoy, A., and Schulman, A. H. (2004). Large retrotransposon derivatives: abundant, conserved but nonautonomous retroelements of barley and related genomes. *Genetics* 166, 1437–1450. doi: 10.1534/genetics.166.3.1437

Lashermes, P., Combes, M. C., Robert, J., Trouslot, P., D'Hont, A., Anthony, F., et al. (1999). Molecular characterisation and origin of the *Coffea arabica* L. genome. *Mol. Gen. Genet.* 261, 259–266. doi: 10.1007/s004380050965

Li, B., Choulet, F., Heng, Y., Hao, W., Paux, E., Liu, Z., et al. (2013). Wheat centromeric retrotransposons: the new ones take a major role in centromeric structure. *Plant J.* 73, 952–965. doi: 10.1111/tpj.12086

Lim, K. B., Yang, T. J., Hwang, Y. J., Kim, J. S., Park, J. Y., Kwon, S. J., et al. (2007). Characterization of the centromere and peri-centromere retrotransposons in *Brassica rapa* and their distribution in related Brassica species. *Plant J.* 49, 173–183. doi: 10.1111/j.1365-313X.2006.02952.x

Liu, Z., Yue, W., Li, D., Wang, R. R. C., Kong, X., Lu, K., et al. (2008). Structure and dynamics of retrotransposons at wheat centromeres and pericentromeres. *Chromosoma* 117, 445–456. doi: 10.1007/s00412-008-0161-9

Llorens, C., Futami, R., Covelli, L., Domínguez-Escribá, L., Viu, J. M., Tamarit, D., et al. (2011). The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. *Nucleic Acids Res.* 39, D70–D74. doi: 10.1093/nar/gkq1061

Llorens, C., Mu-oz-Pomer, A., Bernad, L., Botella, H., and Moya, A. (2009). Network dynamics of eukaryotic LTR retroelements beyond phylogenetic trees. *Biol. Direct.* 4:41. doi: 10.1186/1745-6150-4-41

Lopes, F. R., Jjingo, D., Da Silva, C. R., Andrade, A. C., Marraccini, P., Teixeira, J. B., et al. (2013). Transcriptional activity, chromosomal distribution and

expression effects of transposable elements in *Coffea* genomes. *PLoS ONE* 8:e78931. doi: 10.1371/journal.pone.0078931

Ma, J., and Bennetzen, J. L. (2004). Rapid recent growth and divergence of rice nuclear genomes. *Proc. Natl. Acad. Sci. U.S.A.* 101, 12404–12410. doi: 10.1073/pnas.0403715101

Marques, A., Ribeiro, T., Neumann, P., Macas, J., Novak, P., Schubert, V., et al. (2015). Holocentromeres in *Rhynchospora* are associated with genome-wide centromere-specific repeat arrays interspersed amongst euchromatin. *Proc. Natl. Acad. Sci. U.S.A.* 112, 13633. doi: 10.1073/pnas.1512255112

McCarthy, E. M., and McDonald, J. F. (2003). LTR_STRUC: a novel search and identification program for LTR-retrotransposons. *Bioinformatics* 19, 362–367. doi: 10.1093/bioinformatics/btf878

Mueller, L., Strickler, S. R., Domingues, D. S., Pereira, L. F. P., Andrade, A. A., Marraccini, P., et al. (2015). "Towards a better understanding of the coffea arabica genome structure," in *Proceedings of the 25th International Conference on Coffee Science ASIC*. (Armenia, CO), 42–45.

Nagaki, K., Neumann, P., Zhang, D., Ouyang, S., Buell, C. R., Cheng, Z., et al. (2005). Structure, divergence, and distribution of the CRR centromeric retrotransposon family in rice. *Mol. Biol. Evol.* 22, 845–855. doi: 10.1093/molbev/msi069

Neumann, P., Navrátilová, A., KoblíŽková, A., Kejnovský, E., Hribová, E., Hobza, R., et al. (2011). Plant centromeric retrotransposons: a structural and cytogenetic perspective. *Mob. DNA* 2:4. doi: 10.1186/1759-8753-2-4

Novikova, O. (2009). Chromodomains and LTR-retrotransposons in plants. *Comm. Integr. Biol.* 2, 158–162. doi: 10.4161/cib.7702

Parisod, C., Alix, K., Just, J., Petit, M., Sarilar, V., Mhiri, C., et al. (2010). Impact of transposable elements on the organization and function of allopolyploid genomes. *New Phytol.* 186, 37–45. doi: 10.1111/j.1469-8137.2009.03096.x

Piegu, B., Guyot, R., Picault, N., Roulin, A., Saniyal, A., Kim, H., et al. (2006). Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. *Genome Res.* 16, 1262–1269. doi: 10.1101/gr.5290206

Romano, E., and Brasileiro, A. C. M. (1999). Extração de DNA de plantas: soluções para problemas comumente encontrados. *Biotecnol. Ciência e Desenvolvimento*. 9, 40–43.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M. A., et al. (2000). Artemis: sequence visualization and annotation. *Bioinformatics* 16, 944–945. doi: 10.1093/bioinformatics/16.10.944

Sanei, M., Pickering, R., Kumke, K., Nasuda, S., and Houben, A. (2011). Loss of centromeric histone H3 (CENH3) from centromeres precedes uniparental chromosome elimination in interspecific barley hybrids. *Proc. Natl. Acad. Sci. U.S.A.* 108, E498–E505. doi: 10.1073/pnas.1103190108

SanMiguel, P., Gaut, B. S., Tikhonov, A., Nakajima, Y., and Bennetzen, J. L. (1998). The paleontology of intergene retrotransposons of maize. *Nat. Genet.* 20, 43–45. doi: 10.1038/1695

Sanseverino, W., Hénaff, E., Vives, C., Pinosio, S., Burgos-Paz, W., Morgante, M., et al. (2015). Transposon insertions, structural variations, and SNPs contribute to the evolution of the melon genome. *Mol. Biol. Evol.* 32, 2760–2774. doi: 10.1093/molbev/msv152

Santos, F. C., Guyot, R., Do Valle, C. B., Chiari, L., Techio, V. H., Heslop-Harrison, P., et al. (2015). Chromosomal distribution and evolution of abundant retrotransposons in plants: Gypsy elements in diploid and polyploid Brachiaria forage grasses. *Chromosome Res.* 23, 571–582. doi: 10.1007/s10577-015-9492-6

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternk, S., et al. (2009). The B73 maize genome: complexity, diversity, and dynamics. *Science* 326, 1112–1116. doi: 10.1126/science.1178534

Schwarzacher, T., Ambros, P., and Schweizer, D. (1980). Application of Giemsa banding to orchid karyotype analysis. *Plant Syst. Evol.* 134, 293–297. doi: 10.1007/BF00986805

Schweizer, D., and Loidl, J. (1987). A model for heterochromatin dispersion and the evolution of C band patterns. *Chrom. Today* 9, 61–74. doi: 10.1007/978-94-010-9166-4_7

Sharma, A., and Presting, G. G. (2014). Evolution of centromeric retrotransposons in grasses. *Genome Biol. Evol.* 6, 1335–1352. doi: 10.1093/gbe/evu096

Sonnhammer, E. L., and Durbin, R. (1995). A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* 167:GC1-10. doi: 10.1016/0378-1119(95)00714-8

Tenaillon, M. I., Hufford, M. B., Gaut, B. S., and Ross-Ibarra, J. (2011). Genome size and transposable element content as determined by high-throughput sequencing in maize and *Zea luxurians*. *Genome Biol. Evol.* 3, 219–229. doi: 10.1093/gbe/evr008

The Arabica Coffee Genome Consortium (ACGC) (2014). "Towards a better understanding of the Coffea Arabica Genome Structure," in *Association for Science and Information on Coffee* (International Conference on Coffee Science Cogito), 42–45.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22, 4673–4680. doi: 10.1093/nar/22.22.4673

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., et al. (2007). A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* 8, 973–982. doi: 10.1038/nrg2165

Witte, C. P., Le, Q. H., Bureau, T., and Kumar, A. (2001). Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes. *Proc. Natl. Acad. Sci. U.S.A.* 98, 13778–13783. doi: 10.1073/pnas.241341898

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

Zhang, W., Zuo, S., Li, Z., Meng, Z., Han, J., Song, J., et al. (2017). Isolation and characterization of centromeric repetitive DNA sequences in *Saccharum spontaneum*. *Sci. Rep.* 7:41659. doi: 10.1038/srep41659

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.