HAL
open science

Dynamique et évolution de deux lignées remarquables de rétrotransposons à LTR dans le genre Coffea (famille des Rubiacées)

Mathilde Dupeyron

To cite this version:

Mathilde Dupeyron. Dynamique et évolution de deux lignées remarquables de rétrotransposons à LTR dans le genre Coffea (famille des Rubiacées). Biologie végétale. Université Montpellier, 2017. Français. NNT : 2017MONTT128 . tel-01714153

HAL Id: tel-01714153

https://theses.hal.science/tel-01714153

Submitted on 21 Feb 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THESE POUR OBTENIR LE GRADE DE DOCTEUR DE L'UNIVERSITE DE MONTPELLIER

En Ecophysiologie et Adaptation des Plantes

École doctorale GAIA

Unités de recherche DiADE et IPME

Dynamique et évolution de deux lignées remarquables de rétrotransposons à LTR dans le genre *Coffea* (famille des Rubiacées)

Présentée par Mathilde DUPEYRON
Le 23 novembre 2017

Sous la direction de Perla HAMON
et Romain GUYOT

Devant le jury composé de

Michel LEBRUN, Pr, Université de Montpellier	Président du jury
Marie-Angèle GRANDBASTIEN, DR, INRA Versailles-Grignon	Rapportrice
Frédérique PELSY, DR, INRA Colmar	Rapportrice
Abdelkader AÏNOUCHE, MC, Université de Rennes 1	Examinateur

UNIVERSITÉ DE MONTPELLIER

Remerciements

Je souhaite tout d'abord remercier mes deux directeurs de thèse, Perla Hamon et Romain Guyot, pour m'avoir acceptée en tant que doctorante dans leurs équipes respectives à l'IRD. Merci pour vos conseils tout au long de la thèse et votre soutien. Merci Perla pour tes enseignements riches sur les caféiers, tes partages sur tes expériences de recherche et d'enseignement et bien sûr tes paroles motivantes quand j'en avais besoin. Merci Romain pour ton accueil au Brésil au début de la thèse, tes réponses toujours rapides quand j'avais une question ou un problème bio-informatique et pour le très bon café de Colombie.

Merci à Alain Ghesquière et à Valérie Verdier pour leur accueil au sein des UMR DiADE et IPME. Merci à tous les membres des équipes EvoGeC et CoffeeAdapt. Un vif merci à mon ancien « co-bureau » Emmanuel Couturon qui me racontait ses missions passionnantes en Afrique, à bientôt j'espère ! Merci à Serge Hamon pour m'avoir prêté son ordinateur et ainsi sauvé le présent manuscrit. Merci à Hervé Étienne pour son accueil à CoffeeAdapt et les réunions vivantes avec toute l'équipe.

Je remercie mes rapportrices Marie-Angèle Grandbastien et Frédérique Pelsy, ainsi que mes examinateurs Abdelkader Aïnouche et Michel Lebrun, pour avoir accepté d'évaluer ma thèse.

Un grand merci aux membres de mes comités de suivi de thèse, Dominique This, Marie Mirouze et Cristian Chaparro, pour avoir accepté de faire partie de ce comité. Merci pour tous vos conseils et encouragements pour la thèse.

Merci à Dominique Crouzillat et Alexandre de Kochko, plus largement les membres du consortium ACGC, pour m'avoir fourni les données de séquençage utilisées durant ma thèse.

Merci aux chercheurs qui m'ont chaleureusement accueillie au Brésil au tout début de ma thèse. Un merci particulier à Doug et André pour m'avoir fait découvrir un peu de culture brésilienne. Merci aux doctorantes Renata, Joana et Aline pour m'avoir permis de fêter dignement mon départ de Londrina.

Je remercie évidemment tous les doctorants (et post-doctorants) de l'IRD sans qui les journées de travail et certaines soirées auraient été bien moroses.
Merci à Chloé qui m'a introduite dans le cercle privé des belettes louches. Merci pour ta joie et les week-ends à la montagne (un merci à Lionel au passage pour les histoires de fraises sous la neige me donnant du courage lors d'une certaine randonnée).
Merci à Cécile pour ton accueil parmi les doctorants, les fous rires divers et variés et toutes nos discussions plus ou moins « sérieuses » ! Merci pour le séjour en Allemagne et pour ton coaching sport qui me manque beaucoup.

Merci à Hélène qui partage mon admiration pour les félins d'appartement et pour nous avoir accueillis dans les magnifiques Hautes-Alpes, à Nice et à Quedlinburg. Merci pour tes phrases rigolotes de Banane d'Or et tes critiques « code de la route » !
Merci à Émilie pour les petites attentions qui remontaient le moral, les incompréhensions qui déclenchaient un fou rire contagieux, Globox la tomate et les soirées pizza et barbecue.
Merci à Fabien pour les pauses papotage anti-saturation, ta gentillesse et bien sûr les instants courses qui étaient donc moins ennuyeuses !
Merci à Céline pour les soirées confortables à discuter de plein de choses, les conseils soutenance et post-doc, ainsi que pour m'avoir fait voyager un peu en Belgique.
Merci à mon homonyme première du nom pour la balade à Cassis, tes sourires chaleureux et les récits de tes aventures au Togo et à Ithaca.
Un merci tout particulier aux deux personnes réalisant leur thèse en même temps que moi. Merci Rémi pour les discussions philosophiques, les randonnées (surtout celle qui fut laborieuse), le week-end à Lyon et bien sûr ton soutien pour cette dernière année de thèse. Merci Lucile *alias* Toc-Toc pour tes illogismes rigolos, tes questions parfois gênantes mais mignonnes, les discussions vernis et compagnie et ton rire reconnaissable entre mille.
Merci à Mathieu pour les soirées culinaires en ta demeure, les histoires de tes voyages qui font rêver et ta bonne humeur.
Merci à Jérémy pour le week-end niçois, les débats politico-agronomiques et les soirées qui nous vidaient la tête.
Merci à Marilyne pour tes encouragements, les rires post-moments foufous et ta passion inégalée pour le café (la boisson).
Merci à Maíra pour le soutien rédaction, les conseils soutenance et les soirées pintes.
Merci aux « première année » (presque en deuxième année !), Eoghan pour ta folie attitude et Edith pour ta sympathie.
Merci aux doctorants, stagiaires ou salariés en séjour plus ou moins long à l'IRD : Ialy pour tes bons samoussas, Jackie pour tes compliments, Renata pour nos moments maquillage, Bear pour ton baume anti-douleur, Sinara pour ta gentillesse, Aurore D. pour ta bienveillance, Alix pour les pauses thé ou café, Sunao pour m'avoir rappelé que j'aimerais visiter le Japon, Cyril pour tes « youhou ! » joyeux. Un merci spécial à Aurore R. pour tes encouragements, les sorties bienfaitrices et nos discussions à rallonge.

Je remercie les IRDien(ne)s qui ont rendu mon travail et mon séjour doctoral plus agréable : Carole Bessière, Bruno Barthélémy, Ndomassi Tando, Myriam Collin, Sylvie Doulbeau, Isabelle Hérault, Alexis Dereeper, Nathalie Pujet et celles/ceux que j'oublie sûrement.

Aux autres personnes que j'ai rencontrées à Montpellier : merci à Mélissa, la coach à la place de la coach de Tae-kwon-do, pour les soirées Liar Game, entres autres. Merci à Hélène (deuxième du nom) pour m'avoir fait découvrir J.-P. Jaworski et parlé la langue du thé. Merci à ma petite Berfin pour ta gentillesse, nos sorties ciné et le soutien que j'espère t'apporter également.

Un grand merci à ma marraine du programme de mentorat « Femmes&Sciences » Rosemary Kiernan. Merci pour nos échanges sur la thèse, les doutes et les craintes qu'elle peut provoquer, les post-doctorats et d'autres sujets autour d'un bon déjeuner.

Je n'aurais pas pu aller jusque là sans ma famille. Mille mercis à mes parents, d'un soutien sans faille, qui m'ont toujours poussée à travailler pour pouvoir faire « ce que j'aime » dans la vie. Merci d'avoir toujours été et d'être encore là pour moi. Merci à mon frérot adoré, on est là l'un pour l'autre, même à presque 1000 km de distance. Merci à ma grand-mère pour nos longues discussions téléphoniques ou face à face et son soutien. Merci à mes cousins et leurs escapades montpelliéraines qui m'ont permis de les voir un peu plus que « prévu ». Merci à mes grands-parents maternels partis trop tôt.

Je souhaitais remercier les membres du laboratoire d'Écologie et Biologie des Interactions de Poitiers, sans qui je ne serais pas là non plus. Merci à Clément Gilbert pour m'avoir emportée dans l'addiction aux éléments transposables et permis de faire un excellent stage de Master 2. Merci à Richard Cordaux pour avoir pris le temps de discuter lors de mon stage et plus récemment à l'ICTE. Merci à Isabelle Giraud pour tout ton soutien et ton amitié. Même s'il ne fait plus partie de ce laboratoire, merci à Mathieu Sicard qui m'a « suivie » depuis la L1 jusqu'à l'Université de Montpellier.

Comment ne pas remercier mes plus proches amis qui sont d'un soutien indéfectible ?
Merci aux « Charentaises à l'aise » Katia, Claire et Élise de me supporter depuis le lycée.
Merci évidemment à Camille et Mariette (je n'oublie pas Sylvain et Tony !) pour nos aventures poitevines (entre autres) et toute votre affection dans les moments difficiles comme dans les petites joies du quotidien.
Merci également à Marie S., 15 ans que l'on se connaît !
Merci à Élo (je vais finir par venir te voir à Montréal !), Maëva, Marie P. et Anne pour les moments trop rares que l'on a partagé ces trois dernières années et pour tous les bons souvenirs du lycée.
Merci Thomas pour m'avoir aidée lors de mon emménagement à Montpellier, pour tes soirées avec beaucoup d'ambiances (merci Élodie et petit Lucas !) et pour ton soutien moral.
Merci Caro pour ta joie de vivre, les visites sudistes et nos discussions interminables par sms quand ça ne va pas…mais quand ça va aussi !
Merci Lulu pour nos causeries et soirées geekettes, ainsi que pour m'avoir accueillie régulièrement dans les Yvelines et être venue à Montpellier.
Merci Séverine et Anaïs pour nos riches discussions et balades à Poitiers, Angoulême et Montpellier.

Même s'il ne pourra pas lire ceci, merci à mon chat Gaston (aussi bavard que moi) dont les ronronnements du soir m'apaisent.

Enfin, merci à la ou les personnes que j'ai pu oublier, avec toutes mes excuses.

Sommaire

CHAPITRE 1 – INTRODUCTION GÉNÉRALE 1

1. L'ORIGINE DE L'ETUDE DES ELEMENTS TRANSPOSABLES 1
DECOUVERTE 1
L'IMPORTANCE DU NOMME A TORT « ADN POUBELLE » 2
2. DEFINITION, MOBILITE ET CLASSIFICATION 4
DEFINITION 4
MECANISMES DE TRANSPOSITION 4
TRANSFERTS HORIZONTAUX 6
3. CLASSIFICATION DES ELEMENTS TRANSPOSABLES 7
4. ORIGINE DES ELEMENTS TRANSPOSABLES 10
5. OUTILS DE DETECTION DES ET 11
LA METHODE *DE NOVO* 11
LA METHODE D'HOMOLOGIE DE SEQUENCE 12
LA METHODE D'IDENTIFICATION DE LA STRUCTURE 13
LA METHODE DE GENOMIQUE COMPARATIVE 13
6. IMPACTS DES ET SUR LES GENOMES 14

CHAPITRE 2 – ÉTAT DE L'ART 16

1. COMPOSITION DES GENOMES DES PLANTES EN ET 16
TAILLE DES GENOMES VEGETAUX ET LTR-RT 16
CYCLE DE VIE D'UN LTR-RT 18
LOCALISATION DES LTR-RT 20
2. IMPACTS DES LTR-RT SUR LES GENOMES DES PLANTES 21
VARIATION DE LA TAILLE DES GENOMES 21
REARRANGEMENTS CHROMOSOMIQUES 23
REGULATION DES GENES 25
MECANISMES EPIGENETIQUES ET STRESS 27
3. UTILISATION DES LTR-RT COMME OUTILS MOLECULAIRES 28
4. LES LTR-RT CHEZ LES CAFEIERS 29
LE GENRE *COFFEA* 29
GENOMIQUE ET LTR-RT CHEZ LES CAFEIERS 36
5. PROJETS INTERNATIONAUX G13 ET ACGC 38
6. PRESENTATION DU SUJET DE RECHERCHE 39

CHAPITRE 3 – APPORTS DU SEQUENÇAGE PARTIEL A L'ETUDE DES ELEMENTS TRANSPOSABLES DANS LE GENRE *COFFEA* 42

1. CONTEXTE 42
2. IMPLICATION PERSONNELLE 44
3. CONCLUSIONS ET PERSPECTIVES 45

CHAPITRE 4 - CARACTERISATION D'UNE FAMILLE DE LTR-RETROTRANSPOSONS PEU CONNUE, *DIVO*, CHEZ LES CAFEIERS 46

1. CONTEXTE 46
2. IMPLICATION PERSONNELLE 46
3. CONCLUSIONS ET PERSPECTIVES 48

CHAPITRE 5 - L'ANALYSE DES LTR-RETROTRANSPOSONS *SIRE* INDIQUE UNE PROBABLE ORIGINE PARENTALE OUGANDAISE DE *COFFEA ARABICA* 53

1. CONTEXTE **53**
2. IMPLICATION PERSONNELLE **53**
3. CONCLUSIONS ET PERSPECTIVES **90**

CHAPITRE 6 – ÉVOLUTION DES LTR-RT *SIRE* DANS LE GENRE *COFFEA* **92**

1. CONTEXTE **92**
2. IMPLICATION PERSONNELLE **92**
3. CONCLUSIONS ET PERSPECTIVES **114**

CHAPITRE 7 – DISCUSSION ET PERSPECTIVES **115**

1. CONCLUSION GENERALE **115**
2. DISCUSSION GENERALE ET PERSPECTIVES **117**
ANNOTATION ET CARACTERISATION DES LTR-RT 117
LES *SIRE* ET LEUR DOMAINE *ENVELOPPE* 119
DYNAMIQUE DE *DIVO* ET DES *SIRE* CHEZ LES CAFEIERS 120
C. ARABICA ET POLYPLOÏDIE 122
ÉVOLUTION DU GENRE *COFFEA* 125
3. ÉVOLUTION DES *SIRE* ET DE *DIVO* DANS LE GENRE *COFFEA* **126**

BIBLIOGRAPHIE **128**

ANNEXE 1 – FIGURE ANNEXE 1 ..... **139**

ANNEXE 2 – FIGURE ANNEXE 2 ..... **144**

ANNEXE 3 – VALORISATION SCIENTIFIQUE **150**

ANNEXE 4 – EXPERIENCE D'ENSEIGNEMENT **155**

ANNEXE 5 – FORMATIONS REALISEES **156**

Liste des Figures et Tableau

Figure 1 : Schéma de l'explication de la ségrégation des gènes *A_1* et *Dt* chez le maïs comme observé par M. Rhoades. 11

Figure 2 : Trois différents types de transposition. 13

Figure 3 : Système de classification des éléments transposables proposé par Wicker et al. en 2007. 17

Figure 4 : Schéma des LTR-Rétrotransposons présents dans les génomes des plantes. 25

Figure 5 : Cycle de rétrotransposition d'un LTR-rétrotransposon. 28

Figure 6 : Mécanismes de formation d'un solo-LTR par recombinaisons homologues inégales. 33

Figure 7 : Schémas représentant les endroits où un ET peut s'insérer (flèches noires) et ce qui peut en résulter. 35

Figure 8 : Photographies de 4 espèces de caféiers montrant la différence morphologique des fleurs des *Psilanthus* et des *Coffea*. 39

Figure 9 : Répartition géographique naturelle des espèces de caféiers. 41

Figure 10 : Phylogénie simplifiée des *Coffea* et illustrations des espèces étudiées durant la thèse. 44

Figure 11 : Distribution du contenu en ADN (2C) des génomes des *Coffea* à Madagascar (a) et en Afrique (b). 52

Figure 12 : Arbre en NJ basé sur le domaine RT des Copia de *C. canephora*. 56

Figure 13 : Estimation du nombre de copies de *Divo* dans 24 espèces diploïdes de caféiers. 58

Figure 14 : Analyse phylogénétique (NJ – 100 répétitions de bootstraps) basée sur 1022 domaines RT de *Divo* dans 15 espèces/sous-espèces diploïdes de *Coffea*. 60

Figure 15 : Estimation du nombre de copies des *SIRE* dans un jeu de données 454 de *C. canephora*, *C. arabica* et *C. eugenioides*. 64

Figure 16 : Estimation du nombre de copies de *Del* dans six séquences génomiques de *Coffea*. 129

Figure 17 : Schéma de la présence et de l'absence de *Divo* et des *SIRE* dans les espèces étudiées du genre *Coffea*. 133

<u>Tableau 1</u> : Espèces disponibles, type de séquençage et études dans lesquelles elles ont été utilisées. 50

Liste des abréviations

ACE : Afrique du centre-est

ACGC : Arabica Coffee Genome Consortium

ADN : Acide désoxyribonucléique

AE : Afrique de l'est

AOC : Afrique de l'ouest et du centre

AP : Aspartic protease

ARN : Acide ribonucléique

ARNm : Acide ribonucléique messager

BA : Basse altitude

BAC : Bacterial artificial chromosome

BLAST : Basic local alignment search tool

CBD : Coffee berry disease

CRM : Centromeric retrotransposons of maize

DIRS : Dictyostelium intermediate repeat sequence

EC : Eu-Coffea

ENV : Enveloppe

ET : Élément transposable

FISH : Fluorescent *in situ* hybridisation

G13 : Génomes13

HA : Haute altitude

INT : Intégrase

IOI : îles de l'Océan Indien

LINE : Long interspersed nuclear element

LTR : Long terminal repeats

LTR-RT : Rétrotransposon à LTR

MiSA : Microsatellite identification tool

MITE : Miniature-inverted transposable élément

NGS : New-generation sequencing

NJ : Neighbor-Joining

ORF : Open Reading frame

PAG : Plant and Animal genomes

PBS : Primer binding site

PLE : Penelope-like elements

PPT : Polypurine tract

RdDM : RNA-directed DNA methylation

REMAP : Retrotransposon-microsatellite amplified polymorphism

RFLP : Restriction fragment length polymorphism

RNAseq : RNA sequencing

RH : RNase H

RNase H : Ribo-nucléase H

RPKM : Reads per kilobase per million mapped reads

RT : Reverse transcriptase

SINE : Short interspersed nuclear élément

siRNA : silencing RNA

SSR : Simple sequence repeats

TH : Transfert horizontal

TIR : Terminal inverted repeats

TR-GAG : Terminal-repeat retrotransposons with GAG domain

TSD : Target site duplication

XC : Xeno-Coffea

Chapitre 1 – INTRODUCTION GÉNÉRALE

Depuis les découvertes majeures des XIX et XXème siècles concernant la biologie des organismes, les scientifiques vont toujours plus loin dans leurs questionnements pour comprendre leur complexité. Se basant d'abord sur des observations macroscopiques, l'invention du microscope optique puis des microscopes électroniques leur permit d'étudier le vivant à l'échelle de la cellule. Par la suite et relativement récemment, l'avènement de la génétique autorisa l'étude de cette fabuleuse molécule qu'est l'acide désoxyribonucléique (ADN). Enfin, aujourd'hui, toutes ces connaissances peuvent être organisées, détaillées et complétées par l'examen minutieux des génomes, que l'on peut définir comme l'ensemble du matériel génétique d'une cellule. Le séquençage de ceux-ci ne cesse de s'améliorer et la quantité des données génomiques augmente considérablement, permettant les recherches sur des constituants aussi complexes et mystérieux que les éléments transposables, ces portions d'ADN capables de se déplacer au sein du génome hôte.

1. L'origine de l'étude des éléments transposables

Découverte

Au début du XXème siècle, l'observation de phénotypes variés (notamment au niveau de la couleur) de certaines plantes cultivées alimentaires ou ornementales amena des scientifiques comme Hugo De Vries et R. A. Emerson à rechercher l'origine de telles variations (de Vries and MacDougal 1905; Emerson 1914). Des hypothèses furent soulevées sur l'idée qu'un facteur s'activerait, puis serait inhibé pour s'activer à nouveau chez certains individus, amenant ainsi à différentes « variétés » pour une même espèce. Avec l'ère de la génétique - l'étude des lois de l'hérédité – la notion de « gènes instables » donnant ces phénotypes variés selon des relations de dominance et récessivité ainsi que selon l'environnement, fut mise au jour (Demerec 1935). Marcus M. Rhoades travailla sur la ségrégation des gènes responsables du phénotype « *dotted* » (petites taches colorées) de l'aleurone du maïs (Rhoades 1936). Le gène A_1, responsable de la coloration

de l'aleurone, provoquerait l'absence de coloration sous sa forme récessive a_1 et le phénotype « dotted » se présenterait lors de l'interaction de a_1 avec le gène *Dt* sous sa forme dominante (Rhoades 1938) (Figure 1).

C'est grâce aux travaux révolutionnaires de Barbara McClintock que l'on compris que d'autres composés génétiques que les gènes, activés par certains mécanismes comme la rupture de l'extrémité terminale des chromosomes chez le maïs, pouvaient provoquer les variations phénotypiques observées (McClintock 1950). Très mal reçus dans les années 1950, les travaux de B. McClintock sur la transposition des éléments *Ds* et *Ac* du maïs, furent ignorés par une communauté scientifique attachée à l'idée de « fixité » du génome. Pire, l'idée que des « éléments de contrôle » puissent modifier l'expression des gènes n'était pas envisageable, surtout durant le développement embryonnaire très organisé d'un organisme. Le terme d'élément transposable (ET) était plus facilement accepté, négligeant la notion de contrôle. Ce n'est que dans les années 1960, après la découverte des séquences d'insertion chez les bactéries notamment, que ses travaux commencèrent à être reconnus (Comfort 1999). Les éléments génétiques mobiles furent ensuite découverts et étudiés dans plusieurs organismes, des procaryotes aux eucaryotes incluant les plantes, les champignons et les animaux.

L'importance du nommé à tort « ADN poubelle »

Malgré leur prévalence dans les génomes étudiés, les ET ont longtemps été considérés comme de l'ADN inutile, poubelle ou encore égoïste (Ohno 1972; Doolittle and Sapienza 1980; Orgel and Crick 1980). Il était difficile de concevoir à l'époque que la majorité de l'ADN, humain notamment, ne soit pas ce qu'on appelle des gènes, indispensables au bon fonctionnement de l'organisme. Même encore récemment, les ET furent considérés comme des parasites (Zeh et al. 2009), malgré de nombreux indices de coévolution avec le génome ou la démonstration de gènes humains dérivés d'éléments transposables. Il était important pour moi de préciser ceci, avant de rentrer « dans le vif du sujet » avec ce qui va suivre. Les années 1980 à 2000 ont été riches en découvertes sur ces éléments, allant de concert avec l'amélioration des méthodologies et des

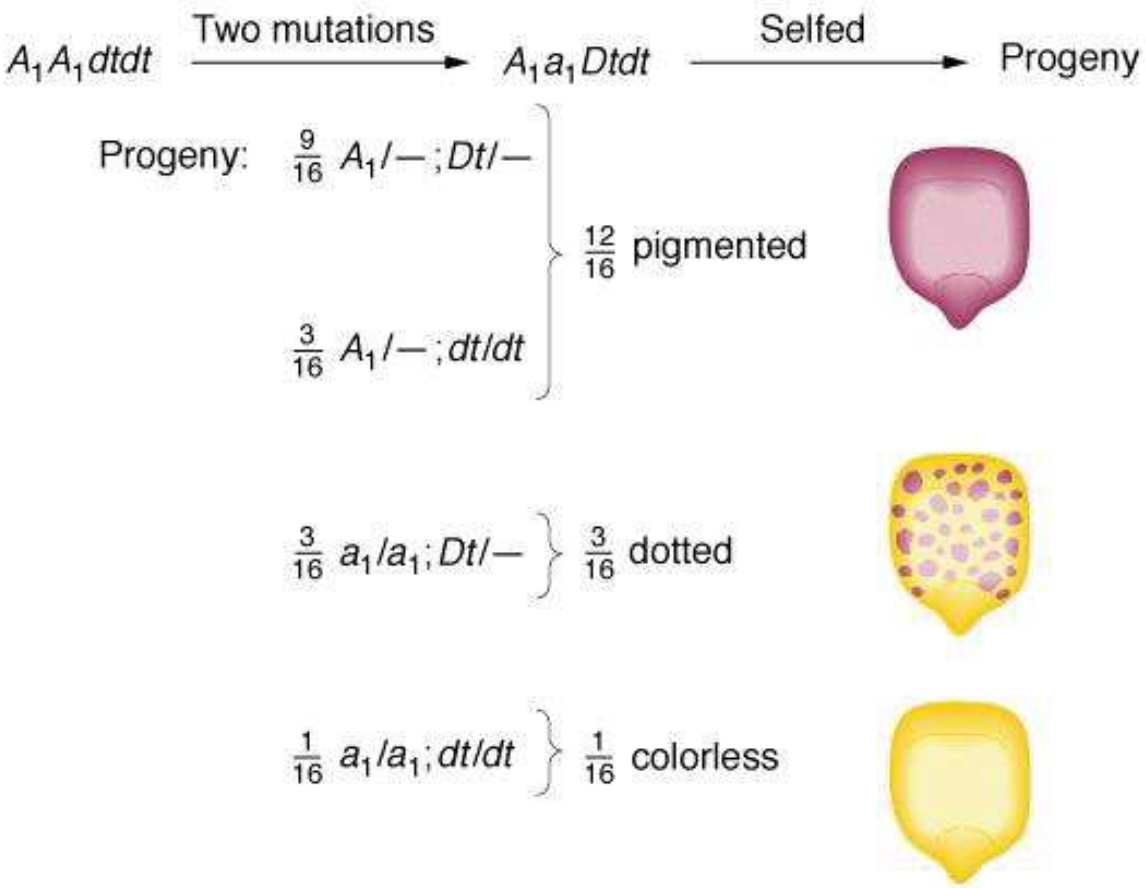


Figure 1 : Schéma de l'explication de la ségrégation des gènes *A_1* et *Dt* chez le maïs comme observé par M. Rhoades. D'après www.bio.miami.edu/dana/250/250SS16_17print.html.

technologies permettant l'étude toujours plus précise des génomes, augmentant également la complexité de l'examen des données obtenues.

2. Définition, mobilité et classification

Définition

On peut définir un ET comme une séquence d'ADN contenant une machinerie enzymatique lui permettant de se déplacer et se dupliquer dans un génome. Il existe plusieurs types d'ET, contenant des enzymes différentes selon le mécanisme de transposition. Les ET ont été détectés en quantité variable dans tous les organismes étudiés, procaryotes comme eucaryotes. Leur mobilité peut avoir de grands impacts sur la structure, la dynamique et l'évolution des génomes, faisant de leur étude une composante indispensable à la compréhension des mécanismes associés ou responsables de l'évolution des espèces hôtes.

Mécanismes de transposition

Durant les années 1990 et 2000, des études de plus en plus nombreuses ont levé le voile non seulement sur la diversité des ET, mais aussi sur le fonctionnement fascinant de la transposition. Hormis le mécanisme de transposition des Hélitrons, plus complexe et décrit plus récemment (Kapitonov and Jurka 2001), il existe deux grands mécanismes de transposition dont les étapes principales sont indiquées dans la Figure 2 (issue de Lisch 2013a).

La rétrotransposition est le premier mécanisme décrit, peut-être parce qu'il est proche du fonctionnement des rétrovirus, et qu'il utilise un système « ADN-ARN-ADN » (Boeke et al. 1985). De manière simplifiée, la séquence du rétrotransposon est transcrite, puis la transcriptase réverse, ou RT, réalise la transcription réverse. Ensuite, l'intégrase (ou INT) insère la nouvelle copie formée à un nouveau site. Ce mécanisme diffère légèrement entre les rétrotransposons à LTR (Long Terminal Repeats – longues répétitions terminales) et ceux sans LTR, notamment au niveau de l'intégration de la nouvelle copie. Les rétrotransposons non-autonomes, c'est-à-dire qui ne contiennent

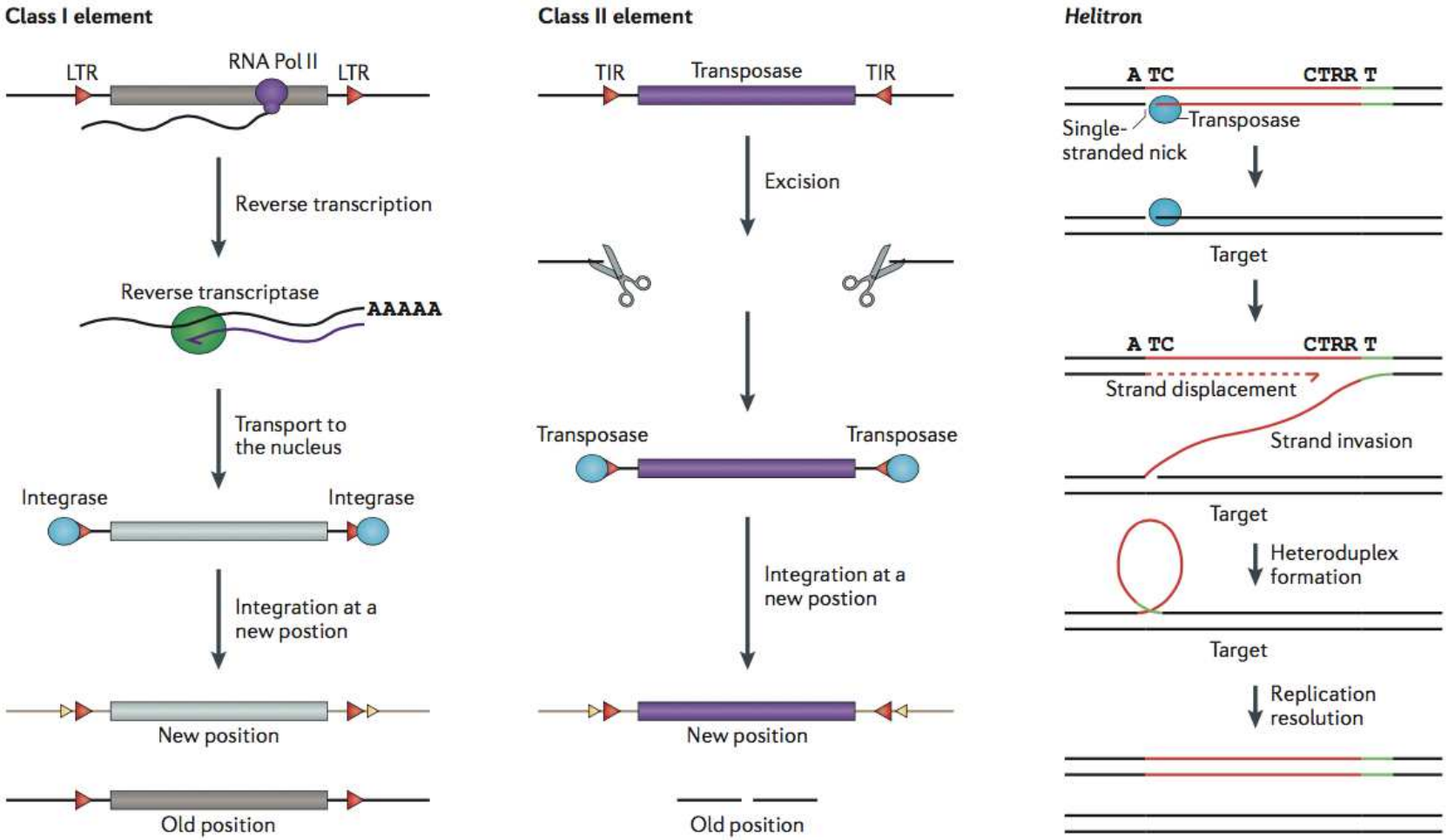


Figure 2 : Trois différents types de transposition. D'après Lisch 2013.
LTR : Long Terminal Repeat, TIR : Terminal Inverted Repeats, R : purines A ou G.

plus les enzymes nécessaires à leur mobilité, requièrent celles contenues dans les rétrotransposons autonomes. On nomme aussi ce mécanisme « copié-collé » car une nouvelle copie est créée à partir d'une copie pré-existante.

La transposition des ET de classe 2, ou « à ADN », consiste en l'excision de la séquence puis son intégration à un nouveau site. Ceci est possible grâce à l'action de la transposase, une enzyme codée par les éléments autonomes. Les éléments non-autonomes dérivent des autonomes. Ils peuvent se mouvoir en utilisant les enzymes produites par les éléments autonomes (Schulman 2012). L'excision de l'élément provoque une cassure double-brin en amont et en aval de l'élément. La cassure est réparée par le mécanisme de « Non homologous end joining » ou par recombinaison homologue (Haber 2000). Ce mécanisme de mobilité est nommé « coupé-collé », par opposition au mécanisme de la rétrotransposition.

Enfin, un mécanisme différent des deux précédents est celui des Hélitrons, qui sont toutefois classés avec les ET de classe 2. Il implique la coupure du brin sens de l'ADN où se trouve la séquence de l'élément, dont le terminus va servir de base à la synthèse du brin d'ADN. Cette nouvelle copie s'insèrera à un nouveau locus, formant ensuite un hétéroduplex dont la réplication conduira à l'insertion.

Transferts horizontaux

Lié à leur capacité de transposition, les ET ont souvent été retrouvés impliqués dans des transferts horizontaux (TH), c'est-à-dire le passage de matériel génétique d'un organisme à un autre, sans intervention du mécanisme de la reproduction (Keeling and Palmer 2008; Schaack et al. 2010). Jusqu'à récemment, la majorité des événements de TH d'ET entre génomes eucaryotes a été décrite chez les animaux ; très peu d'évènements ont été décrits chez les champignons et les plantes. Cela était probablement dû au fait que moins de données génomiques étaient disponibles pour les plantes que pour les animaux à l'époque (mais ce n'est pas le cas pour les champignons), que ces organismes sont peu sujets aux événements de TH ou qu'il y avait un biais dans l'étude des transferts horizontaux en faveur d'espèces animales modèles comme la Drosophile (Wallau et al. 2012). Par exemple, deux des trois critères pouvant amener à l'hypothèse d'événements de TH concernent la répartition des ET considérés au sein du groupe taxonomique étudié, ainsi que la comparaison de la phylogénie des ET

considérés avec celle des organismes impliqués dans les transferts potentiels (Schaack et al. 2010). Si une phylogénie résolue n'est pas disponible pour un groupe d'espèces, il devient difficile, si ce n'est impossible, d'inférer des événements de TH d'ET entre ces espèces.

Dans une étude récente basée sur l'analyse d'un large jeu de données génomiques disponibles, des conservations nucléotidiques importantes d'ET (uniquement des LTR-rétrotransposons), considérées comme des événements potentiels de TH dans de nombreux taxons ont été recherchées (Baidouri et al. 2014). Quarante espèces de plantes séquencées appartenant aux Monocotylédones et aux Dicotylédones ont été examinées. Au total un évènement de TH a été suggéré entre une monocotylédone et une dicotylédone, huit entre ordres différents et 23 entre genres de la même famille. Les événements de TH seraient donc bien plus fréquents chez les plantes que ce que l'on aurait pu penser. Il faut cependant moduler cette conclusion. En effet, aucun mécanisme ou vecteur impliqué dans des TH chez les plantes n'a été découvert à ce jour. Il faut aussi prendre en compte la nécessaire sympatrie des espèces (si on considère que le vecteur agit à courte portée). De plus, d'autres mécanismes de conservation des séquences pourraient aussi être impliqués dans certains TH identifiés. En effet, le phénomène d'introgression, très fréquent chez les plantes, peut montrer les mêmes incongruences que celles provoquées par les événements de TH. La « domestication » d'un ET par le génome, c'est à dire la sélection d'une partie ou de toute sa séquence comme séquence codante, peut amener cet élément sous une sélection purifiante à être très conservé. Un taux d'identité nucléotidique important avec d'autres ET non domestiqués (et non sous sélection purifiante) présents dans des génomes très éloignés, pourraient alors suggérer un évènement de TH (Fortune et al. 2008). Face à des conservations importantes d'ET entre espèces, en l'absence d'un faisceau d'analyses concordant avec un possible événement de TH, il convient de rester prudent.

3. Classification des éléments transposables

En 1989, Finnegan proposa la première classification des ET basée sur les deux mécanismes principaux de transposition : la classe 1 comportant les rétrotransposons et la classe 2 représentée par des ET contenant une transposase, d'autres avec de longues répétitions terminales inversées de taille variable (Finnegan 1989). Avec de plus en plus

de type d'ET découverts, cette classe 1/classe 2 binaire a du être complétée car elle ne suffisait plus à rendre compte de la grande diversité de structure des ET dans les organismes. Capy et al. (1997) ont proposé une classification intégrant de ce fait une troisième classe, correspondant aux ET dont on ne connaissait pas le mode de transposition, en plus des deux premières. Les rétrotransposons à longues répétitions terminales et ceux n'en contenant pas formaient des sous-classes ; de la même manière, les transposons de classe 2 contenant deux types de transposase, étaient aussi divisés en deux sous-classes. Ces sous-classes contenaient les superfamilles, correspondant à plusieurs éléments de structures particulières, elles-mêmes étant composés de plusieurs familles (Lerat 2001). Toujours plus d'ET étant découverts dans les génomes séquencés, un système de classification hiérarchique rassemblant les classifications précédentes, facilitant l'annotation des ET et intégrant les différents mécanismes de transposition décrits jusque-là a été proposé (Wicker et al. 2007).

La Figure 3 présente un résumé de cette classification. Un système de code à trois lettres permet de rapidement les assigner à leurs classes (R pour classe 1 – rétrotransposons et D pour classe 2 – transposons à ADN), ordre et superfamille. Les éléments aujourd'hui appelés « non-autonomes », comme les MITE (miniature inverted transposable elements), sont rattachés à l'une des deux grandes classes puisqu'ils semblent dérivés d'ET autonomes des différentes familles décrites (Wicker et al. 2007). Ce sont ceux qui faisaient partie de la classe 3 dans la classification de Capy et al. (1997). La structure d'un ET détermine donc actuellement sa classification, nécessaire pour les travaux de recherche actuels. Elle indique également quel peut être le mécanisme de transposition de l'élément en question.

Les classifications que je viens de présenter ne correspondent qu'à une partie de ce qui est proposé actuellement par la communauté scientifique, d'autres classifications étant réalisées plus finement sur un type particulier d'ET (Piégu et al. 2015) ou bien selon les enzymes majeures des mécanismes de transposition des procaryotes (Curcio and Derbyshire 2003). Celle de Wicker et al. (2007) peut évidemment être critiquable mais elle a l'avantage d'essayer de rassembler tous les ET existants des organismes eucaryotes en une organisation universelle utile aux chercheurs quel que soit le modèle d'étude. L'augmentation des données génomiques permettront de consolider et d'adapter cette classification, particulièrement pour les séquences d'ET non-autonomes particulièrement difficiles à classer dans des familles d'éléments autonomes.

Classification		Structure	TSD	Code	Occurrence
Order	**Superfamily**				
Class I (retrotransposons)					
LTR	*Copia*	GAG AP INT RT RH	4–6	RLC	P, M, F, O
	Gypsy	GAG AP RT RH INT	4–6	RLG	P, M, F, O
	Bel–Pao	GAG AP RT RH INT	4–6	RLB	M
	Retrovirus	GAG AP RT RH INT ENV	4–6	RLR	M
	ERV	GAG AP RT RH INT ENV	4–6	RLE	M
DIRS	*DIRS*	GAG AP RT RH YR	0	RYD	P, M, F, O
	Ngaro	GAG AP RT RH YR	0	RYN	M, F
	VIPER	GAG AP RT RH YR	0	RYV	O
PLE	*Penelope*	RT EN	Variable	RPP	P, M, F, O
LINE	*R2*	RT EN	Variable	RIR	M
	RTE	APE RT	Variable	RIT	M
	Jockey	ORF1 APE RT	Variable	RIJ	M
	L1	ORF1 APE RT	Variable	RIL	P, M, F, O
	I	ORF1 APE RT RH	Variable	RII	P, M, F
SINE	*tRNA*		Variable	RST	P, M, F
	7SL		Variable	RSL	P, M, F
	5S		Variable	RSS	M, O
Class II (DNA transposons) - Subclass 1					
TIR	*Tc1–Mariner*	Tase*	TA	DTT	P, M, F, O
	hAT	Tase*	8	DTA	P, M, F, O
	Mutator	Tase*	9–11	DTM	P, M, F, O
	Merlin	Tase*	8–9	DTE	M, O
	Transib	Tase*	5	DTR	M, F
	P	Tase	8	DTP	P, M
	PiggyBac	Tase	TTAA	DTB	M, O
	PIF– Harbinger	Tase* ORF2	3	DTH	P, M, F, O
	CACTA	Tase ORF2	2–3	DTC	P, M, F
Crypton	*Crypton*	YR	0	DYC	F
Class II (DNA transposons) - Subclass 2					
Helitron	*Helitron*	RPA Y2 HEL	0	DHH	P, M, F
Maverick	*Maverick*	C-INT ATP CYP POL B	6	DMM	M, F, O

Structural features

Long terminal repeats — Terminal inverted repeats — Coding region — Non-coding region

Diagnostic feature in non-coding region — Region that can contain one or more additional ORFs

Protein coding domains

AP, Aspartic proteinase; APE, Apurinic endonuclease; ATP, Packaging ATPase; C-INT, C-integrase; CYP, Cysteine protease; EN, Endonuclease; ENV, Envelope protein; GAG, Capsid protein; HEL, Helicase; INT, Integrase; ORF, Open reading frame of unknown function; POL B, DNA polymerase B; RH, RNase H; RPA, Replication protein A (found only in plants); RT, Reverse transcriptase; Tase, Transposase (* with DDE motif); YR, Tyrosine recombinase; Y2, YR with YY motif

Species groups

P, Plants; M, Metazoans; F, Fungi; O, Others

Figure 3 : Système de classification des éléments transposables proposé par Wicker et al. en 2007. DIRS : *Dictyostelium* intermediate repeat sequence ; LINE : Long Interspersed Nuclear Element ; PLE : *Penelope*-like éléments : SINE : Short Interspersed Nuclear Element.

4. Origine des éléments transposables

L'origine des ET est particulièrement complexe à déterminer car la diversité de structure des éléments transposables est très importante. Liés à la classification de Finnegan (1989), les deux mécanismes de transposition décrits pour les rétrotransposons et les transposons à ADN ont d'abord suggéré l'existence d'un ancêtre commun à ces deux grandes classes.

Le domaine de la transcriptase inverse (ou reverse transcriptase – RT) montre une grande similarité entre les rétrotransposons et les rétrovirus, ce qui suggère un ancêtre commun à l'ensemble de ces éléments. Des arbres phylogénétiques basés sur ce domaine ont montré l'existence de deux groupes : l'un contenant les introns bactériens de type 2, l'autre formé par les rétrovirus et les LTR-rétrotransposons. De ce fait, il a été suggéré que les virus à ARN et les rétroéléments partageaient un ancêtre commun et plusieurs hypothèses sur l'origine des rétrovirus et des rétrotransposons ont vu le jour (Xiong and Eickbush 1990).

Par la suite, certains motifs, comme la signature DDE (deux acides aspartiques – D – retrouvés à un nombre variable de paires de bases d'écart et un acide glutamique – E – retrouvé 34 à 35 bases après le deuxième acide aspartique), ont permis la proposition de modèles d'évolution plus détaillés des ET (Capy et al. 1997; Eickbush and Malik 2002). Llorens et al. (2008) ont aussi tenté de préciser les relations entre les groupes de rétrotransposons et les rétrovirus, n'utilisant pas tout à fait la classification nouvellement proposée par Wicker et al. (2007). Ceci souligne la difficulté à comprendre l'évolution des ET dans leur ensemble, y compris leur(s) origine(s), malgré les structures communes retrouvées et les nombreuses études phylogénétiques réalisées. Une classification comme celle de Wicker et al. (2007) est nécessaire pour unifier les termes et codes utilisés pour chaque ET dans chaque espèce, mais elle ne représente pas forcément l'histoire évolutive de ces éléments. Il est particulièrement difficile, avec le nombre croissant d'ET découverts qui ne correspondent pas à une seule famille décrite, de comprendre leur(s) origine(s) et leur évolution, donc de proposer une classification satisfaisant toute cette complexité (Wicker 2012).

Le lien entre les virus à ARN et les rétrotransposons font de l'origine de ceux-ci un mystère digne de « qui de l'œuf ou de la poule est arrivé en premier ? ». Ils ont de plus un lien avec les transposons à ADN qui contiennent également la signature DDE et dont

certains semblent liés à des virus. Il est par conséquent impossible encore actuellement de proposer un modèle d'évolution complet et comportant tous les ET décrits à ce jour, surtout que certaines familles ont peu de motifs en commun avec les éléments de classe 1 ou de classe 2. Il est ainsi nécessaire de compléter et d'unifier les classifications proposées tout en laissant la porte ouverte à de nouvelles familles et histoires évolutives (Arensburger et al. 2016).

L'un des défis encore d'actualité sur les éléments transposables concerne leur détection dans les génomes séquencés. Celle-ci nécessite d'être toujours plus précise, notamment pour les raisons évoquées plus haut, ce qui devient possible avec la qualité des séquençages et assemblages et la disponibilité d'outils de plus en plus performants.

5. Outils de détection des ET

Il existe maintenant de nombreux outils pour détecter les ET. Le choix d'en utiliser un plutôt qu'un autre dépend principalement du type de données de séquençage disponibles et de l'organisme considéré. On peut dégager quatre grandes méthodes auxquelles correspondent plusieurs outils (Bergman and Quesneville 2007). Je vais parler de ces quatre méthodes de façon non-exhaustive et présenter les principaux outils existants, notamment ceux que j'ai pu utiliser au cours de mes travaux.

La méthode *de novo*

Cette méthode consiste, comme son nom l'indique, à rechercher des séquences d'ET parmi les répétitions de séquences anonymes dans les génomes, sans se baser sur des caractéristiques particulières (structure) propres aux ET ou sur l'existence de séquences déjà connues et décrites. De ce fait, elle dépend hautement des stratégies et qualité de séquençage et d'assemblage du génome d'étude. L'un des outils connus pour ce genre de méthode est par exemple RepeatFinder, qui utilise la sortie de répétitions exactes (correspondant à des coordonnées dans le génome) donnée par RepeatMatch ou REPuter. Ces coordonnées sont rassemblées selon leur proximité physique puis classées selon le type de répétition (Volfovsky et al. 2001). D'autres outils comme RECON (Bao and Eddy 2002) ou GROUPER (Quesneville et al. 2002) se basent sur des alignements de séquences répétées et leurs bordures dans la séquence d'ADN génomique, puis filtrent

celles qui ne correspondent pas à des ET. Enfin, certains comme RepeatScout (Price et al. 2005) analysent les k-mers (des petites séquences de longueur k, dans les lectures brutes générées par le séquençage) retrouvés à très hautes fréquences, indiquant la présence de séquences répétées potentielles.

La méthode d'homologie de séquence

Cette méthode, très utilisée, consiste à rechercher de nouvelles séquences d'ET d'après leur homologie avec des protéines d'ET déjà connues ou des bases de séquences d'ET annotés dans différentes espèces. Elle est très pratique pour obtenir rapidement la classification des ET détectés (au moins en rétrotransposons ou transposons à ADN) mais est limitée par la présence de régions codantes dans les éléments à annoter ou à une certaine conservation nucléotidique entre les éléments de références et ceux à annoter (Bergman and Quesneville 2007). De ce fait, elle n'est pas adaptée pour les éléments non-autonomes. Censor (Jurka et al. 1996) et RepeatMasker (Smit et al. 1996-2010) sont deux outils populaires de détection des ET par homologie pouvant être utilisés avec différents algorithmes d'alignements, dont BLAST+ (version accessible sur le site NCBI). Censor est implémenté sur le site de Repbase (Bao et al. 2015) qui contient aussi un grand nombre d'ET annotés. Cet outil est également téléchargeable et utilisable en ligne de commande, avec la possibilité d'indiquer sa base personnelle de séquences. RepeatMasker a été développé avec comme première intention de masquer les séquences répétées (ADN « poubelle ») et il est devenu par la suite l'outil de détection de référence pour l'annotation des génomes.

La méthode de détection d'ET par homologie de séquence se révèle très utile, lorsque l'on possède une base de données expertisée et développée dans une espèce proche de celle à analyser. C'est aussi une méthode très rapide avec les algorithmes d'alignement basés sur BLAST. Toutefois, elle est inadaptée pour identifier de nouveaux ET ou de nouvelles structures non-autonomes.

La méthode d'identification de la structure

La plupart des ET possèdent des caractéristiques structurales particulières les rendant détectables grâce à des algorithmes précis. Par exemple, les rétrotransposons à LTR possèdent deux régions dupliquées (les LTR dont la longueur varie entre 100 et 4000 paires de bases), elles même entourées par des répétitions courtes de cinq paires de bases (les TSD ou Target Site Duplication). Proche d'un LTR, dans la région interne de l'élément, se trouve une région similaire à un ARN de transfert cellulaire (ARNt) appelée PBS ou Primer Binding Site et proche de l'autre LTR, se trouve une courte région riche en bases puriques (appelée PPT ou poly-purine tract - Figure 2). Pour les transposons, ce sont des régions dupliquées inversées (de 5 à plus de 1000 paires de bases), bordant la région interne de l'élément et entourées de courtes duplications (TSD) de 2 à 11 paires de bases en fonction des superfamilles (Wicker et al. 2007). Dans certains cas, des régions sub-terminales répétées en tandem ou dupliquées inversées, spécifiques de chaque famille sont présentes. La recherche de ces structures spécifiques à chaque superfamille peut être programmée dans un algorithme de détection et plusieurs programmes sont actuellement disponibles. Ainsi, on peut citer LTR_STRUC (McCarthy and McDonald 2003), LTR_FINDER (Xu and Wang 2007) ou LTRHARVEST (Ellinghaus et al. 2008) pour la détection des rétrotransposons à LTR (LTR-RT), MITE-Hunter (Han and Wessler 2010) pour les transposons non autonomes, et HelitronFinder (Du et al. 2008) pour les Hélitrons.

La méthode de génomique comparative

Cette dernière méthode n'implique pas d'outils spécifiques à la détection d'ET mais nécessite d'avoir des données de séquençage de plusieurs génomes d'espèces proches (Caspi and Pachter 2006). Il s'agit de détecter et d'analyser les sites d'insertion et de délétion d'ET en alignant plusieurs génomes d'espèces proches. Cette méthode se rapproche d'une autre méthodologie de recherche d'insertion d'ET impliquant des techniques de re-séquençage d'individus et de comparaison directe entre des lectures simples ou des lectures par paire avec le génome de référence (Ewing 2015). Considérant le faible coût actuel du séquençage haut-débit et l'accessibilité d'outils bio-informatiques de détection (Fiston-Lavier et al. 2010), cette méthode permet d'analyser la présence d'insertions d'ET dans des centaines d'individus ou de variétés.

Il existe donc de nombreuses méthodes pour identifier les ET dans les génomes. Ces outils de détection plus ou moins précis, peuvent être combinés à des outils permettant la classification des éléments trouvés, pour éviter de le faire manuellement. Leur utilisation et choix dépendent des données que l'on a et de la précision de détection d'ET que l'on recherche (Lerat 2010).

6. Impacts des ET sur les génomes

Les ET sont présents dans tous les organismes avec des variations importantes du nombre de copies : seulement 3% chez la levure (Carr et al. 2012), 20% chez la drosophile (Adams et al. 2000), au moins 35% chez le riz (International Rice Genome Sequencing Project 2005), 45% chez l'Homme (Lander et al. 2001) et jusqu'à 85% chez le maïs (Schnable et al. 2009). L'accumulation des éléments dans les génomes est un mécanisme graduel mais peut aussi être un mécanisme brutal permettant d'augmenter et dans certains cas de doubler la taille d'un génome en quelques millions d'années (Piégu et al. 2006). Leurs capacités à se déplacer ainsi qu'à multiplier le nombre de leurs copies suggèrent que les ET ont de nombreux impacts négatifs ou positifs sur les génomes.

Leur présence dans des régions comportant des gènes peut provoquer des interactions avec ceux-ci ou avec les molécules responsables de leur régulation. Ainsi, une nouvelle insertion peut provoquer des dérégulations menant par exemple à des maladies chez l'Homme (Levin and Moran 2011) ou à des phénotypes problématiques en agronomie, comme la mutation « *mantled* » chez le palmier à huile provoquant l'avortement des fruits, donc de très bas rendements en huile (Ong-Abdullah et al. 2015). Les ET sont aussi des moteurs incroyables de l'évolution, puisqu'ils peuvent apporter des innovations évolutives et fonctionnelles aux génomes, créant de nouveaux gènes cellulaires (Volff 2006; Federoff 2012). Par exemple chez l'Homme, le gène *Gin-1*, vraisemblablement impliqué dans la répression de l'expression d'ET, largement exprimé dans plusieurs tissus et présent chez d'autres mammifères (souris, rat, vache…), est dérivé d'une intégrase d'un rétrotransposon à LTR (Llorens and Marín 2001). Plus impressionnant encore, chez les mammifères placentaires, le domaine *enveloppe* d'éléments rétroviraux endogènes (Gag) a été capturé et « domestiqué » plusieurs fois (phénomène d'évolution convergente) pour former les gènes « *syncytine* » nécessaires

au développement du placenta et à la survie de l'embryon (Lavialle et al. 2013). Beaucoup de rétrotransposons ont aussi été retrouvés dans des introns et sont utilisés comme nouveaux exons de ces gènes (Nekrutenko and Li 2001). Chez la drosophile, le génome ne contient pas de télomérase, enzyme pourtant présente chez la majorité des eucaryotes et impliquée dans la régulation de la longueur des télomères. Trois ETs de la famille *Jockey* (rétrotransposons non-LTRs) jouent ce rôle de maintien des télomères par des rétrotranspositions aux extrémités chromosomiques (Pardue and DeBaryshe 2011). Chez *Arabidopsis thaliana*, un gène très similaire aux transposases de type *hAT* (grande famille d'ETs de classe 2) est indispensable au développement normal de la plante, il provient vraisemblablement d'une transposase « domestiquée » (Bundock and Hooykaas 2005).

Les génomes « hôtes » ne sont pas sans défense face à l'activité des ET. L'activité transcriptomique des ET peut être régulée par des mécanismes épigénétiques, correspondant à des changements dans l'expression des gènes (par extension, des éléments transposables), n'impliquant pas de changements dans la séquence d'ADN (Martienssen and Chandler 2013). La méthylation des cytosines, par l'ajout de groupements méthyle CH_3 empêchant l'accès des enzymes de transcription à l'ADN, est l'un de ces mécanismes. Il existe un mécanisme différent, celui du « RNA silencing » ou siRNA. Il consiste en l'utilisation de petits ARN de 21 à 35 nucléotides pour cibler la dégradation des ARN, comme ceux d'ET (Rigal & Mathieu 2011). Enfin, il existe un mécanisme épigénétique existant uniquement chez les plantes : celui de la méthylation de l'ADN dirigée par des ARN (« RNA-directed DNA methylation – RdDM). Deux ARN polymérases (présentes uniquement chez les plantes) et 24 petits ARN interférants se liant aux protéines de la famille *Argonaute* modifient l'état de méthylation de l'ADN (Fedoroff 2012).

J'ai voulu présenter ici les généralités sur tous les ET, quels que soient les génomes considérés. Ma thèse portant sur les LTR-rétrotransposons dans les génomes de caféiers, je vais maintenant exposer, dans une introduction plus ciblée, les ET (plus particulièrement des LTR-rétrotransposons) dans les génomes des plantes.

Chapitre 2 – ÉTAT DE L'ART

1. Composition des génomes des plantes en ET

Depuis les travaux de B. McClintock sur le génome du maïs, beaucoup de recherches ont été réalisées sur les nombreuses espèces de plantes cultivées. Les ET étant présents dans ces génomes peuvent nous informer sur leur histoire évolutive, si l'on étudie leur diversité et les mécanismes ayant amené à ce que l'on observe actuellement.

Taille des génomes végétaux et LTR-RT

Les génomes des plantes contiennent un grand nombre d'ET, particulièrement des rétrotransposons à LTR (LTR-RT), car leur mécanisme de mobilité peut entraîner un accroissement du nombre de leurs copies. Si l'on considère la variation de la *C*-value (synonyme de génome de base uniquement chez les espèces diploïdes) entre le plus petit et le plus grand génome diploïde de plante connus à ce jour, on obtient une variation de 2300 fois (Bennett and Leitch 2012). Cette prévalence de génomes végétaux de taille impressionnante a été appelée « obésité génomique » et est attribuée à une quantité importante de séquences répétées accumulées et parmi celles-ci d'ET (Bennetzen and Kellogg 1997). En effet, les copies de LTR-RT participent massivement à la taille des génomes des plantes (Kejnovsky et al. 2012) et représentent par exemple 75% du génome du maïs (Schnable et al. 2009).

Les superfamilles *Gypsy* et *Copia* sont les ET les plus présents dans les génomes végétaux (Figure 4). Celles-ci sont formées de plusieurs grandes lignées sur la base de la structure des éléments et sur la phylogénie des domaines codant comme la transcriptase réverse (RT) (Llorens et al. 2008, 2009). Comme toute classification, celle proposée par Llórens peut être discutable, et malgré quelques données manquantes elle sera utilisée

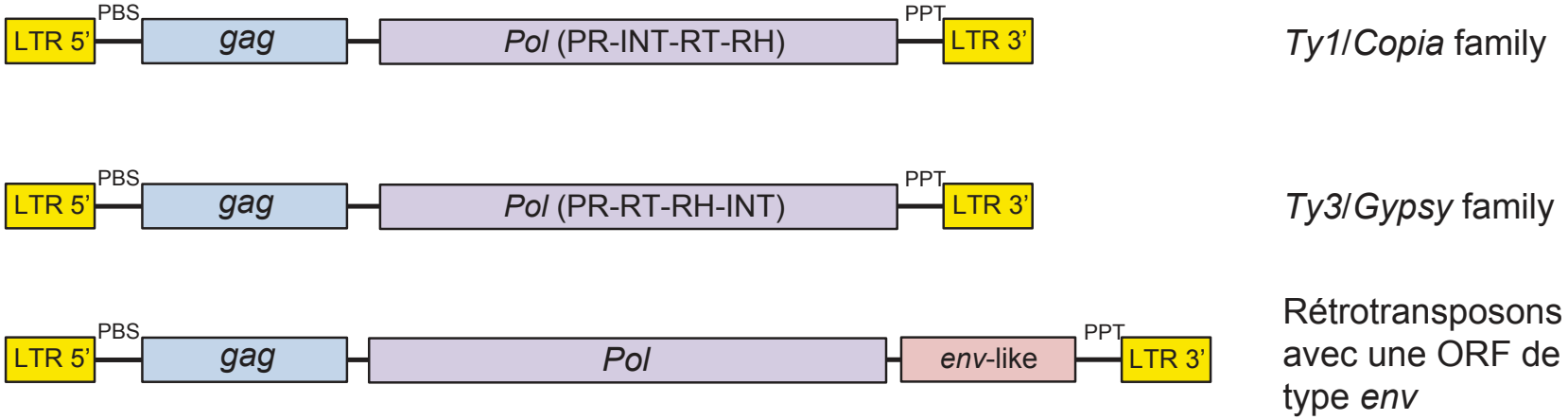


Figure 4 : Schéma des LTR-Rétrotransposons présents dans les génomes des plantes. D'après GyDB.
LTR : Long Terminal Repeats ; PBS : Primer Binding Site ; PR : protéinase ; INT : intégrase ; RT : reverse transcriptase ; RH : RNase H ; PPT : Polypurine Tract ; ORF : Open Reading Frame.

dans les chapitres suivants pour l'identification et l'annotation des LTR-RT.

Chez les eucaryotes, les *Gypsy* sont constitués de 22 lignées, formant deux branches principales – l'une correspondant uniquement aux *Chromovirus*, qui ont un chromodomaine (impliqué dans des interactions avec la chromatine) dans la région C-terminale de l'intégrase. Les *Copia* sont eux composés de 14 lignées, distribuées également dans deux branches majeures. Chez les plantes (dont les algues vertes, considérées comme *Viridiplantae*), les lignées identifiées sont pour les *Gypsy* : *CRM* (Centromeric Retrotransposon of Maize), *Del*, *Galadriel*, *Reina* et *REM-1* (branche 1 – *Chromovirus*) et *Athila* et *Tat* (branche 2) et pour les *Copia* : *Sire*, *Retrofit*, *Tork*, *Oryco* et *Osser* (branche 2). La branche 1 des *Copia* contient des éléments présents uniquement chez des animaux, des diatomées ou des champignons. Ces lignées sont décrites sur le site « GyDB » (http://gydb.org/index.php/Main_Page). Il y a cependant une donnée manquante dans les études de Llórens et collaborateurs. L'étude des LTR-RT *Copia* et leur évolution chez les Triticeae (plus le riz et Arabidopsis) permit de détecter une nouvelle lignée, *Bianca* (Wicker and Keller 2007), qui n'est pas référencée dans la base de données de GyDB. J'ai donc utilisé les séquences de référence de cette étude pour l'inclure dans mes travaux.

Cycle de vie d'un LTR-RT

Le cycle de vie d'un LTR-RT est assez complexe, et réalisé par un ensemble d'enzymes le plus souvent synthétisées par l'élément lui-même, ou dans certains cas, synthétisées par un autre élément. La Figure 5 montre le cycle de rétrotransposition d'un LTR-RT (Schulman 2012).

La séquence de l'élément est transcrite à partir d'un promoteur situé dans le LTR en 5', résultant en un ARNm poly-adenylé (1 et 2). Ce transcrit exporté hors du noyau a deux fonctions : i) la traduction en protéines nécessaires au cycle de rétrotransposition, ii) une matrice nécessaire à la transcription réverse (3). La traduction produit une protéine d'encapsidation, la Gag et une poly-protéine (Pol) contenant la protéinase, la RNase H, la transcriptase réverse (RT) et l'intégrase (INT) (4). L'encapsidation des ARN messagers (ARNm) couplés permet l'initiation des mécanismes de transcription inverse (5). La particule pseudovirale contient également la RNase H, l'INT et la RT (6). Le complexe

protéique encapsidé est ensuite redirigé vers le noyau où l'INT procède à son insertion dans le génome (7 et 8).

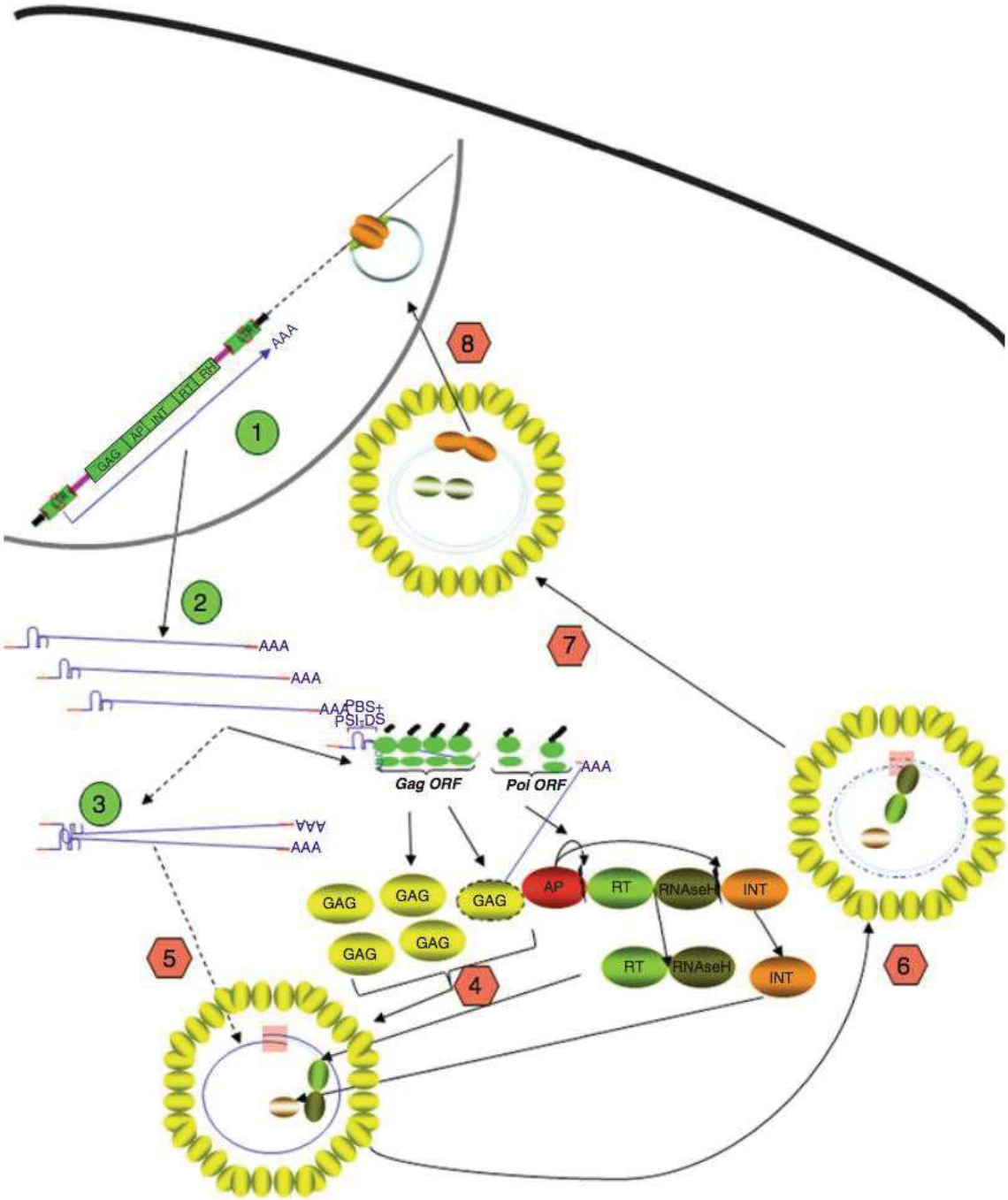


Figure 5 : Cycle de rétrotransposition d'un LTR-rétrotransposon (D'après Schulman 2012). Le trait courbé épais représente la membrane cellulaire, celui plus fin la membrane nucléaire. Chaque chiffre représente les étapes principales de la rétrotransposition, explicitées ci-contre.

Localisation des LTR-RT

La répartition des LTR-RT dans les génomes végétaux semble être relativement commune aux Angiospermes. On les trouve généralement accumulés dans les régions péri-centromériques et distales des chromosomes (Lee and Kim 2014). Cependant, cette répartition globale varie selon les ETs considérés et les génomes. En effet, les *Gypsy* ont une tendance a être plus souvent accumulés dans des régions d'hétérochromatine comme celles précédemment citées, alors que les *Copia* semblent distribués également le long des chromosomes dans les régions hetérochromatiques et euchromatiques. Cette distribution est généralement observée pour les génomes de petite et de moyenne taille. Pour les génomes de très grande taille comme ceux du maïs ou du blé, les LTR-RT sont retrouvés accumulés en très grand nombre dans toutes les régions intergéniques, laissant les gènes regroupés en petit nombre dans des « îlots » géniques au milieu d'un « océan » d'ET (Fedoroff and Bennetzen 2013). Cette structure est observée pour le génome du maïs (Schnable et al. 2009), du sorgho (Paterson et al. 2009) et du blé (Wicker et al. 2001). Les îlots de gènes semblent de plus en plus contractés lorsque les génomes contiennent un très grand nombre d'ET (Fedoroff and Bennetzen 2013). La très grande quantité d'ET dans ces génomes crée des empilements ou « poupées russes » d'ET insérés les uns dans les autres, comme cela a été démontré pour la première fois dans la région *adh1-F* du maïs (SanMiguel et al. 1998). Cette distribution suggère que les régions géniques sont relativement préservées de l'insertion des ET, ou que ces insertions probablement délétères dans ces génomes sont soumises à une forte sélection négative (donc ne sont pas observables).

Les répartitions génomiques sensiblement différentes des *Gypsy* et des *Copia* peuvent aussi faire penser qu'ils ont des sites d'insertion préferentiels (Fedoroff and Bennetzen 2013). Cependant, l'insertion des nouvelles copies requiert la reconnaissance de séquences très courtes par l'intégrase des LTR-RT. De telles séquences statistiquement très fréquentes dans les génomes des plantes contribueraient donc peu aux patrons de distribution observés. Bien sûr, d'autres paramètres comme le niveau de compaction de l'ADN peuvent influencer l'intégration des nouvelles copies des LTR-RT. Chez les éléments *Ty* de la levure, l'intégrase interagit avec des protéines membranaires qui ont le rôle de facteurs de liaison. Ces protéines reconnaissent des marques d'histones spécifiques ou se lient directement à l'ADN selon la conformation de la chromatine. Ceci peut également influencer l'intégration des nouvelles copies (Sultana et

al. 2017). Toujours chez la levure, des insertions préférentielles ont été observées pour *Ty1* dans une fenêtre de 750 paires de bases en amont des gènes transcrits par la Polymérase III. Le génome de la levure étant très dense en régions codantes, ces insertions préférentielles peuvent traduire une pression de sélection contre une insertion délétère dans des régions codantes (Bushman 2003).

Il est donc important de considérer tous ces paramètres lorsque l'on étudie la distribution chromosomique des LTR-RT dans les génomes des plantes. Leur localisation et les potentielles « préférences » de sites d'insertion impactent le génome, tout comme leur capacité à augmenter le nombre de leurs copies.

2. Impacts des LTR-RT sur les génomes des plantes

Le mécanisme de mobilité des LTR-RT peut provoquer non seulement l'augmentation de leur nombre de copies, mais une nouvelle insertion peut mener à des mutations délétères ou une altération de la transcription des gènes pouvant aussi impliquer des modifications épigénétiques de l'ADN environnant (Casacuberta and González 2013). De ce fait, la grande diversité des ET chez les plantes, leurs modes de mobilité et leur accumulation parfois brutale ont un impact important sur la composition, la structure et l'évolution des génomes et des gènes chez les plantes.

Variation de la taille des génomes

L'augmentation du nombre de copies menant à l'« obésité » génomique observée chez beaucoup de plantes peut se faire par activations régulières de l'ensemble des familles de LTR-RT, ou irruptions plus soudaines et intensives (aussi appelées « burst »), de quelques familles de LTR-RT. Les génomes de petites et moyennes tailles semblent posséder les outils permettant de contrôler la transcription et d'éliminer efficacement les ET. Des mécanismes épigénétiques pré et post-transcriptionels (Mirouze et al. 2009) peuvent contrôler la mobilité des éléments alors que les mécanismes de recombinaisons illégitimes (intra ou intermoléculaire) ou de recombinaisons homologues inégales sont responsables du maintien de la taille des génomes du riz ou d'*Arabidopsis* (Kejnovsky et al. 2012; Lee and Kim 2014). Ces mécanismes existent dans les génomes de grande taille comme le maïs ou le blé, mais ne semblent pas avoir été suffisants pour contrôler la

mobilité des éléments. Dans le genre *Oryza*, plusieurs exemples montrent des amplifications et des contrôles de la mobilité des LTR-RT totalement différents. Alors que la mobilité des LTR-RT semblent régulée par des mécanismes post-insertion chez le riz cultivé asiatique (*O. sativa* var *japonica*) (Kejnovsky et al. 2012; Lee and Kim 2014), le plus gros génome diploïde du genre, *Oryza australiensis* (965 Mb) montre une amplification exceptionnelle de trois familles de LTR-RT : *RIRE1* (*Copia*), *Wallabi* et *Kangourou* (*Gypsy*). À eux seuls, ils représentent environ 60% du génome d'*O. australiensis* (Piégu et al. 2006) expliquant la différence de taille avec les riz cultivés asiatiques et africains. Similairement, *O. granulata,* une espèces sauvage de riz asiatique, possède un génome de grande taille (882 Mb), composé à environ 25% par une seule famille de LTR-RT *Gypsy* : *Gran3* (Ammiraju et al. 2007). Chez les espèces de riz diploïdes, la taille des génomes semble corrélée à la quantité de séquences de LTR-RTs (Zuccolo et al. 2007). Il en est généralement de même avec d'autres plantes appartenant aux Poacées comme le maïs, dont le génome a doublé en taille (1200 à 2400 Mb) en quelques millions d'années dû à l'activité des LTR-RTs (SanMiguel et al. 1998). Chez les dicotylédones, les LTR-RT peuvent également jouer un grand rôle dans la taille des génomes. Chez les Brassicacées, *Arabidopsis thaliana* présente une dynamique de rétrotransposition différente de celle d'*A. lyrata* ou *A. alpina*, qui ont de plus gros génomes et une proportion bien plus importante de LTR-RT (Agren and Wright 2015). Les espèces de coton diploïdes montrent une variation importante de la taille de leurs génomes, cette variation étant principalement due aux LTR-RT. Ici, une famille *Gorge3* (*Gypsy*) a subi des amplifications importantes et différentes selon les génomes de *Gossypium* au cours des 5 à 10 derniers millions d'années (Hawkins et al. 2006).

Dans le cas de génomes allopolyploïdes comme ceux du coton cultivé, du tabac ou encore du blé, on observe une variation de taille des génomes et de contenu en ET qui ne correspond pas à la simple addition de celle des génomes progéniteurs et des altérations épigénétiques (Parisod and Senerchia 2012). Ces variations suggèrent des mécanismes de variation du nombre de copies d'ET concomitants au processus de polyploïdisation ('Revolutionary changes') ou *a posteriori* sur un plus long terme ('Evolutionary changes'), que ce soit une perte ou au contraire une augmentation du nombre des copies. Par exemple, plus de 40% des copies de *Tnt1* ont été éliminés dans le sous-génome paternel de *Nicotiana tabacum* (Parisod and Senerchia 2012). Il en est de même chez le blé et le coton avec une perte importante sur le long terme du nombre de copies

de divers ET (Eilam et al. 2008; Hu et al. 2010). Dans le genre *Panax* (Ginseng) une situation contrastée est observée. Dans les génomes diploïdes, plusieurs familles de *Del* (*Gypsy*) sont responsables de la variation de taille des génomes (Lee et al. 2017), mais *PgDel1,* une famille particulière chez *Del* aurait subi une amplification du nombre de copies dans l'espèce allotétraploïde *Panax quinquefolius* et pas dans l'espèce allotétraploïde *P. ginseng*. Ces mécanismes de variation du nombre d'éléments suggèrent que la polyploïdisation pourrait être considérée comme un choc génomique ayant un impact significatif sur les ET (Parisod et al. 2010).

Ces mécanismes évolutifs ont été décrits uniquement chez des plantes annuelles. La dynamique des taille des génomes et l'impact des ET sur l'évolution des génomes ont été peu étudiées chez les plantes pérennes ligneuses (Vicient & Casacuberta 2017). Les quelques études concernant les rétrotransposons dans ce type d'espèces n'ont pas pu mettre clairement en évidence un lien entre les variations de taille de génomes observées (quand de telles variations ont été observées) et l'activité des LTR-RT (Faivre Rampant et al. 2011; Alves et al. 2012; Jiang et al. 2016; Rockinger et al. 2016).

Réarrangements chromosomiques

L'activité des LTR-RT provoque des bouleversements génétiques jusqu'au niveau chromosomique. En effet, des mécanismes tel que la recombinaison homologue inégale amenant à la formation des solo-LTR (Figure 6), peuvent aussi provoquer l'élimination de portions chromosomiques entières dans des zones où sont accumulés beaucoup de rétroéléments. La recombinaison entre deux LTR-RT à des positions chromosomiques différentes provoque une « cassure » des chromosomes. Des rétrotranspositions avortées peuvent éliminer l'ADN adjacent. Des délétions plus ou moins grandes sont ainsi observées dans les génomes d'*A. thaliana*, du blé, du maïs ou encore du riz (Bennetzen 2005).

Ces mécanismes de recombinaison sont en partie responsables de la réorganisation ainsi que des phénomènes de contraction des génomes. Des pertes fréquentes de séquences de petites tailles ont aussi un rôle dans la contraction des petits génomes. Par exemple, *A. thaliana* perd des introns six fois plus vite qu'*A. lyrata*, qui a un génome plus

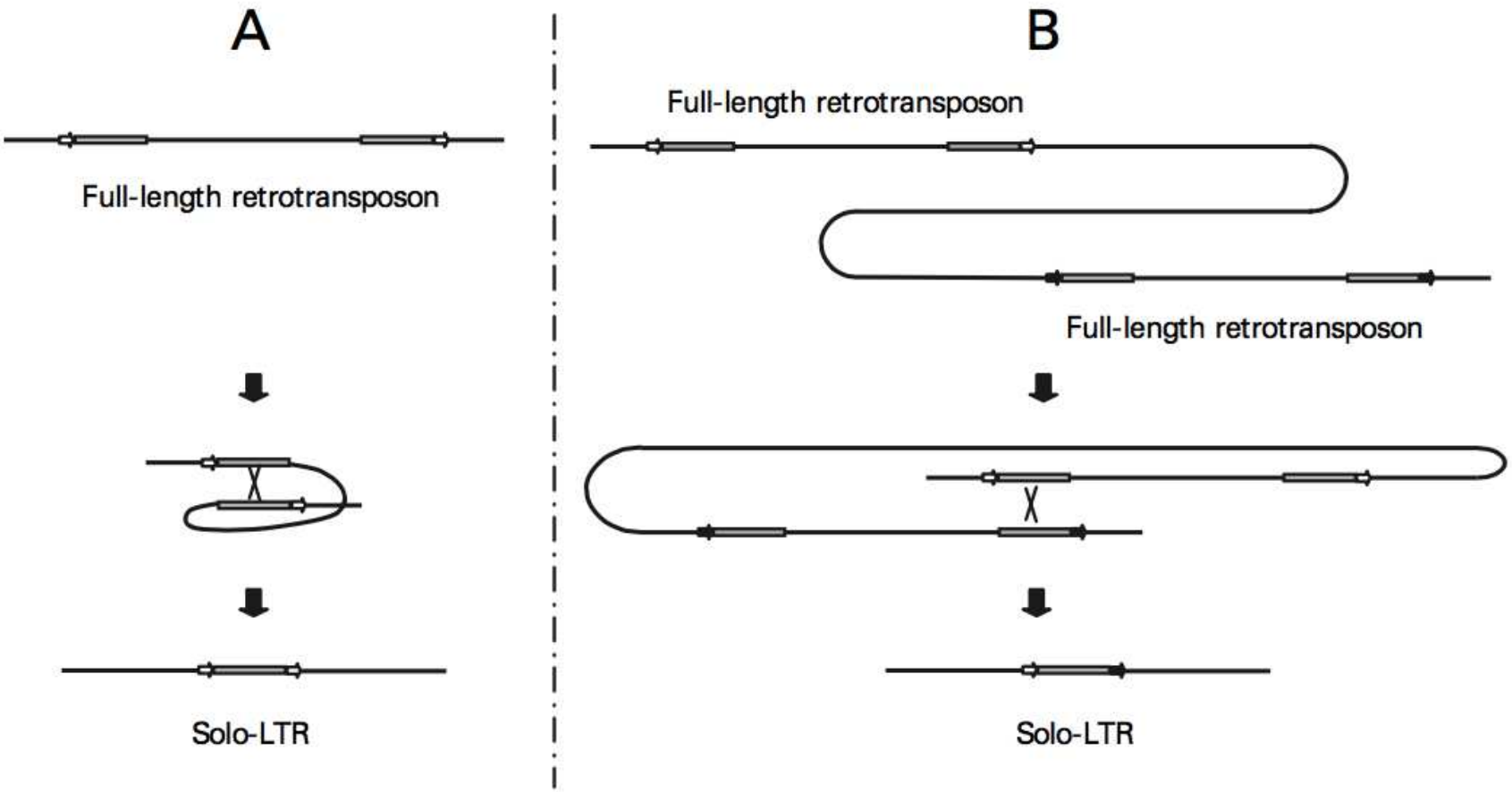


Figure 6 : Mécanismes de formation d'un solo-LTR par recombinaisons homologues inégales. De Vitte & Panaud 2005.

gros et une proportion plus importante de LTR-RTs (Lee and Kim 2014). Chez les allopoplyploïdes, l'hybridation interspécifique provoquerait d'importants réarrangements chromosomiques (Parisod et al. 2010). La perte de séquences par recombinaison illégitime ou encore la mobilisation de certains ET juste après la polyploïdisation mèneraient à une réorganisation des génomes, parfois plus marquée pour l'un des deux sous-génomes parentaux dans le cas des allotétraploïdes (Vicient and Casacuberta 2017). Une mobilisation et une perte assez importante de séquences de plusieurs familles d'ET ont ainsi été observées dans le sous-génome paternel d'*Oryza minuta,* un riz sauvage allotétraploïde originaire d'Asie. Ces réarrangements génomiques ont également un impact indirect sur les gènes, puisque ceux-ci, s'ils se trouvent près d'insertions de rétroéléments, peuvent être « emportés » (du moins en partie) lors d'un processus d'élimination (Parisod et al. 2010).

Régulation des gènes

On peut déterminer différents types d'impacts des ET sur les gènes. Ceux-ci sont résumés en Figure 7. Chaque numéro indique un endroit possible d'insertion d'un LTR-RT près ou dans un gène, puis ce qui peut en résulter au niveau protéique.

Une nouvelle insertion, si elle se situe dans un gène (Figure 7, 1 et 2), peut interrompre sa continuité et donc le rendre non-fonctionnel. Ceci peut être illustré par un exemple récent chez le melon (*Cucumis melo* L.) : quand la plante se situe dans un environnement pauvre en fer, on observe une augmentation de l'expression de gènes codant pour des protéines responsables de la capture du fer. Un mutant spontané est incapable d'induire l'augmentation de l'expression de ces gènes si la plante manque de fer. En fait, l'insertion d'un LTR-RT *Copia* dans le gène correspondant au facteur de transcription régulant ces gènes a provoqué son dysfonctionnement. Il en résulte des protéines tronquées et une incapacité du mutant à augmenter la capture de fer (Ramamurthy and Waters 2017). Il en est de même chez la tomate, dans laquelle l'insertion du rétrotransposon *RIDER* dans le premier exon du gène FER induit un phénotype similaire (Guyot et al. 2005).

L'insertion d'un LTR-RT près ou dans le promoteur d'un gène (Figure 7, 5) peut modifier la régulation de ce gène, ce qui peut avoir une incidence sur le phénotype de la

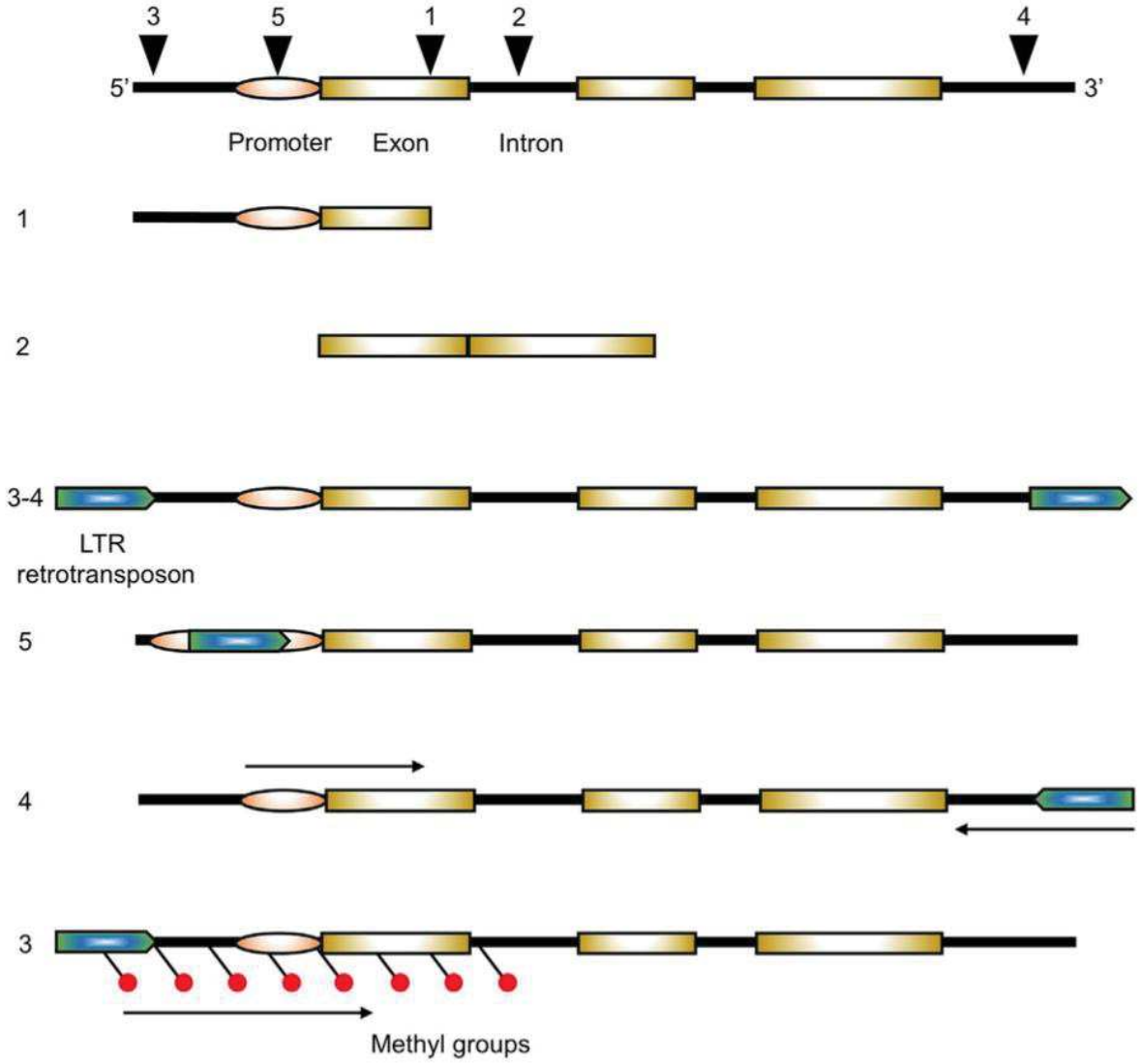


Figure 7 : Schémas représentant les endroits où un ET peut s'insérer (flèches noires) et ce qui peut en résulter. D'après Galindó-Gonzalez et al. 2017. Le trait noir épais représente une séquence ADN contenant un gène (promoteur en rose, exons : rectangles beiges, introns) et le rectangle bleu et vert représente un LTR-RT ou un solo-LTR.

plante. Ces modifications provoquant l'apparition de phénotypes différents peuvent jouer un rôle important dans la sélection humaine des plantes cultivées d'intérêt économique. Par exemple, les différentes couleurs du grain de raisin obtenues selon les cépages pour une même espèce (*Vitis vinifera*) sont provoquées par l'insertion et la perte partielle d'un LTR-RT dans le promoteur d'un gène responsable de la fabrication d'anthocyanines. Quand l'insertion est présente, les anthocyanines ne sont pas synthétisées, donnant la couleur blanche des grains de raisin au lieu de la couleur noire. Si une recombinaison de l'ET (aboutissant à un solo-LTR) a lieu, le gène est régulé différemment et des grains rouges ou roses sont obtenus (This et al. 2007). Un mécanismes similaire est à l'origine de la couleur des pétales des fleurs de la Torénie de Fournier (Nishihara et al. 2014), mais à l'opposé, l'insertion d'un élément active la production d'anthocyanines dans les variétés d'orange sanguine (Butelli et al. 2012).

Parfois, la transcription d'un LTR-RT continue après sa séquence et englobe un gène adjacent entier ou en partie (Figure 7, 4 et insertion en 3). La forme allongée du fruit de la tomate provient de l'insertion de *RIDER* (*Copia*) à un locus particulier, SUN. Sa transcription réverse dépassant le rétrotransposon lui-même et donnant un ADN codant comprenant *Rider*, SUN et 3 autres gènes se sont insérés dans le gène *DEFL1* qui contrôle la forme du fruit. L'accroissement de l'expression des gènes a résulté en la modification de la taille et la forme du fruit (Jiang et al. 2012). Il a aussi été démontré chez le blé que la transcription simultanée d'un gène et d'une séquence d'un LTR-RT dans le sens contraire peut mener à la production d'un ARN double-brin, responsable de l'inactivité du gène (« silencing » post-transcriptionnel) (Kashkush et al. 2002). Selon sa localisation, une nouvelle insertion peut également « créer » un nouveau gène si elle provoque une mutation non-délétère ou apporte une nouvelle séquence codante (Figure 7, 3-4) (Lisch 2013b). Enfin, si un ET réprimé par des mécanismes épigénétiques (groupements méthyles par exemple – 3) s'insère près d'un gène, celui-ci peut à son tour subir ces modifications et ainsi devenir inactif. Ceci a été démontré assez tôt chez le maïs (Lippman et al. 2004).

Mécanismes épigénétiques et stress

Il est observé depuis longtemps que la mobilité des rétrotransposons peut être activée à la suite de stresses biotiques ou abiotiques alors qu'ils sont sous contrôle

épigénétique (« silencing ») (Kumar and Bennetzen 1999). Chez les Diatomées par exemple, le LTR-RT *Copia Blackbeard* s'active lors d'un manque en nitrate (Maumus et al. 2009). Des réactions similaires à des stresses dus à des manques de nutriments ou à la présence de certaines molécules comme l'acide abscissique ont été documentées chez le tabac ou l'orge (Casacuberta and González 2013). Les cultures cellulaires représentent aussi un stress considérable pouvant également provoquer des modifications au niveau du contrôle épigénétique des ET et ayant des effets négatifs par la suite, comme ce fut le cas pour les cultures de palmiers à huile (Ong-Abdullah et al. 2015). Chez le riz, la culture de tissus provoque l'activation de *Tos17*, et son nombre de copies augmente avec la durée de la culture (Hirochika et al. 1996; Hirochika 2001). Chez les polyploïdes comme le blé, des changements dans les empreintes épigénétiques sont observés directement après l'hybridation interspécifique et pourraient impliquer une modification de l'expression et de la mobilité des ET (Parisod et al. 2010). L'activation des ET serait donc une conséquence du relâchement du contrôle épigénétique induit par les stresses, dont semblent faire partie les hybridations interspécifiques conduisant à la polyploïdie.

L'hypothèse actuelle la plus « simple » considère les éléments transposables comme un vecteur induit par des stress permettant une réponse rapide pour une adaptation à l'environnement (Casacuberta and González 2013). Il a été proposé que dans certains cas, la plante pouvait lever le contrôle épigénétique d'ET spécifiques proches de certains gènes pouvant apporter une adaptation particulière (Ito 2012). Toutefois l'hypothèse de Casacuberta and González 2013 semble plus complexe qu'une simple dérégulation du contrôle, impliquant la spécificité des familles d'ET et la spécificité des stresses avec parfois une fonction adaptative qui reste à prouver.

3. Utilisation des LTR-RT comme outils moléculaires

Les LTR-RT étant très répandus chez les plantes et à diverses localisations dans les génomes, ils peuvent être utilisés comme outils pour différentes études de diversité génétique. Un même élément pouvant montrer des différences de sites d'insertion entre plusieurs individus d'une même espèce, différentes techniques (REMAP, S-SAP) ont été utilisées pour visualiser le polymorphisme d'insertion d'ET au sein et entre populations d'une même espèce et comprendre sa dynamique d'insertion. Ceci a été développé chez

l'orge avec *BARE-1* (*Copia)* présent en très grand nombre de copies dans les génomes des céréales (Kumar et al. 1997). Des ET présentant des polymorphismes d'insertion peuvent donc être utilisés afin de comprendre les relations génétiques entre les espèces d'un genre (Mhiri and Grandbastien 2004).

Ils peuvent également être utilisés comme agents mutagènes. Ceci a été réalisé dans plus de 55000 lignées de riz, régénérées par culture cellulaire pour activer l'élément *Tos17* (*Copia*), créer de nouvelles insertions et de nouveaux phénotypes et comprendre la fonction des gènes altérés (Piffanelli et al. 2007).

4. Les LTR-RT chez les caféiers

Le modèle d'étude de l'équipe est le genre *Coffea*, appartenant aux Rubiacées (ordre des Gentianales, Angiospermes). Il est classé dans la sous-famille des Ixoroideae et la tribu des Coffeeae (Robbrecht and Manen 2006).

Le genre *Coffea*

Les caféiers sont représentés par 139 espèces, réparties en Afrique, dans les îles de l'Océan Indien (IOI) et en Asie au sens large (incluant le continent indien et le nord de l'Australie). Ce sont des hôtes naturels des forêts inter-tropicales. Leur distribution géographique indique des adaptations très spécifiques : une capacité à se développer dans des milieux très différents (de très secs à très humides, voire temporairement inondés), une durée du cycle de floraison et de fructification allant de 2 à 13 mois, et une capacité variable à produire de la caféine dans les grains matures (Couturon et al. 2016).

i. Caractérisation botanique et distribution géographique

Les caféiers ont longtemps été séparés en deux genres, notamment par l'observation de différences morphologiques se rapportant aux feuilles et surtout aux fleurs. Le genre *Coffea sensu stricto* (ex-*Coffea* maintenant), représenté aujourd'hui par 119 espèces, présente des fleurs ayant un tube de la corolle court (ne dépassant pas la longueur des pétales), des anthères et un style non inclus. Le genre *Psilanthus*, contenant 20 espèces, montre au contraire des fleurs avec un long tube corollaire (plus long que les pétales) et des anthères incluses toujours situées au-dessus d'un style très court (Figure 8).

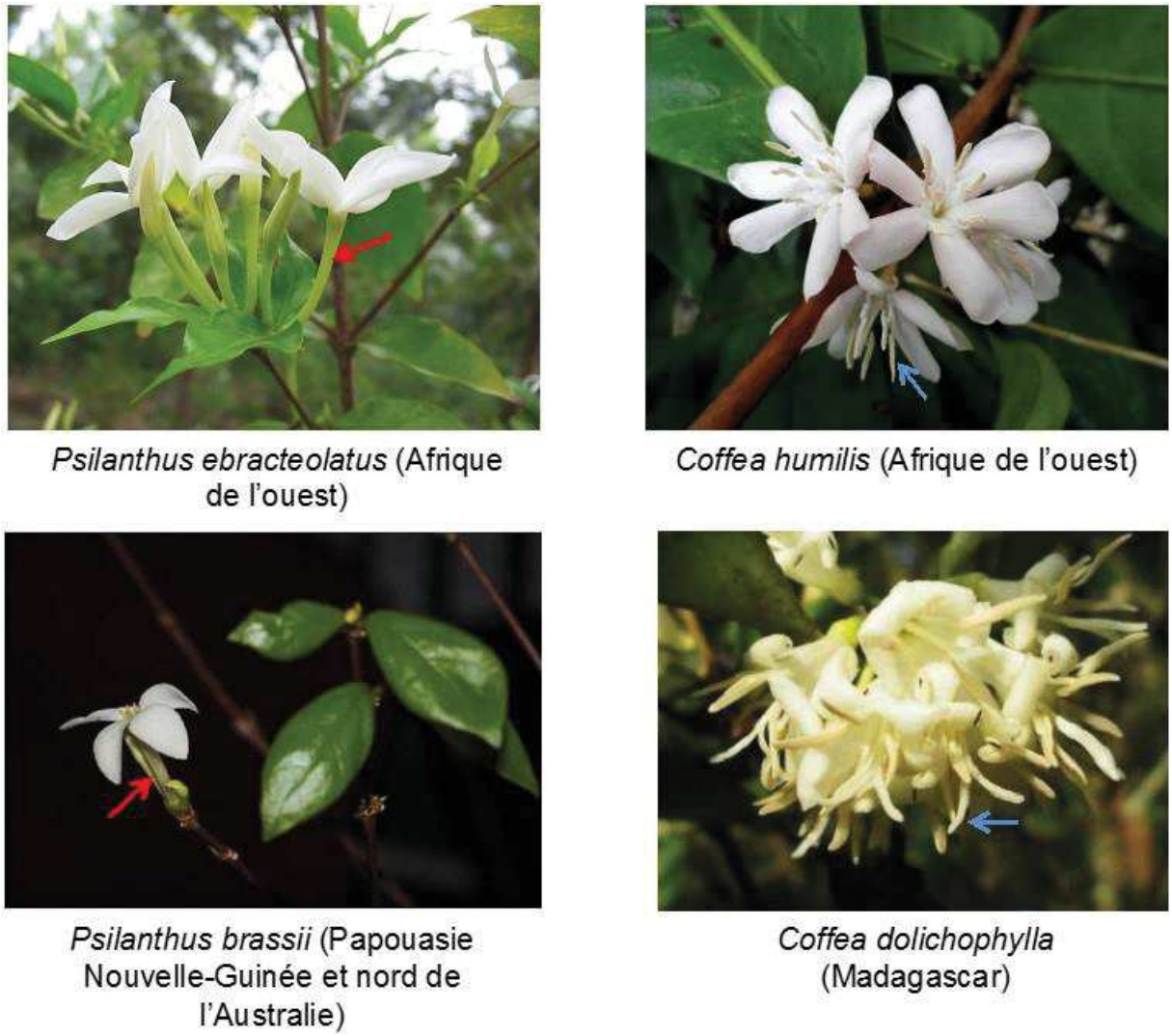


Figure 8 : Photographies de 4 espèces de caféiers montrant la différence morphologique des fleurs des *Psilanthus* et des *Coffea*. Flèches rouges : long tube corollaire, flèches bleues : anthères et style non-inclus.

De plus, les ex-*Coffea* se trouvent sur le continent africain et dans les îles de la région Ouest de l'Océan Indien (IOI), mais sont absents en Asie, alors que les ex-*Psilanthus* sont présents en Afrique et en Asie, mais absents des IOI (Charrier and Berthaud 1985; Davis et al. 2006). La majorité des travaux n'ayant porté que sur les ex-*Coffea*, les espèces du groupe des ex-*Psilanthus* ont été de ce fait très peu étudiées.

Des groupes biogéographiques ont été déterminés pour les ex-*Coffea* sur la base de leur distribution géographique et de l'absence de caféine dans les grains : les Eucoffea en Afrique de l'ouest et du centre, les Mozambicoffea en Afrique de l'est et les Mascarocoffea dans les IOI. Plus récemment, les deux genres ont été rassemblés en un

seul (Davis et al. 2011) formant le nouveau genre *Coffea*. La Figure 9 montre la carte de répartition géographique des espèces du genre *Coffea*. Un endémisme total est observé entre les grandes régions et il est très important au sein même de ces régions. Seulement deux espèces en Afrique (*C. canephora* et *C. liberica*) et à Madagascar (*C. perrieri* et *C. millotii*) ont des distributions géographiques importantes (Davis et al. 2006). Les caféiers montrent des caractères phénotypiques (formes des fleurs, couleurs des fruits, port de l'arbre…) très divers et adaptés à leurs différents milieux de vie, allant des forêts sèches à très humides et poussant sur des sols sableux à latéritiques (Couturon et al. 2016).

ii. Génétique et génomique

Toutes les espèces de caféiers décrites sont diploïdes (2n = 2x = 22 chromosomes) (Bouharmont 1959; Louarn 1976), à l'exception de *Coffea arabica* qui est un allotétraploïde (Carvalho 1952). Ce dernier proviendrait d'un croisement naturel entre *C. canephora* et *C. eugenioides* (Lashermes et al. 1999) s'étant produit assez récemment (0,046 à 0,665 million d'années, (Yu et al. 2011). Malgré le nombre constant de chromosomes des espèces du genre, les tailles des génomes sont variables, allant de 469 Mb pour *C. humblotiana* et *C. mauritiana* à 900 Mb pour *C. humilis*. Ces variations suivent un gradient géographique avec une augmentation de la taille des génomes du nord-est au sud-ouest à Madagascar (Razafinarivo et al. 2012) et de l'est à l'ouest en Afrique (Noirot et al. 2003). Étant donné le manque d'études spécifiques aux ex-*Psilanthus*, exception faite pour *P. ebracteolatus* (même taille de génome que *C. pseudozanguebariae* soit 594 Mb, Cros et al. 1994), de telles informations ne sont pas disponibles pour ces espèces sauvages.

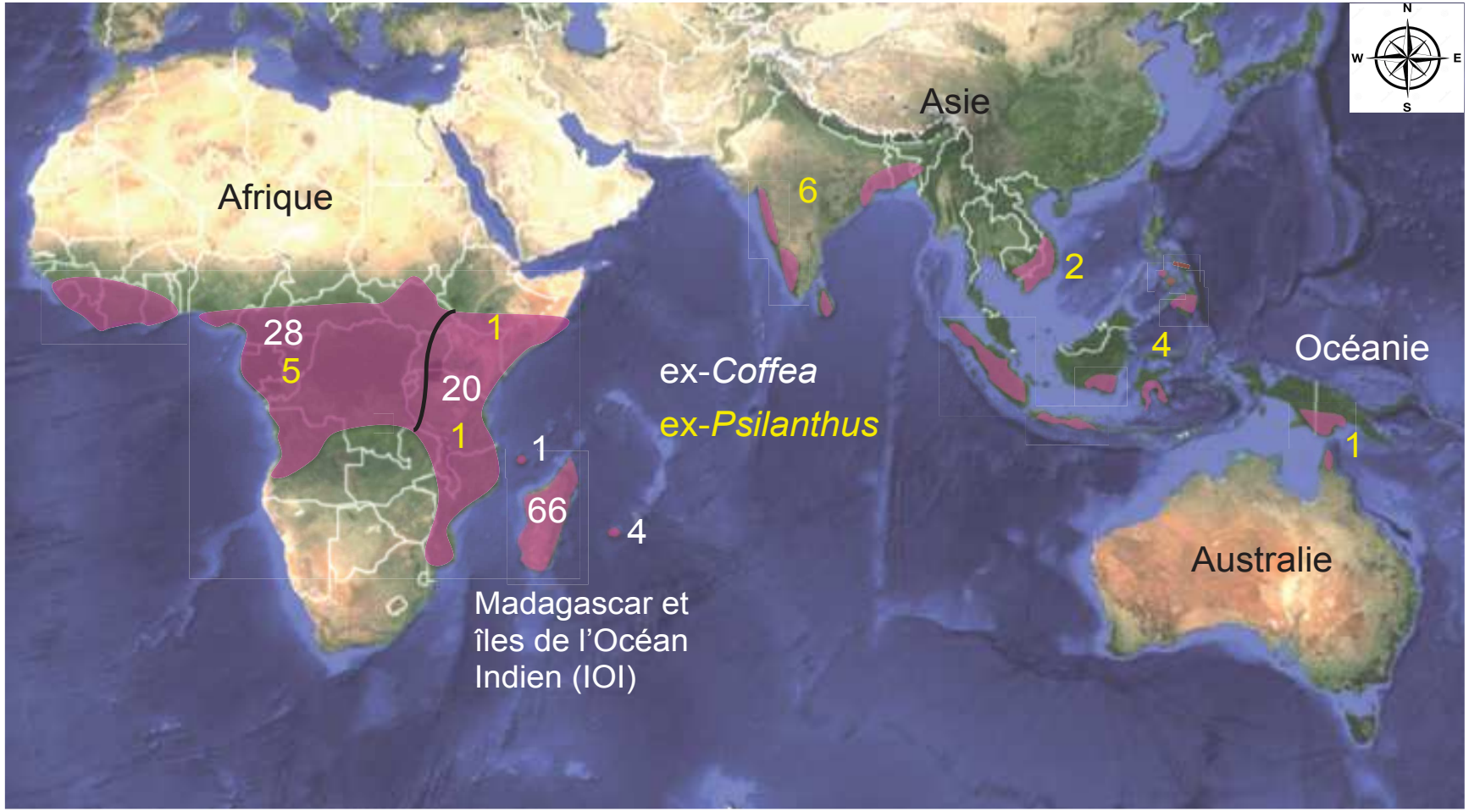


Figure 9 : Répartition géographique naturelle des espèces de caféiers. Les chiffres indiquent le nombre d'espèces présentes dans les régions colorées en rose foncé et la courbe noire symbolise la séparation entre l'Afrique de l'ouest et du centre et l'Afrique de l'est.

La divergence génétique des caféiers a été étudiée *via* les taux de succès de croisements interspécifiques contrôlés, l'utilisation de marqueurs moléculaires nucléaires et l'hybridation fluorescente *in situ* (FISH). S'agissant des hybridations interspécifiques contrôlées, les taux de réussite obtenus en Côte d'Ivoire étaient plus importants entre espèces d'une même région biogéographique (Afrique de l'ouest et du centre, ou Afrique de l'est), qu'entre espèces des deux régions (Louarn 1992). Ils étaient aussi d'autant plus importants que la différence de taille des deux génomes parentaux était petite. De plus, les taux de réussite entre espèces africaines et espèces malgaches étaient extrêmement bas (Charrier 1978). Les études utilisant des marqueurs moléculaires ont rarement concerné les espèces sauvages (De Kochko et al. 2010). Les études réalisées par Razafinarivo et al. (2012) et Andrianasolo et al. (2013) ont montré que les caféiers malgaches sont bien différenciés des caféiers africains. Enfin, les études par FISH montrent un nombre de locus et une localisation différents des ADN ribosomaux 45S et 5S entre les espèces africaines de l'ouest et du centre et celles de l'est (Hamon et al. 2009).

iii. Relations phylogénétiques des caféiers

Les travaux concernant la phylogénie du genre *Coffea* se sont longtemps heurtés à la faible divergence des séquences nucléaires et chloroplastiques utilisées conduisant à l'absence de résolution des arbres obtenus, et à l'incapacité d'établir clairement les relations entre les espèces. Ils montrent cependant l'existence de quelques clades géographiques, compatibles avec les études de diversité génétique et la classification de Chevalier (1942) (Cros et al. 1993; Maurin et al. 2007; Nowak et al. 2012). La première phylogénie entièrement résolue et soutenue, publiée très récemment (Hamon et al. 2017) est basée sur une approche de séquençage nouvelle génération. Les caféiers se distribuent dans deux clades majeurs : l'un, appelé « Xeno-Coffea », contenant tous les ex-*Psilanthus* et un seul ex-*Coffea* (*C. rhamnifolia*), l'autre, nommé « Eu-Coffea » rassemblant tous les autres *Coffea*. À l'intérieur de ces deux grands clades, des sous-clades se distinguent nettement et correspondent aux groupes géographiques précédemment déterminés. Les espèces sont rassemblées par grandes régions et sont bien différenciées, puisqu'aucune n'est dans une autre région que celle dont elle provient excepté *C. humblotiana*, rassemblée avec les espèces malgaches alors qu'elle est

originaire des Comores. On peut ainsi dégager dans les Xeno-Coffea : trois sous-clades, l'un formé de deux espèces sœurs *Coffea rhamnifolia* et *C. neoleroyi* (ex-*Psilanthus)*, les deux autres sous-clades représentant respectivement les ex-*Psilanthus* d'Afrique et les ex-*Psilanthus* d'Asie (sens large).

Dans les Eu-Coffea, excepté *C. charrieriana* en position basale, on observe deux sous-clades majeurs incluant pour l'un les *Coffea* d'Afrique et pour l'autre les *Coffea* des IOI. Le sous-clade africain montre une différenciation marquée entre espèces de basse altitude d'Afrique de l'est, espèces du centre (les deux seules diploïdes autofertiles du genre) et d'altitude de l'est et enfin les espèces de l'ouest et du centre. La Figure 10 montre cette phylogénie. Des représentations en photographies ou planches d'herbiers sont disponibles pour les espèces étudiées dans ce travail en Annexe 1.

iv. Caféiculture

Sur les 139 espèces de *Coffea*, *Coffea arabica* et *C. canephora* (produisant le café connu sous le nom de Robusta) sont les deux espèces de caféiers majoritairement cultivées pour le café-boisson. Aujourd'hui, la caféiculture fait vivre des millions de personnes dans les pays du Sud, exportateurs, pour une consommation croissante de café dans les pays du Nord. L'espèce *Coffea arabica* est majoritairement cultivée, représentant environ 70% de la production mondiale (http://ico.org). À la fin du XIXème et début du XXème siècles, l'espèce *C. liberica* var *dewevrei* (aussi appelée *C. excelsa*) était cultivée en Afrique et en Indonésie (Chevalier 1929). La culture de ce caféier très productif a été abandonnée à cause d'un champignon, *Fusarium xylarioides*, qui a décimé les cultures à partir de la fin des années 1930 (Guillemat 1946). Néanmoins, cette espèce présente d'autres caractéristiques intéressantes. Ainsi, Il existe une variété résistante à la rouille orangée *Hemileia vastatrix* (Berthaud 1986), portant le seul locus majeur de résistance appelé *Sh3. C. liberica* est donc une espèce majeure à étudier pour l'amélioration des caféiers puisque la rouille est responsable de plus de 30% de perte de production lors des crises en Amérique latine (conjonction de climat favorable et de perte de résistance des variétés).

Les menaces telles que les pathogènes (rouille – *Hemileia vastatrix*, Gotera – *Mycena citricolor*, Coffee berry disease (CBD) – *Colletotrichum kahawae*) et les changements climatiques en cours et à venir (liées au développement des pathogènes) pesant sur de nombreuses caféicultures, notamment au Brésil (premier exportateur de café, ico.org) et

Figure 10 : Phylogénie du genre *Coffea*. D'après Hamon et al. 2017. Les branches sont colorées selon les aires de répartition géographiques des espèces : bleu = ex-Psilanthus d'Afrique et d'Asie ; vert = Afrique de l'est et du centre-est ; rouge : Afrique de l'ouest et du centre ; violet : île Maurice, Madagascar et Comores.

en Colombie, font qu'il est nécessaire de mieux connaître les ressources génétiques des caféiers et d'identifier les sources de gènes d'intérêt utiles aux programmes de sélection. Les espèces sauvages représentent la plus grande partie de ces ressources génétiques potentiellement utiles. L'analyse précise des génomes permet de comprendre le fonctionnement de la plante, les interactions potentielles entre gènes et ET, ainsi que l'histoire d'un genre, d'une famille.

Génomique et LTR-RT chez les caféiers

Dans les dix dernières années, des développements significatifs ont été obtenus au niveau génomique pour les caféiers, avec la création de banques BAC (Bacterial Artificial Chromosome) et le séquençage de clones BAC dans des régions d'intérêt, ou encore le séquençage de banques d'expression (EST ou Expressed Sequence Tag) (pour revue : de Kochko et al. 2010).

L'analyse d'un des premiers clones BAC de *C. canephora* a permis l'identification de deux LTR-RT *Nana* et *Divo,* appartenant à la super-famille *Copia*. Ils ont pu être utilisés comme marqueurs moléculaires dans l'étude de caféiers sauvages (africains principalement). L'analyse du polymorphisme d'insertion a indiqué que *Nana* a accompagné la spéciation, alors que l'activation plus récente de *Divo* a accompagné la différenciation génétique au sein de *C. canephora* (Hamon et al. 2011) ; différenciation précédemment observée avec des marqueurs nucléaires RFLP et SSR (Gomez et al. 2009). Plus récemment encore, le polymorphisme d'insertion (REMAP) de quatre LTR-RT (un *Copia -Tork* et trois *Gypsy –Del* et *CRM*), homologues entre *C. canephora* et *C. millotii* de Madagascar, a été analysé pour mieux comprendre les liens phylogénétiques entre les espèces du complexe Millotii. Ils n'ont pas permis de résoudre l'histoire évolutive du complexe Millotii dans son intégralité, suggérant des insertions relativement récentes, après la spéciation. Cependant, les éléments *Gypsy* ont permis une séparation géographique entre les espèces du nord et celles du sud-est de Madagascar, ainsi qu'une différenciation des espèces dans chaque groupe géographique (Roncal et al. 2015).

L'analyse comparative des insertions de LTR-RT de séquences de clones BAC issus de *C. canephora* et de *C. arabica*, a démontré que des insertions de *Copia* spécifiques à *C. arabica* se seraient probablement produites juste après l'allo-tétraploïdisation (Yu et al.

2011). De plus, l'activité des LTR-RT aurait provoqué dans le génome de *C. arabica* de manière préférentielle l'augmentation de la taille du sous-génome *canephora*. Ces résultats sont congruents avec les observations effectuées dans différents génomes allopolyploïdes (Parisod et al. 2010).

L'analyse FISH d'un LTR-RT *Gypsy* a montré qu'il est présent dans sept espèces (d'Afrique de l'ouest, du centre et de l'est) de caféiers (Yuyama et al. 2012). Il serait donc conservé dans le sous-clade africain du clade Eu-Coffea et situé en particulier dans les régions hétérochromatiques. Des données de transcription et des analyses cytogénétiques ont permis d'identifier des fragments de LTR-RT retrouvés dans des transcrits de *C. canephora*, *C. arabica* et *C. racemosa* (espèce est-africaine), suggérant qu'ils participent à la diversité des protéines dans ces espèces (Lopes et al. 2008). Plus récemment, utilisant plus de transcrits, cette équipe a montré que la régulation de l'expression des ET et des gènes contenant des fragments d'ET était différente selon les conditions de culture (notamment la sécheresse) chez *C. arabica* (Lopes et al. 2013).

L'obtention de la séquence du génome de *C. canephora* (710 Mb), bien qu'incomplète (75% du génome séquencé et assemblé), a permis d'analyser sa composition en ET (Denoeud et al. 2014). Celui-ci en contient environ 50%, ce qui correspond aux proportions retrouvées pour les génomes de taille moyenne (par exemple le génome de la vigne (475 Mb) en contient environ 40% et celui de la tomate (950 Mb) en contient 60%). Quarante-deux pourcent de la séquence génomique correspond à des LTR-RT, qui constituent donc la part dominante de la composition du génome de *C. canephora*.

Les LTR-RT sont donc des outils important dans le genre *Coffea* pour mieux comprendre l'histoire évolutive des espèces. Cependant, l'étude exhaustive et plus fine des familles de LTR-RT est nécessaire pour sélectionner les éléments les plus informatifs. Cette sélection requiert l'analyse de séquences génomiques de haute qualité et pour un grand nombre d'espèces. Ainsi, une nouvelle lignée de rétrotransposons non-autonomes, les TR-GAG, constitués d'une région codante avec les seuls gènes *Gap* et *AP*, a été découverte chez *C. canephora* et est présente chez de nombreuses plantes (Chaparro et al. 2015). De nombreuses autres structures inédites restent encore à découvrir et à analyser dans les génomes des caféiers.

5. Projets internationaux G13 et ACGC

Deux projets internationaux sont impliqués dans ce travail, G13 et ACGC.

Génomes13 (G13) est basé sur l'analyse des espèces sauvages de caféiers potentiellement utiles à l'amélioration variétale en utilisant des données NGS (Next Generation Sequencing) et la génomique comparative. En effet, les espèces sauvages, adaptées à différents milieux et conditions de développement, représentent une source potentielle importante de diversité et un réservoir de gènes d'intérêt pour la caféiculture. Un jeu initial de 13 espèces/sous-espèces, puis 24 espèces, complété récement à 69, a été choisi pour sa représentativité en termes de clades phylogénétiques, de variations de taille de génomes et des adaptations environnementales. L'analyse des données doit permettre d'aborder les questions suivantes : la phylogénie chloroplastique, la composition en gènes spécifiques, l'origine des variations de taille du génome, la dynamique évolutive des éléments transposables et des éléments non codants et la diversité allélique globale. Ce consortium international regroupe 13 instituts ou Universités dans huits pays pour 27 chercheurs.

Le second projet, porté par le Consortium pour le séquençage du génome de *C. arabica* (The Arabica Coffee Genome Consortium 2014), vise à produire une séquence du génome de *C. arabica* de très haute qualité avec la technologie des lectures longues (PasBio). L'accession Et39, *C. arabica* sauvage originaire d'Éthiopie, utilisée dans ce projet est un dihaploïde naturel, de telle sorte que l'hétérozygotie observée concerne les deux sous-génomes parentaux. Dans cet objectif, *C. canephora* et *C. eugenioides* sont eux aussi séquencés avec la même technologie afin de faciliter l'identification des deux sous-génomes parentaux. Le re-séquençage additionnel de 36 variétés sauvages et cultivées de *C. arabica* et de plusieurs accessions des deux espèces diploïdes parentales devrait permettre une meilleure compréhension de la diversification récente de cette espèce, et fournir des éléments quant à l'origine géographique des populations à l'origine de *C. arabica* (The Arabica Coffee Genome Consortium 2014). Ce consortium international regroupe 25 instituts ou Universités dans 13 pays pour 60 chercheurs.

6. Présentation du sujet de recherche

Le présent travail se situe dans le cadre des deux projets internationaux, ACGC et G13 décrits au paragraphe précédent et ayant pour but d'étudier : d'une part la structure et l'évolution du caféier cultivé *C. arabica* et ses deux ancêtres diploïdes : *C. canephora* et *C. eugenioides*, d'autre part cherchant à analyser l'évolution des espèces sauvages diploïdes. L'objectif global de mes travaux est de contribuer à la connaissance de la composition et de l'évolution des LTR-RT chez les caféiers et d'utiliser cette connaissance sur les éléments transposables pour mieux comprendre l'évolution des espèces du genre *Coffea*.

L'objectif principal fixé pour mon travail de thèse était la caractérisation fine de deux familles de LTR-RT chez les caféiers. Dans ce but, plusieurs approches ont été utilisées au cours de ce travail.

A) Dans le troisième chapitre, une analyse de la composition en ET de 11 génomes d'espèces sauvages séquencés par la technologie de pyroséquençage (454) a été conduite pour identifier des familles cibles présentant potentiellement des variations du nombre de leurs copies. Ce travail a été publié dans la revue *Molecular Genetics and Genomics* en 2016. Cette analyse a été complétée par l'étude détaillée des résultats de l'annotation des LTR-RT du génome de *C. canephora* (Denoeud et al. 2014).

B) Dans le quatrième chapitre, nous avons abordé l'analyse détaillée de la famille *Divo,* déjà identifiée et utilisée comme marqueur moléculaire en 2011. Nous présentons la caractérisation de *Divo* et de la lignée *Bianca* peu étudiée à ce jour, à laquelle cet élément appartient. Profitant de la disponibilité des séquences génomiques de *C. arabica* et de ses deux ancêtres diploïdes, *C. canephora* et *C. eugenioides* (consortium ACGC), l'activité de *Divo* et sa contribution à la composition et à l'évolution des génomes diploïdes et allotetraploïde ont été étudiées. Ce travail a été publié dans la revue *Molecular Genetics and Genomics* en 2017.

C) Le cinquième chapitre traite de la caractérisation de la lignée *SIRE* dans les génomes de *C. arabica, C. canephora* et *C. eugenioides.* Ce travail est présenté sous la forme d'un article en anglais qui doit être très prochainement soumis.

D) Dans le sixième chapitre, nous abordons l'analyse de la distribution de trois familles d'éléments *SIRE* dans 24 génomes d'espèces sauvages séquencés par la technologie Illumina et par une approche phylogénétique. Cette analyse tente d'approfondir les observations et les hypothèses émises lors du chapitre 3 de ce présent document. Ce travail est présenté sous la forme d'un article en anglais très prochainement soumis.

E) La discussion et la conclusion finales tentent de synthétiser l'ensemble des résultats et leur contribution à la connaissance des ET et l'évolution des espèces du genre *Coffea*.

Plusieurs séquences génomiques sont disponibles pour l'étude des caféiers :

- des génomes de 11 espèces de caféiers cultivées (*C. arabica* et *C. canephora* – trois accessions) et sauvages séquencées avec la technologie 454
- le génome de *C. canephora* (haploïde double HD-200), ainsi que l'identification de novo et l'annotation des ET disponible et visualisable sur le site du génome de *C. canephora* (http://coffee-genome.org).
- Trois génomes séquencés avec la technologie PacBio (Rhoads and Au 2015) : *C. arabica* (accession ET39, sauvage di-haploïde d'Ethiopie) et ses progéniteurs *C. canephora* (HD-200) et *C. eugenioides* (BU-A d'Ouganda). Ils sont en cours d'annotation grâce au consortium ACGC (The Arabica Coffee Genome Consortium 2014).
- 24 génomes séquencés en Illumina de 24 espèces/sous-espèces de caféiers diploïdes représentatives des groupes botaniques et géographiques, ainsi que des variations de taille de génomes observées dans le genre *Coffea* (Projet G13, communication orale PAG 2015).

Le Tableau 1 présente les espèces pour lesquelles des données de séquençage sont disponibles et dans quelles études elles ont été utilisées.

Tableau 1 : Espèces disponibles, type de séquençage et études dans lesquelles elles ont été utilisées.

Espèce	Pays	Type de séquençage	Article 454	Article *Divo*	Article *SIRE* 1	Article *SIRE* 2
C. canephora	Rép. D. du Congo	454, Illumina, PacBio	x	x	x	x
C. canephora	Guinée	454	x			
C. canephora	Ouganda	454, Illumina	x		x	x
C. arabica	cultivé	454	x			
C. arabica	Ethiopie	454, PacBio	x	x	x	
C. eugenioides	Kenya	454, Illumina	x		x	x
C. eugenioides	Ouganda	454, PacBio	x	x	x	
C. neoleroyi (ex-*Psilanthus neoleroyi*)	sud du Soudan	Illumina				x
C. ebracteolata (ex-*P. ebractelatus*)	Côte d'Ivoire	Illumina				x
C. mannii (ex-*P. mannii*)	Cameroun	Illumina				x
C. melanocarpa (ex-*P. melanocarpus*)	Angola	Illumina				x
C. merguensis (ex-*P. merguensis*)	Thaïlande	Illumina				x
C. horsefieldiana (ex-*P. horsefieldianus*)	Indonésie	454, Illumina	x			x
C. benghalensis var. *benghalensis* (ex-*P. benghalensis*)	Inde	Illumina				x
C. benghalensis var. *bababudanii* (ex-*P. bababudanii*)	Inde	Illumina				x
C. brassii (ex-*P. brassii*)	Australie	Illumina				x
C. rhamnifolia	Mozambique	Illumina				x
C. charrieriana	Cameroun	454, Illumina	x			x
C. pseudozanguebariae	Kenya	454, Illumina	x			x
C. racemosa	Mozambique	454, Illumina	x			x
C. mufindiensis	Tanzanie	Illumina				x
C. heterocalyx	Cameroun	454	x			x
C. stenophylla	Côte d'Ivoire	Illumina				x
C. liberica	Côte d'Ivoire	Illumina				x
C. humilis	Côte d'Ivoire	Illumina				x
C. kapakata	Angola	Illumina				x
C. macrocarpa	Île Maurice	Illumina				x
C. humblotiana	Comores	454, Illumina	x			x
C. tetragona	Madagascar	454, Illumina	x			x
C. dolichophylla	Madagascar	454, Illumina	x			x

Chapitre 3 – Apports du séquençage partiel à l'étude des éléments transposables dans le genre *Coffea*

Article reçu le 19 mai 2016, accepté le 25 juillet 2016 et publié en ligne le 28 juillet 2016 dans le journal *Molecular Genetics and Genomics*. DOI : 10.1007/s00438-016-1235-7

1. Contexte

Chez les caféiers, il a été observé un gradient d'accroissement de la taille des génomes nucléaires en Afrique, de l'est vers l'ouest (Noirot et al. 2003) et à Madagascar du Nord vers le Sud (Razafinarivo et al. 2012) (Figure 11). La variation des tailles des génomes pourrait imposer des contraintes sur le développement des plantes, leur phénologie et serait associée au signal phylogénétique (Petrov and Wendel 2006). Afin d'étudier la composition et l'évolution de la taille des génomes nucléaires chez les *Coffea*, une approche par séquençage a été entreprise. La technologie de pyroséquençage était encore en 2013 un séquençage nouvelle génération très utilisé en génomique car permettant d'obtenir des lectures relativement longues (de 400 à 800 pb ; produite par ROCHE - Rothberg and Leamon 2008) avec un coût très inférieur à la technique traditionelle Sanger. Plusieurs articles dans la littérature avaient utilisé l'approche de séquençage partiel en 454 pour obtenir des informations sur la composition des génomes et l'évolution de leur taille, ainsi qu'identifier de nouvelles insertions d'ET chez le pois, le blé, la banane, la vigne et les chauves-souris (Macas and Neumann 2007; Wicker and Keller 2007; Hribová et al. 2010; Pagan et al. 2010). Il faut noter que la technique 454 a été abandonnée en 2016, dépassée en qualité, en quantité de lectures produites et en coût par Illumina et dépassée par la taille des lectures produites par PacBio.

Le séquençage du génome de *C. canephora* a permis l'identification des ET présents dans ce génome (Denoeud et al. 2014) et à la construction d'un premier répertoire à l'aide d'une approche combinée *de novo* (REPET), structurale (LTR_STRUC) et par similarité. Ce répertoire pouvait donc être utilisé dans le cadre d'une analyse sur la

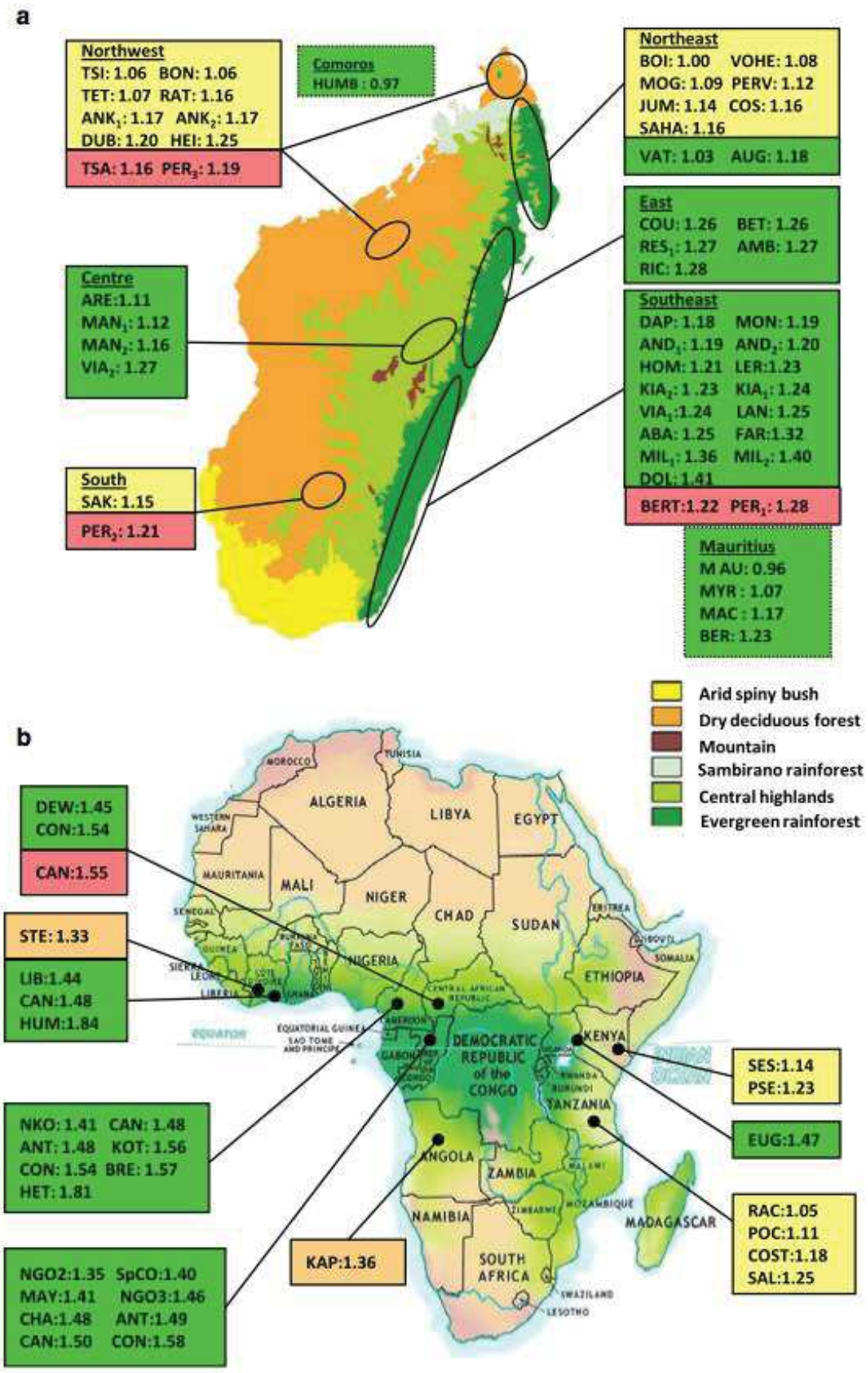


Figure 11 : Distribution du contenu en ADN (2C) des génomes des *Coffea* à Madagascar (a) et en Afrique (b). D'après Razafinarivo et al. 2012.

composition des séquences des génomes d'espèces sauvages.

Un séquençage partiel exploratoire de 15 génomes de 11 espèces de *Coffea* (Tableau 1) acquises avec la collaboration de Nestlé R&D, Tours a été entrepris. Il permet une première estimation de la composition et de l'abondance en ET et en LTR-RT identifiés chez *C. canephora* comme référence dans les autres génomes.

2. Implication personnelle

Pour cet article, l'essentiel du travail a été réalisé par Romain Guyot, Thibaud Darré (stage de Master 2) et Perla Hamon. J'ai utilisé le script perl MiSA pour détecter les microsatellites dans les séquences 454. J'ai recherché d'éventuelles différences en nombre et fréquence de microsatellites entre les différents génomes étudiés. J'ai participé à la relecture de l'article et à la réalisation des Figures Supplémentaires 6 et 7.

Mol Genet Genomics
DOI 10.1007/s00438-016-1235-7

ORIGINAL ARTICLE

Partial sequencing reveals the transposable element composition of *Coffea* genomes and provides evidence for distinct evolutionary stories

Romain Guyot[2] · Thibaud Darré[1] · Mathilde Dupeyron[1] · Alexandre de Kochko[1] · Serge Hamon[1] · Emmanuel Couturon[1] · Dominique Crouzillat[3] · Michel Rigoreau[3] · Jean-Jacques Rakotomalala[4] · Nathalie E. Raharimalala[4] · Sélastique Doffou Akaffou[5] · Perla Hamon[1]

Received: 19 May 2016 / Accepted: 25 July 2016
© Springer-Verlag Berlin Heidelberg 2016

Abstract The *Coffea* genus, 124 described species, has a natural distribution spreading from inter-tropical Africa, to Western Indian Ocean Islands, India, Asia and up to Australasia. Two cultivated species, *C. arabica* and *C. canephora,* are intensively studied while, the breeding potential and the genome composition of all the wild species remained poorly uncharacterized. Here, we report the characterization and comparison of the highly repeated transposable elements content of 11 *Coffea* species representatives of the natural biogeographic distribution. A total of 994 Mb from 454 reads were produced with a genome coverage ranging between 3.2 and 15.7 %. The analyses showed that highly repeated transposable elements, mainly LTR retrotransposons (LTR-RT), represent between 32 and 53 % of *Coffea* genomes depending on their biogeographic location and genome size. Species from West and Central Africa (Eucoffea) contained the highest LTR-RT content but with no strong variation relative to their genome size. At the opposite, for the insular species (Mascarocoffea), a strong variation of LTR-RT was observed suggesting differential dynamics of these elements in this group. Two LTR-RT lineages, SIRE and Del were clearly differentially accumulated between African and insular species, suggesting these lineages were associated to the genome divergence of *Coffea* species in Africa. Altogether, the information obtained in this study improves our knowledge and brings new data on the composition, the evolution and the divergence of wild *Coffea* genomes.

Keywords LTR retrotransposons · Partial genome sequencing · *Coffea* · Genome size · Geographic divergence

Communicated by S. Hohmann.

Electronic supplementary material The online version of this article (doi:10.1007/s00438-016-1235-7) contains supplementary material, which is available to authorized users.

✉ Romain Guyot
romain.guyot@ird.fr

[1] IRD UMR DIADE, EvoGeC, BP 64501, 34394 Montpellier Cedex 5, France

[2] IRD UMR IPME, CoffeeAdapt, BP 64501, 34394 Montpellier Cedex 5, France

[3] Nestlé R&D Tours, 101 AV. G. Eiffel, Notre Dame d'Oe´, BP 49716, 37097 Tours Cedex 2, France

[4] FOFIFA, Ambatobe, Madagascar

[5] University Jean Lorougnon Guédé, Daloa, Ivory Coast

Introduction

Repetitive sequences are major components of plant genomes. Transposable elements (TEs), constituting the mobile part of the genomes, are divided into two main classes (Class I and Class II) according to their mode of transposition. They are hierarchically classified into orders, super-families, lineages, families and individuals within each class (Wicker and Keller 2007; Wicker et al. 2007). Class I elements known as retrotransposons, transpose via an RNA intermediate without movement of the master copy. This 'copy-and-paste' mechanism can theoretically lead to a rapid increase of the frequency of the original copy. Class II, or transposons move following a 'cut-and-paste' mechanism or through DNA replication, resulting to low or moderate new inserted copies. Plant retrotransposons include two major orders: Long Tandem Repeat retrotransposons (LTR-RT) and non-LTR retrotransposons. The first ones include two super-families: *Copia* and *Gypsy* that differ mainly in their coding region organization and

Published online: 28 July 2016

are composed of ancient conserved evolutionary lineages in plants (Wicker and Keller 2007). The second ones includes long and short interspersed nuclear element, LINE and SINE, respectively (Kumar and Bennetzen 1999; Wicker et al. 2007).

During the last 10 years, the accumulation of genomic sequencing data (Michael and Jackson 2013) indicated that TEs are the major component of plant genomes and that their accumulation could be correlated with the genome sizes (Ibarra-Laclette et al. 2013; Kumar and Bennetzen 1999; Lisch 2013). LTR-RTs are the most redundant elements and in extreme cases, they can represent up to 80 % of plant genome sequences, suggesting that their propagation mechanisms are directly responsible of the genome size increase (SanMiguel et al. 1998; Schulman et al. 2004; Bennetzen et al. 2005; Hawkins et al. 2006; Dvořák 2009). Sometimes the propagation mechanisms induce a rapid accumulation, called a "burst", of a few number of LTR-RT families as demonstrated in the wild rice *Oryza australiensis* (Piegu et al. 2006). At the opposite, in maize, the accumulation of LTR-RT families was probably gradual, but leading to a considerable genome size increase when compared to the sorghum or the rice genome. However, a correlation between genome size variations and LTR-RT copy numbers was not established for the *Zea* genus (Meyers et al. 2001), suggesting that the proliferation mechanism of a few LTR-RT families per se cannot explain all genome size variations in plants. The host genome controls the level of transposition of LTR-RTs through epigenetic mechanisms (Bucher et al. 2012; Ito 2013; Ito and Kakutani 2014) This control might be reduced under abiotic stresses (Todorovska 2007; Alzohairy et al. 2014; Kinoshita and Seki 2014), leading to an increase of transposition and suggesting that LRT-RT play a role in the genome adaptation facing environmental changes (Casacuberta and Gonzalez 2013). However, so far no correlation was established between plant genome size and their habitat or phenotypic and life traits (Eilam et al. 2007; Knight and Beaulieu 2008; Slovak et al. 2009; Dušková et al. 2010).

Next generation sequencing technologies provided powerful tools to identify and characterize the repetitive fraction of genomes even in large genomes such as wheat, barley or pea (Macas et al. 2007; Wicker et al. 2009). For these authors, an important advantage of the NGS sequencing lies in the limited bias obtained for the production of the sequences. Low-depth sequencing was effective in identifying the most highly repeated sequences and in estimating their copy numbers in the pea genome (Swaminathan et al. 2007), banana genome (Hribova et al. 2010) and in vesper bats (Pagan et al. 2012) and to study genome evolution at a genus or a family scale (Nystedt et al. 2013). It also allows identifying TEs insertion polymorphism accompanying clonal variation in grape (Carrier et al. 2012). The uneven distribution of TEs between wheat and barley (Wicker et al. 2009), the genome size variation in the allotetraploid species *Nicotiana tabaccum*, (Renny-Byfield et al. 2011) as well the composition and abundance of highly repeated TEs in ten Triticeae taxa (Middleton et al. 2013) were also studied via a 454 pyrosequencing genomic survey.

Ranked fourth among angiosperms, the young Rubiaceae family [90.4 My divergence time, (Bremer and Eriksson 2009)] comprises ca. 600 genera and ca. 13,600 species. This family includes herbs, shrubs and trees growing naturally in overly diverse habitats (from desert to tropical sempervirente forests via temperate areas), altitudes (from sea level to over 2500 m) and soils. In this plant family, diploids are the most common and share the same basic chromosome number [$x = 11$ (Kiehn 1995)]. The *Coffea* genus, member of Rubiaceae, is the most known genus due to its major socio-economic importance worldwide (producers in Southern countries and consumers in Northern countries). Accounting for 124 described species, all diploids with $2n = 2x = 22$ but *C. arabica* (allotetraploid), the natural distribution in inter-tropical forests of Africa and of Western Indian Ocean Islands was recently extended to India, Asia and Australasia (Davis et al. 2011). The recent sequencing of *C. canephora* (also called Robusta) genome showed that no whole genome duplication has occurred after the Asterid clade divergence, some 110 My ago (Denoeud et al. 2014). Moreover, comparative mapping between two divergent African genomes: *C. canephora* and *C. pseudozanguebariae* did not reveal any major chromosomal rearrangements (unpublished data). Despite structural conservations, a notable variation of genome sizes is observed among *Coffea* species. This variation ranges from 469 to 900 Mb with a general pattern of increasing genome sizes from East to West in Africa (Noirot et al. 2003) and from North to South-East in Madagascar (Razafinarivo et al. 2012), suggesting a gradual accumulation of nuclear DNA, under speciation and adaptive processes of the species. Recently, the *C. canephora* genome sequencing allowed the computational identification of TEs (Denoeud et al. 2014). They represent more than half of the available genome sequence, and among them, LTR-RTs are the most frequent order of elements (42 % of the genome). However, outside the *C. canephora* genome, no wide survey of TE composition has been conducted in the *Coffea* genus.

Here, we used a 454 sequencing survey of one tetraploid and ten diploid species representative of the botanical and geographical diversity of the genus *Coffea* to study and compare the composition and abundance of highly repeated transposable elements in their genomes. Using a genome coverage ranging from 3.2 to 15.7 %, the analysis of LTR-RT composition and dynamics shows a clear difference between African and insular *Coffea* species, suggesting an ancient divergence. Contrary to previous hypotheses and

Table 1 454 sequencing data for 11 *Coffea* Species

Species	Accession	Country of origin	Group	No. of 454 reads	Total (bp)	Mean size (bp)	Coverage (%)	Genome size (Mb)
C. arabica	ET39	Ethiopia	EUC	93,194	41,643,904	446	3.2	1300
C. arabica	ET39	Ethiopia	EUC	112,615	51,892,121	460	3.9	1300
C. arabica	ET39	Ethiopia	EUC	140,976	61,729,478	437	4,7	1300
C. canephora	IF410	Ivory Coast	EUC	186,138	85,292,671	458	12.2	700
C. canephora	DH200-94	D. Republic Congo	EUC	98,017	43,037,451	439	6.1	700
C. canephora	BUD15	Uganda	EUC	140,120	64,290,611	458	9.2	700
C. charrieriana	OA22	Cameroon	EUC[a]	136,518	57,405,992	420	7.9	723
C. eugenioides	OUG14	Uganda	EUC	186,449	85,961,094	461	13.3	645
C. eugenioides	DA56	Kenya	EUC	91,834	39,993,235	435	6.2	645
C. heterocalyx	JC65	Cameroon	EUC	123,119	45,633,337	370	5.2	863
C. pseudozangue-bariae	8107	Kenya	MOZ	215,117	91,733,301	426	15.5	593
C. racemosa	IA56	Mozambique	MOZ	173,803	79,199,218	455	15.7	506
C. tetragona	A.252	Madagascar	MAS	147,430	68,881,825	467	13.4	513
C. dolichophylla	A.206	Madagascar	MAS	147,758	70,632,674	478	10.4	682
C. humblotiana	A.230	Comoros	MAS	141,834	62,465,685	440	10.4	469
C. horsfieldiana	HOR	Indonesia	PSI	104,605	44,610,588	426	7.5	593

Botanical groups (Group) are those from Chevalier (1942) with *EUC* Eucoffea (species from West and Central Africa), *MOZ* Mozambicoffea (East Africa), *MAS* Mascarocoffea (species from Western Indian Ocean Islands), *PSI* Paracoffea

[a] The Eucoffea classification for *C. charrieriana* was not established by Chevalier since the species was recently described by Stoffelen et al. (2008). Therefore, its classification was assumed according to its geographical origin. Genome sizes are from Noirot et al. (2003) and Razafinarivo et al. (2012). The genome coverage is given in %

generally admitted idea, our results suggest that the *Coffea* species from Western Indian Ocean Islands and from Asia have diverged independently from their continental counterparts. Furthermore, no strong activation of LTR-RTs was obvious in any species, whatever their genome size, suggesting that other molecular mechanisms or general but limited variation in TE copy numbers are associated to genome size increases in the *Coffea* genus.

Materials and methods

DNA isolation and 454 sequencing

Leaves from Madagascan and Comorian species were obtained from the Kianjavato Coffee Research Station (KCRS) in Madagascar. The African species were sampled from the *Coffea* collection maintained at IRD (Montpellier, France) or Nestlé R&D (Tours, France) greenhouses. The studied species belong to Chevalier's (Chevalier 1942) botanical sections, i.e., Eucoffea (West and Central African species), Mozambicoffea (East African species), Mascarocoffea (species from the Western Indian Ocean Islands) and Paracoffea (species belonging to *Psilanthus* subgenus *Afrocoffea*). In total, we used seven Eucoffea, two Mozambicoffea, three Mascarocoffea and one Paracoffea accessions. Information on the accessions used, their origin and other used data are given in Table 1.

DNA was isolated from fresh or dried leaves using Qiagen DNeasy Plant Mini extraction kits following the manufacturer protocol. Quantity and quality of DNA was measured using a Nanodrop (ND-1000). The libraries construction and Next Generation sequencing were performed at Nestlé R&D laboratory (Tours, France) according to the Roche/454 Life Sciences Sequencing Method using one Roche 454 GS Junior plate per accession. Data were submitted to GenBank, BioProject PRJNA242989. General information on 454-pyrosequencing is available in Table 1.

Sequences analyses

Quality of 454 reads was checked using FASTQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and cleaned using Prinseq v0.20.4 (Schmieder and Edwards 2011).

BLASTX searches (minimum e-value $10e^{-4}$) were first carried out on 454-reads against the RepBase amino acid sequence dataset (Kohany et al. 2006; Jurka et al. 2005)—http://www.girinst.org/repbase/). BLASTN were carried out against *Coffea* coding sequence (CDS, http://coffee-genome.org), the *C. arabica* chloroplast genome (EF044213) and rRNA sequence (X52320 and AY083685)

with a minimum e-value of $10e^{-6}$. BLASTN analyses were also performed against the *C. canephora* repeat database built with REPET (https://urgi.versailles.inra.fr/Tools/REPET) with an e-value of $10e^{-20}$. The goal was to identify the major TE classes, super-families and lineages reported until today at different scales (amino acid and nucleotide) and to obtain their proportion in the investigated genomes. Given the importance of the Class I/LTR-RT in all genomes, BLASTN similarity searches were conducted between 454 reads and a dataset of LTR retrotransposons consensus sequences from *C. canephora* classified according to their Reverse Transcriptase (RT) amino acid similarities (available at the Gypsy Database 2.0). 454 sequences showing similarities with RT domains were classified by phylogenetic analyses. Identified RT domains from 454 datasets were extracted from the nucleotide sequences and translated into amino acids. Amino acid sequences (with a minimum of 150 residues) were aligned (ClustalW) to construct a bootstrapped neighbor-joining tree, edited with FigTree (http://tree.bio.ed.ac.uk/software/figtree/).

Detailed annotation of the SIRE lineage (*Copia*) was performed using LTRFinder (http://tlife.fudan.edu.cn/ltrfinder/, (Xu and Wang 2007). LTR domain sequences were aligned with MUSCLE to build a consensus 100 bootstraps neighbor-joining phylogenetic tree with ClustalW. Complete SIRE elements were annotated with Artemis (Rutherford et al. 2000) and used as references. Structural incongruities (InDels and rearrangements) were searched using graphic alignments (dot-plot, (Sonnhammer and Durbin 1995).

The copy number of SIRE in 454 dataset was estimated as described in (Chaparro et al. 2015) and (Dias et al. 2015). BLASTN searches were carried out with full-length SIRE elements found in the *C. canephora* genome. Reads with more than 90 % of nucleotide identity with the reference sequence over a minimum 90 % of the read lengths were considered as potential fragments of the element. Cumulative lengths of aligned reads were used to extrapolate the contribution of the element to each genome size investigated. For each element family, the potential number of full-length copies is estimated by the division of the estimated size of total members of the element in the genome by the reference sequence length.

De novo detection of repeated sequences

De novo detection of repeated sequences was carried out using RepeatScout (http://bix.ucsd.edu/repeatscout/ (Price et al. 2005)) on 454 sequences for each species. The libraries of repeated sequences were used to mask each 454 dataset using RepeatMasker (http://www.repeatmasker.org). Repetitions were then filtered out according to their minimum redundancy in 454 dataset as follow: 20, 100, 500 and 1000 repetitions.

Searches for microsatellites

Microsatellites were detected on 454 sequences using the MicroSAtellite identification tools (http://pgrc.ipk-gatersleben.de/misa/). The unit size of repetition ranged from 1 to 20 and the number repeated units ranged from 1 to 10.

PCR amplification on *Coffea* DNA

Primers were designed on three full-length SIRE annotated in this analysis (called 36-863, 3-942 and 6-1571) on ENV and LTR domains using Primer3 (http://primer3.ut.ee) (Supplemental data 1A). PCR amplifications were performed in a final volume of 20 μL using the GoTaq DNA polymerase from Promega, according to the manufacturer recommendations: 0.5 ml of dNTP (10 nM), 1 ml of each primer (10 mM), 0.2 U of Taq polymerase (GoTaq, Promega) and 20 ng of DNA matrix. We used the following PCR amplification cycle: 98 °C 2 min.; three steps (98 °C 30 s, 55 °C 30 s, 72 °C 30 s) repeated 35 times followed by a final elongation step (72 °C 5 min). The DNA samples, representative of the biogeographic *Coffea* groups, (Supplemental data 1B) are those used in (Razafinarivo et al. 2013).

Results

454 sequencing in *Coffea*: run reproducibility and characterization of genomes composition

The 454 junior runs were produced for 10 *Coffea* diploid and one tetraploid species. Three independent runs for the same accession (ET39) of the tetraploid species, *C. arabica*, were carried out to check the reproducibility of the runs. In addition, for two diploid species, *C. canephora* and *C. eugenioides*, three (BUD15, HD200 and IF410) and two (DA56 and OUG14) accessions were, respectively, sequenced. The 454 sequencing produced a genome coverage ranging from 3.2 to 4.7 % for *C. arabica* and from 5.2 to 15.7 % for all the diploid species (Table 1). In total, more than 2.2 millions reads, accounting for 994 Mb, were produced and analyzed in this study. The three *C. arabica* replicates gave similar results showing the good reproducibility of the sequencing and enabling to have confidence in the results presented here.

Using BLASTN (CDS, chloroplast genome, rDNA) and BLASTX (transposable elements) we found that protein-coding genes represented between 11 % (*C. heterocalyx*) and 18 % (*C. canephora* acc. DH200-94) of the obtained

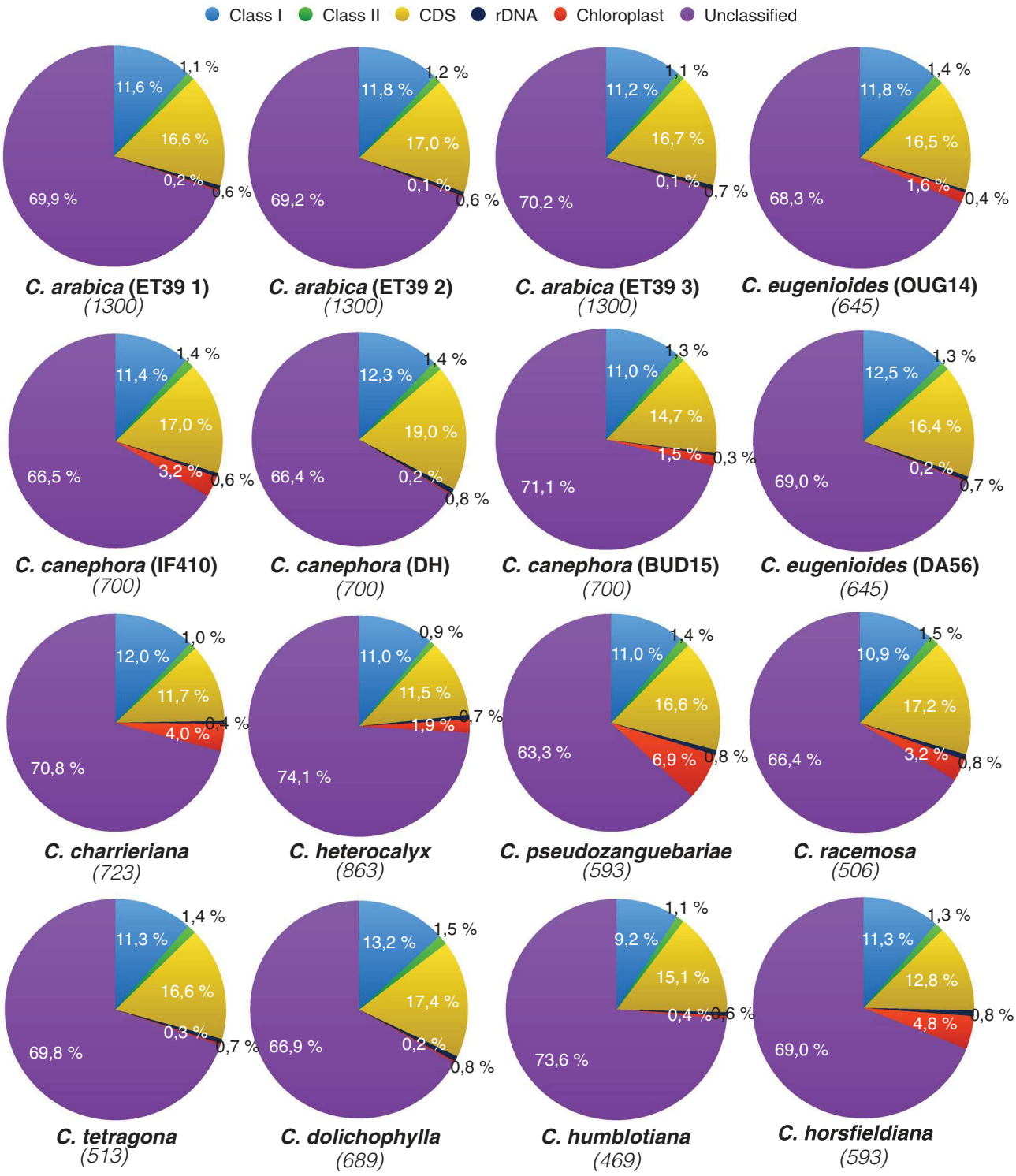


Fig. 1 Composition of 454 reads for 11 *Coffea* species and 14 accessions. *Class I* and *Class II* are known transposable element coding regions, *CDS* cellular coding regions, *rDNAs* ribosomal DNA genes. Name and accession of species were indicated with their respective genome size indicated into *brackets* (in Mb)

data (Fig. 1). A similar percentage to that of *C. canephora* was found for the three *C. arabica* replicates (17 %). However, the proportion of identified chloroplast sequences between species varies between 0.14 % (*C. arabica*) and 7 % (*C. pseudozanguebariae*). Five species showed a percentage of chloroplast sequences larger than 2 % (Fig. 1).

Chloroplast DNA presence may be attributed to the fact that total DNA was extracted for the sequencing and not just the nuclear fraction as in (Carrier et al. 2011) or to different amount of chloroplast DNA inserted into the nuclear genomes according to the studied taxa, such insertions have been observed in the sequenced *C. canephora* genome (Denoeud et al. 2014). Recognizable coding sequences from transposable elements represented a significant proportion ranging from 10 % for *C. humblotiana*, the smallest genome [469 Mb, (Razafinarivo et al. 2012)] to 14 % for *C. dolichophylla*, an average size genome (689 Mb). Interestingly, the genome of *C. heterocalyx* [the biggest one with 863 Mb, (Noirot et al. 2003)] was containing 12 % of transposable element coding genes.

For *C. canephora*, a similar TE coding sequences proportion (Class I and Class II) was found for the three accessions analyzed (BUD15, IF410 and DH200-94) originating from three different geographical areas (respectively, 12.3, 12.8 and 13.7 %). For all the species, most of the identified coding sequences of transposable elements felled into the Class I, as found for the *C. canephora* genome sequence (Denoeud et al. 2014).

To further investigate the composition of repeated sequences in *Coffea* species, we used as reference the *C. canephora* database of consensus transposable elements that was constructed de novo and annotated using the REPET programs. The *C. canephora* database is composed of 4051 consensus sequences for which 1536 and 2023 belonged to the LTR retrotransposons and non-autonomous LTR retrotransposons, respectively. Using this dataset, the proportion of LTR retrotransposons in the 454 reads reached 32 % for *C. humblotiana* and 53 % for *C. heterocalyx* (Supplemental data 2). Interestingly, the amount of 454 reads similar to *C. canephora* LTR retrotransposon consensus sequences was very similar for Eucoffea species whatever their genome size (*C. arabica:* 50–51 %, *C. eugenioides:* 48–50 %, *C. canephora:* 49–52 %, *C. charrieriana*: 48 % *and C. heterocalyx:* 53 %), while a clear lower amount was observed for the Mozambicoffea species (*C. pseudozanguebariae:* 37 %, *C. racemosa:* 39 %), for Mascarocoffea species (*C. tetragona*: 36 %, *C. dolichophylla*: 40 %, and *C. humblotiana*: 32 %) and for Asian Paracoffea (*C. horsfieldiana*: 34 %). These variations between Eucoffea and the three other botanical groups (Mozambicoffea, Mascarocoffea and Paracoffea), appeared independent from the genomes size, at the exception of *C. humblotiana* that showed both the smallest genome (469 Mb) and the lowest percentage of 454 reads containing sequences similar to *C. canephora* LTR retrotransposons (32 %). Such variation could be attributed to the nucleotide divergence of LTR retrotransposons between Eucoffea and the other botanical groups since the nucleotide database of LTR retrotransposons used as reference was established from *C. canephora* (Eucoffea). Altogether our data suggest a noticeable variation of the quantitative LTR-RT content in *Coffea* species genomes.

Abundance of LTR-retrotransposon lineages and their contribution to genome size

To further investigate the quantitative variation of LTR-retrotransposon content, we first classified the REPET consensus sequences into *Copia* and *Gypsy* super-families and, thus, into lineages (*Bianca*, *Oryco*, *Retrofit*, *Sire*, *Tork* for *Copia* and *Athila*, *CRM*, *Del*, *Galadriel*, *Reina* and *TAT* for *Gypsy* (Llorens et al. 2009), according to their similarities to reverse transcriptase (RT) reference domains. In total, LTR-retrotransposon consensus sequences were assigned to 877 families containing RT domains, for which 352 and 525 belong to *Copia* and *Gypsy*, respectively. These 877 families belong to all the different LTR-retrotransposon lineages previously discovered in other plant genomes. Using this dataset, all the *Coffea* species analyzed were found to contain a *Gypsy*/*Copia* ratio ranging from 2.6 to 4.6, suggesting that *Gypsy* represented the most abundant LTR-retrotransposon super-family in *Coffea* species, as previously found in *C. canephora* (Denoeud et al. 2014; Dereeper et al. 2013). The overall proportion of *Copia* and *Gypsy* varied greatly according to Chevalier's botanical classification and increased from Eucoffea to Mascarocoffea (Supplemental data 3). These variations were not noticeable when the 454 reads were translated (using BLASTX analysis against RepBase). Interestingly the *Gypsy*/*Copia* ratio was clearly heterogeneous among Mascarocoffea species. Indeed the proportion of different lineages also varied according to the botanical classification (Fig. 2). Two lineages, *SIRE* from *Copia* and *Del* from *Gypsy* appeared to differ strongly in the 454 reads between Eucoffea, Mozambicoffea, Mascarocoffea and Paracoffea. In Eucoffea, the *SIRE* lineage is present in 4.5–5.1 % of the 454 reads (identified with BLASTN, value $10e^{-20}$), at the exception of *C. charrieriana* for which 3.2 % of reads contained this lineage. Mozambicoffea species contained a lower percentage of *SIRE*, with 2.1 and 2.2 % for *C. pseudozamguebariae* and *C. racemosa*, while *SIRE* sequences were very rare in Mascarocoffea species and Paracoffea (between 1.1 and 1.5 %). Another important variation between botanical groups is observed for the *Del* fraction; going from 16.2 to 14 % in Eucoffea, 10.7 to 11.6 % in Mozambicoffea, 7.3 to 9.9 % in Mascarocoffea and 7.2 % in Paracoffea (Fig. 2; Supplemental data 4). Here also, the lowest percentage in Eucoffea is observed for *C. charrieriana* (13.1 %), contrasting with the other species of this botanical group. The pattern of LTR retrotransposon identified in *C. charrieriana*, suggests that this species differs from all the other Eucoffea species studied here.

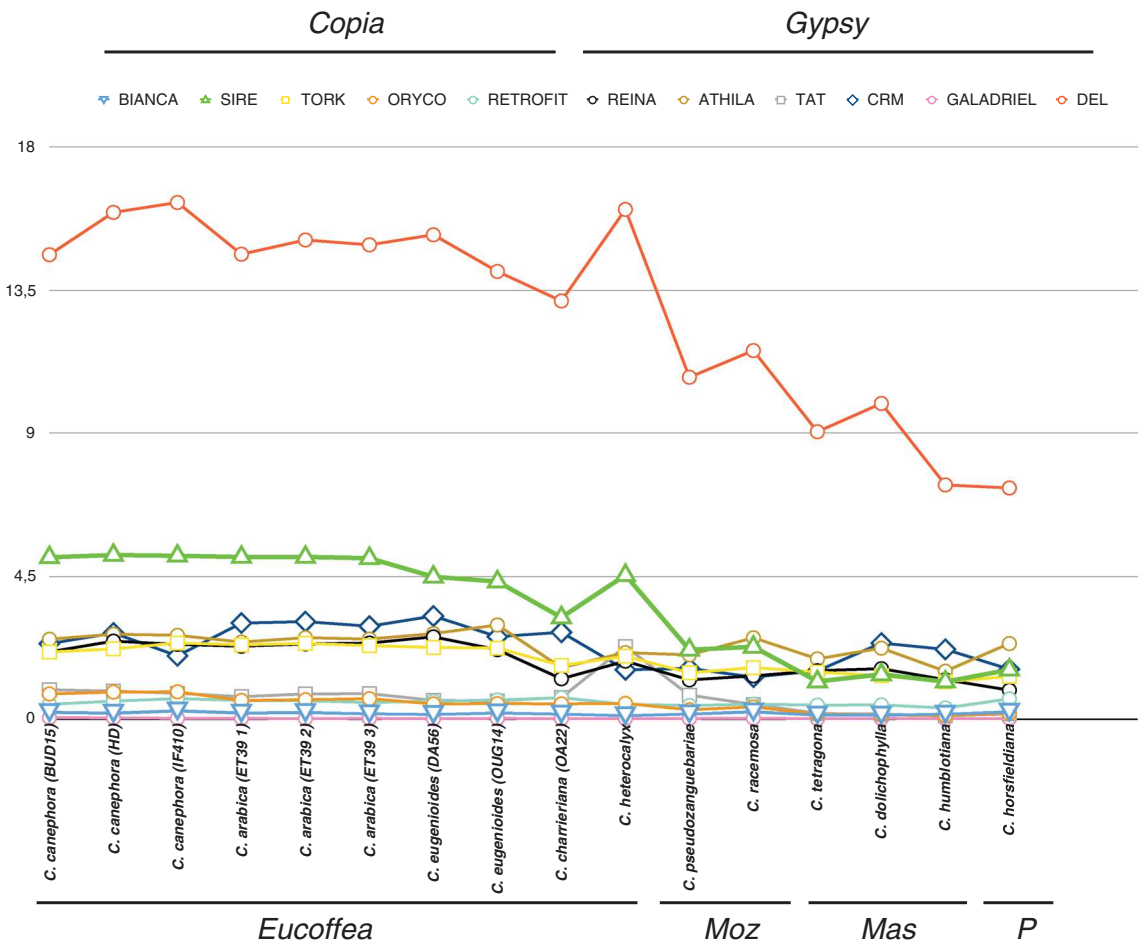


Fig. 2 Composition of 454 reads (in percentage) similar to LTR retrotransposon lineages between 11 *Coffea* species, organized according to their botanical sections: Eucoffea, Mozambicoffea (Moz), Mascarocoffea (Mas) and Paracoffea (P)

Interestingly, no clear relationship was found between the abundance of LTR retrotransposon super-families or lineages and, the genome size variation. However, there is a clear relationship between the abundance of detected elements and the botanical classification of the *Coffea* species.

De novo detection of repeated sequences in *Coffea*

As no clear relationship could be established between the presence of LTR retrotransposons and the genome size variation in *Coffea* genomes, another type of repeated sequences should be involved. For this, we estimated the global number of repeated sequences (excluding microsatellite sequences) presenting more than 20, 100, 500 and 1000 repeats and their proportion in each dataset (Supplemental data 5). Repeated sequences with a minimum of 20 copies represented between 54.4 (*C. heterocalyx*) and 45.6 % (*C. canephora*) of reads for Eucoffea, 46.3 and 41.8 % for Mozambicoffea, 43.4–33.3 % (*C. humblotiana*) for Mascarocoffea and 44.1 % for Paracoffea. A similar pattern was observed for repeated sequences with more than 100, 500 and 1000 copies. *C. heterocalyx* (863 Mb) and *C. canephora* (IF410; 700 Mb) are the two samples with the highest proportion of repeated sequences, while *C. humblotiana*, the smallest genome, has the lower number of repetitions. Among Mascarocoffea, this percentage differs considerably between *C. humblotiana* (469 Mb) and *C. dolichophyla* (682 Mb). Interestingly, some species appears enriched with highly repeated sequences (>500 and >1000 copies), such as *C. heterocalyx* (10.8 % of sequences were repeated more than 500 times), while *C. humblotiana* and *C. horsfieldiana* contained very few highly repeated sequences (Supplemental data 5).

Microsatellites and genome size variation

Different types of microsatellites were identified and their cumulative length was represented on a histogram (Supplemental data 6). No large variation of the microsatellite content was observed among the species analyzed. Indeed,

the amount of microsatellite is higher in *C. arabica*, which is the allotetraploid species, but for the diploid species it doesn't show any variation corresponding to the genome size, whatever the size of the microsatellite motif (Supplemental data 7).

The SIRE LTR retrotransposon lineage and *Coffea* geographic distribution

As LTR retrotransposons represented a significant but variable part of *Coffea* genomes, we assess their relationships from phylogenetic analysis based on their RT domains at the amino acid level. The tree obtained using 2,325 RT domains (with a minimum length of 150 amino acids) (Supplemental data 8) shows clearly an organization into lineages between the two super-families *Gypsy* and *Copia*. For each lineage, it was possible to observe a combination of RT domains from different botanical groups (Eucoffea, Mozambicoffea, Mascarocoffea and Paracoffea). However, one lineage named SIRE, showed a specific pattern with an over-representation of RT sequences from Eucoffea and Mozambicoffea and, very few from Mascarocoffea and Paracoffea. From the 263 RT belonging to the SIRE lineage, five belong to the Indonesian Paracoffea species, and 21, 49 and 188 belong to Mascarocoffea, Mozambicoffea and Eucoffea, respectively. This observation suggests a different dynamics of SIRE elements depending on the botanical group of the species. An in-depth study of this lineage was performed to confirm our observations.

SIRE LTR retrotransposons were identified, annotated and characterized in the *C. canephora* genome (Chaparro et al. 2015). After detailed analysis, a total of 85 full-length SIRE LTR retrotransposons were selected for further analyses. SIRE elements from this dataset showed strong similarities with the SIRE internal coding domains from the Gypsy 2 database, and they had no apparent large insertion. All these predicted SIRE elements showed an overall length around 9–10 kb, with an average LTR length of 1 kb. The internal regions of these sequences included a large open reading frame (ORF1) containing the consensus for the GAG, AP, INT, RT and RNaseH domains. Downstream of ORF1 an additional small ORF (ORF2) showing strong identities with the ENV domain of retroviruses was identified.

These 85 sequences were classified through phylogenetic analysis based on their LTR sequences, into three major clusters (A, B and C) composed of 17, 28 and 40 elements, respectively (Supplemental data 9). For each cluster, one full-length sequence (with highest percentage of LTR identity, and highest overall length) was used as a reference sequences for further analyses (the sequences were named 36-863, 3-942 and 6-1571 for A, B and C cluster, respectively).

The copy number of SIREs elements estimated in the set of species analyzed here and using the three references SIRE sequences previously defined, showed a large variation between botanical groups (Supplemental data 10). The highest number was obtained for the Eucoffea with the exception of *C. charrieriana,* while Mascarocoffea species and Paracoffea showed very few SIRE sequences. The Mozambicoffea showed a moderate number of SIRE copies, whose numbers ranged between that of Eucoffea and Mascarocoffea. To confirm these observations at the molecular level, we conducted a PCR amplification survey of LTR and/or ENV domains based on the three SIRE elements reference over a large panel of species (Supplemental data 1). Amplification products were obtained for nearly all the Eucoffea, while amplifications were obtained for few Mozambicoffea species and almost no amplifications were observed for the Mascarocoffea and Paracoffea (Fig. 3; Supplemental data 11).

Discussion

The objective of this study was to investigate the transposable element composition of diploid and allotetraploid genomes from the *Coffea* genus. In some plant genomes, a clear relationship was established between the number of LTR retrotransposons and the variation of genome size (Piegu et al. 2006; Lee and Kim 2014). Considering a relatively short evolutionary divergence time of the *Coffea* genus [~11 MY; (Tosh et al. 2013)] and a significant variation of genome size observed among species (from 469 to 900 Mb), we focused our study on the identification and the characterization of repeated sequences and more particularly the LTR retrotransposons.

We used the 454 Junior apparatus to produce partial genome sequencing, representing genome coverage from 3.2 to 4.7 % for the allotetraploid *C. arabica* and 5.2–15.7 % for ten diploid *Coffea* species. Such "454 whole genome snapshot" approach has been recently used in plant and animal genomes to study and compare their composition in transposable elements, with similar or even lower (Wicker et al. 2009; Middleton et al. 2013; Sergeeva et al. 2014; Swaminathan et al. 2007; Pagan et al. 2012). No bias of genomic sampling for particular sequence type was noted when using the 454 sequencing procedure (Swaminathan et al. 2007). Indeed using a relatively low genome coverage, only highly repeated transposable elements can be accurately studied and low-copy number repeated sequences will not be represented in our dataset (Macas et al. 2007). Despite the 454 sequencing technology is beginning to be outdated; it generates long reads allowing an accurate identification of genes and transposable elements. Other approaches are now possible to study the

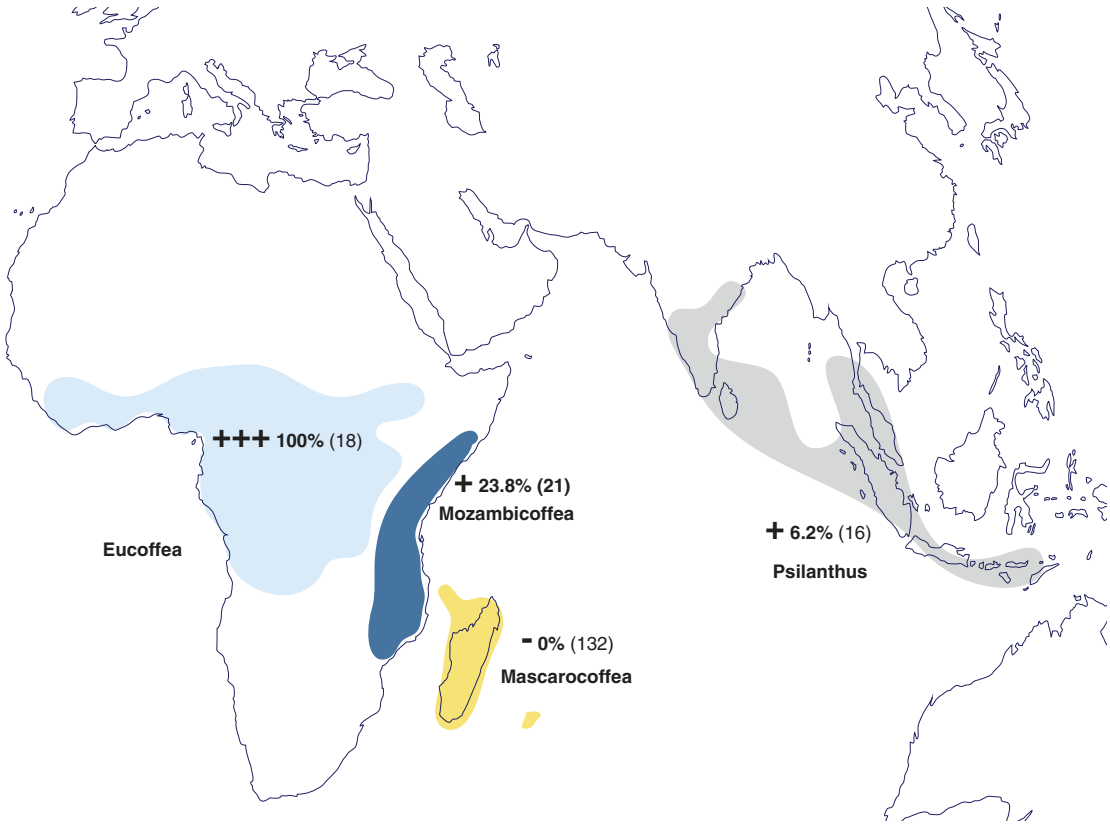


Fig. 3 Geographical distribution of *Coffea* botanical groups and summary of SIRE PCR amplifications. The summary of SIRE PCR amplification is symbolized by the rate of PCR amplification in percentage, and the number of PCR assays performed in parenthesis. 18, 7, 30 and 4 species were used as DNA matrix for, respectively, Eucoffea, Mozambicoffea, Mascarocoffea and Paracoffea (Supplemental data 8, 9 and 10)

transposable element composition and copy numbers using the Illumina platform providing shorter read length but with an unrivaled genome coverage (Ramachandran and Hawkins 2016).

In this study, we confirmed that no bias was observed in the randomness of the sequencing when performing three repetitions of the same accession (*C. arabica*, Et39). So far, few studies concerned the TE abundance and dynamics among species from a single plant genus. Most of them were performed on annual plants, with the exception of the gymnosperm family (Nystedt et al. 2013). However, for perennial angiosperms, the dynamics and the evolutionary history of TE within a genus remain poorly studied.

TE composition reflects the divergence of the botanical groups

For the first time, we conducted a study on TE composition of the genome of eleven *Coffea* species. Our study was based on the analysis of 454 reads for (1) their similarities with known TE proteins in plants and against a library of TE annotated in the *Coffea canephora* genome; and, (2) an ab initio identification of repeated sequences.

We found that the most repeated order of transposable elements are LTR retrotransposons as found in the *C. canephora* genome and in most plant genomes (Lee and Kim 2014). At the TE amino acid level, as curated in Repbase, we found a similar percentage of TE between the generated *C. canephora* 454 reads (ranged from 12.2 to 13.5 %) and a recent and similar analysis of 131,412 BAC End Sequences (BES) from two *C. canephora* (DH200-94) BAC libraries [11.9 % (Dereeper et al. 2013)]. Surprisingly the percentage of known TE coding sequences remains relatively stable whatever the botanical groups, the species and the genome size. Interestingly, the only notable differences concerned *C. dolichophylla* and *C. humblotiana* species showing, respectively, 14.6 and 10.2 % of detected TE coding sequences. Considering the genome size difference (689 and 469 Mb), these species that belong to the Mascarocoffea may have underwent a different history of TE accumulation. This observation was confirmed at the nucleotide level using a *C. canephora de novo* library of TEs using REPET.

Using a detailed classification of LTR-RT REPET consensuses, we also found that some lineages have varying distribution levels among *Coffea* species and botanical groups. For example the *Gypsy* Del lineage identified in higher abundance in African species, decreases from Eucoffea species (14–16 %), to Mozambicoffea (10–11 %), Mascarocoffea and Paracoffea (9–7 %). This suggests an overall increase of the Del LTR-RT westwards; from Indonesian and Malagasy *Coffea* species to eastern and western African species. Another LTR RT lineage, named SIRE (*Copia* super-family) was identified as being significantly numerous in African species (in 5 % of the 454 reads), but almost absent in Indonesian, Madagascan and Comorian species (~1 %), this observation was confirmed by the realization of PCR amplifications (Fig. 3). This indicates that the SIREs proliferated successfully in African species (in Mozambicoffea and especially in Eucoffea) while the copy number remained low, by lack of activity or elimination, in species from insular species.

These two examples of LTR-RT lineages variation, suggesting different history of TE proliferation, reflect independent genome divergences between *Coffea* botanical groups. This result also suggests that geographical differentiation could be associated to independent niches colonization and speciation in Africa, Madagascar and Indonesia. Therefore, quantitative and qualitative TE composition might be used for performing phylogeny analysis and to reinforce a model for the evolution of plant species.

TE composition reflects a different evolution of species within the botanical groups

It is well established that plant genome sizes are directly linked with the proportion of transposable elements. A large amplification of a small number of LTR retrotransposons lineages may cause a dramatic and sudden genome size increase (Piegu et al. 2006). In our study, we found contrasted results between the genome size of *Coffea* species and their TEs composition.

Few variation of TE composition was related to the genome size in Eucoffea, although genome size varies from 645b Mb for *C. eugenioides,* to 863 Mb for *C. heterocalyx* (700 Mb for *C. canephora*). This suggests that no rapid proliferation of few TE families was involved to explain this genome size difference. Particularly the TE proportion is almost identical between *C. canephora* and *C. heterocalyx* with the exception to one *Gypsy* lineage named *TAT*, that varies from 0.9 % in *C. canephora* to 2.2 % in *C. heterocalyx*. However, this recent proliferation in *C. heterocalyx* cannot explain alone the genome size difference between the two species. We, therefore, propose that in Eucoffea the genome size variation would result from a differential accumulation of numerous transposable elements (mainly LTR RT) belonging to a large panel of families.

Similarly, no strong variation of microsatellite copy numbers was detected between species, suggesting that a rapid amplification of some of these simple sequence repeats was not the main mechanisms involved in the *C. heterocalyx* genome size increase as it was observed in *Lupinus* (Martin et al. 2016). Our results are congruent with those of Pinus (Morse et al. 2009), Helianthus (Cavallini et al. 2010), and Lupinus (Martin et al. 2016) both genera showing a large genome size variation (18–40, 3.2–12.3 and, 0.97–2.4 Gb, respectively) but with none element contributing specifically to this variation.

At the opposite, the Mascarocoffea species present more important variations of their TEs composition. The strong contrast in TE content between *C. dolichophylla* and *C. humblotiana* is due to an increase/decrease of the amount of the Del LTR retrotransposon lineage (10 vs 7 %) and a smaller increase/decrease for the remaining LTR RT lineages. *C. humblotiana*, has undergone few proliferation of LTR retrotransposons explaining its small genome size (469 Mb) while *C. dolichophylla* has undergone proliferation of mainly *Del* and several other *Copia* and *Gypsy* LTR-RT lineages. The variation of repeated sequences between *C. dolichophylla* and *C. humblotiana* is also clear with the *de novo* analysis showing a clear increase/decrease in repeated sequences. Since the fully resolved phylogenetic analysis of Mascarocoffea is not yet available, the timescale of the LTR RT proliferation in *C. dolichophylla* cannot be estimated.

Altogether, our analysis demonstrated the power of sequencing at low coverage to study the transposable elements composition of genomes at the genus scale for comparative structural genomics of non-model species. The *C. humblotiana* species represents an interesting genomic model, worth to have its genome completely sequenced. This WGS will allow a better understanding of the mechanisms involved in the decrease or in the control of the proliferation of transposable elements in a genome.

Compliance with ethical standards

Conflict of interest All authors declare they have no conflict of interest.

Funding This research was supported Agropolis Fondation through the "Investissement d'avenir" program (ANR-10-LABX-0001-01) under the reference ID 1002-009.

Ethical approval This article does not contain any studies with human or animals performed by any of the authors.

Data availability The project has been deposited at DDBJ/EMBL/GenBank BioProject ID PRJNA242989.

References

Alzohairy A, Sabir J, Gyulai G, Younis R, Jansen RK, Bahieldin A (2014) Environmental stress activation of plant long-terminal repeat retrotransposons. Funct Plant Biol 41:557–567

Bennetzen JL, Ma J, Devos KM (2005) Mechanisms of recent genome size variation in flowering plants. Ann Bot 95:127–132

Bremer B, Eriksson T (2009) Time tree of Rubiaceae: phylogeny and dating the family, subfamilies, and tribes. Int J Plant Sci 170:766–793

Bucher E, Reinders J, Mirouze M (2012) Epigenetic control of transposon transcription and mobility in Arabidopsis. Curr Opin Plant Biol 15:503–510

Carrier G, Santoni S, Rodier-Goud M, Canaguier A, Kochko A, Dubreuil-Tranchant C, This P, Boursiquot JM, Le Cunff L (2011) An efficient and rapid protocol for plant nuclear DNA preparation suitable for next generation sequencing methods. Am J Bot 98:e13–e15

Carrier G, Le Cunff L, Dereeper A, Legrand D, Sabot F, Bouchez O, Audeguin L, Boursiquot JM, This P (2012) Transposable elements are a major cause of somatic polymorphism in *Vitis vinifera* L. PLoS One 7:10

Casacuberta E, Gonzalez J (2013) The impact of transposable elements in environmental adaptation. Mol Ecol 22:1503–1517

Cavallini A, Natali L, Zuccolo A, Giordani T, Jurman I, Ferrillo V, Vitacolonna N, Sarri V, Cattonaro F, Ceccarelli M, Cionini PG, Morgante M (2010) Analysis of transposons and repeat composition of the sunflower (*Helianthus annuus* L.) genome. Theor Appl Genet 120:491–508

Chaparro C, Gayraud T, de Souza RF, Domingues DS, Akaffou S, Laforga Vanzela AL, Kochko A, Rigoreau M, Crouzillat D, Hamon S, Hamon P, Guyot R (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

Chevalier A (1942) Les caféiers du globe II: Iconographie des caféiers sauvages et cultivés et des Rubiacées prises pour des caféiers. In: Lechevalier P (ed) Encyclopédie Biologique, Paris

Davis AP, Tosh J, Ruch N, Fay MF (2011) Growing coffee: *Psilanthus* (Rubiaceae) subsumed on the basis of molecular and morphological data; implications for the size, morphology, distribution and evolutionary history of *Coffea*. Bot J Linn Soc 167:357–377

Denoeud F, Carretero-Paulet L, Dereeper A, Droc G, Guyot R, Pietrella M, Zheng C, Alberti A, Anthony F, Aprea G, Aury JM, Bento P, Bernard M, Bocs S, Campa C, Cenci A, Combes MC, Crouzillat D, Da Silva C, Daddiego L, De Bellis F, Dussert S, Garsmeur O, Gayraud T, Guignon V, Jahn K, Jamilloux V, Joët T, Labadie K, Lan I, Leclercq J, Lepelley M, Leroy T, Li LT, Librado P, Lopez L, Muñoz A, Noel B, Pallavicini A, Perrotta G, Poncet V, Pot D, Priyono Rigoreau M, Rouard M, Rozas J, Tranchant-Dubreuil C, VanBuren R, Zhang Q, Andrade AC, Argout X, Bertrand B, de Kochko A, Graziosi G, Henry RJ, Jayarama Ming R, Nagai C, Rounsley S, Sankoff D, Giuliano G, Victor A, Albert V, Wincker P, Lashermes P (2014) The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. Science 345:1181–1184

Dereeper A, Guyot R, Tranchant-Dubreuil C, Anthony F, Argout X, de Bellis F, Combes MC, Gavory F, de Kochko A, Kudrna D, Leroy T, Poulain J, Rondeau M, Song X, Wing R, Lashermes P (2013) BAC-end sequences analysis provides first insights into coffee (*Coffea canephora* P.) genome composition and evolution. Plant Mol Biol 83:177–189

Dias ES, Hatt C, Hamon S, Hamon P, Rigoreau M, Crouzillat D, Carareto CM, De Kochko A, Guyot R (2015) Large distribution and high sequence identity of a Copia-type retrotransposon in angiosperm families. Plant Mol Biol 89:83–97

Dušková E, Kolář F, Sklenář P, Rauchová J, Kubešová M, Fér T, Suda J, Marhold K (2010) Genome size correlates with growth form, habitat and phylogeny in the Andean genus *Lasiocephalus* (Asteraceae). Preslia 82:127–148

Dvořák J (2009) Triticeae genome structure and evolution. In: Muehlbauer JG, Feuillet C (eds) Genetics and genomics of the Triticeae. Springer, New York, pp 685–711

Eilam T, Anikster Y, Millet E, Manisterski J, Sag-Assif O, Feldman M (2007) Genome size and genome evolution in diploid Triticeae species. Genome 50:1029–1037

Hawkins JS, Kim H, Nason JD, Wing RA, Wendel JF (2006) Differential lineage-specific amplification of transposable elements is responsible for genome size variation in Gossypium. Genome Res 16:1252–1261

Hribova E, Neumann P, Matsumoto T, Roux N, Macas J, Dolezel J (2010) Repetitive part of the banana (*Musa acuminata*) genome investigated by low-depth 454 sequencing. BMC Plant Biol 10:204

Ibarra-Laclette E, Lyons E, Hernandez-Guzman G, Perez-Torres CA, Carretero-Paulet L, Chang T-H, Lan T, Welch AJ, Juarez MJA, Simpson J, Fernandez-Cortes A, Arteaga-Vazquez M, Gongora-Castillo E, Acevedo-Hernandez G, Schuster SC, Himmelbauer H, Minoche AE, Xu S, Lynch M, Oropeza-Aburto A, Cervantes-Perez SA, de Jesus Ortega-Estrada M, Cervantes-Luevano JI, Michael TP, Mockler T, Bryant D, Herrera-Estrella A, Albert VA, Herrera-Estrella L (2013) Architecture and evolution of a minute plant genome. Nature 498:94–98

Ito H (2013) Small RNAs and regulation of transposons in plants. Genes Genet Syst 88:3–7

Ito H, Kakutani T (2014) Control of transposable elements in *Arabidopsis thaliana*. Chromosome Res 22:217–223

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J (2005) Repbase update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 110:462–467

Kiehn M (1995) Chromosome survey of the Rubiaceae. Ann Mo Bot Gard 82:398–408

Kinoshita T, Seki M (2014) Epigenetic memory for stress response and adaptation in plants. Plant Cell Physiol 55:1859–1863

Knight CA, Beaulieu JM (2008) Genome size scaling through phenotype space. Ann Bot 101:759–766

Kohany O, Gentles AJ, Hankus L, Jurka J (2006) Annotation, submission and screening of repetitive elements in Repbase: Repbase Submitter and Censor. BMC Bioinf 7:474

Kumar A, Bennetzen JL (1999) Plant retrotransposons. Annu Rev Genet 33:479–532

Lee SI, Kim NS (2014) Transposable elements and genome size variations in plants. Genomics Inform 12:87–97

Lisch D (2013) How important are transposons for plant evolution? Nat Rev Genet 14:49–61

Llorens C, Munoz-Pomer A, Bernad L, Botella H, Moya A (2009) Network dynamics of eukaryotic LTR retroelements beyond phylogenetic trees. Biol Direct 4:41

Macas J, Neumann P, Navratilova A (2007) Repetitive DNA in the pea (*Pisum sativum* L.) genome: comprehensive characterization using 454 sequencing and comparison to soybean and Medicago truncatula. BMC Genom 8:427

Martin G, Paris A, Samar M, Keller J, Salmon A, Novak P, Macas J, Aïnouche A (2016) Dramatic lineage-specific accumulation of retrotransposons versus Simple Sequence Repeats across the last 10 million years in Mediterranean and African lupin genomes (Lupinus; Fabaceae). In: International Congress on Transposable elements, Saint Malo, France

Meyers BC, Tingey SV, Morgante M (2001) Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. Genome Res 11:1660–1676

Michael TP, Jackson S (2013) The first 50 plant genomes. Plant Genome 6:1–7

Middleton CP, Stein N, Keller B, Kilian B, Wicker T (2013) Comparative analysis of genome composition in Triticeae reveals strong variation in transposable element dynamics and nucleotide diversity. Plant J 73:347–356

Morse AM, Peterson DG, Islam-Faridi MN, Smith KE, Magbanua Z, Garcia SA, Kubisiak TL, Amerson HV, Carlson JE, Nelson CD, Davis JM (2009) Evolution of genome size and complexity in Pinus. PLoS One 4:e4332

Noirot M, Poncet V, Barre P, Hamon P, Hamon S, De Kochko A (2003) Genome size variations in diploid African Coffea species. Ann Bot (Lond) 92:709–714

Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin YC, Scofield DG, Vezzi F, Delhomme N, Giacomello S, Alexeyenko A, Vicedomini R, Sahlin K, Sherwood E, Elfstrand M, Gramzow L, Holmberg K, Hallman J, Keech O, Klasson L, Koriabine M, Kucukoglu M, Kaller M, Luthman J, Lysholm F, Niittyla T, Olson A, Rilakovic N, Ritland C, Rossello JA, Sena J, Svensson T, Talavera-Lopez C, Theissen G, Tuominen H, Vanneste K, Wu ZQ, Zhang B, Zerbe P, Arvestad L, Bhalerao R, Bohlmann J, Bousquet J, Garcia Gil R, Hvidsten TR, de Jong P, MacKay J, Morgante M, Ritland K, Sundberg B, Thompson SL, Van de Peer Y, Andersson B, Nilsson O, Ingvarsson PK, Lundeberg J, Jansson S (2013) The Norway spruce genome sequence and conifer genome evolution. Nature 497:579–584

Pagan HJ, Macas J, Novak P, McCulloch ES, Stevens RD, Ray DA (2012) Survey sequencing reveals elevated DNA transposon activity, novel elements, and variation in repetitive landscapes among vesper bats. Genome Biol Evol 4:575–585

Piegu B, Guyot R, Picault N, Roulin A, Saniyal A, Kim H, Collura K, Brar DS, Jackson S, Wing RA, Panaud O (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

Price AL, Jones NC, Pevzner PA (2005) De novo identification of repeat families in large genomes. Bioinformatics 21:i351–i358

Ramachandran D, Hawkins JS (2016) Methods for accurate quantification of LTR-retrotransposon copy number using short-read sequence data: a case study in Sorghum. Mol Genet Genomics

Razafinarivo N, Rakotomalala JJ, Brown SC, Bourge M, Hamon S, De Kochko A, Poncet V, Dubreuil-Tranchant C, Couturon E, Guyot R, Hamon P (2012) Geographical gradients in the genome size variation of wild coffee trees (Coffea) native to Africa and Indian Ocean islands. Tree Genet Genomes 8:1345–1358

Razafinarivo NJ, Guyot R, Davis AP, Couturon E, Hamon S, Crouzillat D, Rigoreau M, Dubreuil-Tranchant C, Poncet V, De Kochko A, Rakotomalala JJ, Hamon P (2013) Genetic structure and diversity of coffee (Coffea) across Africa and the Indian Ocean islands revealed using microsatellites. Ann Bot 111:229–248

Renny-Byfield S, Chester M, Kovarik A, Le Comber SC, Grandbastien M-A, Deloger M, Nichols RA, Macas J, Novak P, Chase MW, Leitch AR (2011) Next generation sequencing reveals genome downsizing in allotetraploid *Nicotiana tabacum*, predominantly through the elimination of paternally derived repetitive DNAs. Mol Biol Evol 28:2843–2854

Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B (2000) Artemis: sequence visualization and annotation. Bioinformatics 16:944–945

SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL (1998) The paleontology of intergene retrotransposons of maize. Nat Genet 20:43–45

Schmieder R, Edwards R (2011) Quality control and preprocessing of metagenomic datasets. Bioinformatics 27:863–864

Schulman AH, Gupta PK, Varshney RK (2004) Organization of retrotransposons and microsatellites in cereal genomes. In: Gupta PK, Varshney VR (eds) Cereal genomics. Kluwer Academic, Dordrecht, pp 83–118

Sergeeva EM, Afonnikov DA, Koltunova MK, Gusev VD, Miroshnichenko LA, Vrána J, Kubaláková M, Poncet C, Sourdille P, Feuillet C, Doležel J, Salina EA (2014) Common wheat chromosome 5B composition analysis using low-coverage 454 sequencing. Plant Genome 7:1–16

Slovak M, Vit P, Urfus T, Suda J (2009) Complex pattern of genome size variation in a polymorphic member of the Asteraceae. J Biogeogr 36:372–384

Sonnhammer ELL, Durbin R (1995) A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis (reprinted from Gene Combis, vol 167, pg GC1-GC10, 1995). Gene 167:GC1–GC10

Stoffelen P, Noirot M, Couturon E, Anthony F (2008) A new caffeine-free coffee from Cameroon. Bot J Linn Soc 158:67–72

Swaminathan K, Varala K, Hudson ME (2007) Global repeat discovery and estimation of genomic copy number in a large, complex genome using a high-throughput 454 sequence survey. BMC Genom 8:132

Todorovska E (2007) Retrotransposons and their role in plant-Genome evolution. Biotechnol Biotechnol Equip 21:294–305

Tosh J, Dessein S, Buerki S, Groeninckx I, Mouly A, Bremer B, Smets EF, De Block P (2013) Evolutionary history of the Afro-Madagascan Ixora species (Rubiaceae): species diversification and distribution of key morphological traits inferred from dated molecular phylogenetic trees. Ann Bot 112:1723–1742

Wicker T, Keller B (2007) Genome-wide comparative analysis of copia retrotransposons in Triticeae, rice, and Arabidopsis reveals conserved ancient evolutionary lineages and distinct dynamics of individual copia families. Genome Res 17:1072–1081

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, Paux E, SanMiguel P, Schulman AH (2007) A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8:973–982

Wicker T, Taudien S, Houben A, Keller B, Graner A, Platzer M, Stein N (2009) A whole-genome snapshot of 454 sequences exposes the composition of the barley genome and provides evidence for parallel evolution of genome size in wheat and barley. Plant J 59:712–722

Xu Z, Wang H (2007) LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res 35:W265–W268

Supplemental data 1. PCR amplification on *Coffea* DNA.
A. List of primers used for PCR amplification of three SIRE families. B. List of DNA sample used.

A

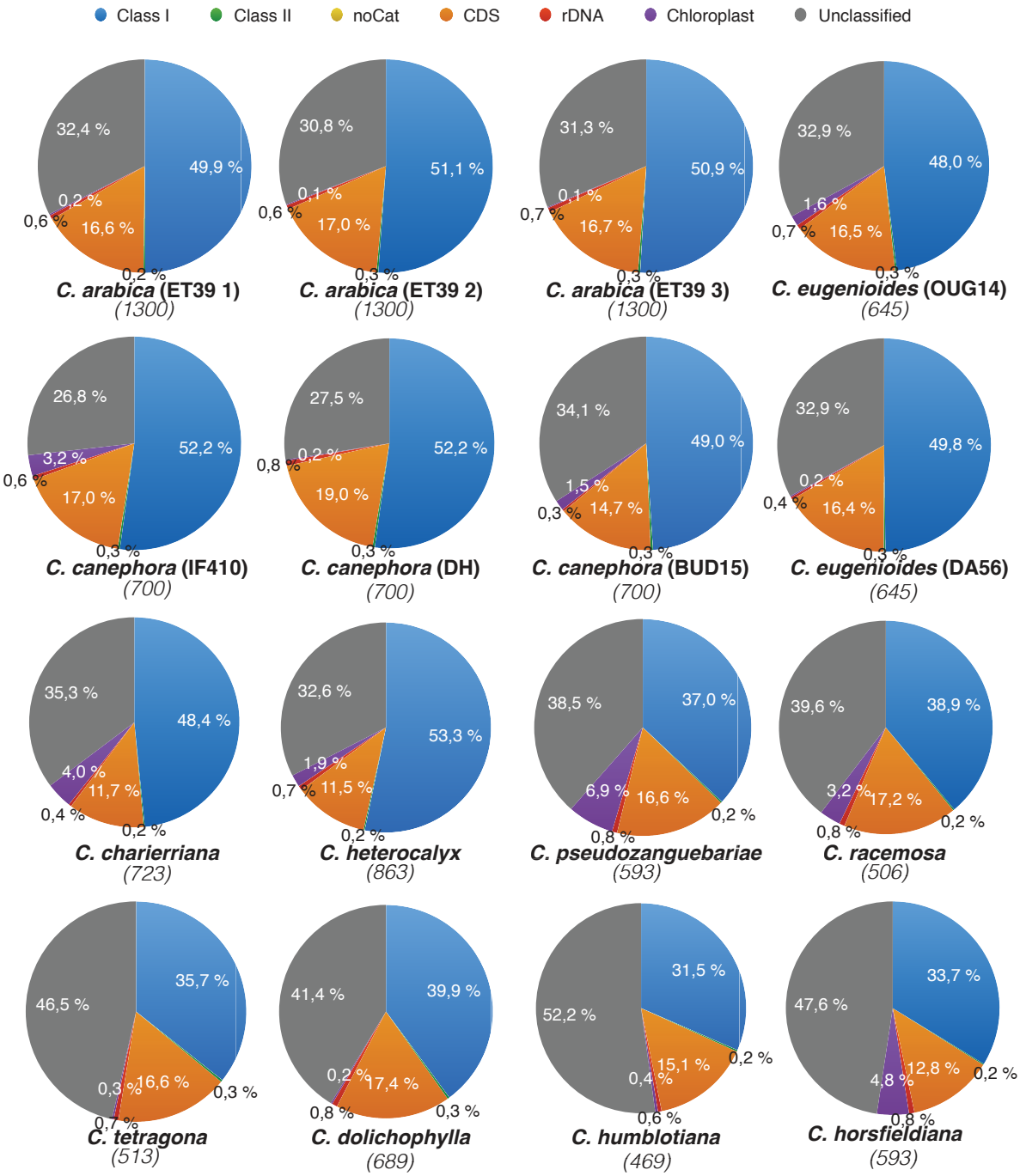
Primers	Sequence (5'-3')	Product Size (bp)
3-942 ENV 5'	CCAAAGTCAATGCGGATTTT	172
3-942 ENV 3'	AATTCAACACCCCTGCTGAG	172
3-942 LTR 5'	CTTCCTTCTTGCCCCAAACC	396
3-942 LTR 3'	GCCGTCCGAAGAAAGTGAAG	396
36-863 LTR 5'	TCTGGTTGATATGCGTCCGA	373
36-863 LTR 3'	TACGATGCTTTGGATGTGGC	373
6-1571 LTR 5'	CGGACGCATGAAAACTGAGT	346
6-1571 LTR 3'	CTTCCTTCTTGCCCCAAACC	346

B

Species	Accession ID	Country of Origin	Classification (Chevalier, 1947)
Coffea kapakata	OK	Angola	Eucoffea
Coffea anthonyi	OD72	Cameroon	Eucoffea
Coffea brevipes	JA54	Cameroon	Eucoffea
Coffea charrieriana	OA22	Cameroon	Eucoffea
Coffea montekupensis	OM11	Cameroon	Eucoffea
Coffea sp. Nkoumbala	OI60	Cameroon	Eucoffea
Coffea sp. Koto	EC61	Cameroon	Eucoffea
Coffea sp. Congo	OB60	Republic of Congo	Eucoffea
Coffea sp. Ngongo2	OF62	Republic of Congo	Eucoffea
Coffea sp. Ngongo3	OG70	Republic of Congo	Eucoffea
Coffea heterocalyx	JC62	Cameroon	Eucoffea
Coffea eugenioides	DA58	Kenya	Eucoffea
Coffea congensis	CB65	Central African Republic	Eucoffea
Coffea liberica var dewevrei	EB51	Central African Republic	Eucoffea
Coffea canephora	BA53	Republic Côte d'Ivoire	Eucoffea
Coffea humilis	G57	Republic Côte d'Ivoire	Eucoffea
Coffea liberica var liberica	EA64	Republic Côte d'Ivoire	Eucoffea
Coffea stenophylla	FB55	Republic Côte d'Ivoire	Eucoffea
Coffea pseudozanguebariae	H52	Kenya	Mozambicoffea
Coffea sessiliflora	PB58	Kenya	Mozambicoffea
Coffea racemosa	IB62	Mozambique	Mozambicoffea
Coffea salvatrix	LA60	Mozambique	Mozambicoffea
Coffea bridsoniae	APD2910	Tanzania	Mozambicoffea
Coffea kihansiensis	APD2922	Tanzania	Mozambicoffea
Coffea mongensis	EF11	Tanzania	Mozambicoffea
Coffea tricalysioides	APD4503	Madagascar	Mascarocoffea
Coffea abbayesii	A.601	Madagascar	Mascarocoffea
Coffea ankaranensis	A.525	Madagascar	Mascarocoffea
Coffea augagneuri	A.966	Madagascar	Mascarocoffea
Coffea bertrandii	A.5	Madagascar	Mascarocoffea
Coffea betamponensis	A.573	Madagascar	Mascarocoffea
Coffea boiviniana	A.980	Madagascar	Mascarocoffea
Coffea bonnieri	A.535	Madagascar	Mascarocoffea
Coffea coursiana	A.570	Madagascar	Mascarocoffea
Coffea dubardii	A.969	Madagascar	Mascarocoffea
Coffea farafanganensis	A.208	Madagascar	Mascarocoffea
Coffea heimii	A.516	Madagascar	Mascarocoffea
Coffea jumellei	A.974	Madagascar	Mascarocoffea
Coffea kianjavatensis	A.602	Madagascar	Mascarocoffea
Coffea lancifolia	A.320	Madagascar	Mascarocoffea
Coffea leroyi	A.315	Madagascar	Mascarocoffea
Coffea andrambovatensis	A.310	Madagascar	Mascarocoffea
Coffea liaudii	A.1013	Madagascar	Mascarocoffea
Coffea mcphersonii	A.977	Madagascar	Mascarocoffea
Coffea millotii	A.222	Madagascar	Mascarocoffea
Coffea dolichophylla	A.206	Madagascar	Mascarocoffea
Coffea mogeneti	A.975	Madagascar	Mascarocoffea
Coffea montis-sacri	A.321	Madagascar	Mascarocoffea
Coffea perrieri	A.12	Madagascar	Mascarocoffea
Coffea ratsimamangae	A.528	Madagascar	Mascarocoffea
Coffea resinosa	A.8	Madagascar	Mascarocoffea
Coffea sakarahae	A.304	Madagascar	Mascarocoffea
Coffea tetragona	A.252	Madagascar	Mascarocoffea
Coffea toshii	A.1000	Madagascar	Mascarocoffea
Coffea tsirananae	A.515	Madagascar	Mascarocoffea
Coffea vatovavyensis	A.830	Madagascar	Mascarocoffea
Coffea mauritiana	BM17/25	Mauritius	Mascarocoffea
Coffea mauritiana	PCH	Reunion	Mascarocoffea
Psilanthus ebracteolatus	PSI11	Republic Côte d'Ivoire	Paracoffea
Psilanthus lebrunianus	15/713	Congo	Paracoffea
Psilanthus travancorensis	PBT2	India	Paracoffea
Psilanthus wightianus	PBT3	India	Paracoffea

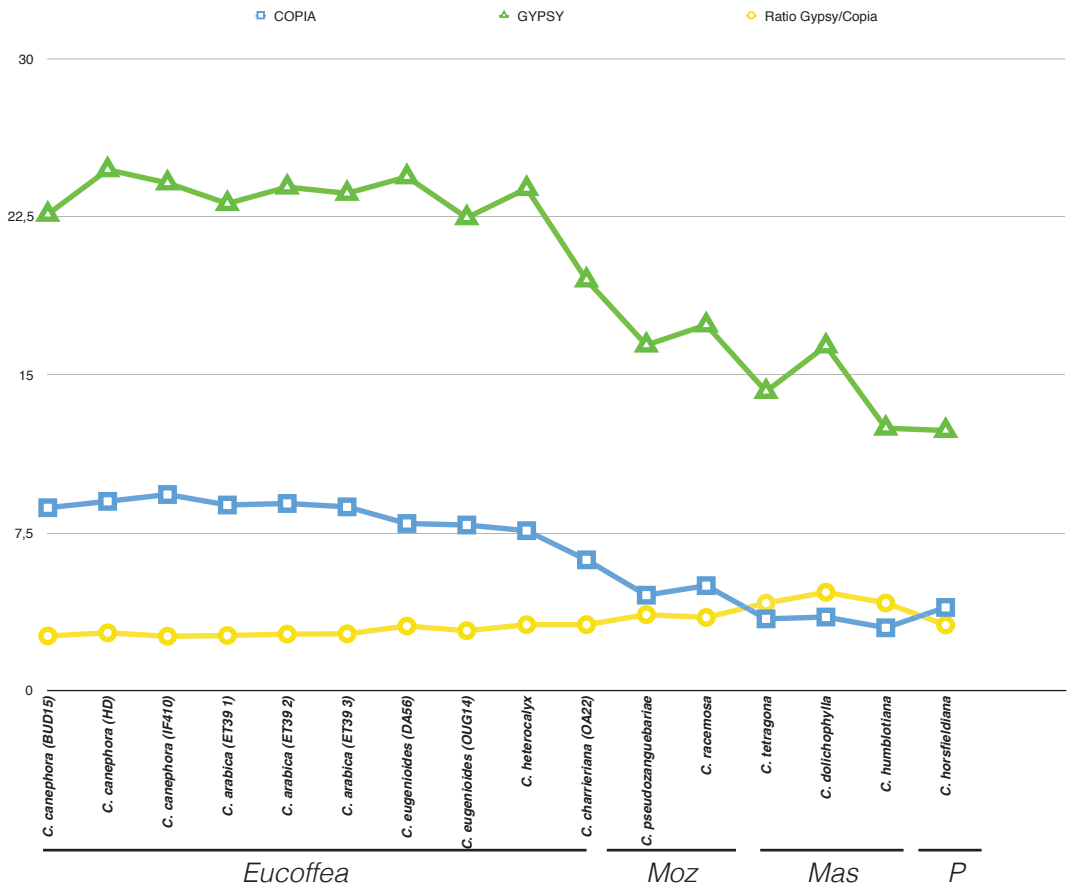
Supplemental data 2. Transposable element composition of 454 reads for 11 *Coffea* species and 14 accessions.

Class I, Class II and No Cat are Transposable Element as found in the *C. canephora* genome and annotated with REPET; CDS are cellular coding regions; rDNAs are ribosomal DNA genes. Name and accession of species were indicated with their respective genome size indicated in brackets (in Mb).

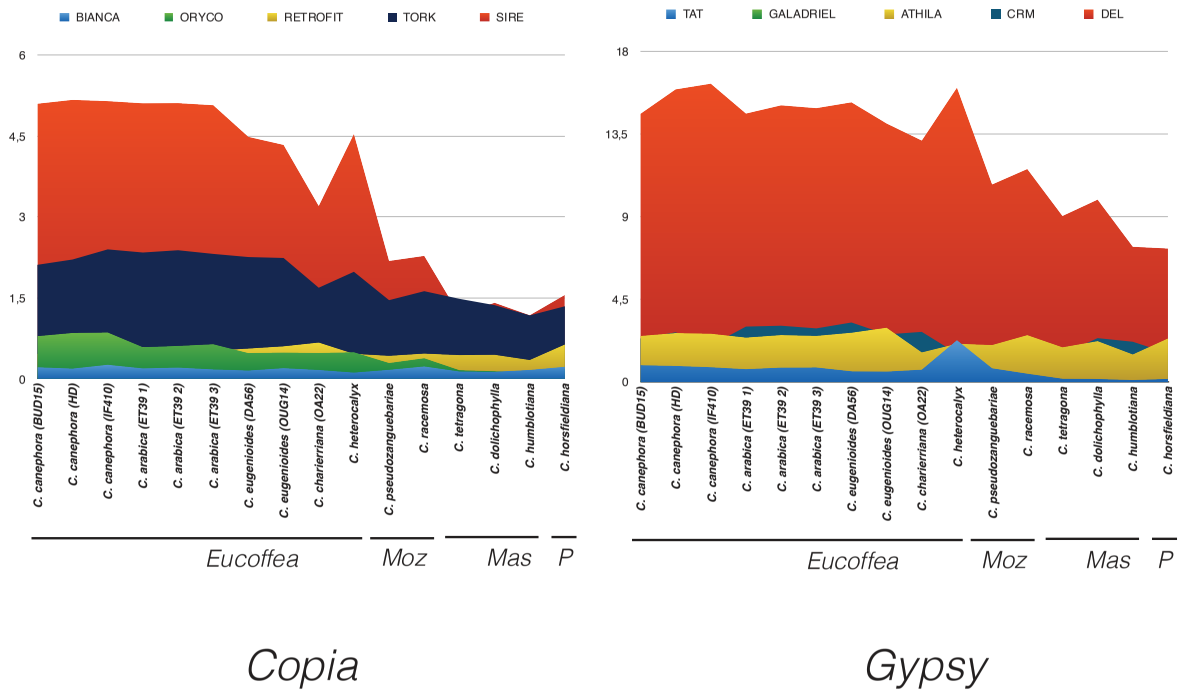


Supplemental data 3. Percentage of *Gypsy* and *Copia* LTR retrotransposons 454 reads for 11 *Coffea* species and 14 accessions.

The TE database used for similarity searches was established in *C. canephora* genome using REPET. The ratio of Gypsy/Copia LTR retrotransposons is also indicated.



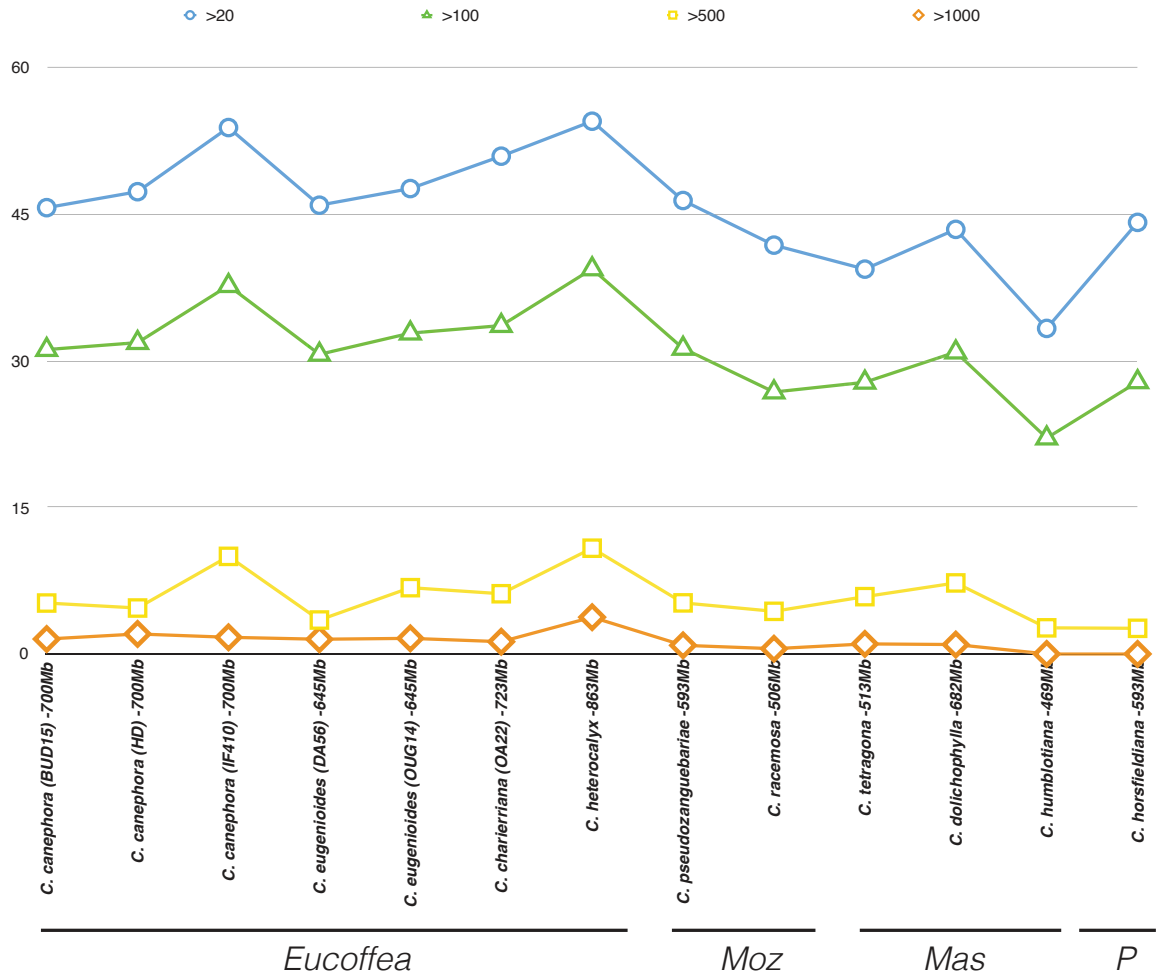
Supplemental data 4. Composition of 454 reads (in percentage) similar to LTR retrotransposon lineages between 11 *Coffea* species, organized according to their botanical sections: Eucoffea, Mozambicoffea (Moz), Mascarocoffea (Mas) and Paracoffea (P).



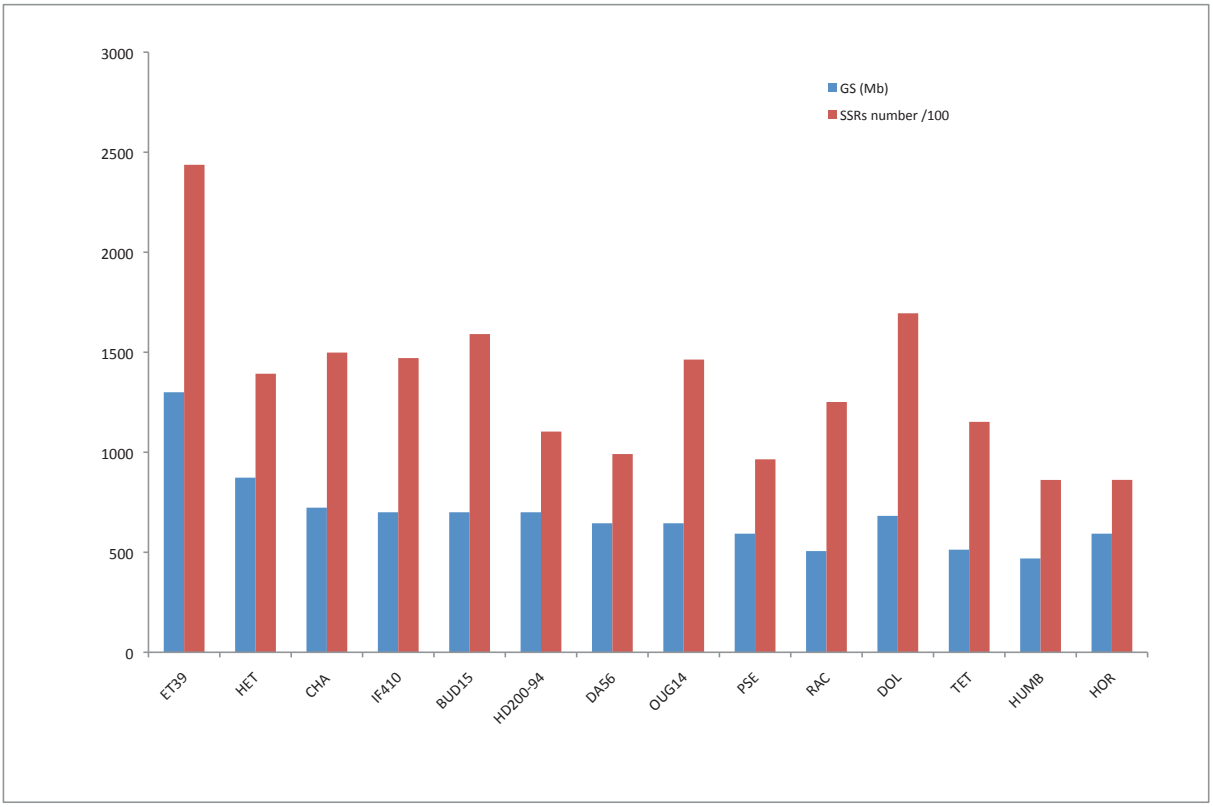
Copia *Gypsy*

Supplemental data 5. Percentage of repeated reads found by RepeatScout in 454 reads for 11 *Coffea* species and 14 accessions.

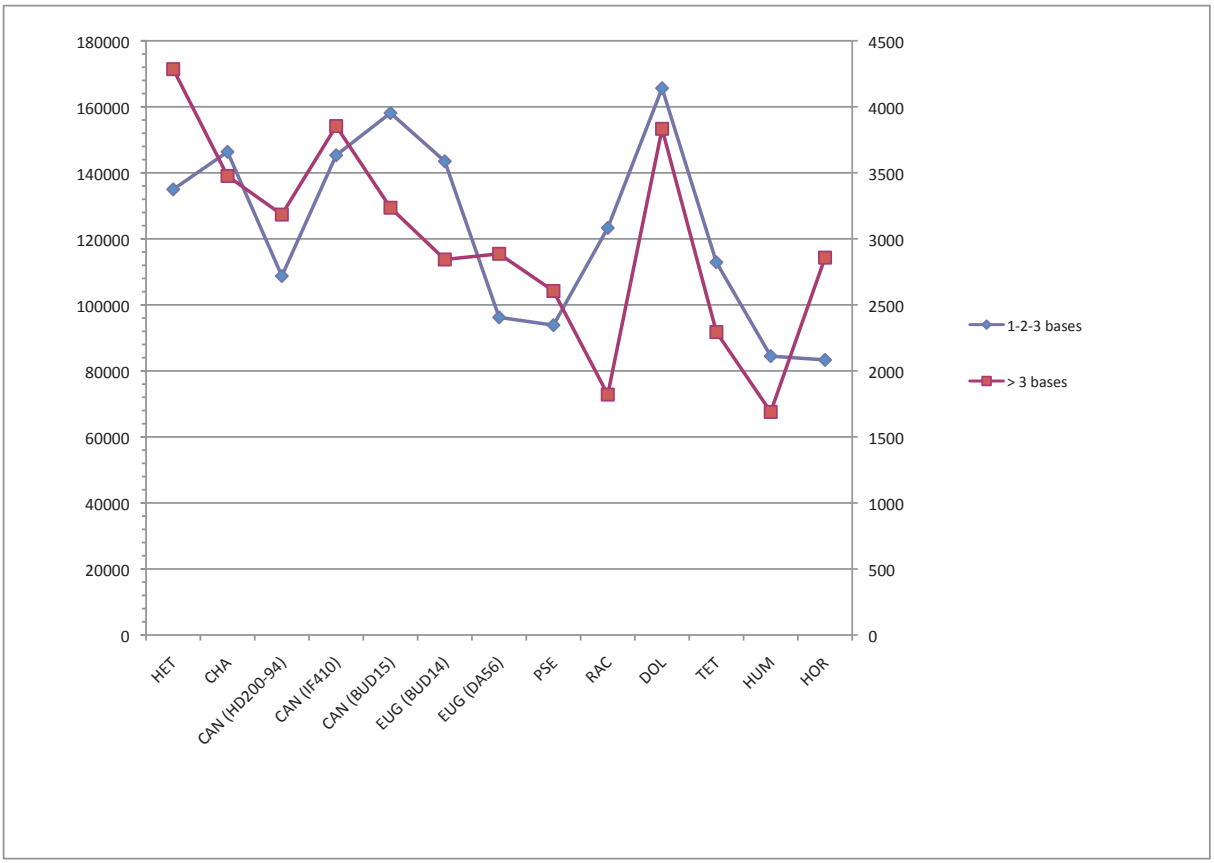
Repeated sequences were classified according to their redundancies: repeats with more than 20 occurrences in 454 dataset (blue), more than 100 occurrences (green), more than 500 occurrences (yellow) and more than 1000 occurrences (orange).



Supplemental data 6. Representation of the Genome size of species analyzed (GS, in blue) and their number of microsatellites (SSR, number/100, in red). *C. arabica* (ET39), *C. canephora* (IF410, BUD15, DH200-94), *C. eugenioides* (DA56, OUG14), *C. charrieriana* (CHA), *C. heterocalyx* (HET), *C. pseudozanguebariae* (PSE), *C. racemosa* (RAC), *C. dolichophylla* (DOL), *C. tetragona* (TET), *C. humblotiana* (HUMB) and *C. horsfieldiana* (HOR).

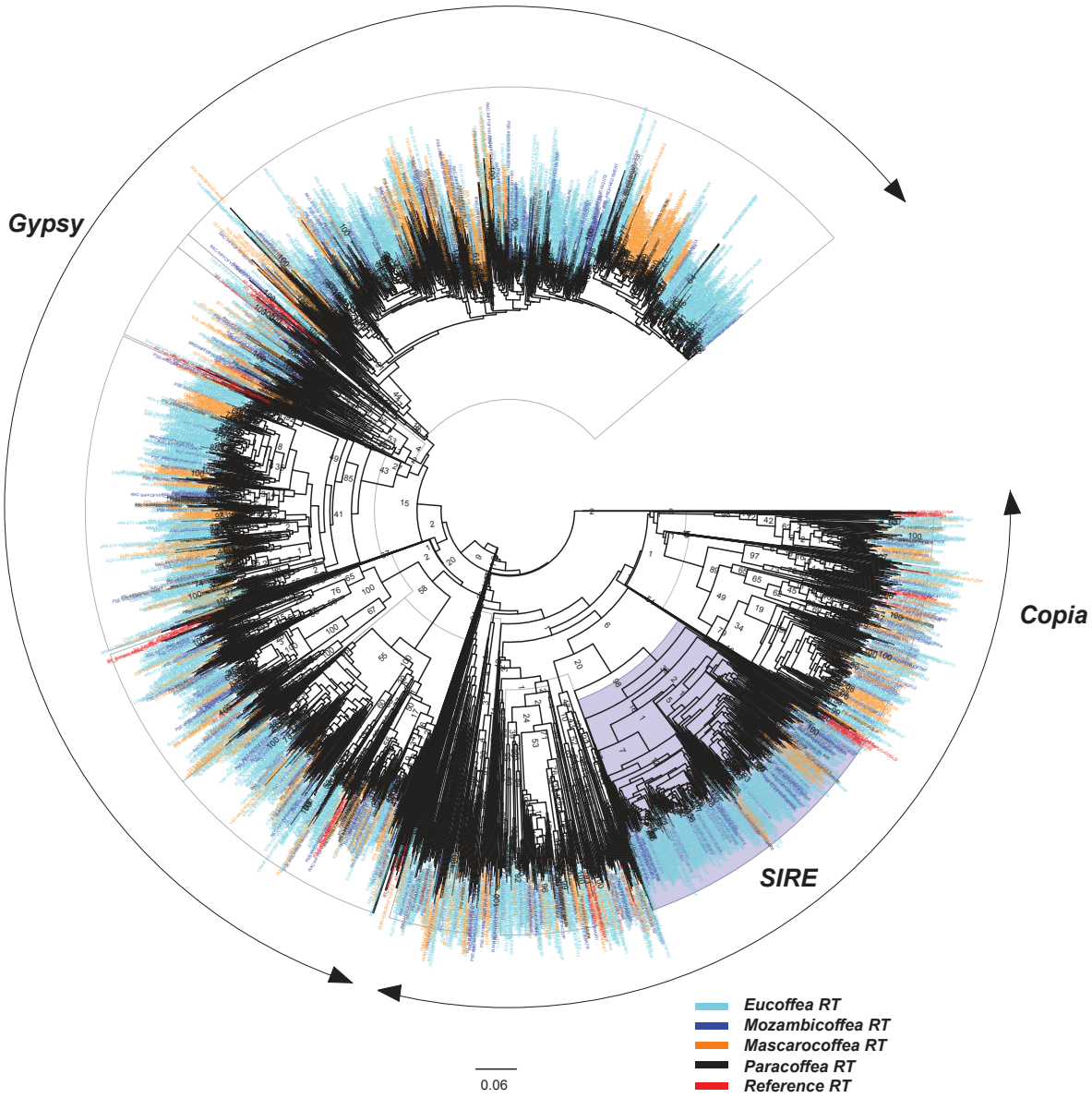


Supplemental data 7. Representation of the number of short microsatellites (repetition units of 1, 2 & 3 bases, in blue), and long microsatellites (repetition units of 4 to 20 bases, in red) in *Coffea* species. *C. arabica* (ET39), *C. canephora* (IF410, BUD15, DH200-94), *C. eugenioides* (DA56, OUG14), *C. charrieriana* (CHA), *C. heterocalyx* (HET), *C. pseudozanguebariae* (PSE), *C. racemosa* (RAC), *C. dolichophylla* (DOL), *C. tetragona* (TET), *C. humblotiana* (HUMB) and *C. horsfieldiana* (HOR).



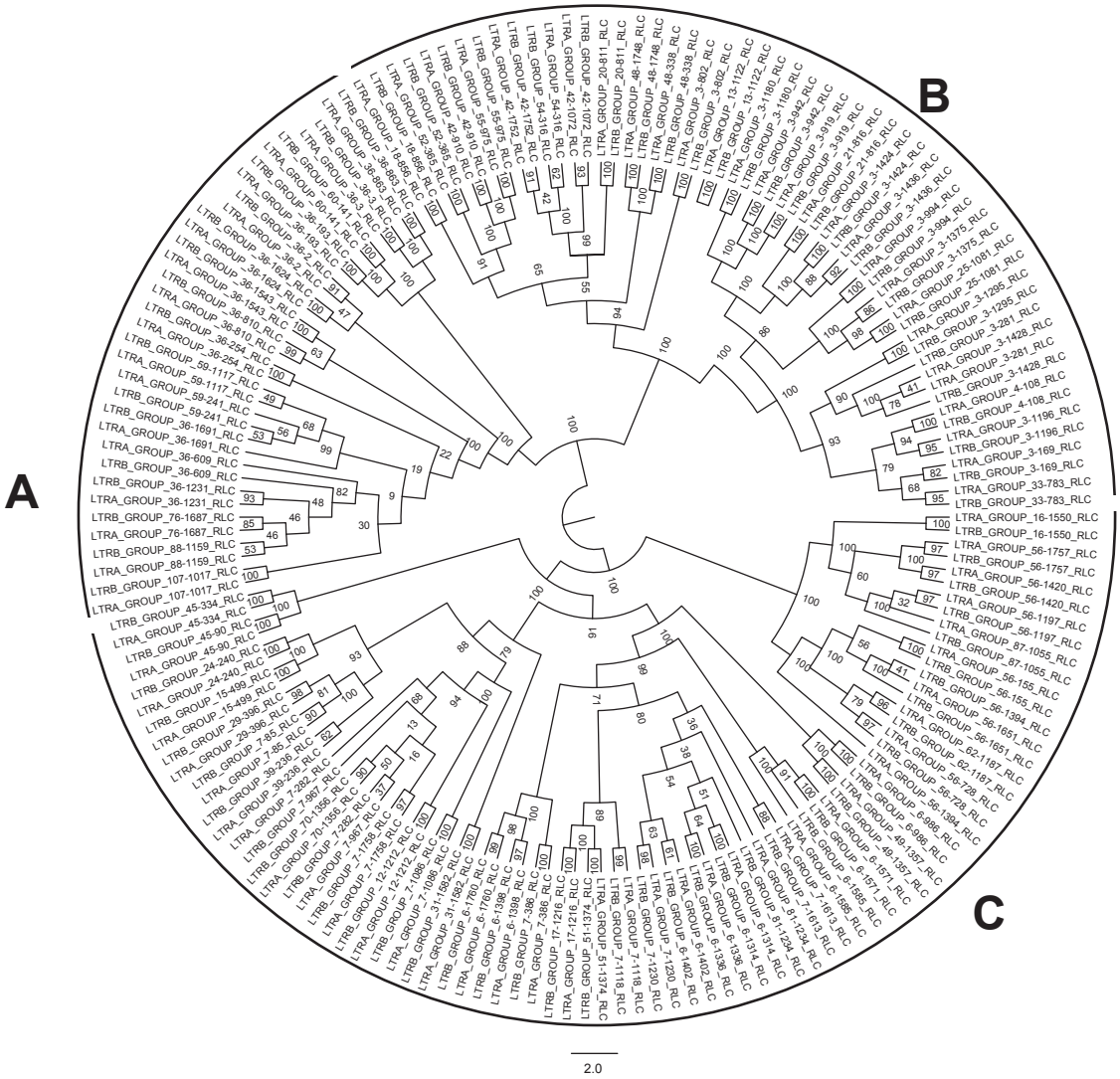
Supplemental data 8. Reverse transcriptase based phylogenetic analyses for 11 *Coffea* species and 14 accessions. 454 reads containing Reverse transcriptase domain for each species were filtered out and translated into amino acids.

RT domains with a minimum of 150 residues were conserved for multiple alignment (with muscle) and for a N-J tree analysis. In red reference domains from Gypsy DB, in light blue RT domains from Eucoffea species, in deep blue RT domains from Mozambicoffea, in orange RT domains from Mascarocoffea and in violet RT domains from Paracoffea. Clades were classified into known LTR retrotransposon lineages and super-families (*Gypsy* and *Copia*). The clade containing the SIRE lineage is colored in violet.

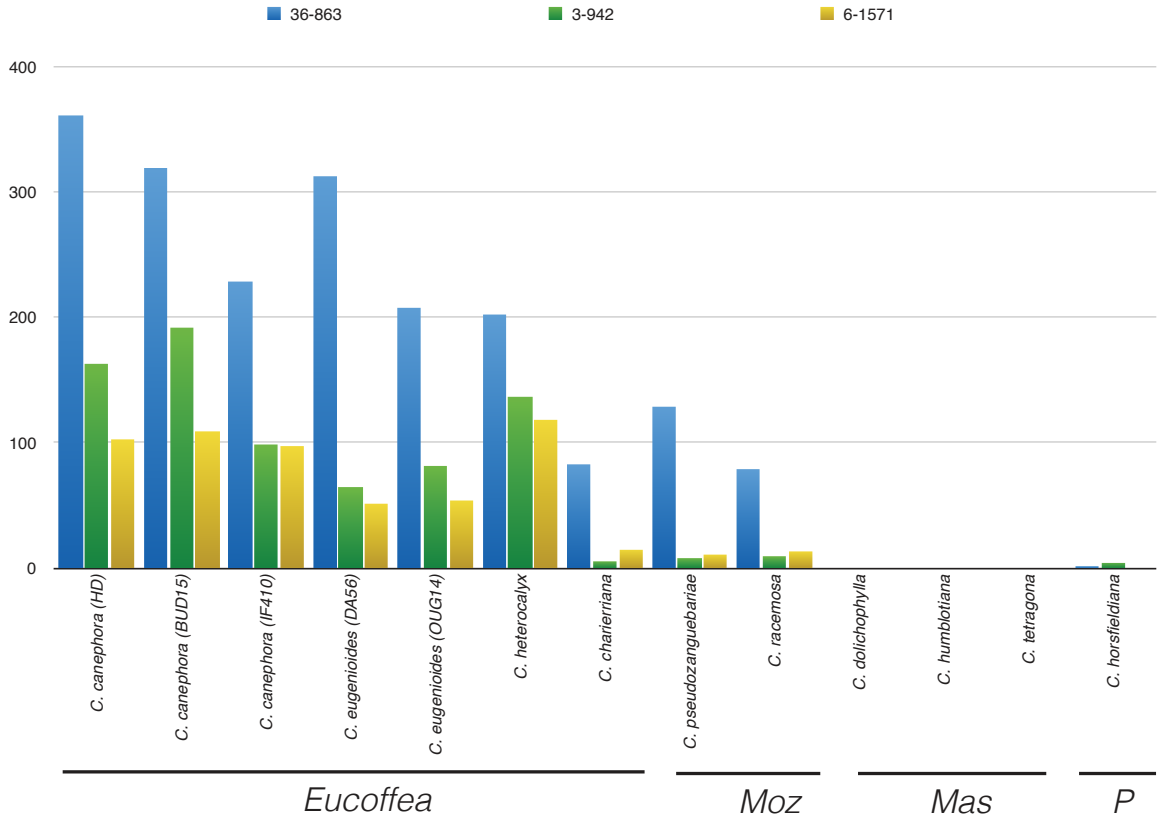


Supplemental data 9. LTR domain based phylogenetic analysis of 85 full-length SIRE elements found by LTR-STRUC in the *C. canephora* draft genome.

Sequences were classified into 3 clades A, B and C.



Supplemental data 10. Number of virtual copies of three SIRE families (36-863, clade A; 3-942, clade B; 6-1571, clade C); found in 454 dataset of 10 diploid Coffea species.



Supplemental data 11. Results of PCR amplification assay with four couple of primers and 62 *Coffea* species. (+ amplification, - no amplification, NA not performed).

Species	Accession ID	Country of Origin	Classification	3-942 ENV	3-942 LTR	36-863 LTR	6-1571 LTR
C. kapakata	OK	Angola	Eucoffea	+	NA	NA	NA
C. anthonyi	OD72	Cameroon	Eucoffea	+	NA	NA	NA
C. brevipes	JA54	Cameroon	Eucoffea	+	NA	NA	NA
C. charrieriana	OA22	Cameroon	Eucoffea	+	NA	NA	NA
C. montekupensis	OM11	Cameroon	Eucoffea	+	NA	NA	NA
C. sp. Nkoumbala	OI60	Cameroon	Eucoffea	+	NA	NA	NA
C. sp. Koto	EC61	Cameroon	Eucoffea	+	NA	NA	NA
C. sp. Congo	OB60	Republic of Congo	Eucoffea	+	NA	NA	NA
C. sp. Ngongo2	OF62	Republic of Congo	Eucoffea	+	NA	NA	NA
C. sp. Ngongo3	OG70	Republic of Congo	Eucoffea	+	NA	NA	NA
C. heterocalyx	JC62	Congo-Cameroon	Eucoffea	+	NA	NA	NA
C. eugenioides	DA58	Kenya	Eucoffea	+	NA	NA	NA
C. congensis	CB65	Central African Republic	Eucoffea	+	NA	NA	NA
C. liberica var dewevrei	EB51	Central African Republic	Eucoffea	+	NA	NA	NA
C. canephora	BA53	Republic Côte d'Ivoire	Eucoffea	+	NA	NA	NA
C. humilis	G57	Republic Côte d'Ivoire	Eucoffea	+	NA	NA	NA
C. liberica var liberica	EA64	Republic Côte d'Ivoire	Eucoffea	+	NA	NA	NA
C. stenophylla	FB55	Republic Côte d'Ivoire	Eucoffea	+	NA	NA	NA
C. pseudozanguebariae	H52	Kenya	Mozambicoffea	+	+	+	-
C. sessiliflora	PB58	Kenya	Mozambicoffea	-	-	-	-
C. racemosa	IB62	Mozambique	Mozambicoffea	-	-	-	-
C. salvatrix	LA60	Mozambique	Mozambicoffea	-	+	+	-
C. bridsoniae	APD2910	Tanzania	Mozambicoffea	-	-	-	-
C. kihansiensis	APD2922	Tanzania	Mozambicoffea	-	-	-	-
C. mongensis	EF11	Tanzania	Mozambicoffea	-	-	-	-
C. tricalysoides	APD4503	Madagascar	Mascarocoffea	-	-	-	-
C. abbayesii	A.601	Madagascar	Mascarocoffea	-	-	-	-
C. ankaranensis	A.525	Madagascar	Mascarocoffea	-	-	-	-
C. augagneuri	A.966	Madagascar	Mascarocoffea	-	-	-	-
C. bertrandii	A.5	Madagascar	Mascarocoffea	-	-	-	-
C. betamponensis	A.573	Madagascar	Mascarocoffea	-	-	-	-
C. boiviniana	A.980	Madagascar	Mascarocoffea	-	-	-	-
C. bonnieri	A.535	Madagascar	Mascarocoffea	-	-	-	-
C. coursiana	A.570	Madagascar	Mascarocoffea	-	-	-	-
C. dubardii	A.969	Madagascar	Mascarocoffea	-	-	-	-
C. farafanganensis	A.208	Madagascar	Mascarocoffea	-	-	-	-
C. heimii	A.516	Madagascar	Mascarocoffea	-	-	-	-
C. jumellei	A.974	Madagascar	Mascarocoffea	-	-	-	-
C. kianjavatensis	A.602	Madagascar	Mascarocoffea	-	-	-	-
C. lancifolia	A.320	Madagascar	Mascarocoffea	-	-	-	-
C. leroyi	A.315	Madagascar	Mascarocoffea	-	-	-	-
C. leroyi (ex andrambo	A.310	Madagascar	Mascarocoffea	-	-	-	-
C. liaudii	A.1013	Madagascar	Mascarocoffea	-	-	-	-
C. mcphersonii	A.977	Madagascar	Mascarocoffea	-	-	-	-
C. millotii	A.222	Madagascar	Mascarocoffea	-	-	-	-
C. millotii (ex dolichoph	A.206	Madagascar	Mascarocoffea	-	-	-	-
C. mogeneti	A.975	Madagascar	Mascarocoffea	-	-	-	-
C. montis-sacri	A.321	Madagascar	Mascarocoffea	-	-	-	-
C. perrieri	A.12	Madagascar	Mascarocoffea	-	-	-	-
C. ratsimamangae	A.528	Madagascar	Mascarocoffea	-	-	-	-
C. resinosa	A.8	Madagascar	Mascarocoffea	-	-	-	-
C. sakarahae	A.304	Madagascar	Mascarocoffea	-	-	-	-
C. tetragona	A.252	Madagascar	Mascarocoffea	-	-	-	-
C. toshii	A.1000	Madagascar	Mascarocoffea	-	-	-	-
C. tsirananae	A.515	Madagascar	Mascarocoffea	-	-	-	-
C. vatovavyensis	A.830	Madagascar	Mascarocoffea	-	-	-	-
C. mauritiana	BM17/25	Mauritius	Mascarocoffea	-	-	-	-
C. mauritiana	PCH	Réunion	Mascarocoffea	-	-	-	-
P. ebracteolatus	PSI11	Rép. Cote d'Ivoire	Psilanthus	-	+	-	-
P. lebrunianus	15/713	Congo	Psilanthus	-	-	-	-
P. travancorensis	PBT2	India	Psilanthus	-	-	-	-
P. wightianus	PBT3	India	Psilanthus	-	-	-	-

3. Conclusions et perspectives

Ces travaux ont permis d'établir que les ET présents dans différents génomes de caféiers n'expliquent pas à eux seuls les variations observées des tailles des génomes nucléaires, à l'exeption de *C. humblotiana* le plus petit génome des *Coffea* et *C. dolichophylla*, le plus gros génome des espèces malgaches. Basé sur une approche de similarité avec le répertoire des ET établi chez *C. canephora*, deux lignées de LTR-RT : les *SIRE* (*Copia*) et *Del* (*Gypsy*), montrent clairement des variations du nombre de lectures détectées entre les espèces d'Afrique de l'ouest et du centre, de l'est, des IOI et les Xeno-Coffea (ex-*Psilanthus*). Ces variations en fonction de l'origine géographique peuvent s'expliquer par une perte de sensibilité de la détection due à une grande divergence nucléotidique des éléments dans les génomes sauvages par rapport à *C. canephora* et aussi par une vraie variation du nombre des séquences. Ces variations peuvent être liées à la divergence des génomes pendant les mécanismes de spéciation ou d'adaptation.

L'approche par séquençage partiel n'apportant pas l'ensemble des détails attendus de la composition des génomes, puis l'émergence de nouvelles générations de séquençage (Illumina) à bas coût, ont encouragé le développement d'une stratégie de séquençage exhaustive afin de mieux comprendre la dynamique de lignées remarquables de LTR-RT.

Chapitre 4 - Caractérisation d'une famille de LTR-rétrotransposons peu connue, *Divo*, chez les caféiers

Article reçu le 26 décembre 2016, accepté le 7 mars 2017 et publié en ligne le 17 mars 2017 dans le journal *Molecular Genetics and Genomics*. DOI : 10.1007/s00438-017-1308-2.

1. Contexte

L'annotation des éléments transposables a été réalisée dans le génome de *Coffea canephora* (Denoeud et al. 2014). Les résultats de l'analyse phylogénétique des LTR-RT avec les domaines RT (Figure 12) ont montré un groupe monophylétique dans la superfamille *Copia* en dehors des groupes correspondant aux grandes lignées décrites dans la base de donnée GyDB (Llorens et al. 2011a). Afin d'éclaircir l'origine de ce groupe une nouvelle analyse a été conduite cette fois avec les génomes nucléaires de *C. canephora* (haploïde doublé HD 200-94), *C. arabica* (accession ET39, sauvage di-haploïde d'Ethiopie) et *C. eugenioides* (BU-A d'Ouganda) séquencés avec la technologie 'long read' PacBio (Rhoads and Au 2015) (The Arabica Coffee Genome Consortium 2014).

2. Implication personnelle

Les auteurs de l'article ont fourni les données de séquençage (D. Crouzillat et A. de Kochko), d'analyses RPKM des transcrits de *C. arabica* (R. F. de Souza) et participé à la rédaction de l'article. J'ai pour ma part réalisé les expérimentations spécifiées dans le Matériel et Méthodes ainsi que participé à la rédaction de l'article. Mes deux directeurs de thèse m'ont apporté leur aide concernant l'analyse des résultats et la rédaction de l'article.

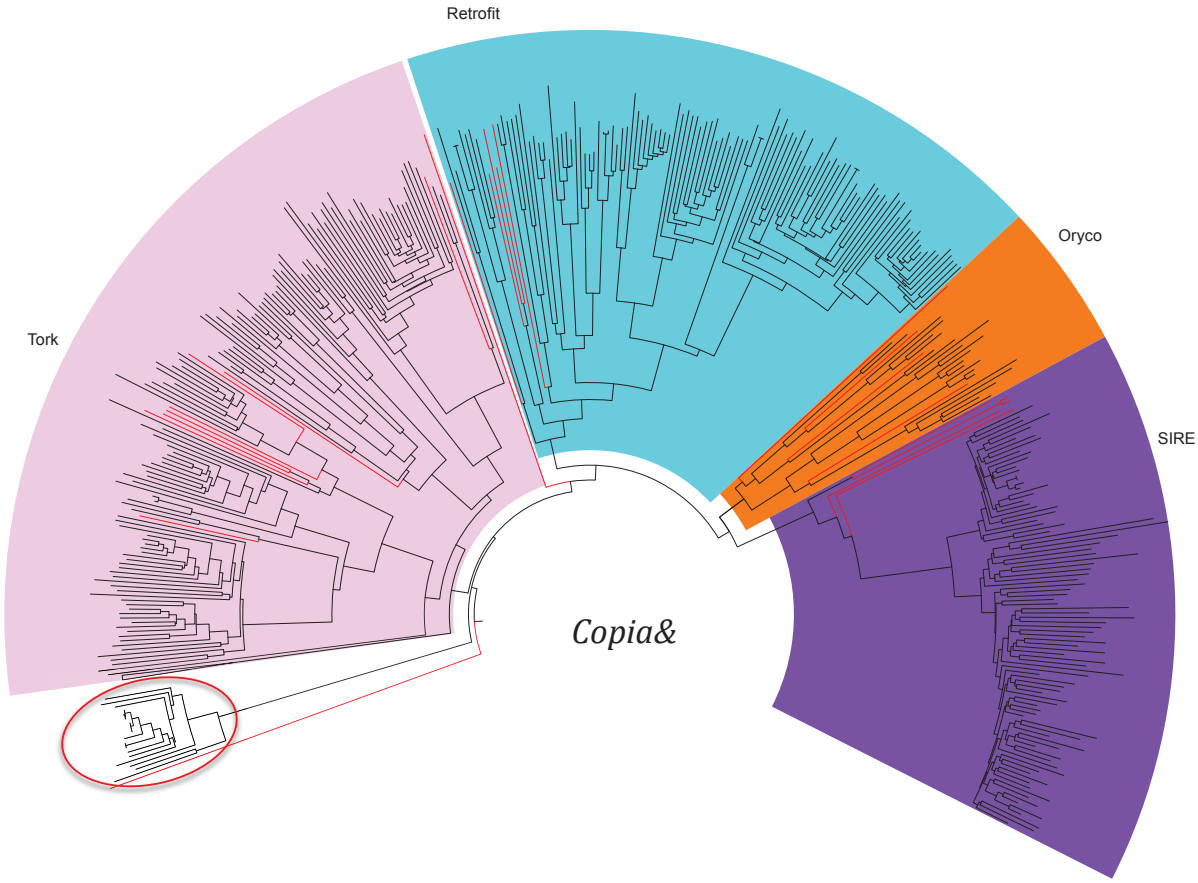


Figure 12 : Arbre en NJ basé sur le domaine RT des Copia de *C. canephora*. Chaque lignée et leurs RT de référence de GyDB (branches rouges) sont identifiées par une couleur différente. Le cercle rouge représente la lignée mal définie incluant l'élément *Divo*.

Mol Genet Genomics
DOI 10.1007/s00438-017-1308-2

ORIGINAL ARTICLE

Distribution of *Divo* in *Coffea* genomes, a poorly described family of angiosperm LTR-Retrotransposons

Mathilde Dupeyron[1,2] · Rogerio Fernandes de Souza[3] · Perla Hamon[1] · Alexandre de Kochko[1] · Dominique Crouzillat[4] · Emmanuel Couturon[1] · Douglas Silva Domingues[5] · Romain Guyot[2]

Received: 26 December 2016 / Accepted: 7 March 2017
© Springer-Verlag Berlin Heidelberg 2017

Abstract *Coffea arabica* (the Arabica coffee) is an allotetraploid species originating from a recent hybridization between two diploid species: *C. canephora* and *C. eugenioides*. Transposable elements can drive structural and functional variation during the process of hybridization and allopolyploid formation in plants. To learn more about the evolution of the *C. arabica* genome, we characterized and studied a new *Copia* LTR-Retrotransposon (LTR-RT) family in diploid and allotetraploid Coffea genomes called *Divo*. It is a complete and relatively compact LTR-RT element (~5 kb), carrying typical Gag and Pol *Copia* type domains. Reverse Trancriptase (RT) domain-based phylogeny demonstrated that *Divo* is a new and well-supported family in the *Bianca* lineage, but strictly restricted to dicotyledonous species. In *C. canephora*, *Divo* is expressed and showed a genomic distribution along gene rich and gene poor regions. The copy number, the molecular estimation of insertion time and the analysis at orthologous locations of insertions in diploid and allotetraploid coffee genomes suggest that *Divo* underwent a different and recent transposition activity in *C. arabica* and *C. canephora* when compared to *C. eugenioides*. The analysis of this novel LTR-RT family represents an important step toward uncovering the genome structure and evolution of *C. arabica* allotetraploid genome.

Keywords *Coffea* · *Copia* LTR-Retrotransposons · *Divo* · *Bianca* · Genomic evolution

Communicated by S. Hohmann.

Electronic supplementary material The online version of this article (doi:10.1007/s00438-017-1308-2) contains supplementary material, which is available to authorized users.

✉ Romain Guyot
romain.guyot@ird.fr

1 IRD UMR DIADE, EvoGec, BP 64501, 34394 Montpellier Cedex 5, France

2 IRD, CIRAD, Univ. Montpellier, IPME, BP 64501, 34394, Montpellier Cedex 5, France

3 Departamento de Biologia Geral, CCB, Universidade Estadual de Londrina,UEL, Londrina, Brazil

4 Nestlé R&D Tours, Notre-Dame d'Oé, Tours, France

5 Department of Botany, Instituto de Biociências, Universidade Estadual Paulista, UNESP, Rio Claro, Brazil

Introduction

Transposable elements (TEs) are mobile genetic elements representing the main components of numerous plant genomes such as rice (35%, International Rice Genome Sequencing Project 2005), grapevine (40%, The French–Italian Public Consortium for Grapevine Genome Characterization 2007), coffee-tree (*Coffea canephora* 50%, Denoeud et al. 2014), orchids (60%, Cai et al. 2015), tomato (60%, Mehra et al. 2015), bread wheat (80%, Brenchley et al. 2012), and maize (80–85%, Schnable et al. 2009). They have the capacity to move from one locus to another within genomes, and for some of them to increase their copy numbers by doing so. Recently, it has been suggested that TEs may also propagate via horizontal transfer mechanisms among genomes of different species or even genera (Feschotte and Pritham 2007; Schaack et al. 2010; Fedoroff 2012; Dias et al. 2015; Gilbert et al. 2016; Lin et al. 2016; Panaud 2016). TEs are also considered as remarkable genome evolution drivers allowing genome adaptation and innovation through chromosome rearrangements, gene expression alterations and sometimes,

Published online: 17 March 2017

generation of new gene functions via molecular domestication of TE domains (Feschotte and Pritham 2007; Fontana 2010). During the allopolyploidy processes, TEs may represent the most dynamic fraction of the genome with major changes in their copy numbers (Parisod et al. 2010).

The faculty of producing large amount of genomic and transcriptomic sequencing data, and the availability of whole-genome sequence data, have promoted the development of bioinformatics tools to identify and to analyze genome components, including TEs (Lerat 2010). The large diversity of TEs led the scientific community to define a hierarchical classification, first separating elements according to their mode of mobility into retrotransposons, or Class 1 elements, and DNA transposons, or Class 2 elements. These classes were further subdivided into orders, super-families, lineages, and families according to their structural features and similarities (Wicker et al. 2007).

Among the class 1 elements, LTR-Retrotransposons (LTR-RTs) are the most abundant TEs in plant genomes. They represent a wide fraction of genomes ranged between 14% in *Arabidopsis thaliana* (The Arabidopsis Genome Initiative 2000), up to 75% in maize (Schnable et al. 2009). LTR-RTs are divided into two super-families: *Copia* and *Gypsy* that differ mainly in their internal coding regions order (Wicker et al. 2007). *Copia* and *Gypsy* are composed of ancient and conserved lineages in plants (Wicker and Keller 2007) that can be phylogenetically classified based on their RT domain (Eickbush and Jamburuthugoda 2007). *Copia* and *Gypsy* LTR-RT may occupy different chromosomal locations as demonstrated by the available sequences of plant genomes (The Arabidopsis Genome Initiative 2000; International Rice Genome Sequencing Project 2005; The French–Italian Public Consortium for Grapevine Genome Characterization 2007; Paterson et al. 2009).

The recently released genome of *C. canephora* also contains an important fraction of LTR-RTs of 42% (Denoeud et al. 2014). The *Gypsy* elements clearly outnumber the *Copia* with 24.1% and 6.8% of the genome sequence, respectively. The remaining 11% is composed of unclassified LTR-RTs and classes small in number like *Bel*/*Pao*, Caulimoviruses, Retroviridae.

Coffea genus belongs to the Rubiaceae family. It contains 124 described species, originating from the intertropic forests of Africa, western Indian Ocean islands, India, Tropical and SouthEast Asia, and Australasia (Davis et al. 2011). All species are diploids with $2n = 2x = 22$ chromosomes (Bouharmont 1959; Louarn 1976), with the exception of the allotetraploid *C. arabica*, one of the two major cultivated species (Carvalho 1952). *C. arabica* has a recent origin (Yu et al. 2011), arising from hybridization between two wild diploid species: *C. canephora*, the other cultivated species (known as Robusta) and *C. eugenioides*, an East African wild species (Lashermes et al. 1999).

Previously, the two first LTR-RT elements identified in sequenced *C. canephora* Bacterial Artificial Clones (BAC) were called *Nana* and *Divo*. They were used to perform RBIP (retrotransposon-based insertion polymorphism) and REMAP (retrotransposon-microsatellite amplified polymorphism) analyses to study the species relationships within *Coffea*. *Divo* was particularly efficient at a low taxonomic level to resolve the genetic diversity within *C. canephora*, suggesting that the mobility of the *Divo* family participated to the *C. canephora* differentiation (Hamon et al. 2011).

In this study, we describe a genomic overview of the *Divo* family, from the *Bianca* lineage, in *C. arabica* and its two diploid progenitors, *C. canephora* and *C. eugenioides*. The *Bianca* lineage has been described in barley, Arabidopsis, and rice (Wicker and Keller 2007) and mentionned in other plant species (Kolano et al. 2013; Marcon et al. 2015; Yin et al. 2015). *Matita*, an element belonging to the *Bianca* lineage, was described more deeply in *Arachis hypogaea*, the cultivated allotetraploid peanut (Nielen et al. 2012). *Matita* appears to be present in peanut genome for a long time, as its insertions have been dated around 3,5 Mya. Its chromosomal distribution has been investigated by FISH experiments, which showed its presence mainly in distal regions of all the chromosomes. The annotated copies did not contain ORFs (stop codons and frameshifts in the putative coding regions) so the potential activity or non-activity of *Matita* has not been studied (Nielen et al. 2012). Since few data or characterizations of LTR-RTs from the *Bianca* lineage are available so far in plants, except in cultivated peanut, we selected this lineage, represented by the family *Divo* in coffee-trees, for further characterization of LTR-RT families in *Coffea*. *Divo* have a relatively short size (5 kb) and a moderated copy number. A RT domain-based phylogenetic analysis demonstrated that *Divo* belongs to the dicotyledonous section of the poorly known *Bianca* lineage. These elements are expressed and quite evenly distributed in the *C. canephora* genome. Differences in the abundance and in the insertion chronology of *Divo* elements were observed among *C. canephora, C. arabica*, and *C. eugenioides* genomes, suggesting different dynamics and impact on diploid and allotetraploid genomes structural evolution.

Materials and methods

Genomic sources

A total of four coffee genome sequences were used in this study: *C. canephora* DH 200−94 (Denoeud et al. 2014), accounting for 568 Mb of scaffolds and assembled into pseudo-molecules, including chromosome 0 (representing

80% of the estimated genome size i.e., 710 Mb); and three genomes sequenced with the single molecule real-time (SMRT, Pacific Biosciences—PacBio) sequencing technology: *C. canephora* (accession DH 200−94), *C. arabica* (accession Et39), and *C. eugenioides* (BU-A) accounting respectively for 679, 1060, and 789 Mb of unordered contigs. The *C. canephora, C. arabica*, and *C. eugenioides* PacBio genome sequences were generated under the Arabica Coffee Genome Consortium (ACGC 2014).

Identification, classification and annotation of LTR-RTs in *C. canephora, C. arabica* and *C. eugenioides* genomes

Potential LTR-RTs were de novo identified using the LTR_STRUC (McCarthy and McDonald 2003) algorithm against the *C. canephora* published genome, and the *C. canephora, C. arabica*, and *C. eugenioides* PacBio genomes. The predicted elements were classified into *Copia* and *Gypsy* super-families according to BLASTX similarities (Altschul et al. 1990) against a database of Gag and Pol domains (available at GyDB, http://www.gydb.org/ Llorens et al. 2011). LTR-RT predicted elements showing no similarity with any GyDB domain were not retained for further analyses.

Reverse transcriptase-based classification of LTR-RTs

The amino-acid RT domain of all LTR-RTs recovered with LTR_STRUC from each genome was extracted as described in Guyot et al. (2016), with a minimum length of 150 amino-acid residues. RT reference domains from GyDB were added to them to understand *Coffea* LTR-RTs affiliations in the *Copia* lineage. Aligned sequences were used to construct a bootstrapped neighbor-joining (NJ) tree (100 bootstrap replicates) edited with Figtree (http://tree.bio.ed.ac.uk/software/figtree/).

Classification, annotation and characterization of the *Bianca* lineage and *Divo* LTR-RT family

The coffee LTR-RTs seqences from the *Bianca* lineage were compared to known LTR-RTs from *C. canephora* and all elements from the *Bianca* lineage in plants (Wicker et al. 2007) using BLASTN. Sequences similar to *Divo*, a previously identified LTR-RT from *C. canephora* (Hamon et al. 2011) were compared using dot-plot (Sonnhammer and Durbin 1996). To search for *Divo* similar elements in publicly available plant genomes, the sequence fragment of *Divo* described in Hamon et al. (2011) (NCBI accession HM755952.1) was used as query for similarity searches on the NCBI website (http://blast.ncbi.nlm.nih.gov/), using a BLASTX and BLASTN e-value cut-off of $1e^{-100}$ and a minimum of 50% of identity over 50% of the query sequence length. Recovered elements were annotated using BLASTX and dot-plot alignments with reference domains (Gypsy Database 2.0 web site) (Sonnhammer and Durbin 1996) and LTR_Finder (Xu and Wang 2007, http://tlife.fudan.edu.cn/ltr_finder/). Final annotations were edited with Artemis (Rutherford et al. 2000). Annotated elements were used for another phylogenetic analysis based on RT amino-acid domains as described in the previous paragraph.

Search for Divo elements in plant genomes

We searched for *Divo* LTR-RTs similar sequences in transposable elements dedicated databases: RepBase (http://www.girinst.org/, (Bao et al. 2015)), the Plant Repeat Database (http://plantrepeats.plantbiology.msu.edu, Ouyang and Buell 2004), and RetrOryza (http://retroryza.fr, Chaparro et al. 2007) using BLASTN. To better understand the evolution of the *Divo* family and its relationships with the *Bianca* lineage, we searched for sequences similar to *Divo* in eukaryote publicly available genome sequences using BLASTN and BLASTX (evalue $< e^{-100}$), using four *Divo* sequences from *C. canephora* (Denoeud et al. 2014 and PacBio), *C. arabica*, and *C. eugenioides* (accessions #: KX767840, KX767841, KX767839 and KX767842). 22 genomic sequences were recovered from 14 angiosperm species and their RT amino-acid domains were used to construct a NJ phylogenetic tree (*Oryza sativa*—accession #AC147802.2, *A. thaliana*—#AP002459, *V. vinifera*—#AM477556.1, *Sorghum bicolor*—#AF466199.1, *Zea mays*—#DQ493648.1, *Rosa rugosa*—#JQ791545.1, *Theobroma cacao* (Jurka 2014—accession #HQ244500), *Fragaria vesca*—#XM_004309244.1, *Ipomoea trifida*—#AY4480105.1, *Beta vulgaris*—#GU057342.1, *Arachis hypogaea*—#HQ637177.1, Oryza rufipogon—#FO681399, *Solanum lycopersicum*—#AAK84483, *M. truncatula*—#CM001223. Additional LTR_RT sequences from TAIR and RetrOryza database and LTR_STRUC output for *A. thaliana, V. vinifera*, and *O. sativa*).

Divo homologous elements were also specifically searched for and characterized from two reference plant genomes: *Arabidopsis thaliana* (GCA_000001735.1) and *Vitis vinifera* (GCA_000003745.2) available from TAIR (https://www.arabidopsis.org) and NCBI (http://www.ncbi.nlm.nih.gov/). First, all potential full-length LTR-RTs were de novo searched with LTR_STRUC and compared by BLASTN with *Divo* elements identified previously. Second, all LTR-RT sequences from Arabidopsis and grapevine previously identified and available in the Plant Repeat Database (http://plantrepeats.plantbiology.msu.edu/search.html) were downloaded and compared by BLASTN with coffee *Divo* elements.

Copy number and insertion time of *Divo* in *C. canephora, C. arabica*, and *C. eugenioides*

Assessment of *Divo* copy number in of *C. canephora* (Denoeud et al. 2014) and the *C. canephora, C. arabica*, and *C. eugenioides* PacBio genomes (ACGC 2014) was carried out with Censor (Kohany et al. 2006). A complete *Divo* element is considered when it contains both ORFs Gag and Pol and a minimum of 99% sequence identity between both LTRs. Such a sequence was found in the *C. canephora* genome and was used as a reference for similarity searches (accession number #KX767841). A copy is considered if it covers a minimum of 80% of the reference sequence with at least 80% of nucleotide identity (Wicker et al. 2007) and a fragmented copy is considered if it covers a minimum of 20% of the reference sequence with at least 80% of nucleotide identity. Full-length copies were also extracted according to the following definition: 80% of nucleotide identity over 100% of the reference sequence length as well as potential solo LTRs (80% of identity over 100% of the LTR sequence length). The genomic distribution of the identified elements in the *C. canephora* pseudochromosomes was established using Circos (Krzywinski et al. 2009).

The insertion time of full-length *Divo* copies was estimated based on the divergence of the 5′- and 3′-LTR sequences of each identified full-length copy. The two LTRs were aligned using Stretcher (EMBOSS), and the divergence (K) was calculated using the Kimura 2-parameter method implemented in Distmat (EMBOSS). The insertion dates (T) were estimated using the formula $T = K/2r$ (SanMiguel et al. 1998) where we used average base substitution rates (r) of $1.3e^{-8}$ established by Ma & Bennetzen (2004).

Presence of *Divo* at orthologous locations in three coffee-trees genomes

Insertion of full-length copies of *Divo* in *C. canephora, C. eugenioides* and *C. arabica* at orthologous locations among the three genomes were compared. As a first step, genomic regions containing full-length *Divo* copies were recovered from the *C. canephora* contigs adding 2 kb upstream and downstream the element. The recovered genomic fragments are then compared as queries using BLASTN (evalue $1e^{-100}$) against the other two genomes. The best results (lowest e-values and highest scores) are then extracted and compared to the queries using dot-plot alignements (Sonnhammer and Durbin 1996). Finally, dot-plot alignments are manually evaluated to classify the orthologous relationships into the following categories: (i) queries are not conserved and so no orthologous regions could be identified; (ii) queries are conserved within an orthologous region but the *Divo* element is not conserved, and (iii) queries are conserved within an orthologous region and the *Divo* element is present at the same insertion site. These steps are repeated for the full-length copies of *Divo* in *C. arabica* and *C. eugenioides*.

Search for *Divo* potential expression in *C. canephora* tissues

RNA sequencing (RNA-seq) data generated under the *C. canephora* genome project (Denoeud et al. 2014) from leaves, roots (*C. canephora* accession #T3518), stamen, and pistil (*C. canephora* accession #BP961) were used to identify the transcriptional pattern of reference sequences. The 130.10^6 RNA-Seq reads were cleaned using prinseq (Schmieder and Edwards 2011) and mapped against 18 *Divo* sequences using Bowtie 2 (Langmead and Salzberg 2012). The number of mapped reads per TE sequence was processed and RPKM (reads per kilo base per million) were calculated. A heatmap representing the expression profiles was computed using Heatmap3 package in RStudio (2012). Differential expression among available RNA-seq libraries was detected using Winflat (Audic and Claverie 1997) with significance threshold of 0.05 and Bonferroni correction. This analysis was performed with IDEG6 software (http://telethon.bio.unipd.it/bioinfo/IDEG6_form/) (Romualdi et al. 2003).

Results

Copia LTR-RTs in *C. canephora, C. arabica* and *C. eugenioides* genomes

Since LTR-RTs represent the main part of the TE fraction found in the *C. canephora* genome, we focused our analyses on these elements, and more specifically on *Copia* LTR-RT lineages and families. LTR_STRUC identified 1799 (588 *Gypsy* and 474 *Copia*), 7363 (2010 *Gypsy* and 999 *Copia*), 4346 (2153 *Gypsy* and 1080 *Copia*) and 3591 (1632 *Gypsy* and 913 *Copia*) LTR-RT elements, for *C. canephora* (Denoeud et al. 2014), *and C. canephora, C. arabica*, and *C. eugenioides* (ACGC), respectively. We specifically screened and filtered out LTR_STRUC potentially complete elements according to similarities with the *Copia*-specific domains. The reverse transcriptase (RT) amino-acid domains of *Copia* recovered sequences were extracted and used for a NJ phylogenetic analysis. The analysis of the resulting NJ trees for *C. canephora, C. arabica*, and *C. eugenioides* shows that coffee RT *Copia* domains were classified into all five *Copia* lineages previously described in plants: *Tork, Oryco, SIRE, Retrofit*, and *Bianca* (Llorens et al. 2009; Wicker and

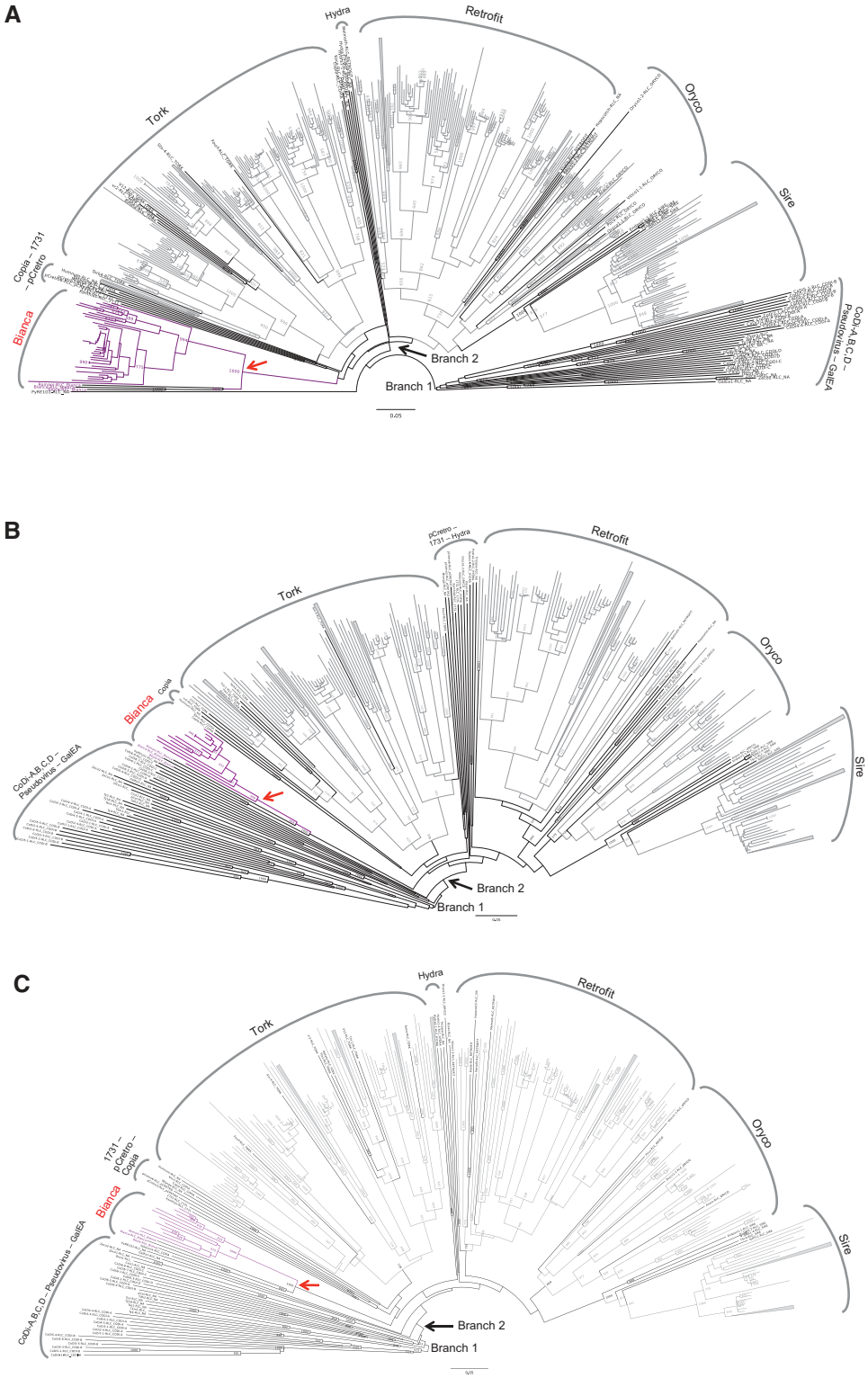


Fig. 1 Phylogenetic analysis of LTR retrotransposons sequences predicted from *C. canephora* (**A**), *C. arabica* (**B**) and *C. eugenioides* (**C**) genomes. Phylogenetic trees were based on amino-acid alignments of the reverse transcriptase (RT) domains; 999, 1080, and 913 amino acids, repectively, from *C. canephora, C. eugenioides*, and *C. arabica* genomes. The classification into lineages was done according to the RT reference domains (*black* branches) downloaded from GyDB. The *Coffea* sequences within the *Bianca* lineage are indicated by a *red arrow*, and lineages are indicated by brackets and names

Keller 2007, Fig. 1). References RT domains from other organisms Diatoms (*CoDI*), Fungi (*Pseudovirus, pCretro*) and Arthropoda (*1731, Hemivirus*) were found clustered outside of plant lineages that include coffee, according to their classification into Branch 1 and 2 (Llorens et al. 2009). The diversity of *Copia* lineages appears very similar between the three species analyzed (Fig. 1). One of the smallest clades called *Bianca* and supported by strong bootstraps (Fig. 1), grouped together 12 sequences from *C. canephora* (Denoeud et al. 2014 and 13 sequences in

PacBio genome), 14 from *C. arabica*, and 12 from *C. eugenioides*.

Divo elements in C. *canephora, C. arabica*, and *C. eugenioides* genomes

In total, 89 full-length elements belonging to the *Bianca* lineage and recognized by LTR_STRUC or BLASTN in the four coffee genome sequences (Supplemental Data 1) were analyzed and annotated. The structure of these elements corresponds to the typical organization of *Copia* elements with two LTRs at each extremity and two ORFs: Gag and Pol containing the protease (PR), integrase (INT), reverse transcriptase (RT), and RNase H (RH) domains, in this specific order. The LTRs were 350 bp long and were terminated with the LTR consensus: 5′TG…CA3′. The overall length of complete elements (i.e., elements carrying two highly conserved LTRs and complete Gag and Pol ORFs) ranged between 5276 bp and 5636 bp (Fig. 2a). The Gag sequence (1065 bp long, separated from Pol by 5 stop codons in all the elements found) presented similarities with the UBN2 family domain (Pfam14223—nucleotide position 718–942). UBN2 is a form of the peptide encoded by the Gag ORF frequently found in the *Copia* LTR-RT superfamily. A Zinc finger amino-acid motif (ZnF_C2HC, nucleotide position 1318–1365), involved in nucleic acids binding, is also found in the peptide encoded by this ORF. The Pol ORF (3501 bp), showed high similarities with Gag_pre-integrase family (Pfam13976, position 2050–2268), Integrase (INT) core domain (Pfam00665, position 2305–2655), Reverse transcriptase (RT) genes (Pfam 07727, position 4645–5085), and RNase-H (RH) domain (position 3637–4374), in this specific order. All these domains show high similarities with *Copia* LTR-RTs.

While the polypurine track motif (PPT, used for the synthesis of the complementary DNA strand) is found upstream the 3′ LTR, the primer-binding site (PBS) presents unusual sequence conservation (Fig. 2b). Among the 89 elements precisely analyzed here, only one (Accession #KX767840) showed a complementary sequence to a tRNA ($tRNA^{Ile}$ (AAT)).

Similarity analyses between coffee full-length elements belonging to the *Bianca* lineage, known *Bianca* elements (Wicker et al. 2007) and known coffee elements showed a relatively good nucleotide conservation with *Divo*, a *Copia* LTR-RT element identified earlier in a *C. canephora* BAC sequence and used to assess insertion site polymorphism (Hamon et al. 2011). Comparisons between *Divo* and a complete and potentially active element in *C. canephora* revealed by LTR_STRUC (Accession #KX767841) indicated an overall percentage of nucleotide identity of 63.6% and a LTR percentage of identity of 58% and 56.7% for the 5′ and 3′ LTR, respectively. This relatively low percentage of nucleotide identity is probably due to the absence of several regions of the *Divo* element identified earlier (Hamon et al. 2011; Fig. 2c). This percentage is similar for all full-length coffee-trees elements. Nevertheless, we named the novel annotated sequences, carrying the new group of RT domains similarly to the initial element discovered earlier: *Divo*. A reference *Divo* element was ascertained for each of the three *Coffea* genomes, based on the most conserved annotated sequence found. These references were used for different analyses when they needed a reference sequence. All the recovered sequences of *Divo* presenting a good conservation and no stop codon in the RT domain were used in RT-based phylogenies, which confirmed their affiliations to the *Bianca* lineage and the *Divo* family (Supplemental data 2).

We also searched for the transcriptional pattern of the *Divo* family using RNAseq reads (Denoeud et al. 2014) from leaves, roots, stamen, and pistil mapped on the 18 *Divo* sequences found in *C. canephora* published genome with LTR_STRUC. Transcriptional pattern suggested transcriptional modulation when vegetative tissues (leaves or roots) are compared to reproductive tissues (stamens or pistils). Seventeen *Divo* exhibited differential expression between leaves or roots versus stamen or pistil, while only seven presented differential expression between leaves and roots and none between pistils and stamen. In addition, a lower degree of expression of these retrotransposons was detected in pistil and stamen when compared to leaves and roots (Supplemental data 3).

Copy number estimation and insertion time of *Divo* elements in *C. canephora, C. arabica* and *C. eugenioides*

One hundred and ninteen, 204, and 132 copies of *Divo* were, respectively, found in *C. canephora* (Denoeud et al. 2014), *C. canephora, C. arabica*, and *C. eugenioides* ACGC sequences (Table 1). Besides looking for highly conserved copies (100% of coverage and ≥80% of identity), less conserved or fragmented copies (80% of identity on at least 20% of the total length) and solo LTRs (Devos et al. 2002) were also detected. Higher copy numbers were obtained for *C. canephora* ACGC sequences, probably due to the completeness of the sequencing technology used. Interestingly, *C. eugenioides* showed a higher *Divo* total copy number when compared to *C. canephora*, but with the notable exception of full-length copies. The allotetraploid genome of *C. arabica* contains the highest total *Divo* copy number. However, for each category, the number of copies in *C. arabica* is lower than the sum of its diploid progenitors. The ratio of solo LTR to full-length or "intact" elements was in a similar order of magnitude for *C. canephora* (4.7:1 and 3.4:1) and *C. arabica* (5.4:1), but three times higher for *C. eugenioides* (16.8:1). In the annotated *C.*

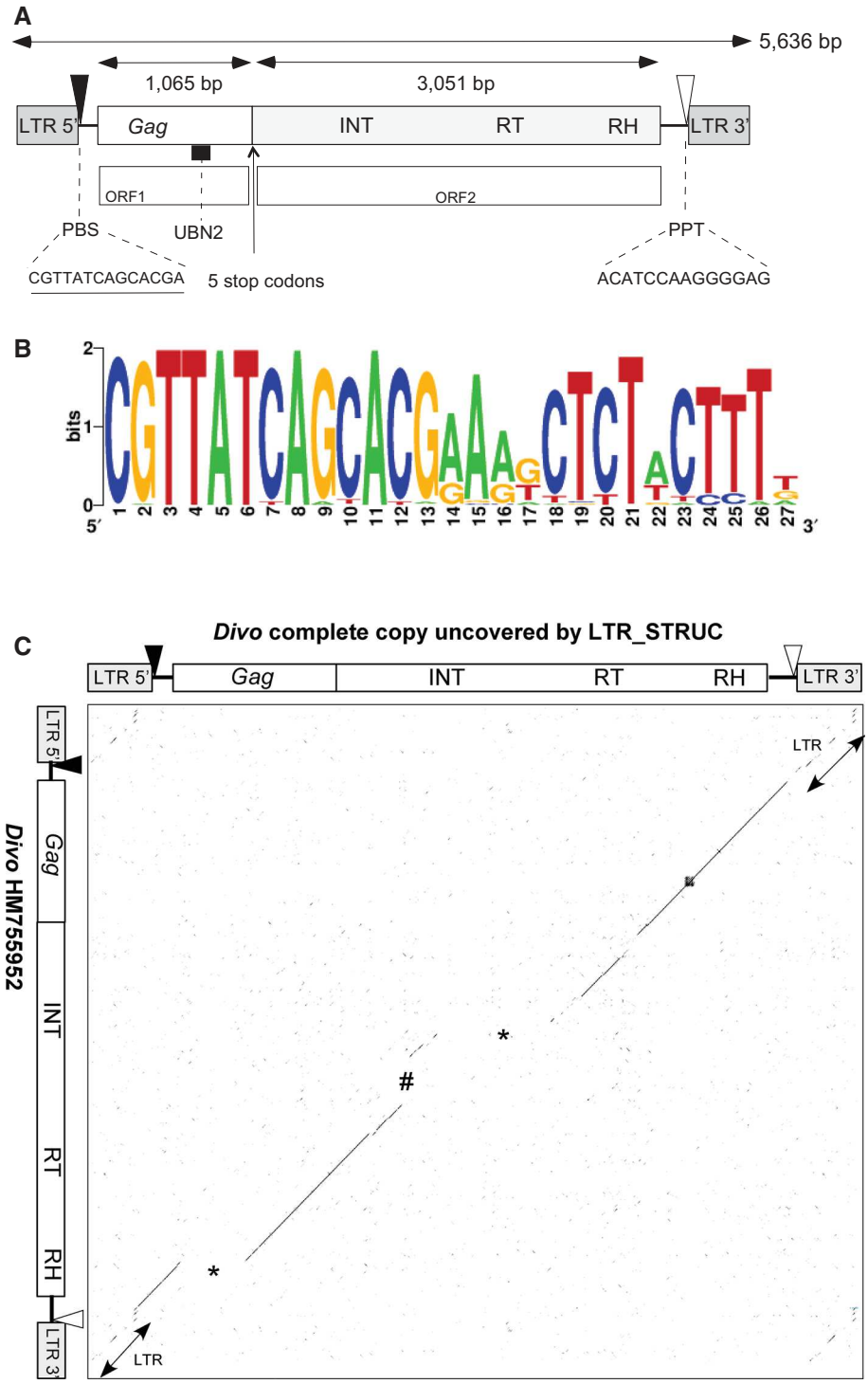


Fig. 2 Structure of the *Copia* LTR-RT *Divo*. **a** Structural features of the *Divo* family. The complete *Divo* element was identified in *C. cenephora* genome (KX767841). *Gag* and *Pol* ORFs are separated by five stop codons. *LTR* long terminal repeats, *PBS* primer-binding site (*black triangle*), *PPT* polypurine tract (*open triangle*), *UBN2* ubinuclein 2 domain, *INT* integrase, *RT* reverse transcriptase, *RH* RNAse H. **b** Web-Logo representation of the PBS of *Divo* full-length copies found in (**c**) *canephora* and *C. arabica*. **c** *Dotter* alignment between the fragmented *Divo (*Hamon et al. 2011, HM755952) and a complete *Divo* element uncovered by LTR_STRUC in *C. canephora* (KX767840). *Asterisks* regions absent in *Divo* but present in the complete element. **#** Regions present in *Divo* but absent in the complete element. The positions of LTR are indicated

canephora pseudo-molecules, the *Divo* family, whatever the status of the copy (full-length, "80–80," fragmented or solo LTR), appears equally distributed along TE-rich and gene-rich regions (Supplemental data 4).

Complete copies LTR sequences (80–100%) were used to calculate their nucleotide divergence and estimate their insertion times in *C. canephora, C. arabica*, and *C. eugenioides* according to the substitution rate established by SanMiguel et al. 1998 (Fig. 3). Our analysis indicates relatively recent insertions of *Divo* in *C. canephora* and in *C. arabica* (at 0–0.5 Mya), while in *C. eugenioides*, two more ancient peaks (at 0.5–1 and 1.5–2 Mya, red line) are detected. Interestingly, the second ancient peak observed in *C. eugenioides* is also detected in a lesser extent in *C.*

Table 1 Estimation of the copy numbers of *Divo* elements in the *C. canephora* genome (*, Denoeud et al. 2014) and *C. canephora, C. arabica*, and *C. eugenioides* genome sequences (§, PacBio)

	Number of intact copies (80–100)	Number of copies (80–80)	Number of partial copies (20–80)	Number of solo LTRs	Solo LTR/ intact copies ratio	Total
*C. canephora**	28	119	199	132	4.7:1	478
C. canephora§	41	129	212	142	3.4:1	524
C. arabica§	37	204	351	201	5.4:1	793
C. eugenioides§	20	132	223	336	16.8:1	711

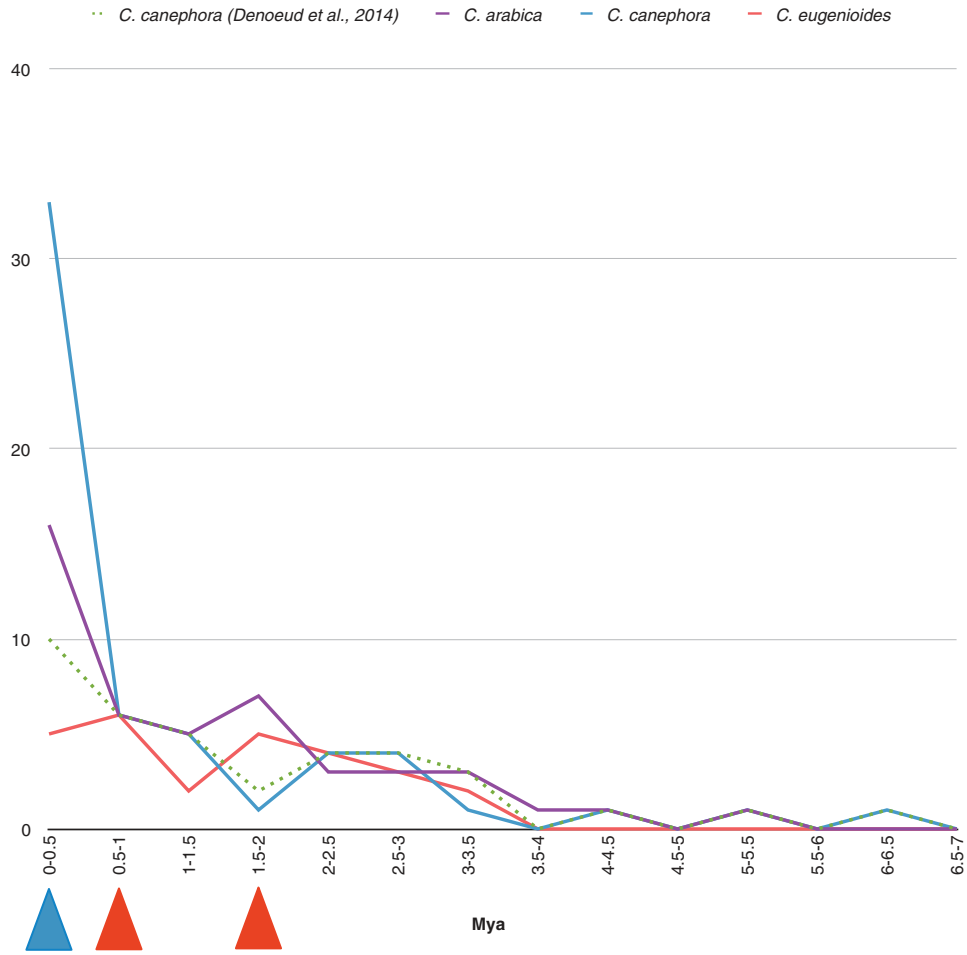


Fig. 3 Estimation of insertion times of *Divo* elements in coffee genome sequences. The LTR sequences of 178 full-length elements uncovered from *C. canephora, C. arabica*, and *C. eugenioides* genomes were used to estimate insertion time using the substitution rate of 1.3×10^{-8} (Ma and Bennetzen 2004). *Blue, red, purple*, and *green lines* represent insertion times respectively in *C. canephora, C. eugenioides, C. arabica*, and *C. canephora* (Denoeud et al. 2014)

arabica (purple line), showing a good conservation of copies from the *C. eugenioides* parental genome in the allotetraploid.

Comparison of orthologous regions of full-length *Divo* insertions between *C. canephora, C. arabica* and *C. eugenioides*

Insertion sites of 39, 37, and 20 *Divo* full-length copies were mined in *C. canephora, C. arabica*, and *C. eugenioides* genomes, respectively, with their location given by Censor (Kohany et al. 2006). 31 specific insertion sites are represented by blue, purple, and red squares, respectively (Fig. 4). In orthologous regions, 16 copy sites are shared between *C. canephora* and *C. arabica* (blue stars), six between *C. arabica* and *C. eugenioides* (red stars), and one between *C. canephora* and *C. eugenioides* (blue and red stars at 0,7 My). Twenty-four copy sites are shared between the three genomes (yellow circles). Copy sites shared between the three genomes are dated from 0.8×10^6 up to 3.2×10^6 years. In *C. canephora*, specific copy insertions are dated from 0 to 3.1×10^6 years.

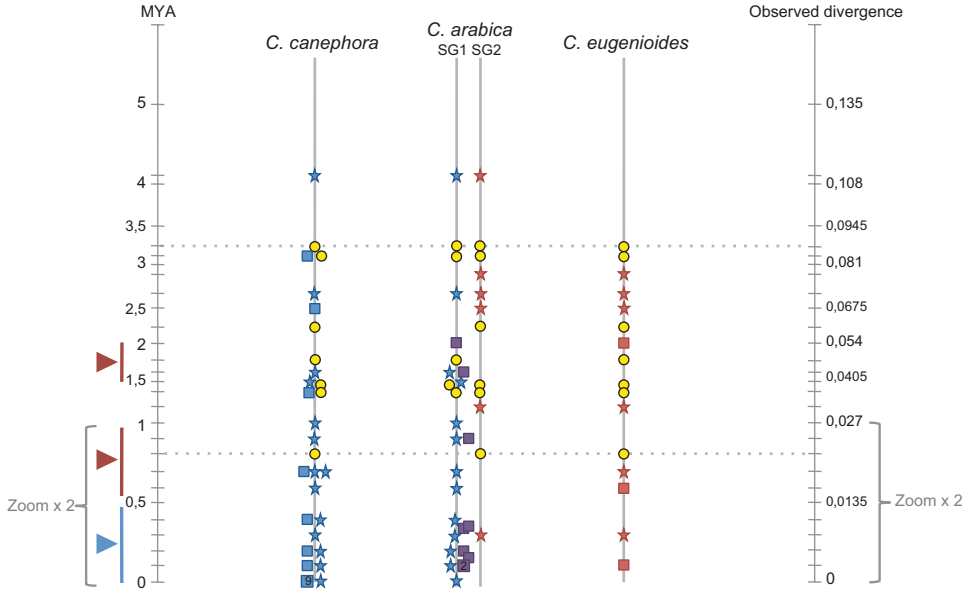


Fig. 4 Timing of insertion of *Divo* and comparative orthologous analysis in *C. canephora, C. eugenioides*, and *C. arabica* ACGC genomes. The *vertical line* on the right shows the divergence scale of LTRs for each element. The *vertical line* on the left shows the insertion times in Mya estimated with the molecular clock of Ma and Bennetzen (2004) (1.3 e^{-8} substitution per site and per year). Peaks of insertions observed in *C. canephora* (0–0.5 Mya) and *C. eugenioides* (0.5–1 and 1.5–2 Mya) relating to Fig. 3 are symbolized by the *blue* and *red triangles*, respectively. The insertion sites are located according to their estimated insertional time. *Yellow circles* represent *Divo* insertions at orthologous sites in the three species. The two horizontal *gray dashed lines* indicate the most recent (0.7 Mya) and the oldest (3.3 Mya) *Divo* elements present in the three species. Noted that for *C. arabica*, the most recent insertion is absent from one sub-genome. Insertions shared between two species are represented in *blue* or *red stars* according to the species involved. The most recent copies shared by *C. eugenioides* and one sub-genome of *C. arabica* in one hand, and *C. canephora* and the other sub-genome of *C. arabica* in the other hand both dated from 0.3 Mya. The oldest copies shared by *C. canephora* and *C arabica* on one hand, and *C. eugeniodies* and *C. arabica* on the other hand, both dated from 2.6 Mya. *Divo* insertions present in only one species are represented by *blue*, *purple*, and *red boxes* respectively for *C. canephora, C. eugenioides*, and *C. arabica* (represented by its two sub-genomes SG1 and SG2). Numbers in *boxes* indicate copy numbers at the site. *Purple boxes* between the two sub-genomes for *C. arabica* indicate unknown sub-genome identification for these insertions

Divo elements in plant genomes

Only one element found in Repbase called *Copia_12* (http://www.girinst.org/2014/vol14/issue9/Copia-12_TC-I.html), showed significant similarity with *Divo* (76% of nucleotide similarity between internal regions and 48.1% between the LTRs). *Copia_12* was annotated in the *Theobroma cacao* genome (Argout et al. 2011), but the element was neither characterized nor classified. Dotplot alignment between *Divo* (Accession #KX767841) and *Copia_12* confirmed the overall conservation of the elements structure with the exception of the LTR regions (only 52% of identity), suggesting that *Copia_12* may belong to the *Divo* family and so that the *Divo* family is not restricted to the *Coffea* genus (Supplemental data 5). We also checked the identity between our sequences of *Divo* from *Coffea* and the *Matita* element from *Arachis duranensis* (accession #JQ040302). The identity between *Matita* and the reference copies of *C. canephora* (Denoeud et al. 2014) and *C. canephora, C. arabica*, and *C. eugenioides* PacBio is of 53.7, 57, 57.2 and 57.1%, respectively. These percentages of identity indicate that *Matita* could effectively be a *Divo* element, but with a different history in *Arachis* genomes, leading to a significant sequence divergence with the *Divo* family from *Coffea*. Moreover, *Matita* is not complete and probably quite degenerate, explaining the weak percentages of identity with complete *Divo* elements.

Using four *Divo* sequences from *C. canephora* (Denoeud et al. 2014 and PacBio), *C. arabica* and *C. eugenioides* (accessions #: KX767840, KX767841, KX767839, and KX767842) as references (best intra-LTR sequence conservation: 97.4, 99.4, 99.4, and 99.7%, respectively, and longest ORF for Pol region). We searched for *Divo* in publicly available plant genomes. 22 genomic sequences were recovered from 14 angiosperm species and their RT amino-acid domains were used to construct a NJ phylogenetic tree (Supplemental data 5). *Divo* from *Coffea* form one monophyletic group inside the *Bianca* lineage. Interestingly, similar sequences to *Divo* found in the previously mentioned plant genomes were separated into two clear clades, corresponding to monocots and dicots, suggesting the *Bianca*

lineage is composed of two families: one for monocots and the other named *Divo* for dicots.

To further characterize *Divo* in dicots, we decided to annotate these elements in two reference genomes: *A. thaliana* (140 Mb) and *V. vinifera* (~500 Mb). A total of 197 and 1,384 potential LTR-RTs were detected in these genomes by LTR_STRUC. Out of these, seven and 44 sequences similar to *Divo* were recovered from the *A. thaliana* and *V. vinifera* genomes, respectively. The overall structure of these sequences is strictly similar to that of the complete *Divo* sequence (#KX767841) (Supplemental data 6), including the total length of the elements (an average of 6,071 bp for *A. thaliana* and 5,824 bp for *V. vinifera*) and the length of LTRs (335 bp on average for *A. thaliana* and 314 bp on average for *V. vinifera*).

In *A. thaliana*, four copies are potentially functional since no frame-shift was present in the ORFs of these elements. One of these (called L34-161, LTRs identity of 98.2%), displays a unique large ORF including the Gag and the Pol regions, as found frequently for *Copia* LTR-RTs (Peterson-Burch and Voytas 2002), but so far unique for all the *Divo* sequences analyzed. In grapevine, three sequences appeared potentially functional. One of them, called L107-1314 (LTRs identity of 96.8%), seems the most conserved as it carries only one stop codon between the Gag and Pol regions, contrary to the two others.

Finally, an analysis of the putative PBS region in 120 *Divo* sequences (from the copies of *C. canephora, C. arabica, C. eugenioides, Arabidopsis thaliana, Vitis vinifera, Brassica rapa, Medicago truncatula* and *Matita*) indicated that only the first 14 bp of the PBS region is conserved, particularly the four nucleotides "TTAT," while the 3′ ends were found more diverse (Fig. 2b).

Altogether these results suggest that *Divo*, the family of LTR-RTs described for the first time from complete elements, is actually conserved among a large panel of dicot plants.

Discussion

A novel LTR-RT family conserved among dicotylodenous plants

We uncovered a novel LTR-RT family called *Divo* in diploid and allotetraploid coffee-tree genomes. This family is related to a degenerated element previously annotated in a *C. canephora* BAC clone and used to study the relationships between 32 *Coffea* species (Hamon et al. 2011). *Divo* was classified into the *Bianca* lineage using a phylogenetic analysis (Fig. 1 and Supplemental data 2) and because it shares the same key structural features with elements from this lineage such as the overall length of the element and LTR sizes (Wicker and Keller 2007; Nielen et al. 2012). However, *Divo*-like homologous sequences were restricted to dicots, suggesting that the *Divo* family evolved specifically since the divergence between dicots and monocots.

Bianca is the most ancient *Copia* lineage as showed by our RT-based phylogenetic analysis (see also Piednoël et al. 2013). *Bianca* elements have been initially detected in Triticeae, rice, Arabidopsis and alfalfa (Wicker and Keller 2007; Wang and Liu 2008). Whereas the *Bianca* lineage was not found in soybean (Du et al. 2010), sugarcane (Domingues et al. 2012) or quinoa (Kolano et al. 2013), it was frequently found in Angiosperm genomes (Piednoël et al. 2013), confirming that this ancient lineage was spread along the Angiosperms divergence. The *Bianca* lineage was also frequently found with a moderated copy number, such as in Arabidopis, rice, peanut, eucalyptus, and poplar (Wicker and Keller 2007; Nielen et al. 2012; Marcon et al. 2015; Natali et al. 2015), with the exception in the pear genome, where *Bianca* represents the highest copy number lineage of all *Copia* elements (Yin et al. 2015).

Similarly to other Angiosperm genomes, *Divo* was found in coffee-trees with a moderate copy number, suggesting that coffee host genomes may apply a control of the copy number of this family,

One of the main characteristics of the *Divo* family is an atypical PBS that did not show any strong complementary sequence to host tRNAs (Fig. 2). A PBS is usually composed of 11 to 18 nucleotides complementary to a host tRNA that primes the reverse transcription of the element (Le Grice 2003). However, the detection of recent *Divo* element insertions based on the LTR divergence suggests potential recent mobility. Further studies, including the detection of circular dsDNA molecules, suggesting replicative forms of the elements (Mirouze et al. 2009), might bring more evidence about the actual transpositional activity of *Divo*.

The comparison of the *Divo* alleged PBS (Fig. 2, CGT TATCAGCACGA) with those of the families *Romani* in Arabidopsis (GTTTATCAGCAC, Wicker and Keller 2007), *Matita* in peanut (TGTTATCAGCAC, Nielen et al. 2012) and *Mtr13* in *Medicago* (CGTTATCAGCACGC, Wang and Liu 2008) suggest that it could be conserved in different families from the *Bianca* lineage. Other groups of LTR-RTs lacking PBS identification were previously characterized in *Aedes aegypti* (Minervini et al. 2009) and in *Dictyostelium*, (Leng et al. 1998), suggesting that these LTR-RTs may not need a functional PBS and/or that they could use another primer to accomplish their replication cycle.

Divo in diploid and allotetraploid coffee-trees genomes

The time of LTR-RTs insertions in genomic sequences can be roughly estimated using the divergence between

LTR sequences of each element, as these regions are supposed to be strictly identical in an active copy at the time of each insertion. Since no specific substitution rate is available for *Coffea*, we used the one estimated by Ma & Bennetzen (2004) for rice LTR-RTs ($1.3e^{-8}$ substitution per site per year), and often applied to other dicots and monocots LTRs divergence analyses (Vitte and Bennetzen 2006). Estimation of LTR-RTs time of insertions in the studied *Coffea* species showed that these elements were differentially amplified in the last 2.5 My. The *C. canephora* ACGC genome contains more recent *Divo* copies than the other genomes and more than the published *C. canephora* genome (Denoeud et al. 2014), which is probably a consequence of the higher quality and completeness reached by the sequencing technology (Fig. 3). Particularly, 18 recent insertions (100% of nucleotide conservation between their LTRs) were observed in *C. canephora*, suggesting that *Divo* was amplified and activated recently in this species, and with a lesser extent in *C. arabica*. On the contrary, almost no recent insertions were detected in *C. eugenioides* (Fig. 4). This result is in agreement with the data obtained by Hamon et al. (2011), where they showed that *Divo* is accompanying the *C. canephora* diversification but not that of the genus *Coffea*, including *C. eugenioides*. As we can observe recent and specific insertion sites in *C. arabica* (Fig. 4), *Divo* could yet also be active or would have been active in the actual *C. canephora* ancestor of *C. arabica*. On the contrary, *C. eugenioides* did not show recent transpositions, while two discrete periods of activity at 0.5–1 and 1.5–2 Mya were evidenced. Furthermore, a high number of solo LTRs were detected in *C. eugenioides*, suggesting that the control of *Divo* copy number may be more efficient in this genome via unequal homologous recombination mechanisms (Bennetzen and Kellogg 1997). The distinct periods of transposition and removal activities of *Divo* between *C. canephora* and *C. eugenioides* indicate a different evolution of the genome structural dynamics of these two diploids. As expected, the insertion periodes of *Divo* elements within *C. arabica* genome share the pattern of both *C. canephora*, with a recent activity (0–0.5 Mya) and *C. eugenioides*, with a secondary and more ancient peak of insertions (1.5–2 Mya; Fig. 3). This pattern (common timing insertion with diploid ancestor, conservation of orthologous copies, and copy number estimation) suggests that the allotetraploid genome of *C. arabica* did not suffer of strong elimination or increase of *Divo* copy number following the allopolyploidization. This result differs from other LTR-RT families in allopolyploid genomes that underwent modifications of their copy numbers after polyploidization (Ainouche et al. 2009; Parisod et al. 2010). Further and wider comparative analysis of LTR-RTs between the *C. arabica* genome and its two diploid progenitors will bring interesting information concerning the consequences of the polyploidization on the LTR-RTs dynamics and control in this model.

An evolutionary scenario for diploid and allotetraploid genomes divergence

We used the complete copies of *Divo* conserved in orthologous regions between the *C. arabica* genome and its two diploid progenitors, *C. canephora* and *C. eugenioides*, to better understand the evolution of their genomes. The relative time of insertion of *Divo* copies allowed us to propose an evolutionary scenario for the divergence time between *C. canephora, C. eugenioides*, and for the formation of *C. arabica*.

The relative time for the *C. canephora* and *C. eugenioides* radiation can be investigated thanks to the conservation of *Divo* copies at orthologous sites, corresponding to *Divo* copies likely inserted in the common ancestor of the two diploid genomes. Such orthologous copies had an estimated time of insertion ranging between 3.1 and 0.8 Mya, suggesting that the two species completely diverged at least 0.8 Mya. However, *Divo* copies were also found specificaly inserted in *C. canpehora* or in *C. eugenioides* in the same time interval, suggesting a long period of radiation into two gene pools to give rise to the two species. The analysis of all *Divo* copies (conserved and non-conserved) that inserted between 3.1 and 0.8 Mya in the two diploid ancestors, showed two waves of insertion (two peaks at 1.5–2 Mya and 0.5–1 Mya) that occurred in *C. eugenioides* but not in *C. canephora*, suggesting a divergence in the activity of *Divo* during the process of radiation. Finally the clear amplification of *Divo* observed in *C. canephora* but not in *C. eugenioides* in the time interval of 0 to 0.5 Mya confirmed that the two species were already differentiated.

The relative time of *C. arabica* polyploidization event may be also estimated using the insertion time of conserved *Divo* at orthologous locations in the two sub-genomes. Since the last common *Divo* insertions at orthologous sites between *C. arabica* and *C. eugenioides* and between *C. arabica* and *C. canephora* were observed in the last 0.3 Mya, we concluded that *C. arabica* is originated from a very recent hybridization, confirming previous estimation (Yu et al. 2011). Interestingly, *Divo* copies showing 100% of identity between the two LTRs (nine copies) were only found in the *C. canephora* genome strongly suggesting that *Divo* remains active in that species in a very recent time.

Coffea arabica is an allotetraploid species originated from a hybridization event that occurred between diploid species and taking place 46,000–665,000 years ago (Yu et al. 2011). Understanding the mechanisms of genome modifications during the allotetraploidization may be of interest. *Divo*, a novel family of the *Bianca* lineage among

the superfamily *Copia*, is present in moderated copy numbers in dicots. Complete and potentially functional *Divo* copies were detected in *C. arabica* and its diploid *C. canephora* and *C. eugenioides* progenitors. The activity of the *Divo* family, and the mechanisms of control of its copy number played certainly a role in the differentiation of *C. canephora* and *C. eugenioides* genomes. Beside strucural impacts on genomes, its precise functional role remains to be elucidated. In the near future, a complete characterization of active transposable elements in *C. arabica* and its diploid progenitors will bring more insights into plant genomes divergence and evolution.

Funding Information R.G. was supported by a Special Visiting Scientist grant from the Ciência sem Fronteiras program under the reference ID 84/2013 (Cnpq/CAPES).

Compliance with ethical standards

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

Ainouche ML, Fortune PM, Salmon A, Parisod C, Grandbastien MA, Fukunaga K, Ricou M, Misset MT (2009) Hybridization, polyploidy and invasion: lessons from Spartina (Poaceae). Biol Invasions 11:1159–1173

Allaire JJ (2012) RStudio: Integrated development environment for R. J Wildl Manage 75:1

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic Local Alignment Search Tool. J Mol Biol 215:403–410

Argout X, Salse J, Aury J-M, Guiltinan MJ, Droc G, Gouzy J, Allegre M, Chaparro C, Legavre T, Maximova SN, Abrouk M, Murat F, Fouet O, Poulain J, Ruiz M, Roguet Y, Rodier-Gout M, Barbosa-Neto JF, Sabot F, Kudrna D, Ammiraju JSS, Schuster SC, Carlson JE, Sallet E, Schiex T, Dievart A, Kramer M, Gelley L, Shi Z, Bérard A, Viot C, Boccara M, Resterucci AM, Guignon V, Sabau X, Axtell MJ, Ma Z, Zhang Y, Brown S, Bourge M, Golser W, Song X, Clement D, Rivallan R, Tahi M, Akaza JM, Pitollat B, Gramacho K, D'Hont A, Brunel D, Infante D, Kebe I, Costet P, Wing R, McCombie WR, Guiderdoni E, Quetier F, Panaud O, Wincker P, Bocs S, Lanaud C (2011) The genome of Theobroma cacao. Nat Genet 43:101–109

Audic S, Claverie J (1997) The Significance of Digital Gene Expression Profiles. Genome Res 7:986–995.

Bao W, Kojima KK, Kohany O (2015) Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA 6:1–6. doi:10.1186/s13100-015-0041-9

Bennetzen JL, Kellogg E (1997) Do Plants Have a One-Way Ticket to Genomic Obesity? Plant Cell 9:1509–1514

Bouharmont J (1959) Recherches sur les affinités chromosomiques dans le genre Coffea. I.N.É.A.C., Montpellier

Brenchley R, Spannagl M, Pfeifer M, Barker GLA, D'Amore R, Allen AM, McKenzie N, Kramer M, Kerhornou Y, Bolser D, Kay S, Waite D, Trick M, Bancroft I, Gu Y, Huo N, Luo MC, Sehgal S, Kianian S, Gill B, Anderson O, Kersey P, Dvorak J, McCombie R, Hall A, Mayer KFX, Edwards KJ, Bevan M, Hall N (2012) Analysis of the bread wheat genome using whole genome shotgun sequencing. Nature 491:705–710

Cai J, Liu X, Vanneste K, Proost S, Tsai WC, Liu KW, Chen LJ, He Q, Xu Q, Bian C, Zheng Z, Sun F, Liu W, Hsiao YY, Pan ZJ, Hsu CC, Yang YP, Hsu YC, Chuang YC, Dievart A, Dufayard JF, Xu X, Wang JY, Wang J, Xiao XJ, Zhao XM, Du R, Zhang GQ, Wang M, Su YY, Xie GC, Liu GH, Li LQ, Huang LQ, Luo YB, Chen HH, Van de Peer Y, Liu ZJ (2015) The genome sequence of the orchid Phalaenopsis equestris. Nat Am 47:65–76.

Carvalho A (1952) Taxonomia de Coffea arabica L. VI - Caracteres morfologicos dos haploides. Bragantia 12:201–212.

Chaparro C, Guyot R, Zuccolo A, Piégu B, Panaud O (2007) RetrOryza: A database of the rice LTR-retrotransposons. Nucleic Acids Res 35:66–70

Davis AP, Tosh J, Ruch N, Fay MF (2011) Growing coffee: Psilanthus (Rubiaceae) subsumed on the basis of molecular and morphological data; implications for the size, morphology, distribution and evolutionary history of Coffea. Bot J Linn Soc 167:357–377

Denoeud F, Carretero-Paulet L, Dereeper A, Droc G, Guyot R, Pietrella M, Zheng C, Alberti A, Anthony F, Aprea G, Aury JM, Bento P, Bernard M, Bocs S, Campa C, Cenci A, Combes MC, Crouzillat D, Da Silva C, Daddiego L, De Bellis F, Dussert S, Garsmeur O, Gayraud T, Guignon V, Jahn K, Jamilloux V, Joët T, Labadie K, Lan T, Leclercq J, Lepelley M, Leroy T, Li LT, Librado P, Lopez L, Muñoz A, Noel B, Pallavicini A, Perrotta G, Poncet V, Pot D, Priyono, Rigoreau M, Rouard M, Rozas J, Tranchant-Dubreuil C, VanBuren R, Zhang Q, Andrade AC, Argout X, Bertrand B, de Kochko A, Graziosi G, Henry RJ, Jayarama, Ming R, Nagai C, Rounsley S, Sankoff D, Giuliano G, Albert VA, Wincker P, Lashermes P (2014) The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. Science 345:1180–1184

Devos KM, Brown JKM, Bennetzen JL (2002) Genome size reduction through illegitimate recombination counteracts genome expansion in arabidopsis. Genome Res 12:1075–1079

Dias ES, Hatt C, Hamon S, Hamon P, Rigoreau M, Crouzillat D, Carareto CMA, de Kochko A, Guyot R (2015) Large distribution and high sequence identity of a Copia-type retrotransposon in angiosperm families. Plant Mol Biol 89:83–97

Domingues DS, Cruz GMQ, Metcalfe CJ, Nogueira FTS, Vicentini R, Alves CS, Van Sluys MA (2012) Analysis of plant LTR-retrotransposons at the fine-scale family level reveals individual molecular patterns. BMC Genomics 13:1–13. doi:10.1186/1471-2164-13-137

Du J, Tian Z, Hans CS, Laten HM, Cannon SB, Jackson SA, Shoemaker RC, Ma J (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

Eickbush TH, Jamburuthugoda VK (2007) The diversity of retrotransposons and the properties of their reverse transcriptases. Mol Cell Biol 134:221–234

Fedoroff NV (2012) Transposable elements, epigenetics, and genome evolution. Science 338:758–767

Feschotte C, Pritham EJ (2007) DNA transposons and the evolution of eukaryotic genomes. Annu Rev Genet 41:331–368

Fontana A (2010) A hypothesis on the role of transposons. Biosystems 101:187–193

Gilbert C, Peccoud J, Chateigner A, Moumen B, Cordaux R, Herniou EA (2016) Continuous influx of genetic material from host to virus populations. PLoS Genet 12:1–21

Guyot R, Darré T, Dupeyron M, de Kochko A, Hamon S, Couturon E, Crouzillat D, Rigoreau M, Rakotomalala JJ, Raharimalala NE, Akaffou SD, Hamon P (2016) Partial sequencing reveals the transposable element composition of Coffea genomes and provides evidence for distinct evolutionary stories. Mol Genet Genomics 291:1979–1990

Hamon P, Duroy PO, Dubreuil-Tranchant C, Costa PMD, Duret C, Razafinarivo NJ, Couturon E, Hamon S, e Kochko A, Poncet V, Guyot R (2011) Two novel Ty1-copia retrotransposons isolated from coffee trees can effectively reveal evolutionary relationships in the Coffea genus (Rubiaceae). Mol Genet Genomics 285:447–460

International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. Nature 436:793–800

Kohany O, Gentles AJ, Hankus L, Jurka J (2006) Annotation, submission and screening of repetitive elements in Repbase: Repbase-Submitter and Censor. BMC Bioinformatics 7:474

Kolano B, Bednara E, Weiss-Schneeweiss H (2013) Isolation and characterization of reverse transcriptase fragments of LTR retrotransposons from the genome of *Chenopodium quinoa* (Amaranthaceae). Plant Cell Rep 32:1575–1588

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA (2009) Circos: an information aesthetic fo comparative genomics. Genome Res 19:1639–1645

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 72:181–204

Lashermes P, Combes MC, Robert J, Trouslot P, D'Hont A, Anthony F, Charrier A (1999) Molecular characterization and origin of the Coffea arabica L. genome. Mol Gen Genet 261:259–266

Le Grice SFJ (2003) "In the beginning": initiation of minus strand DNA synthesis in retroviruses and LTR-containing retrotransposons. BioChemistry 42:14349–14355

Leng P, Klatte DH, Schumann G, Boeke JD, Steck TL (1998) Skipper, an LTR retrotransposon of Dictyostelium. Nucleic Acids Res 26:2008–2015

Lerat E (2010) Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. Heredity 104:520–533

Lin X, Faridi N, Casola C (2016) An ancient transkingdom horizontal transfer of penelopelike retroelements from arthropods to conifers. Genome Biol Evol 8:1252–1266

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

Llorens C, Futami R, Covelli L, Domínguez-Escribá L, Viu JM, Tamarit D, Aguilar-Rodríguez J, Vicente-Ripolles M, Fuster G, Bernet GP, Maumus F, Munoz-Pomer A, Sempere JM, Latorre, Moya A (2011) The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. Nucleic Acids Res 39:D70–D74

Louarn J (1976) Hybrides interspécifiques entre Coffea canephora Pierre et C. eugenioides Moore. Café Cacao Thé 20:33–52

Ma J, Bennetzen JL (2004) Rapid recent growth and divergence of rice nuclear genomes. PNAS 101:12404–12410

Marcon HS, Domingues DS, Silva JC, Borges RJ, Matioli FF, Fonter MRM, Marino CL (2015) Transcriptionally active LTR retrotransposons in Eucalyptus genus are differentially expressed and insertionally polymorphic. BMC Plant Biol 15:198–214

McCarthy EM, McDonald JF (2003) LTR_STRUC: a novel search and identification program for LTR retrotransposons. Bioinformatics 19:362–367

Mehra M, Gangwar I, Shankar R (2015) A deluge of complex repeats: the solanum genome. PLoS One 10:1–38

Minervini CF, Viggiano L, Caizzi R, Marsano RM (2009) Identification of novel LTR retrotransposons in the genome of *Aedes aegypti*. Gene 440:42–49

Mirouze M, Reinders J, Bucher E, Nashimura T, Schneeberger K, Ossowki S, Cao J, Weigel D, Paszkowski J, Mathieu O (2009) Selective epigenetic control of retrotransposition in Arabidopsis. Nature 461:1–5

Nielen S, Vidigal BS, Leal-Bertioli SCM, Ratnaparkhe M, Paterson AH, Garsmeur O, D'Hont A, Guimarães PM, Bertioli DJ (2012) Matita, a new retroelement from peanut: characterization and evolutionary context in the light of the Arachis A–B genome divergence. Mol Genet Genomics 287:21–38

Ouyang S, Buell CR (2004) The TIGR Plant Repeat Databases: a collective resource for the identification of repetitive sequences in plants. Nucleic Acids Res 32:360–363

Panaud O (2016) Horizontal transfers of transposable elements in eukaryotes: The flying genes. Comptes rendus Biol. doi:10.1016/j.crvi.2016.04.013

Parisod C, Alix K, Just J, Petit M, Sarilar V, Mhiri C, Ainouche M, Chalhoub B, Grandbastien MA (2010) Impact of transposable elements on the organization and function of allopolyploid genomes. New Phytol 186:37–45

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, Schmutz J, Spannagl M, Tang H, Wang X, Wicker T, Bharti AK, Chapman J, Fletus FA, Otillar RP, Penning BW, Salamov AA, Wang Y, Zhang L, Carpita NC, Freeling M, Gingle AR, Hash CT, Keller B, Klein P, Kresovich S, McCann MC, Ming R, Peterson DG, Mehboob-ur-Rahman, Ware D, Westhoff P, Mayer KFX, Messing J, Rokhsar DS (2009) The Sorghum bicolor genome and the diversification of grasses. Nature 457:551–556

Peterson-Burch BD, Voytas DF (2002) Genes of the Pseudoviridae (Ty1/copia Retrotransposons). Mol Biol Evol 19:1832–1845

Piednoël M, Carrete-Vega G, Renner SS (2013) Characterization of the LTR retrotransposon repertoire of a plant clade of six diploid and one tetraploid species. Plant J 75:699–709

Romualdi C, Bortoluzzi S, D'Alessi F, Danieli GA (2003) IDEG6: a web tool for detection of differentially expressed genes in multiple tag sampling experiments. Physiol Genomics 12:159–162

Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrel B (2000) Artemis: sequence visualization and annotation. Bioinformatics 16:944–945

SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL (1998) The paleontology of intergene retrotransposons of maize. Nat Genet 20:43–45

Schaack S, Gilbert C, Feschotte C (2010) Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. Trends Ecol Evol 25:537–546

Schmieder R, Edwards R (2011) Quality control and preprocessing of metagenomic datasets. Bioinformatics 27:863–864

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternk S, Liang C, Zhang J, Fulton L, Graves TA, Minx P, Reily AD, Courtney L, Kruchowski SS, Tomlison C, Strong C, Delehaunty K, Fronick C, Courtney B, Rock SM, Belter E, Du F, Kim K, Abbott RM, Cotton M, Levy A, Marchetto P, Ochoa K, Jackson SM, Gillam B, Chen W, Yan L, Higginbotham J, Cardenas M, Waligorski J, Applebaum E, Phelps L, Falcone J, Kanchi K, Thane T, Scimone A, Thane N, Henke J, Wang T, Ruppert J, Shah N, Rotter K, Hodges J, Ingenthron E, Cordes M, Kohlberg S, Sgro J, Delgado BMead K, Chinwalla A, Leonard S, Crouse K, Collura K, Kudrna D, Currie J, He R, Angelova A, Rajasekar S, Mueller T, Lomeli R, Scara G, Ko A, Delaney K, Wissotski M, Lopez G, Campos D, Braidotti M, Ashley E, Golser W, Kim H, Lee S, Lin J, Dujmic Z, Kim W, Talag J, Zuccolo A, Fan C, Sebastian A, Kramer M, Spiegel L, Nascimento L, Zutavern T, Miller B, Ambroise C, Muller S, Spooner W, Narechania A, Ren L, Wei S, Kumari S, Faga B, Levy MJ, McMahan L, Van Buren P, Vaughn MW, Ying K, Yeh CT, Emrich SJ, Jia Y, Kalyanaraman A, Hsia AP, Barbazuk WB, Baucom RS, Brutnell TP, Carpita NC, Chaparro C, Chia JM, Deragon JM, Estill JC, Fu Y, Jeddeloh JA, Han Y, Lee H, Li P, Lish DR, Liu S, Liu Z, Nagel DH, McCann MC, SanMiguel P, Myers AM, Nettleton D, Nguyen J, Penning BW, Ponnala L, Schneider KL, Schwartz DC, Sharma A, Soderlund C, Springer NM, Sun Q, Wang H, Waterman M, Westerman R, Wolfgruber TK, Yang L, Yu Y, Zhang L, Zhou S, Zhu Q, Bennetzen JL, Dawe RK, Jiang J, Jiang N, Presting GG, Wessler SR,

Aluru S, Martienssen RA, Clifton SW, McCombie WR, Wing RA, Wilson RK (2009) The B73 maize genome: complexity, diversity, and dynamics. Science 326:1112–1116

Sonnhammer ELL, Durbin R (1996) A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. Gene 167:1–10

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, pp 42–45

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408:796–815

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

Vitte C, Bennetzen JL (2006) Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution. Proc Nat Acad Sci USA 103:17638–17643.

Wang H, Liu J-S (2008) LTR retrotransposon landscape in Medicago truncatula: more rapid removal than in rice. BMC Genomics 9:382–395

Wicker T, Keller B (2007) Genome-wide comparative analysis of copia retrotransposons in Triticeae, rice, and Arabidopsis reveals conserved ancient evolutionary lineages and distinct dynamics of individual copia families. Genome Res 17:1072–1081

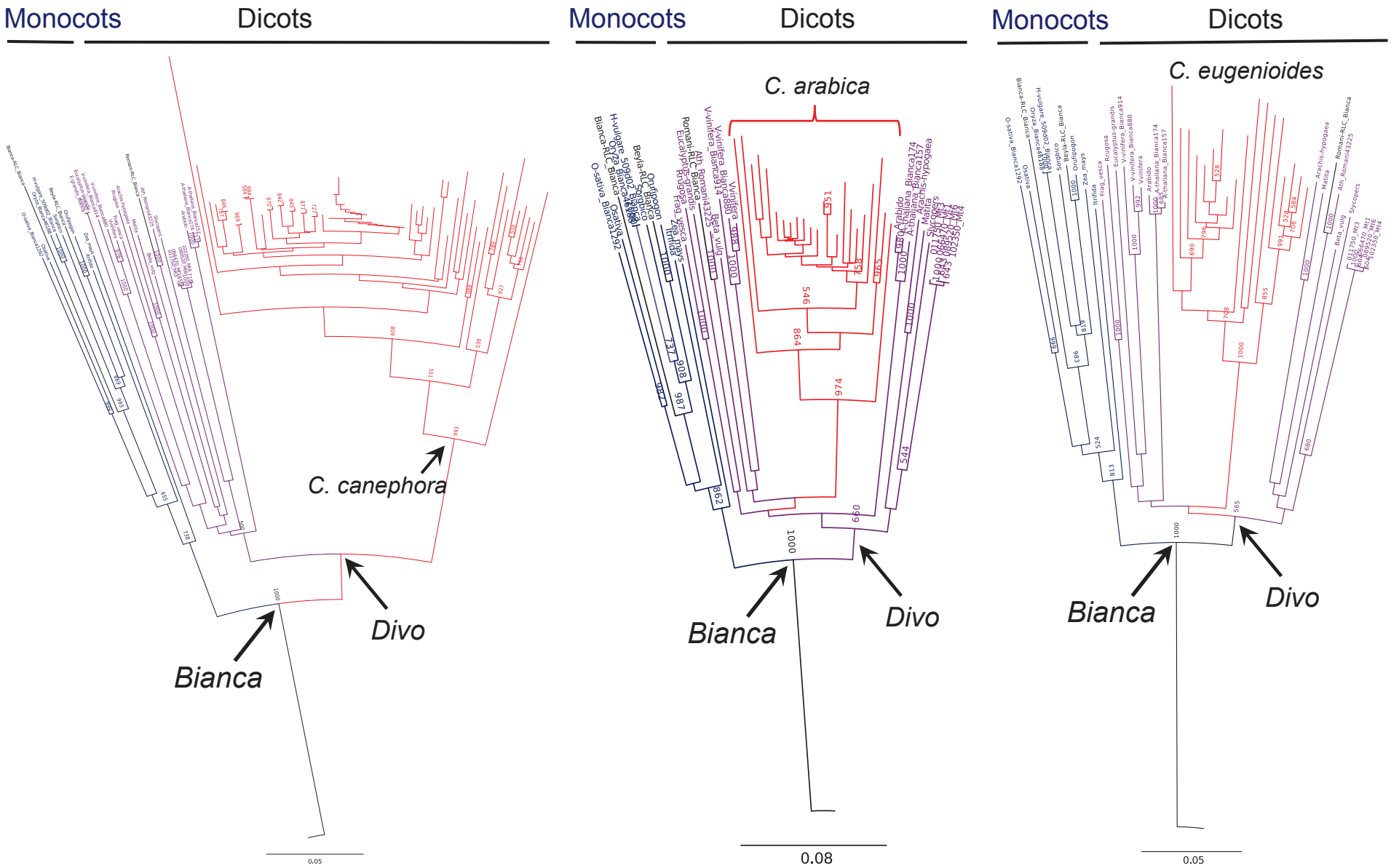
Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, Paux E, SanMiguel P, Schulman AH (2007) A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8:973–982

Xu Z, Wang H (2007) LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res 35:W265–W268

Yin H, Du J, Wu J, Wei S, Xu Y, Tao S, Wu J, Zhang S (2015) Genome-wide annotation and comparative analysis of long terminal repeat retrotransposons between pear species of *P. bretschneideri* and *P. communis*. Sci Rep 5:1–15.

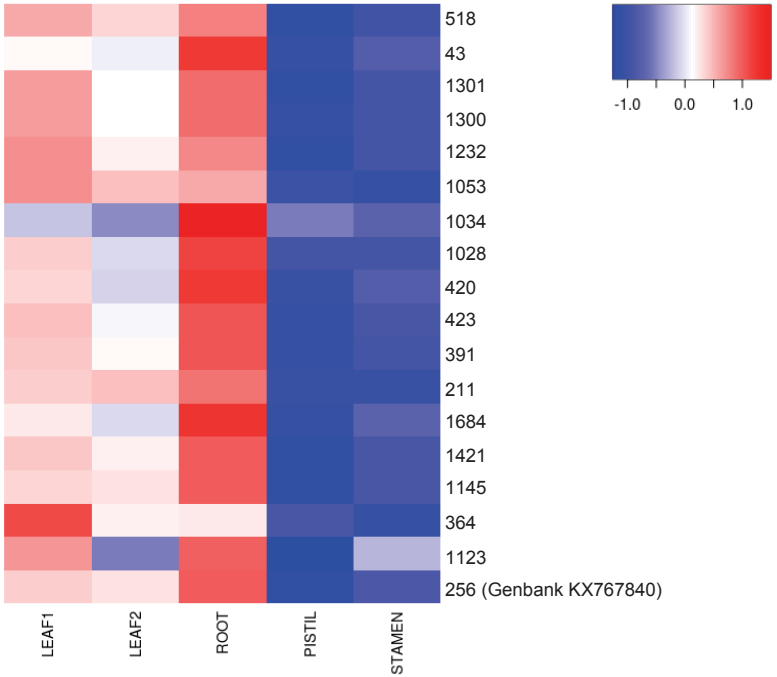
Yu Q, Guyot R, de Kochko A, Byers A, Navajas-Pérez R, Langston BJ, Dubreuil-Tranchant C, Paterson AH, Poncet V, Nagai C, Ming R (2011) Micro-collinearity and genome evolution in the vicinity of an ethylene receptor gene of cultivated diploid and allotetraploid coffee species (Coffea). Plant J 67:305–317

Supplemental data 2: NJ trees of *Bianca* lineage (RT domains) in the three *Coffea* ACGC genomes.
Bootstraps values higher than 50% are indicated. Red branches: *Divo* elements found in coffee trees; purple branches: *Divo* elements from other dicots; black branches: *Bianca* reference RT domains; blue branches: *Bianca* elements from the monocots.



<u>Supplemental data 3</u>: A – Heatmap of normalized RPKM data from *Divo* retroelements expressed in different *C. canephora* tissues. The legend scores represent the deviation of the mean by standard deviation units (blue: no expression – red: expression). **B – Table of differential expression of *Divo* elements in *C. canephora* RNA-seq from leaves, roots, pistil, and stamen.** *Significance threshold of 0.05 with Bonferroni correction = 0.278 (Audic & Claverie, 1997 and Romualdi et al. 2003).

A

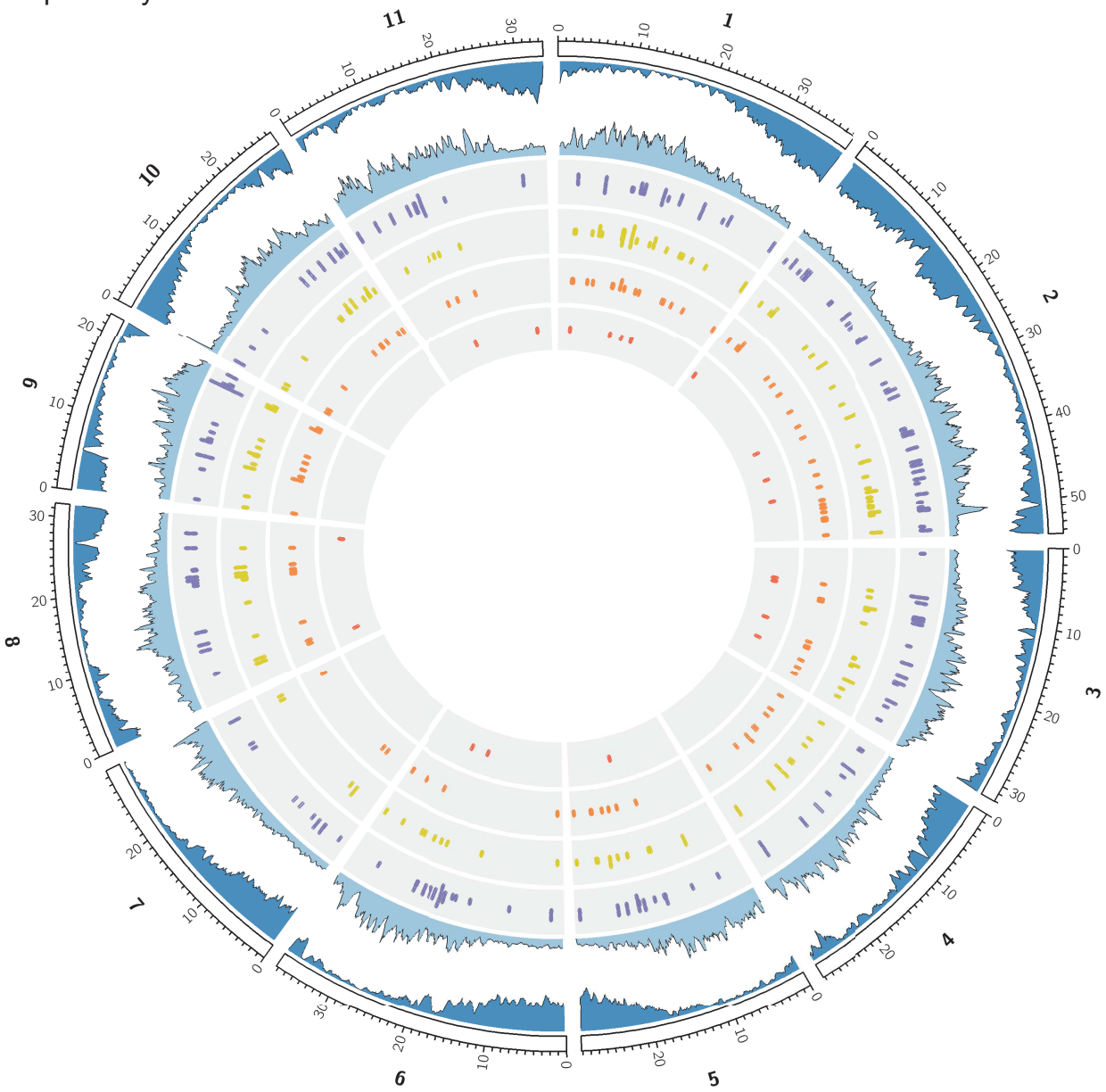


B

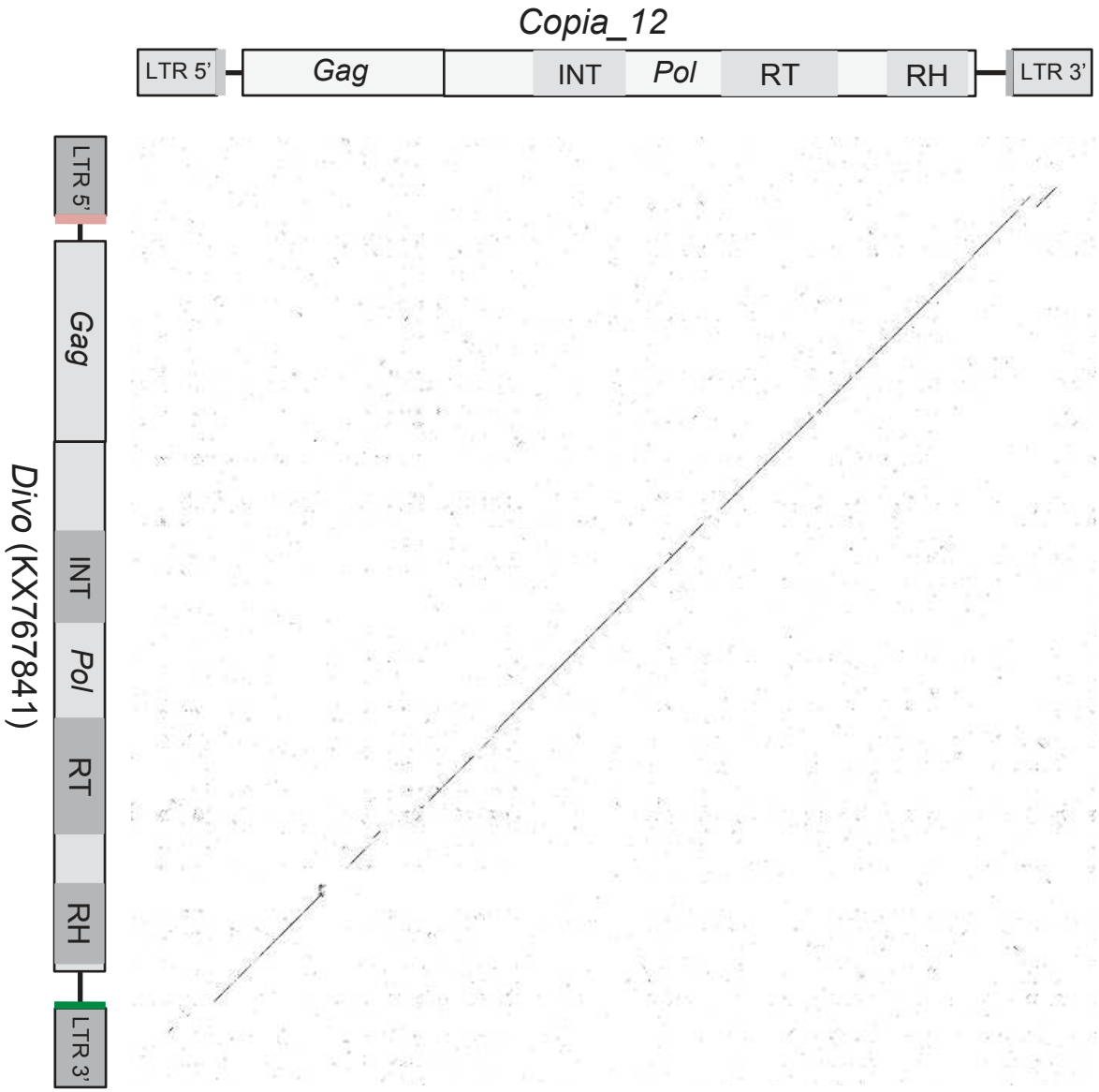
Divo retroelement	Leaf1 x Leaf2	Leaf1 x Root	Leaf2 x Root	Leaf1 x Pistil	Leaf2 x Pistil	Leaf1 x Stamen	Leaf2 x Stamen	Root x Pistil	Root x Stamen	Pistil x Stamen
518	0.030817	0.021228	0.009347	0.000000*	0.000000*	0.000000*	0.000000*	0.000000*	0.000000*	0.053971
43	0.041512	0.000031*	0.000034*	0.000000*	0.000000*	0.000000*	0.000170*	0.000000*	0.000000*	0.012059
1301	0.009202	0.016018	0.000877	0.000000*	0.000000*	0.000000*	0.000000*	0.000000*	0.000000*	0.033008
1300	0.007597	0.017120	0.000770	0.000000*	0.000000*	0.000000*	0.000000*	0.000000*	0.000000*	0.033008
1232	0.013778	0.029336	0.007629	0.000000*	0.000000*	0.000000*	0.000003	0.000000*	0.000000*	0.033008
1053	0.019248	0.016801	0.017304	0.000000*	0.000000*	0.000000*	0.000000*	0.000000*	0.000000*	0.040941
1034	0.039130	0.000000*	0.000000*	0.008204	0.089026	0.000105*	0.011995	0.000000*	0.000000*	0.031611
1028	0.020191	0.001566	0.000080*	0.000000*	0.000000*	0.000000*	0.000078*	0.000000*	0.000000*	0.089444
420	0.009773	0.000074*	0.000001*	0.000000*	0.000000*	0.000000*	0.000167*	0.000000*	0.000000*	0.014746
423	0.019251	0.003395	0.000261*	0.000000*	0.000000*	0.000000*	0.000004*	0.000000*	0.000000*	0.024079
391	0.014631	0.000204*	0.000011*	0.000000*	0.000000*	0.000000*	0.000000*	0.000000*	0.000000*	0.018085
211	0.047635	0.010557	0.015972	0.000000*	0.000000*	0.000000*	0.000000*	0.000000*	0.000000*	0.106358
1684	0.025465	0.000075*	0.000007*	0.000000*	0.000000*	0.000012*	0.001337	0.000000*	0.000000*	0.005634
1421	0.030591	0.003039	0.000844	0.000000*	0.000000*	0.000000*	0.000001*	0.000000*	0.000000*	0.023156
1145	0.040193	0.001819	0.001477	0.000000*	0.000000*	0.000000*	0.000000*	0.000000*	0.000000*	0.011714
364	0.071988	0.044714	0.086770	0.000662	0.018296	0.000482	0.011995	0.013518	0.008602	0.316189
1123	0.004937	0.042520	0.000936	0.000000*	0.005788	0.017307	0.067043	0.000000*	0.006543	0.000800
256 (Genbank KX767840)	0.041209	0.004002	0.002549	0.000000*	0.000000*	0.000000*	0.000004*	0.000000*	0.000000*	0.011423

Supplementary material 4: Circos representation of the locations of *Divo* copies in *C. canephora* published genome.

Complete copies: red; copies: orange; partial copies: yellow and solo-LTRs: purple. Genes and transposable elements repartition is represented in blue and light blue, respectively.

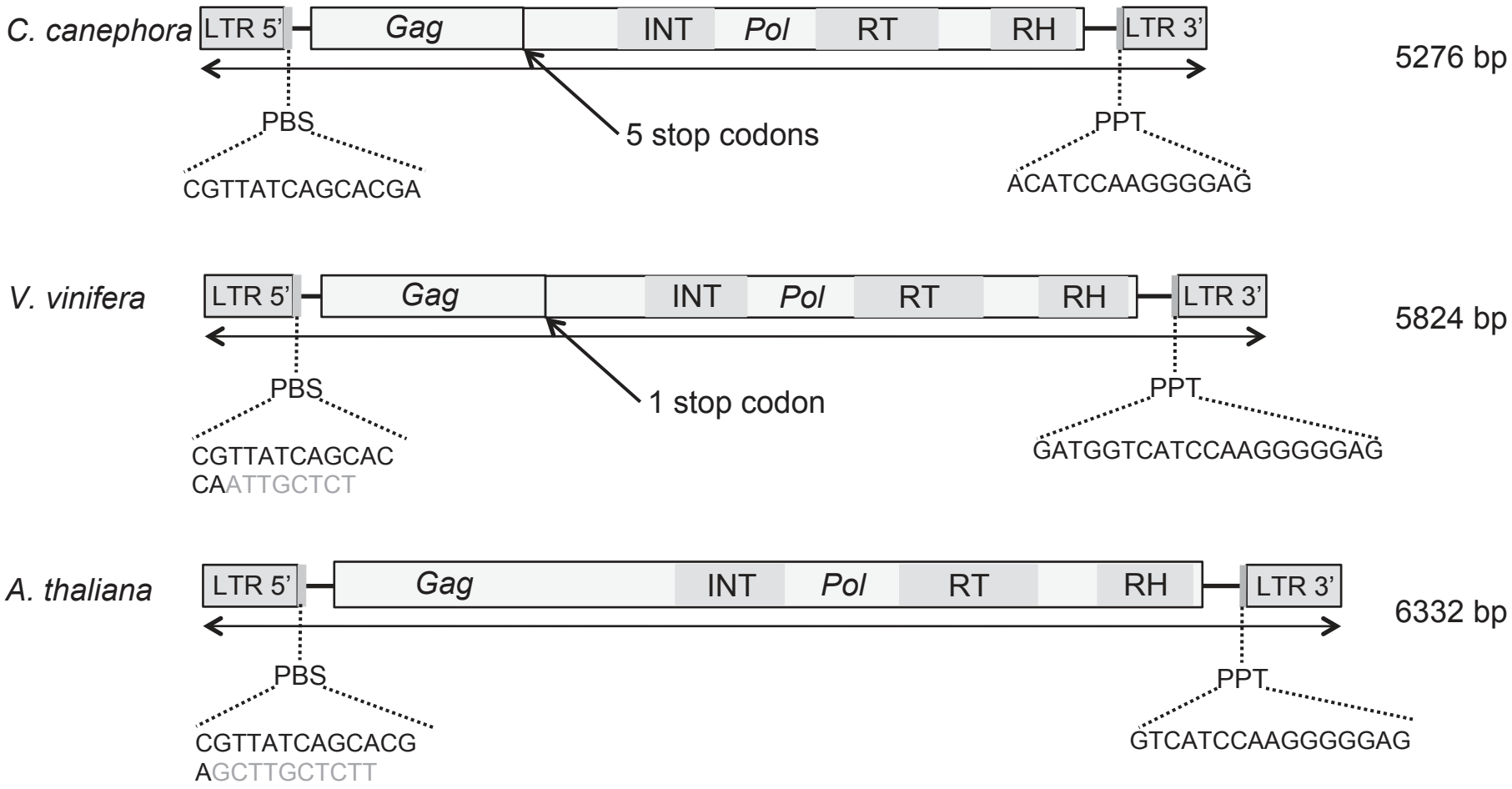


Supplemental data 5: Dot-plot alignment of *Divo*'s reference copy in *C. canephora* and Copia_12 in *Theobroma cacao*.



Supplemental data 6: Structural features of well-conserved copies of *Divo* elements in *C. canephora*, *V. vinifera* and *A. thaliana*.

LTR: Long Terminal Repeats; INT: Integrase; RT: Reverse Transcriptase; RH: RNAse H; PBS: Primer-Binding Site; PPT: PolyPurine tract.



3. Conclusions et perspectives

Dans cet article, nous avons montré que la lignée de *Copia Bianca*, peu connue et retrouvée en faible nombre de copies dans la plupart des génomes où elle a été détectée, est présente chez les caféiers. Elle illustre la difficulté d'utiliser des bases de données comme référence pour l'identification des éléments si celles-ci sont incomplètes (GyDB).

Les analyses phylogénétiques des séquences retrouvées dans plusieurs génomes d'Angiospermes, y compris les caféiers, ont montré que cette lignée de *Copia* peut être séparée en deux grandes familles : les *Bianca* retrouvés chez les Monocotylédones et *Divo*, la famille retrouvée chez les Dicotylédones. Concernant *C. arabica* et ses progéniteurs diploïdes, l'étude précise de *Divo* confirme que celui-ci a participé à la diversification génétique de *C. canephora*, confirmant les résultats de l'étude précédente sur cet élément (Hamon et al. 2011). Les estimations d'âge d'insertion des copies complètes montrent en effet des insertions très récentes et spécifiques à cette espèce. De plus, ces estimations confirment l'origine récente de *C. arabica*. Cette famille de LTR-RT semble avoir une activité moins importante chez *C. eugenioides* qui montre très peu de copies insérées récemment. Il ne semble pas y avoir eu un impact important de la polyploïdisation sur cet élément, étant donné que *C. arabica* montre beaucoup de copies communes à *C. canephora* et/ou *C. eugenioides* et pas d'amplifications récentes.

Divo a donc certainement joué un rôle dans la différenciation des groupes génétiques de *C. canephora*. Il serait intéressant de rechercher des transcrits et d'éventuellement détecter une expression différentielle de *Divo* en fonction de stresses biotique ou abiotique dans les génomes de *C. arabica* et ses progéniteurs. Des données RNAseq de l'équipe CoffeeAdapt peuvent être étudiées avec des outils dédiés à l'analyse de l'expression différentielle des gènes et ET du génome. Je n'ai malheureusement pas eu le temps de commencer ces analyses, mais j'ai testé l'outil « TEtools » (Lerat et al. 2016), qui avec une base de données de séquences d'ET du génome étudié, calcule l'expression différentielle d'ET dans un jeu de données RNAseq. Étant donné les copies très récemment insérées dans le génome de *C. canephora* et en moindre mesure dans

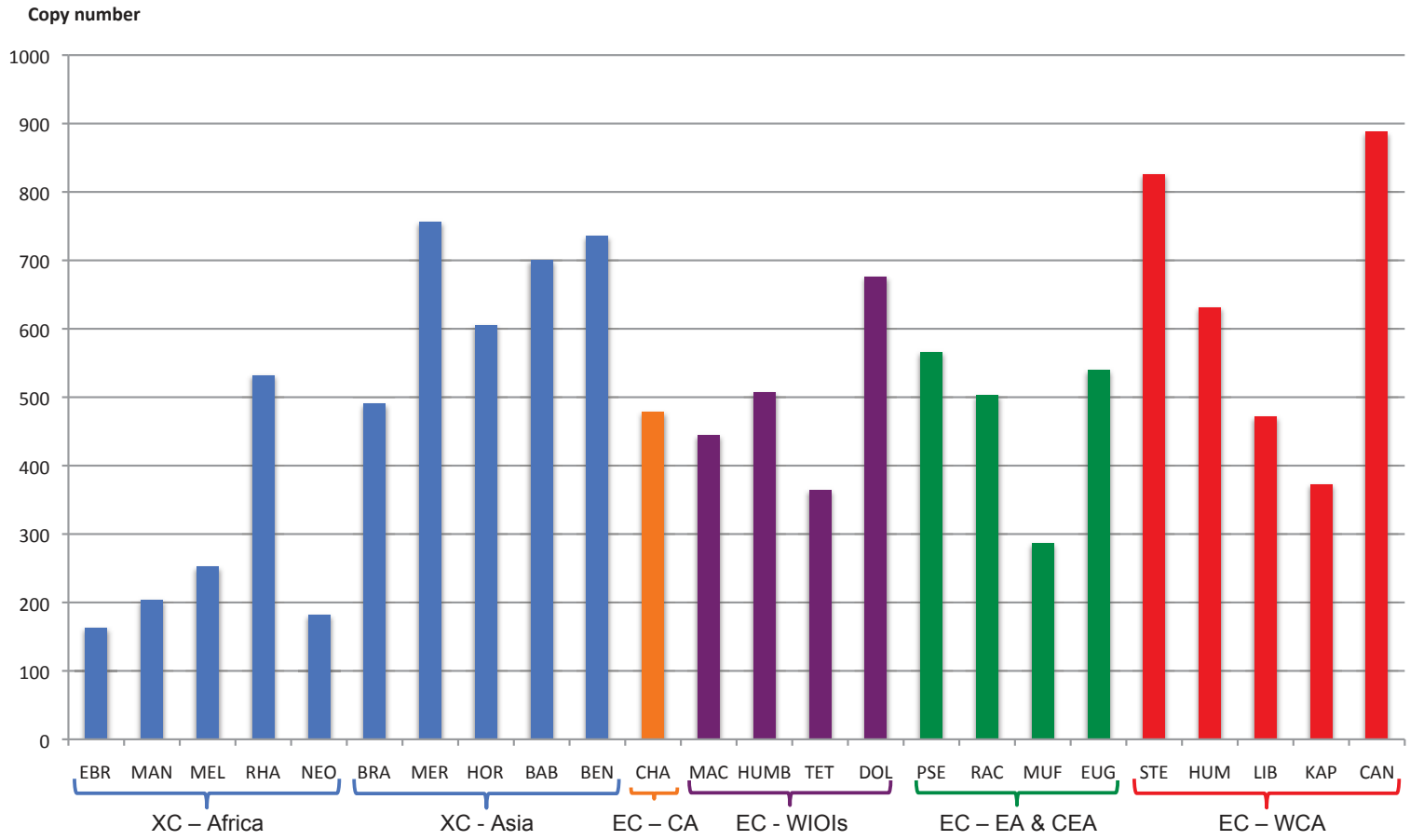


Figure 13 : Estimation du nombre de copies de *Divo* dans 24 espèces diploïdes de caféiers. Les estimations ont été calculées avec Bowtie2 (alignement end-to-end). Les espèces sont ordonnées selon la phylogénie. XC = Xeno-Coffea ; EBR : *Coffea ebracteolata* ; MAN : *C. mannii* ; MEL : *C. melanocarpa* ; RHA : *C. rhamnifolia* ; NEO : *C. neoleroyi* ; BRA : *C. brassii* ; MER : *C. merguinsis* ; HOR : *C. horsefieldiana* ; BAB : *C. benghalensis* var. *bababudani* ; BEN : *C. benghalensis* ; EC = Eu-Coffea ; CHA : *C. charrieriana* ; MAC : *C. macrocarpa* ; HUMB : *C. humblotiana* ; TET : *C. tetragona* ; DOL : *C. dolichophylla* ; PSE : *C. pseudozanguebariae* ; RAC : *C. racemosa* ; MUF : *P. mufindiensis* ; EUG : *C. eugenioides* (DA56 - Kenya) ; STE : *C. stenophylla* ; HUM : *C. humilis* ; LIB : *C. liberica* ; KAP : *C. kapakata* ; CAN : *C. canephora* (BUD15 – Uganda). CA = Afrique centrale ; WIOIs = îles de la région ouest de l'Océan Indien ; EA & CEA = Afrique de l'est et centre-est ; WCA = Afrique de l'ouest et du centre.

celui de *C. arabica*, on peut supposer que *Divo* est actuellement actif dans ces génomes.

Ayant accès à 24 génomes d'espèces/sous-espèces diploïdes de caféiers sauvages séquencés en Illumina (Projet G13, communication orale PAG 2015, Tableau 1), j'ai vérifié la présence de *Divo* dans ces génomes (Figure 13). Il ne semble pas y avoir de grande variation du nombre des copies, mais certaines espèces, notamment celles du clade Xeno-Coffea d'Afrique (excepté *C. rhamnifolia*) semblent en contenir moins que les autres. En effet, *C. rhamnifolia* est originaire d'Afrique de l'est et présente un nombre de copies du même ordre de grandeur que celui des Eu-Coffea d'Afrique de l'est. Paradoxalement, *C. mufindiensis,* qui est d'Afrique de l'est, montre un nombre de copies du même ordre que celui des Xeno-Coffea. Globalement, le nombre de copies estimé ne semble pas varier en fonction de la répartition géographique des espèces, exception faite des Xeno-Coffea d'Afrique. Des caractéristiques propres à ces espèces provoquent peut-être une élimination ou une plus faible activité de *Divo* dans ces génomes.

L'extraction des domaines RT de 16 espèces sur les 24 (à ce moment les génomes de *C. rhamnifolia*, *C. neoleroyi*, *C. melanocarpa*, *C. merguensis*, *C. charrieriana*, *C. mufindiensis*, *C. liberica* et *C. kapakata* n'étaient pas disponibles) a été réalisée (Tableau 1). Malgré la présence d'espèces du clade des Xeno-Coffea (absents dans l'étude de Hamon et al. 2011), l'arbre phylogénétique (NJ) réalisé avec les RT de *Divo* ne montre pas de groupements particuliers selon les espèces ou les zones géographiques (Figure 14). On peut apercevoir de petits groupements (de *C. canephora* notamment) soutenus mais toutes les espèces sont confondues avec peu de support statistique pour les grands clades. On peut cependant observer que certaines branches avec un petit nombre d'espèces différentes sont courtes et soutenues, ce qui pourrait suggérer une activité récente de *Divo* dans ces génomes.

Divo est donc un LTR-RT *Copia* discret, en faible nombre de copies et présent dans les génomes de plantes Dicotylédones uniquement. Il semble présenter une activité ancienne dans toutes les espèces de caféiers (présent dans les formes ancestrales). Son activité a été plus importante dans le génome de *C. canephora*, accompagnant sa diversification et montrant des irruptions d'insertion très récentes. Il pourrait être utilisé comme outil moléculaire pour l'étude de la diversité génétique de *C. canephora*, (d'autres groupes génétiques pourraient exister dans la partie plus à l'est et au nord-est

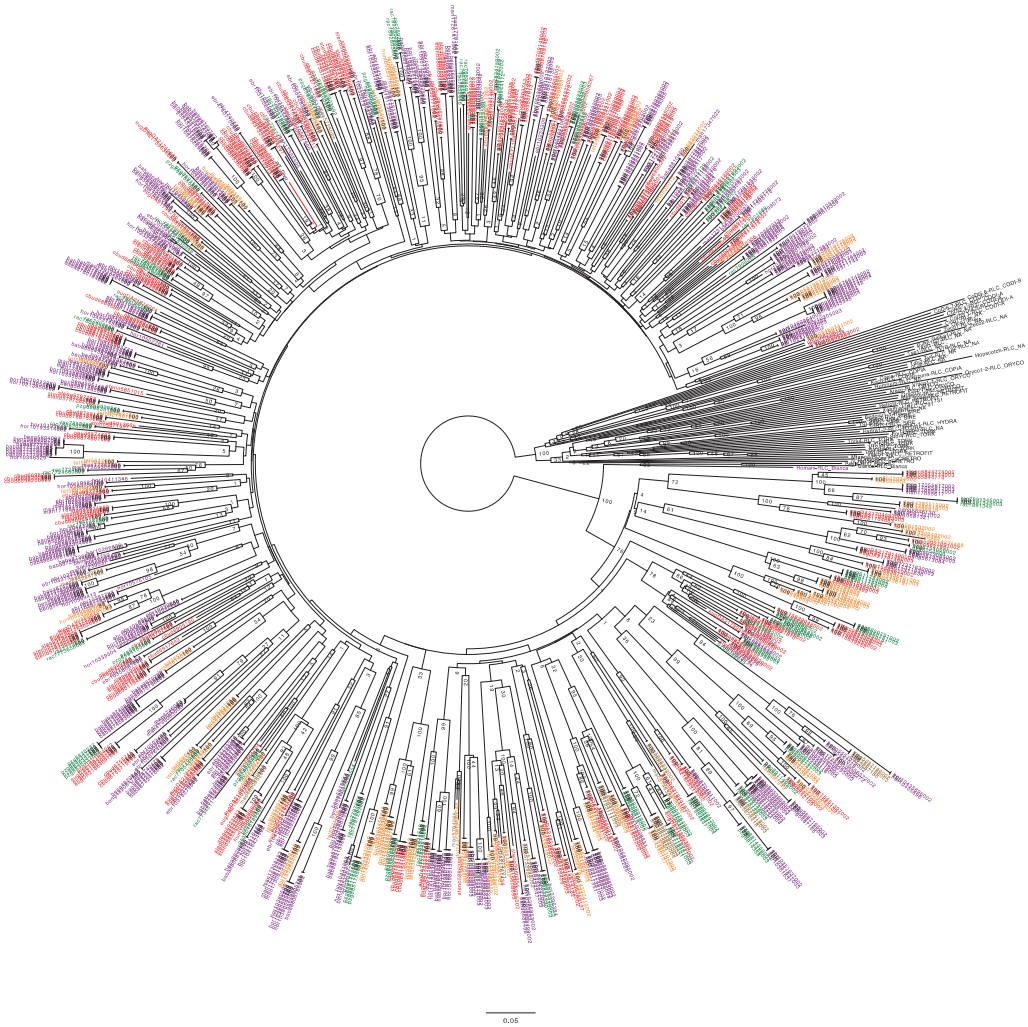


Figure 14 : Analyse phylogénétique (NJ – 100 répétitions de bootstraps) basée sur 1022 domaines RT de *Divo* dans 15 espèces/sous-espèces diploïdes de *Coffea*. Les branches et noms en noir sont les domaines de référence extraits de GyDB et de Wicker et al. 2007 pour *Bianca*. Les espèces sont colorées selon leur répartition : Afrique de l'ouest et du centre en rouge, Afrique de l'est en vert, îles de l'océan indien en orange et marron et Asie en violet.

de sa distribution), ainsi que pour la recherche de l'origine géographique du sous génome *C. canephora* chez *C. arabica*. Les maintiens différentiels de *Divo* et de *Bianca* (uniquement des séquences très dégradées de *Bianca* chez les Monocotylédones et des séquences complètes pour *Divo* chez les Dicotylédones), suggèrent que les *Bianca* ont un cycle de vie et/ou une régulation différents. Ainsi, l'histoire évolutive de *Divo* et de *Bianca* se traduirait aujourd'hui par une famille en perte d'activité chez les Monocotylédones et encore active chez les Dicotylédones.

Chapitre 5 - L'analyse des LTR-rétrotransposons *SIRE* indique une probable origine parentale ougandaise de *Coffea arabica*

1. Contexte

Le chapitre 3 traitait de l'analyse du séquençage partiel et de la composition en ET de 11 espèces de caféiers cultivées (*C. arabica* et *C. canephora*) et sauvages. Deux observations nous ont paru pertinentes. D'une part, la lignée des *SIRE* (*Copia*) montrait une variation du nombre de leurs copies suivant leur appartenance aux groupes bio-géographiques. D'autre part, l'estimation du nombre de copies présentes dans plusieurs accessions d'une même espèce a montré des variations sensibles du nombre des éléments (*C. canephora* : accessions IF410 d'Afrique de l'ouest, HD 200-94 de la République Démocratique du Congo et BUD15 d'Ouganda ; *C. eugenioides* : accessions BU-A d'Ouganda et DA56 du Kenya) (Figure 15).

L'analyse plus précise de cette lignée dans le génome annoté de *Coffea canephora* (Denoeud et al. 2014) a permis de montrer l'existence de trois familles de *SIRE*, A, B et C, dans ce génome (Darré 2014). Ce travail sur les *SIRE* a été affiné par l'analyse des génomes séquencés en PacBio.

2. Implication personnelle

Pour cet article, j'ai repris le travail de Thibaud Darré et comparé les trois familles décrites avec les trois familles que j'ai également détecté dans les génomes PacBio. J'ai par la suite réalisé l'annotation des copies complètes trouvées dans ces génomes et extrait les domaines RT et les LTR pour les analyses phylogénétiques et d'estimation de dates d'insertion des copies complètes. J'ai ensuite recherché les zones génomiques communes à *C. canephora*, *C. arabica* et *C. eugenioides* par dot-plot (Gepard (Krumsiek et al. 2007)) puis vérifié la présence de copies complètes et communes de *SIRE* dans ces régions. J'ai réalisé les analyses BLAST et phylogénétiques détaillées dans la partie

« Methods » de l'article. J'ai ensuite participé à la construction des figures et à la rédaction de l'article.

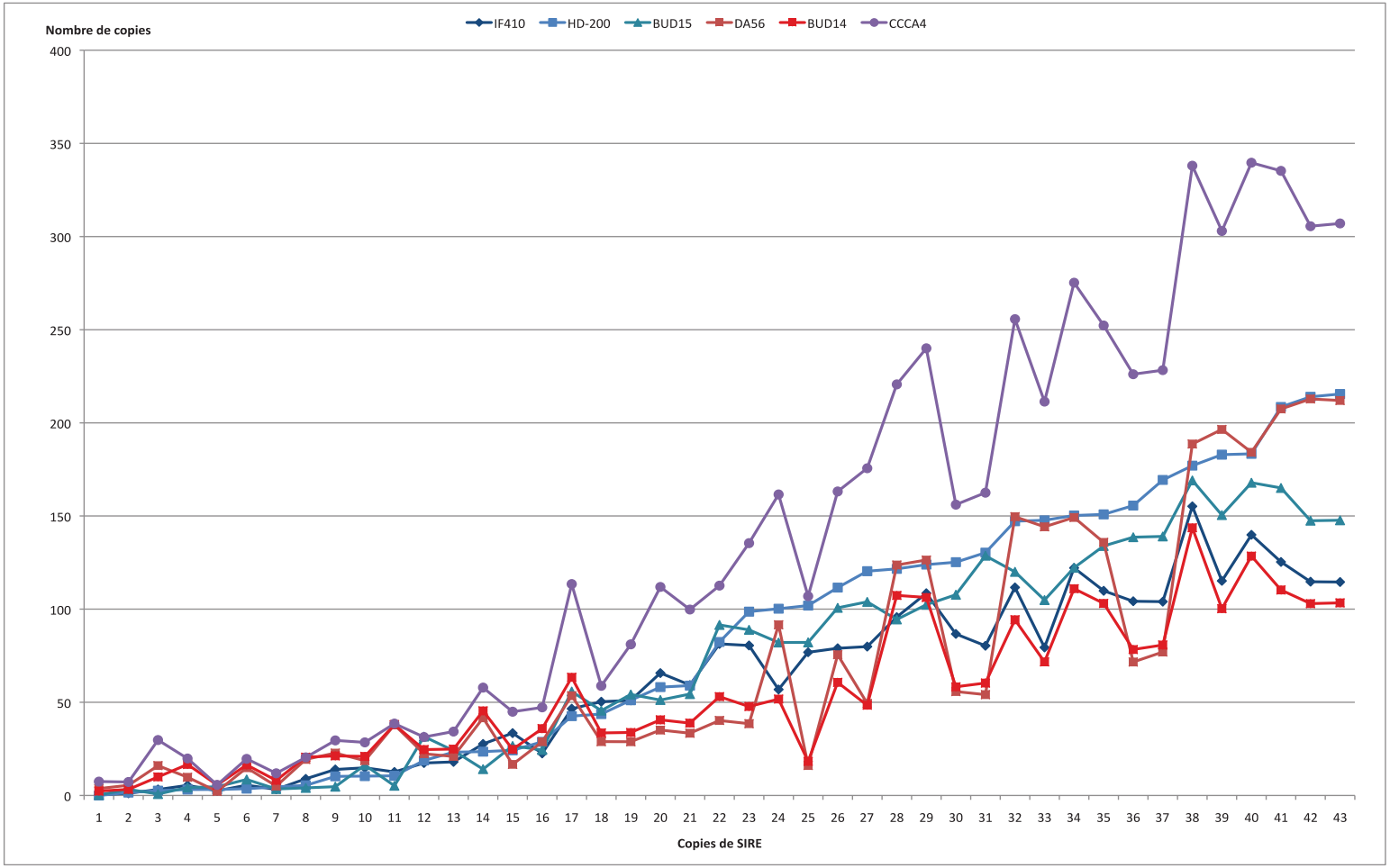


Figure 15 : Estimation du nombre de copies des *SIRE* dans un jeu de données 454 de *C. canephora*, *C. arabica* et *C. eugenioides*.
IF410 : *C. canephora* de Guinée ; HD-200 : *C. canephora* de RDC ; BUD15 : *C. canephora* d'Ouganda ; DA56 : *C. eugenioides* du Kenya ; BUD14 : *C. eugenioides* d'Ouganda ; CCCA4 : *C. arabica* cultivé.

Analysis of *SIRE* LTR-retrotransposons expansion and diversity indicate the likely Ugandese geographic parental origin of *Coffea arabica*.

Mathilde Dupeyron [1,2 *]

Dominique Crouzillat [3]

Alexandre de Kochko [1]

Thibault Darré [1]

Perla Hamon [1]

Romain Guyot [2]

[1] IRD UMR DIADE, EvoGec, BP 64501, 34394 Montpellier Cedex 5, France

[2] IRD UMR IPME, CoffeeAdapt, 911 Avenue Agropolis, 34394, Montpellier cedex 5, France

[3] Nestlé R&D Tours, Notre-Dame d'Oé, Tours, France

*Corresponding Author: Mathilde Dupeyron, Institut de Recherche pour le Développement (IRD), UMR IPME, BP 64501, 34394 Montpellier Cedex 5, France, mathilde.dupeyron@ird.fr

ABSTRACT

Background: In plant genomes, the activity of LTR-retrotransposons has a deep impact on genome structure and evolution, and their proliferation may accompany the diversification of species.

Results: To study the origin of *C. arabica*, we analyzed the *SIRE* LTR-retrotransposons in the *C. arabica* genome and in sequenced accessions of its diploid progenitors: *C. canephora* and *C. eugenioides*. We found that independent *SIRE* activations occurred simultaneously both in *C. canephora* and in *C. eugenioides*. However, no burst of activation following the polyploidization was observed. Phylogenetic analysis of *SIRE* families suggests that the parental genomes of *C. arabica* may be related with *C. canephora* and *C. eugenioides* populations originated from Uganda.

Conclusion: Altogether, our results suggest that the proliferation of the *SIRE* families might be one of the evolutionary forces that act on diploid species in the *Coffea* genus.

Keywords: *SIRE*, LTR-RT, *Coffea arabica*, evolution

BACKGROUND

Mobile genetic elements are DNA components present in all the genomes studied until now. Transposable elements (TEs) are classified according to their mode of mobility: Class I TEs or retrotransposons, use a "copy-and-paste" mechanism, whereas Class II TEs or DNA transposons, use a "cut-and-paste" mechanism. These two classes are then divided into orders, super-families, lineages, families and sometimes sub-families (Wicker et al. 2007). In plant genomes, retrotransposons can be particularly invasive, leading to genomic 'obesity' (Kumar and Bennetzen 1999; Lisch 2013). This propensity to accumulate hundreds or even thousands of copies can be stimulated by the transposition mechanism used, as successful retrotransposition produces for a given element a new DNA copy. Retrotransposons and especially Long Tandem Repeat retrotransposons (LTR-RTs), are the main TEs found in plant genomes. They are composed of *Gypsy*, *Copia*, Bel/Pao, Caulimoviruses and Retroviridae, with *Gypsy* and *Copia* being the two major LTR-RTs super-families in plants. *Copia* and *Gypsy* differ in the organization of the *Gag* and *Pol* regions and can be classified thanks to the conserved domains (especially the reverse transcriptase, RT) found in these regions (Havecker et al. 2004). They appear to be ancient as they are present in a large range of eukaryote genomes (Llorens et al. 2009). A clear relationship can be established between plant genome sizes and the proportion of LTR-RTs detected. Therefore, small genomes like *Utricularia gibba* or *Arabidopsis thaliana* contain only about 2.5 to 4% of LTR-RTs (The Arabidopsis Genome Initiative 2000; Kejnovsky et al. 2012; Ibarra-Laclette et al. 2013), whereas bigger genomes like *Triticum aestivum* or *Zea mays*, can contain them up to 63 and 75% (Schnable et al. 2009; Brenchley et al. 2012). The availability of next-generation sequencing data permits to study more and more complex genomes and to investigate on their structure, functioning and evolution. The huge amount of genomic information for both model and non-model species makes possible an overall understanding of genomes organization and evolution. Part of this information is the increasing interest to investigate on the TEs dynamics.

Coffea genus, belonging to the Rubiaceae family, is a young genus (around 12 My, Hamon et al. 2017) composed of 139 species (Couturon et al. 2016) with a distribution in various habitats in inter-tropical forests of Africa, Western Indian Ocean islands, India, Asia and Australasia (Davis et al. 2011). All species are diploids (2n = 2x = 22 chromosomes) except *C. arabica* (Bouharmont 1959; Louarn 1976). Chromosomal structure, genetic and genomic

studies of diploid species showed that they share a common genome and have discrete chromosomal differences (Hamon et al. 2009, 2015).

Originating from a recent and natural hybridization between *C. canephora* and *C. eugenioides* (Lashermes et al. 1999, Yu et al. 2011), *C. arabica* is an allotetraploid (genome of 1.3 Gb) of high economic importance for many developing countries. The low divergence between the two parental sub-genomes makes hard the assembling and correct separation of homeologous sequences.

The draft sequence of *C. canephora* genome, the diploid cultivated species is composed with 50% TEs with 42% LTR-RTs (Denoeud et al. 2014). A previous study (Hamon et al. 2011) using two *Copia* LTR-retrotransposons (*Divo* and *Nana* identified from a *C. canephora* BAC clone), showed that *Divo* permits the identification of the five *C. canephora* genetic groups previously described by Gomez et al. (2009) while *Nana* is associated to species differentiation.

An in-depth study of the *Divo* family was conducted using the draft genome of *C. canephora* (Denoeud et al. 2014) and PacBio sequencing for *C. canephora*, *C. eugenioides* and *C. arabica* (generated under The Arabica Coffee Genome Consortium, 2014) (Dupeyron et al. 2017). *Divo* elements are only found in dicots and form a sister clade to *Bianca* elements only present in monocots. *Divo* is subjected to small activity peaks at different periods in the two diploid genomes; it did not suffer from deep rearrangements in *C. arabica* following the polyploidization event unlike to reports in polyploid genomes such as *Spartina anglica* or *Nicotiana tabaccum* (Comai et al. 2003; Parisod et al. 2010). Moreover, the last insertion of *Divo* shared by *C. canephora*, *C. eugenioides* and *C. arabica* suggested that the polyploidization would be even more recent than previous high estimations (Yu et al. 2011).

More recently, information on *Coffea* genomes composition in TEs for 11 species was gathered from partial genome sequencing. The *SIRE* lineage (*Copia* super-family) elements of *C. canephora* are distributed into three distinct families. These elements are not equally distributed among *Coffea* species (Guyot et al. 2016). Furthermore, first rough estimates of *SIRE* elements copy number suggested variation between accessions of *C. canephora* (Darré 2014). Among *Copia* lineage, *SIRE* are atypical elements as they are the only ones to carry an envelope-like gene (*ENV*) (Laten et al. 1998). Widespread lineage in plants, they are supposed to be ancient. As they showed recent insertion activity in most of the studied plant genomes (Bousios et al. 2010), they could give valuable information on *Coffea* genome evolution, and especially regarding their dynamics in the allotetraploid *C. arabica*.

In this study, thanks to the *SIRE* repertoire from *C. canephora* (assembled from short-read sequences, *C. canephora* accession HD-200-94, Denoeud et al. 2014) and the forthcoming *C. arabica, C. canephora* and *C. eugenioides* genomes (long read sequencing, PacBio technology, accessions Et39, HD-200-94 and BU-A respectively, ACGC 2014), the *SIRE* elements were mined and analysed. We found that independent *SIRE* activations occurred simultaneously both in *C. canephora* and in *C. eugenioides*. However, no burst of activation following the polyploidization was observed. Phylogenetic analysis of *SIRE* families suggests that the parental genomes of *C. arabica*, may be related with *C. canephora* and *C. eugenioides* populations originated from Uganda. Altogether, our results suggest that the proliferation of the *SIRE* families, might be one of the evolutionary forces that act on diploid species in the *Coffea* genus.

MATERIAL AND METHODS

Genomic sources

Three genomes generated under the Arabica Coffee Genome Consortium (ACGC 2014) sequenced with the single molecule real-time (SMRT, Pacific Biosciences - PacBio) sequencing technology: *C. canephora* (accession DH 200-94), *C. arabica* (accession Et39) and *C. eugenioides* (BU-A) accounting respectively for 679, 1,060 and 789 Mb of unordered contigs) were used in this study. The *C. canephora* genome sequence was assembled into pseudo-molecules corresponding to each chromosome, plus an additional pseudo-molecule ("0") corresponding to the remaining genetic information matching with any chromosome.
In addition, *C. canephora* (accession BUD15) and *C. eugenioides* (accession DA56) genomes sequenced with the Illumina technology were used in this study.

Classification and annotation of *SIRE Copia* LTR-RTs

SIRE sequences were extracted from the LTR_STRUC outputs of the three PacBio genomes as following: from these outputs, potential *SIRE* elements were identified with BLASTX similarities against a database of Gag and Pol domains (available at GyDB, http://www.gydb.org/ Llorens et al. 2011). The amino acid RT domains of all recovered *SIRE* were extracted from each genome as described in Guyot et al. (2016), with a minimum length of 200 amino acid residues.

RT reference domains from GyDB were added to these sequences. All of them were aligned with Muscle (Edgar 2004) and used to construct a bootstrapped neighbor-joining (NJ) tree (1000 replicates) with ClustalW (Thompson et al. 1994) edited with Figtree (http://tree.bio.ed.ac.uk/software/figtree/) and Inkscape (http://www.inkscape.org/).
Precise annotation of *SIRE* complete copies was made using BLASTX and dot-plot alignments with reference domains (Gypsy Database 2.0 web site) (Sonnhammer and Durbin 1996) and LTR_Finder Xu and Wang 2007), http://tlife.fudan.edu.cn/ltr_finder/). Final annotations were edited with Artemis (Rutherford et al. 2000). The presence of cis-acting regulatory elements in the LTRs was mined for the reference sequences for each *SIRE* family and for each genome in PlantCARE (Lescot et al. 2002).

Identification of three families of *SIRE in C. canephora* and *C. eugenioides*

The organization of *SIRE* lineage in three clusters (Guyot et al. 2016) was checked by launching NJ trees of LTR sequences from the *SIRE* detected by LTR_STRUC. The trees were rooted with the four RT domains references of *SIRE* from GyDB (from *Arabidopsis thaliana*, *Glycine max*, *Zea mays* and *Setaria italica*) and edited with Figtree and Inkscape. For one of the three SIRE clusters (named family C), two references were used, as two sub-types of copies are found with short and long LTRs.

Copy number and insertion time of the three *SIRE* families in *C. canephora*, *C. arabica* and *C. eugenioides*

Assessment of *SIRE* copy number in the *C. canephora*, *C. arabica* and *C. eugenioides* PacBio genomes (ACGC 2014) was carried out with Censor (Kohany et al. 2006). A complete element is considered when it contains both ORFs Gag and Pol and a minimum of 99% sequence identity between both LTRs. The complete element (and the most conserved) found in each of the three genomes was used as a reference for the similarity searches. To estimate the copy number of each family separately, one complete and most conserved element for each of the three families was used as reference for Censor. This was not possible for one of the families in *C. arabica*, for which LTR_STRUC found very few copies, so reference sequences of *C. canephora* and *C. eugenioides* were both used as references for Censor in *C. arabica*. Moreover, two types of copies, one with short LTRs, the other with long LTRs, were found for family C in both *C. canephora* and *C. eugenioides* genomes. Therefore, estimations

were done with two reference copies for this family. A "reference" element was determined for each family in each genome. It is the most conserved element found as follow: complete structure with all the enzymatic domains and if possible, LTRs identity ≥ 90% and no stop codons in the ORFs. These *SIRE* copies were then used to launch Censor separately for each family (two times for family C) on the three genomes.

The sequences of each family are often close to each other, so the affiliations to the three families of each copy found by Censor was checked by extracting the RT domain as described in Guyot et al. (2016). For the sequences without RT domain or presenting a too short RT sequence, the LTR sequences were extracted. RT domains in amino acid sequence and LTR were aligned with Muscle (Edgar 2004) and bootstrapped (1,000 replicates) NJ phylogenetic trees were computed with ClustalW (Thompson et al. 1994).

Each LTR sequence of the copies found by LTR_STRUC, when longer than 800 pb, was used for dating estimations. In addition, 27, 6 and 41 LTR sequences from the copies found by Censor were used for *C. canephora*, *C. arabica* and *C. eugenioides*, respectively. Details about the LTR number and the family of *SIRE* they belong are given in Supplementary Material Table S1. The two LTRs were aligned using Stretcher (EMBOSS), and the divergence (K) was calculated using the Kimura 2-parameter method implemented in Distmat (EMBOSS). The insertion dates (T) were estimated using the formula T = K/2r (SanMiguel et al. 1998) where average base substitution rates (r) is of $1.3e^{-8}$ (Ma and Bennetzen 2004).

Comparison of two accessions among *C. canephora* and *C. eugenioides* and organization of the *SIRE* elements in these genomes

A second accession of *C. canephora* (BUD15 from Uganda) is available as well as of *C. eugenioides* (DA56 from Kenya). RT domains of *SIRE* families have been mined in these two genomes by BLASTX analysis with complete copies of *SIRE* families from *C. canephora* (accession HD200-94) as database. Then, NJ trees with 1,000 bootstrap replicates were computed in ClustalW and edited with Figtree and Inkscape.

To observe or not differences between *SIRE* organization between each accession and between *C. canephora*, *C. eugenioides* and *C. arabica*, NJ trees were also computed in the same manner than above with the *SIRE* RT domains found in the three genomes and each accession of *C. canephora* and C. *eugenioides*.

Orthologous locations of the *SIRE* in the three genomes

To search for orthologous regions between the three genomes, we used Gepard (Krumsiek et al. 2007) to align each *C. canephora* pseudo-molecule with *C. eugenioides* scaffolds. Then each shared region was aligned to *C. arabica* scaffolds and only regions shared between the three genomes were kept. Localization of each complete copy of *SIRE* found by Censor in *C. canephora* pseudo-molecules was checked in these shared regions. When copies of *SIRE* were found in orthologous location, the family they belong was checked and their insertional age estimated when the LTRs were long enough (more than 800 bp). Potentially common copies were aligned with Stretcher to obtain their identity percentage. Copies of the same family in orthologous regions with more than 90% identity were considered as shared copies.

Search of *SIRE* transcripts in *C. arabica* sequenced cDNA data

IsoSeq cDNA data (cDNA sequenced via PacBio technology) of *C. arabica* (from ET39 leaves) transcriptome are available under the ACGC consortium, so the presence of complete *SIRE* transcripts was mined by BLASTN searches (e-value $\leq 1e^{-04}$). The results were filtered out according to identity percentage (> 80%) and length (> 2000 bp). Then potential *SIRE* transcripts were extracted and analysed by dot-plot alignments (Sonnhammer and Durbin 1996) and BLASTX searches for detection of the Gag and Pol domains. They were aligned with complete copies of family A and B using Muscle and the alignment was visualized in Seaview.

RESULTS

Structure of *SIRE* elements

LTR_STRUC identified 755 *SIRE* elements in *C. canephora* (HD-200) and 438 in *C. eugenioides* (BU-A) from the *C. canephora* and *C. eugenioides* PacBio sequenced genomes. The structure of the elements is typical for *SIRE lineage*: two LTRs of 1,015 and 1,022 bp for *C. canephora* and *C. eugenioides* respectively; three ORFs containing *Gag* and UBN2 domains, a *Pol* ORF with an integrase, a reverse-transcriptase and a RNAse H in this order and potentially a third ORF corresponding to an *ENV*-like domain (Laten et al. 1998). The overall length is between 7,640 and 10,625 bp on average. The PBS is the same in all of the annotated complete copies: TATCAGAGCTTGGTCTC – Met^{CAT}, and the PPT is present

against the 3' LTR: CAAAAAGGGGGAGAT (**Figure 1**). In *C. arabica* genome, LTR_STRUC identified 251 *SIRE* elements. Their structural characteristics are similar to those from *C. canephora* and *C. eugenioides* at the exception of different length of LTRs. Indeed, an important number of *SIRE* contains elements shorter than usually observed and their LTRs are between 300 and 350 bp in length. *SIRE* elements are classified into 3 families (A, B and C) according to Darré (2014) and Guyot et al. (2016).

Interestingly, No "functional" full-length copy, *e.g*. no frame-shift or stop codons in the ORFs and presence of all the domains required for the retrotransposition, has been found for the C family in the three genomes, whereas "functional" copies are found for families A and B. Moreover, *ENV*-like domain was absent in family C (using LTR_STRUC sequences and Censor complete copies; **Figure 1**).

According to the properties of RT domains, particularly conserved among LTR-RTs, we used them to infer the *SIRE* structure. LTRs were further used for copies insertion time estimation.

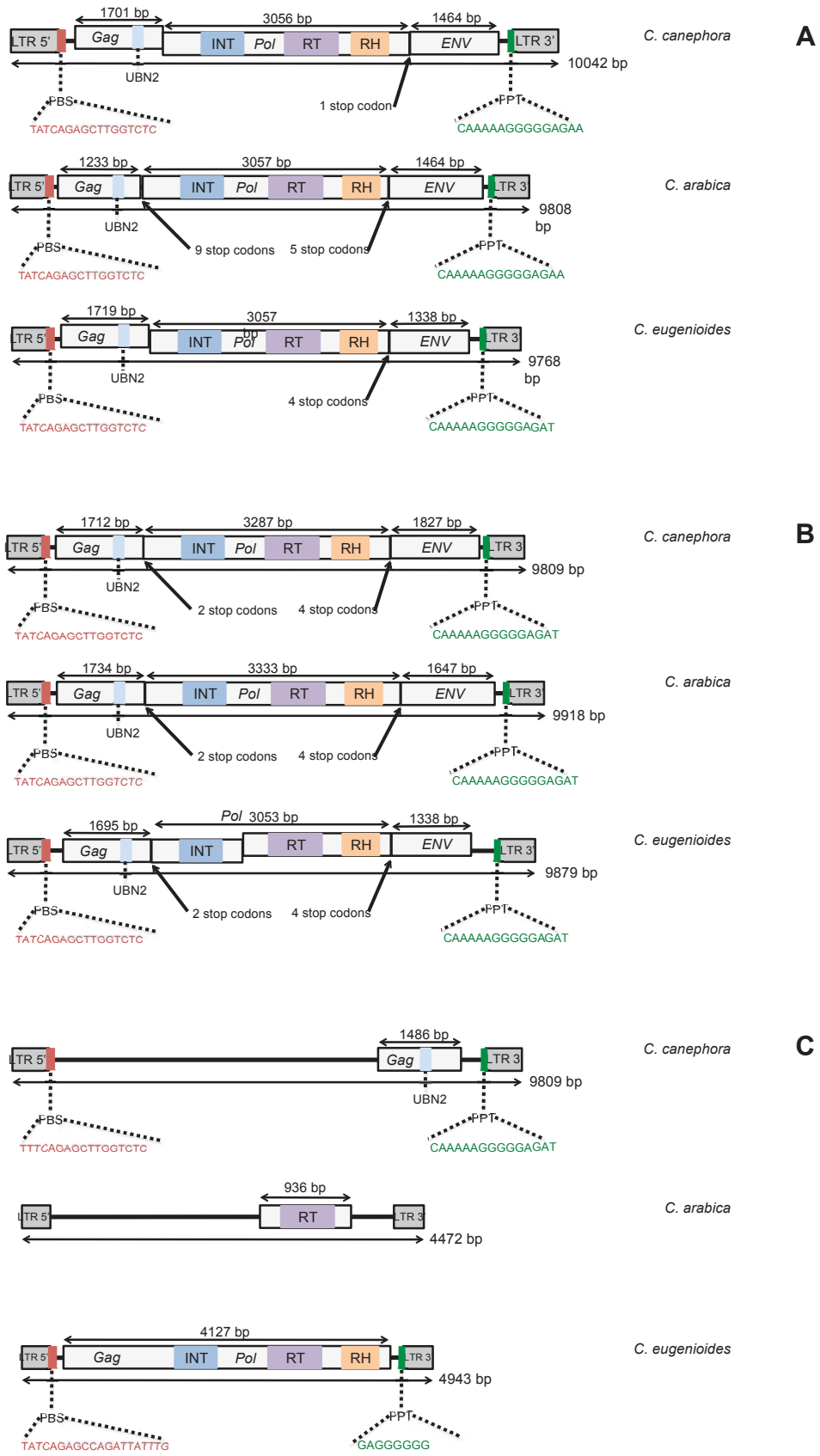


Figure 1. Structure of *SIRE* LTR-retrotransposons found in the PacBio genomes of *C. canephora*, *C. arabica* and *C. eugenioides* by LTR_STRUC. A. *SIRE* family A, B. *SIRE* family B, C. *SIRE* family C.

Phylogenetic trees of *SIRE* in the diploid and allotetraploid coffee genomes.

The NJ phylogenetic trees of *SIRE* RT domains extracted from LTR_STRUC output for *C. canephora* genome (HD-200) showed a distribution into three clusters corresponding to families A, B and C as reported in Guyot et al. (2016) (**Figure 1 and 2**). Three similar clusters were found for the *C. eugenioides* genome (BU-A). Interestingly, other accessions of *C. canephora* (BUD-15) and *C. eugenioides (*DA-56) showed different tree structures (**Figure 2**). *C. canephora*, accession BUD15, showed additional groups (noted in grey) and reduced A and B families, when compared to accession HD-200. The *C. eugenioides* accession DA56 also showed major differences with BU-A, with the absence of differentiation between A and B families and new groups (in grey). These of tree structure variations observed at the intra species level suggested different evolution of the SIRE elements.

In *C. arabica,* the tree structure is similar to those of *C. canephora* (HD-200) and *C. eugenioides* (BU-A), despite few RT domains from the C family were recovered from the genome (**Figure 3A)**. To understand the origin of SIRE RT domains in *C. arabica*, RT from *C. arabica* and two *C. canephora* accessions (HD-200 and BUD15) were mixed and a NJ tree was drawn (**Figure 3B)**. A similar analysis was done with *C. eugenioides* accessions (**Figure 3C)**.

NJ trees with mixed samples of *C. arabica* indicate clearly that most *C. arabica* RT domains clustered with A and B families. Similar observation was done for *C. arabica* and *C. eugenioides* (**Figure 3A and B**).

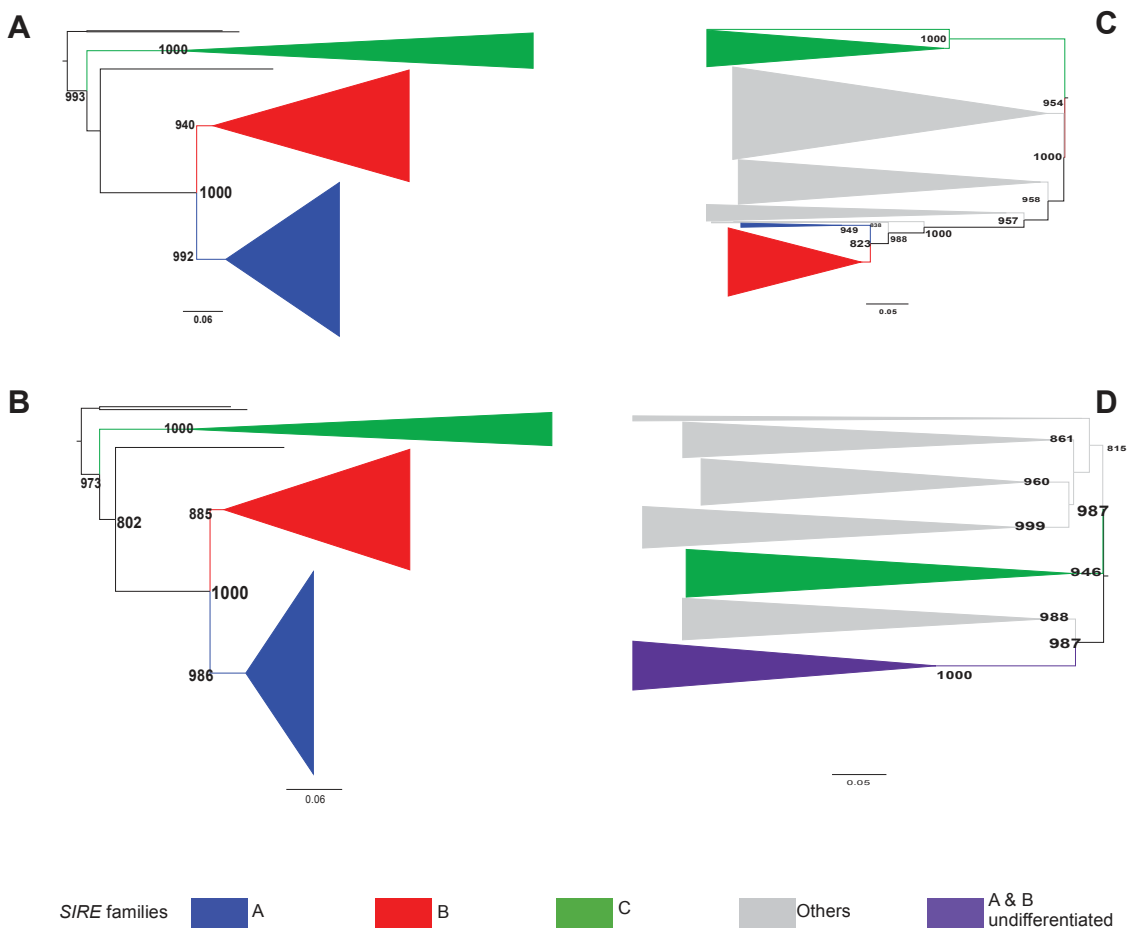


Figure 2. NJ phylogenetic trees of *SIRE* RT domains in *Coffea canephora HD 200* (A), and *C. eugenioides BU-A* (B), *C. canephora* BUD15 (C) and *C. eugenioides DA56* (D).
Black blanches represent outgroups, when used. For (A) and (B), *SIRE* RT domains were extracted from PacBio genome data using LTR_STRUC predictions. For (C) and (D), RT domains were extracted using assembled Illumina data.

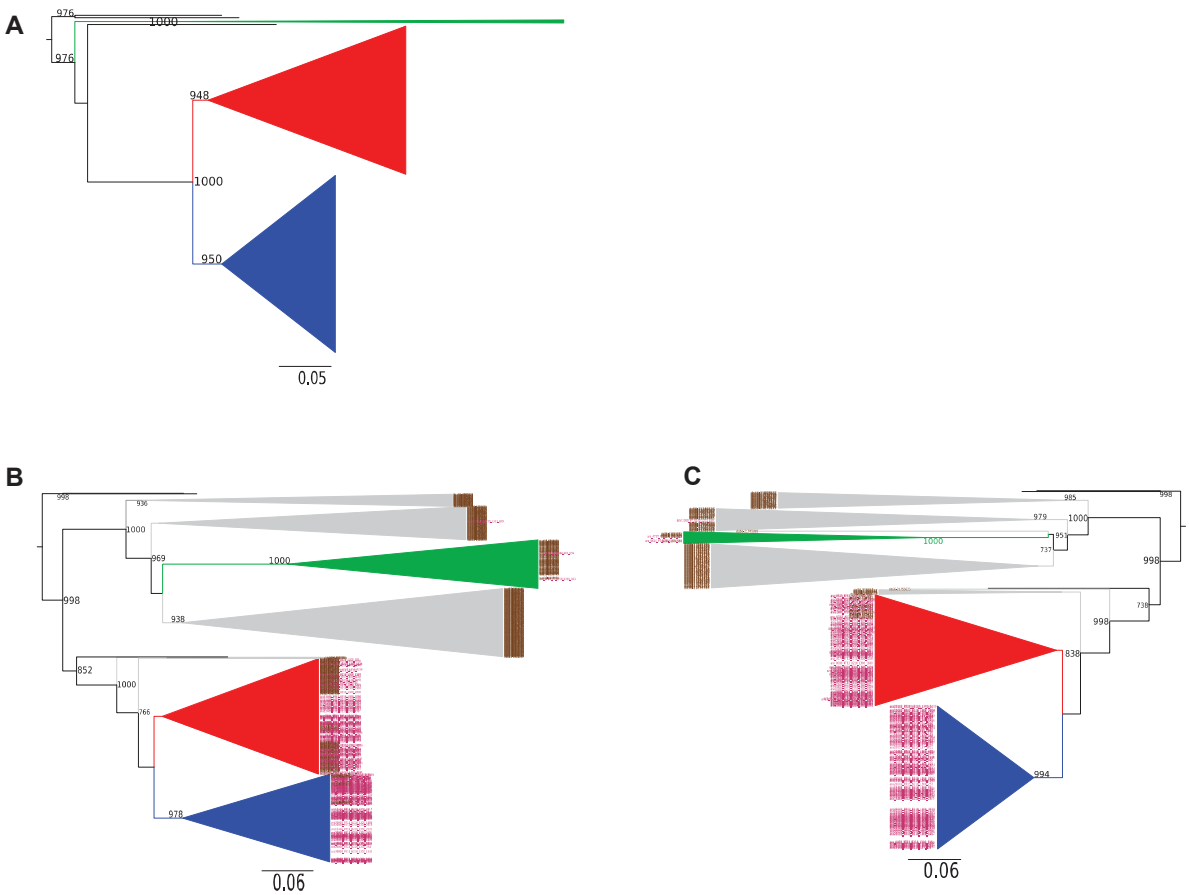


Figure 3. NJ phylogenetic trees of *SIRE* RT domains in *C. arabic*a ET39 (A) and in mixed sample between *C. arabic*a ET39, *C. canephora* HD 200 and *C. canephora* BUD15 (B) and in mixed sample between *C. arabic*a ET39, *C. eugenioides BU-A* and *C. eugenioides DA56* (C).

Copy number estimation of *SIRE* elements in three coffee genomes

The copy number estimates per family and genome as well as the number of solo-LTRs are indicated in **Table 1**. The estimations for the family C showed in **Table 1** are those from the copies with short LTRs, as they represent the highest number of conserved copies. The results for the two types of copies present for family the C are showed in **Supplemental Table S2.**

In *C. canephora*, the estimated total number of copies (intact copies 80-100 and copies 80-80 rules) is almost twice and four-fold higher for the family A than for the families B and C, respectively. However, the number of solo-LTRs is higher for family A, and similar between the families B and C. The ratio between solo-LTRs and intact copies is similar for the families A and B, but it was two-fold higher than for the family C.

For *C. eugenioides*, a different pattern was obtained since the family B copies were predominant (2.5 and 10 -fold more numerous than for the families A and C respectively). Like for *C. canephora*, the most important family in terms of intact copies number was also

the one with the highest number of solo-LTRs. Noteworthy, the ratio between solo-LTRs and intact copies was of the same order of magnitude for families A and B between the two diploid species. However, for the family C, it varies from 9.8, 1 and 704 for *C. canephora*, *C. eugenioides* and *C. arabica,* respectively.

For *C. arabica*, the family A showed a higher number of copies, 3.5 times more than copies of the family B, and solo-LTRs. The family B shows a medium-range copy number but with a high number of solo-LTRs. The family C presents a poor number of copies and partial copies (for copies with short LTRs), with elevated solo-LTRs/intact copies ratio except for *C. eugenioides* (**Table 1** and Supplemental Table S2).

Table 1. Copy number estimation of *SIRE* elements in *C. canephora*, *C. arabica* and *C. eugenioides*. Estimates were made using Censor with the "reference" copies for the three families of *SIRE* found in each genome by LTR_STRUC.

Species	*Coffea canephora*			*Coffea arabica*			*Coffea eugenioides*		
Family	A	B	C	A	B	C	A	B	C
Number of intact copies (80-100)	137	99	42	538	159	1	87	119	2
Number of copies (80-80)	135	48	25	307	85	24	82	289	30
Total	272	147	67	845	244	25	169	408	32
Number of partial copies (20-80)	475	239	309	364	156	31	411	823	49
Number of solo-LTRs	575	429	410	1793	620	704	278	763	2
Ratio solo-LTRs:intact copies	4.2:1	4.3:1	9.8:1	3.3:1	3.9:1	704:1	3.2:1	6.4:1	1

SIRE families underwent a recent activity in the three coffee genomes

LTR-based time insertion revealed that *SIRE* families underwent a burst in the two diploid genomes in the last 1 million year (**Figure 4 A**) and no specific burst was observed for *C. arabica*.

At the family level, recent activities (< 1 My) have been detected, but with specific patterns for each family (**Figure 4B, C and D**). The family A showed very recent activities for all coffee genomes, while the family B, showed more ancient activities and unstable recent activities. For the family C (no useful LTR copies have been found for *C. arabica*), an interesting variation of the activity is observable between *C. canephora* and *C. eugenioides*.

Altogether, very few copies have shown 100% of nucleotide identity between intra-element LTRs (data not shown), suggesting that very few copies could have been inserted in the modern past. However, transcriptomic activities have been detected using BLASTN searches against *C. arabica* sequenced cDNA. Two transcripts of the family C were detected (Figure S1).

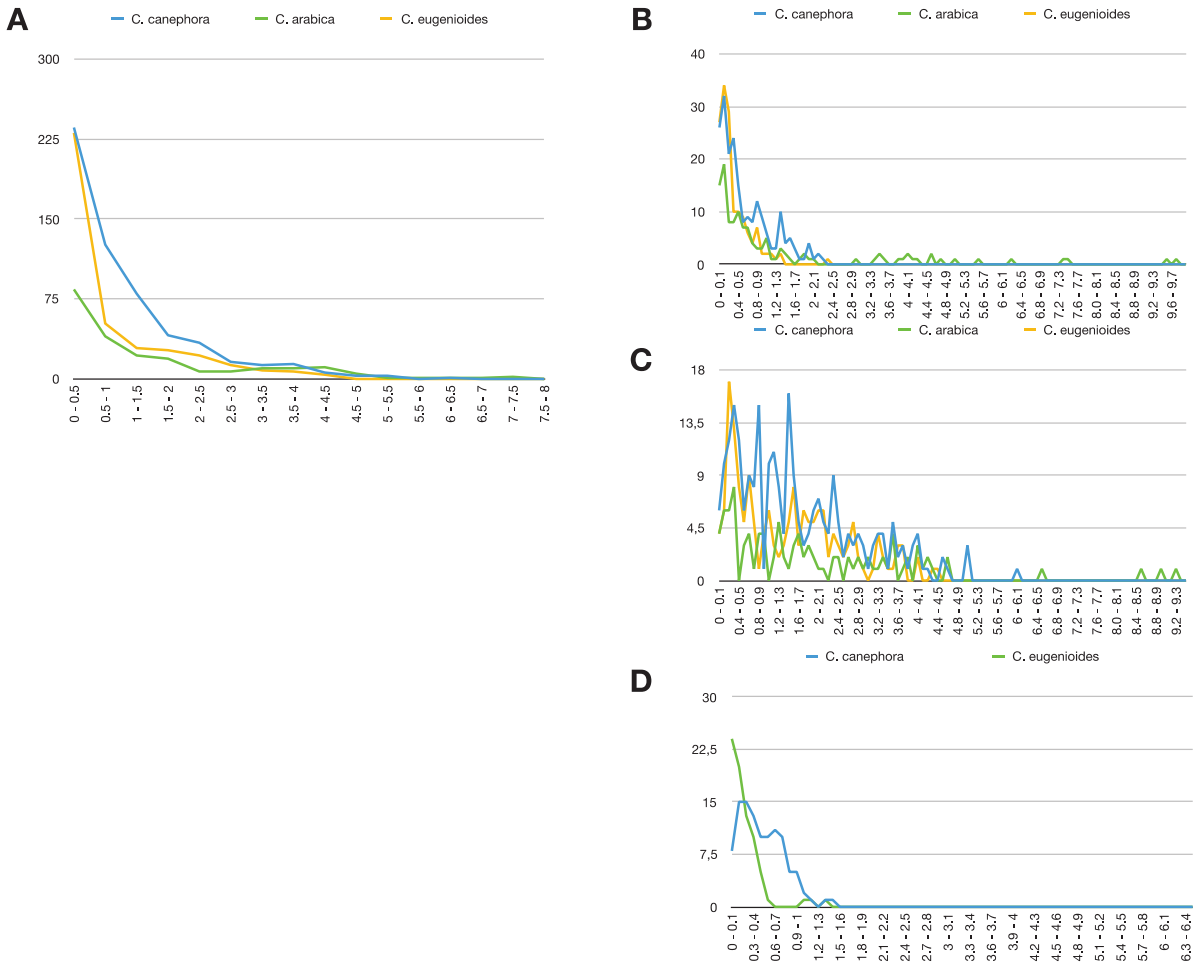


Figure 4. Insertion time analysis of the *SIRE* families in *Coffea canephora*, *C. arabica* and *C. eugenioides*. A. All families, B. Family A, C. Family B and D. Family C.

DISCUSSION

In this paper, based on PacBio and Illumina sequencing of three coffee species, we worked on the evolutionary analysis of only one lineage of LTR-retrotransposons called *SIRE*, an ancient lineage in plant genomes (Miguel et al. 2008; Bousios et al. 2010; Bousios and Darzentas 2013). *SIRE* in coffee are well separated into three families, called hereafter A, B and C, based on their similarity, phylogenetic and structural analyses (Guyot et al. 2016). A recent expansion of the three *SIRE* families can be observed, but with different patterns in diploid

and allotetraploid coffee trees. *SIRE* are largely eliminated in *C. arabica*. However, different composition and diversity of *SIRE* at the intra-species level suggest a deep influence of these elements on the diversification of diploid species.

C. arabica is a recent allopolyploid resulting from a natural hybridization between *C. canephora* and *C. eugenioides* (Lashermes et al. 1999). However, the precise geographic origin of the diploid progenitors was not yet identified, and the evolution of *C. arabica*, just after the interspecific hybridization, remains largely unknown.

LTR-retrotransposons are major players in the evolution of plant genomes, since their activities may impact deeply the genome structure and size and also influence gene expression. It was also reported that interspecific hybridization and polyploidization may also influence the LTR-retrotransposons activities (Parisod et al. 2010).

Our analysis suggests that expansions of *SIRE* occurred in the last 1.5 My, the family A and C showed a burst in all species before 1.5 My, while the family B showed a recent expansion but with unstable activities going to up to 5 My. The time frame of proliferation of *SIRE* indicates they started to expand before the hybridizations that get rise to *C. arabica*, 0.04 to 0.6 Mya.

In *C. arabica*, no new insertion of *SIRE* (i.e. < 100,000 years) has been detected in the polyploid when compared to its diploid modern progenitors. However, a high rate elimination of *SIRE* leading to solo-LTR via unequal recombination (Bennetzen et al. 2005; Grover and Wendel 2010) can be suggested. The ratio between solo-LTRs and complete copies, ranging between 3.2 and 6.4 are detected for the families A and B, counterbalancing expansions of these families. For the family C, the ratio between solo-LTRs and complete copies for elements with long LTRs (704:1) indicates a strong elimination process of these elements in *C. arabica*. Despite the insertion time of solo-LTR cannot be accurately estimated, these results suggest that *SIRE* have probably not participated to the diversification of *C. arabica* accessions.

By comparing the evolution of SIRE families in diploid and tetraploid coffee trees, we found a very contrasted situation at the intra-species level in *C. canephora* and *C. eugenioides*, and in *C. arabica* suggesting that the evolution of SIRE participated to the divergence of diploid species.

C. canephora is a species with a wide range of habitats and wide natural distribution, from sea level in Ivory Coast to altitude in Uganda and from humid forests of West Africa to woody savannah in Central African Republic (Berthaud 1986). *C. canephora* showed a large diversity and genetic differentiation (Musoli et al. 2009; Gomez et al. 2009), and the activity

of transposable elements could be associated to its capacity of adaptation to new and contrasted ecological niches. Overall variations of transposable elements composition between accessions of *C. canephora* were also observed using partial sequencing for HD-200 and BUD15 accessions (Guyot et al. 2016), suggesting that *SIRE* were not the only elements having participated to the diversification of this species. To further understand the *SIRE* and overall TE diversity in *C. canephora*, deep sequencing of representatives of different genetic groups (Gomez et al. 2009) could be undertaken in the future.

C. eugenioides is a species with a more reduced geographical repartition area than *C. canephora*, going from Kivu region in Democratic Republic of the Congo to West Kenya (Berthaud 1986). Very few molecular and diversity studies targeted *C. eugenioides,* and the diversity of this species remains largely unknown. Similarly to *C. canephora*, an outstanding diversity of *SIRE* was present in *C. eugenioides*. *C. eugenioides* accessions BU-A and DA56 are originated from Uganda and Kenya, respectively. These countries are separated by the Great African rift, an important geological feature that may be a source of ecological barriers (White et al. 1994), possibly resulting in intra-species differentiation.

By comparing *SIRE* differentiation in diploid species with *C. arabica*, we identified unexpected similar patterns. The NJ trees of *SIRE* RT domains from *C. arabica* and *C. canephora* accession BUD15 and *C. arabica* and *C. eugenioides* accession BU-A show similar phylogenetic patterns. *C. arabica* appears closer to *C. eugenioides* accession BU-A than accession DA56 (Ethiopia, Uganda and Kenya, respectively, **Figure 5**). Altogether, our results suggest that the parental genomes of *C. arabica*, may be related with *C. canephora* and *C. eugenioides* populations originated from Uganda.

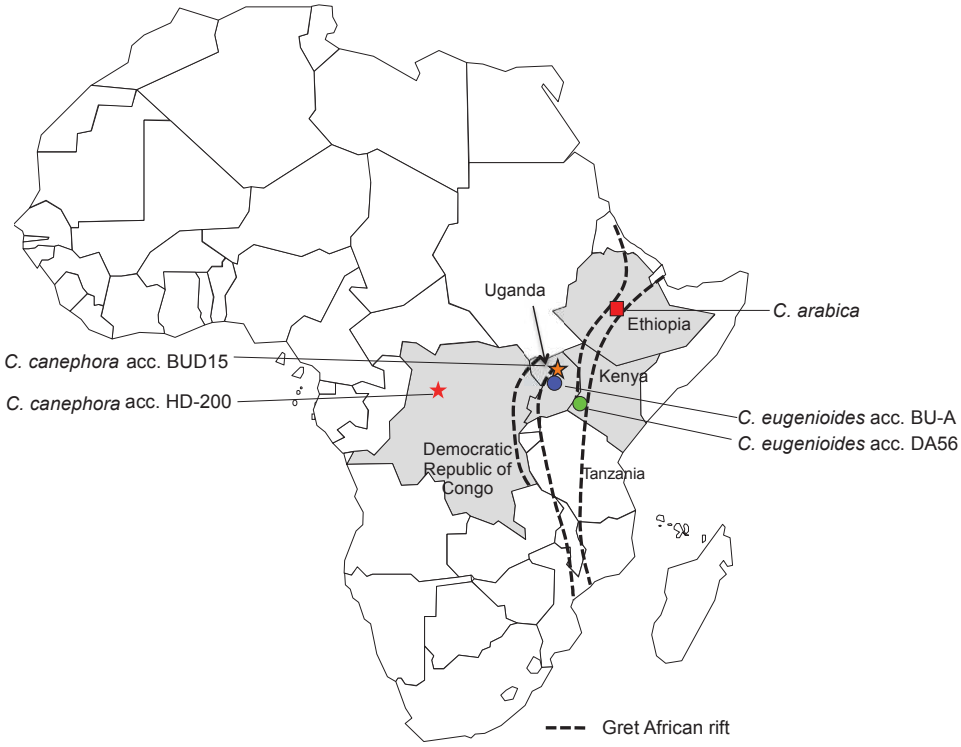


Figure 5. Geographical localization of *Coffea canephora* accessions HD 200 and BUD15, *C. arabica* and *C. eugenioides.*

CONCLUSION

Detailed analysis of the LTR-retrotransposons *SIRE* families in coffee trees provides important information to elucidate the parental geographic origin of *C. arabica* as well as molecular markers to study the diversification of its diploid progenitors: *C. canephora* and *C. eugenioides*. The proliferation of retrotransposons such as the *SIRE* families, might be one of the evolutionary forces that act on diploid species in the *Coffea* genus. To confirm the role of *SIRE* families and to elucidate the diversification at the genus level, further genome sequencing of numerous wild coffee species are now required.

REFERENCES

Adams MD, Celniker SE, Holt RA, et al (2000) The Genome Sequence of Drosophila melanogaster. Genetics 287:2185–2195. doi: 10.1126/science.287.5461.2185

Agren JA, Wright SI (2015) Selfish genetic elements and plant genome size evolution. Trends Plant Sci 1–2. doi: 10.1016/j.tplants.2015.03.007

Alipour A, Tsuchimoto S, Sakai H, et al (2013) Structural characterization of copia-type retrotransposons leads to insights into the marker development in a biofuel crop,

Jatropha curcas L. Biotechnol Biofuels 6:1–13. doi: 10.1186/1754-6834-6-129

Alves S, Ribeiro T, Inácio V, et al (2012) Genomic organization and dynamics of repetitive DNA sequences in representatives of three Fagaceae genera. Genome 55:348–359. doi: 10.1139/g2012-020

Ammiraju JSS, Zuccolo A, Yu Y, et al (2007) Evolutionary dynamics of an ancient retrotransposon family provides insights into evolution of genome size in the genus Oryza. Plant J 52:342–351. doi: 10.1111/j.1365-313X.2007.03242.x

Andrianasolo DN, Davis AP, Razafinarivo NJ, et al (2013) High genetic diversity of in situ and ex situ populations of Madagascan coffee species: Further implications for the management of coffee genetic resources. Tree Genet Genomes 9:1295–1312. doi: 10.1007/s11295-013-0638-4

Arensburger P, Piégu B, Bigot Y (2016) The future of transposable element annotation and their classification in the light of functional genomics - what we can learn from the fables of Jean de la Fontaine? Mob Genet Elements. doi: 10.1080/2159256X.2016.1256852

Baidouri M El, Carpentier MC, Cooke R, et al (2014) Widespread and frequent horizontal transfers of transposable elements in plants. Genome Res 24:831–838. doi: 10.1101/gr.164400.113

Bao W, Kojima KK, Kohany O (2015) Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA 6:1–6. doi: 10.1186/s13100-015-0041-9

Bao Z, Eddy SR (2002) Automated de novo identification of repeat sequence families in sequenced genomes. Genome Res 12:1269–1276. doi: 10.1101/gr.88502

Bennett MD, Leitch IJ (2012) Plant DNA C-values database. http://www.kew.org/cvalues/.

Bennetzen JL (2005) Transposable elements, gene creation and genome rearrangement in flowering plants. Curr Opin Genet Dev 15:621–627. doi: 10.1016/j.gde.2005.09.010

Bennetzen JL, Kellogg E (1997) Do Plants Have a One-Way Ticket to Genomic Obesity? Plant Cell 9:1509–1514.

Bennetzen JL, Ma J, Devos KM (2005) Mechanisms of recent genome size variation in flowering plants. Ann Bot 95:127–32. doi: 10.1093/aob/mci008

Bergman CM, Quesneville H (2007) Discovering and detecting transposable elements in genome sequences. Brief Bioinform 8:382–392. doi: 10.1093/bib/bbm048

Berthaud J (1986) Les ressources génétiques pour l'amélioration des caféiers africains diploïdes. Paris Sud

Boeke JD, Garfinkel DJ, Styles CA, Fink GR (1985) Ty elements transpose through an RNA

intermediate. Cell 40:491–500. doi: 10.1016/0092-8674(85)90197-7

Bouharmont J (1959) Recherches sur les affinités chromosomiques dans le genre Coffea. I.N.É.A.C., Montpellier

Bousios A, Darzentas N (2013) Sirevirus LTR retrotransposons: phylogenetic misconceptions in the plant world. Mob DNA 4:1–5. doi: 10.1186/1759-8753-4-9

Bousios A, Darzentas N, Tsaftaris A, Pearce SR (2010) Highly conserved motifs in non-coding regions of Sirevirus retrotransposons: the key for their pattern of distribution within and across plants? BMC Genomics 11:1–14. doi: 10.1186/1471-2164-11-89

Bousios A, Kourmpetis YAI, Pavlidis P, et al (2012) The turbulent life of Sirevirus retrotransposons and the evolution of the maize genome: More than ten thousand elements tell the story. Plant J 69:475–488. doi: 10.1111/j.1365-313X.2011.04806.x

Brenchley R, Spannagl M, Pfeifer M, et al (2012) Analysis of the bread wheat genome using whole genome shotgun sequencing. Nature 491:705–710. doi: 10.1038/nature11650.Analysis

Bundock P, Hooykaas P (2005) An Arabidopsis hAT-like transposase is essential for plant development. Nature 436:282–284. doi: 10.1038/nature03667

Bushman FD (2003) Targeting survival: Integration site selection by retroviruses and LTR-Retrotransposons. Cell 115:135–138. doi: 10.1016/S0092-8674(03)00760-8

Butelli E, Licciardello C, Zhang Y, et al (2012) Retrotransposons Control Fruit-Specific, Cold-Dependent Accumulation of Anthocyanins in Blood Oranges. Plant Cell 24:1242–1255. doi: 10.1105/tpc.111.095232

Cai Z, Liu H, He Q, et al (2014) Differential genome evolution and speciation of Coix lacryma-jobi L. and Coix aquatica Roxb. hybrid guangxi revealed by repetitive sequence analysis and fine karyotyping. BMC Genomics 15:1–16. doi: 10.1186/1471-2164-15-1025

Capy P, Langin T, Higuet D, et al (1997) Do the integrases of LTR-retrotransposons and class II element transposases have a common ancestor? Genetica 100:63–72. doi: 10.1023/A:1018300721953

Carr M, Bensasson D, Bergman CM (2012) Evolutionary Genomics of Transposable Elements in Saccharomyces cerevisiae. PLoS One 7:50978–50993. doi: 10.1371/journal.pone.0050978

Carvalho A (1952) Taxonomia de Coffea Arabica L. VI - Caracteres morfologicos dos haploides. Bragantia 12:201–212.

Carvalho M, Ribeiro T, Viegas W, et al (2010) Presence of env-like sequences in Quercus

suber retrotransposons. J Appl Genet 51:461–467. doi: 10.1007/BF03208875

Casacuberta E, González J (2013) The impact of transposable elements in environmental adaptation. Mol Ecol 22:1503–1517. doi: 10.1111/mec.12170

Caspi A, Pachter L (2006) Identification of transposable elements using multiple alignments of related genomes. Genome Res 16:260–270. doi: 10.1101/gr.4361206

Chaparro C, Gayraud T, Souza RF De, 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

Charrier A (1978) La Structure génétique des caféiers spontanées de la région malgache (Mascarocoffea). Leur relations avec les caféiers d'origine africaine (Eucoffea).

Charrier A, Berthaud J (1985) Botanical Classification of Coffee. Coffee Bot Biochem Prod Beans Beverage 13–47. doi: 10.1007/978-1-4615-6657-1

Chevalier A (1929) Principes d'arboriculture fruitière applicables aux caféiers. Encycl. Biol. 1–27.

Comai L, Madlung A, Josefsson C, Tyagi A (2003) Do the different parental "heteromes" cause genomic shock in newly formed allopolyploids? Philos Trans R Soc London Biol Sci 358:1149–1155. doi: 10.1098/rstb.2003.1305

Comfort NC (1999) "The real point is control": The reception of Barbara McClintock's controlling elements. J Hist Biol 32:133–162.

Couturon E, Raharimalala NE, Rakotomalala J-J, et al (2016) Caféiers sauvages - Un trésor en péril au coeur des forêts tropicales ! Montpellier

Cristancho MA, Botero-Rozo DO, Giraldo W, et al (2014) Annotation of a hybrid partial genome of the coffee rust (Hemileia vastatrix) contributes to the gene repertoire catalog of the Pucciniales. Front Plant Sci 5:1–11. doi: 10.3389/fpls.2014.00594

Cros J, Gavalda MC, Chabrillange N, et al (1994) Variations in the total nuclear DNA content in african Coffea species (Rubiaceae). Café Cacao Thé XXXVIII:3–10.

Cros J, Lashermes P, Marmey P, et al (1993) Molecular analysis of genetic diversity and phylogenetic relationships in Coffea. In: Quinzième colloque scientifique international sur le café. Association Scientifique Internationale du Café (ASIC), Montpellier, pp 41–46

Curcio MJ, Derbyshire KM (2003) The outs and ins of transposition: from Mu to Kangaroo. Nat Rev Mol Cell Biol 4:865–877. doi: 10.1038/nrm1241

Darré T (2014) Evolution et diversité du genre Coffea à travers l'étude des rétroéléments à LTR. Université Montpellier II

Davis AP, Govaerts R, Bridson DM, Stoffelen P (2006) An annotatated taxonomic of the genus coffea (Rubiaceae). Bot J Linn Soc 152:465–512.

Davis AP, Tosh J, Ruch N, Fay MF (2011) Growing coffee: Psilanthus (Rubiaceae) subsumed on the basis of molecular and morphological data; implications for the size, morphology, distribution and evolutionary history of Coffea. Bot J Linn Soc 167:357–377. doi: 10.1111/j.1095-8339.2011.01177.x

De Kochko A, Akaffou S, Andrade AC, et al (2010) Advances in Coffea Genomics. Adv Bot Res 53:23–63. doi: 10.1016/S0065-2296(10)53002-7

de Vries H, MacDougal DT (1905) Species and Varieties: Their Origin by Mutation. The Plant World 8:86–90.

Demerec M (1935) Unstable Genes. Bot Rev 1:233–248.

Denoeud F, Carretero-Paulet L, Dereeper A, et al (2014) The Coffee Genome Provides Insight into the Convergent Evolution of Caffeine Biosynthesis. Science (80-) 345:1180–1184. doi: 10.1126/science.1255274

Dias ES, Hatt C, Hamon S, et al (2015) Large distribution and high sequence identity of a Copia-type retrotransposon in angiosperm families. Plant Mol Biol 89:83–97. doi: 10.1007/s11103-015-0352-8

Doolittle WF, Sapienza C (1980) Selfish genes, the phenotype paradigm and genome evolution. Nature 284:601–603.

Du C, Caronna J, He L, Dooner HK (2008) Computational prediction and molecular confirmation of Helitron transposons in the maize genome. BMC Genomics 9:51. doi: 10.1186/1471-2164-9-51

Dupeyron M, de Souza RF, Hamon P, et al (2017) Distribution of Divo in Coffea genomes, a poorly described family of angiosperm LTR-Retrotransposons. Mol Genet Genomics 1–14. doi: 10.1007/s00438-017-1308-2

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–7. doi: 10.1093/nar/gkh340

Eickbush TH, Malik HS (2002) Origins and Evolution of Retrotransposons. ASM Press, Washington DC

Eilam T, Anikster Y, Millet E, et al (2008) Nuclear DNA amount and genome downsizing in natural and synthetic allopolyploids of the genera Aegilops and Triticum. Genome 51:616–627. doi: 10.1139/G08-043

Ellinghaus D, Kurtz S, Willhoeft U (2008) LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics 9:1–14. doi:

10.1186/1471-2105-9-18

Emerson RA (1914) The Inheritance of a Recurring Somatic Variation in Variegated Ears of Maize. Am Nat 48:87–115.

Ewing AD (2015) Transposable element detection from whole genome sequence data. Mob DNA 6:24–32. doi: 10.1186/s13100-015-0055-3

Faivre Rampant P, Lesur I, Boussardon C, et al (2011) Analysis of BAC end sequences in oak, a keystone forest tree species, providing insight into the composition of its genome. BMC Genomics 12:292. doi: 10.1186/1471-2164-12-292

Federoff N V. (2012) Transposable Elements, Epigenetics, and Genome Evolution. Science (80-) 338:758–767.

Fedoroff N V., Bennetzen JL (2013) Transposons , Genomic Shock , and Genome Evolution. In: Fedoroff N V. (ed) Plant Transposons and Genome Dynamics in Evolution. John Wiley & Sons, Inc., pp 181–201

Fedoroff N V (2012) Transposable Elements, Epigenetics, and Genome Evolution. Science (80-) 338:758–767.

Finnegan DJ (1989) Eucaryotic transposable elements and genome evolution. Trends Genet 5:103–107.

Fiston-Lavier A-S, Carrigan M, Petrov DA, Gonzalez J (2010) T-lex: a program for fast and accurate assessment of transposable element presence using next-generation sequencing data. Nucleic Acids Res 39:1–10. doi: 10.1093/nar/gkq1291

Fortune PM, Roulin A, Panaud O (2008) Horizontal transfer of transposable elements in plants. Commun Integr Biol 1:74–77. doi: 10.4161/cib.1.1.6328

Gilbert C, Schaack S, Pace II JK, et al (2010) A role for host-parasite interactions in the horizontal transfer of DNA transposons across animal phyla. Nature 464:1347–1350.

Gomez C, Dussert S, Hamon P, et al (2009) Current genetic differentiation of Coffea canephora Pierre ex A. Froehn in the Guineo-Congolian African zone: cumulative impact of ancient climatic changes and recent human activities. BMC Evol Biol 9:1–19. doi: 10.1186/1471-2148-9-167

Grandbastien M-A, Spielmann A, Caboche M (1989) Tnt1, a mobile retroviral-like transposable element of tobacco isolated by plant cell genetics. Nature 337:376–380.

Grandbastien MA (2015) LTR retrotransposons, handy hitchhikers of plant regulation and stress response. Biochim Biophys Acta 1849:403–416. doi: 10.1016/j.bbagrm.2014.07.017

Grover CE, Wendel JF (2010) Recent Insights into Mechanisms of Genome Size Change in

Plants. J Bot 2010:1–8. doi: 10.1155/2010/382732

Guillemat J (1946) Quelques observations sur la Trachéomycose du “Coffea excelsa.” Rev Int Bot appliquée d’agriculture Trop 542–550.

Guyot R, Cheng X, Su Y, et al (2005) Complex organization and evolution of the tomato pericentromeric region at the FER gene locus. Plant Physiol 138:1205–1215. doi: 10.1104/pp.104.058099

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

Haber JE (2000) Repairing a Double-Strand Break. Trends Genet 16:259–264.

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

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

Hamon P, Hamon S, Razafinarivo NJ, et al (2015) Coffea Genome Organization and Evolution. In: Coffee in Health and Disease Prevention. pp 29–37

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

Han Y, Wessler SR (2010) MITE-Hunter: A program for discovering miniature inverted-repeat transposable elements from genomic sequences. Nucleic Acids Res 38:e199. doi: 10.1093/nar/gkq862

Havecker ER, Gao X, Voytas DF (2004) The diversity of LTR retrotransposons. Genome Biol 5:225.1-225.6.

Hawkins JS, HyeRan K, Nason JD, et al (2006) Differential lineage-specific amplification of transposable elements is responsible for genome size variation in Gossypium. Genome Res 16:1252–1261. doi: 10.1101/gr.5282906.1

Hirochika H (2001) Contribution of the Tos17 retrotransposon to rice functional genomics. Curr Opin Plant Biol 4:118–122. doi: 10.1016/S1369-5266(00)00146-1

Hirochika H, Sugimoto K, Otsuki Y, et al (1996) Retrotransposons of rice involved in

mutations induced by tissue culture. Proc Natl Acad Sci U S A 93:7783–7788. doi: 10.1073/pnas.93.15.7783

Hribová E, Neumann P, Matsumoto T, et al (2010) Repetitive part of the banana (Musa acuminata) genome investigated by low-depth 454 sequencing. BMC Plant Biol 10:204–214. doi: 10.1186/1471-2229-10-204

Hu G, Hawkins JS, Grover CE, Wendel JF (2010) The history and disposition of transposable elements in polyploid Gossypium. Genome 53:599–607. doi: 10.1139/g10-038

Ibarra-Laclette E, Lyons E, Hernández-Guzmán G, et al (2013) Architecture and evolution of a minute plant genome. Nature 498:94–99. doi: 10.1038/nature12132

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

Ito H (2012) Small RNAs and transposon silencing in plants. Dev Growth Differ 54:100–107. doi: 10.1111/j.1440-169X.2011.01309.x

Jiang N, Visa S, Wu S, Van Der Knaap E (2012) Rider Transposon Insertion and Phenotypic Change in Tomato. In: Grandbastien M-A, Casacuberta JM (eds) Plant Transposable Elements. Springer-Verlag, pp 297–312

Jiang S, Cai D, Sun Y, Teng Y (2016) Isolation and characterization of putative functional long terminal repeat retrotransposons in the Pyrus genome. Mob DNA 7:1–10. doi: 10.1186/s13100-016-0058-8

Jurka J, Klonowski P, Dagman V, Pelton P (1996) Censor--A program for identification and elimination of repetitive elements from DNA sequences. Comput Chem 20:119–121. doi: 10.1016/S0097-8485(96)80013-1

Kapitonov V V., Jurka J (2001) Rolling-circle transposons in eukaryotes. Proc Natl Acad Sci 98:8714–8719. doi: 10.1073/pnas.151269298

Kashkush K, Feldman M, Levy AA (2002) Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat. Nat Genet 33:102–106. doi: 10.1038/ng1063

Keeling PJ, Palmer JD (2008) Horizontal gene transfer in eukaryotic evolution. Nat Rev Genet 9:605–618.

Kejnovsky E, Hawkins JS, Feschotte C (2012) Plant Transposable Elements: Biology and Evolution. In: Wendel JF, Greilhuber J, Dolezel J, Leitch IJ (eds) Plant Genome Diversity Volume 1. Springer Vienna, Vienna, pp 17–34

Kohany O, Gentles AJ, Hankus L, Jurka J (2006) Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. BMC Bioinformatics

7:474. doi: 10.1186/1471-2105-7-474

Kolano B, Bednara E, Weiss-Schneeweiss H (2013) Isolation and characterization of reverse transcriptase fragments of LTR retrotransposons from the genome of Chenopodium quinoa (Amaranthaceae). Plant Cell Rep 32:1575–1588. doi: 10.1007/s00299-013-1468-4

Krumsiek J, Arnold R, Rattei T (2007) Gepard: a rapid and sensitive tool for creating dotplots on genome scale. Bioinformatics 23:1026–1028. doi: 10.1093/bioinformatics/btm039

Kumar A, Bennetzen JL (1999) Plant Retrotransposons. Annu Rev Genet 33:479–532.

Kumar A, Pearce SR, McLean K, et al (1997) The Ty1-copia group of retrotransposons in plants: genomic organisation, evolution, and use as molecular markers. Genetica 100:205–217. doi: 10.1023/A:1018393931948

Lander E, Linton L, Birren B, et al (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921.

Lashermes P, Combes MC, Robert J, et al (1999) Molecular characterization and origin of the Coffea arabica L. genome. Mol Gen Genet 261:259–266.

Laten HM, Havecker ER, Farmer LM, Voytas DF (2003) SIRE1, an endogenous retrovirus family from Glycine max, is highly homogeneous and evolutionarily young. Mol Biol Evol 20:1222–1230. doi: 10.1093/molbev/msg142

Laten HM, Majumbar A, Gaucher EA (1998) SIRE-1, a copia/Ty1-like retroelement from soybean, encodes a retroviral envelope-like protein. Proc Natl Acad Sci U S A 95:6897–6902.

Lavialle C, Cornelis G, Dupressoir A, et al (2013) Paleovirology of “syncytins”, retroviral env genes exapted for a role in placentation. Philos Trans R Soc London Biol Sci 368:1–10. doi: 10.1098/rstb.2012.0507

Lee J, Waminal NE, Choi H-I, et al (2017) Rapid amplification of four retrotransposon families promoted speciation and genome size expansion in the genus Panax. Sci Rep 7:1–9. doi: 10.1038/s41598-017-08194-5

Lee S, Kim N (2014) Transposable Elements and Genome Size Variations in Plants. Genomics Inform 12:87–97.

Lerat E (2010) Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. Heredity (Edinb) 104:520–533. doi: 10.1038/hdy.2009.165

Lerat E (2001) Comparaison de séquences d’éléments transposables et de gènes d’hôte chez cinq espèces : A. thaliana, C. elegans, D. melanogaster, H. sapiens et S. cerevisiae.

Université Claurde Bernard - Lyon I

Lerat E, Fablet M, Modolo L, et al (2016) TEtools facilitates big data expression analysis of transposable elements and reveals an antagonism between their activity and that of piRNA genes. Nucleic Acids Res gkw953. doi: 10.1093/nar/gkw953

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

Levin HL, Moran J V. (2011) Dynamic interactions between transposable elements and their hosts. Nat Rev Genet 12:615–627. doi: 10.1038/nrg3030.Dynamic

Lippman Z, Gendrel A-V, Black M, et al (2004) Role of transposable elements in heterochromatin and epigenetic control. Nature 430:471–476. doi: 10.1038/nature02651

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

Lisch DR (2013b) Transposons in Plant Gene Regulation. In: Fedoroff N V. (ed) Plant Transposons and Genome Dynamics in Evolution. John Wiley & Sons, Inc., pp 93–116

Llorens C, Fares M a, Moya A (2008) Relationships of gag-pol diversity between Ty3/Gypsy and Retroviridae LTR retroelements and the three kings hypothesis. BMC Evol Biol 8:276. doi: 10.1186/1471-2148-8-276

Llorens C, Futami R, Covelli L, et al (2011a) The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. Nucleic Acids Res 39:D70-4. doi: 10.1093/nar/gkq1061

Llorens C, Futami R, Covelli L, et al (2011b) The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. Nucleic Acids Res 39:D70–D74. doi: 10.1093/nar/gkq1061

Llorens C, Marín I (2001) A Mammalian Gene Evolved from the Integrase Domain of an LTR Retrotransposon. Mol Biol Evol 18:1597–1600.

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

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

Lopes FR, Jjingo D, Da Silva CRM, et al (2013) Transcriptional activity, chromosomal distribution and expression effects of transposable elements in Coffea genomes. PLoS One. doi: 10.1371/journal.pone.0078931

Louarn J (1976) Hybrides interspécifiques entre Coffea canephora Pierre et C. eugenioides Moore. 20:33–52.

Louarn J (1992) La fertilité des hybrides interspécifiques et les relations génomiques entre caféiers diploïdes d'origines africaine (Genre Coffea L. sous-genre Coffea). Université Paris-Sud, centre d'Orsay

Ma J, Bennetzen JL (2004) Rapid recent growth and divergence of rice nuclear genomes. PNAS 101:12404–12410.

Macas J, Neumann P (2007) Ogre elements — A distinct group of plant Ty3/gypsy-like retrotransposons. Gene 390:108–116. doi: 10.1016/j.gene.2006.08.007

Martienssen RA, Chandler VL (2013) Molecular Mechanisms of Transposon Epigenetic Regulation. In: Fedoroff N V. (ed) Plant Transposons and Genome Dynamics in Evolution. John Wiley & Sons, Inc., pp 71–92

Maurin O, Davis AP, Chester M, et al (2007) Towards a phylogeny for Coffea (Rubiaceae): Identifying well-supported lineages based on nuclear and plastid DNA sequences. Ann Bot 1–19. doi: 10.1093/aob/mcm257

McCarthy EM, McDonald JF (2003) LTR_STRUC: a novel search and identification program for LTR retrotransposons. Bioinformatics 19:362–367. doi: 10.1093/bioinformatics/btf878

McClintock B (1950) The origin and behavior of mutable loci in maize. PNAS 36:344–355.

Mhiri C, Grandbastien M-A (2004) Éléments transposables et analyse de la biodiversité végétale. In: Morot-Gaudry J, Briat J (eds) La génomique en biologie végétale. INRA, Paris, pp 377–401

Miguel C, Simões M, Oliveira MM, Rocheta M (2008) Envelope-like retrotransposons in the plant kingdom: Evidence of their presence in gymnosperms (Pinus pinaster). J Mol Evol 67:517–525. doi: 10.1007/s00239-008-9168-3

Mirouze M, Reinders J, Bucher E, et al (2009) Selective epigenetic control of retrotransposition in Arabidopsis. Nature 461:1–5. doi: 10.1038/nature08328

Musoli P, Cubry P, Aluka P, et al (2009) Genetic differentiation of wild and cultivated populations: diversity of Coffea canephora Pierre in Uganda. Genome 52:634–646. doi: 10.1139/G09-037

Natali L, Cossu RM, Mascagni F, et al (2015) A survey of Gypsy and Copia LTR-retrotransposon superfamilies and lineages and their distinct dynamics in the Populus trichocarpa (L.) genome. Tree Genet Genomes 11:107–120. doi: 10.1007/s11295-015-0937-z

Nekrutenko A, Li W (2001) Transposable elements are found in a large number of human protein-coding genes. Trends Genet 17:619–621.

Nielen S, Vidigal BS, Leal-Bertioli SCM, et al (2012) Matita, a new retroelement from peanut: characterization and evolutionary context in the light of the Arachis A – B genome divergence. Mol Genet Genomics 287:21–38. doi: 10.1007/s00438-011-0656-6

Nishihara M, Yamada E, Saito M, et al (2014) Molecular characterization of mutations in white-flowered torenia plants. BMC Plant Biol 14:86–98. doi: 10.1186/1471-2229-14-86

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

Nowak MD, Davis AP, Yoder AD (2012) Sequence Data from New Plastid and Nuclear COSII Regions Resolves Early Diverging Lineages in Coffea (Rubiaceae). Syst Bot 37:995–1005. doi: 10.1600/036364412X656482

Ohno S (1972) So much "junk" DNA in our genome. Evol Genet Syst 23:366–370.

Ong-Abdullah M, Ordway JM, Jiang N, et al (2015) Loss of Karma transposon methylation underlies the mantled somaclonal variant of oil palm. Nature 525:533–550. doi: 10.1038/nature15365

Orgel LE, Crick FHC (1980) Selfish DNA: the ultimate parasite. Nature 284:604–607.

Pagan HJT, Smith JD, Hubley RM, Ray D a (2010) PiggyBac-ing on a primate genome: novel elements, recent activity and horizontal transfer. Genome Biol Evol 2:293–303. doi: 10.1093/gbe/evq021

Pardue M-L, DeBaryshe PG (2011) Retrotransposons that maintain chromosome ends. Proc Natl Acad Sci U S A 108:20317–20324. doi: 10.1073/pnas.1100278108

Parisod C, Alix K, Just J, 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

Parisod C, Senerchia N (2012) Responses of transposable elements to polyploidy. In: Grandbastien M-A, Casacuberta JM (eds) Plant Transposable Elements. Springer-Verlag, pp 147–168

Paterson AH, Bowers JE, Bruggmann R, et al (2009) The Sorghum bicolor genome and the diversification of grasses. 457:551–556. doi: 10.1038/nature07723

Pearce SR (2007) SIRE-1, a putative plant retrovirus is closely related to a legume TY1-copia retrotransposon family. Cell Mol Biol Lett 12:120–126. doi: 10.2478/s11658-006-0053-z

Petrov DA, Wendel JF (2006) Evolution of eukaryotic genome structure. In: Fox CW, B. WJ (eds) Evolutionary genetics: Concepts and case studies. Oxford University Press,

Piednoël M, Carrete-Vega G, Renner SS (2013) Characterization of the LTR retrotransposon repertoire of a plant clade of six diploid and one tetraploid species. Plant J 75:699–709.

doi: 10.1111/tpj.12233

Piégu B, Bire S, Arensburger P, Bigot Y (2015) A survey of transposable element classification systems - A call for a fundamental update to meet the challenge of their diversity and complexity. Mol Phylogenet Evol 86:90–109. doi: 10.1016/j.ympev.2015.03.009

Piégu B, Guyot R, Picault N, 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.of

Piffanelli P, Droc G, Mieulet D, et al (2007) Large-scale characterization of Tos17 insertion sites in a rice T-DNA mutant library. Plant Mol Biol 65:587–601. doi: 10.1007/s11103-007-9222-3

Price AL, Jones NC, Pevzner PA (2005) De novo identification of repeat families in large genomes. Bioinformatics 21:351–358. doi: 10.1093/bioinformatics/bti1018

Quesneville H, Nouaud D, Anxolabéhère D (2002) Detection of new transposable element families in Drosophila melanogaster and Anopheles gambiae genomes. J Mol Evol 57:50–59. doi: 10.1007/s00239-003-0007-2

Ramamurthy RK, Waters BM (2017) Mapping and characterization of the fefe gene that controls iron uptake in Melon (Cucumis melo L.). Front Plant Sci 8:1–13. doi: 10.3389/fpls.2017.01003

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

Rhoades MM (1938) Effect of the Dt gene on the mutability of the a1 allele in maize. Genetics 23:377–397.

Rhoades MM (1936) The effect of varying gene dosage on aleurone colour in maize. J Genet 23:347–354.

Rhoads A, Au KF (2015) PacBio Sequencing and Its Applications. Genomics, Proteomics Bioinforma 13:278–289. doi: 10.1016/j.gpb.2015.08.002

Robbrecht E, Manen J-F (2006) The major evolutionary lineages of the coffee family (Rubiaceae, angiosperms). Combined analysis (nDNA and cpDNA) to infer the position of Coptosapelta and Luculia, and supertree construction based on rbcL, rps16, trnL-trnF and atpB-rbcL data. A new class. Syst Geogr Plants 76:85–146.

Rockinger A, Sousa A, Carvalho FA, Renner SS (2016) Chromosome number reduction in the sister clade of carica papaya with concomitant genome size doubling. Am J Bot

103:1082–1088. doi: 10.3732/ajb.1600134

Roncal J, Guyot R, Hamon P, et al (2015) Active transposable elements recover species boundaries and geographic structure in Madagascan coffee species. Mol Genet Genomics 291:155–168. doi: 10.1007/s00438-015-1098-3

Rothberg JM, Leamon JH (2008) The development and impact of 454 sequencing. Nat Biotechnol 26:1117–1124. doi: 10.1038/nbt1485

Rutherford K, Parkhill J, Crook J, et al (2000) Artemis : sequence visualization and annotation. Bioinformatics 16:944–945.

SanMiguel P, Gaut BS, Tikhonov A, et al (1998) The paleontology of intergene retrotransposons of maize. Nat Genet 20:43–45. doi: 10.1038/1695

Schaack S, Gilbert C, Feschotte C (2010) Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. Trends Ecol Evol 25:537–546.

Schnable P, Ware D, Fulton R, et al (2009) The B73 Maize Genome: Complexity, Diversity, and Dynamics. Science (80-) 326:1112–1115.

Schulman AH (2012) Hitching a ride: nonautonomous retrotransposons and parasitism as a lifestyle. In: Grandbastien M-A, Casacuberta JM (eds) Plant Transposable Elements. Springer-Verlag, pp 71–88

Senerchia N, Felber F, Parisod C (2014) Contrasting evolutionary trajectories of multiple retrotransposons following independent allopolyploidy in wild wheats. New Phytol 202:975–985. doi: 10.1111/nph.12731

Smit AFA, Hubley R, Green P RepeatMasker.

Sonnhammer ELL, Durbin R (1996) A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. Gene 167:1–10.

Sultana T, Zamborlini A, Cristofari G, Lesage P (2017) Integration site selection by retroviruses and transposable elements in eukaryotes. Nat Rev Genet 18:292–308. doi: 10.1038/nrg.2017.7

Talhinhas P, Batista D, Diniz I, et al (2017) Pathogen profile The coffee leaf rust pathogen Hemileia vastatrix : one and a half centuries around the tropics. Mol Plant Pathol 18:1039–1051. doi: 10.1111/mpp.12512

Terzian C, Ferraz C, Demaille J, Bucheton A (2000) Evolution of the Gypsy Endogenous Retrovirus in the Drosophila melanogaster Subgroup. Mol Biol Evol 17:908–914.

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, pp 42–45

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408:796–815.

This P, Lacombe T, Cadle-Davidson M, Owens CL (2007) Wine grape (Vitis vinifera L .) color associates with allelic variation in the domestication gene VvmybA1. Theor Appl Genet 114:723–730. doi: 10.1007/s00122-006-0472-2

Thompson JD, Higgins DG, Gibson TJ (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.

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

Vitte C, Panaud O (2005) LTR retrotransposons and flowering plant genome size: emergence of the increase/decrease model. Cytogenet Genome Res 110:91–107. doi: 10.1159/000084941

Volff J-N (2006) Turning junk into gold: Domestication of transposable elements and the creation of new genes in eukaryotes. BioEssays 28:913–922. doi: 10.1002/bies.20452

Volfovsky N, Haas BJ, Salzberg SL (2001) A clustering method for repeat analysis in DNA sequences. Genome Biol 2:1–11. doi: 10.1186/gb-2001-2-8-research0027

Wallau GL, Ortiz MF, Loreto E (2012) Horizontal Transposon Transfer in Eukarya: Detection, Bias, and Perspectives. Genome Biol Evol 4:801–811.

Weber B, Wenke T, Fr??mmel U, et al (2010) The Ty1-copia families SALIRE and Cotzilla populating the Beta vulgaris genome show remarkable differences in abundance, chromosomal distribution, and age. Chromosom Res 18:247–263. doi: 10.1007/s10577-009-9104-4

White SE, Habera LF, Wessler SR (1994) Retrotransposons in the flanking regions of normal plant genes: A role for copia-like elements in the evolution of gene structure and expression. Proc Natl Acad Sci U S A 91:11792–11796.

Wicker T (2012) So many repeats and so little time: how to classify transposable elements. In: Grandbastien M-A, Casacuberta JM (eds) Plant Transposable Elements. Springer-Verlag, Zurich, pp 1–15

Wicker T, Keller B (2007) Genome-wide comparative analysis of copia retrotransposons in Triticeae, rice, and Arabidopsis reveals conserved ancient evolutionary lineages and distinct dynamics of individual copia families. Genome Res 17:1072–1081. doi: 10.1101/gr.6214107.Because

Wicker T, Sabot F, Hua-Van A, et al (2007) A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8:973–82. doi: 10.1038/nrg2165

Wicker T, Stein N, Albar L, et al (2001) Analysis of a contiguous 211 kb sequence in diploid wheat (Triticum monococcum L.) reveals multiple mechanisms of genome evolution. Plant J 26:307–316.

Xiong Y, Eickbush TH (1990) Origin and evolution of retroelements based upon their reverse transcriptase sequences. EMBO J 9:3353–3362.

Xu Z, Wang H (2007) LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res 35:W265–W268. doi: 10.1093/nar/gkm286

Yin H, Du J, Wu J, et al (2015) Genome-wide Annotation and Comparative Analysis of Long Terminal Repeat Retrotransposons between Pear Species of P. bretschneideri and P. Communis. Sci Rep 5:1–15. doi: 10.1038/srep17644

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

Yuyama PM, Protasio Pereira LF, Benedito dos Santos T, et al (2012) FISH using a gag-like fragment probe reveals a common Ty3-gypsy-like retrotransposon in genome of Coffea species. Genome 55:825–833. doi: 10.1139/gen-2012-0081

Zeh DW, Zeh JA, Ishida Y (2009) Transposable elements and an epigenetic basis for punctuated equilibria. BioEssays 31:715–726. doi: 10.1002/bies.200900026

Zuccolo A, Sebastian A, Talag J, et al (2007) Transposable element distribution, abundance and role in genome size variation in the genus Oryza. BMC Evol Biol 7:1–15. doi: 10.1186/1471-2148-7-152

Supplementary Table S1: Number of LTRs used for insertion estimation dating of the three families of *SIRE* in *C. canephora*, *C. arabica* and *C. eugenioides*.

Species	*Coffea canephora*			*C. arabica*		*C. eugenioides*		
Family	A	B	C	A	B	A	B	C
LTR number	208	256	107	125	109	171	146	76

Supplementary Table S2: Copy number estimation of *SIRE* elements from family C in *C. canephora*, *C. arabica* and *C. eugenioides*. These estimations have been made using Censor with the two most conserved copies (short and long LTRs) found in each genome by LTR_STRUC.

Species	*Coffea canephora*		*Coffea arabica*		*Coffea eugenioides*	
LTR	Short	Long	Short	Long	Short	Long
Number of intact copies (80-100)	42	1	1	1	2	1
Number of copies (80-80)	25	0	24	23	30	0
Total	67	1	25	24	32	1
Number of partial copies (20-80)	309	670	31	30	49	956
Number of solo-LTRs	410	58	704	1096	2	474
Ratio solo-LTRs:intact copies	9.8:1	58:1	704:1	1096:1	1	474:1

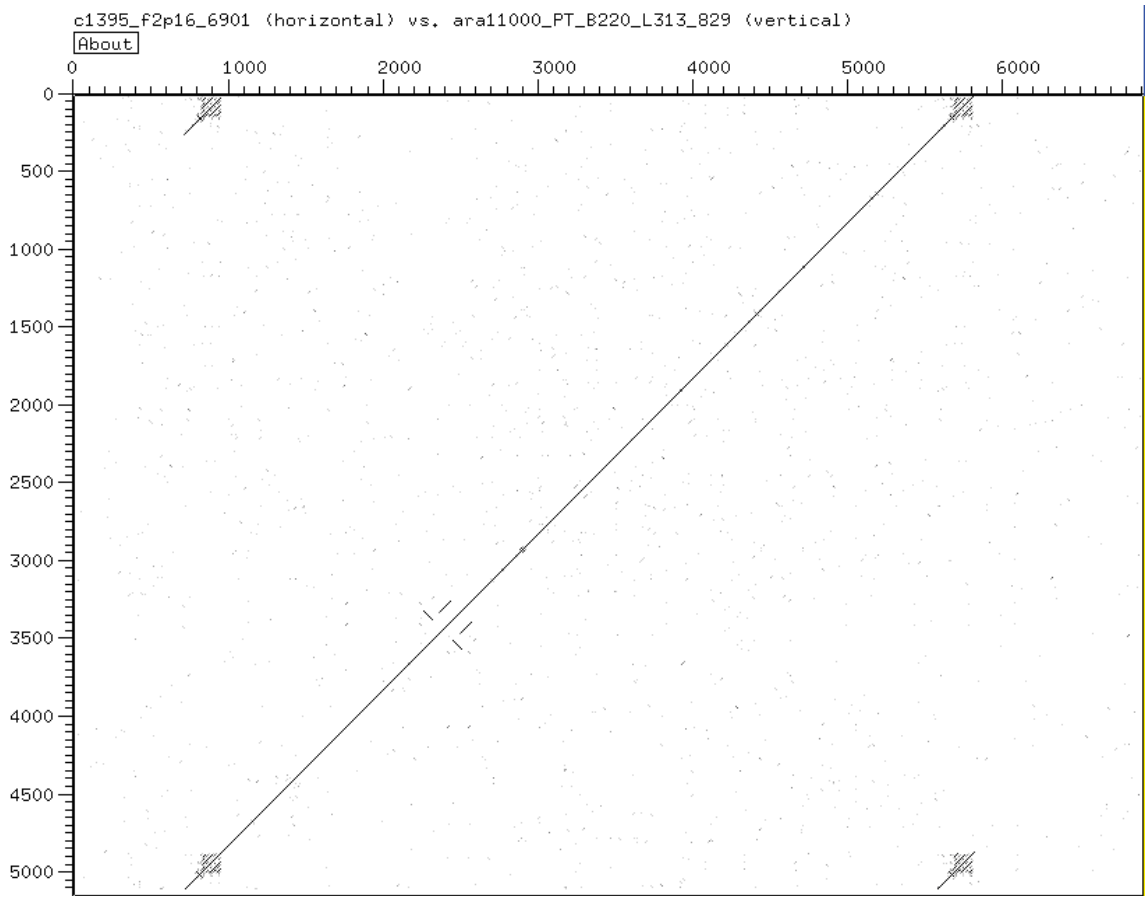


Figure S1 : dot-plot alignment of C. arabica SIRE transcript and corresponding sequence on BLAST results from LTR_STRUC output.

3. Conclusions et perspectives

Dans cette étude, nous avons pu déterminer que l'origine géographique des populations de *C. canephora* et *C. eugenioides* à l'origine de *C. arabica* est vraisemblablement localisée en Ouganda. En 2010, une prospection de caféiers effectuée par l'IRD dans les forêts primaires de moyenne altitude de la région de Zoka (Nord-Ouest de l'Ouganda, ~1000 m d'altitude) a permis d'observer les espèces *C. canephora* et *C. eugenioides*, vivant en sympatrie (A. de Kocho, communication personelle). Ce sont ces individus qui ont été séquencés dans le cadre de ce projet (BUD15 et BU-A). Les trois familles d'éléments *SIRE* identifiées et caractérisées montrent une histoire évolutive différente, ayant potentiellement influencé inégalement l'évolution des génomes.

Je me suis particulièrement intéressée à la famille C, qui présente des caractéristiques singulières n'étant pas présentes dans les familles A et B. Premièrement, elle ne contient pas de domaine *enveloppe* caractéristique de nombreuses familles de *SIRE*. L'*enveloppe* peut intervenir dans le transfert des copies de LTR-RT entre les cellules, comme c'est le cas pour les particules virales des rétrovirus chez les animaux. Par exemple chez la Drosophile, des éléments *Gypsy* avec *enveloppe* seraient impliqués dans des TH (Terzian et al. 2000). Deuxièmement, elle présente deux types de copies différant par la longueur des LTR ; copies avec des LTR de taille « classique » (environ 1 kb) et copies avec des LTR courts (de 300 à 400 pb). Ces deux types de séquences ne montrent pas le même nombre de copie chez *C. canephora* et *C. eugenioides* : il y a plus de séquences à LTR courts chez *C. canephora* que chez *C. eugenioides*.

L'origine des LTR courts des éléments de la famille C reste énigmatique. Ce sont des séquences très conservées (entre 90 et 100%), suggérant des insertions récentes des éléments et donc une fonctionnalité des séquences LTR, contenant les motifs régulateurs nécessaires au bon déroulement de la rétrotransposition. Le site PlantCARE (Lescot et al. 2002) donne accès à des outils permettant de rechercher des motifs particuliers tels que des TATA box ou des motifs de régulation de certains gènes dans ces LTR (Grandbastien 2015). La recherche de motifs régulateurs dans les LTR des copies de référence des trois familles de *SIRE* de chacun des trois génomes (pour la famille C, seulement chez *C. canephora* et *C. eugenioides* et les deux types de LTR, longs et courts) a montré qu'ils ont tous, y compris les LTR plus courts de la famille C, les

TATA et CAAT box, éléments impliqués dans la transcription en tant que promoteurs. Ils ont chacun plusieurs motifs cis-régulateurs impliqués notamment dans les réponses à certains stress comme celui de forte chaleur, ainsi que pour la plupart, un motif de 10 nucléotides impliqué dans la régulation du cycle circadien (Figure Annexe 2). Des séquences aussi longues que 10 nucléotides ont une probabilité faible (un peu moins d'une chance sur un million) d'être rencontrée au hasard dans des séquences de 10 kb. On peut donc supposer que ces séquences régulatrices présentes dans les LTR des *SIRE* ont un impact sur la régulation de certains gènes. Étant donné les caractéristiques de la famille C : i) la plus ancienne de la lignée des *SIRE* ; ii) un nombre de copies inférieur à celui des deux autres familles et iii) quasiment absente du génome de *C. arabica*, l'hypothèse d'un contrôle fort de leur activité et donc une élimination progessive en cours dans les génomes des caféiers pourrait être envisagée. Cependant, les insertions récentes que nous avons détectées chez *C. canephora* et *C. eugenioides* et les motifs régulateurs présents dans les LTR suggèrent que des copies de cette famille sont nécessaires au génome et sont donc conservées en petit nombre.

Les *SIRE* sont donc présents et structurés en trois familles dans les génomes de *C. arabica* et ses progéniteurs diploïdes. Cette structuration en trois familles n'est pas aussi poussée dans les accessions de *C. canephora* BUD15 et *C. eugenioides* DA56, suggérant une origine récente des familles A et B et une évolution différentielle des *SIRE* selon les populations de ces deux espèces. Ils peuvent donc être de bons marqueurs pour la différenciation des populations de *C. canephora*, mais aussi de *C. eugenioides*. S'ils montrent une structuration différente également dans les génomes d'autres espèces de caféiers, ils peuvent apporter des informations intéressantes sur l'évolution du genre *Coffea*, ce qui est le sujet du chapitre suivant.

Chapitre 6 – Évolution des LTR-RT *SIRE* dans le genre *Coffea*

1. Contexte

D'après les analyses précédentes (sur le jeu de données 454, la description des *SIRE* dans le génome publié de *C. canephora* et de leur description dans le génome de *C. arabica* et ses progéniteurs), il semblait judicieux de détailler l'étude des *SIRE* dans les génomes d'espèces sauvages maintenant disponibles en séquençage Illumina (projet G13). Ces séquences génomiques concernent 24 espèces, incluant des ex-*Psilanthus* pour lesquels nous n'avions pas de séquences dans les jeux de données précédents. Ceci est d'autant plus intéressant à la lumière des informations apportées par la phylogénie moléculaire du genre *Coffea*.

2. Implication personnelle

Pour cette étude, les auteurs ont participé à la génération des données de séquençage (Perla Hamon pour le projet G13), à leur assemblage (Nestlé et Romain Guyot) et à la relecture de l'article. J'ai effectué les analyses BLAST et les mapping et analysé les résultats. J'ai extrait les domaines RT et ENV des 24 espèces et j'ai réalisé les analyses phylogénétiques. J'ai réalisé les figures et rédigé l'article, que mes directeurs de thèse ont relu et corrigé.

The evolution and diversity of *SIRE* LTR-retrotransposons are associated to the diversification of wild diploid *Coffea* species

Mathilde Dupeyron [1,2 *]

Dominique Crouzillat [3]

Alexandre de Kochko [1]

Perla Hamon [1]

Romain Guyot [2]

[1] IRD UMR DIADE, EvoGec, BP 64501, 34394 Montpellier Cedex 5, France

[2] IRD UMR IPME, CoffeeAdapt, 911 Avenue Agropolis, 34394, Montpellier cedex 5, France

[3] Nestlé R&D Tours, Notre-Dame d'Oé, Tours, France

*Corresponding Author: Mathilde Dupeyron, Institut de Recherche pour le Développement (IRD), UMR IPME, BP 64501, 34394 Montpellier Cedex 5, France, mathilde.dupeyron@ird.fr

ABSTRACT

LTR-retrotransposons (LTR-RTs) are the main components of plant genomes. Since their discovery, accumulation of evidence showed their implication as formidable tools in creating genome diversities. Their abilities to replicate and increase their copy numbers, upon biotic and abiotic stress, suggest that LTR-RTs are involved in genome plasticity for rapid adaptation leading to speciation. The *Coffea* genus comprises 139 diverse species with a high adaptation to different tropical and sub-tropical environments. They also possess a high diversity of LTR-RT, representing 42% of the *C. canephora* genome, the sequenced genome reference for the *Coffea* genus.

Using the LTR-RTs *SIRE* lineage elements, we showed that they are present in all 24 wild species studied here, but with remarkable copy number amplitudes. Presence and copy number variations of three *SIRE* families are associated to phylogenetic clades, species diversification and probably adaptation to theirs environments. *SIRE* RT- and ENV-based phylogenetic analyses indicate a higher diversity in wild species than previously expected based on the *C. canephora* genome. Our results showed how *SIRE* elements might be involved in the *Coffea* genus diversification.

INTRODUCTION

Since the inclusion of *Psilanthus* species, the genus *Coffea* (Rubiaceae family) is composed of 139 species (Couturon et al. 2016) of which *C. arabica* and *C. canephora* (producing the Robusta coffee) are of major socio-economic importance worldwide. Therefore, wild coffee-trees are found in the inter-tropical forests of Africa, Western Indian Ocean Islands (WIOIs), Tropical and Southeast Asia (Indian sub-continent included) and Australasia (Davis et al. 2011). All of them are diploids (Bouharmont 1959; Louarn 1976) with the exception of *C. arabica* (allotetraploid – Carvalho 1952). Three botanical sections were determined based on their geographical distribution (West and Central Africa, East Africa) and/or their biochemical peculiarity for WIOIs species (caffeine-free species commonly named Mascarocoffea). Cross-fertilization tests, molecular markers and *in situ* hybridization studies (Charrier 1978; Louarn 1992; Hamon et al. 2009; De Kochko et al. 2010; Razafinarivo et al. 2012; Andrianasolo et al. 2013) supported these geographical groups and highlighted their genetic divergence.

Former *Psilanthus* species markedly differ from the former *Coffea* ones, mainly by the floral morphology with a long corolla tube, short style and fully or partially included anthers, whereas former *Coffea* species are long-styled with exerted anthers. They also differ by their geographical distribution since none *Psilanthus* is found in WIOIs while none *Coffea* is present in Asia (broad sense). Numerous studies failed to depict the phylogenetic relationships among *Coffea* and *Psilanthus* species (Cros et al. 1993; Robbrecht and Manen 2006; Maurin et al. 2007; Hamon et al. 2009; Nowak et al. 2012). The first fully resolved and robust phylogeny has just been published (Hamon et al. 2017) thanks to the Genotyping-By-Sequencing (GBS) methodology. It supports the geographic differentiation previously reported and shows clearly independent diversification in each main region as none species from one region is nested in another region except for the Comorian *C. humblotiana* that originates from North Madagascar. The phylogeny also showed two main clades: one called "Xeno-Coffea", composed with all ex-*Psilanthus* species and one *Coffea* from Somalia (*C. rhamnifolia*, showing morphological traits of ex-*Psilanthus* species but with *Coffea*-like flowers). The other *Coffea* called "Eu-Coffea" includes all remaining *Coffea* species. Among each clade, phylogenetic relationships appear clearly (supplemental Figure S1), so it is possible to use it for an analysis of the evolution of particular traits as it has been done with caffeine content evolution.

LTR-retrotransposons (LTR-RTs) are the most prevalent TEs in plant genomes (Kumar and Bennetzen 1999; Lisch 2013) and are composed by 2 long terminal repeats (LTRs) and one or two open reading frames (ORFs) containing the Gag and Pol regions. In these ORFs, reside the coding regions essential to their mobility: the integrase (INT), the reverse transcriptase (RT) and the RNase H (RH) (Havecker et al. 2004). For some LTR-RT, an additional ORF is observed, containing an *envelope*-like gene. This coding region is found in retroviruses and is involved in the capacity of cell-to-cell infection (Eickbush and Malik 2002). These domains are well conserved and can be used to study their evolution and diversity in the genomes they reside in. LTR-RTs represent also good markers of diversity depending on the impact of their dynamics and evolution in these genomes. Such study has previously been done in 18 Eu-Coffea and one Xeno-Coffea species, with two LTR-RTs (*Nana* and *Divo*) discovered in *C. canephora* (Hamon et al. 2011). *Nana* appeared associated to the species differentiation while *Divo* (belonging to the *Bianca* lineage, Dupeyron et al. 2017) followed the genetic differentiation *among C. canephora* that was previously observed with nuclear microsatellite markers (Gomez et al. 2009). Recent analyses of *Divo* confirmed its presence in all 24 wild *Coffea* genomes representatives of the main clades (unpublished data), suggesting that it belongs to an ancient family pre-dating the *Coffea* differentiation.

In order to better understand the LTR-RTs composition of *Coffea* genomes and their differentiation, partial genome sequencing from ten diploid species, representatives of the bio-geographic groups (Xeno-Coffea from Africa and Australasia and Eu-Coffea from West, Central and East Africa, and Indian Ocean islands) were used. The global composition in LTR-RTs is similar for all the ten species, excepted for the *Del* lineage of the *Gypsy* Super-family and the *SIRE* lineage of the super-family *Copia*. *SIRE* elements were first discovered in *Glycine max* and they are the only *Copia* elements carrying an *envelope*-like gene (or ENV domain, Laten et al. 1998) suggesting a similar structure to retroviruses. As *SIRE* are widespread among plant genomes and generally in a huge copy number, the ENV gene – if functional - could favoured *SIRE* expansion in the host genome (Laten and Gaston 2012). Firstly believed as a recent lineage, the study of highly conserved motifs in non-coding regions but with high nucleotide divergence between *SIRE* sequences from various plant lineages shows that in fact, *SIRE* is an ancient lineage, with recent insertion activity in most of the plant lineages studied until now (Bousios et al. 2010). Interestingly, using partial sequencing and PCR amplifications in *Coffea*, *SIRE* are not detected in genomes from WIOIs and from Indonesia, whereas it is highly detected in West and Central African species and quite well detected in East African (Guyot et al. 2016, supplemental Figure S2). An in-depth

analysis of complete genome sequencing is required to confirm and to detail the distribution pattern of *SIRE* and to understand their dynamics in genomes of the *Coffea* genus.

Here, with the availability of high coverage Illumina sequencing data of 24 wild diploid *Coffea* species/sub-species, we attempted to first characterize *SIRE* LTR-RTs content in these genomes. Their complete absence in species of the WIOIs is not fully confirmed, since a very low copy number of elements were detected from species in Madagascar and a complete absence in *C. humblotiana*.

Phylogenetic study of a broad *Coffea* genome sampling highlights the dynamics of *SIRE* in these genomes. The evolution of *Coffea* genus is discussed as well.

MATERIAL AND METHODS

Genomic resources

24 accessions corresponding to 24 species/sub-species were studied. Information on their origin and classification (from Hamon et al. 2017) are given in **Table 1**. The twenty-four wild coffee-trees genome sequences generated under the G13 Consortium (Oral communication PAG 2015) were used in this study in addition to *C. canephora* and *C. eugenioides* genomes sequencing obtained from the Arabica Coffee Genome Consortium (ACGC 2014). Genomic data information for each studied species is reported in **Table 1**.

The genome of *C. canephora* (accession DH 200-94) generated under the ACGC Consortium (ACGC 2014) with the PacBio technology was also used in this study.

Table 1: Information on *Coffea* species examined in this study

Species (ex-Psilanthus name)	Accession	Country of origin	Germplasm collection source	Phylogenetic clade[a]	Raw sequences number	Estimated sequence size	Estimated depth of coverage	Contig number n:500	N50 (bp)	Cumulative lentgh (Mb)	Genome size (Mb) *estimated
C. rhamnifolia (Chiov.) Bridson	P04003534	Mozambique	MNHN - P	Xeno-Coffea	86 M x 2	17 Gb	24,66144	51356	25642	412	* 700
C. neoleroyi A. P. Davis (ex-*Psilanthus neoleroyi*)	APD 6008	South Sudan	K	Xeno-Coffea	49 M x 2 61 M x 2	10 Gb 12 Gb	30,653564	45840	30346	447	* 700
C. ebracteolata (Hiern) Brenan (ex-*Psilanthus ebractelatus*)	PSI11	Ivory Coast	BRC	Xeno-Coffea	118 M x 2 85 M x 2	23 Gb 17 Gb	28,65140903	196214	2931	374	562
C. mannii (Hook.f.) A. P. Davis (ex-*Psilanthus mannii*)	2003 1365-45 (BR)	Cameroon	BR	Xeno-Coffea	46 M x 2 60 M X 2 45 M x 2 55 M X 2	8 Gb 10 Gb 8 Gb 11 Gb	46,98801231	292874	3989	665	* 700
C. melanocarpa (ex-*Psilanthus melanocarpus*)	P00128349	Angola	MNHN - P	Xeno-Coffea	33 M x 2 33 M x 2 32 M x 2 32 M x 2	8 Gb 8 Gb 8 Gb 8 Gb	46,76398786	283166	5923	749	* 700
C. merguensis (ex-*Psilanthus merguensis*)	P04551601	Thailand	MNHN - P	Xeno-Coffea	26 M x 2 25,5 M x 2 31 M x 2 30 M x 2	6 Gb 6 Gb 7,7 Gb 7,6 Gb	40,22165143	162418	9017	580	* 700
C. horsefieldiana Miq. (ex-*Psilanthus horsefieldianus*)	HOR	Indonesia	ICCRI	Xeno-Coffea	61 M x 2 81 M x 2	12 Gb 16 Gb	27,84498233	163725	9059	768	593
C. benghalensis B. Heyne ex Schult. var. *benghalensis* (ex-*Psilanthus benghalensis*)	PBTA (CCRI)	India	CBI	Xeno-Coffea	81 M x 2 59 M x 2	16 Gb 11 Gb	74,08662246	94529	13659	467	* 700
C. benghalensis var. *bababudanii* (Sivar., Biju & P.Mathew) A.P.Davis (ex-*Psilanthus bababudanii*)	PBT1 (CCRI)	India	CBI	Xeno-Coffea	82 M x 2 64 M x 2	16 Gb 10 Gb	67,29426407	116782	13519	508	* 700
C. brassii (ex-*Psilanthus brassii*)	D. Crayn 1196 (CNS)	Australia	CNS	Xeno-Coffea	72 M x 2	12 Gb	26,68763511	137171	823	113	* 700
C. charrieriana Stoff. & F. Anthony	OA22	Cameroon	BRC	Eu-Coffea	78 M x 2	15,5 Gb	21,45671657	77203	15850	519	724
C. pseudozanguebariae Bridson	H53	Kenya	BRC	Eu-Coffea	55 M x 2 75 M x 2	11 Gb 15 Gb	47,42322676	170071	7614	593	601
C. racemosa Lour.	IB62	Mozambique	BRC	Eu-Coffea	55 M x 2 72 M x 2	11 Gb 14 Gb	50,64657242	130383	12593	596	513
C. mufindiensis Hutch. Ex-Bridson subsp. *mufindiensis*	APD 2917	Tanzania	K	Eu-Coffea	51 M x 2 49 M x 2	10 Gb 10 Gb	30,900956	151504	5001	430	* 650
C. eugenioides S. Moore	DA56	Kenya	BRC	Eu-Coffea	127 M x 2 76 M x 2	25 Gb 14 Gb	37,12008634	128164	3927	277	645
C. stenophylla G. Don.	FB55	Ivory Coast	BRC	Eu-Coffea	75 M x 2 54 M x 2	15 Gb 10 Gb	32,52574448	102938	12025	500	650
C. humilis A. Chevalier	G57	Ivory Coast	BRC	Eu-Coffea	59 M x 2 71 M x 2	11 Gb 14 Gb	29,40954516	227202	5056	599	900
C. liberica Bull. var. *liberica*	EA61	Ivory Coast	BRC	Eu-Coffea	32 M x 2	8 Gb	11,39634348	85240	1414	103	704
C. kapakata A. Chev.	KAP	Angola	BRC	Eu-Coffea	36,5 M x 2	9 Gb	14,16307652	141096	2349	245	* 650
C. canephora A. Froehner	BUD15	Uganda	BRC	Eu-Coffea	16 M x 2	3 Gb	4,507042254	181027	9387	619	710
C. macrocarpa A. Rich.	PET (P, K)	Mauritius	BRC	Eu-Coffea	76 M x 2 95 M x 2	13 Gb 18 Gb	59,43231303	62275	16832	393	548
C. humblotiana Baill.	BM19/20	Comoros	BRC	Eu-Coffea	67 M x 2 51 M x 2	13 Gb 10 Gb	50,46572344	71337	16797	429	474
C. tetragona Jum. & H. Perrier	A252	Madagascar	KCRS	Eu-Coffea	94 M x 2 64 M x 2	18 Gb 12 Gb	49,50984768	92743	14288	487	521
C. dolichophylla J.-F. Leroy	A206	Madagascar	KCRS	Eu-Coffea	75 M x 2	15 Gb	36,33816676	189324	6661	608	689

a: according to Hamon et al. (2017)

Estimation of *SIRE* copy number in 24 *Coffea* genomes

SIRE copy number was previously mined in ten *Coffea* genomes partially sequenced with the 454 sequencing technology. Here, with the 24 genomes sequenced by the Illumina technology, a more precise estimation can be done with mapping techniques. We used Bowtie2 (Langmead and Salzberg 2012) on raw-data of 24 wild *Coffea* species (Table 1). These data are in FASTQ format and are generated randomly. We choose to use only one million sequences, so the four first millions of lines of each file in FASTQ format (representing 100 Mb of sequences). RT domain of the most conserved *SIRE* copy found in *C. canephora* (Dupeyron et al. in preparation) was used as a reference. Then Bowtie 2 was launched with these parameters: end-to-end mode for a more precise alignment during the mapping, no unaligned sequences in the output file, very fast and output file in SAM format (Sequence Alignment/Map format). The counts obtained were adjusted according to the number of base pair in the raw-data file and the genome size of each *Coffea* genome (some of

them are estimated, Table 1). The same analyse was launched with the *envelope*-like domain found in *C. canephora* (from families A and B, see Dupeyron et al. in preparation).

Mining of the three *SIRE* families in 24 *Coffea* genomes

Full set of Illumina data for each species was used for genome assembly, using MaSuRCA (Zimin et al. 2013, Version 3.2.2) or Abyss (Simpson et al. 2009 - **Table 1**). RT domains of LTR-RTs were searched in 24 assembled genomes of the G13 project by one by one BLASTx (Altschul et al. 1990) analyses against GyDB database reference RT domains, with a minimum e-value of $1.10e^{-4}$ and only one sequence targeted per hit found. Then, a Bash script containing GeneWise (Birney et al. 2004) was launched to extract RT domains (> 200 amino acid residues, excepted for *C. brassii*: > 180 amino acid residues) corresponding to LTR-RTs that were recognized in the BLASTx analysis. If the presence of *SIRE* RT domains was confirmed, the same analysis was launched with this time representative RT domains of *C. canephora* (complete copies of the three families and copies of the family 1 found by Censor (Dupeyron et al. in preparation) as references. Once extracted, *SIRE* RT domains of each genome were aligned with those of *C. canephora* and a NJ phylogenetic tree was computed in ClustalW (Thompson et al. 1994) with 1,000 bootstraps replicates for each species.

The same procedure was used to extract and compute NJ phylogenetic trees with the *envelope*-like domains (> 180 amino acid residues). The references used were the domains from families A and B from *C. canephora* (accession HD-200) and from the tomato *SIRE* element *Tortl-1* available in GyDB. NJ phylogenetic trees of *SIRE* ENV domains have also been computed according to the *Coffea* molecular phylogeny from Hamon et al. (2017) and rooted with the four *SIRE* ENV references from GyDB (*Endovir1-1* from *A. thaliana*, *Opie-2* from *Zea mays*, *SIRE1-4* from *Glycine max* and *TorTL1* from *Lycopersicon esculentum*). ENV domain references of families A and B from *C. canephora* and *C. eugenioides* have been added to the alignments. *C. brassii* was not used in the ENV domains analysis because of the small contig length of its genome assembly.

RESULTS

***SIRE* are present in the 23 *Coffea* diploid species, but in variable copy numbers depending on the species**

In order to estimate the copy number of the *SIRE* in the 24 species/sub-species used in this study, Bowtie 2 was launched on the raw-data of 24 wild *Coffea* species with the RT domains of *SIRE* elements annotated in *C. canephora* as reference. *SIRE* elements were detected in all species, with the exception of *C. humblotiana* (**Figure 1**). Considering an eastward species distribution from West Africa to North-Australia, the highest copy numbers are obtained for WCA species plus *C. eugenioides* and *C. mufindiensis*. Intermediate copy numbers are observed for the EA clade, *C. charrieriana,* and the Xeno-Coffea clade, whatever their origin (African or Asian) as observed previously (Guyot et al. 2016). The WIOIs species present the lowest *SIRE* content, and no copy were detected in the Comorian *C. humblotiana*.

The same analysis based on the detection of the *envelope*-like domains shows global similar patterns of estimated copy numbers among the 24 species. The highest copy numbers are obtained for the Central-East clade and the WCA African species exception to *C. charrieriana*. Low copy numbers are obtained for *C. neoleroyi*, EA and Asian ex-*Psilanthus* (part of Xeno-Coffea). None *envelope*-like domain has been detected in *C. rhamnifolia*, WIOIs and Asian Xeno-Coffea species (Suppl. material Figure S2).

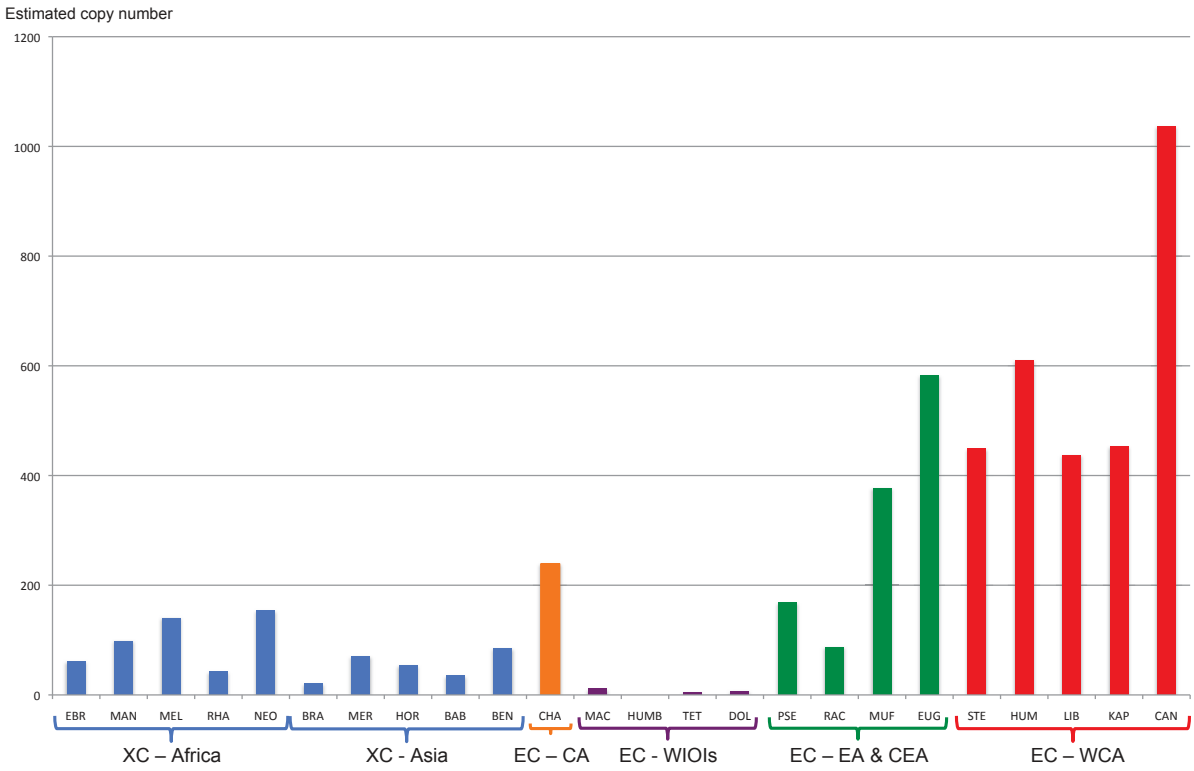


Figure 1: Copy number estimation of *SIRE* RT domains in 24 species of coffee-trees classified according to their geographical origin.
XC: Xeno-Coffea; EC-CA: Eu-Coffea – Central Africa; WIOIs: Western Indian Ocena Islands; EA & CEA: East and Centre-East Africa; WCA: West and Central Africa. EBR: *C. ebracteolata*; MAN: *C. mannii*; MEL: *C. melanocarpa*; RHA: *C. rhamnifolia*; NEO: *C. neoleroyi*; BRA: *C. brassii*; MER: *C. merguensis*; HOR: *C. horsefieldiana*; BAB: *C. bababudanii*; BEN: *C. benghalensis*; CHA: *C. charrieriana*; MAC: *C. macrocarpa*; HUMB: *C. humblotiana*; TET: *C. tetragona*; DOL: *C. dolichophylla*; PSE: *C. pseudozanguebariae*; RAC: *C. racemosa*; MUF: *C. mufindiensis*; EUG: *C. eugenioides* (DA56, Kenya); STE=*Coffea stenophylla*; HUM=*C. humilis*; LIB=*C. liberica*; KAP: *C. kapakata*; CAN: *C. canephora* (BUD15, Uganda).

***SIRE* elements are differentially structured in *Coffea* genomes**

To compare the results with simple mapping analysis, NJ trees were performed with RT amino-acid domains recovered from each genome assembly. NJ trees performed with *SIRE* RT domains were presented with concatenated branches obtained for the different main clades together with the complete *Coffea* molecular phylogeny tree as a frame obtained from Hamon et al. (2017), and enlarged with the addition of three new ex-*Psilanthus* species (one African and two Asian) (**Figure 2**). *SIRE* that were organized in 3 families (A, B and C) according to NJ trees of LTR sequences from the *SIRE* detected by LTR_STRUC in *C. canephora,* were

used as references (Dupeyron et al. in preparation). Due to the large data set used, NJ trees are presented clade by clade: Xeno-Coffea, WIOIs, EA, CEA and WCA (**Figure 2**).

All the trees generated from each clade showed an overall similar pattern. The family C of *SIRE* is at a basal position (in green) whereas B and C families (in red and blue) showed shorter branch lengths. One overall interesting observation is that the majority of *SIRE* RT domains formed new families (in grey) in the branches leading to C or to A and B families, suggesting more complex relationships at the amino-acid level in wild species. In addition, while some RT domains fell into the family C clade (but present in all species studied), very few or none elements were found associated to A or B families, identified originally in *C. canephora*.

Concerning more specifically the Xeno-Coffea clade with species from Africa (**Figure 2 (A)**), only *C. melanocarpa* shows several copies included to A and B families (turquoise names). *C. rhamnifolia* and *C. neoleroyi* didn't get any A and B family RTs. This contrasted situation may suggest that the presence of the A and B RT families is relatively versatile among Xeno-Coffea species. For *C. charrieriana*, the formed clades are very similar to the Xeno-Coffea *C. rhamnifolia* and *C. neoleroyi* species with the complete absence of A and B families. Species from the WIOIs also exhibit a very similar pattern to *C. charrieriana,* Xeno-Coffea *C. rhamifolia* and *C. neoleroyi* species.

NJ tree from the EA and CEA species (**Figure 2 (B)**) showed a very similar clade organization between each other, with RT belonging to C and B families and numerous extra clades in the C and B branches. The exception is for *C. mufindiensis* that shows the same tree organization than WCA species (**Figure 2 (C)**).

NJ trees were also generated using the ENV domains only present in *SIRE* families A and B (**Figure 3**). Concerning the Xeno-Coffea clade, *C. rhamnifolia* and *C. neoleroyi* exhibited similar topologies together, with the formation of new families (in grey) and ENV domains falling into the B family but not in the A family (at the exception of one ENV sequence in *C. neoleroyi*). This result differs with the RT-based NJ trees, where no RT domain is present for family A and B for *C. rhamnifolia* and *C. neoleroyi*. *C. melanocarpa* from African Xeno-Coffea (**Figure 3 -1)** and Eu-coffea species from WCA, show *ENV* domains clustering with A and B families. Eu-coffea species from East and Central-East Africa contain B family ENV domains and only *C. mufindiensis* contains A family ENV domains.

Finally, Asian Xeno-Coffea (**Figure 3 -2**), *C. charrieriana*, and species from the WIOIs do not contain any ENV domains into A or B families in concordance with the absence of A and

B families RT domains. Presence and absence of RT and ENV domains in A, B and C clades are summarized in **Figure 4**.

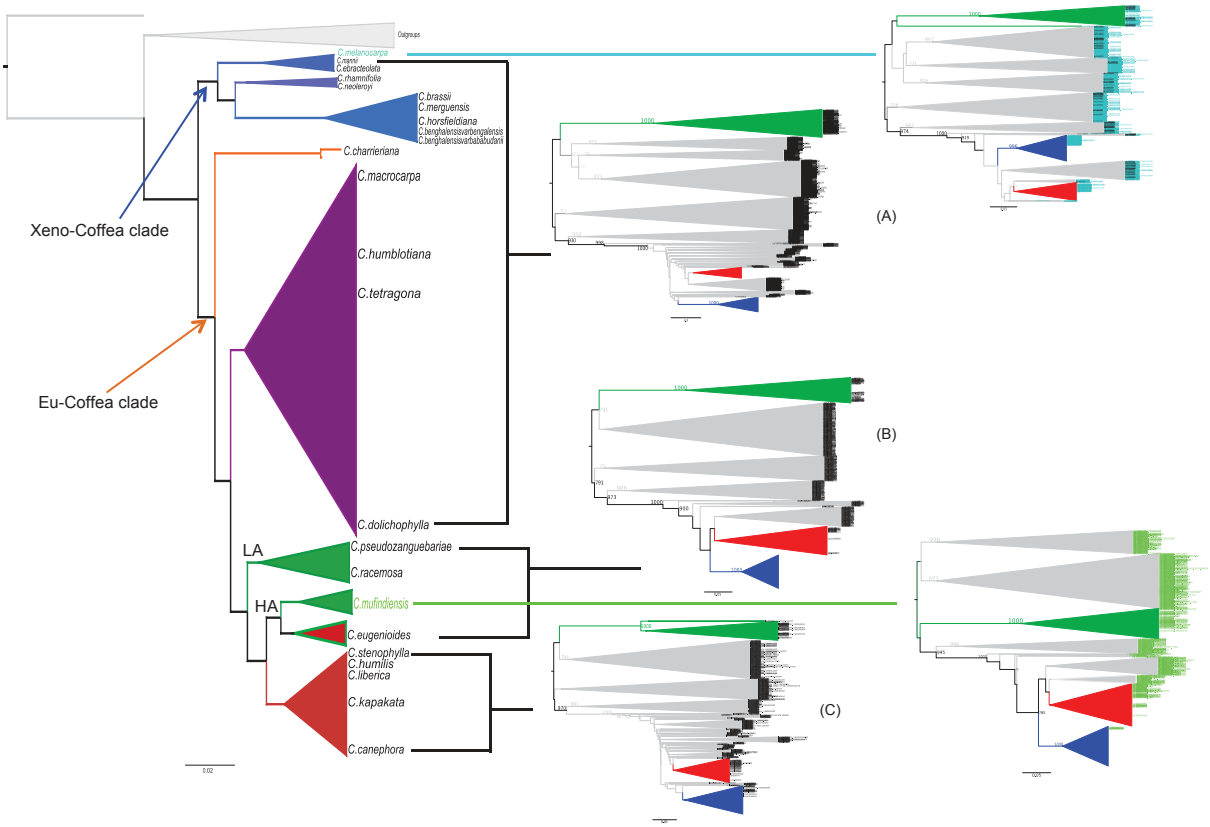


Figure 2: Phylogenetic tree of *Coffea* genus and *SIRE* RT domains NJ trees of *Coffea* diploid species.
(A) Representative trees of species without families A and B (excepting *C. melanocarpa*, turquoise names); (B) Representative tree of species with family C and B only (excepting *C. mufindiensis*, green names); (C) Representative tree of species with the three families. Blue: family A; Red: family B; Green: family C; Grey: other families or sub-families; LA: low altitude species; HA: high altitude species.

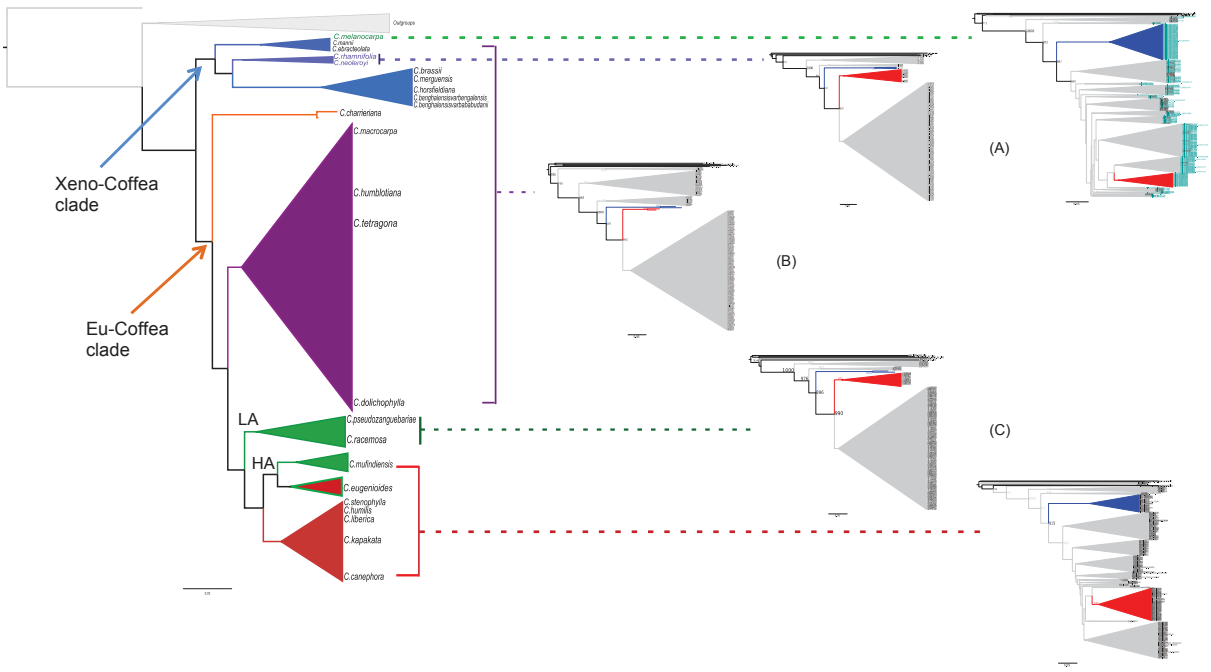


Figure 3: Phylogenetic tree of *Coffea* genus and *SIRE* ENV domains NJ trees of *Coffea* diploid species.
(A) Representative trees of species without family A (excepting *C. melanocarpa*, turquoise names); (B) Representative tree of species without families A and B; (C) Representative tree of species with the two families. Blue: family A; Red: family B; Grey: other families or sub-families; LA: low altitude species; HA: high altitude species.

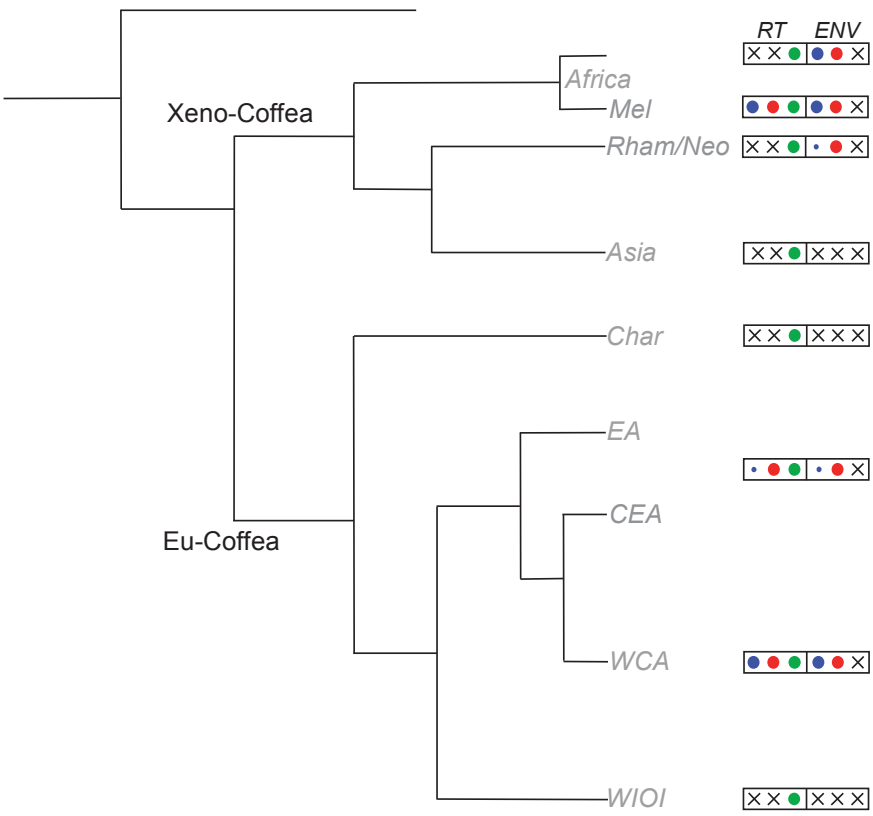


Figure 4: Presence/absence summary of *SIRE* RT and ENV domains in the main *Coffea* clades. The size of the dots indicates the more or less important detected copy number.

DISCUSSION

SIRE elements are present in all _Coffea_ genomes with contrasted amplitudes.

The analysis of partial 454 sequencing of ten diploid *Coffea* species and specific PCR amplifications suggested that *SIRE* were absent from the genomes of WIOI species (supplemental Figure S2 – data from Guyot et al. 2016). Read mapping using Illumina data from 24 diploid *Coffea* species and sub-species revealed that a small number of reads can be detected in WIOIs genomes at the exception of the Comorian species *C. humblotiana*. However, *SIRE* RT domains were identified in all assembled genomes, including *C. humblotiana*. These apparent contradictions suggest that each result should be carefully interpreted.

454 partial sequencing gave an approximation of TE content in genomes, but partial sequencing (i.e. < 10% of the genome) is only a good method for evaluating high copy number and conserved TEs. TEs in small copy numbers are not statistically accurately evaluated, as seen for *SIRE* elements in the WIOIs species. Read mapping against a reference genome or sequence is a robust and stringent method when the genome is sequenced with an elevated coverage. However, if the reference sequences are incomplete or don't represent all the diversity found in genomes, target sequences might be discarded from the results by the stringency of the mapping algorithm parameters. Here, BLASTx algorithm allows more flexibility to search for slightly divergent amino-acid sequences when compared to the reference, but works well with large sequences such as contigs. The clear demonstration is the recovery of numerous domains (RT and ENV) that do not clustered with defined families in *C. canephora*, using LTR nucleotide based phylogeny.

Taking into account all methods, the results of *SIRE* elements dynamics agree fully with a strong divergence in the Eu-coffea clade between African and WIOI species, as observed previously (Charrier 1978; Razafinarivo et al. 2013; Hamon et al. 2017), but also highlight clear separation between African, Asian and *C. rhamnifolia*/*C. neoleoroyi* in Xeno-Coffea (**Figure 4**).

Full long read sequencing and complete assembly of one reference species in each clade will give an important resource to study diversity of *SIRE* elements and described new families and as well study genomes divergence.

SIRE structuring is different according to *Coffea* clades

SIRE elements were organized into three families (or families) in *C. canephora* reference genome (HD200-94; noted A, B and C, Dupeyron et al. in preparation). RT and ENV domains of these families have been searched in 24 assembled genome sequences of wild coffee-trees representatives of the Xeno- and Eu-Coffea clades and the *Coffea* genus phylogeny (Hamon et al. 2017).

Elements of the C family, the only one without any ENV domains in *C. canephora* (Dupeyron et al. in preparation), *C. eugenioides* and *C. arabica*, were identified in all genomes analysed whatever the clade. Together with its basal position in NJ trees, this information suggests that an overall stability of the family C elements and probably it might be present in the ancestor of all *Coffea* species. The A and B families, when present in studied species, showed shorter branch lengths. This suggest that A and B families arose later than the C families from a common ancestor element. The presence of A and B families in Xeno and Eu-Coffea support this hypothesis, but lead to a complex evolutionary history, with successful diversification in Africa and a complete lost in WIOI species (Figure 4). At this step of our analysis, we cannot exclude that the presence of A and B families both in African Xeno- (*C. melanocarpa*) and Eu-Coffea species from CEA might be due to horizontal transfer mechanisms between species (Dias et al. 2015) or ancestral interspecific homoploid hybridizations.

The Eu-Coffea clade from East and Central-East African species is sub-divided in two families with East African species of low altitude (LA Figures 2 and 3), represented by *C. racemosa* and *C. pseudozanguebariae* and with East and Central-East African species of high altitudes (HA **Figures 2** and **3**), represented by *C. mufindiensis* and *C. eugenioides*. These two species cluster close to the WCA clade in the *Coffea* genus phylogeny (Figure S1) and their NJ tree topologies appear similar, suggesting a different history between East and Central-East African species of high altitude and East African species of low altitude. This different history may be related to the presence of the Great Rift Valley, where different flora and fauna were observed according to each side of the rift (White 1983). More genomic data on East African species are needed to detail their place in the phylogeny and TE content, as 17 species were described but only four are available to date and in the present study.

SIRE dynamics and species evolution

Different African species such as *C. melanocarpa* (Xeno-Coffea), *C. charrieriana* (basal to Eu-Coffea), *C. mufindiensis* from East Africa (Eu-Coffea) and *C. canephora* from West

Africa (Eu-Coffea) showed individual *SIRE* phylogenetic structures (Figures 2 and 3) different from the clade they belong.

C. charrieriana, found in South East Cameroon, is the most ancient Eu-Coffea (> 11 My, Hamon et al. 2017). Together with its lack of caffeine, its lowest sucrose content and lowest LTR-RTs genome composition, *C. charrieriana* represents a unique *Coffea* species in Africa (Hamon et al. 2017). This basal species need to be more investigated at the molecular level to understand the origin of their characteristic similarities with *C. melanocarpa* and *C. mufindiensis*.

Previous taxonomic studies placed *C. melanocarpa* in *Psilanthus* subgenus Afrocoffea (*Psilanthus melanocarpus*), but weakly supported due to an atypical morphological shape compared to other ex-*Psilanthus* species in this subgenus (presence of filaments and submedifixed anthers) (Davis et al. 2006). Moreover, contrary to the two other Xeno-Coffea species from Africa studied here, *C. ebracteolata* and *C. mannii* occurring in WCA, *C. melanocarpa* naturally resides in the northern forests of Angola, Central-South Africa, with diverse climates and different types of forests, particularly in the North-West of the country (USAID 2008). The special features of Angolian forests might have impacted *C. melanocarpa* adaptation to its environment, leading to differences in flowers shape and so in specific evolution of transposable element such as *SIRE*. The diversification of *SIRE* of *C. melanocarpa* when compare to *C. mannii* and *C. ebracteolata* supports its classification into a different sub-clade (Hamon et al. 2017).

C. mufindiensis, belonging to East and Central-East African Eu-Coffea, is the only species of this clade that contains both family A RT and ENV domains. This species found in Tanzania is in the clade close to WCA Eu-Coffea, containing the three *SIRE* families. *C. mufindiensis* resides in humid evergreen forests between 1,200 and 2,150 meters, similarly to *C. eugenioides*, whereas *C. pseudozanguebariae* and *C. racemosa* are rather dispersed in lower altitudes areas in seasonally dry and often littoral evergreen forests (Davis et al. 2006). Did these clear habitat differences influence the *SIRE* differentiation during their adaptation to different environments? Deep analyses at the level of population for each species are now necessary to better understand the impact of the environment to transposable elements diversification.

CONCLUSION

SIREs have been found as a very active lineage of elements in all studied plant species. In *Coffea* genus, they are present in all species but with different copy number amplitudes following phylogenetic clades, species diversification and probably adaptation to their environments. RT and ENV based phylogenetic analysis highlight a higher diversity of *SIRE* in wild species than previously expected based on the *C. canephora* genome. An accurate re-classification of *SIRE* families using wild species sequencing resources will bring more detailed information of how *Coffea* species diversified in Africa and Western Indian Ocean Islands and their role in ecological adaptation of species.

REFERENCES

Altschul SF, Gish W, Miller W, et al (1990) Basic Local Alignment Search Tool. J Mol Biol 215:403–410.

Andrianasolo DN, Davis AP, Razafinarivo NJ, et al (2013) High genetic diversity of in situ and ex situ populations of Madagascan coffee species: Further implications for the management of coffee genetic resources. Tree Genet Genomes 9:1295–1312. doi: 10.1007/s11295-013-0638-4

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

Bouharmont J (1959) Recherches sur les affinités chromosomiques dans le genre Coffea. I.N.É.A.C., Montpellier

Bousios A, Darzentas N, Tsaftaris A, Pearce SR (2010) Highly conserved motifs in non-coding regions of Sirevirus retrotransposons: the key for their pattern of distribution within and across plants? BMC Genomics 11:1–14. doi: 10.1186/1471-2164-11-89

Carvalho A (1952) Taxonomia de Coffea Arabica L. VI - Caracteres morfologicos dos haploides. Bragantia 12:201–212.

Charrier A (1978) La Structure génétique des caféiers spontanées de la région malgache (Mascarocoffea). Leur relations avec les caféiers d'origine africaine (Eucoffea).

Couturon E, Raharimalala NE, Rakotomalala J-J, et al (2016) Caféiers sauvages - Un trésor en péril au coeur des forêts tropicales ! Montpellier

Cros J, Lashermes P, Marmey P, et al (1993) Molecular analysis of genetic diversity and phylogenetic relationships in Coffea. In: Quinzième colloque scientifique international sur le café. Association Scientifique Internationale du Café (ASIC), Montpellier, pp 41–

46
Davis AP, Govaerts R, Bridson DM, Stoffelen P (2006) An annotatated taxonomic of the genus coffea (Rubiaceae). Bot J Linn Soc 152:465–512.
Davis AP, Tosh J, Ruch N, Fay MF (2011) Growing coffee: Psilanthus (Rubiaceae) subsumed on the basis of molecular and morphological data; implications for the size, morphology, distribution and evolutionary history of Coffea. Bot J Linn Soc 167:357–377. doi: 10.1111/j.1095-8339.2011.01177.x
De Kochko A, Akaffou S, Andrade AC, et al (2010) Advances in Coffea Genomics. Adv Bot Res 53:23–63. doi: 10.1016/S0065-2296(10)53002-7
Dias ES, Hatt C, Hamon S, et al (2015) Large distribution and high sequence identity of a Copia-type retrotransposon in angiosperm families. Plant Mol Biol 89:83–97. doi: 10.1007/s11103-015-0352-8
Dupeyron M, de Souza RF, Hamon P, et al (2017) Distribution of Divo in Coffea genomes, a poorly described family of angiosperm LTR-Retrotransposons. Mol Genet Genomics 1–14. doi: 10.1007/s00438-017-1308-2
Eickbush TH, Malik HS (2002) Origins and Evolution of Retrotransposons. ASM Press, Washington DC
Gomez C, Dussert S, Hamon P, et al (2009) Current genetic differentiation of Coffea canephora Pierre ex A. Froehn in the Guineo-Congolian African zone: cumulative impact of ancient climatic changes and recent human activities. BMC Evol Biol 9:1–19. doi: 10.1186/1471-2148-9-167
Guyot R, Darré T, Dupeyron M, et al (2016) Partial sequencing reveals the transposable element composition of Coffea genomes and provides evidence for distinct evolutionary stories. Mol Genet Genomics 291:1979–1990. doi: 10.1007/s00438-016-1235-7
Hamon P, Duroy PO, Dubreuil-Tranchant C, et al (2011) Two novel Ty1-copia retrotransposons isolated from coffee trees can effectively reveal evolutionary relationships in the Coffea genus (Rubiaceae). Mol Genet Genomics 285:447–460. doi: 10.1007/s00438-011-0617-0
Hamon P, Grover CE, Davis AP, et al (2017) Genotyping-by-sequencing provides the first well-resolved phylogeny for coffee (Coffea) and insights into the evolution of caffeine content in its species. Mol Phylogenet Evol 109:351–361. doi: 10.1016/j.ympev.2017.02.009
Hamon P, Siljak-Yakovlev S, Srisuwan S, et al (2009) Physical mapping of rDNA and heterochromatin in chromosomes of 16 Coffea species: A revised view of species

differentiation. Chromosom Res 17:291–304. doi: 10.1007/s10577-009-9033-2

Havecker ER, Gao X, Voytas DF (2004) The diversity of LTR retrotransposons. Genome Biol 5:225.1-225.6.

Kumar A, Bennetzen JL (1999) Plant Retrotransposons. Annu Rev Genet 33:479–532.

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 72:181–204. doi: 10.1038/nature13314.A

Laten HM, Gaston GD (2012) Plant endogenous retroviruses? A case of mysterious ORFs. In: Grandbastien M-A, Casacuberta JM (eds) Plant Transposable Elements. Springer-Verlag, pp 89–112

Laten HM, Majumbar A, Gaucher EA (1998) SIRE-1, a copia/Ty1-like retroelement from soybean, encodes a retroviral envelope-like protein. Proc Natl Acad Sci U S A 95:6897–6902.

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

Louarn J (1976) Hybrides interspécifiques entre Coffea canephora Pierre et C. eugenioides Moore. 20:33–52.

Louarn J (1992) La fertilité des hybrides interspécifiques et les relations génomiques entre caféiers diploïdes d'origines africaine (Genre Coffea L. sous-genre Coffea). Université Paris-Sud, centre d'Orsay

Maurin O, Davis AP, Chester M, et al (2007) Towards a phylogeny for Coffea (Rubiaceae): Identifying well-supported lineages based on nuclear and plastid DNA sequences. Ann Bot 1–19. doi: 10.1093/aob/mcm257

Nowak MD, Davis AP, Yoder AD (2012) Sequence Data from New Plastid and Nuclear COSII Regions Resolves Early Diverging Lineages in Coffea (Rubiaceae). Syst Bot 37:995–1005. doi: 10.1600/036364412X656482

Razafinarivo NJ, Guyot R, Davis AP, et al (2013) Genetic structure and diversity of coffee (Coffea) across Africa and the Indian Ocean islands revealed using microsatellites. Ann Bot 111:229–248. doi: 10.1093/aob/mcs283

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

Robbrecht E, Manen J-F (2006) The major evolutionary lineages of the coffee family (Rubiaceae, angiosperms). Combined analysis (nDNA and cpDNA) to infer the position of Coptosapelta and Luculia, and supertree construction based on rbcL, rps16, trnL-trnF

and atpB-rbcL data. A new class. Syst Geogr Plants 76:85–146.

Simpson JT, Wong K, Jackman SD, et al (2009) ABySS : A parallel assembler for short read sequence data. Genome Res 1117–1123. doi: 10.1101/gr.089532.108

Thompson JD, Higgins DG, Gibson TJ (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.

USAID (2008) 118/119 Biodiversity and Tropical Forest Assessment for Angola.

White F (1983) The vegetation of Africa. A descriptive memoir to accompany the Unesco/AEIFAT/UNSO vegetation map of Africa. Paris

Zimin A V., Marçais G, Puiu D, et al (2013) The MaSuRCA genome assembler. Bioinformatics 29:2669–2677. doi: 10.1093/bioinformatics/btt476

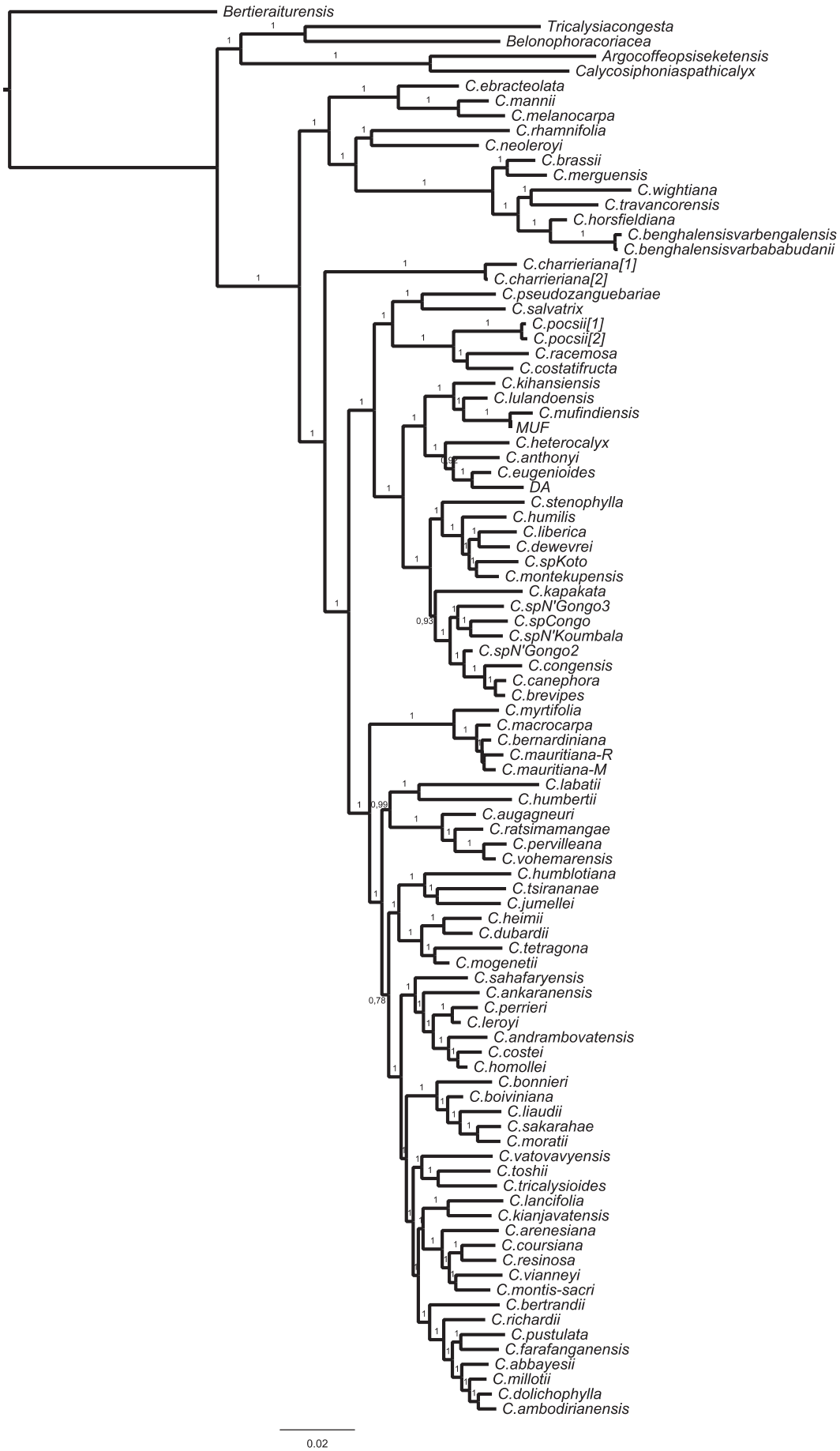


Figure S1: Phylogeny of the *Coffea* genus, from Hamon et al. 2017.

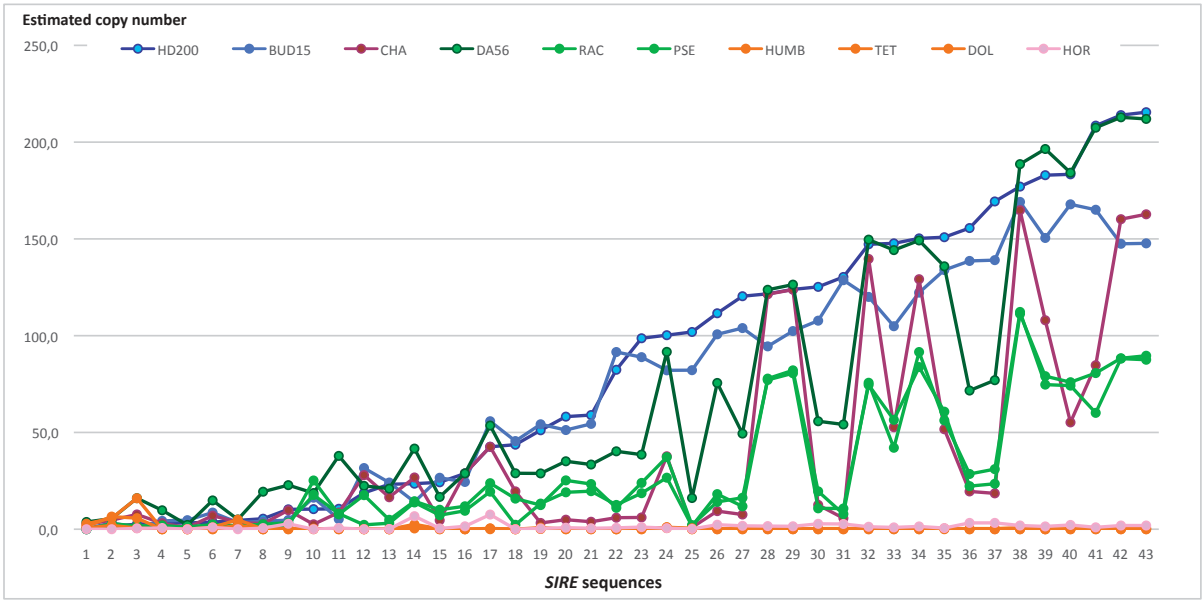


Figure S2: detection of *SIRE* elements in ten *Coffea* genome sequences obtained with 454 sequencing technology. HD200: *C. canephora* (Democratic Republic of the Congo); CHA: C. *charrieriana*; BUD15: *C. canephora* (Uganda); DA56: *C. eugenioides* (Kenya); RAC: *C. racemosa*; PSE: C. *pseudozanguebariae*; HUMB: *C. humblotiana*; DOL: *C. dolichophylla*; TET: *C. tetragona*; HOR: *C. horsefieldiana*.

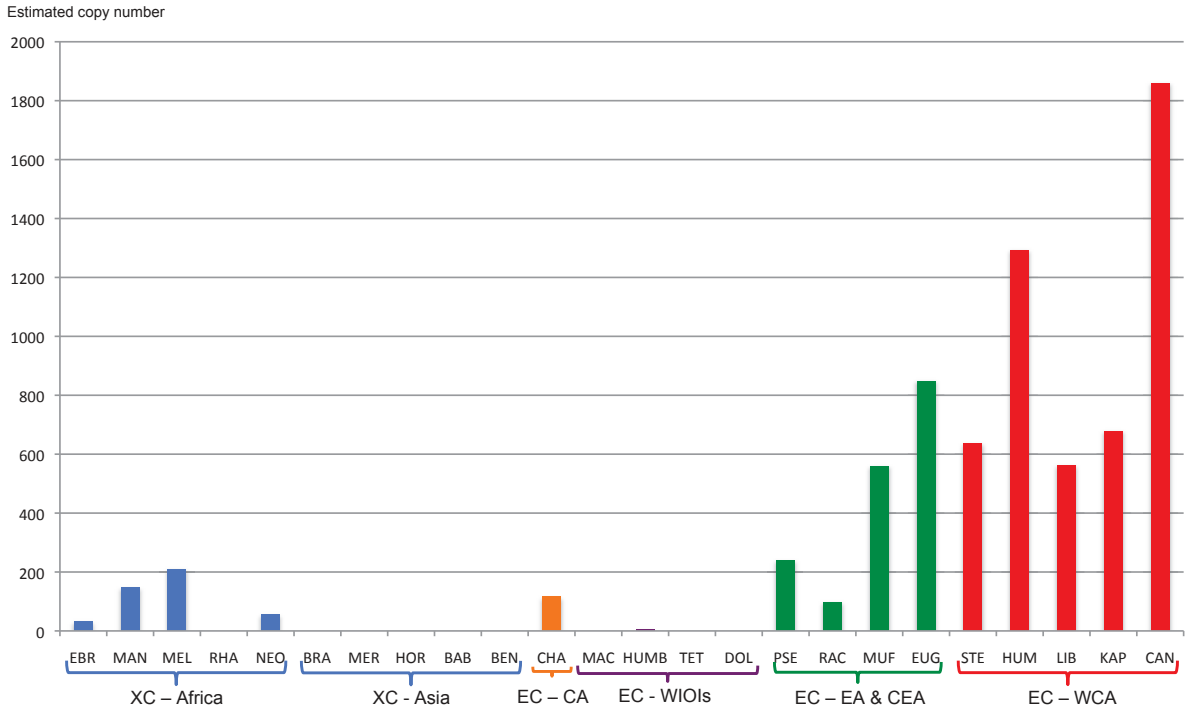


Figure S3: Copy number estimation of *SIRE* ENV domains in 24 species of coffee-trees classified according to their geographical origin.
XC: Xeno-Coffea; EC-CA: Eu-Coffea – Central Africa; WIOIs: Western Indian Ocean Islands; EA & CEA: East and Centre-East Africa; WCA: West and Central Africa. EBR: *C. ebracteolata*; MAN: *C. mannii*; MEL: *C. melanocarpa*; RHA: *C. rhamnifolia*; NEO: *C. neoleroyi*; BRA: *C. brassii*; MER: *C. merguensis*; HOR: *C. horsefieldiana*; BAB: *C. bababudanii*; BEN: *C. benghalensis*; CHA: *C. charrieriana*; MAC: *C. macrocarpa*; HUMB: *C. humblotiana*; TET: *C. tetragona*; DOL: *C. dolichophylla*; PSE: *C. pseudozanguebariae*; RAC: *C. racemosa*; MUF: *C. mufindiensis*; EUG: *C. eugenioides* (DA56, Kenya); STE=*Coffea stenophylla*; HUM=*C. humilis*; LIB=*C. liberica*; KAP: *C. kapakata*; CAN: *C. canephora* (BUD15, Uganda).

3. Conclusions et perspectives

L'étude présentée dans cet article a donc confirmé l'utilité des éléments *SIRE* comme outils moléculaires pour comprendre l'évolution du genre *Coffea*. Contrairement à ce que l'étude précédente avait montré, les *SIRE* ne sont pas absents ou quasiment absents de certains génomes. Ils sont présents mais sous des formes divergentes selon les clades considérés, par rapport aux éléments de références décrits chez *C. canephora*. Leur détection dans ces espèces distantes nécessite donc de varier des méthodologies les plus strictes (ou stringeantes) vers des approches plus relâchées.

On peut conclure des études sur cette lignée que l'activité et la divergence des éléments *SIRE* a accompagné l'évolution du genre *Coffea* et la diversification des espèces selon leur distribution géographique. Certaines espèces montrent une structure des *SIRE* différente de celles des espèces du clade auquel elles appartiennent (*C. melanocarpa* pour les Xeno-Coffea d'Afrique, *C. mufindiensis* pour les Eu-coffea d'Afrique de l'est), suggérant une évolution particulière de leur génome. Une analyse plus fine du contenu et de l'évolution de ces génomes permettra certainement de comprendre leur histoire évolutive.

Chapitre 7 – Discussion et perspectives

1. Conclusion générale

Mon travail de thèse avait pour objectif principal d'analyser la dynamique des LTR-RT au sein du genre *Coffea*, en se focalisant sur l'étude de l'impact d'une famille et d'une lignée de LTR-RT sur l'unique évènement de polyploïdisation du genre et sur son évolution. Quatre articles ont mis en valeur les travaux réalisés : l'un sur les données de séquençage partiels pour analyser le contenu global en LTR-RT de 11 espèces de caféiers, le deuxième sur la famille *Divo*, appartenant à la lignée peu connue *Bianca* et son impact sur l'allopolyploïdie, le troisième sur la structuration et la dynamique de la lignée *SIRE* dans l'allotétraploïde *C. arabica* et ses deux progéniteurs et enfin le quatrième sur l'évolution des *SIRE* dans le genre *Coffea*. Ceci a pu être réalisé grâce aux données de séquençage PacBio de *C. canephora*, *C. arabica* et *C. eugenioides* (ACGC), ainsi qu'aux données de séquençage avec la technologie Illumina de 24 génomes d'espèces sauvages (projet G13).

Les résultats principaux de ces études sont les suivants :

- Les analyses des données de séquençage 454 ont permis de montrer que les variations de taille des génomes observées pour les 11 espèces étudiées ne sont pas expliquées par la composition globale de ces génomes en ET. Cependant, l'analyse plus fine de la lignée des *SIRE* avec les séquences de référence de *C. canephora* a montré une disparité du nombre de copies selon les groupes bio-géographiques considérés. Ces données de séquençage partiel sont donc utiles à une première analyse de la composition des génomes en LTR-RT, mais ont nécessité d'être complétées pour comprendre plus finement leur dynamique et leurs impacts sur les génomes des *Coffea*.
- La famille *Divo*, appartenant à la lignée *Copia* : *Bianca*, est, contrairement aux *SIRE,* présente chez toutes les espèces de caféiers étudiées sans grande variation du nombre de copies. Cet élément montre une activité significative chez *C. canephora*, ce qui n'est pas le cas chez *C. eugenioides* qui semble contrôler

l'activité de cette famille. Cette dynamique différente entre les deux génomes à l'origine de *C. arabica* ne semble pas avoir provoqué de grands changements chez l'allotétraploïde, où *Divo* est présent et semble avoir eu une activité récente discrète, correspondant aux deux génomes parentaux. Ceci suggère que l'activité de *Divo* est indépendante des évènements à l'origine de la polyploïdie ou d'éventuels évènements post-polyploïdie.

- La lignée des *SIRE* est particulièrement structurée chez *C. canephora* et *C. eugenioides*, mais pas dans toutes les populations de ces deux espèces, suggérant une dynamique différente de ces éléments au cours de leur différenciation. Concernant la polyploïdisation, là encore aucun grand bouleversement génomique n'est observé en lien avec l'activité des *SIRE*. Cependant, la famille C, la plus ancienne des *SIRE,* semble avoir été éliminée du génome de *C. arabica* (présence de nombreux solo-LTR témoins de recombinaisons homologues inégales) alors que la famille A, la plus récente, a montré une forte activité vraisemblablement post-polyploïdisation. La structuration des *SIRE* en trois familles est donc différente selon les populations de *C. canephora* et *C. eugenioides*, suggérant un lien avec leur adaptation à différents environnements et la dynamique de ces trois familles est également différente, notamment dans le génome de *C. arabica*.
- Enfin, l'analyse des trois familles des *SIRE,* au travers du séquençage Illumina de 24 espèces/sous-espèces de caféiers sauvages, précise et complète les observations faites à partir des données de séquençage partiel. Les *SIRE* sont en réalité présents dans tous les génomes étudiés, mais toutes les familles telles que définies dans le génome de *C. canephora* (accession HD 200-94) ne sont pas retrouvées dans tous les génomes. L'évolution des *SIRE* est donc fortement associée à la divergence des espèces, avec des différences à un niveau plus fin pour certaines espèces comme dans le clade des Xeno-Coffea.

2. Discussion générale et perspectives

Annotation et caractérisation des LTR-RT

L'une des problématiques soulevée par notre étude sur *Divo* et sur la lignée des *SIRE* est celle de la détection plus générale des ET dans les génomes et leur classification. En effet, lorsqu'un élément n'a été que peu de fois détecté et de ce fait, est peu voire pas caractérisé, il est difficile de le retrouver dans les bases de données spécifiques. Si la détection et la classification se basent sur des approches par similarité, des erreurs et des manques dans les bases de données peuvent engendrer une mauvaise détection et classification. Dans le cas de *Divo* et de la lignée *Bianca*, le faible nombre de copies retrouvées en général a sûrement participé à cette difficulté de détection. Les premiers éléments découverts de cette lignée étaient partiels (Wicker et al. 2007; Hamon et al. 2011), suggérant que *Bianca* et *Divo* étaient des *Copia* anciens et en cours d'élimination et aucun domaine de référence n'était disponible dans la base de données GyDB. De plus, chaque *Bianca* de chaque espèce a été nommé spécifiquement : *Bianca* pour *Hordeum vulgare*, *Beyla* pour *Oryza sativa* et *Romani* pour *Arabidopsis thaliana* (Wicker and Keller 2007), *Matita* pour l'arachide (Nielen et al. 2012). Le manque de données sur *Bianca*, ces noms différents selon les espèces et des éléments seulement partiels décrits ont participé à la difficulté d'assigner *Divo* à une des lignées de *Copia* existantes, les recherches par BLAST ne montrant pas de similarités avec des *Bianca* connus placés dans NCBI. Une autre problématique est la détection bio-informatique des éléments complets et fonctionnels, particulièrement pour les familles de LTR-RT à petits nombres de copies comme *Divo*. L'absence de copie complète de *Bianca* décrite dans *A. thaliana* est surprenante, même si les *Romani* de la base de données TAIR (https://www.arabidopsis.org) sont relativement complets. Pourtant, nous avons détecté une copie complète et potentiellement fonctionnelle dans le génome d'*A. thaliana*, c'est-à-dire comportant les domaines complets Gag et Pol, sans codon stop. Dans *O. sativa*, un plus grand nombre de copies, mais toutes dégradées (délétions, codons stop) ont aussi été détectées, en accord avec ce qui a été observé chez *Hordeum vulgare* (Wicker and Keller 2007). Les autres études recherchant la présence de *Bianca* dans des génomes variés comme ceux de l'herbe à chapelets (*Coix lacrima-jobi*) de la pourghère (*Jatropha curcas*) ou du quinoa ont été basées sur la détection par

hybridation in situ (FISH). Cependant, les LTR-RT n'ont pas été caractérisés et le nombre de leurs copies n'a pas été estimé (Alipour et al. 2013; Kolano et al. 2013; Cai et al. 2014).

Bianca a semblé être une lignée de *Copia* disparaissant peu à peu des génomes et donc sans grand intérêt pendant longtemps. Cependant, elle a été retrouvée en grand nombre de copies (le plus grand nombre de copies parmi les lignées de *Copia*) dans le génome de la poire (Yin et al. 2015) et en nombre de copies modéré chez les caféiers. Ainsi, il semblerait que l'activité de *Bianca* soit différente selon la nature monocotylédone ou dicotylédone de l'organisme considéré et dans tous les cas étudiés, supérieure dans les dicotylédones. La faible activité transcriptionnelle et insertionnelle d'une famille de LTR-RT associée à l'évitement des systèmes de contrôle des génomes garantirait sa persistance. Au sein des dicotylédones, *Bianca* garderait donc une activité au minimum modérée.

Concernant la lignée des *SIRE*, la difficulté de leur détection dans certaines espèces de caféiers rejoint la problématique de détection et classification des ET dans les génomes. En effet, nous avons vu que selon les méthodes utilisées, les *SIRE* spécifiques de *C. canephora* ne sont pas détectables dans les espèces des IOI et peu détectables chez les Xeno-Coffea d'Asie particulièrement (Chapitres 3 et 6). L'idéal pour étudier une lignée de LTR-RT dans un genre entier est donc de rechercher des domaines amino-acides conservés de cette lignée avec des méthodes moins strictes comme le BLAST, plutôt que des méthodes utilisant l'homologie de séquence très précises comme le mapping des lectures nucléotidiques. Ce type de spécificité des *SIRE* dans certains clades des *Coffea* n'est pas ou très peu observable par exemple avec *Divo*, détecté même par mapping dans tous les génomes *Coffea* étudiés. L'analyse des *SIRE* a montré que leur détection faible dans les génomes des espèces des IOI et d'Asie n'était pas due à un faible nombre de copies, mais à une divergence importante des séquences de *C. canephora* par rapport aux copies dans ces espèces.

Dans le cas de *Divo* et des *SIRE*, la présence de nombreux éléments partiels, dégénérés et recombinés (solo-LTR) participe aussi à la difficulté de la classification et de l'annotation des LTR-RT. La création d'une base de donnée expertisée des ET chez les caféiers doit sûrement être entreprise pour faciliter l'analyse des séquences génomiques qui seront générées dans les prochaines années chez les Rubiacées. Plus globalement, la qualité de séquençage et d'assemblage des génomes est indispensable à une bonne

détection et caractérisation d'anciens LTR-RT encore actifs dans certains génomes, comme le sont *Bianca* et les *SIRE*.

Les *SIRE* et leur domaine *enveloppe*

Les *SIRE* sont des LTR-RT *Copia* bien plus connus et étudiés que les *Bianca*. Cela n'empêche pas de soulever des questions à leur propos encore aujourd'hui, notamment sur la présence du domaine *enveloppe* (ENV), son utilité au cycle de vie des *SIRE* et à leur évolution. Jusqu'à maintenant, les *SIRE* ont été étudiés principalement dans des plantes annuelles de type Monocotylédone et Dicotylédone (Laten et al. 2003; Pearce 2007; Weber et al. 2010; Bousios et al. 2012). Moins d'études ont concerné les plantes pérennes ligneuses, et elles se basaient surtout sur la détection du domaine *enveloppe* (Miguel et al. 2008; Carvalho et al. 2010), limitant la possibilité d'étudier des *SIRE* sans *enveloppe* comme ceux de la famille C des caféiers (Chapitres 5 et 6). Présents également chez les Gymnospermes, les *SIRE* sont sans aucun doute une lignée très ancienne. Étant donné le manque de connaissances et de techniques d'étude précises sur la mobilisation des LTR-RT, on ne peut qu'émettre des hypothèses quant à l'acquisition de l'*enveloppe* et l'histoire des *SIRE* dans les génomes des caféiers et plus largement des plantes. Ce domaine ENV est présent dans beaucoup de séquences de *SIRE*, y compris chez les Gymnospermes (Miguel et al. 2008). Chez les caféiers, l'arbre raciné avec des « outgroups » indique que la famille la plus ancienne ne contient pas le domaine *ENV*, contrairement aux familles A et B, qui semblent plus récentes. Les profils d'arbres obtenus à partir des séquences RT ou ENV au regard de la phylogénie des caféiers semblent en accord avec cette constatation. Par contre, la présence de séquences ENV chez *C. rhamnifolia* et *C. neoleroyi*, toutes deux appartenant au clade des Xeno-Coffea, semble indiquer que l'enveloppe a pu être acquise très tôt chez un ancêtre commun aux caféiers, conduisant à une population d'éléments SIRE avec et sans *enveloppe* ou être acquise plus récemment mais de manière indépendante par l'ancêtre des deux espèces sœurs pré-citées et par les espèces africaines du clade Eu-Coffea. Ces deux types d'éléments auraient alors évolué indépendamment, donnant la famille C sans gène *ENV* (présente chez tous les caféiers) et de nouvelles familles (en gris - Chapitre 6), puis les familles A et B présentes essentiellement dans les espèces d'Afrique de l'ouest et du centre et de nouvelles familles (différentes ou non des précédentes) avec *enveloppe*.

L'acquisition de l'*enveloppe* peut représenter une innovation évolutive ayant facilité l'activité des séquences, menant à ce que nous observons aujourd'hui : une famille C en petit nombre de copies et en cours d'élimination de certains génomes comme celui de *C. arabica*, des familles (A et B ou autres) toujours actives et en plus grand nombre de copies. Chez les rétrovirus, l'*enveloppe* est nécessaire au passage des particules virales entre les cellules (Eickbush and Malik 2002). Ceci est possible chez les animaux, mais n'a pas été démontré chez les plantes, pour lesquelles la paroi des cellules est une barrière importante au passage des particules virales. Il reste donc à comprendre l'utilité de ce domaine supplémentaire chez les LTR-RT des plantes. L'origine ancienne des *SIRE* suggère leur présence dans les autres genres de la tribu des *Coffeeae*, voire la famille des Rubiacées dans son ensemble. L'étude conjointe de la structuration des *SIRE* en parallèle à l'établissement d'une phylogénie moléculaire robuste des Rubiacées permettrait sans doute d'appréhender des tranches de l'histoire évolutive de cette famille. L'analyse du génome de *C. humblotiana*, séquencé récemment avec la technologie PacBio (D. Crouzillat, communication personnelle) pourrait permettre l'identification de séquences de *SIRE* complètes spécifiques à ce génome et à ceux des espèces des îles de l'océan Indien. Ces éléments serviraient alors à mieux comprendre l'origine de la différenciation des *SIRE* chez les caféiers.

Dynamique de *Divo* et des *SIRE* chez les caféiers

Divo est vraisemblablement ancien et peu actif globalement au niveau du genre *Coffea*, alors que les *SIRE*, même s'ils sont anciens, sont plus actifs et se différencient en plusieurs familles. Par contre, au niveau spécifique, l'espèce qui semble être le siège d'une activité particulièrement importante à la fois de *Divo* et des *SIRE* est sans aucun doute *C. canephora*. En fait, c'est l'espèce ayant la plus grande répartition géographique naturelle allant d'ouest en est de la Guinée Conakry à l'Ouganda et du nord au sud, de la République Centrafrique à l'Angola (Berthaud 1986). C'est l'espèce la plus différenciée génétiquement avec 6 groupes identifiés jusqu'à maintenant (Musoli et al. 2009; Gomez et al. 2009). *Divo* est associé à cette différenciation (Hamon et al. 2011; Dupeyron et al. 2017). Il ne nous a pas été possible d'analyser des représentants des différents groupes génétiques de *C. canephora* dans notre étude des *SIRE*. Néanmoins, la structuration des *SIRE* dans les deux accessions disponibles de *C. canephora* indique une différenciation

moins poussée dans l'accession d'Ouganda avec la présence de familles n'étant pas A, B, ou C. Quoi qu'il en soit, les LTR-RT *Copia* et en particulier *Divo* et *SIRE* ont pu jouer un rôle dans/ou être associés à sa diversification.

La présence de *Divo* près de gènes chez *C. canephora* pourrait être une indication de son rôle potentiel dans la régulation des gènes, au moins chez cette espèce. L'expression différentielle de *Divo* et des *SIRE* (étendue à tous les LTR-RT détectés dans *C. canephora*) dans les génomes de *C. canephora*, *C. arabica* et *C. eugenioides* à partir de données RNAseq, sous condition de différents stresses biotiques et abiotiques (disponibles au sein de l'équipe CoffeeAdapt) devra être entreprise. Ainsi, comme il a été démontré chez le riz avec *Tos17*, en faible nombre de copies mais s'exprimant dans le cas du stress induit par la culture cellulaire (Piffanelli 2007), *Divo* pourrait avoir une augmentation de son activité insertionnelle en cas de stress. Les *SIRE* en plus grand nombre de copies comme *Tnt1* chez le tabac (Grandbastien et al. 1989), pourraient également être exprimés dans certaines conditions de stress, biotiques notamment (Bui & Grandbastien 2012). Certaines copies complètes de *Divo* se situent près de gènes chez *C. canephora* (Chapitre 4). On peut supposer la même chose pour les *SIRE*, pour lesquels je n'ai pas eu le temps de vérifier leur localisation chez *C. canephora*. Les données RNAseq disponibles au sein de l'équipe CoffeeAdapt concernent plusieurs conditions d'expérimentations (luminosité, altitude, température, stress hydrique, cycle circadien, infection par des pathogènes, variétés hybrides ou sauvages) pouvant avoir une influence sur l'expression des gènes chez les caféiers cultivés. L'outil TEtools (Lerat et al. 2016) pourra être utilisé spécifiquement pour analyser l'expression différentielle des ET. Les LTR-RT détectés par LTR_STRUC dans les génomes de *C. canephora*, *C. arabica* et *C. eugenioides* pourront être triés et mis sous forme de liste pour le fichier de référence nécessaire au fonctionnement de l'outil. Selon la condition de stress considérée, il n'est pas inconsidéré de penser que certains LTR-RT auront un patron d'expression différentiel qui pourrait permettre de mieux comprendre le fonctionnement de certains gènes, notamment ceux proches de copies complètes de LTR-RT. Leur expression différentielle indiquerait également que des mécanismes de « silencing » sont levés quand la plante est stressée, peut-être pour que l'activation de certains ET module l'expression de gènes spécifiques, comme il a déjà été proposé chez Arabidopsis (Ito 2012).

C. arabica et polyploïdie

Comme pour *Divo*, l'évènement de polyploïdisation ne semble pas avoir eu de forts impacts sur les *SIRE*, du moins globalement (Chapitres 4 et 5). Il me semble important de souligner le fait que quasiment aucune séquence de la famille C n'a pu être détectée dans *C. arabica* et que de nombreux solo-LTR ont été détecté (Chapitre 5). Par ailleurs, la famille A semble avoir récemment subi une amplification très importante étant donné le nombre estimé de copies. On peut ainsi penser que la polyploïdisation a été associée à l'élimination de la plupart des copies de la famille C, déjà en cours chez *C. eugenioides* et en moindre mesure chez *C. canephora*. Un mécanisme de contrôle du nombre de copies chez *C. arabica* après l'évènement de polyploïdisation pourrait aussi avoir limité l'augmentation de la taille du génome. La polyploïdisation a aussi pu provoquer la levée des contraintes épigénétiques des copies de la famille A, permettant une amplification de celle-ci. De tels évènements ont été observés chez les blés sauvages (*Aegilops*, Senerchia et al. 2014) et l'orobanche (Piednoël et al. 2013).

L'étude de *Divo* et des *SIRE* a permis de préciser certaines observations faites précédemment avec des données de séquençage partiel dans certaines espèces de caféiers, ainsi que de donner plus d'informations sur la formation de *C. arabica*. Néanmoins, l'identification plus précise de l'origine des populations parentales aurait nécessité une étude plus exhaustive des différentes populations de *C. canephora* et *C. eugenioides* en gardant à l'esprit que les populations actuelles ne sont pas obligatoirement représentatives de leurs formes ancestrales. Les données de séquençage partiel des génomes d'un jeu d'espèces sauvages ont permis d'estimer leur composition et abondance en LTR-RT auparavant identifiés chez *C. canephora* (Guyot et al. 2016 – Chapitre 3). Ainsi d'autres éléments, tels que ceux de la lignée *Del* (LTR-RT *Gypsy*), ont montré des différences significatives du nombre de copies selon les espèces considérées (données non publiées – Figure 16). L'analyse fine de ces éléments pourrait permettre de mieux comprendre les dynamiques différentes des LTR-RT chez les caféiers et notamment leur impact sur ou leur évolution propre sous le prisme de l'évolution du genre *Coffea* et de l'unique évènement de polyploïdisation naturelle dans ce genre.

Contrairement à ce qui a été observé dans de nombreuses plantes annuelles (Vicient and Casacuberta 2017), le génome de *C. arabica* ne semble pas avoir subi de réarrangements drastiques ou bien d'augmentation ou diminution significative de la

taille de son génome. De plus, il présente les mêmes lignées de *Copia* et *Gypsy* que ses deux progéniteurs, sans grands changements (excepté peut-être pour la famille C des *SIRE* et des petites modifications potentielles dans d'autres lignées de LTR-RT). Dans les malheureusement peu nombreuses études concernant l'évolution des LTR-RT dans des génomes de plantes pérennes ligneuses, ces mêmes conclusions ont été formulées. Ceci suggère que la différence majeure entre les plantes annuelles et pérennes, à savoir leur cycle de vie, est probablement en liaison avec l'importance des restructurations et donc de l'évolution de ces génomes. Pour faire face à un stress, une plante annuelle doit très vite s'adapter et être en mesure de se reproduire faute de quoi, la survie de la population ou de l'espèce peut être remise en cause. Au niveau individuel, la plante sans cette adaptation pourrait perdre la capacité à transmettre ses gènes. Une plante pérenne ligneuse, au contraire, n'a pas besoin de se reproduire tous les ans puisqu'elle va vivre plusieurs années : un ou plusieurs stresses empêchant la reproduction une année ne se traduira pas obligatoirement par la mort directe de la plante et surtout, par son impossibilité à participer à la constitution des générations suivantes. La plante pérenne pourrait donc réagir de manière moins brutale que la plante annuelle. C'est peut être cette différence d'intensité de réactions qui aurait des conséquences différentes par exemple sur l'intensité d'activation des LTR-RT. Ce type de différence entre plantes pérennes et annuelles a été soulignée dans l'étude sur la dynamique des *Copia* et des *Gypsy* dans le génome du peuplier (Natali et al. 2015). Plus d'études sur l'activité et l'évolution des LTR-RT dans des génomes de plantes pérennes ligneuses sont nécessaires pour confirmer cette hypothèse. Avec l'accessibilité croissante de données génomiques, l'étude comparative entre génomes issus d'espèces ligneuses pérennes et d'espèces annuelles est possible. L'étude détaillée des LTR-RT devrait fournir des éléments de réponse quant au comportement différentiel entre ces deux types de plantes.

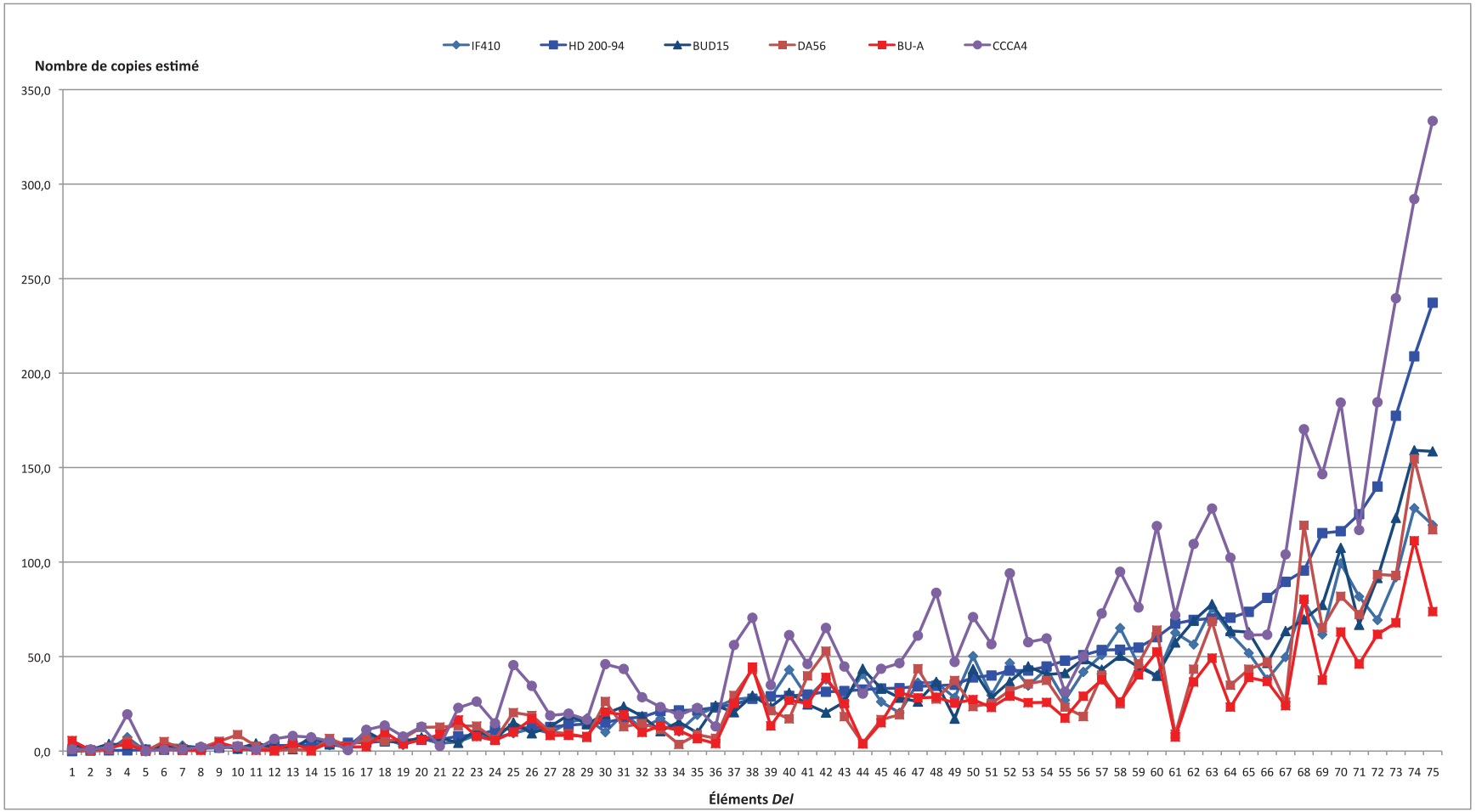


Figure 16 : Estimation du nombre de copies de *Del* dans six séquences génomiques de *Coffea*. IF410 : *C. canephora* de Guinée ; HD-200 : *C. canephora* de RDC ; BUD15 : *C. canephora* d'Ouganda ; DA56 : *C. eugenioides* du Kenya ; BUD14 : *C. eugenioides* d'Ouganda ; CCCA4 : *C. arabica* cultivé.

Évolution du genre *Coffea*

Les études réalisées durant ma thèse ont montré que le genre *Coffea* contient deux lignées LTR-RT ayant une dynamique différente, mais présentes toutes les deux depuis longtemps chez les plantes. Certaines espèces comme *C. melanocarpa* et *C. mufindiensis* montrent une structuration des *SIRE* commune à celle des caféiers d'Afrique de l'ouest et du centre, alors qu'elles font partie du clade des Xeno-Coffea pour l'une et des Eu-Coffea d'Afrique du centre-est pour l'autre. Des analyses plus poussées de *Divo* montreraient peut-être des particularités chez ces deux espèces, faisant d'elles des cas particuliers au sein du genre *Coffea,* à étudier plus en détails.

Maintenant qu'une phylogénie résolue du genre *Coffea* est disponible, il serait possible de caractériser d'éventuels évènements de transferts horizontaux (TH). Aucune horloge moléculaire n'est malheureusement disponible pour les caféiers, on ne pourrait donc pas estimer la date d'éventuels TH, mais on peut tout de même essayer d'identifier de tels évènements. Par exemple, Dias et al. (2015) suggèrent une dynamique évolutive d'un élément *Copia* issu de *C. canephora* chez les angiospermes, impliquant entre autres un possible transfert horizontal. On peut aussi rechercher des éléments particulièrement conservés entre *C. canephora* ou *C. arabica* et l'un de leurs pathogènes. En effet, un événement de TH n'est possible que si des individus de populations, espèces ou genres différents sont géographiquement et écologiquement proches. Un pathogène peut donc représenter un déclencheur ou un vecteur idéal pour un ETH (Gilbert et al. 2010). La rouille du caféier (*Hemileia vastatrix*) est un champignon microscopique infectant les caféiers par les stomates et se développant de façon extra- et intra-cellulaire dans ceux-ci (Talhinhas et al. 2017). Sa proximité physique avec les cellules des caféiers infectés est telle que le transfert de matériel génétique entre ce pathogène et ses hôtes est envisageable. De plus, la rouille semble contenir une grande quantité d'ET, plus de 74% (Cristancho et al. 2014). Ceux-ci pourraient être impliqués dans la régulation de gènes favorisant la pathogénicité du champignon. Les identifier et identifier d'éventuelles insertions dans le génome des caféiers par ETH pourraient apporter des informations très importantes pour la lutte contre la rouille du caféier.

Enfin, les analyses réalisées pendant cette thèse pourront être complétées chez d'autres espèces de caféiers sauvages, notamment celles des IOI. En effet, les séquences

Illumina de 45 génomes d'espèces malgaches sont maintenant disponibles, ainsi que deux séquençages PacBio de deux espèces sans caféine, *C. homollei* et *C. humblotiana*, le plus petit génome du genre (projet G13). Trente-six accessions de *C. arabica* sauvages et cultivés ont été re-séquencées dans le cadre du projet ACGC. Leur étude permettrait de confirmer l'origine génétique (un ou plusieurs évènements de polyploïdisation) et géographique de *C. arabica*.

3. Évolution des *SIRE* et de *Divo* dans le genre *Coffea*

Avec les données obtenues au cours de cette thèse, nous pouvons résumer nos principaux résultats dans la Figure 17. Dans cette figure, les éléments *SIRE* non-affiliés aux familles A, B ou C présents dans la quasi-totalité des espèces, notamment dans les espèces des clades les plus anciens (Hamon et al. 2017), incluent des éléments avec et sans *enveloppe*. Ceci laisse à penser que l'ancêtre commun des caféiers le plus récent comporte ces deux types de *SIRE*. La différenciation très poussée des *SIRE* (trois familles chez *C. canephora* et probablement d'aurtes familles à définir) est à mettre en parallèle avec celle observée chez les caféiers et conduisant aux deux clades frères AOC et ACE. Une forte différenciation des *SIRE* a aussi été observée chez un seul Xeno-Coffea, à savoir *C. melanocarpa*. Cette espèce n'est pas particulière qu'à cet égard, elle présente aussi d'autres caractéristiques qui en font une exception au sein des ex-*Psilanthus* (Chapitre 6). L'origine de ces particularités reste à définir.

Divo, également présent chez l'ancêtre commun le plus récent des caféiers, bien que transmis à toutes les espèces, a eu une activité indépendante de la spéciation. Pour une raison qui reste à déterminer, il a été actif dans la différenciation de *C. canephora*, sans pour autant s'être différencié en sous-familles. Cette espèce ayant une adaptation à des environnements différents, il serait intéressant d'étudier *Divo* chez *C. congensis*. C'est une espèce génétiquement très proche de *C. canephora*, puisque 80% des hypbrides F1 entre ces deux espèces sont fertiles (Louarn 1992), alors que leurs niches écologiques sont exclusives : *C. congensis* se développe dans des zones temporairement inondées (au bord des fleuves), là où ne peut pas se développer *C. canephora* (Berthaud 1986).

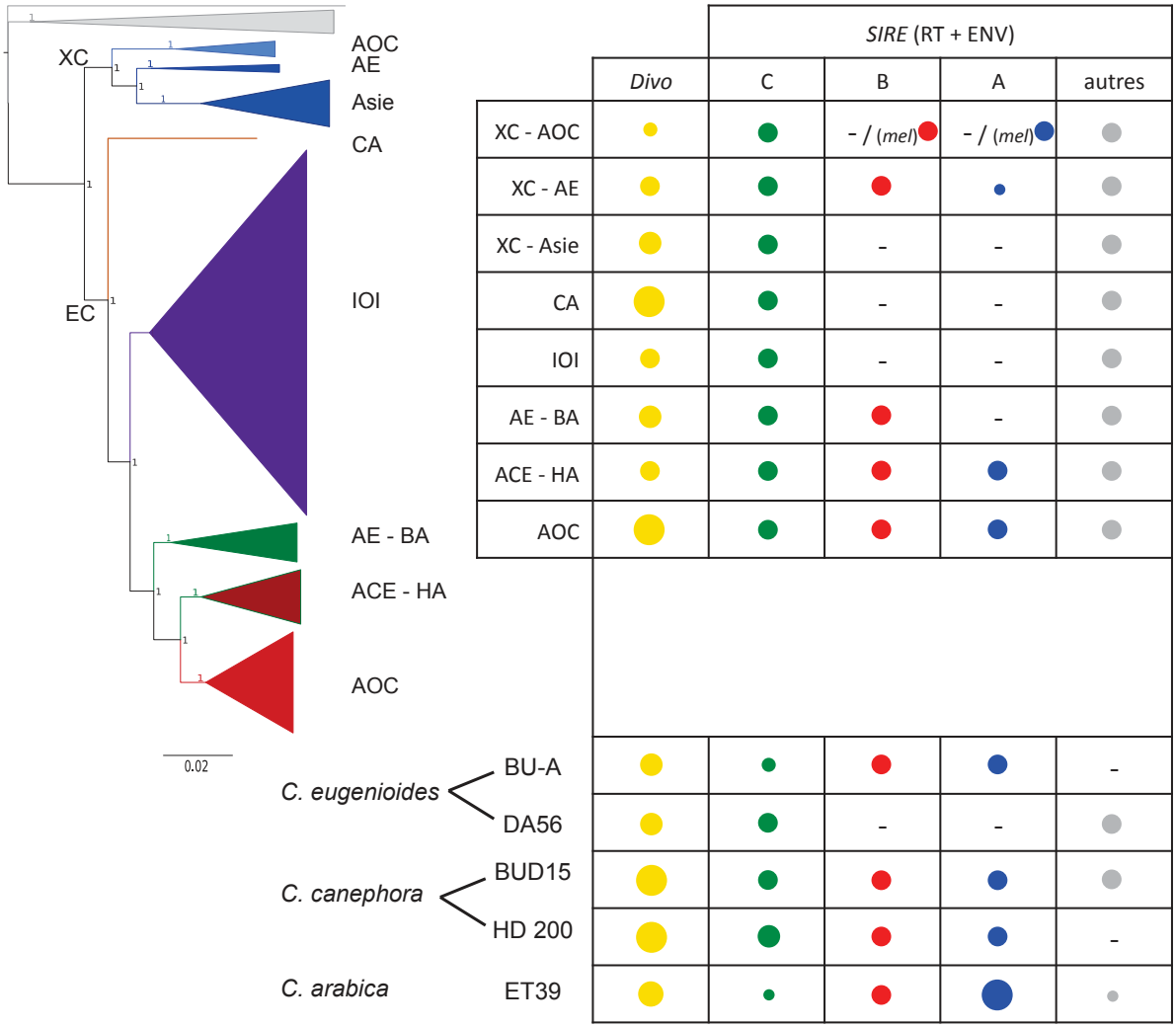


Figure 17 : Schéma de la présence et de l'absence de *Divo* et des *SIRE* dans les espèces étudiées du genre *Coffea*. Les points sont proportionnels au nombre de copies trouvées, quand estimé. - : absence ; XC : Xeno-Coffea ; EC : Eu-Coffea ; AOC : Afrique de l'ouest et du centre ; AE : Afrique de l'est ; CA : centre Afrique ; IOI : îles de l'ouest de l'océan Indien ; BA : basse altitute ; HA : haute altitude ; ACE : Afrique du centre-est.

Bibliographie

Adams MD, Celniker SE, Holt RA, et al (2000) The Genome Sequence of Drosophila melanogaster. Genetics 287:2185–2195. doi: 10.1126/science.287.5461.2185

Agren JA, Wright SI (2015) Selfish genetic elements and plant genome size evolution. Trends Plant Sci 1–2. doi: 10.1016/j.tplants.2015.03.007

Alipour A, Tsuchimoto S, Sakai H, et al (2013) Structural characterization of copia-type retrotransposons leads to insights into the marker development in a biofuel crop, Jatropha curcas L. Biotechnol Biofuels 6:1–13. doi: 10.1186/1754-6834-6-129

Alves S, Ribeiro T, Inácio V, et al (2012) Genomic organization and dynamics of repetitive DNA sequences in representatives of three Fagaceae genera. Genome 55:348–359. doi: 10.1139/g2012-020

Ammiraju JSS, Zuccolo A, Yu Y, et al (2007) Evolutionary dynamics of an ancient retrotransposon family provides insights into evolution of genome size in the genus Oryza. Plant J 52:342–351. doi: 10.1111/j.1365-313X.2007.03242.x

Andrianasolo DN, Davis AP, Razafinarivo NJ, et al (2013) High genetic diversity of in situ and ex situ populations of Madagascan coffee species: Further implications for the management of coffee genetic resources. Tree Genet Genomes 9:1295–1312. doi: 10.1007/s11295-013-0638-4

Arensburger P, Piégu B, Bigot Y (2016) The future of transposable element annotation and their classification in the light of functional genomics - what we can learn from the fables of Jean de la Fontaine? Mob Genet Elements. doi: 10.1080/2159256X.2016.1256852

Baidouri M El, Carpentier MC, Cooke R, et al (2014) Widespread and frequent horizontal transfers of transposable elements in plants. Genome Res 24:831–838. doi: 10.1101/gr.164400.113

Bao W, Kojima KK, Kohany O (2015) Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA 6:1–6. doi: 10.1186/s13100-015-0041-9

Bao Z, Eddy SR (2002) Automated de novo identification of repeat sequence families in sequenced genomes. Genome Res 12:1269–1276. doi: 10.1101/gr.88502

Bennett MD, Leitch IJ (2012) Plant DNA C-values database. http://www.kew.org/cvalues/.

Bennetzen JL (2005) Transposable elements, gene creation and genome rearrangement in flowering plants. Curr Opin Genet Dev 15:621–627. doi: 10.1016/j.gde.2005.09.010

Bennetzen JL, Kellogg E (1997) Do Plants Have a One-Way Ticket to Genomic Obesity? Plant Cell 9:1509–1514.

Bennetzen JL, Ma J, Devos KM (2005) Mechanisms of recent genome size variation in flowering plants. Ann Bot 95:127–32. doi: 10.1093/aob/mci008

Bergman CM, Quesneville H (2007) Discovering and detecting transposable elements in genome sequences. Brief Bioinform 8:382–392. doi: 10.1093/bib/bbm048

Berthaud J (1986) Les ressources génétiques pour l'amélioration des caféiers africains diploïdes. Paris Sud

Boeke JD, Garfinkel DJ, Styles CA, Fink GR (1985) Ty elements transpose through an RNA intermediate. Cell 40:491–500. doi: 10.1016/0092-8674(85)90197-7

Bouharmont J (1959) Recherches sur les affinités chromosomiques dans le genre Coffea. I.N.É.A.C., Montpellier

Bousios A, Darzentas N (2013) Sirevirus LTR retrotransposons: phylogenetic misconceptions in the plant world. Mob DNA 4:1–5. doi: 10.1186/1759-8753-4-9
Bousios A, Darzentas N, Tsaftaris A, Pearce SR (2010) Highly conserved motifs in non-coding regions of Sirevirus retrotransposons: the key for their pattern of distribution within and across plants? BMC Genomics 11:1–14. doi: 10.1186/1471-2164-11-89
Bousios A, Kourmpetis YAI, Pavlidis P, et al (2012) The turbulent life of Sirevirus retrotransposons and the evolution of the maize genome: More than ten thousand elements tell the story. Plant J 69:475–488. doi: 10.1111/j.1365-313X.2011.04806.x
Brenchley R, Spannagl M, Pfeifer M, et al (2012) Analysis of the bread wheat genome using whole genome shotgun sequencing. Nature 491:705–710. doi: 10.1038/nature11650.Analysis
Bundock P, Hooykaas P (2005) An Arabidopsis hAT-like transposase is essential for plant development. Nature 436:282–284. doi: 10.1038/nature03667
Bushman FD (2003) Targeting survival: Integration site selection by retroviruses and LTR-Retrotransposons. Cell 115:135–138. doi: 10.1016/S0092-8674(03)00760-8
Butelli E, Licciardello C, Zhang Y, et al (2012) Retrotransposons Control Fruit-Specific, Cold-Dependent Accumulation of Anthocyanins in Blood Oranges. Plant Cell 24:1242–1255. doi: 10.1105/tpc.111.095232
Cai Z, Liu H, He Q, et al (2014) Differential genome evolution and speciation of Coix lacryma-jobi L. and Coix aquatica Roxb. hybrid guangxi revealed by repetitive sequence analysis and fine karyotyping. BMC Genomics 15:1–16. doi: 10.1186/1471-2164-15-1025
Capy P, Langin T, Higuet D, et al (1997) Do the integrases of LTR-retrotransposons and class II element transposases have a common ancestor? Genetica 100:63–72. doi: 10.1023/A:1018300721953
Carr M, Bensasson D, Bergman CM (2012) Evolutionary Genomics of Transposable Elements in Saccharomyces cerevisiae. PLoS One 7:50978–50993. doi: 10.1371/journal.pone.0050978
Carvalho A (1952) Taxonomia de Coffea Arabica L. VI - Caracteres morfologicos dos haploides. Bragantia 12:201–212.
Carvalho M, Ribeiro T, Viegas W, et al (2010) Presence of env-like sequences in Quercus suber retrotransposons. J Appl Genet 51:461–467. doi: 10.1007/BF03208875
Casacuberta E, González J (2013) The impact of transposable elements in environmental adaptation. Mol Ecol 22:1503–1517. doi: 10.1111/mec.12170
Caspi A, Pachter L (2006) Identification of transposable elements using multiple alignments of related genomes. Genome Res 16:260–270. doi: 10.1101/gr.4361206
Chaparro C, Gayraud T, Souza RF De, 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
Charrier A (1978) La Structure génétique des caféiers spontanées de la région malgache (Mascarocoffea). Leur relations avec les caféiers d'origine africaine (Eucoffea).
Charrier A, Berthaud J (1985) Botanical Classification of Coffee. Coffee Bot Biochem Prod Beans Beverage 13–47. doi: 10.1007/978-1-4615-6657-1
Chevalier A (1929) Principes d'arboriculture fruitière applicables aux caféiers. Encycl. Biol. 1–27.
Comai L, Madlung A, Josefsson C, Tyagi A (2003) Do the different parental "heteromes" cause genomic shock in newly formed allopolyploids? Philos Trans R Soc London Biol Sci 358:1149–1155. doi: 10.1098/rstb.2003.1305

Comfort NC (1999) “The real point is control”: The reception of Barbara McClintock’s controlling elements. J Hist Biol 32:133–162.
Couturon E, Raharimalala NE, Rakotomalala J-J, et al (2016) Caféiers sauvages - Un trésor en péril au coeur des forêts tropicales ! Montpellier
Cristancho MA, Botero-Rozo DO, Giraldo W, et al (2014) Annotation of a hybrid partial genome of the coffee rust (Hemileia vastatrix) contributes to the gene repertoire catalog of the Pucciniales. Front Plant Sci 5:1–11. doi: 10.3389/fpls.2014.00594
Cros J, Gavalda MC, Chabrillange N, et al (1994) Variations in the total nuclear DNA content in african Coffea species (Rubiaceae). Café Cacao Thé XXXVIII:3–10.
Cros J, Lashermes P, Marmey P, et al (1993) Molecular analysis of genetic diversity and phylogenetic relationships in Coffea. In: Quinzième colloque scientifique international sur le café. Association Scientifique Internationale du Café (ASIC), Montpellier, pp 41–46
Curcio MJ, Derbyshire KM (2003) The outs and ins of transposition: from Mu to Kangaroo. Nat Rev Mol Cell Biol 4:865–877. doi: 10.1038/nrm1241
Darré T (2014) Evolution et diversité du genre Coffea à travers l’étude des rétroéléments à LTR. Université Montpellier II
Davis AP, Govaerts R, Bridson DM, Stoffelen P (2006) An annotatated taxonomic of the genus coffea (Rubiaceae). Bot J Linn Soc 152:465–512.
Davis AP, Tosh J, Ruch N, Fay MF (2011) Growing coffee: Psilanthus (Rubiaceae) subsumed on the basis of molecular and morphological data; implications for the size, morphology, distribution and evolutionary history of Coffea. Bot J Linn Soc 167:357–377. doi: 10.1111/j.1095-8339.2011.01177.x
De Kochko A, Akaffou S, Andrade AC, et al (2010) Advances in Coffea Genomics. Adv Bot Res 53:23–63. doi: 10.1016/S0065-2296(10)53002-7
de Vries H, MacDougal DT (1905) Species and Varieties: Their Origin by Mutation. The Plant World 8:86–90.
Demerec M (1935) Unstable Genes. Bot Rev 1:233–248.
Denoeud F, Carretero-Paulet L, Dereeper A, et al (2014) The Coffee Genome Provides Insight into the Convergent Evolution of Caffeine Biosynthesis. Science (80-) 345:1180–1184. doi: 10.1126/science.1255274
Dias ES, Hatt C, Hamon S, et al (2015) Large distribution and high sequence identity of a Copia-type retrotransposon in angiosperm families. Plant Mol Biol 89:83–97. doi: 10.1007/s11103-015-0352-8
Doolittle WF, Sapienza C (1980) Selfish genes, the phenotype paradigm and genome evolution. Nature 284:601–603.
Du C, Caronna J, He L, Dooner HK (2008) Computational prediction and molecular confirmation of Helitron transposons in the maize genome. BMC Genomics 9:51. doi: 10.1186/1471-2164-9-51
Dupeyron M, de Souza RF, Hamon P, et al (2017) Distribution of Divo in Coffea genomes, a poorly described family of angiosperm LTR-Retrotransposons. Mol Genet Genomics 1–14. doi: 10.1007/s00438-017-1308-2
Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–7. doi: 10.1093/nar/gkh340
Eickbush TH, Malik HS (2002) Origins and Evolution of Retrotransposons. ASM Press, Washington DC
Eilam T, Anikster Y, Millet E, et al (2008) Nuclear DNA amount and genome downsizing in natural and synthetic allopolyploids of the genera Aegilops and Triticum. Genome 51:616–627. doi: 10.1139/G08-043

Ellinghaus D, Kurtz S, Willhoeft U (2008) LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics 9:1–14. doi: 10.1186/1471-2105-9-18

Emerson RA (1914) The Inheritance of a Recurring Somatic Variation in Variegated Ears of Maize. Am Nat 48:87–115.

Ewing AD (2015) Transposable element detection from whole genome sequence data. Mob DNA 6:24–32. doi: 10.1186/s13100-015-0055-3

Faivre Rampant P, Lesur I, Boussardon C, et al (2011) Analysis of BAC end sequences in oak, a keystone forest tree species, providing insight into the composition of its genome. BMC Genomics 12:292. doi: 10.1186/1471-2164-12-292

Federoff N V. (2012) Transposable Elements, Epigenetics, and Genome Evolution. Science (80-) 338:758–767.

Fedoroff N V., Bennetzen JL (2013) Transposons , Genomic Shock , and Genome Evolution. In: Fedoroff N V. (ed) Plant Transposons and Genome Dynamics in Evolution. John Wiley & Sons, Inc., pp 181–201

Fedoroff N V (2012) Transposable Elements, Epigenetics, and Genome Evolution. Science (80-) 338:758–767.

Finnegan DJ (1989) Eucaryotic transposable elements and genome evolution. Trends Genet 5:103–107.

Fiston-Lavier A-S, Carrigan M, Petrov DA, Gonzalez J (2010) T-lex: a program for fast and accurate assessment of transposable element presence using next-generation sequencing data. Nucleic Acids Res 39:1–10. doi: 10.1093/nar/gkq1291

Fortune PM, Roulin A, Panaud O (2008) Horizontal transfer of transposable elements in plants. Commun Integr Biol 1:74–77. doi: 10.4161/cib.1.1.6328

Gilbert C, Schaack S, Pace II JK, et al (2010) A role for host-parasite interactions in the horizontal transfer of DNA transposons across animal phyla. Nature 464:1347–1350.

Gomez C, Dussert S, Hamon P, et al (2009) Current genetic differentiation of Coffea canephora Pierre ex A. Froehn in the Guineo-Congolian African zone: cumulative impact of ancient climatic changes and recent human activities. BMC Evol Biol 9:1–19. doi: 10.1186/1471-2148-9-167

Grandbastien M-A, Spielmann A, Caboche M (1989) Tnt1, a mobile retroviral-like transposable element of tobacco isolated by plant cell genetics. Nature 337:376–380.

Grandbastien MA (2015) LTR retrotransposons, handy hitchhikers of plant regulation and stress response. Biochim Biophys Acta 1849:403–416. doi: 10.1016/j.bbagrm.2014.07.017

Grover CE, Wendel JF (2010) Recent Insights into Mechanisms of Genome Size Change in Plants. J Bot 2010:1–8. doi: 10.1155/2010/382732

Guillemat J (1946) Quelques observations sur la Trachéomycose du “Coffea excelsa.” Rev Int Bot appliquée d’agriculture Trop 542–550.

Guyot R, Cheng X, Su Y, et al (2005) Complex organization and evolution of the tomato pericentromeric region at the FER gene locus. Plant Physiol 138:1205–1215. doi: 10.1104/pp.104.058099

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

Haber JE (2000) Repairing a Double-Strand Break. Trends Genet 16:259–264.

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

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

Hamon P, Hamon S, Razafinarivo NJ, et al (2015) Coffea Genome Organization and Evolution. In: Coffee in Health and Disease Prevention. pp 29–37

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

Han Y, Wessler SR (2010) MITE-Hunter: A program for discovering miniature inverted-repeat transposable elements from genomic sequences. Nucleic Acids Res 38:e199. doi: 10.1093/nar/gkq862

Havecker ER, Gao X, Voytas DF (2004) The diversity of LTR retrotransposons. Genome Biol 5:225.1-225.6.

Hawkins JS, HyeRan K, Nason JD, et al (2006) Differential lineage-specific amplification of transposable elements is responsible for genome size variation in Gossypium. Genome Res 16:1252–1261. doi: 10.1101/gr.5282906.1

Hirochika H (2001) Contribution of the Tos17 retrotransposon to rice functional genomics. Curr Opin Plant Biol 4:118–122. doi: 10.1016/S1369-5266(00)00146-1

Hirochika H, Sugimoto K, Otsuki Y, et al (1996) Retrotransposons of rice involved in mutations induced by tissue culture. Proc Natl Acad Sci U S A 93:7783–7788. doi: 10.1073/pnas.93.15.7783

Hribová E, Neumann P, Matsumoto T, et al (2010) Repetitive part of the banana (Musa acuminata) genome investigated by low-depth 454 sequencing. BMC Plant Biol 10:204–214. doi: 10.1186/1471-2229-10-204

Hu G, Hawkins JS, Grover CE, Wendel JF (2010) The history and disposition of transposable elements in polyploid Gossypium. Genome 53:599–607. doi: 10.1139/g10-038

Ibarra-Laclette E, Lyons E, Hernández-Guzmán G, et al (2013) Architecture and evolution of a minute plant genome. Nature 498:94–99. doi: 10.1038/nature12132

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

Ito H (2012) Small RNAs and transposon silencing in plants. Dev Growth Differ 54:100–107. doi: 10.1111/j.1440-169X.2011.01309.x

Jiang N, Visa S, Wu S, Van Der Knaap E (2012) Rider Transposon Insertion and Phenotypic Change in Tomato. In: Grandbastien M-A, Casacuberta JM (eds) Plant Transposable Elements. Springer-Verlag, pp 297–312

Jiang S, Cai D, Sun Y, Teng Y (2016) Isolation and characterization of putative functional long terminal repeat retrotransposons in the Pyrus genome. Mob DNA 7:1–10. doi: 10.1186/s13100-016-0058-8

Jurka J, Klonowski P, Dagman V, Pelton P (1996) Censor--A program for identification and elimination of repetitive elements from DNA sequences. Comput Chem 20:119–121. doi: 10.1016/S0097-8485(96)80013-1

Kapitonov V V., Jurka J (2001) Rolling-circle transposons in eukaryotes. Proc Natl Acad Sci 98:8714–8719. doi: 10.1073/pnas.151269298

Kashkush K, Feldman M, Levy AA (2002) Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat. Nat Genet 33:102–106. doi: 10.1038/ng1063

Keeling PJ, Palmer JD (2008) Horizontal gene transfer in eukaryotic evolution. Nat Rev Genet 9:605–618.

Kejnovsky E, Hawkins JS, Feschotte C (2012) Plant Transposable Elements: Biology and Evolution. In: Wendel JF, Greilhuber J, Dolezel J, Leitch IJ (eds) Plant Genome Diversity Volume 1. Springer Vienna, Vienna, pp 17–34

Kohany O, Gentles AJ, Hankus L, Jurka J (2006) Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. BMC Bioinformatics 7:474. doi: 10.1186/1471-2105-7-474

Kolano B, Bednara E, Weiss-Schneeweiss H (2013) Isolation and characterization of reverse transcriptase fragments of LTR retrotransposons from the genome of Chenopodium quinoa (Amaranthaceae). Plant Cell Rep 32:1575–1588. doi: 10.1007/s00299-013-1468-4

Krumsiek J, Arnold R, Rattei T (2007) Gepard: a rapid and sensitive tool for creating dotplots on genome scale. Bioinformatics 23:1026–1028. doi: 10.1093/bioinformatics/btm039

Kumar A, Bennetzen JL (1999) Plant Retrotransposons. Annu Rev Genet 33:479–532.

Kumar A, Pearce SR, McLean K, et al (1997) The Ty1-copia group of retrotransposons in plants: genomic organisation, evolution, and use as molecular markers. Genetica 100:205–217. doi: 10.1023/A:1018393931948

Lander E, Linton L, Birren B, et al (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921.

Lashermes P, Combes MC, Robert J, et al (1999) Molecular characterization and origin of the Coffea arabica L. genome. Mol Gen Genet 261:259–266.

Laten HM, Havecker ER, Farmer LM, Voytas DF (2003) SIRE1, an endogenous retrovirus family from Glycine max, is highly homogeneous and evolutionarily young. Mol Biol Evol 20:1222–1230. doi: 10.1093/molbev/msg142

Laten HM, Majumbar A, Gaucher EA (1998) SIRE-1, a copia/Ty1-like retroelement from soybean, encodes a retroviral envelope-like protein. Proc Natl Acad Sci U S A 95:6897–6902.

Lavialle C, Cornelis G, Dupressoir A, et al (2013) Paleovirology of “syncytins”, retroviral env genes exapted for a role in placentation. Philos Trans R Soc London Biol Sci 368:1–10. doi: 10.1098/rstb.2012.0507

Lee J, Waminal NE, Choi H-I, et al (2017) Rapid amplification of four retrotransposon families promoted speciation and genome size expansion in the genus Panax. Sci Rep 7:1–9. doi: 10.1038/s41598-017-08194-5

Lee S, Kim N (2014) Transposable Elements and Genome Size Variations in Plants. Genomics Inform 12:87–97.

Lerat E (2010) Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. Heredity (Edinb) 104:520–533. doi: 10.1038/hdy.2009.165

Lerat E (2001) Comparaison de séquences d’éléments transposables et de gènes d’hôte chez cinq espèces : A. thaliana, C. elegans, D. melanogaster, H. sapiens et S. cerevisiae. Université Claurde Bernard - Lyon I

Lerat E, Fablet M, Modolo L, et al (2016) TEtools facilitates big data expression analysis of transposable elements and reveals an antagonism between their activity and that of piRNA genes. Nucleic Acids Res gkw953. doi: 10.1093/nar/gkw953

Lescot M, Déhais P, Thijs G, et al (2002) PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. Nucleic Acids Res 30:325–327. doi: 10.1093/nar/30.1.325
Levin HL, Moran J V. (2011) Dynamic interactions between transposable elements and their hosts. Nat Rev Genet 12:615–627. doi: 10.1038/nrg3030.Dynamic
Lippman Z, Gendrel A-V, Black M, et al (2004) Role of transposable elements in heterochromatin and epigenetic control. Nature 430:471–476. doi: 10.1038/nature02651
Lisch D (2013a) How important are transposons for plant evolution? Nat Rev Genet 14:49–61. doi: 10.1038/nrg3374
Lisch DR (2013b) Transposons in Plant Gene Regulation. In: Fedoroff N V. (ed) Plant Transposons and Genome Dynamics in Evolution. John Wiley & Sons, Inc., pp 93–116
Llorens C, Fares M a, Moya A (2008) Relationships of gag-pol diversity between Ty3/Gypsy and Retroviridae LTR retroelements and the three kings hypothesis. BMC Evol Biol 8:276. doi: 10.1186/1471-2148-8-276
Llorens C, Futami R, Covelli L, et al (2011a) The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. Nucleic Acids Res 39:D70-4. doi: 10.1093/nar/gkq1061
Llorens C, Futami R, Covelli L, et al (2011b) The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. Nucleic Acids Res 39:D70–D74. doi: 10.1093/nar/gkq1061
Llorens C, Marín I (2001) A Mammalian Gene Evolved from the Integrase Domain of an LTR Retrotransposon. Mol Biol Evol 18:1597–1600.
Llorens C, Muñoz-Pomer A, Bernad L, et al (2009) Network dynamics of eukaryotic LTR retroelements beyond phylogenetic trees. Biol Direct 4:41. doi: 10.1186/1745-6150-4-41
Lopes FR, Carazzolle MF, Pereira G a G, et al (2008) Transposable elements in Coffea (Gentianales: Rubiacea) transcripts and their role in the origin of protein diversity in flowering plants. Mol Genet Genomics 279:385–401. doi: 10.1007/s00438-008-0319-4
Lopes FR, Jjingo D, Da Silva CRM, et al (2013) Transcriptional activity, chromosomal distribution and expression effects of transposable elements in Coffea genomes. PLoS One. doi: 10.1371/journal.pone.0078931
Louarn J (1976) Hybrides interspécifiques entre Coffea canephora Pierre et C. eugenioides Moore. 20:33–52.
Louarn J (1992) La fertilité des hybrides interspécifiques et les relations génomiques entre caféiers diploïdes d'origines africaine (Genre Coffea L. sous-genre Coffea). Université Paris-Sud, centre d'Orsay
Ma J, Bennetzen JL (2004) Rapid recent growth and divergence of rice nuclear genomes. PNAS 101:12404–12410.
Macas J, Neumann P (2007) Ogre elements — A distinct group of plant Ty3/gypsy-like retrotransposons. Gene 390:108–116. doi: 10.1016/j.gene.2006.08.007
Martienssen RA, Chandler VL (2013) Molecular Mechanisms of Transposon Epigenetic Regulation. In: Fedoroff N V. (ed) Plant Transposons and Genome Dynamics in Evolution. John Wiley & Sons, Inc., pp 71–92
Maumus F, Allen EA, Mhiri C et al (2009) Potential impact of stress activated retrotransposons on genome evolution in a marine diatom. BMC Genomics 10:624.
Maurin O, Davis AP, Chester M, et al (2007) Towards a phylogeny for Coffea (Rubiaceae):

Identifying well-supported lineages based on nuclear and plastid DNA sequences. Ann Bot 1–19. doi: 10.1093/aob/mcm257
McCarthy EM, McDonald JF (2003) LTR_STRUC: a novel search and identification program for LTR retrotransposons. Bioinformatics 19:362–367. doi: 10.1093/bioinformatics/btf878
McClintock B (1950) The origin and behavior of mutable loci in maize. PNAS 36:344–355.
Mhiri C, Grandbastien M-A (2004) Éléments transposables et analyse de la biodiversité végétale. In: Morot-Gaudry J, Briat J (eds) La génomique en biologie végétale. INRA, Paris, pp 377–401
Miguel C, Simões M, Oliveira MM, Rocheta M (2008) Envelope-like retrotransposons in the plant kingdom: Evidence of their presence in gymnosperms (Pinus pinaster). J Mol Evol 67:517–525. doi: 10.1007/s00239-008-9168-3
Mirouze M, Reinders J, Bucher E, et al (2009) Selective epigenetic control of retrotransposition in Arabidopsis. Nature 461:1–5. doi: 10.1038/nature08328
Musoli P, Cubry P, Aluka P, et al (2009) Genetic differentiation of wild and cultivated populations: diversity of Coffea canephora Pierre in Uganda. Genome 52:634–646. doi: 10.1139/G09-037
Natali L, Cossu RM, Mascagni F, et al (2015) A survey of Gypsy and Copia LTR-retrotransposon superfamilies and lineages and their distinct dynamics in the Populus trichocarpa (L.) genome. Tree Genet Genomes 11:107–120. doi: 10.1007/s11295-015-0937-z
Nekrutenko A, Li W (2001) Transposable elements are found in a large number of human protein-coding genes. Trends Genet 17:619–621.
Nielen S, Vidigal BS, Leal-Bertioli SCM, et al (2012) Matita, a new retroelement from peanut: characterization and evolutionary context in the light of the Arachis A – B genome divergence. Mol Genet Genomics 287:21–38. doi: 10.1007/s00438-011-0656-6
Nishihara M, Yamada E, Saito M, et al (2014) Molecular characterization of mutations in white-flowered torenia plants. BMC Plant Biol 14:86–98. doi: 10.1186/1471-2229-14-86
Noirot M, Poncet V, Barre P, et al (2003) Genome size variations in diploid African Coffea species. Ann Bot 92:709–714. doi: 10.1093/aob/mcg183
Nowak MD, Davis AP, Yoder AD (2012) Sequence Data from New Plastid and Nuclear COSII Regions Resolves Early Diverging Lineages in Coffea (Rubiaceae). Syst Bot 37:995–1005. doi: 10.1600/036364412X656482
Ohno S (1972) So much "junk" DNA in our genome. Evol Genet Syst 23:366–370.
Ong-Abdullah M, Ordway JM, Jiang N, et al (2015) Loss of Karma transposon methylation underlies the mantled somaclonal variant of oil palm. Nature 525:533–550. doi: 10.1038/nature15365
Orgel LE, Crick FHC (1980) Selfish DNA: the ultimate parasite. Nature 284:604–607.
Pagan HJT, Smith JD, Hubley RM, Ray D a (2010) PiggyBac-ing on a primate genome: novel elements, recent activity and horizontal transfer. Genome Biol Evol 2:293–303. doi: 10.1093/gbe/evq021
Pardue M-L, DeBaryshe PG (2011) Retrotransposons that maintain chromosome ends. Proc Natl Acad Sci U S A 108:20317–20324. doi: 10.1073/pnas.1100278108
Parisod C, Alix K, Just J, 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

Parisod C, Senerchia N (2012) Responses of transposable elements to polyploidy. In: Grandbastien M-A, Casacuberta JM (eds) Plant Transposable Elements. Springer-Verlag, pp 147–168
Paterson AH, Bowers JE, Bruggmann R, et al (2009) The Sorghum bicolor genome and the diversification of grasses. 457:551–556. doi: 10.1038/nature07723
Pearce SR (2007) SIRE-1, a putative plant retrovirus is closely related to a legume TY1-copia retrotransposon family. Cell Mol Biol Lett 12:120–126. doi: 10.2478/s11658-006-0053-z
Petrov DA, Wendel JF (2006) Evolution of eukaryotic genome structure. In: Fox CW, B. WJ (eds) Evolutionary genetics: Concepts and case studies. Oxford University Press,
Piednoël M, Carrete-Vega G, Renner SS (2013) Characterization of the LTR retrotransposon repertoire of a plant clade of six diploid and one tetraploid species. Plant J 75:699–709. doi: 10.1111/tpj.12233
Piégu B, Bire S, Arensburger P, Bigot Y (2015) A survey of transposable element classification systems - A call for a fundamental update to meet the challenge of their diversity and complexity. Mol Phylogenet Evol 86:90–109. doi: 10.1016/j.ympev.2015.03.009
Piégu B, Guyot R, Picault N, 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.of
Piffanelli P, Droc G, Mieulet D, et al (2007) Large-scale characterization of Tos17 insertion sites in a rice T-DNA mutant library. Plant Mol Biol 65:587–601. doi: 10.1007/s11103-007-9222-3
Price AL, Jones NC, Pevzner PA (2005) De novo identification of repeat families in large genomes. Bioinformatics 21:351–358. doi: 10.1093/bioinformatics/bti1018
Quesneville H, Nouaud D, Anxolabéhère D (2002) Detection of new transposable element families in Drosophila melanogaster and Anopheles gambiae genomes. J Mol Evol 57:50–59. doi: 10.1007/s00239-003-0007-2
Ramamurthy RK, Waters BM (2017) Mapping and characterization of the fefe gene that controls iron uptake in Melon (Cucumis melo L.). Front Plant Sci 8:1–13. doi: 10.3389/fpls.2017.01003
Razafinarivo NJ, Rakotomalala JJ, Brown SC, et al (2012) Geographical gradients in the genome size variation of wild coffee trees (Coffea) native to Africa and Indian Ocean islands. Tree Genet Genomes 8:1345–1358. doi: 10.1007/s11295-012-0520-9
Rhoades MM (1938) Effect of the Dt gene on the mutability of the a1 allele in maize. Genetics 23:377–397.
Rhoades MM (1936) The effect of varying gene dosage on aleurone colour in maize. J Genet 23:347–354.
Rhoads A, Au KF (2015) PacBio Sequencing and Its Applications. Genomics, Proteomics Bioinforma 13:278–289. doi: 10.1016/j.gpb.2015.08.002
Robbrecht E, Manen J-F (2006) The major evolutionary lineages of the coffee family (Rubiaceae, angiosperms). Combined analysis (nDNA and cpDNA) to infer the position of Coptosapelta and Luculia, and supertree construction based on rbcL, rps16, trnL-trnF and atpB-rbcL data. A new class. Syst Geogr Plants 76:85–146.
Rockinger A, Sousa A, Carvalho FA, Renner SS (2016) Chromosome number reduction in the sister clade of carica papaya with concomitant genome size doubling. Am J Bot 103:1082–1088. doi: 10.3732/ajb.1600134
Roncal J, Guyot R, Hamon P, et al (2015) Active transposable elements recover species boundaries and geographic structure in Madagascan coffee species. Mol Genet

Genomics 291:155–168. doi: 10.1007/s00438-015-1098-3
Rothberg JM, Leamon JH (2008) The development and impact of 454 sequencing. Nat Biotechnol 26:1117–1124. doi: 10.1038/nbt1485
Rutherford K, Parkhill J, Crook J, et al (2000) Artemis : sequence visualization and annotation. Bioinformatics 16:944–945.
SanMiguel P, Gaut BS, Tikhonov A, et al (1998) The paleontology of intergene retrotransposons of maize. Nat Genet 20:43–45. doi: 10.1038/1695
Schaack S, Gilbert C, Feschotte C (2010) Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. Trends Ecol Evol 25:537–546.
Schnable P, Ware D, Fulton R, et al (2009) The B73 Maize Genome: Complexity, Diversity, and Dynamics. Science (80-) 326:1112–1115.
Schulman AH (2012) Hitching a ride: nonautonomous retrotransposons and parasitism as a lifestyle. In: Grandbastien M-A, Casacuberta JM (eds) Plant Transposable Elements. Springer-Verlag, pp 71–88
Senerchia N, Felber F, Parisod C (2014) Contrasting evolutionary trajectories of multiple retrotransposons following independent allopolyploidy in wild wheats. New Phytol 202:975–985. doi: 10.1111/nph.12731
Smit AFA, Hubley R, Green P RepeatMasker.
Sonnhammer ELL, Durbin R (1996) A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. Gene 167:1–10.
Sultana T, Zamborlini A, Cristofari G, Lesage P (2017) Integration site selection by retroviruses and transposable elements in eukaryotes. Nat Rev Genet 18:292–308. doi: 10.1038/nrg.2017.7
Talhinhas P, Batista D, Diniz I, et al (2017) Pathogen profile The coffee leaf rust pathogen Hemileia vastatrix : one and a half centuries around the tropics. Mol Plant Pathol 18:1039–1051. doi: 10.1111/mpp.12512
Terzian C, Ferraz C, Demaille J, Bucheton A (2000) Evolution of the Gypsy Endogenous Retrovirus in the Drosophila melanogaster Subgroup. Mol Biol Evol 17:908–914.
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, pp 42–45
The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408:796–815.
This P, Lacombe T, Cadle-Davidson M, Owens CL (2007) Wine grape (Vitis vinifera L .) color associates with allelic variation in the domestication gene VvmybA1. Theor Appl Genet 114:723–730. doi: 10.1007/s00122-006-0472-2
Thompson JD, Higgins DG, Gibson TJ (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.
Vicient CM, Casacuberta JM (2017) Impact of transposable elements on polyploid plant genomes. Ann Bot 120:195–207. doi: 10.1093/aob/mcx078
Vitte C, Panaud O (2005) LTR retrotransposons and flowering plant genome size: emergence of the increase/decrease model. Cytogenet Genome Res 110:91–107. doi: 10.1159/000084941
Volff J-N (2006) Turning junk into gold: Domestication of transposable elements and the creation of new genes in eukaryotes. BioEssays 28:913–922. doi: 10.1002/bies.20452
Volfovsky N, Haas BJ, Salzberg SL (2001) A clustering method for repeat analysis in DNA

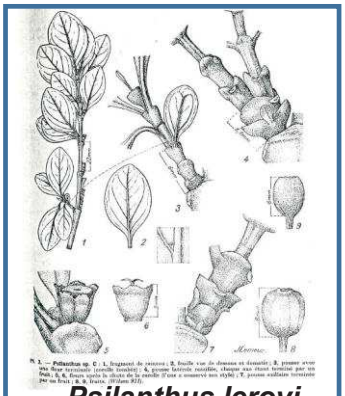
sequences. Genome Biol 2:1–11. doi: 10.1186/gb-2001-2-8-research0027
Wallau GL, Ortiz MF, Loreto E (2012) Horizontal Transposon Transfer in Eukarya: Detection, Bias, and Perspectives. Genome Biol Evol 4:801–811.
Weber B, Wenke T, Fr??mmel U, et al (2010) The Ty1-copia families SALIRE and Cotzilla populating the Beta vulgaris genome show remarkable differences in abundance, chromosomal distribution, and age. Chromosom Res 18:247–263. doi: 10.1007/s10577-009-9104-4
White SE, Habera LF, Wessler SR (1994) Retrotransposons in the flanking regions of normal plant genes: A role for copia-like elements in the evolution of gene structure and expression. Proc Natl Acad Sci U S A 91:11792–11796.
Wicker T (2012) So many repeats and so little time: how to classify transposable elements. In: Grandbastien M-A, Casacuberta JM (eds) Plant Transposable Elements. Springer-Verlag, Zurich, pp 1–15
Wicker T, Keller B (2007) Genome-wide comparative analysis of copia retrotransposons in Triticeae, rice, and Arabidopsis reveals conserved ancient evolutionary lineages and distinct dynamics of individual copia families. Genome Res 17:1072–1081. doi: 10.1101/gr.6214107.Because
Wicker T, Sabot F, Hua-Van A, et al (2007) A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8:973–82. doi: 10.1038/nrg2165
Wicker T, Stein N, Albar L, et al (2001) Analysis of a contiguous 211 kb sequence in diploid wheat (Triticum monococcum L.) reveals multiple mechanisms of genome evolution. Plant J 26:307–316.
Xiong Y, Eickbush TH (1990) Origin and evolution of retroelements based upon their reverse transcriptase sequences. EMBO J 9:3353–3362.
Xu Z, Wang H (2007) LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res 35:W265–W268. doi: 10.1093/nar/gkm286
Yin H, Du J, Wu J, et al (2015) Genome-wide Annotation and Comparative Analysis of Long Terminal Repeat Retrotransposons between Pear Species of P. bretschneideri and P. Communis. Sci Rep 5:1–15. doi: 10.1038/srep17644
Yu Q, Guyot R, de Kochko A, et al (2011) Micro-collinearity and genome evolution in the vicinity of an ethylene receptor gene of cultivated diploid and allotetraploid coffee species (Coffea). Plant J 67:305–317. doi: 10.1111/j.1365-313X.2011.04590.x
Yuyama PM, Protasio Pereira LF, Benedito dos Santos T, et al (2012) FISH using a gag-like fragment probe reveals a common Ty3-gypsy-like retrotransposon in genome of Coffea species. Genome 55:825–833. doi: 10.1139/gen-2012-0081
Zeh DW, Zeh JA, Ishida Y (2009) Transposable elements and an epigenetic basis for punctuated equilibria. BioEssays 31:715–726. doi: 10.1002/bies.200900026
Zuccolo A, Sebastian A, Talag J, et al (2007) Transposable element distribution, abundance and role in genome size variation in the genus Oryza. BMC Evol Biol 7:1–15. doi: 10.1186/1471-2148-7-152

Annexe 1 – Figure annexe 1

Photographies et/ou planches d'herbiers des espèces de caféiers de cette étude.

Psilanthus ebracteolatus

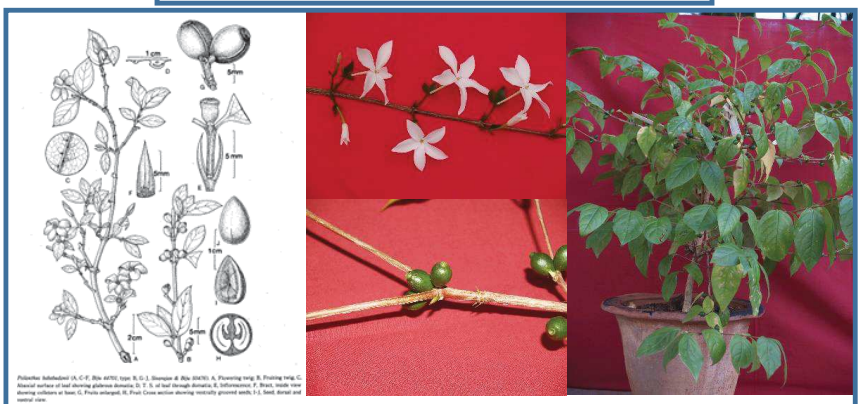
Psilanthus mannii (photo fruit: O. Maurin)



Psilanthus leroyi

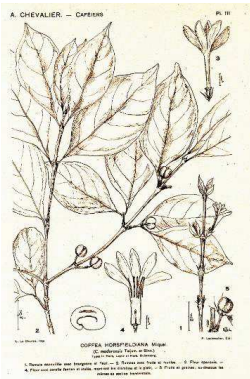
Psilanthus melanocarpus

Psilanthus benghalensis



Psilanthus benghalensis var bababudanii

Psilanthus merguensis

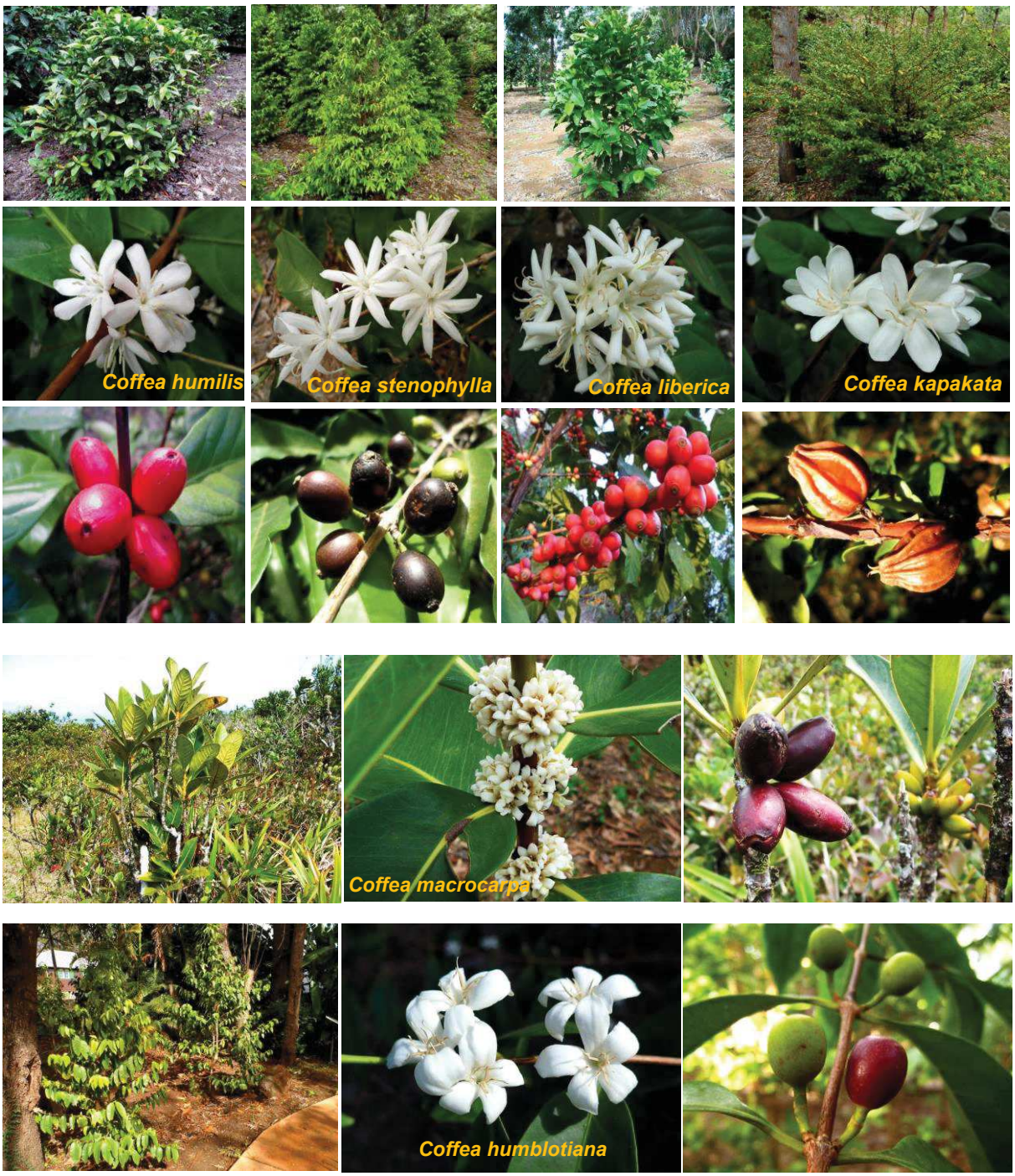


Psilanthus horsfieldianus

Psilanthus brassii

Coffea arabica
Coffea canephora
Coffea eugenioides

Coffea charrieriana
Coffea pseudozanguebariae
Coffea racemosa


Coffea humilis
Coffea stenophylla
Coffea liberica
Coffea kapakata
Coffea macrocarpa
Coffea humblotiana

Coffea tetragona

Coffea dolichophylla

Annexe 2 – Figure annexe 2

Copie des résultats de PlantCARE pour les motifs dans l'un des deux LTR de *C. canephora* pour les trois séquences de référence pour chaque famille de *SIRE*. En premier le LTR court de la séquence de référence courte de la famille C. Suivant : séquence du LTR long de la famille C, puis séquence du LTR de la famille A et enfin séquence de référence de la famille B.

```
>L90_710 332nt
+ GTTATGAAAT AACAAAAATG TGGATGTCCC TTTGTATTCC CAAGGAGATT AATTCATGAT TTTATGGCTA
- CAATACTTTA TTGTTTTTAC ACCTACAGGG AAACATAAGG GTTCCTCTAA TTAAGTACTA AAATACCGAT

+ CTTATGTATA GAAGTTACAA TTCATAAAAT CCATTCATGT AGTGGAGAAA CCGTTTCATA GGTTTCTTCA
- GAATACATAT CTTCAATGTT AAGTATTTTA GGTAAGTACA TCACCTCTTT GGCAAAGTAT CCAAAGAAGT

+ TATTATTGTA GTAATGGATG TTTAATAAGA TTTATGTACC TTTAATCTTG TAGTGCCTAT ATATAGAGGT
- ATAATAACAT CATTACCTAC AAATTATTCT AAATACATGG AAATTAGAAC ATCACGGATA TATATCTCCA

+ ACATTGAGAT CCTTTGGGAG TACTTTTGTA TCTTGTTGGA ATACAAGAAT ATCATTCTCC ATTGCTCTAC
- TGTAACTCTA GGAAACCCTC ATGAAAACAT AGAACAACCT TATGTTCTTA TAGTAAGAGG TAACGAGATG

+ TTTTCCTCTA ATATCTCTAT TCATATTATA GTAGCTTATT AGTTTTATAA C
- AAAAGGAGAT TATAGAGATA AGTATAATAT CATCGAATAA TCAAAATATT G
```

CAAT-box

Site Name	Organism	Position	Strand	Matrix score.	sequence	function
CAAT-box	Glycine max	88	+	5	CAATT	common cis-acting element in promoter and enhancer regions
CAAT-box	Hordeum vulgare	213	-	4	CAAT	common cis-acting element in promoter and enhancer regions
CAAT-box	Hordeum vulgare	145	-	4	CAAT	common cis-acting element in promoter and enhancer regions
CAAT-box	Hordeum vulgare	271	-	4	CAAT	common cis-acting element in promoter and enhancer regions

TATA-box

Site Name	Organism	Position	Strand	Matrix score.	sequence	function
TATA-box	Lycopersicon esculentum	60	+	5	TTTTA	core promoter element around -30 of transcription start
TATA-box	Nicotiana tabacum	73	-	9	tcTATAAAta	core promoter element around -30 of transcription start
TATA-box	Helianthus annuus	75	-	6	TATACA	core promoter element around -30 of transcription start
TATA-box	Arabidopsis thaliana	77	+	4	TATA	core promoter element around -30 of transcription start
TATA-box	Lycopersicon esculentum	95	-	5	TTTTA	core promoter element around -30 of transcription start
TATA-box	Glycine max	141	-	5	TAATA	core promoter element around -30 of transcription start
TATA-box	Glycine max	163	+	5	TAATA	core promoter element around -30 of transcription start
TATA-box	Arabidopsis thaliana	169	-	9	taTATAAAtc	core promoter element around -30 of transcription start
TATA-box	Oryza sativa	171	-	8	TACATAAA	core promoter element around -30 of transcription start
TATA-box	Arabidopsis thaliana	196	-	9	taTATAAAgg	core promoter element around -30 of transcription start
TATA-box	Arabidopsis thaliana	198	-	8	TATATATA	core promoter element around -30 of transcription start
TATA-box	Brassica napus	199	-	6	ATATAT	core promoter element around -30 of transcription start
TATA-box	Arabidopsis thaliana	200	-	4	TATA	core promoter element around -30 of transcription start
TATA-box	Arabidopsis thaliana	202	-	4	TATA	core promoter element around -30 of transcription start
TATA-box	Oryza sativa	234	-	7	TACAAAA	core promoter element around -30 of transcription start
TATA-box	Glycine max	289	+	5	TAATA	core promoter element around -30 of transcription start
TATA-box	Glycine max	304	-	5	TAATA	core promoter element around -30 of transcription start
TATA-box	Brassica napus	305	+	6	ATTATA	core promoter element around -30 of transcription start
TATA-box	Arabidopsis thaliana	306	-	5	TATAA	core promoter element around -30 of transcription start
TATA-box	Arabidopsis thaliana	307	-	4	TATA	core promoter element around -30 of transcription start
TATA-box	Glycine max	317	-	5	TAATA	core promoter element around -30 of transcription start
TATA-box	Arabidopsis thaliana	323	-	7	TATAAAA	core promoter element around -30 of transcription start
TATA-box	Arabidopsis thaliana	324	-	6	TATAAA	core promoter element around -30 of transcription start
TATA-box	Arabidopsis thaliana	325	-	5	TATAA	core promoter element around -30 of transcription start
TATA-box	Arabidopsis thaliana	326	-	4	TATA	core promoter element around -30 of transcription start

circadian

Site Name	Organism	Position	Strand	Matrix score.	sequence	function
circadian	Lycopersicon esculentum	254	+	6	CAANNNNATC	cis-acting regulatory element involved in circadian control

\>L89_1374 1049nt

```
+ GTTGACCTTG GTTGATCAGT CCTTGGTTTT GATGATTAAC AAACCAAAAT GTAGATTTGG ACTAATGTTT
- CAACTGGAAC CAACTAGTCA GGAACCAAAA CTACTAATTG TTTGGTTTTA CATCTAAACC TGATTACAAA

+ TTATGTGAGA AATTTGTTAG AACAGGTCAT TGATCTAAAA GAGAATCGAA AAGATTTGAA TCAAAGAGAA
- AATACACTCT TTAAACAATC TTGTCCAGTA ACTAGATTTT CTCTTAGCTT TTCTAAACTT AGTTTCTCTT

+ GTTAAGAATT TTGAAGGTTC GGACGTAGAG GATCGGACGC TCGATAGAGG ATCTGACGTC CGACAAACGA
- CAATTCTTAA AACTTCCAAG CCTGCATCTC CTAGCCTGCG AGCTATCTCC TAGACTGCAG GCTGTTTGCT

+ CGATACACTT CAGACGCTTG GATTGGACGC TCATTCATTG CTTCGGCCGT CCGACACAAA AAGAATGAAA
- GCTATGTGAA GTCTGCGAAC CTAACCTGCG AGTAAGTAAC GAAGCCGGCA GGCTGTGTTT TTCTTACTTT

+ GTTGAACATT CTGACTTCTA TCGGACGCAG GGTTCGGACG CAGAACAAAT GGTTCGGACG CCCGATAGGT
- CAACTTGTAA GACTGAAGAT AGCCTGCGTC CCAAGCCTGC GTCTTGTTTA CCAAGCCTGC GGGCTATCCA

+ GGTTCGGACG CATGGTTCGG ACGCAGAGCA AATGGTTCGG AGGTCCGACA GGCCAACGGC TAGTTTCGAC
- CCAAGCCTGC GTACCAAGCC TGCGTCTCGT TTACCAAGCC TCCAGGCTGT CCGGTTGCCG ATCAAAGCTG

+ AGCATTTAAT ATTTGACCGT TGGAGAGTCT TTTGGAGCCA TTTCTCACCT TCTATAAATA CCCCAAAGCA
- TCGTAAATTA TAAACTGGCA ACCTCTCAGA AAACCTCGGT AAAGAGTGGA AGATATTTAT GGGGTTTCGT

+ CAAGAAGACA TAGACTTTTT GCCAACACTA TATACAAGCT TACAAGTGAG ATCTTTGAGT AGAAATATTC
- GTTCTTCTGT ATCTGAAAAA CGGTTGTGAT ATATGTTCGA ATGTTCACTC TAGAAACTCA TCTTTATAAG

+ TTTGTTGGTT GTATAAGGGT TGGGAAAGTT GGGTTGTGAG GTTGCTCAAG TGAAGGTTAC TCTCTAGGAG
- AAACAACCAA CATATTCCCA ACCCTTTCAA CCCAACACTC CAACGAGTTC ACTTCCAATG AGAGATCCTC

+ GAGTAAAACC TTGGTGAAGG TGAACCTATC ACTTGGAGTG GTAAAACCTT GATAAAGGTT GCCTCACTTG
- CTCATTTTGG AACCACTTCC ACTTGGATAG TGAACCTCAC CATTTTGGAA CTATTTCCAA CGGAGTGAAC

+ TATAAAGTAG TTCTTAATTG AGTGGGTAAG CTTTCAATTG TAACTAGTTG AGGGTTTACT TGGAGAGTAG
- ATATTTCATC AAGAATTAAC TCACCCATTC GAAAGTTAAC ATTGATCAAC TCCCAAATGA ACCTCTCATC

+ TAAAATTCCT TTACTCAACC AAAGTATGTT TGGGGTGAGG AAGGAGTGAG CCTTCACTTG TTCAAGTTGG
- ATTTTAAGGA AATGAGTTGG TTTCATACAA ACCCCACTCC TTCCTCACTC GGAAGTGAAC AAGTTCAACC

+ TTAGCACTAC CATCAATTGA AGAAACTATT TGGTGGATTC AACTCCAAGT TTGGAAGAAG CTTGGGAGTT
- AATCGTGATG GTAGTTAACT TCTTTGATAA ACCACCTAAG TTGAGGTTCA AACCTTCTTC GAACCCTCAA

+ TGATTGGTTT GAATTTCTTT TACTTTATTG TTACTCTATT TTTGTACTTT AAAATTGCTT ATATTCTTTG
- ACTAACCAAA CTTAAAGAAA ATGAAATAAC AATGAGATAA AAACATGAAA TTTTAACGAA TATAAGAAAC

+ TCATTGTATT TACTTGACTA CTTGTCTGGT TGAGTATTTT GGTTGTATTT GGTTGCATCG GGGCTTTC
- AGTAACATAA ATGAACTGAT GAACAGACCA ACTCATAAAA CCAACATAAA CCAACGTAGC CCCGAAAG
```

ARE

Site Name	Organism	Position	Strand	Matrix score.	sequence	function
ARE	Zea mays	24	+	6	TGGTTT	cis-acting regulatory element essential for the anaerobic induction
ARE	Zea mays	915	+	6	TGGTTT	cis-acting regulatory element essential for the anaerobic induction
ARE	Zea mays	41	-	6	TGGTTT	cis-acting regulatory element essential for the anaerobic induction

Box-W1

Site Name	Organism	Position	Strand	Matrix score.	sequence	function
Box-W1	Petroselinum crispum	2	+	6	TTGACC	fungal elicitor responsive element
Box-W1	Petroselinum crispum	433	+	6	TTGACC	fungal elicitor responsive element

C-box

Site Name	Organism	Position	Strand	Matrix score.	sequence	function
C-box	Arabidopsis thaliana	193	+	9	CTGACGTCAG	cis-acting regulatory element involved in light responsiveness

CAAT-box

Site Name	Organism	Position	Strand	Matrix score.	sequence	function
CAAT-box	Brassica rapa	55	-	5	CAAAT	common cis-acting element in promoter and enhancer regions
CAAT-box	Brassica rapa	82	-	5	CAAAT	common cis-acting element in promoter and enhancer regions
CAAT-box	Hordeum vulgare	99	-	4	CAAT	common cis-acting element in promoter and enhancer regions
CAAT-box	Brassica rapa	124	-	5	CAAAT	common cis-acting element in promoter and enhancer regions
CAAT-box	Arabidopsis thaliana	232	-	5	CCAAT	common cis-acting element in promoter and enhancer regions
CAAT-box	Hordeum vulgare	247	-	4	CAAT	common cis-acting element in promoter and enhancer regions
CAAT-box	Brassica rapa	326	+	5	CAAAT	common cis-acting element in promoter and enhancer regions
CAAT-box	Brassica rapa	379	+	5	CAAAT	common cis-acting element in promoter and enhancer regions
CAAT-box	Brassica rapa	431	-	5	CAAAT	common cis-acting element in promoter and enhancer regions
CAAT-box	Petunia hybrida	519	+	7	TGCCAAC	common cis-acting element in promoter and enhancer regions
CAAT-box	Glycine max	716	-	5	CAATT	common cis-acting element in promoter and enhancer regions
CAAT-box	Hordeum vulgare	717	-	4	CAAT	common cis-acting element in promoter and enhancer regions
CAAT-box	Glycine max	735	+	5	CAATT	common cis-acting element in promoter and enhancer regions
CAAT-box	Glycine max	736	-	5	CAATT	common cis-acting element in promoter and enhancer regions
CAAT-box	Hordeum vulgare	737	-	4	CAAT	common cis-acting element in promoter and enhancer regions
CAAT-box	Glycine max	854	+	5	CAATT	common cis-acting element in promoter and enhancer regions
CAAT-box	Glycine max	855	-	5	CAATT	common cis-acting element in promoter and enhancer regions
CAAT-box	Hordeum vulgare	856	-	4	CAAT	common cis-acting element in promoter and enhancer regions
CAAT-box	Brassica rapa	868	-	5	CAAAT	common cis-acting element in promoter and enhancer regions
CAAT-box	Arabidopsis thaliana	913	-	5	CCAAT	common cis-acting element in promoter and enhancer regions
CAAT-box	Hordeum vulgare	937	-	4	CAAT	common cis-acting element in promoter and enhancer regions
CAAT-box	Glycine max	963	-	5	CAATT	common cis-acting element in promoter and enhancer regions
CAAT-box	Hordeum vulgare	984	-	4	CAAT	common cis-acting element in promoter and enhancer regions
CAAT-box	Hordeum vulgare	983	-	4	CAAT	common cis-acting element in promoter and enhancer regions
CAAT-box	Brassica rapa	1027	-	5	CAAAT	common cis-acting element in promoter and enhancer regions

CCGTCC-box

Site Name	Organism	Position	Strand	Matrix score.	sequence	function
CCGTCC-box	Arabidopsis thaliana	257	+	6	CCGTCC	cis-acting regulatory element related to meristem specific activation

CGTCA-motif

Site Name	Organism	Position	Strand	Matrix score.	sequence	function
CGTCA-motif	Hordeum vulgare	194	-	5	CGTCA	cis-acting regulatory element involved in the MeJA-responsiveness

G-box

Site Name	Organism	Position	Strand	Matrix score.	sequence	function
G-box	Larix laricina	997	-	9	TACACGTAGT	cis-acting regulatory element involved in light responsiveness

HSE

Site Name	Organism	Position	Strand	Matrix score.	sequence	function
HSE	Brassica oleracea	79	-	9	AAAAAATTTC	cis-acting element involved in heat stress responsiveness

MBS

Site Name	Organism	Position	Strand	Matrix score.	sequence	function
MBS	Zea mays	434	-	6	CGGTCA	MYB Binding Site

MSA-like

Site Name	Organism	Position	Strand	Matrix score.	sequence	function
MSA-like	Catharanthus roseus	452	+	8.5	(T/C)C(T/C)AACGG(T/C)(T/C)A	cis-acting element involved in cell cycle regulation
MSA-like	Catharanthus roseus	436	-	9	TCCAACGGT	cis-acting element involved in cell cycle regulation
MSA-like	Catharanthus roseus	434	-	9	(T/C)C(T/C)AACGG(T/C)(T/C)A	cis-acting element involved in cell cycle regulation

Skn-1_motif

Site Name	Organism	Position	Strand	Matrix score.	sequence	function
Skn-1_motif	Oryza sativa	96	+	5	GTCAT	cis-acting regulatory element required for endosperm expression
Skn-1_motif	Oryza sativa	980	+	5	GTCAT	cis-acting regulatory element required for endosperm expression

TATA-box

Site Name	Organism	Position	Strand	Matrix score.	sequence	function
TATA-box	Lycopersicon esculentum	69	+	5	TTTTA	core promoter element around -30 of transcription start
TATA-box	Lycopersicon esculentum	186	-	5	TTTTA	core promoter element around -30 of transcription start
TATA-box	Glycine max	437	+	5	TAATA	core promoter element around -30 of transcription start
TATA-box	Nicotiana tabacum	471	+	10	taTATAAAta	core promoter element around -30 of transcription start
TATA-box	Ac	473	+	7	TATAAAT	core promoter element around -30 of transcription start
TATA-box	Arabidopsis thaliana	519	+	4	TATA	core promoter element around -30 of transcription start
TATA-box	Arabidopsis thaliana	521	+	4	TATA	core promoter element around -30 of transcription start
TATA-box	Helianthus annuus	570	-	6	TATACA	core promoter element around -30 of transcription start
TATA-box	Arabidopsis thaliana	572	-	4	TATA	core promoter element around -30 of transcription start
TATA-box	Lycopersicon esculentum	634	-	5	TTTTA	core promoter element around -30 of transcription start
TATA-box	Lycopersicon esculentum	672	-	5	TTTTA	core promoter element around -30 of transcription start
TATA-box	Helianthus annuus	699	-	8	TATACA	core promoter element around -30 of transcription start
TATA-box	Arabidopsis thaliana	781	+	6	TATAAA	core promoter element around -30 of transcription start
TATA-box	Lycopersicon esculentum	771	-	5	TTTTA	core promoter element around -30 of transcription start
TATA-box	Lycopersicon esculentum	828	+	5	TTTTA	core promoter element around -30 of transcription start
TATA-box	Oryza sativa	950	-	7	TACAAAA	core promoter element around -30 of transcription start
TATA-box	Lycopersicon esculentum	960	-	5	TTTTA	core promoter element around -30 of transcription start
TATA-box	Arabidopsis thaliana	969	-	5	TATAA	core promoter element around -30 of transcription start
TATA-box	Arabidopsis thaliana	970	-	4	TATA	core promoter element around -30 of transcription start

TC-rich repeats

Site Name	Organism	Position	Strand	Matrix score.	sequence	function
TC-rich repeats	Nicotiana tabacum	858	-	9	ATTTTCTTCA	cis-acting element involved in defense and stress responsiveness

circadian

Site Name	Organism	Position	Strand	Matrix score.	sequence	function
circadian	Lycopersicon esculentum	538	-	9	CAAAGATATC	cis-acting regulatory element involved in circadian control
circadian	Lycopersicon esculentum	912	-	9	CAANNNNATC	cis-acting regulatory element involved in circadian control

L363_2468

```
+ GTTATTCTTG GTTTTGATGA TCACAAAAGT TTGTTTATAC AGACCAAGAA AGTGAACACT TTGATCAGGT
- CAATAAGAAC CAAAACTACT AGTGTTTTCA AACAAATATG TCTGGTTCTT TCACTTGTGA AACTAGTCCA

+ CTGTTGGAAC AAAGAAAGCA TATGTTCGTT TATCTCCTGA CTACGTTGCT TTGGATGTGG CAAACAAGAT
- GACAACCTTG TTTCTTTCGT ATACAAGCAA ATAGAGGACT GATGCAACGA AACCTACACC GTTTGTTCTA

+ TGGAATAATA TGAATGAATT TAGTCATTGT TTTGTTTCGA TCAAAGATAT TGTCTTTTAA GTATGTCTAG
- ACCTTATTAT ACTTACTTAA ATCAGTAACA AAACAAAGCT AGTTTCTATA ACAGAAAATT CATACAGATC

+ TTTGTCATTC TTGGTACTTT GAAATATTTG TCTAACCTTC TTCAAATACA AGTGTTTTTG TCTATCCAAG
- AAACAGTAAG AACCATGAAA CTTTATAAAC AGATTGGAAG AAGTTTATGT TCACAAAAAC AGATAGGTTC

+ AATCAGGTTC AAATATGTCA TGAAATGCAA AACAACCTAA ATGAGAAGCA AAAACAGGAC AATCAGGTCG
- TTAGTCCAAG TTTATACAGT ACTTTACGTT TTGTTGGATT TACTCTTCGT TTTTGTCCTG TTAGTCCAGC

+ GACGGCCGAA AGGAAACTGT CGGACGTCCG AAAAGATAAA GTACATCAGG AAGGAAGCAC CGTCGGACGC
- CTGCCGGCTT TCCTTTGACA GCCTGCAGGC TTTTCTATTT CATGTAGTCC TTCCTTCGTG GCAGCCTGCG

+ TCACATGGAA GCATCGGACG TCCGGAAGGA TCGGACGCTT GCCATCGGAC GCTCATCAAC CGCATCGGAC
- AGTGTACCTT CGTAGCCTGC AGGCCTTCCT AGCCTGCGAA CGGTAGCCTG CGAGTAGTTG GCGTAGCCTG

+ GTCCGAAAAA TTCCAAGATA ACTTGCTAGC TCTCGGCCCA AGGTCGGACG CTGTGCTGTC GGACGACAAA
- CAGGCTTTTT AAGGTTCTAT TGAACGATCG AGAGCCGGGT TCCAGCCTGC GACACGACAG CCTGCTGTTT

+ AAACTCATCG GACGTCCGAA CCTCAACTTG GCTATCATCG GACGATTGGT TCGGACGATG ATTCGATCGT
- TTTGAGTAGC CTGCAGGCTT GGAGTTGAAC CGATAGTAGC CTGCTAACCA AGCCTGCTAC TAAGCTAGCA

+ CGGACGTCCG ACAGCCCCAA CGGCTAGTTG ACTCTTCAGC TGCCTTCTAT CCGTTGGAAG CATTGATAAA
- GCCTGCAGGC TGTCGGGGTT GCCGATCAAC TGAGAAGTCG ACGGAAGATA GGCAACCTTC GTAACTATTT

+ GCCCATTTCT GGATCCCTTT AAAATCAAAC TGGTCTGAAC CAAGAAGGGA CTTTTGCACA CTTTGTTTAC
- CGGGTAAAGA CCTAGGGAAA TTTTAGTTTG ACCAGACTTG GTTCTTCCCT GAAAACGTGT GAAACAAATG

+ AAGTTCTCAA GAGATATTTT AGCTAGAAAA TAGTCTCCAA AGCAGATTTG TTTTAAAGTG GTGTGAATTT
- TTCAAGAGTT CTCTATAAAA TCGATCTTTT ATCAGAGGTT TCGTCTAAAC AAAATTTCAC CACACTTAAA

+ CTGTTGAGCA TTTCTTTTGT GGTTGAAAGG ATCTTTAGTG TAGCTTTGTT GAGGGTTTTC TGAGTGATTG
- GACAACTCGT AAAGAAAACA CCAACTTTCC TAGAAATCAC ATCGAAACAA CTCCCAAAAG ACTCACTAAC

+ TAAAACTTCT TGGCTTGACT AAGTGAGGCT TAGGGCAAGA AGGAAGTGCT CCCTCCATTG TACATCTAGT
- ATTTTGAAGA ACCGAACTGA TTCACTCCGA ATCCCGTTCT TCCTTCACGA GGGAGGTAAC ATGTAGATCA

+ TGATCTTCTT TCATCAAAGA GAAGTTGCTT TTCCTAGTGT TTGGTCTTCA AGTTTGAGGA TAGCTTGGAA
- ACTAGAAGAA AGTAGTTTCT CTTCAACGAA AAGGATCACA AACCAGAAGT TCAAACTCCT ATCGAACCTT

+ GACACTTGGT TTGATTCCCT ATTTTACTTT TCTGTTTAAT TAAAATTCTC ATTGCTTATC TATACTTGTT
- CTGTGAACCA AACTAAGGGA TAAAATGAAA AGACAAATTA ATTTTAAGAG TAACGAATAG ATATGAACAA

+ TTTCTGATCA ATATTGCTCC TATCTTCTAA TTTACTTAGT TGATCATTAC TAGAAAAGAA GGTAAATTTT
- AAAGACTAGT TATAACGAGG ATAGAAGATT AAATGAATCA ACTAGTAATG ATCTTTTCTT CCATTTAAAA

+ TATTAAAGAA AAAGTGCATA AATTTGGTTA AGATTTTTAT CAACCCAATT CACCCCCCCT CTTGGTTGTC
- ATAATTTCTT TTTCACGTAT TTAAACCAAT TCTAAAATTA GTTGGGTTAA GTGGGGGGGA GAACCAACAG

+ TTTGGGACTT AC
- AAACCCTGAA TG
```

3-AF1 binding site

Site Name	Organism	Position	Strand	Matrix score.	sequence	function
3-AF1 binding site	Solanum tuberosum	779	+	11	AAGAGATATTT	light responsive element

ACE

Site Name	Organism	Position	Strand	Matrix score.	sequence	function
ACE	Petroselinum crispum	110	+	9	ACTACGTTGG	cis-acting element involved in light responsiveness

ARE

Site Name	Organism	Position	Strand	Matrix score.	sequence	function
ARE	Zea mays	9	+	6	TGGTTT	cis-acting regulatory element essential for the anaerobic induction
ARE	Zea mays	1057	+	6	TGGTTT	cis-acting regulatory element essential for the anaerobic induction

Box I

Site Name	Organism	Position	Strand	Matrix score.	sequence	function
Box I	Pisum sativum	228	-	7	TTTCAAA	light responsive element

CAAT-box

Site Name	Organism	Position	Strand	Matrix score.	sequence	function
CAAT-box	Arabidopsis thaliana	139	-	5	CCAAT	common cis-acting element in promoter and enhancer regions
CAAT-box	Hordeum vulgare	166	-	4	CAAT	common cis-acting element in promoter and enhancer regions
CAAT-box	Hordeum vulgare	189	-	4	CAAT	common cis-acting element in promoter and enhancer regions
CAAT-box	Brassica rapa	236	-	5	CAAAT	common cis-acting element in promoter and enhancer regions
CAAT-box	Brassica rapa	253	+	5	CAAAT	common cis-acting element in promoter and enhancer regions
CAAT-box	Brassica rapa	290	+	5	CAAAT	common cis-acting element in promoter and enhancer regions
CAAT-box	Hordeum vulgare	340	+	4	CAAT	common cis-acting element in promoter and enhancer regions
CAAT-box	Arabidopsis thaliana	605	-	5	CCAAT	common cis-acting element in promoter and enhancer regions
CAAT-box	Hordeum vulgare	692	-	4	CAAT	common cis-acting element in promoter and enhancer regions
CAAT-box	Brassica rapa	816	-	5	CAAAT	common cis-acting element in promoter and enhancer regions
CAAT-box	Hordeum vulgare	907	-	4	CAAT	common cis-acting element in promoter and enhancer regions
CAAT-box	Hordeum vulgare	967	-	4	CAAT	common cis-acting element in promoter and enhancer regions
CAAT-box	Hordeum vulgare	1101	-	4	CAAT	common cis-acting element in promoter and enhancer regions
CAAT-box	Hordeum vulgare	1129	+	4	CAAT	common cis-acting element in promoter and enhancer regions
CAAT-box	Hordeum vulgare	1133	-	4	CAAT	common cis-acting element in promoter and enhancer regions
CAAT-box	Brassica rapa	1212	-	5	CAAAT	common cis-acting element in promoter and enhancer regions
CAAT-box	Arabidopsis thaliana	1235	+	5	CCAAT	common cis-acting element in promoter and enhancer regions
CAAT-box	Glycine max	1236	+	5	CAATT	common cis-acting element in promoter and enhancer regions

CCAAT-box

Site Name	Organism	Position	Strand	Matrix score.	sequence	function
CCAAT-box	Hordeum vulgare	648	+	6	CAACGG	MYBHv1 binding site
CCAAT-box	Hordeum vulgare	681	-	6	CAACGG	MYBHv1 binding site

CCGTCC-box

Site Name	Organism	Position	Strand	Matrix score.	sequence	function
CCGTCC-box	Arabidopsis thaliana	350	-	6	CCGTCC	cis-acting regulatory element related to meristem specific activation

ERE

Site Name	Organism	Position	Strand	Matrix score.	sequence	function
ERE	Dianthus caryophyllus	228	-	8	ATTTCAAA	ethylene-responsive element

G-Box

Site Name	Organism	Position	Strand	Matrix score.	sequence	function
G-Box	Pisum sativum	420	+	10	CACACATGAA	cis-acting regulatory element involved in light responsiveness

G-box

Site Name	Organism	Position	Strand	Matrix score.	sequence	function
G-box	Solanum tuberosum	422	+	7	CACATGG	cis-acting regulatory element involved in light responsiveness

GARE-motif

Site Name	Organism	Position	Strand	Matrix score.	sequence	function
GARE-motif	Brassica oleracea	70	+	7	TCTGTTG	gibberellin-responsive element
GARE-motif	Brassica oleracea	1081	-	7	AAACAGA	gibberellin-responsive element
GARE-motif	Brassica oleracea	840	+	7	TCTGTTG	gibberellin-responsive element

GATA-motif

GCN4_motif

Site Name	Organism	Position	Strand	Matrix score.	sequence	function
GCN4_motif	Oryza sativa	921	-	7	CAAGCCA	cis-regulatory element involved in endosperm expression

GT1-motif

Site Name	Organism	Position	Strand	Matrix score.	sequence	function
GT1-motif	Arabidopsis thaliana	1216	+	6	GGTTAA	light responsive element

LTR

Site Name	Organism	Position	Strand	Matrix score.	sequence	function
LTR	Hordeum vulgare	356	+	6	CCGAAA	cis-acting element involved in low-temperature responsiveness
LTR	Hordeum vulgare	493	+	6	CCGAAA	cis-acting element involved in low-temperature responsiveness
LTR	Hordeum vulgare	378	+	6	CCGAAA	cis-acting element involved in low-temperature responsiveness

MRE

Site Name	Organism	Position	Strand	Matrix score.	sequence	function
MRE	Petroselinum crispum	314	+	7	AACCTAA	MYB binding site involved in light responsiveness

Skn-1_motif

Site Name	Organism	Position	Strand	Matrix score.	sequence	function
Skn-1_motif	Oryza sativa	163	+	5	GTCAT	cis-acting regulatory element required for endosperm expression
Skn-1_motif	Oryza sativa	297	+	5	GTCAT	cis-acting regulatory element required for endosperm expression
Skn-1_motif	Oryza sativa	214	+	5	GTCAT	cis-acting regulatory element required for endosperm expression

Sp1

Site Name	Organism	Position	Strand	Matrix score.	sequence	function
Sp1	Zea mays	1243	+	5	CC(G/A)CCC	light responsive element
Sp1	Zea mays	1244	+	5	CC(G/A)CCC	light responsive element

TATA-box

Site Name	Organism	Position	Strand	Matrix score.	sequence	function
TATA-box	Arabidopsis thaliana	34	-	6	TATAAA	core promoter element around -30 of transcription start
TATA-box	Lycopersicon esculentum	911	-	5	TTTTA	core promoter element around -30 of transcription start
TATA-box	Lycopersicon esculentum	195	+	5	TTTTA	core promoter element around -30 of transcription start
TATA-box	Lycopersicon esculentum	1188	+	5	TTTTA	core promoter element around -30 of transcription start
TATA-box	Arabidopsis thaliana	36	+	4	TATA	core promoter element around -30 of transcription start
TATA-box	Lycopersicon esculentum	1091	-	5	TTTTA	core promoter element around -30 of transcription start
TATA-box	Lycopersicon esculentum	787	+	5	TTTTA	core promoter element around -30 of transcription start
TATA-box	Lycopersicon esculentum	1224	+	5	TTTTA	core promoter element around -30 of transcription start
TATA-box	Arabidopsis thaliana	35	-	5	TATAA	core promoter element around -30 of transcription start
TATA-box	Lycopersicon esculentum	1072	+	5	TTTTA	core promoter element around -30 of transcription start
TATA-box	Lycopersicon esculentum	720	-	5	TTTTA	core promoter element around -30 of transcription start
TATA-box	Glycine max	1191	-	5	TAATA	core promoter element around -30 of transcription start
TATA-box	Glycine max	146	+	5	TAATA	core promoter element around -30 of transcription start
TATA-box	Arabidopsis thaliana	1111	-	4	TATA	core promoter element around -30 of transcription start
TATA-box	Lycopersicon esculentum	821	+	5	TTTTA	core promoter element around -30 of transcription start

TCA-element

Site Name	Organism	Position	Strand	Matrix score.	sequence	function
TCA-element	Brassica oleracea	1075	-	9	CAGAAAAGGA	cis-acting element involved in salicylic acid responsiveness

circadian

Site Name	Organism	Position	Strand	Matrix score.	sequence	function
circadian	Lycopersicon esculentum	182	+	9	CAAAGATATC	cis-acting regulatory element involved in circadian control
circadian	Lycopersicon esculentum	477	+	6	CAANNNNATC	cis-acting regulatory element involved in circadian control

```
>L111_1370 1060nt
+ GTTTATCTCA ACTCATATCT TTCGATGATT ACAAAACAAA TGGATATTTC TAATACCTCT TGTAGAAATC
- CAAATAGAGT TGAGTATAGA AAGCTACTAA TGTTTTGTTT ACCTATAAAG ATTATGGAGA ACATCTTTAG

+ CTTTTCAGAA ATGAACCAAA CAGGTCTGAG GCTTTAAAGC AGATAAAGGA GATCAAAAAG TATGAAGATT
- GAAAAGTCTT TACTTGGTTT GTCCAGACTC CGAAATTTCG TCTATTTCCT CTAGTTTTTC ATACTTCTAA

+ GCTGAGGATG ATGTCGGACG CACAAAAGGA AAGTATCGGA CGTCCGATAT CCTTTGAGCA TCTTCGGACG
- CGACTCCTAC TACAGCCTGC GTGTTTTCCT TTCATAGCCT GCAGGCTATA GGAAACTCGT AGAAGCCTGC

+ AGTAAAGAAC TCAGACTCGG ACGTCCGAAG AAGACAAGCC AGAGGTTCGA TGACTACTAA TCATGATCGG
- TCATTTCTTG AGTCTGAGCC TGCAGGCTTC TTCTGTTCGG TCTCCAAGCT ACTGATGATT AGTACTAGCC

+ ACGCTCAACA TCTGTGTATC GGACGTCCGA CAAGTATTGC TGCTCTTCGG ACACAACACT CTGAACGCAT
- TGCGAGTTGT AGACACATAG CCTGCAGGCT GTTCATAACG ACGAGAAGCC TGTGTTGTGA GACTTGCGTA

+ CGGCCGTCCG AAGGATTTCA AGAAAACTTC CAGAACTCAT TGATCTCGTT CGGACGCAGA AAATCCAACC
- GCCGGCAGGC TTCCTAAAGT TCTTTTGAAG GTCTTGAGTA ACTAGAGCAA GCCTGCGTCT TTTAGGTTGG

+ ATCGGACGTC CGACAGCACC AACGGCTAGT TGACTCTTCA GCTGCTTTCT ACCCGTTAAC AGCATTAATT
- TAGCCTGCAG GCTGTCGTGG TTGCCGATCA ACTGAGAAGT CGACGAAAGA TGGGCAATTG TCGTAATTAA

+ GAGGAATTCT CTGGTCTCCT ATAAAAGGAA CAAAGTCAAC CACCTCAAGA CAACTTTGTA CATCAAGCTT
- CTCCTTAAGA GACCAGAGGA TATTTTCCTT GTTTCAGTTG GTGGAGTTCT GTTGAAACAT GTAGTTCGAA

+ ACATCAGATT GTGAGTGATT CAAGTGCTAG AATACTCAAA AGAAACATTT GTATTCCTTG CATGTGCAAT
- TGTAGTCTAA CACTCACTAA GTTCACGATC TTATGAGTTT TCTTTGTAAA CATAAGGAAC GTACACGTTA

+ CTTCGTGAAG CTGTTTTTTC AAGTGTGATA TAGCTTCTTC AATAGTGTAG CATAGTGAGG GTTTGCGAGT
- GAAGCACTTC GACAAAAAAG TTCACACTAT ATCGAAGAAG TTATCACATC GTATCACTCC CAAACGCTCA

+ GTATGTAAAA CTTCCTTGCT TGACCAAGTG TGTTTTGGGG CAAGAAGGAA GTGATCCCTT CCTTGTACAC
- CATACATTTT GAAGGAACGA ACTGGTTCAC ACAAAACCCC GTTCTTCCTT CACTAGGGAA GGAACATGTG

+ ATAAGATTGG TTGCAAGTCT ATTCAGCTTG AAGTAACTTG GTATGAAATA GAGGTGTTCA AACATCAGTT
- TATTCTAACC AACGTTCAGA TAAGTCGAAC TTCATTGAAC CATACTTTAT CTCCACAAGT TTGTAGTCAA

+ GTGTTTGAAG CTTCGTTTGG TCCTTAACTC TCATTTACTG CTTTTCTATA TAAACTGCTC TTCTCCTCAT
- CACAAACTTC GAACCAAACC AGGAATTGAG AGTAAATGAC GAAAAGATAT ATTTGACGAG AAGAGGAGTA

+ CTCACTAATA CCTGTGCTAC TATATTATCT TGTTCATTGA GAAGCATTTT GAAGAAGAAG GGCTGTCTGT
- GAGTGATTAT GGACACGATG ATATAATAGA ACAAGTAACT CTTCGTAAAA CTTCTTCTTC CCGACAGACA

+ CCAAAAAAGG TTGAATATTT ACTAGCAGGT TTTTGAAAGC CTAATTCACC CCCCCTCTTA GGTTGTCTTC
- GGTTTTTTCC AACTTATAAA TGATCGTCCA AAAACTTTCG GATTAAGTGG GGGGGAGAAT CCAACAGAAG

+ GATCCTTAC
- CTAGGAATG
```

ARE

Site Name	Organism	Position	Strand	Matrix score.	sequence	function
ARE	Zea mays	953	+	6	TGGTTT	cis-acting regulatory element essential for the anaerobic induction

ATGCAAAT motif

Site Name	Organism	Position	Strand	Matrix score.	sequence	function
ATGCAAAT motif	Oryza sativa	607	-	8	ATACAAAT	cis-acting regulatory element associated to the TGACTCA motif

Site Name	Organism	Position	Strand	Matrix score.	sequence	function
Box I	Pisum sativum	1012	-	7	TTTCAAA	light responsive element

Box-W1

Site Name	Organism	Position	Strand	Matrix score.	sequence	function
Box-W1	Petroselinum crispum	720	+	6	TTGACC	fungal elicitor responsive element

CAAT-box

Site Name	Organism	Position	Strand	Matrix score.	sequence	function
CAAT-box	Brassica rapa	37	+	5	CAAAT	common cis-acting element in promoter and enhancer regions
CAAT-box	Hordeum vulgare	548	-	4	CAAT	common cis-acting element in promoter and enhancer regions
CAAT-box	Glycine max	487	-	5	CAATT	common cis-acting element in promoter and enhancer regions
CAAT-box	Arabidopsis thaliana	778	-	5	CCAAT	common cis-acting element in promoter and enhancer regions
CAAT-box	Hordeum vulgare	316	-	4	CAAT	common cis-acting element in promoter and enhancer regions
CAAT-box	Hordeum vulgare	627	+	4	CAAT	common cis-acting element in promoter and enhancer regions
CAAT-box	Hordeum vulgare	489	-	4	CAAT	common cis-acting element in promoter and enhancer regions
CAAT-box	Hordeum vulgare	946	-	4	CAAT	common cis-acting element in promoter and enhancer regions
CAAT-box	Hordeum vulgare	138	-	4	CAAT	common cis-acting element in promoter and enhancer regions
CAAT-box	Brassica rapa	607	-	5	CAAAT	common cis-acting element in promoter and enhancer regions
CAAT-box	Hordeum vulgare	389	-	4	CAAT	common cis-acting element in promoter and enhancer regions
CAAT-box	Hordeum vulgare	670	+	4	CAAT	common cis-acting element in promoter and enhancer regions

CCAAT-box

Site Name	Organism	Position	Strand	Matrix score.	sequence	function
CCAAT-box	Hordeum vulgare	440	+	6	CAACCG	HBEBw1 binding site

CCGTCC-box

Site Name	Organism	Position	Strand	Matrix score.	sequence	function
CCGTCC-box	Arabidopsis thaliana	334	+	6	CCGTCC	cis-acting regulatory element related to meristem specific activation

GCN4_motif

Site Name	Organism	Position	Strand	Matrix score.	sequence	function
GCN4_motif	Oryza sativa	245	+	7	CAAGTCA	cis-regulatory element involved in endosperm expression

MBS

Site Name	Organism	Position	Strand	Matrix score.	sequence	function
MBS	Arabidopsis thaliana	834	-	6	CAACTG	MYB binding site involved in drought-inducibility

MRE

Site Name	Organism	Position	Strand	Matrix score.	sequence	function
MRE	Petroselinum crispum	1538	-	7	AACCTAA	MYB binding site involved in light responsiveness

MSA-like

Site Name	Organism	Position	Strand	Matrix score.	sequence	function
MSA-like	Catharanthus roseus	439	+	8.5	(T/C)(T/C)(T/C)AACNG(T/C)(T/C)(T/C)A	cis-acting element involved in cell cycle regulation

O2-site

Site Name	Organism	Position	Strand	Matrix score.	sequence	function
O2-site	Zea mays	147	+	9	GATGATGTGG	cis-acting regulatory element involved in zein metabolism regulation

P-box

Site Name	Organism	Position	Strand	Matrix score.	sequence	function
P-box	Oryza sativa	143	-	7	CCTTTTG	gibberellin-responsive element

Skn-1_motif

Site Name	Organism	Position	Strand	Matrix score.	sequence	function
Skn-1_motif	Oryza sativa	240	-	5	GTCAT	cis-acting regulatory element required for endosperm expression

Sp1

Site Name	Organism	Position	Strand	Matrix score.	sequence	function
Sp1	Zea mays	1529	+	5	CC(G/A)CCC	light responsive element
Sp1	Zea mays	1530	+	5	CC(G/A)CCC	light responsive element

TATA-box

Site Name	Organism	Position	Strand	Matrix score.	sequence	function
TATA-box	Oryza sativa	30	+	7	TACAAAA	core promoter element around -30 of transcription start
TATA-box	Brassica oleracea	888	+	6	ATATAA	core promoter element around -30 of transcription start
TATA-box	Lycopersicon esculentum	512	-	5	TTTTA	core promoter element around -30 of transcription start
TATA-box	Glycine max	513	-	5	TAATA	core promoter element around -30 of transcription start
TATA-box	Arabidopsis thaliana	558	+	9	ccTATAAAac	core promoter element around -30 of transcription start
TATA-box	Glycine max	916	+	5	TAATA	core promoter element around -30 of transcription start
TATA-box	Lycopersicon esculentum	756	-	5	TTTTA	core promoter element around -30 of transcription start
TATA-box	Arabidopsis thaliana	889	+	6	TATAAA	core promoter element around -30 of transcription start
TATA-box	Glycine max	51	+	5	TAATA	core promoter element around -30 of transcription start
TATA-box	Arabidopsis thaliana	931	-	4	TATA	core promoter element around -30 of transcription start
TATA-box	Arabidopsis thaliana	659	-	4	TATA	core promoter element around -30 of transcription start
TATA-box	Arabidopsis thaliana	510	+	6	TATAAA	core promoter element around -30 of transcription start
TATA-box	Arabidopsis thaliana	887	-	4	TATA	core promoter element around -30 of transcription start

TCA-element

Site Name	Organism	Position	Strand	Matrix score.	sequence	function
TCA-element	Brassica oleracea	69	-	9	CAGAAAAGGA	cis-acting element involved in salicylic acid responsiveness
TCA-element	Brassica oleracea	161	+	9	CAGAAAAGGA	cis-acting element involved in salicylic acid responsiveness

W box

Site Name	Organism	Position	Strand	Matrix score.	sequence	function
W box	Arabidopsis thaliana	720	+	6	TTGACC	

Annexe 3 – Valorisation scientifique

Publications

Guyot R., Darré T., Dupeyron M., De Kochko A., Hamon S., Couturon E., Crouzillat D., Rigoreau M., Rakotomalala J.-J., Raharimalala N. E., Doffou Akaffou S., Hamon P. *Partial sequencing reveals the transposable element composition of Coffea genomes and provides evidence for distinct evolutionary stories*. Mol. Genet. Genomics (2016) **291**(5):1979-90.

Dupeyron M., Fernandez de Souza R., Hamon P., De Kochko A., Crouzillat D., Couturon E. and Guyot, R. *Distribution of Divo in Coffea genomes, a poorly described family of Angiosperm LTR-Retrotransposons*. Mol Genet Genomics (2017) **292**(4):741-754. doi: 10.1007/s00438-017-1308-2. Epub 2017 Mar 17.

Dupeyron M., De Kochko A., Crouzillat D., Hamon P. and Guyot R. *Analysis of SIRE LTR-retrotransposons expansion and diversity indicate the geographic origin of Coffea arabica.* In prep.

Dupeyron M., De Kochko A., Crouzillat D., Hamon P. and Guyot R. *The evolution and diversity of SIRE lTR-retrotransposons are associated to the diversification of wild diploid Coffea species.* In prep.

Note sur le blog de BioMed Central à propos de l'ICTE à Saint-Malo – section « On Biology ». http://blogs.biomedcentral.com/on-biology/2016/05/19/bright-days-saint-malo-transposable-element-science/

Présentation orale

Distribution of *Divo* in *Coffea* genomes, a poorly described family of Angiosperm LTR-Retrotransposons. Lors du colloque DynaGeV à Montpellier (8 et 9 juin 2017).

Posters

Dupeyron M., Hamon P., De Kochko A., Crouzillat D., Hamon S. and Guyot, R. *Divo, a new LTR-Retrotranspons family in Coffea genomes*.
Présenté à l'ICTE 2016 (Internationational Congress on Transposable Elements) à Saint-Malo (16-19 avril 2016) et au colloque DynaGeV (Dynamique des Génomes Végétaux) à AgroParisTech (7 et 8 juillet 2016).
Présenté par Perla Hamon au congrès « VII International Rubiaceae and Gentianales Conference » à Copenhague, Danemark (11-14 septembre 2017).

Dupeyron M., Hamon P., De Kochko A., Crouzillat D. and Guyot, R. *SIRE LTR-retrotranspons and their evolution across wild Coffea diploid species*.
Présenté à DynaGeV au Cirad à Montpellier (8 et 9 juin 2017).

Dupeyron M., Hamon P., De Kochko A., Crouzillat D., Hamon S. and Guyot, R. *SIRE LTR-retrotranspons organization is congruent with molecular Coffea phylogenetic clades divergence.*
Présenté par Perla Hamon au congrès « VII International Rubiaceae and Gentianales Conference » à Copenhague, Danemark (11-14 septembre 2017).

Participation à des congrès/séminaires

- GDR Éléments Transposables à Paris les 1er et 2 décembre 2014 ;
- Journées des doctorants (PhDays), 30 et 31 mars 2015, 7 et 8 avril 2016 ;
- Journées du réseau DynaGeV (Dynamique des Génomes Végétaux), 28 et 29 mai 2015, INRA Centre de Recherche de Toulouse – 7 et 8 juillet 2016, AgroParisTech – 8 et 9 juin 2017, CIRAD à Montpellier ;
- ICTE (International Congress on Transposable Elements) du 16 au 19 avril 2016 à Saint-Malo ;
- ASIC (Association for Science and Information on coffee – International Conference on Coffee Science) du 13 au 19 novembre 2016, Kunming, Chine. Bourse obtenue.

Divo, a new family of LTR retrotransposons in *Coffea*.

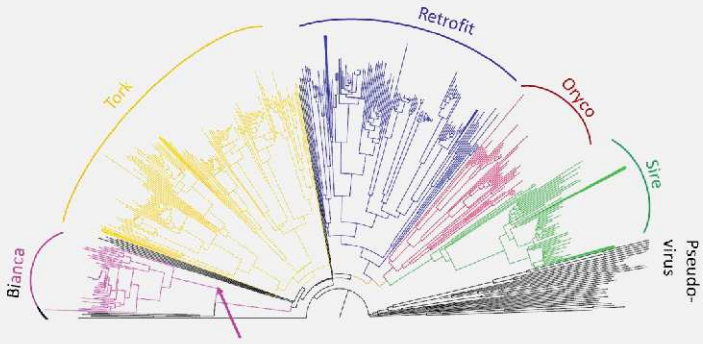
Dupeyron M[‡]*, Hamon P*, de Kochko A*, Crouzillat D[§], Hamon S* and Guyot R[‡]

*Evolution des Génomes des Caféiers , UMR DIADE; ‡CoffeeAdapt, UMR IMPE; IRD, centre IRD de Montpellier, BP 64501, Montpellier Cedex 5, France; §Nestlé R&D Tours, Notre Dame d'Oé, Tours, France

Abstract

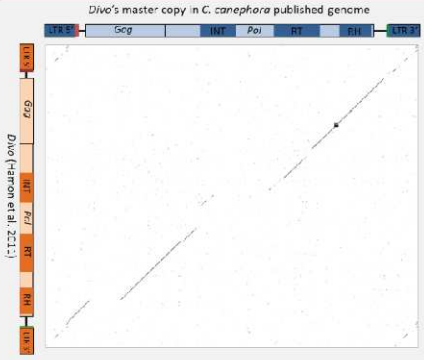
LTR retrotransposons (LTR-RTs) are the main components of plant genomes. Numerous lineages and families have been described leading to a well-established classification of elements, based on their overall structure and Reverse-Transcriptase based phylogeny. With the availability of bioinformatics tools dedicated to the LTR-RTs identification and analyses, it became reasonable to perform annotation of such transposable elements at whole-genome scale. Here, in the frame of the Arabica Coffee Genome Consortium (ACGC), we performed the LTR-RT annotation of fully sequenced allotetraploid *C. arabica* genome and its diploid progenitors. In particular, we described a novel family of *Copia* LTR-RTs named *Divo*, identified previously in *C. canephora* and used as molecular marker to study the genetic diversity among *Coffea* (Hamon *et al.*, 2011). *Divo* elements are complete and relatively compact (~5kb) carrying typical GAG and POL *Copia* domains. However, Reverse Tranciptase (RT) and Integrase (INT) domain-based phylogenetic analyses demonstrated that *Divo* forms a new and well supported family within the *Bianca* lineage. So far, *Divo* was restricted to dicotyledonous genomes. In coffee trees, as well as in *Arabidopsis* and grapevine, *Divo* was present in relatively low copy numbers, but in *C. arabica* and its diploid progenitors, the presence of recently inserted and complete copies and the detection of RNAseq transcription suggest that *Divo* might be active. The contribution of *Divo* to the allotetraploid *C. arabica* genome structure is analysed. Altogether our results indicated that *Divo* is a novel *Copia* LTR-RT family, ubiquitous in dicotyledonous genomes.

A new *Copia* LTR-RT family in *C. arabica* genome and its diploid progenitors



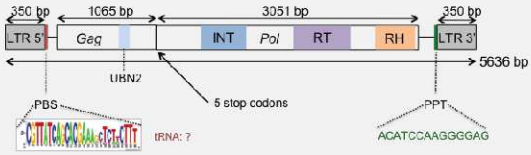
RT-based phylogenetic tree of *Copia* LTR-RTs in *C. arabica* genome.
Full length LTR-RTs were first predicted with LTR_STRUC in *C. arabica* genome. The RT domains were extracted and classified with Neighbor-Joining tree (100 bootstraps) according to the RT reference domains obtained from GyDB. The putative new family is indicated by a purple arrow in the *Bianca* lineage. **These results suggest that a new family, belonging to the *Bianca* lineage, is present in *C. arabica* and its diploid progenitors.**

Divo, a new family in Coffee trees



Dot-plot alignment of *Divo* (Hamon *et al.*, 2011) and a conserved copy of the new family present in *C. canephora* published genome.
Red rectangles: Primer Binding Site (PBS); Green rectangles: Polypurine Tract (PPT). **The new family is named *Divo*, since full-length annotated elements are similar to a partial element used as molecular marker (REMAP, Hamon *et al.*, 2011).**

Divo: a typical structure of *Copia* LTR-RTs



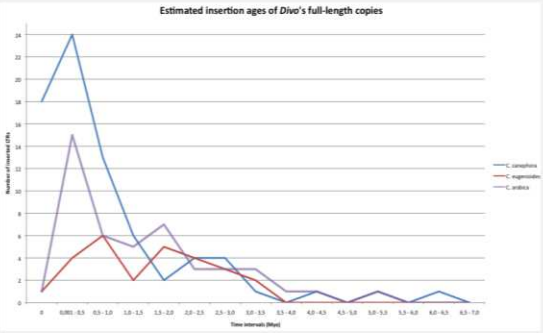
Structural characterization of *Divo*.
The Primer Binding Site (PBS) presents an ambigous sequence not corresponding to known tRNAs.
All these characteristics confirm that *Divo* is a transposable element displaying the typical structure of *Copia*.

Divo copy number in *C. arabica* and its progenitors

	Full-length copies (80% ID – 100% length)	Copies (80% ID – 80% length)	Partial copies (20% ID – 80% length)	Solo-LTRs	Total
C. canephora	41	129	212	142	524
C. arabica	37	204	351	201	793
C. eugenioides	20	132	223	336	711

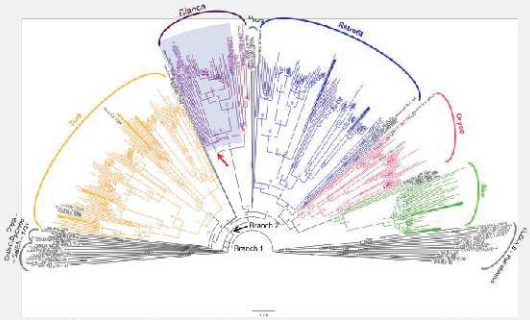
Assessment of the copy number of *Divo* in ACGC *C. canephora*, *C. arabica* and *C. eugenioides* draft genome sequences carried out with Censor (Kohany *et al.*, 2006).
The copy number differences between *C. canephora* and *C. eugenioides* (41 vs 20 and 142 vs 336) suggest different evolution of *Divo* in these two species. *C. arabica* doesn't present the addition of the copy number of its two progenitors, suggesting variation of *Divo* copy numbers during or after *C. arabica* formation both in the two diploids and in the allotetraploid.

Divo's insertions are recent



The insertion time of full-length copies was estimated based on the divergence of the 5'- and 3'-LTR sequences of *C. canephora*, *C. arabica* and *C. eugenioides* draft genomes (ACGC). The insertion dates were estimated using an average base substitution rate of 1.3 e^{-8} (Ma & Bennetzen 2004). ***C. canephora* presents much more younger *Divo* copies than *C. arabica* and *C. eugenioides*. *C. arabica* presents an insertion time pattern at the transition between its two progenitors. As some *Divo*'s insertions seem to be recent, this family could still be active.**

Divo is distributed among *Coffea* species and other Dicots



RT-based phylogenetic tree of *Copia* LTR-RTs in *C. arabica* together with *Divo* copies in *C. arabica* and other plant genus.
The presence of *Divo*'s copies in other species was attested using BLAST researches on NCBI, TAIR and RetrOryza websites. The RT domains corresponding to all these TEs were extracted and classified in a Neighbor-Joining tree (100 bootstraps, black names for the RT reference domains extracted from GyDB). *Divo* family is indicated by a red arrow and highlited in the *Bianca* lineage. **This suggests that the *Bianca* lineage contains at least two families, *Divo* and *Bianca*. *Bianca* seems present only in Monocots, whereas *Divo* has been found only in Dicots. This result raises questions about the differential evolution of this TE lineage across plant evolution.**

Conclusion: A novel repetitive element, called *Divo*, displaying the typical structure of *Copia* LTR-RTs was identified in the genome of *C. arabica* and its diploid progenitors *C. canephora* and *C. eugenioides*. Divo is a new family within the Bianca lineage only present in dicots while "original" Bianca would be only in monocots. With complete copies and recent inserted copies detected, *Divo* could be active and might have played a role in the evolution of *Coffea* genus.

References : Denoeud, F., *et al.* (2014) The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. *Science* **345**:6201. Hamon *et al.*, (2011) Two novel Tyl-*copia* retrotransposons isolated from coffee trees can effectively reveal evolutionary relationships in the *Coffea* genus (Rubiaceae). *Mol Genet Genomics* **285**:447-460.

SIRE LTR-retrotransposons and their evolution across wild *Coffea* diploid species

Dupeyron M[‡]*, Hamon P*, de Kochko A*, Crouzillat D[§], and Guyot R[‡]

*Evolution des Génomes des Caféiers , UMR DIADE; [‡]CoffeeAdapt, UMR IMPE; IRD, centre IRD de Montpellier, BP 64501, Montpellier Cedex 5, France; [§]Nestlé R&D Tours, Notre Dame d'Oé, Tours, France

mathilde.dupeyron@ird.fr

Abstract

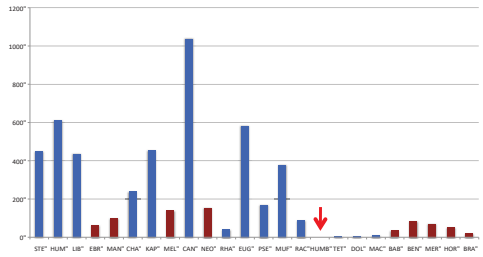
LTR-retrotransposons (LTR-RTs) are major components of plant genomes. Numerous lineages and families have been described and nowadays, the availability of bioinformatics tools allows us to study these transposable elements (TEs) at whole-genome scale.

The genome of *Coffea canephora* (Robusta coffee) has been released recently and contains about 50% of TEs, among which 42% are LTR-RTs with *Gypsy* and *Copia* being the predominant super-families. Besides *Coffea canephora*, 24 wild *Coffea* species, representative of the taxonomic groups now well resolved, are available to understand the variation of transposable element composition in the genus.

SIRE LTR-RTs are *Copia* elements discovered in *Glycine max* and containing an envelope-like domain. They are widespread in plant genomes and present conserved motifs in their sequence, suggesting an ancient origin of this lineage. Previous analyses of TE composition in 11 *Coffea* species partially sequenced with the 454 technology showed a different repartition of *SIRE* according to their geographic distribution, suggesting different dynamics according to taxonomic groups. A deeper insight in *SIRE* composition and evolution is needed to confirmed or deny our previous observations in *Coffea* species. We used deep Illumina sequencing data available for the 24 wild diploid *Coffea* species to study their *SIRE* composition and evolution.

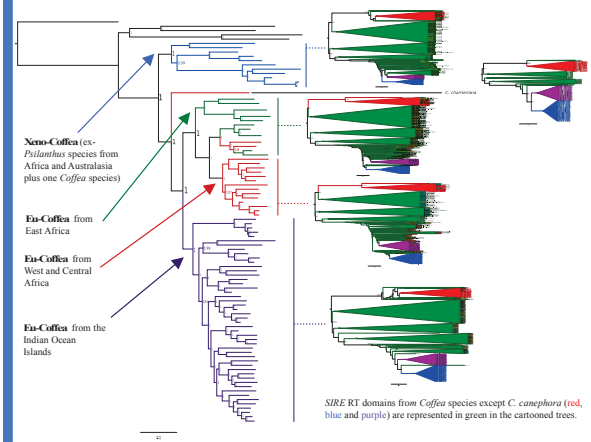
We showed that *SIRE* are present in all the coffee-trees studied but with a difference in copy number. This difference seems due to a high divergence of *SIRE* sequences according to the taxonomic groups and not to a high difference in copy number. Moreover, the *SIRE* LTR-RTs underwent a diversification in three families in *Coffea* of West and Central Africa, whereas this diversification is subtler in *Coffea* from East Africa and very different in *Coffea* from Indian Ocean Islands and Australasia.

The *SIRE* lineage is detected in most of *Coffea* species



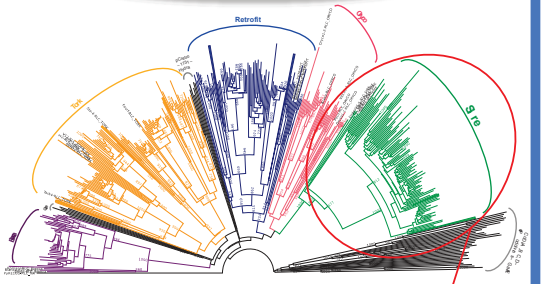
Copy number estimation (Bowtie2) of SIRE RT domains in 24 species of coffee-trees classified according to their geographical areas. STE=*Coffea stenophylla*; HUM=*C. humilis*; LIB=*C. liberica*; EBR: *C. ebracteolata*; MAN: *C. mannii*; CHA: *C. charrieriana*; KAP: *C. kapakata*; MEL: *C. melanocarpa*; CAN: *C. canephora* (BUD15, Uganda); NEO: *C. neoleroyi*; RHA: *C. rhamnifolia*; EUG: *C. eugenioides* (DA56, Kenya); PSE: *C. pseudozanguebariae*; MUF: *C. mufindiensis*; RAC: *C. racemosa*; HUMB: *C. humblotiana*; TET: *C. tetragona*; DOL: *C. dolichophylla*; MAC: *C. macrocarpa*; BAB: *C. babaudani*; BEN: *C. benghalensis*; MER: *C. mergensis*; HOR: *C. horsefieldiana*; BRA: *C. brassii*. The ex-*Psilanthus* species are colored in burgundy.

SIRE underwent different diversifications in the four major taxonomic groups



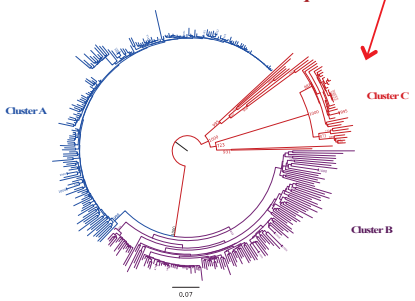
***Coffea* phylogenetic NJ trees of *SIRE* RT domains according to the four main taxonomic groups of the *Coffea* genus (from Hamon et al. 2017).** RT domains of *SIRE* have been extracted from 23 species after a BLASTx search against *SIRE* RT references of the three families (C red, A blue and purple) from *C. canephora*. NJ trees of these RT domains have been computed (1000 boot straps) with *C. canephora* RT references and according to each taxonomic group (in different colors in *Coffea* phylogeny).

SIRE lineage in *C. canephora*



RT-based phylogenetic tree of *Copia* LTR-RTs in *C. canephora* genome. Full length LTR-RTs were first predicted with LTR_STRUC in *C. canephora* genome. The RT domains were extracted and classified with Neighbor-Joining tree (1000 boot straps) according to the RT reference domains (black branches) from GyDB.

Three families of *SIRE* in *C. canephora*



The *SIRE* lineage contains three families, A, B and C in *C. canephora*. Longer branches for family C suggest that it is older than the two others.

Results & Discussion

SIRE lineage is present in all the species studied here. The copy number estimation shows that it is in high number in *Coffea* from West and Central Africa. The copy number decreases in species from East Africa and in a more important way in Xeno-Coffea. Finally, they are very poorly detected in *Coffea* from the Indian Ocean Islands (IOIs).

As three clusters (families) have been identified in *C. canephora*, the most conserved copies of each has been extracted and mined in *C. canephora*. The complete copies were used as references for BLASTx searches of their RT domains in the 24 diploid *Coffea* species (Illumina sequencing). Matching RT domains of at least 200 amino-acids were extracted and NJ trees were computed with *C. canephora*'s *SIRE* RT domains. First, contrary to what is shown by copy number estimations, the BLASTx searches identified *SIRE* RT domains in the same proportions in all the 24 species (depending on sequencing and assembly qualities). Mapping analysis did not detected any *SIRE* in ***C. humblotiana*** whereas RT domains were found in the BLASTx analysis. This can be due to an important sequence divergence between *SIRE* in *C. canephora* and *SIRE* in *Coffea* from the IOIs. When comparing the presence of the three families in the 24 *Coffea* species analysed, we can precisely observe a different diversification of *SIRE* according to the four major taxonomic groups: Xeno-Coffea contains cluster C and *SIRE* close to cluster B but no RT domains belonging to cluster A. This is pretty the same for *C. charrieriana* which is branched beforehand the split in different clades of Eu-Coffea. Eu-Coffea from East Africa show a more important diversification in clade A and B but not as important as the Eu-Coffea from West and Central Africa (among which *C. canephora* still shows the most important *SIRE* diversification). Finally, the *Coffea* from the IOIs contain few RT domains in cluster C. The majority of *SIRE* is close to clusters A and B but none of them really belong to these clusters, suggesting a different diversification of *SIRE* in other clusters. This is in agreement with the hypothesis of a sequence divergence making the mapping inefficient to detect copies of *SIRE* in IOIs' *Coffea*, as only one sequence from *C. canephora* was used as a reference.

So *SIRE* seem to have underwent a diversification in three families in Eu-Coffea from West and Central Africa, particularly in *C. canephora*. This diversification happened differently or did not happen in the other *Coffea* diploid species.

Conclusion

The study of *SIRE* lineage in wild *Coffea* diploid species allowed us to observe a different dynamics of these LTR-RTs according to the four major taxonomic groups of the *Coffea* genus. An important divergence is showed by RT domains phylogenies between species from Africa and species from IOIs and Australasia. Moreover, the only species having an important geographical repartition area, *C. canephora*, is also the species showing the most important *SIRE* differentiation. We can suggest that *SIRE* were present in *Coffea* genomes before the emergence of the *Coffea* genus and they followed its diversification according to the geographical repartition of the species.

References:

Denoeud, F., *et al.* (2014) The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. *Science* **345**:6201.

Arabica Coffee Genome Consortium (ACGC 2014).

Hamon P. et al. (2017). GBS coffee phylogeny and the evolution of caffeine content. *Mol Phylogenet Evol* **109**:351-361.

Guyot R. et al. (2017). Partial sequencing reveals the transposable element composition of *Coffea* genomes and provides evidence for distinct evolutionary stories. *Mol Genet Genomics* 1-12.

Bousios A. et al. (2010) Highly conserved motifs in non-coding regions of *Sireevirus* retrotransposons: the key for their pattern of distribution within and across plants? *BMC Genomics* **11**:89

SIRE LTR-retrotransposons organization is congruent with molecular *Coffea* phylogenetic clades divergence.

Dupeyron M‡*, Hamon P*, de Kochko A*, Crouzillat D§, Hamon S* and Guyot R‡

*Evolution des Génomes des Caféiers , UMR DIADE; ‡CoffeeAdapt, UMR IMPE; IRD, centre IRD de Montpellier, BP 64501, Montpellier Cedex 5, France; §Nestlé R&D Tours, Notre Dame d'Oé, Tours, France

Abstract

LTR-retrotransposons (LTR-RTs) are major components of plant genomes. Numerous lineages and families have been described and nowadays, the availability of bioinformatics tools allows us to study these transposable elements (TEs) at whole-genome scale.

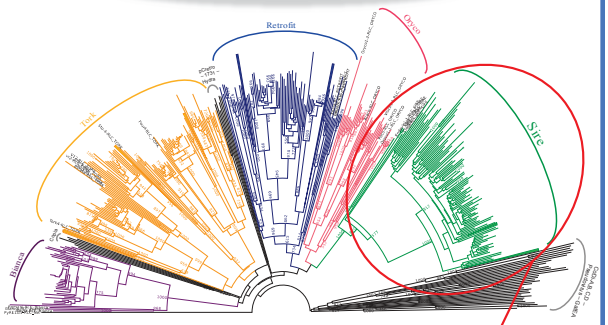
The genome of *Coffea canephora* (Robusta coffee) has been released recently and contains about 50% of TEs, among which 42% are LTR-RTs with *Gypsy* and *Copia* being the predominant super-families. Besides *Coffea canephora,* 24 wild *Coffea* species, representative of the major biogeographic groups, are available to understand the variation of transposable element composition in the genus.

SIRE LTR-RTs are *Copia* elements discovered in *Glycine max*. They are widespread in plant genomes and present conserved motifs in their sequence, suggesting an ancient origin of this lineage. Previous analyses of TE composition in 11 *Coffea* species partially sequenced showed a different repartition of *SIRE* according to their geographic distribution, suggesting different dynamics.

To get a deeper insight in *SIRE* composition and evolution, here we used deep Illumina sequencing data available for 24 wild diploid *Coffea* species.

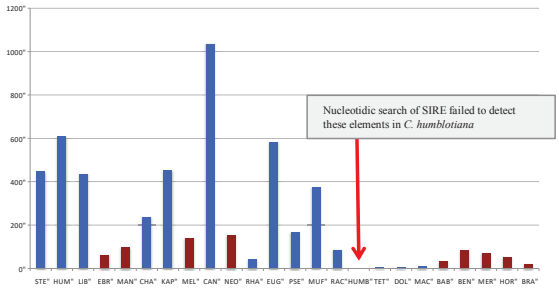
We showed that *SIRE* are present in all the studied species but with different copy numbers. This difference more reflects the high divergence of *SIRE* sequences between the biogeographic groups than real high difference in copy numbers. Moreover, the *SIRE* LTR-RTs underwent its differentiation in three families in Eu-*Coffea* from West and Central Africa (apart *C. charrieriana*), whereas this diversification is subtler in East African species and occurred differently and independently in WIOIs.

SIRE lineage in *C. canephora*



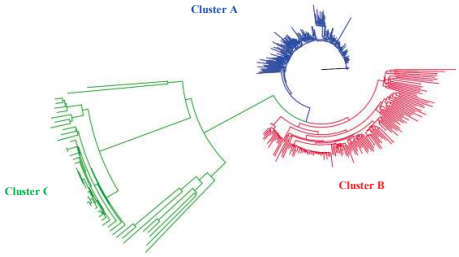
RT-based phylogenetic tree of *Copia* LTR-RTs in *C. canephora* genome.
Full length LTR-RTs were first predicted with LTR_STRUC in *C. canephora* genome. The RT domains were extracted and classified with Neighbor-Joining tree (1000 bootstraps) according to the RT reference domains (black branches) from GyDB.

The *SIRE* lineage is detected in most of *Coffea* species



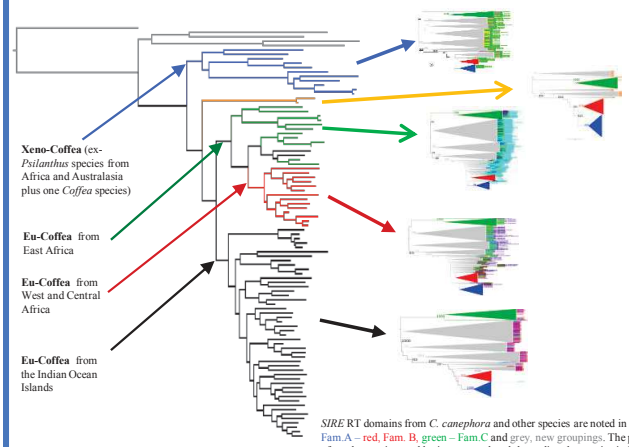
Copy number estimation (Bowtie2) of SIRE RT domains in 24 species of coffee-trees classified according to their geographical areas. STE=*Coffea stenophylla*; HUM=*C. humilis*; LIB=*C. liberica*; EBR: *C. ebracteolata*; MAN: *C. mannii*; CHA: *C. charrieriana*; KAP: *C. kapakata*; MEL: *C. melanocarpa*; CAN: *C. canephora* (BUD15, Uganda); NEO: *C. neoleroyi*; RHA: *C. rhamnifolia*; EUG: *C. eugenioides* (DA56, Kenya); PSE: *C. pseudozanguebariae*; MUF: *C. mufindiensis*; RAC: *C. racemosa*; HUMB: *C. humblotiana*; TET: *C. tetragona*; DOL: *C. dolichophylla*; MAC: *C. macrocarpa*; BAB: *C. babahudanii*; BEN: *C. benghalensis*; MER: *C. merguensis*; HOR: *C. horsefieldiana*; BRA: *C. brassii*.
The ex-*Psilanthus* species are colored in burgundy.

Three families of *SIRE* in *C. canephora*



In *C. canephora*, the *SIRE* lineage is composed with three families, A, B and C.

SIRE underwent different diversifications in the four major taxonomic groups



SIRE RT domains from *C. canephora* and other species are noted in blue, Fam.A – red, Fam. B, green – Fam.C and grey, new groupings. The presence of an element is noted by its name colored depending the species it derived.

NJ trees of *SIRE* RT domains according to four main biogeographic groups among the *Coffea* genus (phylogeny adapted from Hamon et al. 2017).
SIRE RT domains were extracted from 23 species after a BLASTx search against *SIRE* RT references of the three families (A, B and C) from *C. canephora*. NJ trees of these RT domains have been computed (1000 bootstraps) with *C. canephora* RT references and according to each biogeographic group (in different colors in the *Coffea* phylogeny).

Results & Discussion

SIRE lineage is present in all the species studied here. The copy number estimates based on nucleotidic sequences show high numbers in West and Central Africa, decreasing numbers in East Africa and in a more important way in Xeno-Coffea clade. They are very poorly detected in the Western Indian Ocean Islands (WIOIs).

The most conserved copies of each family (A, B and C) identified in *C. canephora* were used as references for BLASTx searches (based on amino acids sequences) of their RT domains in the 24 diploid *Coffea* species. Matching RT domains of at least 200 a.a were extracted and NJ trees computed using *C. canephora's SIRE* RT domains as references.

In that case, the BLASTx searches identified *SIRE* RT domains in all species including *C. humblotiana*. This can be due to high *SIRE* divergence at the nucleotidic level between *C. canephora* and *C. humblotiana*.

Looking at the structuring of *SIRE* in the 24 *Coffea* species, we can note:

* All species contains *SIRE* belonging to the family C.
* Xeno-Coffea contains also other new groupings but only *P. melanocarpus* has representatives of families A & B.
* *C. charrieriana* in intermediate position between XC- and EC- clades is similar to XC-clade for *SIRE* organization. This observation is one more peculiar feature that set it apart from EC-clade (all *ex-Coffea* but *C. rhamnifolia*).
* Eu-Coffea from East Africa and especially *C. eugenioides* from Uganda and *C. mufindiensis* have elements belonging to families A and B.
* The most important *SIRE* diversification is observed in West and Central Africa.
* *SIRE* structuring in WIOIs shows similar pattern than in XC-clade.
* Finally, *SIRE* would underwent its diversification in three families (A, B and C) in Eu-Coffea from West and Central Africa, particularly in *C. canephora*.

On an evolutionary point of view, these results support the recent activation of families A and B in the African Eu-Coffea clade (exception to *C. charrieriana*), that occurred after African and WIOIs divergence.

Assumptions on the presence of *SIRE* in ancestral *Coffea* genomes rise the question on their presence and level of differentiation in the Coffeeae tribe and related Rubiaceae tribes.

Conclusion

In *C. canephora*, *SIRE* lineage is composed with three well defined and supported families. The study of this lineage in wild *Coffea* diploids allowed us to observe its different dynamics following major biogeographic *Coffea* groups differentiation. Therefore, the major divergence is showed within the EC-clade, between African and non-African species. Intriguing, *C. charrieriana* from Cameroon, WIOIs- and XC- clades show similar pattern of *SIRE* organization. We assume that *SIRE* was present in ancestral *Coffea* genomes and its own diversification followed the genus evolution at the high level (basal branches) of the *Coffea* phylogeny.

References:

Denoeud, F., *et al*. (2014) The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. *Science* **345**:6201.

Arabica Coffee Genome Consortium (ACGC 2014).

Hamon P. et al. (2017). GBS coffee phylogeny and the evolution of caffeine content. *Mol Phylogenet Evol* **109**:351-361.

Guyot R. et al. (2017). Partial sequencing reveals the transposable element composition of *Coffea* genomes and provides evidence for distinct evolutionary stories. *Mol Genet Genomics* 1-12.

Bousios A. et al. (2010) Highly conserved motifs in non-coding regions of *Sirevirus* retrotransposons: the key for their pattern of distribution within and across plants? *BMC Genomics* **11**:89

Annexe 4 – Expérience d'enseignement

Monitorat

Un an de charge d'enseignement au sein du département Biologie-Écologie de l'Université de Montpellier (correspondant à 45h de formation) :

- Travaux Pratiques (TP) de 3h et Travaux Dirigés (TD) de 1h30 de Biologie Intégrative (HLBE 101) Licence première année (L1) au premier semestre de l'année scolaire 2015/2016
- TD de 2h de Cycle de Vie 1 (HLBE 201) L1 au deuxième semestre de l'année scolaire 2015/2016

Annexe 5 – Formations réalisées

Modules scientifiques

- Accueil et Rencontre des Doctorants (11 décembre 2014), **3 heures** validées par le Collège Doctoral Languedoc Roussillon, catégorie : *Méthodologie et outils de la thèse* ;
- Initiation à la bioinformatique au Brésil (du 27 au 31 octobre 2014), **30 heures** validées par l'École doctorale GAIA (anciennement SIBAGHE), catégorie : *Formation aux logiciels* ;
- Formation « Bien commencer avec R » (du 23 au 25 juin 2015), **18 heures** validées Collège Doctoral Languedoc Roussillon, catégorie : *Méthodologie et outils de la thèse*.

Modules d'ouverture

- Organisation d'un séminaire (PhDays, 30 et 31 mars 2015), **13 heures** validées par l'École doctorale GAIA (anciennement SIBAGHE), catégorie : *Doctoriales* ;
- Prise de parole en public, pédagogie interactive niveau 1 (4, 5 et 11 mai 2015), **21 heures** validées par le Collège Doctoral Languedoc Roussillon, catégorie : *Enseignement*.
- Présentation d'un poster à un congrès international (ICTE – du 16 au 19 avril 2016), **4 heures** validées par l'École doctorale GAIA, catégorie : *Séminaires* ;
- Doctorat et poursuite de carrière, association PhDOOC (Massive Open Online Courses – MOOC – décembre 2016 et janvier 2017), **18 heures** validées par l'École doctorale GAIA, catégorie : *Préparation à l'arpès-thèse*.

Total formations : 107 heures / 7 modules.

Résumé

Les éléments transposables (ET) sont des portions d'ADN capables de se déplacer et d'augmenter le nombre de leurs copies dans les génomes. Deux grands types de transposition, correspondant à deux grandes classes d'ET, sont retrouvés chez la quasi-totalité des génomes étudiés à ce jour. Les rétrotransposons à LTR (Long Terminal Repeats, LTR-RT), appartenant à la Classe 1, sont les composants majoritaires des génomes des plantes. Leur prolifération peut avoir un impact important sur l'organisation, la variation de taille, l'évolution des génomes et l'activité des gènes.
Le café, largement consommé dans le monde et produit uniquement par des pays du Sud, est issu de deux espèces cultivées d'origine africaine : *Coffea arabica* et *C. canephora*. Le genre *Coffea* est constitué de 139 espèces occupant des habitats très variés en Afrique, dans les îles de l'ouest de l'océan Indien, l'Inde, l'Asie tropicale et du sud-est et au nord de l'Australie. Toutes les espèces son diploïdes, à l'exception notable de *C. arabica*, allotétraploïde, issu d'une hybridation interspécifique récente entre les deux espèces diploïdes : *C. canephora* et *C. eugenioides*. Pour autant, la taille des génomes des espèces diploïdes varie du simple au double. Les nombreuses données génomiques aujourd'hui disponibles au sein du genre *Coffea* permettent d'étudier la dynamique des LTR-RT constituant au minimum 42% du génome de *C. canephora,* l'espèce séquencée et disponible dans les bases de données publiques.
Dans ce travail, deux lignées remarquables de LTR-RT, *Bianca* et *SIRE*, ont été étudiées par des approches bio-informatiques. *Bianca sensu stricto*, présente uniquement chez les monocotylédones, est représentée chez les dicotylédones par la famille *Divo*, très peu étudiée à ce jour. L'activation récente de *Divo* sans induire sa propre structuration, est étroitement associée à la différenciation génétique de *C. canephora*. Par contre, tout en étant présente dans toutes les espèces de caféiers étudiées, l'activation semble sporadique. À l'opposé, les éléments *SIRE*, la seule lignée de LTR-RT de la superfamille des *Copia* contenant un domaine *enveloppe* comme les rétrovirus, montre des variations structurales importantes entre les accessions des espèces diploïdes à l'origine de *C arabica* et plus globalement, et en parallèle de l'évolution du genre.
Nos travaux montrent que la compréhension de la dynamique des LTR-RT dans un genre peut permettre de mieux appréhender son histoire évolutive, chaque famille de LTR-RT pouvant apporter un éclairage différent. Nos résultats indiquent qu'à la fois les clades biogéographiques (phylogénie moléculaire des caféiers) mais aussi certaines accessions d'espèces diploïdes ont des histoires particulières. Celles-ci seraient vraisemblablement liées à la colonisation de nouvelles niches et à la dynamique des LTR-RT composant les génomes des *Coffea*.

Mots clés : Rétrotransposons à LTR *Copia*, *Divo*, *SIRE,* dynamique évolutive*, Coffea*.

Summary

Transposable elements (TEs) are DNA fragments that are able to move and to increase their copy numbers. Two transposition mechanisms corresponding to the two main TE classes are found in almost all organisms. LTR retrotransposons (Long Terminal Repeats, LTR-RTs), belonging to Class 1, are the main components of plant genomes. Genome organisation, size variation, evolution and gene activity can be strongly impacted by their proliferation.
Worldwide consumed and produced by South countries, coffee is obtained from two African cultivated species: *Coffea arabica* and *C. canephora*. The *Coffea* genus includes 139 species occurring in diverse habitats in Africa, Madagascar, Mascarene Islands, Comoros, India, Southeast and Tropical Asia and North Australia. All the species are diploids, except the noteworthy allotetraploid *C. arabica*, originated from a recent inter-specific hybridisation between two diploids: *C. canephora* and *C. eugenioides*. However, genome size of diploid species can vary for up to two folds. Today, the numerous genomic data available for *Coffea* allows the study of LTR-RTs, constituting at least 42% of *C. canephora* genome, the sequenced species available in public databases.
In this work, two notable LTR-RT lineages, *Bianca* and *SIRE*, have been studied by bioinformatics approaches. *Bianca s.s.*, is present only in Monocots and it is represented in Dicots by the *Divo* family, poorly studied nowadays. The recent activation of *Divo,* without leading to its own structuring, is closely associated to the genetic differentiation of *C. canephora*. However, this activation seems sporadic as being present in all the coffee-trees species studied here. On the opposite, *SIRE* elements, which are the only *Copia* LTR-RTs carrying an envelope-like gene as retroviruses, show an important structuring variation between accessions among *C. arabica* progenitors, and in parallel to the genus evolution.
Our work shows that understanding the LTR-RTs dynamics in a genus allows a better perception of its evolutionary history, with the possibility of different evolutionary timing given by different LTR-RTs families. Our results also indicate that both the biogeographic clades (coffee molecular phylogeny) and also some diploid accessions have peculiar histories, probably related to the colonisation of new ecological niches and to the LTR-RTs dynamics.

Keywords: *Copia* LTR retrotransposons, *Divo*, *SIRE*, evolutionary dynamics, *Coffea*.