UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Câmpus de São José do Rio Preto

Elaine Silva Dias

Análise de diversidade e expressão de retrotransposons ativos em espécies de *Coffea*

São José do Rio Preto
2015

Elaine Silva Dias

Análise de diversidade e expressão de retrotransposons ativos em espécies de *Coffea*

Tese em cotutela apresentada como parte dos requisitos para obtenção do título de Doutor em Genética, junto ao Programa de Pós-Graduação em Genética, do Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista "Júlio de Mesquita Filho", Campus de São José do Rio Preto, Brasil; e de Doutor em *Biologie intégrative des plantes* junto ao SIBAGHE - *Systèmes Intégrés en Biologie, Agronomie, Géosciences, Hydrosciences, Environnement*, da Universidade de Montpellier, França.

Orientador: Prof.ª Dr.ª Claudia Marcia Aparecida Carareto
Orientador: Dr. Alexandre de Kochko
Coorientador: Dr. Romain Guyot

São José do Rio Preto
2015

Dias, Elaine Silva.
Análise de diversidade e expressão de retrotransposons ativos em espécies de Coffea / Elaine Silva Dias. -- São José do Rio Preto, 2015
175 f. : il., tabs.

Orientador: Claudia Marcia Aparecida Carareto
Orientador: Alexandre de Kochko
Coorientador: Romain Guyot
Tese (doutorado com dupla titulação) – Universidade Estadual Paulista "Júlio de Mesquita Filho", Instituto de Biociências, Letras e Ciências Exatas e Universidade de Montpellier

1. Genética molecular. 2. Genomas. 3. Café - Genética. 4. Plantas - Evolução. 5. Elementos de DNA transponíveis. I. Carareto, Claudia Marcia Aparecida. II. Kochko, Alexandre de. III. Guyot, Romain. IV. Universidade Estadual Paulista "Júlio de Mesquita Filho". Instituto de Biociências, Letras e Ciências Exatas. V. Universidade de Montpellier. VI. Título.

CDU – 575.113

Ficha catalográfica elaborada pela Biblioteca do IBILCE
UNESP - Câmpus de São José do Rio Preto

Elaine Silva Dias

Análise de diversidade e expressão de retrotransposons ativos em espécies de *Coffea*

Tese em cotutela apresentada como parte dos requisitos para obtenção do título de Doutor em Genética, junto ao Programa de Pós-Graduação em Genética, do Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista “Júlio de Mesquita Filho”, Campus de São José do Rio Preto, Brasil; e de Doutor em *Biologie intégrative des plantes* junto ao SIBAGHE - *Systèmes Intégrés en Biologie, Agronomie, Géosciences, Hydrosciences, Environnement*, da Universidade de Montpellier, França.

Comissão Examinadora

Prof.ª Dr.ª Claudia Marcia Aparecida - Orientador
UNESP – São José do Rio Preto

Prof. Dr. Alexandre de Kochko - Orientador
IRD – Instituto de Pesquisa para o Desenvolvimento, França

Prof.ª Dr.ª Diana Fernandez
IRD – Instituto de Pesquisa para o Desenvolvimento, França

Prof. Dr. Gustavo Kuhn
UFMG – Universidade Federal de Minas Gerais

Prof.ª Dr.ª Maria Pilar Garcia Guerreiro
UAB – Universidade Autônoma de Barcelona, Espanha

Prof.ª Dr.ª Maura Helena Manfrin
USP – Universidade de São Paulo

São José do Rio Preto - 07 de julho de 2015

A presente Tese foi realizada em Cotutela entre a UNIVERSIDADE ESTADUAL PAULISTA “JÚLIO DE MESQUITA FILHO” e a “UNIVERSIDADE DE MONTPELLIER” sob orientação da Profa. Dra. Claudia Marcia Aparecida Carareto (UNESP - Brasil) e do Dr. Alexandre de Kochko (IRD, Institut de recherche pour le développement - França). O presente trabalho foi realizado no Departamento de Biologia do Instituto de Biociências Letras e Ciências Exatas de São José do Rio Preto, da UNESP, no Laboratório de Evolução Molecular sob responsabilidade da Profa. Dra. Claudia M. A. Carareto, com bolsa de Doutorado no País da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP, processo 2011/18226-0) e bolsa de doutorado-sanduíche Programa CAPES-Fundação Agropolis de uma parceria da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES, processo 9127-11-9) com a Fundação Agropolis. Na Universidade de Montpellier, o presente trabalho foi desenvolvido junto à equipe EVOGEC (Evolution du Génome des Caféiers) da unidade DIADE (Diversité, Adaptation et Développement des Plantes) no IRD.

Dedico este trabalho

À minha “vó” Rosa (*in memoriam*)... sempre em meus pensamentos.

“O saber a gente aprende com os mestres e os livros.
A sabedoria se aprende é com a vida e com os humildes.”

Cora Coralina - poetisa e contista brasileira

AGRADECIMENTOS

E quatro anos e alguns poucos meses se passaram... tempo em que tanta coisa foi vivida, tempo em que novas e antigas pessoas fizeram tanta diferença. É chegada a hora de tentar lembrar-se de todas e agradecê-las, pois certamente essa etapa não seria cumprida sem cada uma das pessoas que de alguma forma contribuíram e que, ao longo desse tempo, me sustentaram nessa jornada.

"[...] Nobody said it was easy
No one ever said it would be this hard [...]"

The Scientist – Coldplay

Agradeço imensamente à minha orientadora pela UNESP, Professora Claudia Marcia Aparecida Carareto, não somente por esses quatro anos, mas sim, por mais de oito anos de parceria. Muito obrigada por ter sido, para mim, um exemplo de conduta, de ética e de orientação, e, sobretudo, nesses quatro anos, por me ensinar não somente a questionar e a ouvir, mas também a acreditar no e a defender o meu trabalho. Muito obrigada, quaisquer palavras que eu escrevesse aqui seriam poucas para lhe agradecer por tudo o que aprendi com a senhora.

"Diga-me eu esquecerei, ensina-me e eu poderei lembrar, envolva-me e eu aprenderei."
Benjamin Franklin

Ao Doutor Alexandre De Kochko, meu orientador pela Universidade de Montpellier (UM), obrigada por ter aberto as portas de sua equipe, foram, talvez, passos lentos e pequenos tropeços, mas que mudaram completamente o rumo dessa jornada e, principalmente, a minha visão sobre a pesquisa e o mundo científico. Muito obrigada pelas conversas e por partilhar comigo algumas das experiências que viveu em alguns recantos do mundo, eu apreciei deveras cada uma delas e me esforcei para aprender com 'as linhas e as entrelinhas'.

Ao Doutor Romain Guyot, meu supervisor pela UM, agradeço-lhe imensamente pelos diversos ensinamentos.

Aos pesquisadores, Dra. Perla Hamon, Dr. Serge Hamon, Dra. Valerie Poncet, Dra Julissa Roncal, Christine Tranchant-Dubreuil, dentre outros do Instituto de Pesquisa para o Desenvolvimento (IRD do Francês, L'Institut de recherche pour le développement), pelo apoio e pelas ricas discussões científicas que pude participar, recebam o meu muito obrigada.

Às pesquisadoras, Dra. Marie-Angèle Grandbastien, Dra. Frédérique Pelsy, Dra. Dominique This e Dra. Marie Mirouze pelas valiosas contribuições durante meus comitês de tese pela UM.

Aos Pesquisadores, Dr. Douglas Domingues, Dr. Pierre Marraccini, Dr. Alan Andrade, Dr. Oliveiro Guerreiro Filho, Dr. A. D'Hont, Dr. Herman E. Taedoumg, *Dr P. De Block, Dr J-J. Rakotomalala e Dr. Philippe Lashermes por disponibilizarem amostras utilizadas neste trabalho, muito obrigada.*

A todos do IRD que me ajudaram, me ensinaram e fizeram da minha estadia no instituto uma experiência tão rica e agradável, obrigada pela ajuda, pelas conversas, pela presença. Um obrigado especial aos chefs da cozinha que sempre me trataram com muito carinho.

Aos funcionários da Universidade de Montpellier que me auxiliaram nas mais diferentes etapas, muito obrigada.

Às Professoras Mary Massumi Itoyama e Lilian Madi-Ravazzi, pelas contribuições durante minha qualificação pela UNESP.

À Joice Matos Biselli Périco, pelo apoio técnico e presença amiga durante o desenvolvimento do projeto.

Aos funcionários da seção de Pós-Graduação do IBILCE/UNESP, em especial à Silvinha, à Rosemar e ao Alex, que me socorreram e sempre estiveram lá por mim, muito obrigada.

A todos, absolutamente todos os funcionários do IBILCE/UNESP, que me presentearam por todos esses quase nove anos com 'bom dia's, sorrisos e boas conversas, recebam o meu muito obrigada, vocês fizeram meus dias muito mais agradáveis. Em especial ao pessoal do polo computacional que tantas vezes me ajudaram.

Aos meus alunos que continuaram a me ensinar ao longo desses anos, mesmo eu estando afastada das salas de aula.

"Existem pessoas que tornam nossa caminhada mais significativa…
pela companhia…
pelo apoio…
pelo carinho…
e porque nos tornam melhores…"

Aos meus amigos que ao longo da minha existência me estruturaram e me mantiveram de pé...

A Andres Gutierrez, Andres Mauricio Villegas Hincapie, Christine Tranchant-Dubreuil, Clémence Hatt, Emira Cherif, Julie Orjuela, Julissa Ronal, Karina Castillo, Kim Fooyontphanich, Kadidia Koita, Laura Mihaela Stefan, Mohammad Salma, Nhung Ta, Roberto Bobadilla Landey, Zvjezdana Markovic, vocês foram minha família por um ano e serão meus amigos para sempre... obrigada por tudo que vivi ao lado de vocês e pelos inúmeros sorrisos que compartilhamos. Julie e Gutierrez, obrigada pela companhia nos jogos de futebol. Vocês me tornaram uma pessoa melhor. Sinto falta de todos vocês.

"[…] Oh, did you want me to change?
Well I'd change for good,
And I want you to know that you'll always get your way
And I wanted to say […]"

Shiver - Coldplay

Aos meus amigos do laboratório de Evolução Molecular, Adriana Granzotto, Camila Vieira, Maryanna Simão, Raduan Soleman, Marjorie Silva, Priscila Pasquetto, Yuri C. Grandinete e Wellington Santos, obrigada por simplesmente estarem lá e tornarem tudo muito mais legal. Dri, seria impossível, em algumas palavras, agradecer tudo o que já fez por mim, obrigada por estar lá e por voltar quando precisei, muito obrigada, Amiga. Mary, obrigada pelas risadas e pelo carinho. Camila, obrigada pelo otimismo e pela força nesses últimos tempos.

À Juliana Faria dos Santos, Fernanda Martins e Esther Salgueiro, as quatro mosqueteiras, obrigada por todo o carinho e amizade, e por me darem meus sobrinhos Bia, minha princesa, Davi e o mais novo peixinho que vem por aí. Amo todas vocês e amo absurdamente esses pequenos.

A todos meus companheiros de turma e de pós-graduação, pelos breves encontros e sorrisos. Em especial ao Paulo Souza, pelos escassos cafés, obrigada pelo apoio e carinho, à Bianca Facchim pelas palavras de incentivo. Á Maysa Succi e Laila Toniol Cardin pelo carinho.

À Alessandra Paulino e Claudia Garcia, minhas amigas desde sempre, muito obrigada por tudo.

Às minhas primas e nossa pequena Ana Isabelle, obrigada por todo carinho. Aos meus tios e tias que torceram por mim. À Tia Antônia, que assumiu o papel de minha avó, e torceu e rezou por mim.

À Juliana, minha irmã, aqui, definitivamente me faltam palavras... minha irmã, amiga, parceira, exemplo... amo-te, obrigada por estar sempre ao meu lado.

Aos meus pais, Dora e Gilberto, obrigada por tudo, por, como já disse anteriormente, fazerem dos estudos o meu caminho e não o meu destino, por toda compreensão diante de minha ausência, por me apoiarem incondicionalmente e por fazerem me sentir amada, sempre. Muito, muito obrigada.

Por fim, a Deus, por me dar o direito de escolha... espero ter mais acertado do que errado até aqui.

“Se você quer construir um navio, não angarie homens para juntar madeira ou atribua-lhes tarefas e trabalho, mas sim os ensine a desejar a infinita imensidão do oceano.”

Antoine de Saint-Exupery

RESUMO

A história evolutiva das angiospermas é marcada por rápida e ampla diversificação, cujo *background* deve-se à ação de numerosos fatores; dentre esses, os elementos de transposição (TEs) têm sido considerados como um dos agentes mais importantes. TEs podem compor grandes porções do genoma de plantas e, dessa forma, desempenhar um papel importante na promoção da diversidade genética. O objetivo deste estudo foi investigar a dinâmica evolutiva de TEs ativos em espécies do gênero *Coffea*. Uma ampla análise da distribuição e evolução de um retrotransposon com LTR (LTR-RT), o elemento *Copia25*, foi realizada em genomas de plantas. *Copia25* é amplamente distribuído na família Rubiaceae, e, está presente em espécies distantemente relacionadas pertencentes às subclasses Asteridae e Rosidae, e à classe das monocotiledôneas. Em particular, foi observada uma incongruência envolvendo sequências *Copia25* de espécies do gênero *Musa*, uma monocotiledônea, e do gênero *Ixora*, uma dicotiledônea (Rubiaceae), que seria devido a um evento de transferência horizontal (HT) entre essas espécies ou entre suas linhagens ancestrais. *Copia25* apresenta dinâmica evolutiva complexa em angiospermas, cuja história incluiria conservação de sequências, perdas estocásticas e HT. Dez LTR-RTs foram anotados no genoma de *C. canephora* e tiveram seus perfis insercionais obtidos, usando os métodos de IRAP e REMAP, em genótipos das espécies progenitoras, *C. canephora* e *C. eugenioides*, e do alotetraploide, *C. arabica*. Perdas de inserções teriam ocorrido no alotetraploide, sendo essas mais significativas em cinco das dez famílias investigadas, e observou-se, ainda, a ocorrência de alterações direcionais nos subgenomas, sendo mais frequentes as ocorridas no subgenoma maternal, *C. eugenioides*. O presente trabalho contribui para o entendimento da evolução dos LTR-RTs nos genomas, da colonização de novos genomas por esses elementos, bem como, da sua dinâmica evolutiva em um genoma recém-originado.

Palavras-chave: Elementos de transposição. Reorganização genômica. Conservação de sequências. Transferência horizontal. Plantas com flor. Alopoliploide. Café.

ABSTRACT

The evolutionary history of the angiosperms is characterized by its rapid and broad diversification, whose background is due to the action of numerous factors; among them, the transposable elements (TEs) have been considered as one of the most important agents. TEs might compose large portions of the plant genomes, and play an important role in promoting genetic diversity. The aim of this study was to investigate the evolutionary dynamics of active TEs in the Coffea *species. In the first chapter, are presented the results of an extensive analysis of the distribution and evolution in plant genomes of a retrotransposon with LTR (LTR-RT), the* Copia25*.* Copia25 *is widely distributed in the Rubiaceae family, and is present in distantly related species belonging to the Rosidae and Asteridae subclasses, and the class of monocotyledons. In particular, it was observed an incongruity involving* Copia25 *sequences of* Musa *species, a monocot, and* Ixora *species, a dicot (Rubiaceae), which could be due to horizontal transfer (HT) between these species or their ancestral lineages.* Copia25 *has a complex evolutionary dynamics in angiosperms, whose history could include conservation sequences, stochastic loss and HT. Ten LTR-RTs were annotated in* C. canephora *genome and had their insertional profiles obtained, using IRAP and REMAP methods, in genotypes of the parental species,* C. canephora *and* C. eugenioides*, and the allotetraploid,* C. arabica*. Losses of insertions could have occurred in the allotetraploid, these being more significant in five out of ten families studied, and also was observed the occurrence of directional changes in progenitors subgenomes, being more frequent those occurred in maternal subgenome*, C. eugenioides. *This study contributes to the understanding of the evolution of LTR-RTs within genomes, the colonization of new genomes for these elements as well as its evolutionary dynamics in a newly originated genome.*

Keywords: Transposable elements. Genomic reorganization. Sequence conservation. Horizontal transfer. Flowering plant. Allopolyploidy. Coffee.

LONG RÉSUMÉ

Les éléments transposables (ET, ou TEs en Anglais, pour transposable elements) sont des séquences d'ADN répétées capables de se mouvoir d'un endroit à un autre dans un génome. Ils peuvent accroître leur nombre de copies lors de ce processus. A de rares exceptions près, les ET se trouvent dans tous les génomes d'eucaryotes analysées jusqu'ici mais aussi dans celui de certains procaryotes. On les trouve des bactéries jusqu'aux êtres humains. De part leur intervention dans la modulation des génomes, leurs capacités évolutives sont très importantes (KIDWELL; LISCH, 2000; 2001; GOTEA; MAKALOWSKI, 2006; MAKALOWSKI; TODA, 2007). Les rétrotransposons à LTR (en Anglais, Long Terminal Repeat) sont les éléments les plus couramment trouvés chez les eucaryotes. En raison de leur mécanisme de transposition, qui suit un modèle « copier-coller », les rétrotransposons connaissent une phase de transcription en ARN qui peut être lui-même reverse transcrit et l'ADN en résultant pourra être inséré en nouveau site du génome. Ce model réplicatif peut rapidement accroître le nombre de copies de l'élément, et par conséquent entrainer une augmentation de la taille du génome (KUMAR, 1996; KUMAR et al., 1997; SANMIGUEL; BENNETZEN, 1998). Les rétrotransposons sont omniprésents dans les génomes de toutes les plantes. Ils peuvent constituer jusqu'à 80% du génome de ces organismes (TENAILLON et al., 2011).

Le genre *Coffea*, appartenant à la famille des Rubiacées, comporte 125 espèces reconnues. Parmi elles deux sont largement cultivées, *Coffea canephora* (qui donne le café Robusta) et *C. arabica* (café Arabica). L'espèce *C. arabica* est la seule du genre a être tétraploïde (2n = 4x = 44). Le génome de *C. canephora* a été récemment séquencé, il a été montré que les ET constituent environ 50% de son génome. Parmi eux, 80% sont des rétrotransposons (DENOEUD et al., 2014). L'espèce *C. arabica* est issue d'un croisement relativement récent, il y a moins d'un million d'années, entre *C. canephora* et *C. eugenioides*, une espèce sauvage d'Afrique de l'Est (LASHERMES et al., 1999; YU et al., 2011). *C. canephora* et *C. eugenioides* auraient divergé il y a environ 4,2 millions d'années au maximum (YU et al., 2011).

La transmission des ET se produit généralement verticalement et leurs histoires évolutives peuvent être marquées chez différentes espèces entre autres par des pertes. Dans quelques cas, ces séquences peuvent aussi être transmises horizontalement et coloniser de nouveaux génomes. Grâce à une analyse de séquences 454 (shotgun) et d'extrémités de clones BAC de *C. canephora*, des rétrotransposons présents dans ce génome ont pu être reconstruits artificiellement et identifiés. Parmi ces éléments, un, dénommé *Copia25*, a montré une identité étonnement élevée avec un élément de la tomate (*Rider*; JIANG et al., 2009). Une analyse approfondie de ce rétrotransposon toujours actif chez *C. canephora* dans 41 génomes de plantes, montre sa présence avec une identité de séquence étonnement élevée et ce aussi bien dans des espèces dicotylédones (Astérides et Rosides) que monocotylédones groupes ayant divergé il y a environ 150 millions d'années. Ce résultat montre bien la dynamique complexe de l'évolution de cet élément ancien, qui prédate la divergence des deux grands groupes des angiospermes. Cette dynamique évolutive doit faire appel à plusieurs processus dont la conservation de séquences, un renouvellement rapide, des pertes stochastiques et des transferts horizontaux.

Une situation particulière concerne l'identité remarquable de *Copia25* entre les espèces du genre *Musa* (monocotylédone-Zingiberidae) et celles du genre *Ixora* (dicotylédone-Asteridae). Le genre *Ixora* fait également partie de la famille des Rubiacées. De part la position dans l'arbre phylogénétique de la séquence de *Copia25* de *Musa* dans le clade des Rubiacées au voisinage d'*Ixora*, la seule hypothèse pouvant expliquer cette proximité est celle du transfert horizontal entre ces deux genres dont les ancêtres partageaient la même région géographique (Asie du Sud-Est) entre 30 et 50 millions d'années (LIU et al., 2010; CHRISTELOVÁ et al., 2011; LORENCE et al., 2007; TOSH et al., 2013). Ces résultats constituent le premier chapitre de cette thèse sous forme d'un manuscrit qui a été soumis à la revue « Plant Molecular Biology » et qui se trouve dans sa phase d'approbation finale après avoir été accepté avec modifications en première lecture.

Des amplifications et des pertes de familles d'ET ont été observées au cours de la réorganisation de génomes suivant des évènements de polyploïdisation résultants soit de duplications du génome soit d'hybridations (PARISOD; SENERCHIA, 2012). De tels évènements de polyploïdisation (avec ou sans hybridations) sont des facteurs qui contribuent à la diversification des angiospermes (BAACK; RIESEBERG, 2007; SOLTIS; SOLTIS, 2009; JIAO et al., 2011). Une allopolyploidisation peut se caractériser par des événements qui se produisent avant, comme l’absence de réduction chromosomique à la méiose, et/ou après l’hybridation comme la duplication du génome hybride. Les premières générations qui suivent la polyploïdisation sont caractérisées par une grande instabilité génomique qui s’accompagne d’une réorganisation structurelle et où l’épigénétisme joue un rôle certain. Au cours des générations suivantes, la structure du génome peut tendre vers une restauration de l’état diploïde. Au cours de ces générations on assiste à l’accumulation de mutations ponctuelles et à des réarrangements structuraux impliquant des régions homologues (PARISOD; SENERCHIA, 2012; CHANG et al., 2010). Les ET sont associés à deux événements majeurs après les événements de plolyploïdisation, les modifications épigénétiques et la réorganisation structurale du génome (MCCLINTOCK, 1984).

Les résultats de l'analyse comparative du polymorphisme d’insertion de 10 rétrotransposons à LTR (LTR-RT) dans le génome de *C. arabica* et dans celui des espèces parentales, suggèrent une restructuration du génome accompagnée d’une perte sélective d’ET chez l’allotetraploïde. Aucune des familles d’ET incluses dans cette analyse n’a présenté d'amplification dans le génome de l’hybride, ce qui suggère que la polyploïdisation ne les a pas réactivés. Le contrôle épigénétique n’aurait pas été affecté. Les résultats obtenus semblent plutôt indiquer une perte de copies chez l’allotétraploïde. Cette perte serait plus importante pour cinq des dix familles d’ET étudiées. Ils indiquent également la présence de modifications spécifiques en fonction des sous-génomes. Le sous-génome provenant de *C. eugenioides*, serait plus souvent impliqué dans cette réorganisation que celui provenant de *C. canephora*. La réorganisation du génome lors des premières générations, que certains auteurs appellent changements révolutionnaires, serait dépendante de la distance génétique séparant les deux espèces parentales. L’alloploïdisation se produisant plus souvent entre espèces proches comme *C. eugenioides* et *C. canephora* (LASHERMES et al., 1999).

Mis à part la réorganisation génomique, les résultats obtenus permettent de tirer des conclusions sur l'évolution des 10 familles de LTR-RTs chez trois espèces de

Coffea. *C. canephora* l'espèce originellement la plus répandue en Afrique, présente une structuration génétique en au moins 7 groupes distincts (GOMEZ et al., 2009; PONCET *com. pers*). Toutefois, les résultats provenant de l'analyse du polymorphisme des sites d'insertion de familles de LTR-RTs par Analysées en Coordonnées Principales (PCoAs), montrent une population homogène incluant les génotypes des deux espèces parentales, *C. eugenioides* et *C. canephora* avec les génotypes de *C. arabica* formant un groupe distinct pour la plupart des familles de ETs. Sauf si l'on considère les possibilités de transfert horizontal et/ou de pertes stochastiques. Les ET sont transmis très généralement verticalement. Il sont hérités d'un ancêtre commun et sont présents, ou non, dans un génome actuel en raison d'une série de facteurs liés à l'élément lui même, à l'hôte et à l'interaction ET-hôte (LE ROUZIC et al., 2007). La tribu Coffeae a divergé relativement récemment, environ 15 millions d'années (BREMER; ERICKSON, 2009) quant aux espèces parentales de *C. arabica*, elles n'auraient divergé qu'il y a 4,2 millions d'années au grand maximum, et 0.6 million d'année au minimum (YU et al., 2011). Le groupe homogène formé par les espèces *C. canephora* et *C. eugenioides* quant à la distribution des 10 familles de LTR-RTs analysés, pourrait résulter de la récente divergence de ces espèces. Le patron de distribution des sites d'insertion observé de nos jours résulterait de la situation préexistante chez l'ancêtre commun. L'estimation de l'âge d'insertion des différents éléments dans les espèces parentales et les résultats fournis par les réseaux complexes renforcent cette suggestion. Ces populations LTR-RTs, en raison de la récente divergence, ne seraient pas ont fusionné pour former les populations isolées. Les résultats obtenus par l'analyse des 10 familles de LTR-RTs constituent un manuscrit qui sera prochainement soumis et constituent le deuxième chapitre de cette thèse.

L'ensemble des résultats présentés dans cette thèse contribue à la compréhension de l'évolution et la dynamique des LTR-RTs dans le génome de quelques espèces du genre *Coffea* mais aussi plus généralement chez les angiospermes. Ils décrivent des mécanismes de colonisation de nouveaux génomes par ces éléments et leur dynamique d'évolution dans un génome nouvellement formé.

References

BAACK, E. J.; RIESEBERG, L. H. A genomic view of introgression and hybrid speciation. **Current Opinion in Genetics & Development**, v. 17, n. 6, p. 513-518, 2007.

BREMER, B.; ERIKSSON, T. Time tree of Rubiaceae: Phylogeny and dating the family, subfamily, and tribes. **International Journal of Plant Sciences**, v. 170, n. 6, p. 766-793, 2009.

CHANG, P. L. et al. Homoeolog-specific retention and use in allotetraploid Arabidopsis suecica depends on parent of origin and network partners. **Genome Biology**, v. 11, n. 12, p. R125, 2010.

CHRISTELOVA, P. et al. A multi gene sequence-based phylogeny of the Musaceae (banana) family. **Bmc Evolutionary Biology**, v. 11, p. 103, 2011.

DENOEUD, F. et al. The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. **Science**, v. 345, n. 6201, p. 1181-4, 2014.

GOMEZ, C. et al. 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 Evolutionary Biology**, v. 9, p. 167, 2009.

GOTEA, V.; MAKALOWSKI, W. Do transposable elements really contribute to proteomes? **Trends in Genetics**, v. 22, n. 5, p. 260-267, 2006.

JIAO, Y. et al. Ancestral polyploidy in seed plants and angiosperms. **Nature**, v. 473, n. 7345, p. 97-100, 2011.

JIANG, N. et al. Genome organization of the tomato sun locus and characterization of the unusual retrotransposon Rider. **The Plant Journal**, v. 60, p. 181–193, 2009.

KIDWELL, M. G.; LISCH, D. R. Perspective: Transposable elements, parasitic DNA, and genome evolution. **Evolution**, v. 55, n. 1, p. 1-24, 2001.

KIDWELL, M. G.; LISCH, D. R. Transposable elements and host genome evolution. **Trends in Ecology & Evolution**, v. 15, n. 3, p. 95-99, 2000.

KUMAR, A. et al. The Ty1-copia group of retrotransposons in plants: genomic organisation, evolution, and use as molecular markers. **Genetica**, v. 100, n. 1-3, p. 205-217, 1997.

KUMAR, A. The adventures of the Ty1- copia group of retrotransposons in plants. **Trends In Genetics**, v. 12, n. 2, p. 41–43, 1996.

LASHERMES, P.et al. Molecular characterisation and origin of the Coffea arabica L. genome. **Molecular and General Genetics**, v. 261, n. 2, p. 259-266, 1999.

LE ROUZIC, A.; BOUTIN, T. S.; CAPY, P. Long-term evolution of transposable elements. **Proc. Natl. Acad. Sci. U.S.A.**, v. 104, p. 19375–80, 2007.

LIU, A.; KRESS, W.; LI, D. Phylogenetic analyses of the banana family (Musaceae) based on nuclear ribosomal (ITS) and chloroplast (trnL-F) evidence. **Taxon**, v. 59, n. 1, p. 20-28, 2010.

LORENCE, D. et al. Revision of Ixora (Rubiaceae) in the Marquesas Islands (French Polynesia). **Botanical Journal of The Linnean Society**, v. 155, n. 4, p. 581–597, 2007.

MAKALOWSKI, W.; TODA, Y. Modulation of host genes by mammalian transposable elements. **Genome dynamics**, v. 3, p. 163-74, 2007.

MCCLINTOCK, B. The significance of responses of the genome to challenge. **Science**, v. 226, n. 4676, p. 792-801, 1984.

PARISOD, C.; SENERCHIA, N. Responses of transposable elements to polyploidy. In: **Plant Transposable Elements** (Eds. Grandbastien & Casacuberta), Topics in Current Genetics, Springer, v. 24, p. 147-168, 2012.

SANMIGUEL, P.; BENNETZEN, J. L. Evidence that a recent increase in maize genome size was caused by the massive amplification of intergene retrotransposons. **Annals of Botany**, v. 82, p. 37-44, 1998.

SOLTIS, P. S.; SOLTIS, D. E. The role of genetic and genomic attributes in the success of polyploids. **Proceedings of the National Academy of Sciences**, v. 97, n. 13, p. 7051-7, 2000.

TENAILLON, M. I. et al. Genome size and transposable element content as determined by high-throughput sequencing in maize and Zea luxurians. **Genome Biology and Evolution**, v. 3, p. 219-29, 2011.

TOSH, J. et al. Evolutionary history of the Afro-Madagascan Ixora species (Rubiaceae): species diversification and distribution of key morphological traits inferred from dated molecular phylogenetic trees. **Annals of Botany**, v. 112, n. 9, p. 1723-1742, 2013.

YU, Q. et al. Micro-collinearity and genome evolution in the vicinity of an ethylene receptor gene of cultivated diploid and allotetraploid coffee species (Coffea). **Plant Journal**, v. 67, n. 2, p. 305-17, 2011.

SUMÁRIO

1 INTRODUÇÃO

1 INTRODUÇÃO GERAL

1. 1 Elementos de Transposição

Elementos de transposição (TEs, do Inglês, Transposable Elements) são sequências repetitivas de DNA capazes de se mobilizarem de um local para outro no genoma e de aumentarem seu número de cópias durante esse processo. Com raras exceções, os TEs são encontrados em todos os genomas eucariotos analisados até então, desde bactérias até humanos, e em cerca de 80% dos procariotos analisados (TOUCHON; ROCHA, 2007); e constituem forças evolutivas importantes na modulação dos genomas (KIDWELL; LISCH, 2000; 2001; GOTEA; MAKALOWSKI, 2006; MAKALOWSKI; TODA, 2007). Essas sequências repetitivas podem compor grande parte do genoma dos organismos, variando entre 1% (fungo *Fusarium graminearum*, CUOMO et al., 2007) até 85% (milho *Zea mays*, TENAILLON et al., 2011), constituindo uma parcela importante do genomas de plantas.

Os elementos de transposição são agrupados em duas Classes que se diferenciam de acordo com a presença ou não de uma etapa de transcrição reversa durante sua mobilização. Os elementos de Classe I (Retrotransposons) realizam essa etapa, assim, o RNA intermediário é transcrito reversamente e a nova molécula é inserida no genoma (mecanismo *copy-and-paste*). Já os elementos de Classe II (Transposons de DNA) clivam a fita dupla do DNA e a reinserem em outro local do genoma (mecanismo *cut-and-paste*). Os elementos de Classe I podem ser agrupados em cinco ordens cujos representantes se diferenciam em dois grupos quanto à presença ou não de longas repetições terminais diretas, as LTRs (do Inglês, Long Terminal Repeats) – os Retrotransposons com LTRs (LTR-RTs), os elementos DIRS-*like* e os Penelope-*like* – e aqueles sem as repetições terminais (non-LTRs) – os LINEs (do Inglês, Long Interspersed Nuclear Elements) e os SINEs (do Inglês, Short Interspersed Nuclear Elements). Os elementos da Classe II, por sua vez, são agrupados em duas subclasses, dependendo do número de fitas de DNA clivadas durante a mobilização: Subclasse I, os elementos que clivam a fita dupla de DNA (Ordens TIR e Crypton), e Subclasse II, os que se movem a partir do deslocamento da fita simples de DNA formando um *loop* com posterior clivagem e reintegração no genoma, os Helitrons, e os que se movem a partir da fita simples de DNA durante a duplicação, os Mavericks (WICKER et al., 2007) (Figura 1).

Adicionalmente à classificação baseada no mecanismo de mobilização, os TEs também podem ser classificados quanto à sua autonomia nesse processo. Os autônomos, que produzem as proteínas responsáveis por sua mobilização (mobilização *in cis*), e os não autônomos, que perderam a capacidade de produzir essas proteínas e dependem daquelas produzidas pelos TEs autônomos (mobilização *in trans*). Elementos não autônomos estão presentes em ambas as Classes, I e II. Dentro da Classe I, os LARDs (do Inglês, Large Retrotransposon Derivatives) são elementos que apresentam grandes LTRs e grande região interna com os *cores* proteicos, mas não produzem as proteínas para sua mobilização (KALENDAR et al., 2004); os TRIMs (do Inglês, Terminal-Repeat Retrotransposons In Miniature), que apresentam LTRs pequenas e perderam completamente os domínios proteicos internos (WITTE et al., 2001); e os SINEs, dentre os retrotransposons sem LTRs, que são derivados de eventos de retrotransposição de transcritos de RNA (normalmente tRNAs) e podem ou não apresentar na região 3’ sequência similar a regiões 3’ de elementos LINEs relacionados (KRAMEROV; VASSETZKY, 2005). Os MITEs (do Inglês, Miniature Inverted Repeat Transposable Elements) são elementos não autônomos presentes na Classe II. Eles são geralmente derivados de transposons autônomos, tendo sido originados de deleções internas ocorridas nesses, remanescendo apenas as porções terminais invertidas, e, em alguns casos, regiões internas não codificantes presentes entre as TIRs (do Inglês, Inverted Terminal Repeats) (FESCHOTTE et al., 2002).

Em cada ordem, os TEs podem ainda ser agrupados em superfamílias e famílias. As superfamílias são classificadas de acordo com a homologia compartilhada em nível proteico (WICKER, 2012) e com a organização proteica ou domínios não codificantes, bem como outras características estruturais, como a presença e o tamanho dos TSDs (do Inglês, Target Site Duplications) (WICKER et al., 2007). As famílias, por sua vez, são definidas pela conservação na sequência de DNA de porções de regiões codificantes, sendo classificadas, dentro de uma mesma família, sequências que compartilham ao menos 80% de identidade nucleotídica sobre 80% do tamanho total de uma sequência de ao menos 80 pb. Essa classificação é conhecida como regra 80-80-80, proposta por Wicker e colaboradores em 2007 (do Inglês, “80-80-80” rule, WICKER et al., 2007). Biologicamente, o compartilhamento de 80% de identidade sobre 80% da extensão sugere que as sequências envolvidas teriam sido originadas de uma cópia mãe ancestral em um tempo evolutivo recente (WICKER, 2012). Agrupamentos em níveis inferiores à família são identificados com base em sua clara segregação em análises filogenéticas (WICKER et al., 2007).

Os retrotransposons com LTRs são os elementos mais comumente encontrados nos eucariotos. As duas principais superfamílias são *Ty1/Copia* e *Ty3/Gypsy*, que se segregam tanto pela homologia de suas regiões proteicas quanto pelo arranjo desses domínios na ORF *pol* (do Inglês, Open Read Frame; *pol*, poliprotein). Ambas as superfamílias apresentam duas ORFs, a *gag*, responsável por codificar as proteínas estruturais do capsídeo, e a *pol*, uma poliproteína que codifica as enzimas responsáveis pela mobilização do LTR-RT, tais como protease aspártica (AP), integrase (INT), transcriptase reversa (RT) e RNase-H (RH). Essas superfamílias diferem no arranjo dos domínios RT e INT. A protease aspártica é responsável pelo processamento pós-tradução da poliproteína inicial que resulta em dois polipeptídeos, um com a RT e a RH, responsável pela transcrição reversa, e um da INT, que será responsável por inserir a nova cópia no genoma (SABOT; SCHULMAN, 2006). Alguns elementos, relacionados à superfamília *Ty3/Gypsy*, apresentam uma terceira ORF, *env*, que codifica uma proteína do envelope similar às presentes no retrovírus. Devido ao seu mecanismo de transposição, os retrotransposons geram um grande número de moléculas de RNAs que, quando transcritas reversamente, podem representar uma nova cópia em potencial. Esse modo replicativo pode rapidamente aumentar o número de cópias do elemento, o que, por sua vez, pode aumentar o tamanho do genoma (KUMAR, 1996; KUMAR et al., 1997; SANMIGUEL; BENNETZEN, 1998).

Retrotransposons são ubíquos em plantas e podem constituir até 75% do genoma desses organismos (TENAILLON et al., 2011). A superfamília *Ty1/Copia* ocorre amplamente desde algas unicelulares até plantas superiores e elementos da superfamília *Ty3/Gypsy* são encontrados em gimnospermas e angiospermas (KUMAR; BENNETZEN, 2003). O número de cópias de elementos pertencente a cada superfamília varia de apenas poucas cópias, como do *Ty1/Copia Bs1* (JIN; BENNETZEN, 1989), que apresenta entre 1 a 5 cópias em milho; até mais de 50.000 cópias, como o *BARE-1*, outro *Ty1/Copia*, em cevada. Tal variação também é observada para a superfamília *Ty3/Gypsy*, desde 10 inserções do elemento *RIRE3* em arroz (KUMEKAWA et al., 1999) a até cerca de 20.000 cópias do retrotransposon *Cinful-1* em milho (SANMIGUEL; BENNETZEN, 1998). Essa variação no número de cópias ocorre até mesmo em um único elemento em espécies relacionadas, como o *Ogre*, um *Ty3/Gypsy*, que varia de apenas cerca de 100 cópias em *Vicia faba* até cerca de 100.000 inserções em *V. pannonica* (HODSON; BRYANT, 2012).

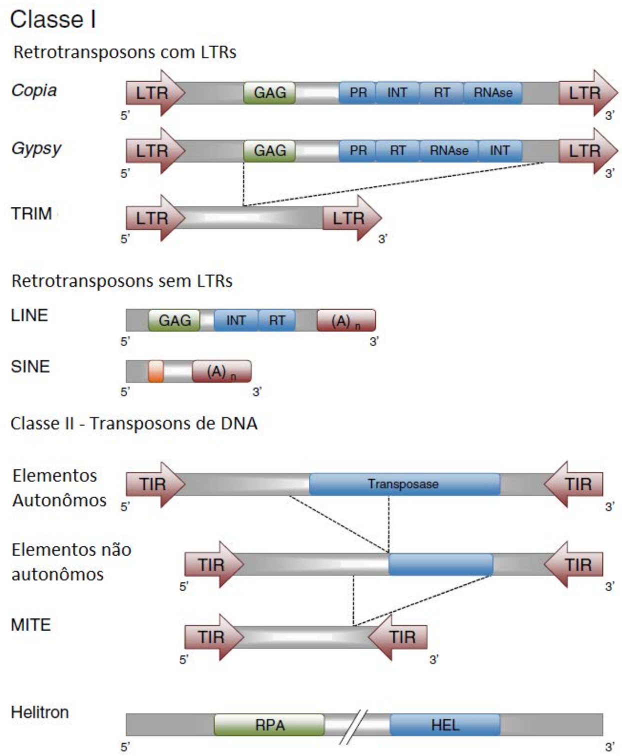


Figura 1. Esquema da classificação simplificada dos TEs. Adaptado de Parisod et al. (2009).

1. 2 Conservação de sequências

Os TEs, constituindo parte do aparato genômico do hospedeiro, são transferidos verticalmente, sendo transmitidos, portanto, dos parentais para os descendentes. Uma vez presentes em um genoma, os TEs podem persistir por inúmeras gerações e, inclusive, estarem sujeitos a eventos que levam à sua divergência (MARUYAMA; HARTL, 1991). Durante o processo de replicação, sobretudo dos retrotransposons, erros durante a transcrição, realizada pela RNA polimerase, e a transcrição reversa, realizada pela transcriptase reversa, resultam em cópias similares, mas não idênticas à cópia mãe. Esse acúmulo de mutações ao longo do tempo, associado a outras mutações estruturais as quais os TEs estão sujeitos, como *indels* e recombinação ilegítima, levam a sua inativação em longo prazo (BROOKFIELD; BADGE, 1997; PINSKER et al., 2001). Além disso, não é esperado que cópias individuais de retrotransposons estejam sob ação de seleção natural e sejam mantidas ao longo do tempo. Em uma cópia ativa, mutações que levam à sua inativação poderiam ser seletivamente neutras ou seletivamente favoráveis ao hospedeiro, por minimizar os efeitos mutagênicos da transposição, enquanto que mutações em cópias inativas seriam, presumivelmente, seletivamente neutras (LOHE et al., 1995). Entretanto, a seleção natural pode atuar ao longo da sequência de cópias ativas, sobretudo, nas regiões essenciais para a mobilização, como IN e RT, mantendo a sequência e, assim, o sucesso replicativo do elemento ao menos até o surgimento de um sistema de controle efetivo por parte do hospedeiro (JORDAN; MCDONALD; 1998). Como resultado desse sucesso, ter-se-ia a formação de metapopulações de TEs (HANSKI, 1998; LE ROUZIC et al., 2007).

A mobilização, evolução e permanência de um TE em um genoma estão associadas a diversos fatores – como taxa de mutação, taxa de recombinação, sistema de controle epigenético, deriva e seleção natural, relacionados tanto ao TE quanto ao hospedeiro – fazendo com que diferentes famílias de TEs em um mesmo genoma e uma mesma família em genomas diferentes estejam em fases distintas de sua evolução. Dessa forma, a dinâmica dos TEs em longo prazo é afetada por fatores do TE, do hospedeiro e daqueles oriundos da interação TE-hospedeiro (LE ROUZIC et al., 2007). Relatos recentes de alta conservação envolvendo sequências de TEs entre grupos distantemente relacionados exemplificam essa complexidade. Se por um lado a conservação e a ampla presença do TRIM *Cassandra* (desde samambaias até angiospermas), que apresenta um domínio RNA 5S em sua LTR – o qual lhe conferiria capacidade transcricional independente, e de evasão de silenciamento por metilação, devido à presença de um promotor pol III nesse domínio –, seria o resultado de sua exaptação por parte dos hospedeiros (KALENDAR et al., 2008); por outro lado, a conservação do LTR-RT *Tvv1*, em espécies que divergiram há pelo menos 100 milhões de anos, seria resultado da manutenção de uma replicação eficiente e da reduzida perda de sequências por recombinação do elemento (MOISY et al., 2014).

A conservação de sequências de um TE entre táxons, próximos ou distantemente relacionados, pode ser decorrente da transferência horizontal dessas sequências, posterior à divergência dos táxons a partir de seu ancestral comum. A identificação de eventos de HT constitui um desafio que tem sido nos últimos tempos transposto devido à disponibilização de sequências e à identificação de TEs em diversas espécies. Sendo transmitido verticalmente e compartilhando uma sequência ancestral comum, a história evolutiva de um TE, proposta por sua filogenia, deveria seguir a filogenia das espécies que o porta. Incongruências entre essas filogenias sugerem a ocorrência de eventos como introgressão, polimorfismo ancestral, domesticação ou transferência horizontal.

A sugestão de HT, além da incongruência filogenética, pode ser inferida quando ocorre uma identidade inesperada entre os TEs dos táxons analisados, que pode estar associada à distribuição irregular do TE entre os diferentes táxons (LORETO et al., 2008). Adicionalmente, mecanismos outros que poderiam explicar tais ocorrências – como polimorfismo ancestral, domesticação, conservação de sítios funcionais, taxa evolutiva similar entre as espécies envolvidas e seleção purificadora, as chamadas hipóteses alternativas de HT – devem ser ponderados e analisados (CAPY et al., 1994; CUMMINGS, 1994; SCHAACK et al., 2010; WALLAU et al., 2012). Outros pontos importantes devem ser levados em conta antes da sugestão de HT. Para que a hipótese de HT possa ser proposta, as espécies envolvidas devem ter sobreposição geográfica em um mesmo período de tempo e compartilhamento de nicho ecológico. Essa convivência resultaria em oportunidades para que transferências pudessem ocorrer. Dois mecanismos de transferência têm sido sugeridos para plantas: de modo direto, o planta-para-planta (do Inglês, *plant-to-plant*), envolveria principalmente plantas em relações parasíticas; e de modo indireto, como também sugerido para animais, mediada por um vetor. Neste, bactérias, fungos ou vírus poderiam capturar e transmitir uma sequência entre espécies, como também insetos e pássaros poderiam intermediar essa transferência. Contudo, embora plausíveis, esses cenários permanecem como especulação, sendo poucos os casos descritos onde foi possível identificar os organismos envolvidos (Figura 2).

Embora os números de transferência horizontal de TEs propostos tenham aumentado nos últimos anos, há um desequilíbrio no que concerne aos organismos envolvidos nesses eventos, sendo, a maioria, em animais, e em plantas, um número reduzido (DIAO et al., 2006; FORTUNE et al., 2008; ROULIN et al., 2008; CHENG et al., 2009; EL BAIDOURI et al., 2014). Parte desse desequilíbrio pode ser explicada pela disponibilidade bem mais recente de genomas de plantas do que de animais, e pela complexidade dos genomas das mesmas, o que dificulta a comparação em larga escala que viabiliza a proposição de HT. Não obstante, os genomas de plantas possuem uma natural propensão à transferência de material genético devido à facilidade de intercruzamento e à autonomia de sua linhagem germinativa. Além disso, o grande conteúdo de TEs, sequências naturalmente móveis, e particularmente, os retrotransposons – que apresentam uma molécula intermediária estável (fita dupla de DNA), realizam parte de seu ciclo de vida no citoplasma, e, em alguns casos, codificam uma proteína envelope-*like* que pode lhes conferir capacidade infectante similar a dos retrovírus – viabilizaria a troca e a captura de material genético, aumentando a oportunidade de ocorrência de HTs.

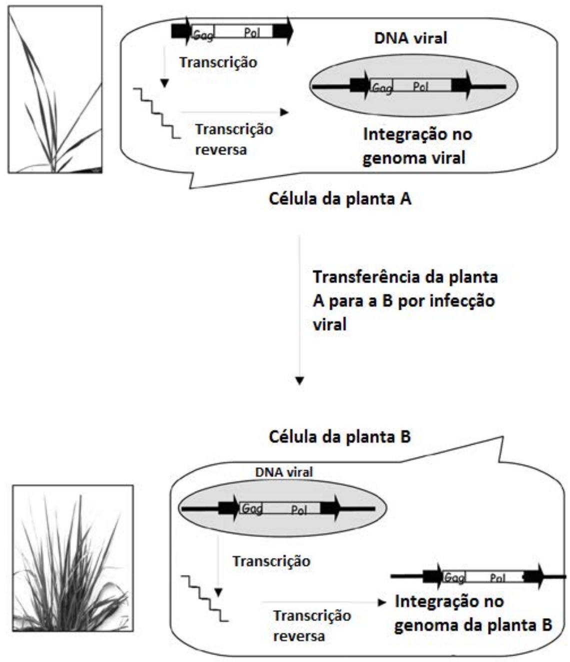


Figura 2. Transferência horizontal mediada por um vetor. Adaptado de Fortune et al. (2008).

1. 3 Alopoliploidização

O fato de as plantas terem facilidade em promover intercruzamento, além de propiciar a troca de material genético também pode resultar no aumento da diversidade biológica com o surgimento de novas espécies. A tolerância à hibridização e à poliploidia (com ou sem a hibridização) são fatores que levaram a diversificação das angiospermas (BAACK; RIESEBERG, 2007; SOLTIS; SOLTIS, 2009; JIAO et al., 2011). Em híbridos, pode ocorrer a introgressão de genes e de TEs, e seus genomas podem se tornar estáveis a ponto de originar uma nova espécie, sobretudo, se a hibridização for acompanhada de poliploidia (OLIVER et al., 2013). Estima-se que entre 30 e 80% das espécies de angiospermas tenham passado por um ou mais eventos de poliploidização (STEBBIS, 1947; MASTERSON, 1994; WENDEL, 2000) que estariam envolvidos em ao menos 15% dos eventos de especiação (WOOD et al., 2009).

A poliploidia se caracteriza pela presença de mais de dois genomas por célula, sendo os alopoliploides originados a partir de um evento de hibridização interespecífico seguido por duplicação cromossômica (não necessariamente nessa ordem) (OLIVER et al., 2013). Com o evento de alopoliploidização, tem-se a formação de uma nova espécie, visto que os híbridos originados são, com frequência, reprodutivamente isolados de seus progenitores devido a diferenças na ploidia. Geneticamente, os alopoliploides se caracterizam por heterozigosidade fixa (um locus homólogo dominante em um subgenoma parental e recessivo no outro é obrigatoriamente heterozigoto no gameta do híbrido), formação de bivalentes e segregação dissômica durante a meiose, com os cromossomos homólogos pareando-se preferencialmente entre si (SOLTIS; SOLTIS, 2000). Três principais pontos conferem ao poliploide alguma vantagem: a heterosigozidade, a redundância gênica e a perda da autoincompatibilidade e ganho de reprodução assexual. A heterozigosidade, que nos diploides tende a diminuir ao longo do tempo devido a recombinações entre cromossomos homólogos, tende a se manter no alopoliploide devido à segregação de dissômica que previne a recombinação entre homeólogos (cromossomos derivados das diferentes espécies progenitoras e que são relacionados por compartilharem uma ancestralidade). A redundância gênica gerada, por sua vez, além de reduzir a frequência de homozigotos recessivos, também aumenta a diversidade gênica por disponibilizar cópias que evoluem sem afetar o *background* original, podendo resultar em um aumento da diversidade funcional. A autoincompatibilidade, em alguns casos de poliploidia, é desfeita nos poliploides que passam a ser autoférteis, ou seja, ocorre a produção de semente como resultado da polinização pelo pólen da mesma flor ou de flores do mesmo indivíduo (COMAI et al., 2005). Contudo, algumas desvantagens também são associadas à poliploidia, como os efeitos prejudiciais do aumento nuclear e celular, a propensão de produzir células aneuploides (células com um número cromossômico diferente do normal da espécie), bem como a instabilidade epigenética, que influencia na regulação gênica (COMAI et al., 2005). Embora os poliploides frequentemente sofram o efeito de gargalo de garrafa (*bottleneck*) devido a essas dificuldades (COMAI et al., 2005), e muitas linhagens recém-formadas falhem em persistir – analisando neopoliploides observa-se que poliploides apresentam taxa de diversificação inferior a dos diploides, ou seja, a taxa de extinção é superior a de especiação –, eles possuem potencial para obter sucesso evolutivo em longo prazo (MAYROSE et al., 2011).

A alopoliploidização pode ser caracterizada por quatro diferentes estágios que envolvem etapas que acontecem em momentos anteriores e posteriores à formação do híbrido. No estágio 1, ocorre a divergência entre as espécies parentais, com ambas adaptando-se ao seu ambiente específico e adotando estratégias próprias de cruzamento e reprodução. Nessa fase, a seleção direcional pode contribuir para a fixação de mutações espécie-específicas em regiões regulatórias e codificantes, enquanto que mutações deletérias podem ser fixadas por deriva. Nos estágios 2 e 3, as espécies divergentes hibridizam e aumentam a ploidia, o que permite o pareamento correto durante a meiose. A hibridização, em geral, resulta em instabilidade fenotípica, rearranjos genômicos generalizados, silenciamento epigenético e *splicing* alternativo. Nessa fase, o híbrido passa por um rápido ajuste intragenômico com mudanças em um curto prazo ocorrendo imediatamente após a poliploidização – essas seriam mudanças revolucionárias. Por outro lado, o estágio 4 é marcado pela evolução em longo prazo dos genes homeólogos e caracteriza-se por ocorrer mais lentamente na escala de tempo evolutivo – seriam mudanças evolutivas (Figura 3). Embora os rearranjos cromossômicos e a reprogramação epigenética possam afetar ambos subgenomas, na maioria dos casos, eles afetam diferencialmente, e o acúmulo dessas ocorrências pode aumentar a divergência entre os subgenomas (FELDMAN; LEVY, 2002; CHANG et al., 2010; PARISOD; SENERCHIA, 2012) (Figura 3). Essa preferência por um dos subgenomas sugere que as interações núcleo-citoplasma representam incompatibilidades cruciais a serem superadas após a emergência do novo genoma, mas que seriam importantes (JOSEFSSON et al., 2006), sobretudo, resolvendo conflitos derivados das espécies geneticamente divergentes (PARISOD; SENERCHIA, 2012).

Devido à ampla presença de TEs nos genomas eucariotos, atingindo proporções particularmente grandes no genoma de plantas, é esperado que essas sequências tenham um importante papel nos eventos subsequentes à alopoliploidização. Os TEs são associados a duas principais ocorrências após eventos de alopoliploidização, alterações epigenéticas e reorganização genômica, e sua influência em ambas se entremeiam fazendo com que a distinção entre uma ou outra nem sempre seja clara. De acordo com a hipótese do “Genome Shock”, proposta por Barbara McClintock, o estresse genômico decorrente da hibridização, dentre diversos outros estresses que também poderiam resultar em respostas semelhantes do genoma, poderia reativar TEs silenciados e induzir reorganizações genômicas em um curto período de tempo que poderiam ser a base para a formação de novas espécies (MCCLINTOCK, 1984). Essa reorganização influenciaria o genoma por duas vias, na promoção de rearranjos genômicos, levando a perdas e duplicações, os quais, por sua vez, alterariam o contexto epigenético do genoma hospedeiro (TEIXEIRA et al., 2009; PARISOD; SENERCHIA, 2012).

Alterações epigenéticas poderiam reativar famílias de TEs e resultar em *bursts* (do Inglês, explosões) de transposição após a alopoliploidização. Eventos dessa natureza foram reportados em alotetraploides sintéticos de trigo, cuja alopoliploidização teria levado a reativação do retrotransposon Wis 2-1A, sugerida pela presença de transcritos no híbrido e pela ausência nos progenitores (KASHKUSH et al., 2002). Ocorrência similar, envolvendo um retrotratransposon *Ty1/Copia* e dois transposons de DNA da subfamília

Sunfish, foi reportada para alotetraploide sintético de *Arabdopsis thaliana* a partir de análises de *microarray*. Neste caso, observou-se que os elementos *Sunfish* seriam metilados nos parentais autotetraploides, mas demetilados e reativados nos alotetraploide (MADLUNG et al., 2005). Os TEs também estão relacionados à reorganização cromossômica após alopoliploidização. Devido ao seu caráter repetitivo, essas sequências podem atuar como substratos para recombinação (HEDGES; DEININGER, 2007), envolvendo tanto cromossomos homólogos (recombinação homóloga desigual) quanto somente pequenas regiões homólogas (recombinação ilegítima) (DEVOS et al., 2002). Restruturação e, em alguns casos, perda de porções de TEs foram reportadas em estudos que analisaram alopoliploides *Brassica* (GAETA et al., 2007), trigo (SHAKED et al., 2001), *Tragopogon* (TATE et al., 2006) e *Spartina* (PARISOD et al., 2009). Esses estudos mostram perdas de sequências de DNA, dentre essas TEs, nesses alopoliploides, sobretudo, em um curto período de tempo após o evento de hibridização. Os rearranjos genômicos que resultam nessa perda e reorganização estariam relacionados com a divergência das espécies envolvidas na hibridização. Os casos exemplificados envolveram espécies proximamente relacionadas, por outro lado, em eventos envolvendo espécies mais distantes essa reorganização não foi observada (JACKSON; CHENG, 2010) – alotetraploide sintético entre *A. thaliana* e *Cardaminopsis arenosa* (COMAI et al., 2000); e alotetraploide entre espécies de algodão (*Gossypium*) que divergiram entre 7 e 8 milhões de anos (LIU et al., 2001). Quanto mais distantes as espécies fossem menor seria a possibilidade de pareamento entre homeólogos, e maior a estabilidade do híbrido, de outra forma, os alopoliploides poderiam passar por um rápido processo de diploidização, envolvendo rearranjos genômicos (JACKSON; CHENG, 2010). Os TEs teriam assim esse papel duplo na reorganização genômica após a alopoliploidização, afetando tanto características estruturais quanto estados epigenético do genoma (TEIXEIRA et al., 2009).

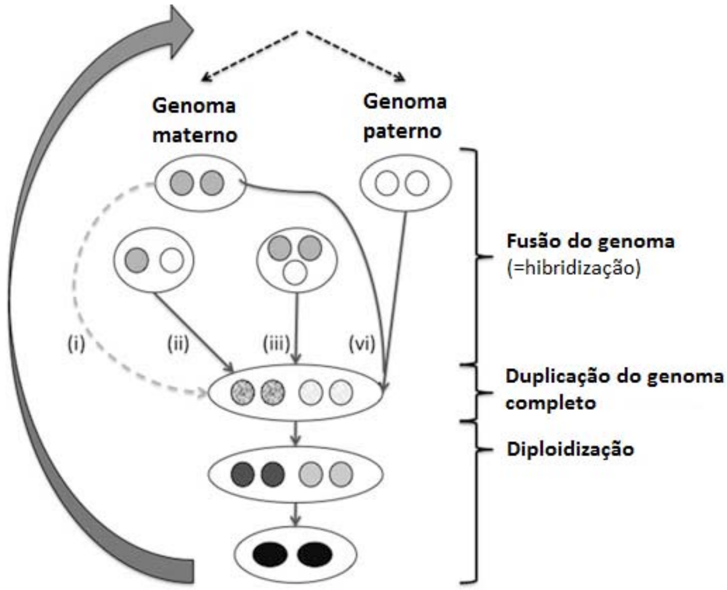


Figura 3. Evolução dos poliploides naturais. Adaptado de Parisod e Senerchia (2012).

1. 4 Gênero *Coffea*

O gênero *Coffea* pertence à família Rubiaceae, a quarta mais numerosa família de Angiospermas, com mais de 13.000 espécies distribuídas em cerca de 650 gêneros que ocupam regiões tropicais em sua maior parte. A família Rubiaceae pertence à superclasse Asteridae que divergiu do clado Rosidae entre 114 e 125 milhões de anos (WIKSTRÖM et al., 2001), estes compreendem dois terços das espécies de angiospermas. Ambas são dicotiledôneas, grupo que compartilha um ancestral comum com as monocotiledôneas há cerca de 150 milhões de anos. Devido à sua importância econômica, *Coffea* é o gênero mais estudado da família Rubiaceae. Esse gênero pertence à tribo Coffeeae da subfamília Ixoroideae, e teria originado e se diversificado recentemente no continente africano (CROS et al., 1998), há cerca de 15 milhões de anos (BREMER; ERICKSON, 2009) (Figura 4(A)).

Duas espécies, dentre as mais de cem que constituem o gênero, são cultivadas em todas as regiões tropicais do mundo: *C. canephora*, espécie diploide (2n = 22) e autoincompatível que prefere ambientes úmidos e várzeas, e *C. arabica*, espécie alotetraploide (única do gênero, 2n = 4x = 44) e autofértil que cresce em regiões mais frias, muitas vezes, em altitudes de até 2.000 metros. Evidências botânicas indicam que *C. arabica* foi originada no platô da Etiópia Central, onde cresce em estado selvagem até hoje, provavelmente, pelo cruzamento relativamente recente, há menos de 1 milhão de anos atrás, entre *C. canephora* e *C. eugenioides* (LASHERMES et al., 1999; YU et al., 2011; CENCI et al., 2012). *Coffea canephora* e *C. eugenioides* divergiram há cerca de 4,2 milhões de anos (YU et al., 2011), e estudos de segmentos do genoma cloroplástico (cpDNA) revelam que um ancestral de *C. eugenioides* ou uma espécie próxima seria o ancestral materno de *C. arabica* (TESFAYE et al., 2007) (Figura 4(B)).

TEs constituem ~ 50% do genoma de *C. canephora*, sendo que desses cerca de 80% são retrotransposons (DENOEUD et al., 2014). Estudos dessas sequências repetitivas no transcriptoma de três espécies de *Coffea* (*C. arabica*, *C. canephora* e *C. racemosa*) sugerem a presença de retrotransposons ativos (LOPES et al., 2008; 2013). No presente trabalho, os TEs foram analisados sob duas vertentes. Primeiramente, foi realizada uma ampla análise comparativa envolvendo diversas espécies de angiospermas de um elemento específico, o *Copia25*, identificado no genoma de *C. canephora*, e que apresentou, em análises preliminares, alta similaridade com elementos de espécies distantemente relacionadas. Em um segundo momento, foi realizada uma análise comparativa no que concerne à dinâmica evolutiva de 10 retrotransposons estruturalmente completos identificados no genoma do parental *C. canephora* entre os parentais diploides e o híbrido alotetraploide. Os resultados obtidos permitiram a proposição de hipóteses de eventos que teriam ocorrido em um tempo evolutivo recente, i.e., nas mudanças genômicas ocorridas no híbrido após sua alopoliploidização no último milhão de anos, bem como em um tempo evolutivo distante, nos eventos de transferência horizontal e na alta conservação de sequências envolvendo espécies que divergiram há 150 milhões de anos.

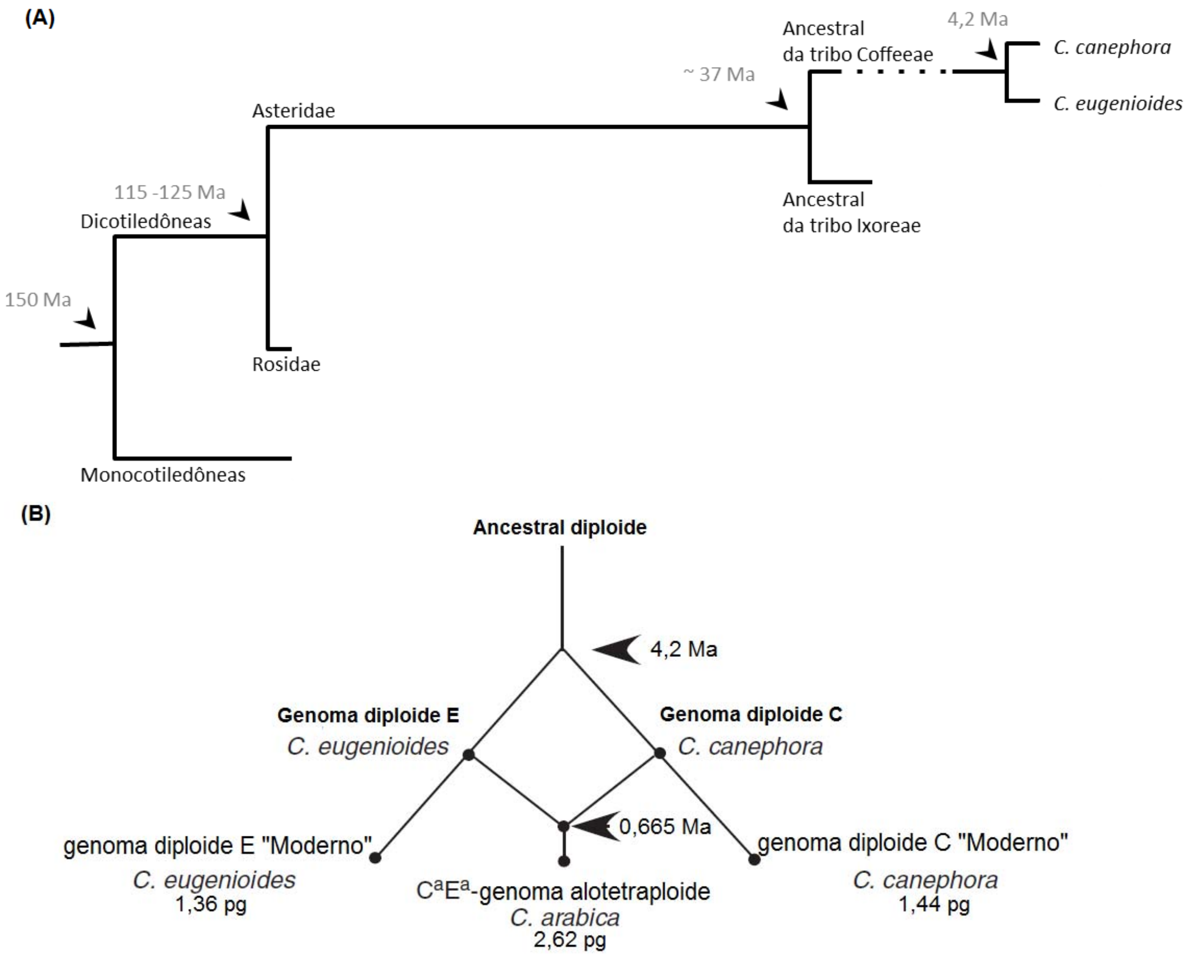


Figura 4. Relações entre as espécies envolvidas neste trabalho. (A) Esquema representando a história evolutiva dos principais grupos envolvidos neste trabalho. (B) Esquema da origem do alotetraploide *C. arabica*; Adaptado de Yu et al. (2011). Ma = Milhões de anos.

2 OBJETIVOS

2 OBJETIVOS

A presente tese teve como objetivo geral analisar a dinâmica evolutiva de LTR-RTs identificados no genoma de *C. canephora*. Essa análise foi realizada em duas frentes: i) na inferência da história evolutiva de um LTR-RT, em particular, o *Copia25*, em vários genomas de plantas; e, ii) na avaliação do uso de retrotransposons ativos como marcadores moleculares em três espécies de *Coffea*, as espécies parentais, *C. canephora* e *C. eugenioides*, e o híbrido alotetraplóide, *C. arabica*, para estimar a variabilidade dos RTs nessas espécies, e sua herdabilidade e dinâmica evolutiva em *C. arabica*.

A busca pelo objetivo geral foi propiciada pelos seguintes objetivos específicos:

1 - Anotar o LTR-RT *Copia25* no genoma de *C. canephora* e realizar uma busca por sequências pertencentes à mesma família nos genomas disponíveis de 41 espécies de plantas;

2 - Analisar evolutiva e filogeneticamente a família *Copia25*.

3 - Anotar 10 LTR-RTs potencialmente ativos no genoma de *C. canephora*;

4 - Determinar o polimorfismo de inserção e o padrão de herança dos 10 RTs selecionados nas espécies parentais (*C. canephora* e *C. eugenioides*) e no híbrido (*C. arabica*);

5 - Estimar a diversidade genética (abundância, distribuição e polimorfismo insercional) dos LTR-RTs em 18 genótipos de *C. canephora*, 5 de *C. eugenioides* e 21 de *C. arabica*.

Plant Molecular Biology

Large distribution and high sequence identity of a Copia-type retrotransposon in angiosperm families

--Manuscript Draft--

Manuscript Number:	PLAN-D-15-00076R2
Full Title:	Large distribution and high sequence identity of a Copia-type retrotransposon in angiosperm families
Article Type:	Manuscript
Keywords:	Copia25, transposable element, genome dynamics, sequence conservation, horizontal transfer, Rubiaceae.
Corresponding Author:	Romain Guyot, Ph.D Institut de Recherche pour le Développement Montpellier, FRANCE
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Institut de Recherche pour le Développement
Corresponding Author's Secondary Institution:	
First Author:	Elaine Silva Dias
First Author Secondary Information:	
Order of Authors:	Elaine Silva Dias
	Clemence Hatt
	Serge Hamon
	Perla Hamon
	Michel Rigoreau
	Dominique Crouzillat
	Claudia Aparecida Carareto
	Alexandre de Kochko
	Romain Guyot, Ph.D
Order of Authors Secondary Information:	
Funding Information:	
Abstract:	Retrotransposons are the main component of plant genomes. Recent studies have revealed the complexity of their evolutionary dynamics. Here, we have identified Copia25 in Coffea canephora, a new plant retrotransposon belonging to the Ty1-Copia superfamily. In the Coffea genomes analyzed, Copia25 is present in relatively low copy numbers and transcribed. Similarity sequence searches and PCR analyses show that this retrotransposon with LTRs (Long Terminal Repeats) is widely distributed among the Rubiaceae family and that it is also present in other distantly related species belonging to Asterids, Rosids and monocots. A particular situation is the high sequence identity found between the Copia25 sequences of Musa, a monocot, and Ixora, a dicot species (Rubiaceae). Our results reveal the complexity of the evolutionary dynamics of the ancient element Copia25 in angiosperm, involving several processes including sequence conservation, rapid turnover, stochastic losses and horizontal transfer.
Response to Reviewers:	Dear editor, please find below our point by point answers to Reviewer1.

Large distribution and high sequence identity of a *Copia*-type retrotransposon in angiosperm families

Authors and Affiliations

Elaine Silva Dias[1,3] (elainedias_bio@yahoo.com.br)

Clémence Hatt[1] (clemhatt@gmail.com)

Serge Hamon[1] (serge.hamon@ird.fr)

Perla Hamon[1] (perla.hamon@ird.fr)

Michel Rigoreau[2] (michel.rigoreau@rdto.nestle.com)

Dominique Crouzillat[2] (dominique.crouzillat@rdto.nestle.com)

Claudia Marcia Aparecida Carareto[3] (carareto@ibilce.unesp.br)

Alexandre de Kochko[1] (alexandre.dekochko@ird.fr)

Romain Guyot[4]* (romain.guyot@ird.fr)

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

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

[3]UNESP – Univ. Estadual Paulista, Department of Biology, São José do Rio Preto, SP, Brazil.

[4]IRD UMR IPME, COFFEEADAPT, BP 64501, 34394 Montpellier Cedex 5, France

*Corresponding Author: Romain Guyot, Institut de Recherche pour le Développement (IRD), UMR IPME, BP 64501, 34394 Montpellier Cedex 5, France, +33467416455, romain.guyot@ird.fr

Data deposition: KM439056 to KM439101

Abstract

Retrotransposons are the main component of plant genomes. Recent studies have revealed the complexity of their evolutionary dynamics. Here, we have identified *Copia25* in *Coffea canephora*, a new plant retrotransposon belonging to the *Ty1-Copia* superfamily. In the *Coffea* genomes analyzed, *Copia25* is present in relatively low copy numbers and transcribed. Similarity sequence searches and PCR analyses show that this retrotransposon with LTRs (Long Terminal Repeats) is widely distributed among the Rubiaceae family and that it is also present in other distantly related species belonging to Asterids, Rosids and monocots. A particular situation is the high sequence identity found between the *Copia25* sequences of *Musa*, a monocot, and *Ixora*, a dicot species (Rubiaceae). Our results reveal the complexity of the evolutionary dynamics of the ancient element *Copia25* in angiosperm, involving several processes including sequence conservation, rapid turnover, stochastic losses and horizontal transfer.

Keywords

Copia25, transposable element, genome dynamics, sequence conservation, horizontal transfer, Rubiaceae.

Introduction

Transposable elements (TEs) are the major component of plant genomes. TEs are typically "vertically" transmitted from parent to offspring. If a new insertion occurs in germ cells tissues, the new copy will be transmitted to the progeny. In certain cases, TEs can be horizontally transferred (HT) between reproductively isolated species. Although more than 200 cases of HT have been reported most of them involve animals (Schaack et al. 2010), mainly insects (mostly *Drosophila*), and few potential cases have been reported in plants (Cheng et al. 2009; Diao et al. 2006; Fortune et al. 2008; Roulin et al. 2008) with the exception of a very recent observation (El Baidouri et al. 2014). The HTs concern both Class I (or Retrotransposon) and Class II (or Transposons) elements, and the mechanisms underlying TE HTs remain speculative in most of the cases (vectors could be pathogens, intracellular parasites, insects, etc.). Because TEs play a major role in the dynamics of genomes, their direct introduction into a "naïve" genome through HT may induce important consequences in chromosomal and genomic evolution. However, the detection of potential HT of TEs in complete genomes is relatively complex and requires highly sensitive methods to differentiate between unresolved sequence conservation and HT events (de Carvalho and Loreto 2012). In the absence of a clear mechanism underlying HT, cases of outstanding sequence conservation of TEs between evolutionarily distant plant species living in separate geographical areas have raised questions as to the existence of other mechanisms leading to this conservation (Moisy et al. 2014). The recent availability of plant genome sequences (Michael and Jackson 2013) gave new opportunities to identify and to characterize transposable elements and to gain a higher understanding of the evolutionary dynamics of these elements and their conservation between distantly related species.

The coffee genus (*Coffea*) that belongs to the Rubiaceae family, comprises 124 species, originating from Africa, Madagascar, the Mascarene Islands, Asia and Oceania (Davis 2010; Davis 2011). *Coffea* species are diploids (2n = 2x = 22) and generally allogamous. The notable exception is the self-fertilizing allotetraploid *Coffea arabica* (2n = 4x = 44), native to the Ethiopian highlands and originating from a recent hybridization of two different diploid ancestors, *C. canephora* and *C. eugenioides* (Lashermes et al. 1999; Yu et al. 2011). The current possibility of accessing genomic and transcriptomic sequences of *Coffea* species has made it possible to expand our knowledge of the composition and behavior of TEs in these important species. The analysis of the *C. canephora* genome showed that these sequences contained about 50% of transposable elements (Denoeud et al. 2014). The vast majority of them (85%) are retrotransposons with LTRs (LTR-RTs). The study of TEs in *Coffea* is very recent and the few individual TEs investigated to date show different dynamics between closely related coffee species (Hamon et al. 2011; Yuyama et al. 2012).

In this study, LTR-RTs were identified in the *C. canephora* genome using BAC-end sequences (BESs) and 454 sequences. One of them, a *Ty1-Copia* element named *Copia25*, was characterized and analyzed under different aspects of its evolution because its nucleotide sequence showed unusually high similarities with distantly related plant genomes. Furthermore, *Copia25* was found quite similar to *Rider*, an active retrotransposon identified in the tomato with a rather unique evolutionary history. *Rider* activity has played a role in the origin of at least three different phenotypes of this species (Jiang et al. 2009; Jiang et al. 2012; Xiao et al. 2008). Since it is absent in *Solanum tuberosum*, it has been suggested that *Rider* appeared in the tomato by HT from *Arabidopsis thaliana* (Cheng et al. 2009). The similarity shared between *Copia25* and *Rider* makes the TE identified in *C. canephora* interesting to investigate, particularly for its activity and evolutionary dynamics. In the current study, we show that *Copia25* is an active element in *Coffea*, widely present in Rubiaceae species. In addition, a phylogenetic analysis indicates outstanding conservation of *Copia25* in coffee trees and in distantly related species, such as banana (*Musa* genus), a monocot. The different processes that can lead to high conservation of *Copia25* in Angiosperms are discussed.

Materials and Methods

Genome sequencing

The Next-Generation Sequencing (NGS – by Genomic 454 Pyrosequencing - GS Junior System Roche) was performed in two accessions of *C. canephora* Pierre ex A. Froehner (HD200-94 a double haploid from the Congolese diversity group, also used for whole genome sequencing – Denoeud et al. 2014, http://coffee-genome.org –, and BUD15 from Uganda), as well as in one accession from each of the following taxa: *C. arabica* L. (ET39 from Ethiopia), *C. eugenioides* S. Moore (DA56 from Kenya), *C. pseudozanguebariae* Bridson (08107 from Kenya), *C. heterocalyx* Stoff (JC65 from Cameroon), *C. racemosa* Lour (IA56 from Mozambique), *C. humblotiana* Baill (A.230 from Comoros), *C. millotii* J.-F. Leroy (ex-*dolichophylla*, A.206 from Madagascar) and *C. tetragona* Jum. & H. Perrier (A.252 from Madagascar), *Coffea* (ex-*Psilanthus*) *horsfieldiana* (Miq.) J.-F. Leroy (HOR from Indonesia) and *Craterispermum* Sp. Novo Kribi (from Cameroon) (Chevalier 1946; Maurin et al. 2007). The cultivars and the above-mentioned sequenced accessions grow in the IRD greenhouses (Montpellier, France), at the Kianjavato research station (Madagascar) or in the Nestlé R&D greenhouses (Tours, France). The total genomic DNA was extracted from young leaves using the Qiagen DNeasy Plant Mini Kit following the manufacturer’s protocol. The library and sequencing for the NGS were performed at the Nestlé R&D laboratory according to the

Roche/454 Life Sciences Sequencing Method. Data were submitted to GenBank, BioProject PRJNA242989.

Sequence Analyses

We used 131,412 BAC end sequences (BESs) (Dereeper et al. 2013) obtained by Sanger sequencing and 106,459 sequences obtained by 454 Roche-NGS technology, both derived from the *C. canephora* HD200-94 accession. All sequences (Sanger and 454 Roche) were used for the assembly using AAARF (Assisted Automated Assembler of Repeat Families - DeBarry et al. 2008). The following parameters for the BLAST analyses and the Minimally Covered Sequences (MCS) construction and controlling "build" extensions were applied: minimum hit length: 150; minimum hit identity: 0.89; minimum coverage depth: 4; required MCS length: 150; maximum E-value: $1e^{-25}$; required coverage length: 150; minimum hit number: 2; required overlap between MCS and new query: 90; and maximum times a number sequence is used in each direction: 13. These parameters were those that gave best assembly results after several modification and assembly testing.

AAARF "builds" were analyzed using BLASTx (min E-value $1e^{-4}$) against public protein sequence databases (uniprot_sprot; http://www.uniprot.org/), and transposable element databases available in Repbase (Jurka et al. 2005 – http://www.girinst.org/repbase/) and Gypsy DB 2.0 (http://gydb.org - Llorens et al. 2011). The graphical dot-pot (Dotter - Sonnhammer and Durbin 1995) was also performed. The final annotations of each "build" were edited in Artemis (Carver et al. 2005). Validation of LTR-RT "build" structures was performed by comparative analysis with public Coffee BAC sequences, from the NCBI and the genome of *C. canephora* (Denoeud et al. 2014 - coffee-genome.org). Five BAC clones for *C. canephora* (EU164537, HQ696512, HQ696507, HQ696513 and HM635075) and 12 BAC clones for *C. arabica* (GU123896, GU123899, GU123898, GU123894, GU123897, GU123895, HQ696508, HQ696510, HQ696509, HQ696511, HQ834787 and HQ832564) were downloaded from GenBank, accounting for a total of 3,023 Mb. BLASTN searches (E-value < $1e^{-150}$) against public Expressed Sequenced Tags (ESTs) databases from *C. canephora* and *C. arabica* were used to evaluate the transcription of the builds.

Estimation of the *Copia25* copy number using 454 sequencing survey

BLASTN searches were carried out with the full-length *Copia25* sequence (from BAC HQ696507) as query. Reads with more than 90% of nucleotide identity with *Copia25* over a minimum of 80% of the read lengths were considered as potential fragments of the element. Cumulative lengths of aligned reads to *Copia25* were used to extrapolate the contribution of the element to each genome size investigated.

Identification of *Copia25* in plant genomes

The sequence trimmed from AAARF was blasted against the *C. canephora* genome, as well as against 40 angiosperm and one non-angiosperm genome sequences available in the public databases of NCBI, Phytozome and Gramene (Table S1). BLASTN was used to search for the complete nucleotide sequence or the coding region of *Copia25* in the genomes. The retrieved sequences were analyzed using LTRharvest (Ellinghaus et al. 2008) in order to recover only the sequences with a structure similar to retrotransposons. These sequences were compared to the amino acid sequence of the *Copia25* reverse transcriptase (RT) using TBLASTN and against the *Ty1-Copia* retrotransposon databases of plants (Repbase http://www.girinst.org) resulting in 98 sequences from 34 species (Table S2).

Molecular analysis

The DNA of 24 Rubiaceae species (Table S3, Fig. S1) was extracted by using DNeasy Plant mini-kit (QIAGEN). The DNA of the *Musa* species was donated by Dr. A. D'hont (CIRAD, France). Primers were designed on intact RT region of *C. canephora Copia25* genomic sequences using Primer3 (http://bioinfo.ut.ee/primer3-0.4.0/primer3/) (*Forward*: 5' GGG GTT GAA GAT GCA AGG TA 3'; *Reverse*: 5' AGC TGC TCC CAA ATC TTT CA 3'). For the reaction, 0.625 unit of Taq polymerase (Invitrogen), 20 ng genomic DNA, 1 mM of $MgCl_2$, 1 X buffer, 0.08 mM of dNTPs and 0.4 mM of each primer were used for a final volume of 25 μL. PCR conditions were as follows: initial denaturation (94 °C, 120 s); followed by 40 cycles of denaturation (94 °C, 30 s), annealing (55 °C, 30 s) and extension (72 °C, 180 s). Each PCR product was analyzed by gel electrophoresis on 1.2% agarose gel, purified (DNA GFX DNA & Gel Band, GE) and cloned (TOPO XL Cloning kit, Invitrogen) according to the manufacturer specifications. The plasmids extracted were sequenced using the specific primers. The *Copia25* sequences were registered under the GenBank Accession Numbers KM439056 to KM439101. For the reverse transcription polymerase chain reaction (RT-PCR) 1 μg of the total RNA from leaves of *C. canephora, C. eugenioides* and *C. arabica* was treated with RQ1 RNase-Free DNase (Promega) and reverse-transcribed using ImProm-II™ Reverse Transcription System (Promega). The synthesized cDNA served as templates for RT-PCR. DNA contamination was checked using the primers of the gene sucrose synthase (SUS10/SUS11 - Marraccini et al. 2011). RT-PCR was performed using the same specific primers according to the protocol described as before, with 50 ng of cDNA.

Evolutionary Analyses

Phylogenetic analyses were performed with MEGA 5.2 (Kumar et al. 2008) on sequence datasets aligned with the MAFFT program. Each phylogeny was reconstructed using the best model using Find Best DNA/Protein Model (Maximum Likelihood) in Mega 6 (Tamura et al. 2013), with 1000 replicates; the bootstrap consensus tree inferred is taken to represent the evolutionary history of the taxa analyzed. All positions containing gaps and missing data were eliminated. As rates of synonymous substitution are not available for Rubiaceae (genes or TEs), and because LTR sequences (non-coding regions) and those from the RT domain (coding region) may evolve differently, two rates, estimated for grasses and palms, were used. The age of insertion of *Copia25* within *C. canephora* genome was estimated using the molecular clock equation, as previously described (Moisy et al. 2014; SanMiguel et al. 1998; Wicker and Keller 2007), where *k* was the Kimura 2-parameter distance between both LTRs of the same copy, and *r* is 1.3 x 10^{-8} base substitutions per site per year (Ma and Bennetzen 2004). The Kimura 2-parameter method of distance estimation of non-coding nucleotide sequences was used for LTR distance estimation (SanMiguel et al. 1996). However, gene conversion between LTR of the same element could be a source of errors in estimating insertion time. This putative error is not taken into account in our analysis since conversion of LTR remains poorly understood in plant genomes. The age of the ancestor of the *Copia25* sequences was also estimated using the molecular clock equation, using *Ks* (number of synonymous substitutions per synonymous site) and the rate of synonymous substitutions as 6.5 x 10^{-9} base substitutions per site per year (Gaut et al. 1996) for the RT domain (Vitte et al. 2007).

In order to investigate whether *Copia25* was under selective pressure a codon substitution model was used to estimate ω (Ka/Ks). The ω ratio measures the direction and the magnitude of selection on amino acid changes, with values of $\omega < 1$, $= 1$, and > 1 indicating negative purifying selection, neutral evolution, and positive selection, respectively. To estimate ω two approaches were used: (i) the Ka/Ks pairwise ratio for species with the full-length polyprotein sequence available (coffee, potato, tobacco and banana); and (ii) likelihood ratio tests (LRTs) for a simplified phylogeny (Fig. S2) containing species representatives of each of the Rubiaceae tribes and potato, tobacco and banana, using 315 nt of the RT domain. Premature stop codons were removed from the sequences for both analyses. For the pairwise Ka/Ks, the reference sequences of the *Copia25* Subfamilies 1 and 2 (chr7_16264485-16269785 and chr8_8081742-8086630 respectively) were compared with their homologous sequences in potato, tobacco and banana. *Ka* and *Ks* were obtained using DnaSP v5 (Librado and Rozas 2009). Selective pressure acting on COSII (conserved orthologs group) genes of potato, banana and coffee (Wu et al. 2006) was also investigated. The COSII sequences in potato and *C. canephora* are available on the Sol Genomics Network website (http://solgenomics.net). 515 COSII accessions present in single copy in potato and coffee were blasted (BLASTn) against the *Musa acuminata* CDSs (D'Hont et al. 2012 - http://banana-genome.cirad.fr/) in order to obtain the *Musa* COSII sequences. Seven COSII sequences showing the highest sequence identity were used to calculate the Ka/Ks ratio and nucleotide identity (Table S4). The second approach used different ω ratio parameters for different branches on the phylogeny (Anisimova and Ziheng 2007; Yang and Nielsen 1998). To estimate the log likelihood values (LRT), a one-ratio model was used. This model assumes the same ω free or fixed (ω = 1) parameter for the entire tree, Model I and Model II, respectively. A two-ratio model was used to estimate the LRTs for specific clades on the phylogeny, since we assumed that the sequence group of interest (separately for *Ixora*, Model III = ω free, and Model IV = ω fixed; and, for *Musa*, Model V = ω free, and Model VI = ω fixed) has a different ω_F from that of the ω_B background. For the pairs of models (I *vs* II, III *vs* IV, V *vs* VI)), the log likelihood values were compared in a hypothesis test (X^2). These analyses were implemented using the codeml program in the PAML package (Yang 1997).

Results

Assembly of repeated sequences with BAC-end Sanger sequences and 454 random reads from *C. canephora*

Sanger and 454 sequences from *C. canephora* (accession HD200-94) were used to identify and characterize the TEs. Two bacterial artificial chromosome (BAC) libraries were recently constructed from the same plant and a total of 134,827 Sanger sequences (mean size 683 bp) were generated from BAC-end sequences (BES) and released (Dereeper et al. 2013). In addition, 106,459 random 454 Roche reads (mean size 423 bp) were also generated from the same plant (Table S5).

In all, Sanger and 454 sequences represent 137,104,866 bp (241,286 sequences), giving an estimated coverage of 19.5% of the *C. canephora* genome (710 Mb). They were used together to assemble repeated sequences using the Assisted Automated Assembler of Repeat Families Algorithm (AAARF, DeBarry et al. 2008). A total of 1,306 "builds" (also called contigs) were generated with a length ranging from 135 to 24,745 bp, and a mean length of 1,306 bp. Most of them (45%) have a length comprised between 0.5 and 1 kb. In total, 317 builds showed similarities with TE proteins available in public databases after translating the assembled sequences. Fifty-two of them, showing sizes larger than 3 kb, were selected for the subsequent analysis. Forty-nine out of 52 showed strong similarity to LTR-RT proteins (Table S6 and Table S7). Over the 49 contigs, 12 elements were removed due to non-canonical (complex) structure, suggesting incorrect assembly, and in a significant number of builds, manual corrections were made (Table S6, 10 builds labeled with ‡), following the same procedure as described in De Barry et al. (2008). The 37 remaining builds with canonical TE structures showed exclusively similarities with LTR-RT proteins, suggesting that it may represent the main abundant transposable element family in the *C. canephora* genome (Table 1). These 37 potential retrotransposon builds, were manually annotated, and incomplete structures of all them were found (Fig. S3). According to the structural annotation, the were classified as “LTR-I-LTR” when the internal region and both complete or partial LTRs were present; as, “I” if only an internal region was present, as “LTR-I” with complete or partial 5’ LTR with an internal region, and, “I-LTR” with an internal region and complete or partial 3’ LTR (Table S6).

The 37 LTR-RT builds were used as query for similarity search (BLASTn) for complete or partial copies present in the available *Coffea* BAC clones sequences (Table S6). Ten LTR-RT builds showed high levels of nucleotide conservation with nine *C. canephora* (4) and *C. arabica* (5) BAC sequences (BLAST E-value cutoff: $10e^{-100}$; Table S6). Moreover, some builds showed similarities with *Coffea* transcriptomic sequences. Indeed, 15 and four LTR-RT builds were found in *C. canephora* and *C. arabica* ESTs, respectively (Table S7).

Characterization of *Copia25*, a *Ty1-Copia* LTR retrotransposon in Coffee trees

Among the retrotransposons identified in *C. canephora* sequences (accession HD200-94), the sequence of one *Ty1-Copia* element, hereafter named *Copia25*, showed high BLASTN scores across various distantly related plant genomes, suggesting that *Copia25* has a singular evolutionary history. *Copia25* also showed an overall structure similarity to *Rider* (EU195798), an active retrotransposon found conserved between distant dicot species (Cheng et al. 2009; Jiang et al. 2012), as indicated by dot-plot alignment (not shown). The *Copia25* reassembled contig was blasted (BLASTN $10e^{-100}$, Table S6) against *C. arabica* and *C. canephora* BAC sequences. It was found in *C. canephora* but with an uncommon arrangement, which appears to be a tandem of two elements sharing one LTR sequence in the median of the structure (accession HQ696507). In *C. arabica*, in turn, a complete sequence of 5,382 bp was found. This sequence is flanked by two perfect 5-bp TSDs (5'-GGAAC-3'), and its two LTRs are both 530 bp long and show high sequence identity (99.2%) (accession HQ832564 - Fig. S4). This copy is localized on a homologous region to *C. canephora*, most probably the *C. canephora* sub-genome within *C. arabica*, but it is absent in the syntenic region in both 126 (Moschetto et al. 1996; Yu et al. 2011) and HD200-94 *C. canephora* genotypes (Denoeud et al. 2014).

A search was also made for the *Copia25* contig (using Censor) in the *C. canephora* genome (Denoeud et al. 2014) and 72 full-length copies were identified. All of them showed premature stop codons in the *pol* coding region, indicating that none of them is potentially functional. Nonetheless, similarity searches showed high sequence identity between *Copia25* and Expressed Sequence Tags (98 and 99% of nucleotide identity with DV679393 and GT681881, respectively). In addition, the *Copia25* RT regions were successfully amplified by RT-PCR on RNA extracted from *C. canephora*, *C. arabica* and *C. eugenioides* leaves (Fig. S5).

Full-length *Copia25* copies exist throughout the *C. canephora* genome mainly in gene-poor and LTR-RTs rich areas. The majority of them are located in the non-anchored set of scaffolds (pseudo-chromosome "0") (Fig. 1a; Table S8). The sharing of structural characteristics among group of sequences of a TE family might indicate the occurrence of subfamilies. In such cases, the different groups have different most recent ancestral copy – i.e. different mother (or master) copy –, which independently originated copies. A Maximum

Likelihood with the distance corrected by General Time Reversible model and 1000 replicates phylogenetic tree was produced using the *pol* (2,640 nt) nucleotide sequence of the 72 full-length *Copia25* copies. Based on the tree topology, two clusters were segregated (Fig. 1b). Following Wicker's parameters (Wicker et al. 2007) segregating criterion they are hereafter considered as subfamilies, one harboring 44 copies (Subfamily 1) and the other 28 (Subfamily 2). Only one copy did not group with either of the two clusters; this copy was discarded from further analyses. In each subfamily, the sequence with the perfect structure (based on the best conservation of both LTRs and the presence of an intact or few stop codons in the ORF coding for the polyprotein) was chosen as a reference sequence for the subfamily (Subfamily 1: chr7_16264485-16269785; Subfamily 2: chr8_8081742-8086630). These two sequences are 87.8% identical, and have 9.8% of InDels. The differences between them are mainly concentrated in the LTR region, where the identity is only 71%, and InDels reach 15%, resulting in only 59% of overlap. Such difference results in poor LTR alignment of the 72 copies. Additionally, Subfamily 2 presents a 208 bp deletion in the UTL 5' (Untranslated Leader) region. The corrected distances (Tamura-3 parameters) within each subfamily are 0.123 and 0.138 respectively, for Subfamily 1 and 2, and 0.222 between subfamilies (overall mean of 0.174). The divergence between the two LTRs of each copy was calculated and an insertion time was inferred. Subfamily 1 showed a mean time of insertion of 2.97 ± 0.204 Mya (minimum: 0.5, maximum: 5.2 Mya) and Subfamily 2 showed a mean time of insertion of 4.53 ± 0.399 Mya (minimum: 1.3, maximum 10.1 Mya) (Fig. 2, Table S8).

Presence of *Copia25* in the Rubiaceae family

In order to investigate the evolution of *Copia25*, sequence similarity searches and PCR amplifications were used to search for its presence in the *Coffea* genus and in other Rubiaceae species. First, 11 genotypes representing 10 *Coffea* species (including ex-*Psilanthus*) and *Craterispermum sp. Novo kribi* were surveyed using high-throughput 454 Roche sequencing. The number of bases produced for each species and the estimated genome coverage according the genome sizes are shown in the Table 2. The 454 sequences were used to survey the presence of highly conserved *Copia25* sequences, using as criteria: 90% minimal nucleotide identity over 80% of the sequence length. The number of *Copia25* conserved sequences found for each species and their respective cumulative length according to the genome size are available in the Table 2. Sequences fitting these criteria were present in all *Coffea* genomes studied here, but not in *Craterispermum*. The cumulative length of *Copia25* reads was estimated to range from 186 to 1,513 kb of estimated cumulative sequences in diploid species and 842 kb in the allotetraploid *C. arabica* (Table 2).

The presence of *Copia25* was also investigated by PCR amplification and sequencing of the product in 13 *Coffea* and 11 other Rubiaceae species (Table S3, Fig. S1). The *Copia25* RT region was amplified and sequenced in 13 *Coffea* species, three from West Africa (*C. stenophylla*, *C. humilis* and *C. ebracteolatus*), one from West/Central Africa (*C. canephora*), three from East Africa (*C. costatifructa*, *C. pseudozanguebariae* and *C. eugenioides*), one from Northeast Africa (*C. arabica*), and five from Indian Ocean Islands (*C. millotii – ex-dolichophylla –*, *C. perrieri*, *C. resinosa*, *C. tetragona* and *C. vianneyi*) (Chevalier 1946; Maurin et al. 2007). The same region was also amplified and sequenced in 11 other Rubiaceae species: *Bertiera iturensis*, *Tricalysia congesta*, *Oxyanthus formosus*, *Ixora* sp., *I. coccínea*, *I. finlaysoniana*, *I. foliicalyx*, *Polysphaeria parvifolia*, *Coptosperma* sp., *Pyrostria* sp., and *Craterispermum schwenfurthii*. The final dataset contains 319 nucleotides, and the nucleotide identity varied from 62% to 100% among different sequences comparisons (Table S9).

Copia25 distribution among monocots and dicots

Besides the Rubiaceae species, similar *Copia25* sequences were sought among the 40 available plant sequences representing the angiosperm clades, and one non-angiosperm species using BLASTN. Similar *Copia25* sequences were found in 34 species but not in the remaining eight ones, as follows: *Arabidopsis lyrata*, *Carica papaya*, *Cucumis sativus*, *Fragaria vesca*, *Linum usitatissimum*, *Selaginella moellendorffii*, *Phoenix dactylifera* and *Zea mays* (Table S1).

In the 34 genomes where sequences similar to *Copia25* were found, these latter were extracted for further phylogenetic analysis. Using a fragment of 750 bp from the RT region, a phylogeny was reconstructed using Maximum Likelihood, with the distance corrected by Tamura 3-parameter and 1000 replicates in order to investigate the relationships among the *Copia25* sequences (Fig. 3, Fig. S6 and Tables S10 and S11). One well-supported (95% bootstrap value) phylogenetic clade was found to include *C. canephora Copia25* and sequences belonging to four dicotyledonous species: *Nicotiana benthamiana*, *N. tabacum*, *S. tuberosum* (Solanaceae) and *Ricinus communis* (Euphorbiaceae), and more surprisingly, three monocotyledonous species, *Musa accuminata and M. balbisiana* (Musaceae), and, in a basal position, *Eleais guineensis* (Arecaceae). These sequences were considered homologous to *Copia25* because they share over 80% sequence identity over 80% of their length in the reverse transcriptase domain (Wicker et al. 2007), except for *R. communis* and *E. guinensis*. Since these two species cluster within the clade and share, with *Copia25*, over 70% of identity they were considered to belong to the same family.

Besides the *Copia25* clade, additional *Ty1-Copia* sequences related to it, clustered in strongly-supported clades composed of species of the same family, which supports a hypothesis of vertical inheritance (Fig. 3). It is the case of the elements found in the monocotyledonous family of Poaceae where all of them cluster in a clade with a 94% bootstrap value. A similar occurrence was found in the Malvaceae species (100%) and in Fabaceae (98%) species, but it is also weakly supported among Brassicaceae (79%). The exceptions in this context are the particular strongly-supported relationships between *Medicago truncatula* (Fabaceae) and *Mimulus guttatus* (Phrymaceae) (94%), among *Populus trichocarpa* (Salicaceae), *Gossypium hirsutum* (Malvaceae) and *Malus domestica* (Rosaceae) (100%), and finally between *Solanum lycopersicum* (Solanaceae) and *Arabidopsis thaliana* (Brassicaceae) (100%).

The reconstructed phylogeny using only sequences recovered from public databases (Fig. 3) did not show a clear relationship between the sequences from coffee tree and those from other species in the clade. In an effort to better understand the relationships of *Copia25* among the species present in the *Copia25* clade, we reconstructed a new Maximum Likelihood phylogeny (with the distance corrected by Tamura 3-parameter and 1000 replicates), adding RT sequences obtained from several Rubiaceae species and three Musaceae species (*M. accuminata*, *M. balbisiana* and *M. boman*) (Fig. 4 and Fig. S7). As shown in Fig. 4, the unrooted phylogenetic tree revealed that *Copia25-Musa* is nested into the Rubiaceae species as shown by a closer well-supported relationship (bootstrap value 92%) between *Copia25-Musa* and *Copia25-Ixora* and between *Craterispermum* sp. and all Rubiaceae and Musaceae species (bootstrap value 65%). Rubiaceae and Musaceae *Copia25* are clearly separated from Solanaceae by high bootstrap value (92) and a topology structure. This result suggested that Rubiaceae and Musaceae *Copia25* constitute a unique evolutionary lineage (Fig. 4).

To further confirm the close relationship between *Copia25-Coffea* and *Copia25-Musa*, we first aligned each *Copia25-Coffea* sequence (*Copia25 C. canephora* reference sequence of Subfamily 1 and 2) with the *Copia25-Musa* (*M. balbisiana* AC186755). The alignments showed an overall nucleotide identity of 74.1% and 79.6% for Subfamily 1 and 2, respectively, and an overall amino acid sequence identity rate of 81.7% (similarity: 79.8%) with Subfamily 1, and 81.60% (similarity: 80.1%) with Subfamily 2 (Fig. 5a). Their LTRs were also extracted and aligned, showing a high identity rate (53.9% between *Musa* and the reference sequence of Subfamily 1; and 59.4% with the Subfamily 2 reference sequence) (Fig. 5b). This level of identity is indeed quite significant for non-coding regions and considering the species divergence, i.e. about 150 Mya (Chaw et al. 2004; Wikstrom et al. 2001). Homologous sequences to *Copia25-Musa* from the *M. balbisiana* genome (B genome) were also found in the sequenced *M. acuminata* genome (A genome; D'Hont et al. 2012). These homologous sequences show high sequence identity (e.g. Chr9: 16119963-16124880; 91.1% of identity) between the two banana genomes that diverged by about 4.6 Mya (Lescot et al. 2008).

Evolution of *Copia25* in monocots and dicots

To investigate the evolution of *Copia25* in detail, we used the nucleotide sequences of *Copia25* from *M. balbisiana*, *C. canephora*, *S. tuberosum* and *N. benthamiana* for pairwise sequence comparisons. The results summarized in Supplementary Table S12 show higher identity between the *Copia25* of coffee and banana than between all the other species. We compared the identity of *Copia25* with the identities of seven COSII sequences showing the highest sequence identity between banana and coffee. These genes share an average of 74.7% of identity between banana and coffee, while the coding region of *Copia25* shows 85%. For the *Copia25* polyprotein and these seven COSII genes, we performed a pairwise Ka/Ks (non-synonymous per synonymous substitution ratio) analysis by comparison of banana, potato, tobacco and coffee sequences. Both COSII and *Copia25* were under purifying selection, however they were found more relaxed in *Copia25* (minimum: 0.233, maximum: 0.287) than in COSII (minimum: 0.038, maximum: 0.215) sequences.

The LRT results reinforce the proposition of the purifying selection acting on the *Copia25* sequences (Table 3). The log likelihood values using a one-ratio model (Model I: ω free, and Model II: ω fixed) for the entire phylogenetic tree (Fig S2) were significantly lower than the neutral expectation, indicating purifying selection (0.191, $2\Delta\ell = 239.308$, $p < 0.01$). The LRTs of the *Ixora* and *Musa* clades were estimated separately. For these, a two-ratio model was applied, since we assumed that the sequence group of interest has a different ω_F from that of the ω_B background (Model III: ω free, and Model IV: ω fixed, for *Ixora* clade; and Model V: ω free, and Model VI: ω fixed, for the *Musa* clade). Purifying selection was also detected for *Ixora* clade (0.127, $2\Delta\ell = 33.568$, $p < 0.01$), while for the *Musa* clade the ω value did not differ from neutral evolution (Table 3). The negative selective pressure would explain the narrow relationship between the coffee and banana sequences. However, the negative selection for *Copia25* and COS, and the neutrality for *Copia25* in *Musa* clade indicate that this alone does not explain their clustering in the phylogeny.

The divergence time of two sequences harbored by two species from their common ancestral sequence was estimated by using both COSII and *Copia25*. The estimated divergence time using *Copia25* sequences for *Musa* and *Coffea* is much lower than for COSII sequences. While the latter ones ranged from 94.5 to 181.8 Mya, when using *Copia25* the time was 35.5 and 31.7 Mya. Indeed, the estimated divergence time using the *Copia25* from banana and the Solanaceae species is similar to that found for coffee, tobacco and potato. The high similarity and the *Ks* values for the comparisons between coffee and banana with the other Solanaceae species indicate that the *Copia25* sequence could be a recent guest in banana species genome.

Discussion

Copia25 in the Rubiaceae family

In this study, we identified an expressed *Ty1-Copia* in the *C. canephora* genome, *Copia25*, and analyzed it under various aspects, providing a broad insight into its evolution. *Copia25* was found distributed in several species of the *Coffea* genus from Africa, the Indian Ocean Islands and Indonesia. The occurrence of *Copia25* in these species denotes that it could be present in the ancestor of this phylogenetic group and has been inherited by the derived lineages. Our proposition of its presence in the *Coffea* lineage ancestor is reinforced by the occurrence of *Copia25* in at least two of the three subfamilies of the Rubiaceae family, Rubioideae (*Craterispermum schwenfurthii*) and Ixoroideae (*Coffea* spp., *Ixora* spp., *Bertiera iturensis*, *Coptosperma* sp., *Oxyanthus formosus*, *Polysphaeria parvifolia*, *Pyrostria* sp., *Tricalysia cloneongesta*), also suggesting its ancient evolutionary history in Rubiaceae. Altogether these data suggest the presence of *Copia25* in both of the Rubiaceae subfamilies preceding their ancient divergence.

High sequence identity of *Copia25* of over 150 My of plant genome evolution

Our similarity searches and molecular biology approaches revealed patchy conservation of *Copia25*. They show high sequence identity between a monocot genus of the Musaceae family and two different dicotyledonous families in Asteridae: the Rubiaceae and Solanaceae families. While monocot and dicot species diverged about 150 Mya, the Asteridae and Rosidae lineages diverged ~114 Mya. More recently, Rubiaceae and Solanaceae diverged from their common ancestor about 83 Mya (Chaw et al. 2004; Wikstrom et al. 2001). This discontinuous and incongruent distribution in dicots and monocots highlights a complex evolutionary history of *Copia25* in plants that could be traced back to the origin of angiosperms.

Copia25-Coffea clusters in a strongly supported clade (100% bootstrap value) with homologous sequences from three Solanaceae species, *S. tuberosum*, *N. tabacum* and *N. benthamiana*, and Musaceae species, *Musa* spp.. However, the nucleotide identity between *Copia25-Coffea* and *Copia25-Musa* is higher than the one observed between *Coffea* and potato and tobacco, and even in the comparison between *Musa* and Solanaceae (*S. tuberosum*: 77.4%; *N. benthamiana*: 77.2%). When the seven orthologous (COSII) genes showing the highest sequence conservation are compared among the same species, the nucleotide identity between *C. canephora* and *M. balbisiana* ranged from 67.8% to 80.2%, less than the *Copia25* polyprotein identity for the same species comparison (Subfamily 1: 84.5% and Subfamily 2: 85.5%). Equivalent identities were also found in the *gag* region. Such outstandingly high conservation raises questions about the molecular mechanisms, which are at its origin.

Conservation of TEs between distantly related genera could be the result of different and non-exclusive processes (Capy et al. 1994; Cummings 1994; Schaack et al. 2010; Wallau et al. 2011) such as: (i) domestication, (ii) conservation of functional sites, (iii) similarity of evolutionary rates, (iv) purifying selection and (v) horizontal transfer. The first two scenarios cannot explain the conservation of *Copia25* across genera, since only portions of the TE are generally domesticated and because the mechanisms of conserving functional sites exclusively involve coding regions. High sequence identity was found for the full-length sequences of *Copia25,* including non-coding LTR regions. Similar TE evolutionary rate in distinct species is an attractive hypothesis to explain the conservation observed in *Copia25.* However, the TE evolutionary rate depends on multiple parameters such as the specific TE activity and the efficiency of TE host control mechanisms. Such a scenario remains unlikely since these evolutionary mechanisms should be identical in several distantly-related species. The fourth process, a purifying selection, would explain the high identity of a given TE between distantly related species. The Ka/Ks ratio estimated for pairwise comparisons of *Copia25* between *Musa* and *Coffea* sequences is low (< 0.3), denoting purifying selection and explaining the conservation and the activity (at least until very recently) of this particular element. However, the *Ks* values between *Coffea* and Solanaceae, *Musa* and Solanaceae and *Musa* and *Coffea* species are at least twice as low for *Copia25* as for COSII sequences. This observation suggests that other evolutionary processes besides purifying selection might be involved in *Copia25* conservation. Finally, HTs of TEs, an occurrence suggested but rarely confirmed in plants (Diao et al. 2006; El Baidouri et al. 2014; Fortune et al. 2008) may explain the strong conservation level in coding and non-coding regions, and the sparse distribution of TEs. However, HT scenarios first require ecological, chronological, and geographical distribution overlapping between the species involved in the potential transfer to be seriously considered. These requirements are not expected for *Musa* and *Coffea*, but a chronological and geographical distribution overlap might have existed for the *Musa* and *Ixora* species. The *Ixora* genus belongs to the Ixoroideae subfamily of the Rubiaceae family such as the *Coffea* genus, but both belong to different tribes, Ixoreae and Coffeeae (Fig. S1). The genus *Musa* evolved and diversified in tropical Asia (Liu et al. 2010), and the *Musa* lineage ancestor originated ~50 Mya (Christelova et al. 2011). Likewise, the *Ixora* genus originated in South-East Asia, in Borneo in particular (Lorence et al. 2007), and its ancestral lineage originated 30 to 50 Mya (Tosh et al. 2013). Therefore, the ancestors of *Musa* and *Ixora* could have shared the same period and geographical origin. The hypothesis of the HT of *Copia25* between the ancestors of *Ixora* and *Musa* is therefore supported by the chronological and geographical distribution of species. This hypothesis is also supported by the high global sequence identity as well as by the *Ks* values, which are much lower for *Copia25* than for the COSII, suggesting that its presence is recent in the *Musa* genome. Furthermore, the phylogeny of *Copia25* RT including the *Musa* and Rubiaceae species sequences clearly indicates a strong relationship between *Copia25-Musa* and *Copia25-Ixora* (Fig. 4). This relationship does not result from similar selective pressure acting in both groups (as showed by LRT analyses, which exclude purifying selection as the process responsible for sequence similarity) and thus reinforces the proposition of HT. The putative period of *Copia25* transfer from *Ixora* to *Musa* can be estimated by the molecular clock equation using the RT sequences (375 nt; *Ks* ranged from 0.25 to 0.56). The estimated age range from 19 to 43 Mya is congruent with the period when the ancestors of both genera shared geographical distribution. This estimation must be considered with caution because of the short sequence used for establishing the time of divergence and because the molecular clock used is not calibrated for Rubiaceae. Our results thus suggest a potential and ancestral HT of *Copia25* from *Ixora* to *Musa* (Fig. S8).

With the facility for plants to inter-cross and given the autonomy of their germ line, plant genomes have a natural propensity to transfer genetic material. They also have a high content of LTR-RTs, elements whose cytoplasmic multiplication phase heightens the likelihood of being captured and exchanged among other species, thus favoring potential HT. Thanks to the fast-growing number of data sequences available, more studies are being conducted involving several species. Their results reveal scenarios of complex evolution, particularly those concerning TEs. Here, our detailed analyses of *Copia25* in angiosperms disclose the complexity of the evolutionary dynamics of this ancient element, involving several processes including sequence conservation, rapid turnover, stochastic losses and horizontal transfer. Additional information on the presence and the activity of *Copia25* in angiosperms is required to precisely identify the mechanism involved in such remarkable conservation of a transposable element harbored by large and divergent groups of plant species.

Acknowledgments

This research was supported Agropolis Fondation through the “Investissement d’avenir” program (ANR-10-LABX-0001-01) under the reference ID 1002-009 and 1102-006, CAPES (Grants 01/2010 to CMAC and fellowship 9127-11-9 to ESD), Brazilian agencies FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo - Grant 2013/15070-4 to CMAC and fellowship 2011/18226-0 to ESD) and CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico - Grant 306493/2013-6 to CMAC) and French agency ANR (Agence Nationale de la Recherche; Genoplante ANR-08- GENM-022-001). Acknowledgments to Dr. A. D’Hont for providing *Musa* spp. DNA samples; Herman E. Taedoumg for providing *Craterispermum* samples; Dr. P. De Block for providing Rubiaceae samples; Dr. J-J. Rakotomalala for providing Mascarocoffea samples. Acknowledgements to Philippe Lashermes and the Coffee Genome Consortium for the availability of the *C. canephora* BAC-end sequences and draft genome.

Conflict of Interest The authors declare that they have no competing interests.

Electronic supplementary material

The paper contains supplementary material, File 1.

Figure legends

Fig. 1 Distribution and phylogenetic relationship of the copies of *Copia25* identified in the *C. canephora* genome. a Distribution of full-length copies (black lines) and fragmented copies of *Copia25* (red dashes) along the 11 *C. canephora* pseudo-molecules. The gene density along pseudo-molecules is represented in grey while the LTR retrotransposons are represented in red in a separate layer. Fragmented copies are defined as a minimum of 90% nucleotide conservation and 10 to 80% coverage of full-length copies. **b** Phylogeny reconstructed using the *pol* of the full-length copies of *Copia25*. The phylogeny was reconstructed using Neighbor joining, with the distance corrected by General Time Reversible model, and 1000 replicates. All positions containing gaps and missing data were eliminated. There were a total of 2,640 nucleotides in the final dataset. Only the bootstrap values over 70 are shown. Represented in blue are the sequences of Subfamily 1, and in red, Subfamily 2.

Fig. 2 Estimation of the insertion time distribution (in millions of years) of the 72 full-length *Copia25* (Subfamily 1 and 2) copies identified in the *C. canephora* genome. The insertion time was estimated using the Kimura 2-parameter between both LTRs of the same copy and the following molecular clock equation with $r = 1.3 \times 10^{-8}$ (Ma and Bennetzen 2004).

Fig. 3 Phylogeny of the RT domain from sequences similar to the *Copia25* elements in the 29 plant genomes analyzed. The phylogeny was reconstructed using Maximum Likelihood, with the distance corrected by Tamura 3-parameter, and 1000 replicates; the bootstrap consensus tree inferred is taken to represent the evolutionary history of the taxa analyzed. All positions containing gaps and missing data were eliminated. There were a total of 602 nucleotide sites in the final dataset; and a total of 98 nucleotide sequences. A discrete Gamma distribution was used to model evolutionary rate differences among sites (2 categories (+G, parameter = 1.7864)). The rate variation model allowed for some sites to be evolutionarily invariable ([+I], 10.1863% sites). The highlighted clade corresponds to the *Copia25* family; in blue, the monocot species in *Copia25* clade; the number in parentheses is the number of sequences collapsed in the tree. Species abbreviation: *S. tuberosum Solanum tuberosum* (potato), *N. tabacum: Nicotamia tabacum* (tobacco), *N. benthamiana: Nicotamia benthamiana*, *C. canephora*: *Coffea canephora* (coffee), *R. communis; Ricinus communis* (castor oil), *E. guineensis: Elaeis guineensis* (African oil palm), *S. italica: Setaria italica* (Foxtail millet), *S. bicolor: Sorghum bicolor* (sorghum), *O. sativa*: *Oryza sativa* (rice), *T. aestivum: Triticum aestivum* (wheat), *H. vulgare: Hordeum vulgare* (barley), *B. distachyon: Brachypodium distachyon*, *V. vinifera: Vitis vinifera (grape)*, *Gossypium* (cotton), *A. trichopoda: Amborella trichopoda*, *G. max: Glycine max* (soybean), *P. vulgaris: Phaseolus vulgaris* (common bean), *C. cajan: Cajanus cajan* (pigeon pea), *L. japonicus: Lotus japonicus*, *M. truncatula: Medicago truncatula*, *E. grandis: Eucalyptus grandis*, *T. cacao: Theobroma cacao*, *F. ananasa: Fragaria x ananasa* (strawberry), *P. trichocarpa: Populus trichocarpa*, *G. hirsutum: Gossypium hirsutum*, *M. domestica: Malus domestica* (apple), *A. thaliana: Arabidopsis thaliana*, *S. lycopersicum: Solanum lycopersicum* (tomato), *M. guttatus: Mimulus guttatus*, *C. sinensis: Clementina sinensis*, *B. rapa: Brassica rapa*.

Fig. 4 Phylogenetic analysis of *Copia25* RT domain homologs. The phylogeny was reconstructed using Maximum Likelihood, with the distance corrected by Tamura 3-parameter, and 1000 replicates; the tree with the highest log likelihood (-4739.5265) is shown. A discrete Gamma distribution was used to model evolutionary rate differences among sites (2 categories (+G, parameter = 1.1187)). The tree is drawn to scale, with branch lengths measured by number of substitutions per site. All positions containing gaps and missing data were eliminated. There were a total of 313 positions in the final dataset; and a total of 69 nucleotide sequences. Only the bootstrap values over 50% are shown. In green, the clade corresponding to the cluster between *Copia25 Musa* and *Ixora* sequences; in blue, the monocot species. The number of collapsed sequences is indicated in parentheses. Species abbreviation: *S. tuberosum Solanum tuberosum* (potato), *N. tabacum: Nicotamia tabacum* (tobacco), *N. benthamiana Nicotamia benthamiana*, *R. communis; Ricinus communis* (castor oil), *E. guineensis: Elaeis guineensis* (African oil palm) and *C.* means *Coffea.*

Fig. 5 Comparison between *Copia25* and *Copia25-Musa*. a/b Dot plot alignment between the full-length copy of *Copia25* (reference sequences, Subfamilies 1 **(a)** and 2 **(b)**) and the *Copia25-Musa* found in a genomic segment of the *Musa balbisiana* BAC clone (horizontal axis; AC186755 100804-105774). **c** Nucleotide alignment of 5’ LTR of *Copia25* Subfamily 2 and *Copia25-Musa*.

Table 1 Summary of the AAARF assembly. Only contigs larger than 3 Kb (52 over 317) and with a correct assembly structure (37 over 52) were analyzed.

TE classification	Number of identified contigs (> 3Kb)	Number of contigs with EST similarity (E-value <$10e^{-100}$)
Class I LTR retrotransposons	37	26
Class I LTR retrotransposons, *Ty3-Gypsy*	28	22
Class I LTR retrotransposons, *Ty1-Copia*	9	4
Class II transposons	0	0
Total	37	26

Table 2 Estimation of the *Copia25* copy number in *Coffea* genomes using 454 sequencing survey. Only 454 reads with a minimum of 90% of nucleotide identity and over 80% of the read length were considered.

Species	Ploidy level	Estimated genome size (Mb)	#454 sequences	Produced bases (Mb)	Genome coverage %	# of *Copia25* reads	Cumulative length of aligned reads (Kb)	Estimated length in genomes (Kb)
C. canephora (HD94-200)	2x	710	106459	45.05	6.40	70	31,189	487,3
C. canephora (BUD15)	2x	710	149196	67.08	9,58	102	47,092	491,5
C. arabica	4x	1,240	122258	54.5	4.39	85	36,980	842,3
C. eugenioides	2x	645	101309	42.1	6.52	71	30,171	462,7
C. heterocalyx	2x	863	194300	60.51	2.25	42	13,732	610,3
C. racemosa	2x	506	88498	34.19	5.7	179	86,284	1513,7
C. pseudozanguebariae	2x	593	215117	91.7	15.4	68	28,669	186,1
C. humblotiana	2x	469	160479	67.99	14.49	102	45,373	313,3
C. tetragona	2x	513	160107	72.66	14.10	199	97,927	694,5
C. millotii	2x	682	163873	76.65	11.23	95	43,173	384,4
C. horsfieldiana	2x	593*	112793	46.25	7.8	72	29,593	379,3
Craterispermum sp. Novo Kribi	2x	748	49789	19.44	2.59	0	0	0

*: mean value estimates from other ex-*Psilanthus* accessions in absence of clear data for *C. horsfieldiana*.

Table 3 Likelihood ratio test for testing models of sequence evolution for Copia25 retrotransposons.

Model		**Parameter**	**ℓ**	**2Δℓ**	**ω_B**	**ω_F**	**Conclusion**
One-ratio	Model I	ω free	-2469.160	239.308**	0.191	-	Purifying selection in the *Copia25* tree
	Model II	ω = 1	-2588.814		-	-	
Two-ratio	Model III	ω free	-2468.462	33.568**	0.198	0.127	Purifying selection in the *Ixora Copia25* clade
	Model IV	ω = 1	-2485.246		0.198	1	
	Model V	ω free	-2463.734	2.526	0.169	0.552	Neutral evolution in the *Musa Copia25* clade
	Model VI	ω = 1	-2464.998		0.168	1	

Critical values of X^2, 1 df: *: 3.84; **: 6.63; $2\Delta\ell = 2\ (l_1 - l_0)$

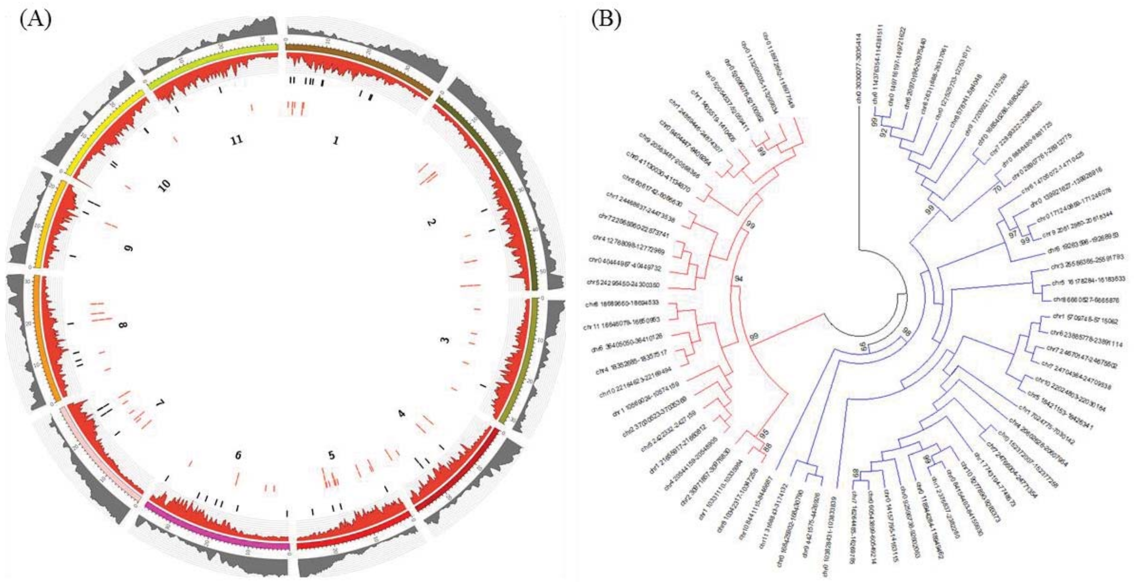


Fig. 1 Distribution and phylogenetic relationship of the copies of *Copia25* identified in the *C. canephora* genome. a Distribution of full-length copies (black lines) and fragmented copies of *Copia25* (red dashes) along the 11 *C. canephora* pseudo-molecules. The gene density along pseudo-molecules is represented in grey while the LTR retrotransposons are represented in red in a separate layer. Fragmented copies are defined as a minimum of 90% nucleotide conservation and 10 to 80% coverage of full-length copies. **b** Phylogeny reconstructed using the *pol* of the full-length copies of *Copia25*. The phylogeny was reconstructed using Neighbor joining, with the distance corrected by General Time Reversible model, and 1000 replicates. All positions containing gaps and missing data were eliminated. There were a total of 2,640 nucleotides in the final dataset. Only the bootstrap values over 70 are shown. Represented in blue are the sequences of Subfamily 1, and in red, Subfamily 2.

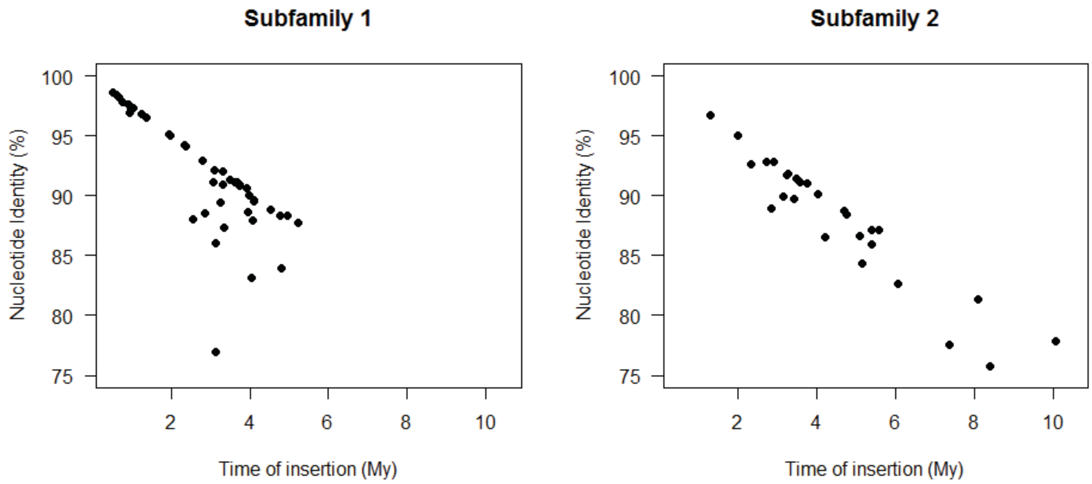


Fig. 2 Estimation of the insertion time distribution (in millions of years) of the 72 full-length *Copia25* (Subfamily 1 and 2) copies identified in the *C. canephora* genome. The insertion time was estimated using the Kimura 2-parameter between both LTRs of the same copy and the following molecular clock equation with r = 1.3 x 10^{-8} (Ma and Bennetzen 2004).

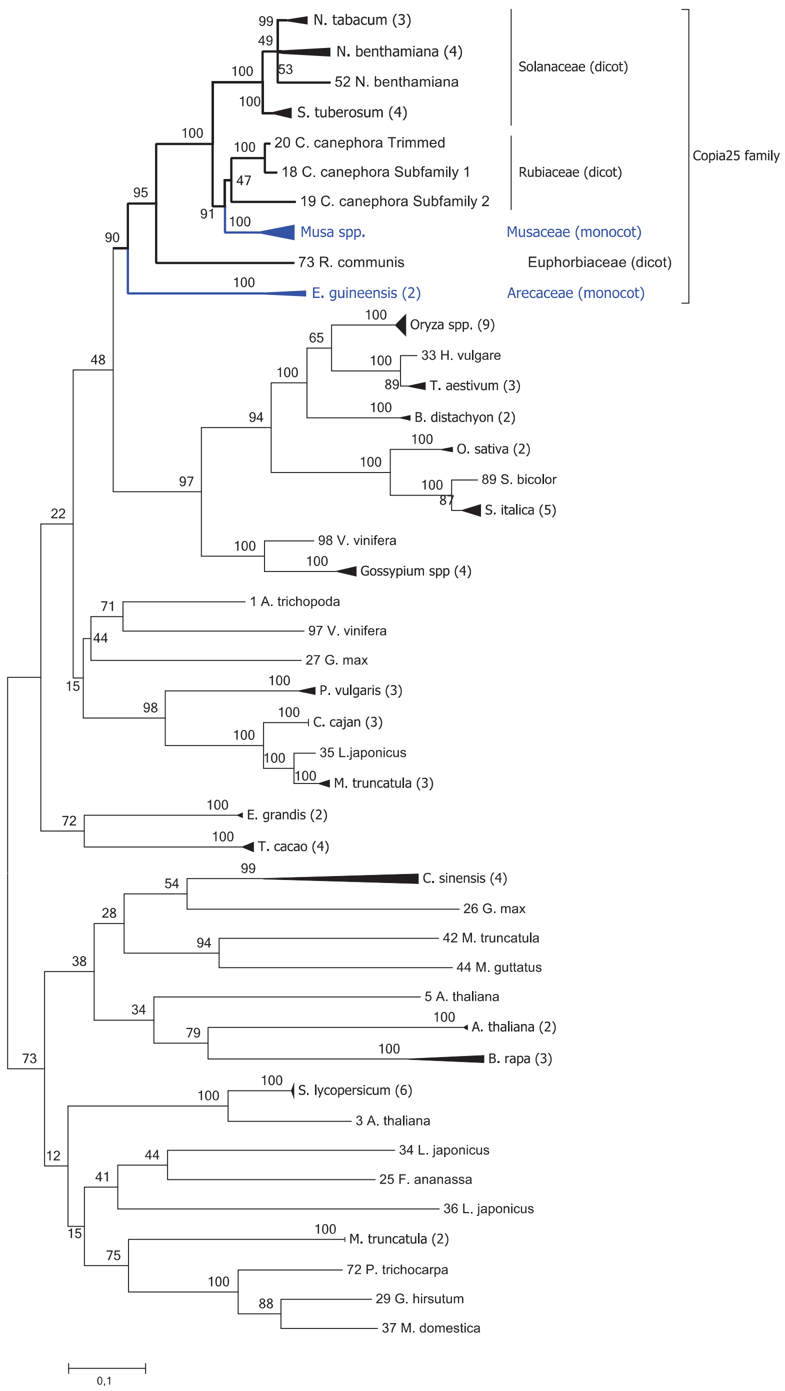


Fig. 3 Phylogeny of the RT domain from sequences similar to the *Copia25* elements in the 29 plant genomes analyzed. The phylogeny was reconstructed using Maximum Likelihood, with the distance corrected by Tamura 3-parameter, and 1000 replicates; the bootstrap consensus tree inferred is taken to represent the evolutionary history of the taxa analyzed. All positions containing gaps and missing data were eliminated. There were a total of 602 nucleotide sites in the final dataset; and a total of 98 nucleotide sequences. A discrete Gamma distribution was used to model evolutionary rate differences among sites (2 categories (+G, parameter = 1.7864)). The rate variation model allowed for some sites to be evolutionarily invariable ([+I], 10.1863% sites). The highlighted clade corresponds to the *Copia25* family; in blue, the monocot species in *Copia25* clade; the number in parentheses is the number of sequences collapsed in the

tree. Species abbreviation: *S. tuberosum Solanum tuberosum* (potato), *N. tabacum: Nicotamia tabacum* (tobacco), *N. benthamiana: Nicotamia benthamiana*, *C. canephora*: *Coffea canephora* (coffee), *R. communis; Ricinus communis* (castor oil), *E. guineensis: Elaeis guineensis* (African oil palm), *S. italica: Setaria italica* (Foxtail millet), *S. bicolor: Sorghum bicolor* (sorghum), *O. sativa*: *Oryza sativa* (rice), *T. aestivum: Triticum aestivum* (wheat), *H. vulgare: Hordeum vulgare* (barley), *B. distachyon: Brachypodium distachyon, V. vinifera: Vitis vinifera (grape), Gossypium* (cotton), *A. trichopoda: Amborella trichopoda, G. max: Glycine max* (soybean), *P. vulgaris: Phaseolus vulgaris* (common bean), *C. cajan: Cajanus cajan* (pigeon pea), *L. japonicus: Lotus japonicus*, *M. truncatula: Medicago truncatula*, *E. grandis: Eucalyptus grandis*, *T. cacao: Theobroma cacao*, *F. ananasa: Fragaria x ananasa* (strawberry), *P. trichocarpa: Populus trichocarpa*, *G. hirsutum: Gossypium hirsutum*, *M. domestica: Malus domestica* (apple), *A. thaliana: Arabidopsis thaliana*, *S. lycopersicum: Solanum lycopersicum* (tomato), *M. guttatus: Mimulus guttatus*, *C. sinensis: Clementina sinensis*, *B. rapa: Brassica rapa*.

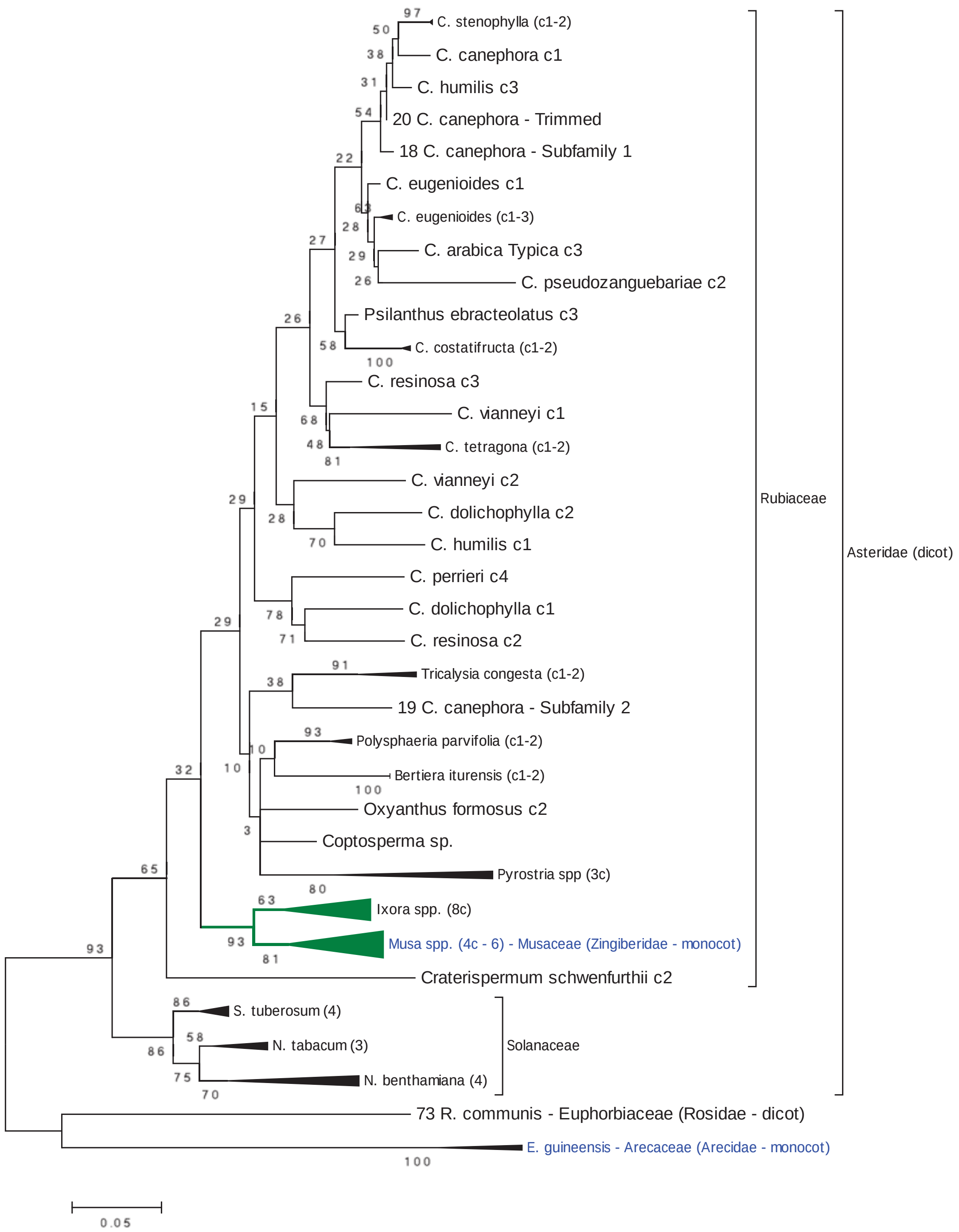


Fig. 4 Phylogenetic analysis of *Copia25* RT domain homologs. The phylogeny was reconstructed using Maximum Likelihood, with the distance corrected by Tamura 3-parameter, and 1000 replicates; the tree with the highest log likelihood (-4739.5265) is shown. A discrete Gamma distribution was used to model evolutionary rate differences among sites (2 categories (+G, parameter = 1.1187)). The tree is drawn to scale, with branch lengths measured by number of substitutions per site. All positions containing gaps and missing data were eliminated. There were a total of 313 positions in the final dataset; and a total of 69 nucleotide sequences. Only the bootstrap values over 50% are shown. In green, the clade corresponding to the cluster between *Copia25 Musa* and *Ixora* sequences; in blue, the monocot species. The number of collapsed sequences is

indicated in parentheses. Species abbreviation: *S. tuberosum Solanum tuberosum* (potato), *N. tabacum: Nicotamia tabacum* (tobacco), *N. benthamiana Nicotamia benthamiana*, *R. communis; Ricinus communis* (castor oil), *E. guineensis: Elaeis guineensis* (African oil palm) and *C.* means *Coffea.*

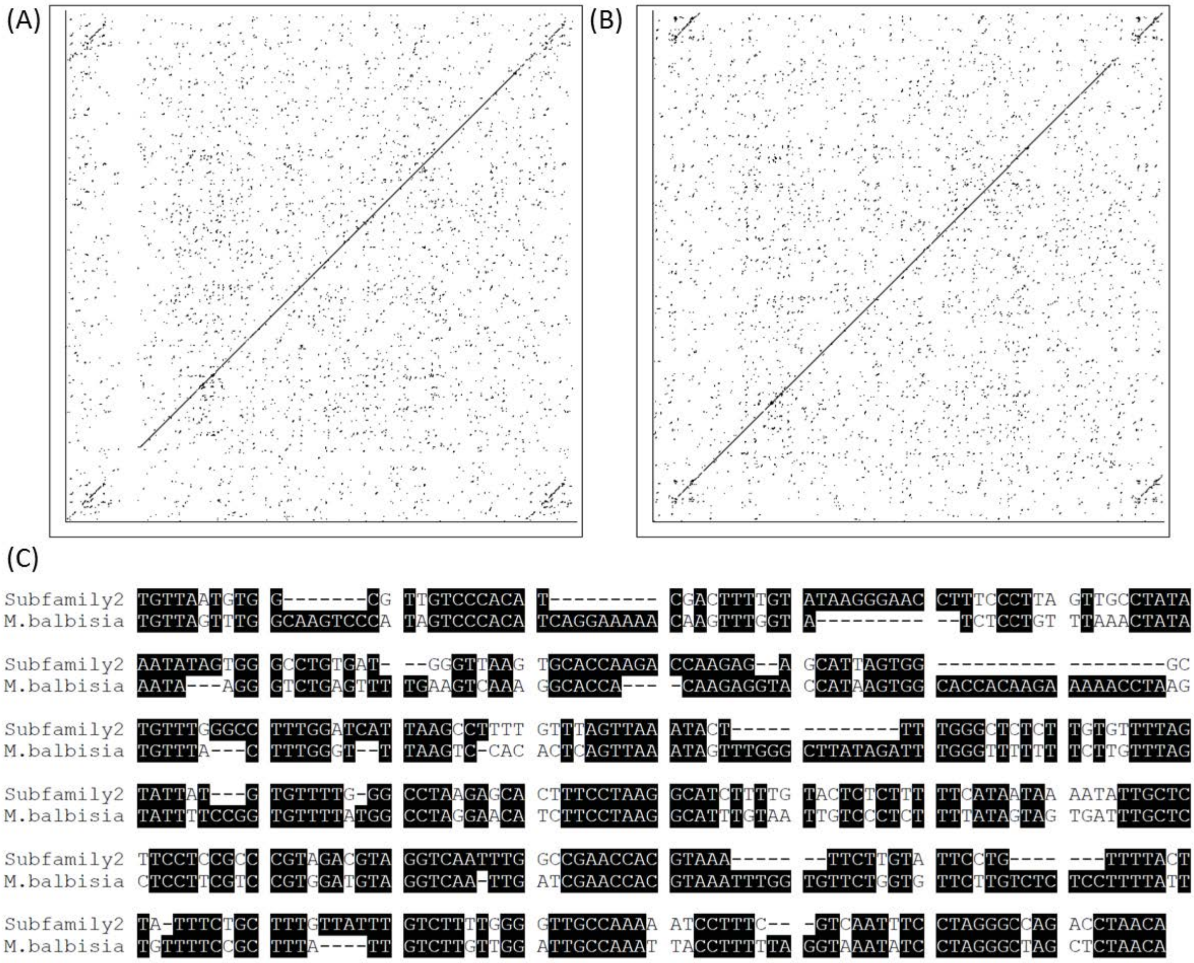


Fig. 5 Comparison between *Copia25* and *Copia25-Musa*. Dot plot alignment between the full-length copy of *Copia25* (reference sequences, **a** Subfamilies 1 and **b** 2) and the *Copia25-Musa* found in a genomic segment of the *Musa balbisiana* BAC clone (horizontal axis; AC186755 100804-105774). **c** Nucleotide alignment of 5’ LTR of *Copia25* Subfamily 2 and *Copia25-Musa*.

References

Anisimova M, Ziheng Y (2007) Multiple Hypothesis Testing to Detect Lineages under Positive Selection that Affects Only a Few Sites. Mol Biol Evol 24:1219–1228

Capy P, Anxolabehere D, Langin T (1994) The strange phylogenies of transposable elements: are horizontal transfers the only explantation? Trends Genet 10:7-12

Carver TJ, Rutherford KM, Berriman M, Rajandream MA, Barrell BG, Parkhill J (2005) ACT: the Artemis Comparison Tool. Bioinformatics 21:3422-3423

Chaw SM, Chang CC, Chen HL, Li WH (2004) Dating the monocot-dicot divergence and the origin of core eudicots using whole chloroplast genomes. J Mol Evol 58:424-441

Cheng X, Zhang D, Cheng Z, Keller B, Ling HQ (2009) A new family of Ty1-copia-like retrotransposons originated in the tomato genome by a recent horizontal transfer event. Genetics 181:1183-1193

Chevalier A (1946) Ecologie et distribution géographique des caféiers sauvages et cultivés. Rev Bot Appl Agric Trop 26:81-94

Christelova P, Valarik M, Hribova E, De Langhe E, Dolezel J (2011) A multi gene sequence-based phylogeny of the Musaceae (banana) family. BMC Evol Biol 11:103

Cummings MP (1994) Transmission patterns of eukaryotic transposable elements: arguments for and against horizontal transfer. Trends Ecol Evol 9:141-145

D'Hont A, Denoeud F, Aury J-M, et al (2012) The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. Nature 488:213–217

Davis AP (2010) Six species of Psilanthus transferred to Coffea (Coffeeae, Rubiaceae). Phytotaxa 10:41-45

Davis AP (2011) Psilanthus mannii, the type species of Psilanthus, transferred to Coffea. Nordic Journal of Botany 29:471-472

de Carvalho MO, Loreto EL (2012) Methods for detection of horizontal transfer of transposable elements in complete genomes. Genetics and molecular biology 35:1078-1084

DeBarry JD, Liu R, Bennetzen JL (2008) Discovery and assembly of repeat family pseudomolecules from sparse genomic sequence data using the Assisted Automated Assembler of Repeat Families (AAARF) algorithm. BMC Bioinformatics 9:235

Denoeud F, Carretero-Paulet L, Dereeper A, et al (2014) The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. Science 345:1181–4

Dereeper A, Guyot R, Tranchant-Dubreuil C, et al (2013) BAC-end sequences analysis provides first insights into coffee (Coffea canephora P.) genome composition and evolution. Plant Mol Biol 83:177–89

Diao X, Freeling M, Lisch D (2006) Horizontal transfer of a plant transposon. PLoS Biol 4:e5

El Baidouri M, Carpentier M-CC, Cooke R, et al (2014) Widespread and frequent horizontal transfers of transposable elements in plants. Genome Res 24:831–8

Ellinghaus D, Kurtz S, Willhoeft U (2008) LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics 9:18

Fortune PM, Roulin A, Panaud O (2008) Horizontal transfer of transposable elements in plants. Commun Integr Biol 1:74-77

Gaut BS, Morton BR, McCaig BC, Clegg MT (1996) Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene Adh parallel rate differences at the plastid gene rbcL. Proc Natl Acad Sci U S A 93:10274-10279

Hamon P, Duroy P-OO, 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 285:447–60

Jiang N, Gao D, Xiao H, van der Knaap E (2009) Genome organization of the tomato sun locus and characterization of the unusual retrotransposon Rider. Plant J 60:181-193

Jiang N, Visa S, Wu S, van der Knaap E (2012) Rider Transposon Insertion and Phenotypic Change in Tomato. Topics in Current Genetics 24:297-312

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

Kumar S, Nei M, Dudley J, Tamura K (2008) MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. Brief Bioinform 9:299-306

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

Lescot M, Piffanelli P, Ciampi A, et al (2008) Insights into the Musa genome: Syntenic relationships to rice and between Musa species. BMC Genomics 9:58.

Librado P, Rozas J (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25:1451-1452

Liu A, Kress W, Li D (2010) Phylogenetic analyses of the banana family (Musaceae) based on nuclear ribosomal (ITS) and chloroplast (trnL-F) evidence. Taxon 59:20-28

Llorens C, Futami R, Covelli L, et al. (2010) The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. Nucleic Acids Res 39:D70-74

Lorence D, Wagner W, Mouly A, Florence J (2007) Revision of Ixora (Rubiaceae) in the Marquesas Islands (French Polynesia). Botanical Journal of The Linnean Society 155:581–597

Ma J, Bennetzen J (2004) Rapid recent growth and divergence of rice nuclear genomes. Proceedings of the National Academy of Sciences of the United States of America 101:12404–12410

Marraccini P, Freire LP, Alves GS, et al (2011) RBCS1 expression in coffee: Coffea orthologs, Coffea arabica homeologs, and expression variability between genotypes and under drought stress. BMC Plant Biol 11:85

Maurin O, Davis AP, Chester M, Mvungi EF, Jaufeerally-Fakim Y, Fay MF (2007) Towards a Phylogeny for Coffea (Rubiaceae): identifying well-supported lineages based on nuclear and plastid DNA sequences. Ann Bot (Lond) 100:1565-1583

Michael TP, Jackson S (2013) The First 50 Plant Genomes. The Plant Genome 6:1-7

Moisy C, Schulman AH, Kalendar R, et al (2014) The Tvv1 retrotransposon family is conserved between plant genomes separated by over 100 million years. Theor Appl Genet 127:1223-35

Moschetto D, Montagnon C, Guyot B, Perriot JJ, Leroy T, Eskes A (1996) Studies on the effect of genotype on cup quality of Coffea canephora. Tropical Science 36:18-31

Roulin A, Piegu B, Wing RA, Panaud O (2008) Evidence of multiple horizontal transfers of the long terminal repeat retrotransposon RIRE1 within the genus Oryza. Plant J 53:950-959

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

SanMiguel P, Tikhonov A, Jin YK, et al. (1996) Nested retrotransposons in the intergenic regions of the maize genome. Science 274: 765-768.

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

Sonnhammer EL, Durbin R (1995) A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. Gene 167:GC1-10

Tamura K, Stecher G, Peterson D, Filipski A (2013) MEGA6: molecular evolutionary genetics analysis version 6.0. Mol Biol Evol 30:2725–2729

Tosh J, Dessein S, Buerki S, et al (2013) Evolutionary history of the Afro-Madagascan Ixora species (Rubiaceae): species diversification and distribution of key morphological traits inferred from dated molecular phylogenetic trees. Annals of Botany 112:1723-1742

Vitte C, Panaud O, Quesneville H (2007) LTR retrotransposons in rice (Oryza sativa, L.): recent burst amplifications followed by rapid DNA loss. BMC Genomics 8:218

Wallau GL, Hua-Van A, Capy P, Loreto EL (2011) The evolutionary history of mariner-like elements in Neotropical drosophilids. Genetica 139:327-338

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, et al (2007) A unified classification system for eukaryotic transposable elements. Nature Reviews Genetics 8:973–982

Wikstrom N, Savolainen V, Chase MW (2001) Evolution of the angiosperms: calibrating the family tree. Proc Biol Sci 268:2211-2220

Wu F, Mueller LA, Crouzillat D, Petiard V, Tanksley SD (2006) Combining bioinformatics and phylogenetics to identify large sets of single-copy orthologous genes (COSII) for comparative, evolutionary and systematic studies: a test case in the euasterid plant clade. Genetics 174:1407-1420

Xiao H, Jiang N, Schaffner E, Stockinger EJ, van der Knaap E (2008) A retrotransposon-mediated gene duplication underlies morphological variation of tomato fruit. Science 319:1527-1530

Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13:555-556

Yang Z, Nielsen R (1998) Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J Mol Evol 46:409-418

Yu Q, Guyot R, Kochko A de, 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 67:305–17

Yuyama PM, Pereira LF, Santos TB dos, et al (2012) FISH using a gag-like fragment probe reveals a common Ty3-gypsy-like retrotransposon in genome of Coffea species. Genome 55:825–33

Supplementary Tables

Table S1 Plant genomes and sequences available analyzed for the presence of sequences similar to *Copia25*.

Species	Classification	Download date	Version	Source	Sequences similar to *Copia25*
Amborella trichopoda	-	-	v. 1.0 - scaffold00202	amborella.org	Yes
Arabidopsis lyrata	Dicot	Nov-13	1.0.20	Gramene	No
Arabidopsis thaliana	Dicot	Oct-13	167	Phytozome	Yes
Brassica rapa	Dicot	Oct-13	197	Phytozome	Yes
Cajanus cajan	Dicot	Oct-13	-	NCBI	Yes
Carica papaya	Dicot	Oct-13	113	Phytozome	No
Citrus sinensis	Dicot	Oct-13	154	Phytozome	Yes
Cucumis sativus	Dicot	Oct-13	122	Phytozome	No
Eucalyptus grandis	Dicot	Oct-13	201	Phytozome	Yes
Fragaria vesca	Dicot	Oct-13	226	Phytozome	No
Fragaria× ananassa	Dicot	-	FJ871121	NCBI	Yes
Glycine max	Dicot	Nov-13	1.0.20	Gramene	Yes
Gossypium hirsutum	Dicot	-	AC243164/AC187141	NCBI	Yes
Gossypium raimondii	Dicot	Oct-13	221	Phytozome	Yes
Linum usitatissimum	Dicot	Oct-13	200	Phytozome	No
Lotus japonicus	Dicot	Nov-13	PRJNA10747	NCBI	Yes
Malus X domestica	Dicot	Oct-13	196	Phytozome	Yes
Mediucago truncatula	Dicot	Oct-13	198	Phytozome	Yes
Mimulus guttatus	Dicot	Oct-13	140	Phytozome	Yes
Nicotiana benthamiana	Dicot	Nov-13	0.4.4	SolNetwork	Yes
Nicotiana tabacum	Dicot	-	website	SolNetwork	Yes
Phaseolus vulgaris	Dicot	Oct-13	218	Phytozome	Yes
Populus trichocarpa	Dicot	Oct-13	210	Phytozome	Yes
Ricinus communis	Dicot	Oct-13	119	Phytozome	Yes
Selaginella moellendorffii	Dicot	Oct-13	91	Phytozome	No
Solanum lycopersicum	Dicot	Oct-13	225	Phytozome	Yes
Solanum tuberosum	Dicot	Oct-13	206	Phytozome	Yes
Theobroma cacao	Dicot	Oct-13	233	Phytozome	Yes
Vinis vinifera	Dicot	Oct-13	145	Phytozome	Yes
Brachypodium distachyon	Monocot	Nov-13	1.0.20	Gramene	Yes
Elaeis guineensis	Monocot	Oct-13	GCA_000442705.1	NCBI	Yes
Hordeum vulgare	Monocot	Nov-13	2.20	Gramene	Yes
Musa accuminata	Monocot	Nov-13	Version 1	cirad fr	Yes
Musa balbisiana	Monocot	-	AC186755	NCBI	Yes
Oryza glaberrima	Monocot	-	AC210484	NCBI	Yes
Oryza sativa	Monocot	Oct-13	204	Phytozome	Yes
Phoenix dactylifera	Monocot	Oct-13	r20101206	NCBI	No
Setaria italica	Monocot	Nov-13	2.0.20	Gramene	Yes
Sorghum bicolor	Monocot	Nov-13	1.20	Gramene	Yes
Triticum aestivum	Monocot	Nov-13	IWGSP1.20	Gramene	Yes
Zea mays	Monocot	Nov-13	3.20	Gramene	No

Table S2 Information on the sequences similar to *Copia25* obtained from 34 plant genomes [RM indicates the retroelements obtained from RepeatMarker database (Giri)].

Identification	Species	Genomic localization or Accession database	Sequence Start	Sequence End	RT Start	RT End
1_A.trichopoda	*Amborella trichopoda*	scaffold00202			3113	3847
2_A.thaliana	*Arabidopsis thaliana*	AB005236			3242	3991
3_A.thaliana	*Arabidopsis thaliana*	AL138663			2962	3711
4_A.thaliana	*Arabidopsis thaliana*	RM ATCOPIA56-I-Copia-Arabidopsis-thaliana			2656	3405
5_A.thaliana	*Arabidopsis thaliana*	RM Copia-4-AT-I-Copia-Arabidopsis-thaliana			2645	3394
6_B.distachyon	*Brachypodium distachyon*	chr1	24306651	24311008	2649	3395
7_B.distachyon	*Brachypodium distachyon*	chr1	37527611	37532031	2717	3430
8_B.rapa	*Brassica rapa*	A01	18910371	18916227	3283	4032
9_B.rapa	*Brassica rapa*	A02	2321990	2316155	3292	4041
10_B.rapa	*Brassica rapa*	A06	17120601	17114901	3303	4052
11_C.cajan	*Cajanus cajan*	CcLG04	3885043	3879934	3067	3816
12_C.cajan	*Cajanus cajan*	Scaffold134199	231135	226018	3066	3815
13_C.cajan	*Cajanus cajan*	Scaffold135625	126929	121836	3048	3797
14_C.sinensis	*Citrus sinensis*	RM Citrus-sinensis-CIRE1.1-Copia-like-AM040263			3018	3761
15_C.sinensis	*Citrus sinensis*	scaffold00012	1165765	1168526	904	1665
16_C.sinensis	*Citrus sinensis*	scaffold00128	299481	296631	3061	3822
17_C.sinensis	*Citrus sinensis*	scaffold00261	101537	99271	2953	3714
18_C.canephora	*Coffea canephora*	chr7	16264485	16269785	1	750
19_C.canephora	*Coffea canephora*	chr8	8081742	8086630	1	750
20_C.canephora	*Coffea canephora*	Trimmed*				
21_E.guineensis	*Elaeis guineensis*	scaffold530537133	35194741	35189675	3143	3892
22_E.guineensis	*Elaeis guineensis*	scaffold530537138	11453672	11458682	3209	3958
23_E.grandis	*Eucalyptus grandis*	scaffold-7	3112959	3105459	4948	5697
24_E.grandis	*Eucalyptus grandis*	scaffold-8	23923876	23929933	3549	4298
25_F.ananassa	*Fragaria× ananassa*	RM Fragaria-x-ananassa-FaRE1-*Ty1*-*Copia*-FJ871121			3061	3810
26_G.max	*Glycine max*	chr9	4345567	4350223	2871	3620
27_G.max	*Glycine max*	RM Copia-20-GM-I-Copia-Glycine-max			2443	3174
28_G.hirsutum	*Gossypium hirsutum*	AC187141			2880	3629
29_G.hirsutum	*Gossypium hirsutum*	AC243157			3163	3915
30_G.hirsutum	*Gossypium hirsutum*	AC243164			2910	3659
31_G.raimondii	*Gossypium raimondii*	Chr02	15738397	15733670	2909	3658
32_G.raimondii	*Gossypium raimondii*	Chr09	48762919	48767369	2508	3257
33_H.vulgare	*Hordeum vulgare*	AC249518			3332	4078
34_L.japonicus	*Lotus japonicus*	AP004495			2839	3588
35_L.japonicus	*Lotus japonicus*	LjChr6	59550618	59543727	3997	4746
36_L.japonicus	*Lotus japonicus*	RM MERE1-LIKE-1-TY-Copia-FJ544857			3158	3895
37_M.domestica	*Malus domestica*	RM Copia-61-Mad-I-Copia-Malus-x-domestica			2682	3434
38_M.truncatula	*Medicago truncatula*	AC127428			2980	3729
39_M.truncatula	*Medicago truncatula*	chr4	19436504	19441021	2814	3563
40_M.truncatula	*Medicago truncatula*	chr7	2556074	2552130	1454	2203
41_M.truncatula	*Medicago truncatula*	RM COP18-I-MT-Copia-Medicago-truncatula			2576	3325
42_M.truncatula	*Medicago truncatula*	RM COP7-I-MT-Copia-Medicago-truncatula			2504	3247
43_M.truncatula	*Medicago truncatula*	RM SHACOP17-I-MT-Copia-Medicago-truncatula			2586	3335
44_M.guttatus	*Mimulus guttatus*	scaffold-401			3217	3966
45_M.acuminata	*Musa acuminata*	CAIC01013562			2957	3706
46_M.acuminata	*Musa acuminata*	chr11	13506541	13517285	2857	3606
47_M.acuminata	*Musa acuminata*	chr8	16753559	16764303	2672	3421
48_M.acuminata	*Musa acuminata*	chrUn-random	1,4E+08	1,4E+08	2932	3681
49_M.acuminata	*Musa acuminata*	chrUn-random	86982745	86993489	2961	3710
50_M.balbisiana	*Musa balbisiana*	AC186755			2984	3733
51_N.benthamiana	*Nicotiana benthamiana*	Niben044Scf00010752	122311	116818	3349	4098
52_N.benthamiana	*Nicotiana benthamiana*	Niben044Scf00037632			110462	1150
53_N.benthamiana	*Nicotiana benthamiana*	Niben044Scf00037679			14784	10293
54_N.benthamiana	*Nicotiana benthamiana*	Niben044Scf00041320			90776	85091
55_N.tabacum	*Nicotiana tabacum*	scaffold114537			1214	1963
56_N.tabacum	*Nicotiana tabacum*	scaffold203761			158	907
57_N.tabacum	*Nicotiana tabacum*	scaffold212962			157	852
58_O.glaberrima	*Oryza glaberrima*	AC210484			3281	4027
59_O.sativa	*Oryza sativa*	AC134048			3245	3991
60_O.sativa	*Oryza sativa*	AP003571			3492	4238
61_O.sativa	*Oryza sativa*	Chr10	9367237	9362589	2786	3532
62_O.sativa	*Oryza sativa*	Chr11	2723167	2717995	3251	3997
63_O.sativa	*Oryza sativa*	Chr12	21824822	21829897	3242	3988
64_O.sativa	*Oryza sativa*	Chr5	24501204	24496123	3242	3988

65_O.sativa	*Oryza sativa*	Chr5	25132942	25138014	3243	3989
66_O.sativa	*Oryza sativa*	Chr6	21442052	21436988	3229	3975
67_O.sativa	*Oryza sativa*	RM COPI1-I-Copia-Oryza-sativa			2537	3283
68_O.sativa	*Oryza sativa*	RM COPIA3-I-OS-Copia-Oryza-sativa-Indica-Group			2970	3716
69_P.vulgaris	*Phaseolus vulgaris*	Chr03	14665482	14670011	3053	3802
70_P.vulgaris	*Phaseolus vulgaris*	Chr10	32070463	32065365	3054	3803
71_P.vulgaris	*Phaseolus vulgaris*	Chr10	32708547	32703567	2940	3689
72_P.trichocarpa	*Populus trichocarpa*	AC212926			3632	4384
73_R.communis	*Ricinus communis*	scaffold29815	55324	60437	3058	3807
74_S.italica	*Setaria italica*	scaffold-2	1689101	1683932	3022	3768
75_S.italica	*Setaria italica*	scaffold-3	38393820	38388423	3211	3957
76_S.italica	*Setaria italica*	scaffold-6	295998	290807	3316	4062
77_S.italica	*Setaria italica*	scaffold-8	34825793	34831166	3150	3896
78_S.italica	*Setaria italica*	scaffold-8	39553498	39559002	3320	4066
79_S.lycopersicum	*Solanum lycopersicum*	AC210359			2905	3654
80_S.lycopersicum	*Solanum lycopersicum*	C02.40-contig15	214976	210319	2807	3544
81_S.lycopersicum	*Solanum lycopersicum*	C05.11-contig15	177859	182529	2807	3556
82_S.lycopersicum	*Solanum lycopersicum*	Copia-2-SL-I-Copia-Solanum-lycopersicum			2507	3256
83_S.lycopersicum	*Solanum lycopersicum*	gi|196192			2910	3659
84_S.lycopersicum	*Solanum lycopersicum*	SL2.40ch12	3162175	3157505	2807	3556
85_S.tuberosum	*Solanum tuberosum*	chr02*			3016	3765
86_S.tuberosum	*Solanum tuberosum*	chr04	3628215	3624135	3124	3873
87_S.tuberosum	*Solanum tuberosum*	chr05	24789816	24784119	3740	4489
88_S.tuberosum	*Solanum tuberosum*	chr01	42417329	42412060	2975	3832
89_S.bicolor	*Sorghum bicolor*	chr7	6045027	6035801	5006	5752
90_T.cacao	*Theobroma cacao*	JN127773			4965	5714
91_T.cacao	*Theobroma cacao*	scaffold-4	24774071	24779084	3025	3744
92_T.cacao	*Theobroma cacao*	scaffold-7	12101319	12106330	2994	3743
93_T.cacao	*Theobroma cacao*	scaffold-7	5599154	5604165	2992	3741
94_T.aestivum	*Triticum aestivum*	DQ890165			3431	4177
95_T.aestivum	*Triticum aestivum*	IWGSC-CSS-5AS-scaff-1517400	1	10586	2406	3152
96_T.aestivum	*Triticum aestivum*	IWGSC-CSS-6DS-scaff-2125824	1	7176	1021	1767
97_V.vinifera	*Vitis vinifera*	chr17	1264226	1269134	2912	3658
98_V.vinifera	*Vitis vinifera*	Copia-79-VV-I-Copia-Vitis-vinifera			2603	3352

Table S3 *Musa* and *Coffea* species used in the PCR analyses (Chevalier, 1946; Maurin et al. 2007).

Class	Family	Subfamily	Tribe	Genus	Species	Botanical Group	Source
Dicot	Rubiaceae	Ixoroideae	1 Bertiereae	*Bertiera*	*B. iturensis*	-	BGM
			2 Coffeeae	*Coffea*	*C. arabica*	Eucoffea	IAC
					C. canephora	Eucoffea	IAC
					C. eugenioides	Eucoffea	IAC
					C. humilis	Eucoffea	IRD
					C. stenophylla	Eucoffea	IRD
					C. millotii (ex-dolichophylla)	Mascarocoffea	FOFIFA
					C. perrieri	Mascarocoffea	FOFIFA
					C. resinosa	Mascarocoffea	FOFIFA
					C. tetragona	Mascarocoffea	FOFIFA
					C. vianneyi	Mascarocoffea	FOFIFA
					C. costatifructa	Mozambicoffea	FOFIFA
					C. pseudozanguebariae	Mozambicoffea	FOFIFA
					C. ebracteolatus (ex *Psilanthus*)	-	IRD
				Tricalysia	*T. congesta*	-	BGM
			3 Gardenieae	*Oxyanthus*	*O. formosus*	-	BGM
			4 Ixoreae	*Ixora*	*I. coccinea*	-	IBILCE
					I. finlaysoniana	-	IBILCE
					I. foliicalyx	-	BGM
					Ixora. spp	-	IBILCE
			5 Octotropideae	*Polysphaeria*	*P. parvifolia*	-	BGM
			6 Pavetteae	*Coptosperma*	*Coptosperma spp*	-	BGM
			7 Vanguerieae	*Pyrostria*	*Pyrostria spp*	-	BGM
		Rubioideae	8 Craterispermeae	*Craterispermum*	*C. schwenfurthii*	-	BGM
Monocot	Musaceae	-		*Musa*	*M. boman*	-	CIRAD
		-			*M. acuminata*	-	CIRAD
		-			*M. balbisiana*	-	CIRAD

BGM: Botanic Garden Meise (Belgium), IAC: Instituto Agronômico de Campinas (Brazil), FOFIFA: National Center for Research Applied to Rural Development (Madagascar)

Table S4 COSII gene accessions used for identity calculation, *Ks* and Ka/Ka estimations.

COSII	Species	Sequence identification	Number of aligned sites
Aspartate-semialdehyde_dehydrogenase	*C. canephora*	chr11_23570374_23566410	474 nt - 158 aa
	M. accuminata	GSMUA_Achr10T18110_001	
	N. benthamiana	Niben044Scf00038253:14188..20177	
	S. tuberosum	gi|565396910	
Biotin synthase	*C. canephora*	chr3_25234051_25234051	279 nt - 93 aa
	M. accuminata	GSMUA_Achr11T08830_001	
	N. benthamiana	Niben044Scf00002035:357..15138	
	S. tuberosum	gi|565361050	
Copper amine oxidase 1-like	*C. canephora*	chr11_32046106_32041371	2100 nt 700 aa
	M. accuminata	GSMUA_Achr10T16890_001	
	N. benthamiana	Niben044Scf00036077:26707..35213	
	S. tuberosum	gi|565382570	
Deoxycytidylate deaminase	*C. canephora*	chr9_3394518_3388836	594 nt - 198 aa
	M. accuminata	GSMUA_Achr11T10880_001	
	N. benthamiana	Niben044Scf00006050:62772..67165	
	S. tuberosum	gi|565369691	
Dynein light chain 1 cytoplasmic	*C. canephora*	chr3_1212408_1211986	171 nt - 57 aa
	M. accuminata	GSMUA_Achr11T11710_001	
	N. benthamiana	Niben044Scf00011546:65168..69639	
	S. tuberosum	gi|565351748	
Glucose 6 phosphate isomerase 1	*C. canephora*	chr10_2235193_2236918	390 nt - 130 aa
	M. accuminata	GSMUA_Achr10T28640_001	
	N. benthamiana	Niben044Scf00015167:103938..129393	
	S. tuberosum	gi|568214480	
Mannosyl-oligosaccharide 12-alpha-mannosidase MNS3	*C. canephora*	chrUn_random_69750499_69749992	399 nt - 133 aa
	M. accuminata	GSMUA_Achr10T14310_001	
	N. benthamiana	Niben044Scf00061720:6187..13106	
	S. tuberosum	gi|565372777	

Table S5 Summary statistics of the two sequencing data sets used from the *C. canephora* DH200-94 accession.

	BESs	**454**	**Total**
Bases	92,046,566	45,058,300	137,104,866
Sequences	134,827	106,459	241,286
Average length (bp)	683	423	568
Min. Length (bp)	60	40	40
Max. Length (bp)	976	764	976
% GC	38.26	39.13	38.64

Table S6 Analysis of the 52 assembled builds showing similarities to RT-LTRs in Repbase (size > 3,000 bp); list of identified builds with their similarities onto *Coffea* BAC sequences and their structural features: 5'LTR-I-LTR3': complete elements, 5'LTR-I, I or I-LTR3': partial elements. In red the *Copia25* LTR-RT.

Name	Family	TE Name	AAARF Contig Name_ *Length*	BLASTn vs *Coffea* BAC E-value < e-100	Build size (bp)	Element size after analysis and correction (bp)	Identified LTR-RT structure
Build#2	**1**	***GYPSY#2***	***HAQIRCC01A7X9T 24745 ‡***	/	**24745**	**11130**	**5'LTR-I-LTR3'**
Build#20	1	*GYPSY#20*	HAQIRCC01ATL3M_6423	/	6423	6423	5'LTR-I
Build#29	1	*GYPSY#20*	HAQIRCC01B3L5T_5684	HQ696507/HQ696509/HQ696510/GU123896	5684	5684	I
Build#3	**1**	***GYPSY#3***	***HAQIRCC01A9HSK_16012 ‡***	/	**16012**	**10316**	**5'LTR-I-LTR3'**
Build#36	**1**	***GYPSY#36***	**HAQIRCC01BGZT3_7900**	/	**7900**	**7900**	**5'LTR-I-LTR3'**
Build#41	1	*GYPSY#41*	HAQIRCC01BMSFQ_3712	/	3712	3712	I-LTR3'
Build#46	**1**	***GYPSY#46***	**HAQIRCC01BSGUF_5337**	**GU123896/HQ696507/HQ696509/HQ696510**	**5337**	**5337**	**I**
Build#5	1	*GYPSY#5*	HAQIRCC01AD2OF_5390	/	5390	5390	I-LTR3'
Build#11	**2**	***GYPSY#11***	***HAQIRCC01AOZMX_13036 ‡***	/	**13036**	**6414**	**I-LTR3'**
Build#32	**2**	***GYPSY#32***	***HAQIRCC01BBGLU_11215 ‡***	/	**11215**	**7943**	**5'LTR-I-LTR3'**
Build#44	**2**	***GYPSY#44***	***HAQIRCC01BQSZZ_14518 ‡***	/	**14518**	**8238**	**5'LTR-I-LTR3'**
Build#47	**2**	***GYPSY#47***	**HAQIRCC01BSTZ2_6194**	/	**6194**	**6194**	**I-LTR3'**
Build#52	**2**	***GYPSY#52***	**HAQIRCC01BXZNV_4535**	/	**4535**	**4535**	**I-LTR3'**
Build#48	3	***GYPSY#48***	HAQIRCC01BTFB3_7038	/	7038	7038	I-LTR3'
Build#33	4	*GYPSY#33*	HAQIRCC01BC7EQ_3250	/	3250	3250	I
Build#9	4	*GYPSY#9*	HAQIRCC01AJR3X_6645	/	6645	6645	I
Build#17	5	*GYPSY#17*	HAQIRCC01ARO50_6446	HQ696507/HQ696509/HQ696510/GU123896	6446	6446	I
Build#13	6	*GYPSY#13*	HAQIRCC01AP9IV_4453	/	4453	4453	I-LTR3'
Build#43	6	*GYPSY#43*	HAQIRCC01BOSZP_6272	GU123897	6272	6272	I-LTR3'
Build#12	**7**	***COPIA#12***	***HAQIRCC01AP281_10765 ‡***	/	**10765**	**5996**	**5'LTR-I-LTR3'**
Build#15	**8**	***COPIA#15***	**HAQIRCC01APR3X_5351**	/	**5351**	**5351**	**I-LTR3'**
Build#21	**8**	***COPIA#21***	**HAQIRCC01AUVBO_3137**	/	**3137**	**3137**	**I**
Build#25	**8**	***COPIA#25***	**HAQIRCC01B0V44_3585**	**HQ696507/HQ832564**	**3585**	**3585**	**I**
Build#42	**9**	***GYPSY#42***	**HAQIRCC01BORW9_4961**	/	**4961**	**4961**	**I**
Build#49	**9**	***GYPSY#49***	**HAQIRCC01BUUFE_3530**	/	**3530**	**3530**	**I**
Build#6	**9**	***GYPSY#6***	**HAQIRCC01AD9P1_4060**	/	**4060**	**4060**	**I**
Build#24	**10**	***GYPSY#24***	**HAQIRCC01B0S4D_9536**	/	**9536**	**9536**	**5'LTR-I**
Build#50	**11**	***GYPSY#50***	**HAQIRCC01BV3ED_4819**	**GU123898**	**4819**	**4819**	**5'LTR-I**
Build#37	12	*GYPSY#37*	HAQIRCC01BJ2DL_3070	HQ696509/GU123894/HQ696510	3070	3070	I
Build#40	**12**	***GYPSY#40***	**HAQIRCC01BLFND_4763**	**GU123894/HQ696509**	**4763**	**4763**	**I**
Build#30	**13**	***COPIA#30***	**HAQIRCC01BA548_4534**	/	**4534**	**4534**	**I**
Build#18	**14**	***COPIA#18***	***HAQIRCC01ASBNT_16928 ‡***	***GU123897***	**16928**	**10629**	**5'LTR-I-LTR3'**
Build#38	14	*COPIA#38*	*HAQIRCC01BL57D_8463 ‡*	/	8463	4094	5'LTR-I-LTR3'
Build#8	14	*COPIA#8*	*HAQIRCC01AHRBL_5573 ‡*	/	5573	4420	5'LTR-I-LTR3'
Build#35	15	*GYPSY#35*	HAQIRCC01BG1JZ_4363	/	4363	4363	5'LTR-I
Build#22	**16**	***COPIA#22***	***HAQIRCC01AWPCQ_4834 ‡***	***GU123895***	**4834**	**4154**	**I-LTR3'**
Build#45	17	*GYPSY#45*	HAQIRCC01BS3HL_3250	/	3250	3250	I-LTR3'
Build#1	/	/	HAQIRCC01A24DQ_5543†	/	5543	/	No
Build#10	/	/	HAQIRCC01AOR7U_5170†	/	5170	/	No
Build#14	/	/	HAQIRCC01APQN0_4330†	/	4330	/	No
Build#16	/	/	HAQIRCC01AR8BG_5051†	/	5051	/	No
Build#19	/	/	HAQIRCC01ASEDX_3085†	/	3085	/	No
Build#23	/	/	HAQIRCC01AYZUV_9903†	/	9903	/	No
Build#26	/	/	HAQIRCC01B1L0V_3437	/	3437	/	No
Build#27	/	/	HAQIRCC01B1L72_17548†	/	17548	/	No
Build#28	/	/	HAQIRCC01B2GPW_10949†	/	10949	/	No
Build#31	/	/	HAQIRCC01BAEQF_7685†	/	7685	/	No
Build#34	/	/	HAQIRCC01BD83A_3689†	/	3689	/	No
Build#39	/	/	HAQIRCC01BLANA_4530	/	4530	/	No
Build#4	/	/	HAQIRCC01AB8JZ_8723	/	8723	/	No
Build#51	/	/	HAQIRCC01BWSAB_3214†	/	3214	/	No
Build#7	/	/	HAQIRCC01AG1PD_3286†	/	3286	/	No

Table S7 Analysis of the 52 assembled builds with similarities with RT-LTRs Repbase (size > 3,000 bp); list of identified builds with their similarities against LTR retrotransposon domains (GAG: capsid, RT: Reverse transcriptase, INT: Integrase, RNAseH, AP: Aspartic protease) and against available ESTs.

Name	TE Name	BLASTx vs REPBASE protein	E-value	GAG	E-value	RT	E-value	INT	E-value	RNaseH	E-value	AP	E-value	Best EST can	E-value
Build#2	***GYPSY*#2**	***Gypsy*-26_ST-I_1p:ClassI:LTR:*Gypsy***	**e-74**	**Del**	**e-07**	**Tma**	**e-16**	**Retrosat-**	**e-35**	**Retrosat-**	**e-07**	**/**	**/**	**GW345818**	**0.0**
Build#20	*GYPSY*#20	*Gypsy*-5_CP-I_1p:ClassI:LTR:*Gypsy*	e-12	Del	e-10	/	/	/	/	/	/	/	/	DV666953	0.0
Build#29	*GYPSY*#20	*Gypsy*-93_ZM-I_2p:ClassI:LTR:*Gypsy*	e-20	Retrosat-2	e-15	/	/	/	/	/	/	/	/	DV665686	0.0
Build#3	***GYPSY*#3**	***Gypsy*-49_Mad-I_1p:ClassI:LTR:*Gypsy***	**e-75**	**Retrosat-2**	**e-05**	**Del**	**e-19**	**Retrosat-**	**e-38**	**Peabody**	**e-17**	**/**	**/**	**DV710342**	**0.0**
Build#36	***GYPSY*#36**	***Gypsy*63-ZM_I_2p:ClassI:LTR:*Gypsy***	**e-97**	**/**	**/**	**Legolas**	**e-49**	**Retrosat-**	**e-69**	**Tma**	**e-17**	**/**	**/**	**DV675167**	**0.0**
Build#41	*GYPSY*#41	GYPSODE1_I_1p:ClassI:LTR:*Gypsy*	e-144	/	/	Legolas	e-25	Tma	e-70	Peabody	e-40	/	/	/	/
Build#46	***GYPSY*#46**	***Gypsy*-6_CP-I_1p:ClassI:LTR:*Gypsy***	**0**	**/**	**/**	**Legolas**	**e-46**	**Peabody**	**e-51**	**Peabody**	**e-49**	**Del**	**e-07**	**DV711786**	**0.0**
Build#5	*GYPSY*#5	*Gypsy*-17_SB-I_1p:ClassI:LTR:*Gypsy*	e-97	/	/	Retrosat-	e-51	Tma	e-41	Bagy-1	e-31	/	/	DV678974	0.0
Build#11	***GYPSY*#11**	***Gypsy*-29_PTr-I_1p:ClassI:LTR:*Gypsy***	**e-99**	**/**	**/**	**CRM**	**e-22**	**CRM**	**e-41**	**CRM**	**e-11**	**CRM**	**e-06**	**GW345937**	**e-157**
Build#32	***GYPSY*#32**	***Gypsy*-27_PTr-I_2p:ClassI:LTR:*Gypsy***	**e-72**	**/**	**/**	**/**	**/**	**Cereba**	**e-41**	**CRM**	**e-13**	**/**	**/**	**DV665803**	**e-172**
Build#44	***GYPSY*#44**	***Gypsy*-27_PTr-I_2p:ClassI:LTR:*Gypsy***	**e-76**	**CRM**	**e-10**	**CRM**	**e-32**	**Beetle1**	**e-30**	**CRM**	**e-14**	**Beetle1**	**e-06**	**GW345826**	**0.0**
Build#47	***GYPSY*#47**	***Gypsy*-168_ZM-I_1p:ClassI:LTR:*Gypsy***	**e-151**	**CRM**	**e-41**	**CRM**	**e-36**	**CRM**	**e-65**	**CRM**	**e-40**	**CRM**	**e-22**	**DV704226**	**0.0**
Build#52	***GYPSY*#52**	***Gypsy*-33_ST-I_1p:ClassI:LTR:*Gypsy***	**e-125**	**/**	**/**	**CRM**	**e-41**	**CRM**	**e-58**	**CRM**	**e-15**	**Beetle1**	**e-22**	**GW345815**	**0.0**
Build#48	***GYPSY*#48**	*Gypsy*-6_CP-I_1p:ClassI:LTR:*Gypsy*	e-06	/	/	/	/	/	/	/	/	/	/	DV673458	0.0
Build#33	*GYPSY*#33	Ogre-SD1_ORF2+3:ClassI:LTR:*Gypsy*	e-32	ORF2_Ogre	e-27	/	/	/	/	/	/	/	/	/	/
Build#9	*GYPSY*#9	Ogre-LE1_ORF1#2:ClassI:LTR:*Gypsy*	e-13	ORF2_Ogre	e-07	/	/	/	/	/	/	/	/	EE198708	e-103
Build#17	*GYPSY*#17	*Gypsy*-5_CP-I_1p:ClassI:LTR:*Gypsy*	e-12	Retrosat-2	e-10	/	/	/	/	/	/	/	/	DV672304	0.0
Build#13	*GYPSY*#13	*Gypsy*-6_CP-I_1p:ClassI:LTR:*Gypsy*	e-47	/	/	/	/	Peabody	e-39	/	/	/	/	GW345831	0.0
Build#43	*GYPSY*#43	GYPSODE1_I_1p:ClassI:LTR:*Gypsy*	e-44	/	/	/	/	Peabody	e-28	Peabody	e-14	/	/	DV704203	0.0
Build#12	***COPIA*#12**	**COP2_I_MT_1p:ClassI:LTR:*Copia***	**e-110**	**/**	**/**	**Mtanga**	**e-07**	**/**	**/**	**/**	**/**	**/**	**/**	**/**	**/**
Build#15	***COPIA*#15**	***Copia*-2_PTri-I_1p:ClassI:LTR:*Copia***	**e-162**	**/**	**/**	**Sto-4**	**e-18**	**Sto-4**	**e-18**	**Sto-4**	**e-39**	**/**	**/**	**DV667278**	**0.0**
Build#21	***COPIA*#21**	***Copia*-93_ST-I_1p:ClassI:LTR:*Copia***	**e-164**	**/**	**/**	**Tork4**	**e-69**	**Tork4**	**e-39**	**Tork4**	**e-13**	**/**	**/**	**/**	**/**
Build#25	***COPIA*#25**	**COP18_I_MT_p#2:ClassI:LTR:*Copia***	**e-50**	**/**	**/**	**Tork4**	**e-20**	**Tork4**	**e-19**	**Tork4**	**e-34**	**Tnt-1**	**e-06**	**/**	**/**
Build#42	***GYPSY*#42**	***Gypsy*-34_VV-I_2p:ClassI:LTR:*Gypsy***	**e-138**	**/**	**/**	**Tft2**	**e-22**	**Ogre**	**e-24**	**/**	**/**	**/**	**/**	**DV673844**	**0.0**
Build#49	***GYPSY*#49**	***GYPSY*3-I_MT_2p:ClassI:LTR:*Gypsy***	**e-138**	**/**	**/**	**Ogre**	**e-75**	**/**	**/**	**Ogre**	**e-30**	**/**	**/**	**/**	**/**
Build#6	***GYPSY*#6**	**Ogre-**	**e-59**	**/**	**/**	**Tft2**	**e-18**	**Ogre**	**e-14**	**Ogre**	**e-05**	**/**	**/**	**DV666858**	**0.0**
Build#24	***GYPSY*#24**	***Gypsy*-32_ST-I_2p:ClassI:LTR:*Gypsy***	**e-78**	**/**	**/**	**Athila4-**	**e-30**	**Cyclops-2**	**e-33**	**Cyclops-2**	**e-12**	**/**	**/**	**DV678819**	**0.0**
Build#50	***GYPSY*#50**	**ATLANTYS1p1:ClassI:LTR:*Gypsy***	**e-58**	**RetroSor1**	**e-09**	**Cinful-1**	**e-22**	**/**	**/**	**Tat4-1**	**e-08**	**RetroSor1**	**e-15**	**/**	**/**
Build#37	*GYPSY*#37	DIASPORA_I_1p#2:ClassI:LTR:*Gypsy*	e-86	Diaspora	e-55	/	/	/	/	/	/	Diaspora	e-19	/	/
Build#40	***GYPSY*#40**	***Gypsy*-34_Mad-I_2p:ClassI:LTR:*Gypsy***	**e-86**	**Diaspora**	**e-15**	**Athila4-**	**e-19**	**/**	**/**	**Diaspora**	**e-45**	**/**	**/**	**DV676094**	**0.0**
Build#30	***COPIA*#30**	***Copia*7-ZM_I_1p:ClassI:LTR:*Copia***	**e-110**	**/**	**/**	**Sto-4**	**e-12**	**Sto-4**	**e-15**	**Vitico1-2**	**e-26**	**/**	**/**	**/**	**/**
Build#18	***COPIA*#18**	***Copia*-34_SB-I_3p:ClassI:LTR:*Copia***	**e-58**	**/**	**/**	**Opie-2**	**e-40**	**SIRE1-4**	**e-29**	**TSI-9**	**e-11**	**ToRTL1**	**e-8**	**/**	**/**
Build#38	*COPIA*#38	Gm*COPIA*10_1p#2:ClassI:LTR:*Copia*	e-8	/	/	/	/	/	/	/	/	/	/	GW345808	0.0
Build#8	*COPIA*#8	Gm*COPIA*10_1p#2:ClassI:LTR:*Copia*	e-16	ToRTL1	e-10	/	/	/	/	/	/	/	/	DV702511	0.0
Build#35	*GYPSY*#35	*Gypsy*-6_CP-I_1p:ClassI:LTR:*Gypsy*	e-12	Del	e-06	/	/	/	/	/	/	/	/	DV668142	0.0
Build#22	***COPIA*#22**	***Copia*-94_VV-I_1p:ClassI:LTR:*Copia***	**e-122**	**/**	**/**	**Tork4**	**e-11**	**Xanthias**	**e-16**	**Melmoth**	**e-15**	**/**	**/**	**EE193974**	**0.0**
Build#45	*GYPSY*#45	GYLES1_I_1p:ClassI:LTR:*Gypsy*	e-48	/	/	/	/	Galadriel	e-35	Galadriel	e-13	/	/	/	/
Build#1	/	*Gypsy*-2_Pru-I_1p:ClassI:LTR:*Gypsy*	e-36	/	/	/	/	/	/	/	/	/	/	/	/
Build#10	/	Gm*COPIA*10_1p#2:ClassI:LTR:*Copia*	e-27	/	/	/	/	/	/	/	/	/	/	/	/
Build#14	/	Gret1_I_1p#2:ClassI:LTR:*Gypsy*	e-24	/	/	/	/	/	/	/	/	/	/	/	/
Build#16	/	*Gypsy*-6_CP-I_1p:ClassI:LTR:*Gypsy*	e-5	/	/	/	/	/	/	/	/	/	/	/	/
Build#19	/	*Gypsy*-4_CP-I_1p:ClassI:LTR:*Gypsy*	e-32	/	/	/	/	/	/	/	/	/	/	/	/
Build#23	/	*Gypsy*-25_ST-I_1p:ClassI:LTR:*Gypsy*	e-46	/	/	/	/	/	/	/	/	/	/	/	/
Build#26	/	Helitron-	e-7	/	/	/	/	/	/	/	/	/	/	/	/
Build#27	/	*Gypsy*-31_ST-I_1p:ClassI:LTR:*Gypsy*	e-6	/	/	/	/	/	/	/	/	/	/	/	/
Build#28	/	*Gypsy*-4_CP-I_1p:ClassI:LTR:*Gypsy*	e-5	/	/	/	/	/	/	/	/	/	/	/	/
Build#31	/	*Gypsy*-5_CP-I_1p:ClassI:LTR:*Gypsy*	e-12	/	/	/	/	/	/	/	/	/	/	/	/
Build#34	/	*Gypsy*-46_Mad-I_3p:ClassI:LTR:*Gypsy*	e-38	/	/	/	/	/	/	/	/	/	/	/	/
Build#39	/	Helitron-	e-11	/	/	/	/	/	/	/	/	/	/	/	/
Build#4	/	Helitron-	e-38	/	/	/	/	/	/	/	/	/	/	/	/
Build#51	/	COP7_I_MT_1p:ClassI:LTR:*Copia*	e-17	/	/	/	/	/	/	/	/	/	/	/	/
Build#7	/	Gm*GYPSY*11_I_1p:ClassI:LTR:?	e-84	/	/	/	/	/	/	/	/	/	/	/	/

Table S8 Characteristics of the *Copia25* copies found in *C. canephora* genome: in red, Sub-family 1 and in blue, Sub-family 2.

Identification/Localization	Element Length	LTR Length 5'	LTR Length 3'	Subfamily	LTR Identity (%)	Gap (%)	Distance (K2p)	Age of Insertion (Mya)
chr0_14157795-14163115	5321	511	511	1	98.6	0.0	0.014	0.53
chr0_60543899-60549214	5316	522	522	1	98.5	0.0	0.016	0.60
chr0_92596738-92602063	5326	519	519	1	98.3	0.0	0.018	0.68
chr7_16264485-16269785	5301	509	509	1	97.8	0.2	0.020	0.77
chr1_7743194-7748573	5380	553	553	1	97.6	0.0	0.024	0.92
chr10_9277890-9283373	5484	552	552	1	96.9	0.7	0.024	0.93
chr1_2376837-2382280	5444	573	573	1	97.2	0.3	0.025	0.97
chr0_84154493-84159930	5438	572	572	1	97.4	0.0	0.027	1.03
chr4_20602628-20607954	5327	509	509	1	96.9	0.0	0.032	1.24
chr0_118944284-118949462	5179	467	467	1	96.6	0.0	0.035	1.36
chr0_152372007-152377258	5252	496	496	1	95.2	0.0	0.050	1.94
chr7_24766004-24771354	5351	527	527	1	95.1	0.0	0.052	1.98
chr5_18421153-18426341	5189	416	416	1	94.2	0.0	0.061	2.33
chr0_103828431-103833839	5409	577	577	1	94.1	0.0	0.062	2.38
chr5_16178284-16183633	5350	542	542	1	88.0	6.1	0.066	2.56
chr9_20612980-20618344	5365	556	556	1	93.0	0.2	0.073	2.80
chr7_24704364-24709538	5175	576	576	1	88.5	4.9	0.074	2.86
chr7_22859322-22864620	5299	520	520	1	91.2	1.5	0.080	3.07
chr6_23885778-23891114	5337	575	575	1	92.2	0.3	0.080	3.09
chr0_168540286-168545362	5077	235	235	1	77.0	16.6	0.081	3.13
chr7_24670147-24675502	5356	575	575	1	86.4	6.8	0.082	3.14
chr11_3168843-3174132	5290	540	540	1	89.4	3.0	0.084	3.24
chr0_28907761-28912775	5015	175	175	1	92.0	0.0	0.086	3.30
chr1_7024775-7030142	5368	506	506	1	90.9	1.2	0.086	3.32
chr8_6660527-6665876	5350	537	537	1	87.3	5.0	0.086	3.33
chr0_171240859-171246078	5220	556	556	1	91.4	0.2	0.091	3.49
chr0_168425902-168430790	4889	68	68	1	91.2	0.0	0.094	3.62
chr0_139921627-139926918	5292	556	556	1	91.2	0.0	0.096	3.68
chr3_25586385-25591793	5409	555	555	1	91.0	0.2	0.097	3.72
chr9_17209921-17215259	5339	526	526	1	90.9	0.2	0.097	3.73
chr10_22024803-22030184	5382	575	575	1	91.0	0.2	0.097	3.74
chr0_3030077-3035414	5338	511	511	1	90.6	0.0	0.102	3.92
chr6_14705072-14710425	5354	548	548	1	88.7	2.0	0.103	3.96
chr10_8441115-8446687	5573	414	414	1	90.1	0.5	0.103	3.97
chr0_114376354-114381511	5158	398	398	1	83.2	8.0	0.105	4.05
chr1_5709748-5715062	5315	533	533	1	88.0	2.6	0.106	4.07
chr6_20970196-20975440	5245	509	509	1	89.6	0.8	0.107	4.10
chr6_26311688-26317061	5374	523	523	1	89.7	0.6	0.107	4.10
chr9_4421575-4426926	5352	539	539	1	88.9	0.6	0.118	4.54
chr0_9886490-9891725	5236	417	533	1	68.8	22.8	0.120	4.62
chr0_127525733-127531017	5285	548	548	1	88.3	0.5	0.124	4.78
chr0_149716197-149721622	5426	578	578	1	83.9	5.5	0.125	4.79
chr6_19263596-19268953	5358	543	543	1	88.4	0.0	0.129	4.96
chr6_578741-584048	5308	546	546	1	87.7	0.2	0.136	5.22
chr1_24468637-24473538	4902	405	405	2	96.8	0.0	0.033	1.28
chr8_8081742-8086630	4889	404	404	2	95.0	0.0	0.052	2.00
chr1_24869446-24874307	4862	407	407	2	92.6	1.7	0.061	2.33
chr0_118972652-118977549	4898	407	407	2	92.9	0.5	0.070	2.71
chr4_12768098-12772969	4872	406	406	2	88.9	4.4	0.074	2.85
chr1_1405519-1410405	4887	407	407	2	92.9	0.0	0.076	2.91
chr9_20563487-20568366	4880	406	406	2	89.9	2.7	0.082	3.14
chr0_40444987-40449732	4746	255	255	2	91.8	0.4	0.084	3.23
chr0_41130030-41134870	4841	407	407	2	91.9	0.2	0.085	3.26
chr7_22868860-22873741	4882	408	408	2	89.7	2.2	0.089	3.42
chr0_52096076-52100952	4877	407	407	2	91.4	0.2	0.090	3.47
chr0_52054537-52059411	4875	408	408	2	91.2	0.2	0.093	3.58
chr0_9404447-9409264	4818	401	401	2	91.0	0.0	0.098	3.75
chr8_18689660-18694533	4874	405	405	2	90.1	0.2	0.105	4.04
chr5_2422332-2427159	4828	395	395	2	86.6	3.8	0.109	4.21
chr4_18352685-18357517	4833	374	374	2	88.8	0.3	0.122	4.71
chr5_24295450-24300350	4901	408	408	2	88.5	0.5	0.124	4.77
chr1_10569024-10574159	5136	406	406	2	86.7	1.5	0.132	5.09
chr11_16646079-16650953	4875	408	408	2	84.3	4.2	0.134	5.15
chr2_37030523-37035369	4847	405	405	2	87.2	0.2	0.140	5.38
chr6_36405050-36410126	5077	398	398	2	85.9	1.8	0.140	5.39
chr10_22164623-22169494	4872	404	404	2	87.1	0.0	0.145	5.58
chr4_20544159-20548905	4747	380	380	2	82.6	3.9	0.158	6.06
chr1_21655917-21660812	4896	430	430	2	77.7	7.2	0.192	7.37
chr2_30971857-30976830	4974	407	407	2	81.3	1.0	0.211	8.11
chr8_10342317-10347258	4942	352	352	2	76.7	7.1	0.218	8.40
chr1_10331110-10335964	4855	390	390	2	77.9	0.8	0.262	10.06
chr0_113295035-113299634	4600	406	-	2	-	-	-	-

Table S9 Nucleotide identity (%) of the Copia25 sequences. A total of 383 nucleotides of the RT region were used; all positions containing gaps and missing data were eliminated. There were a total of 319 positions in the final dataset.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
1_18_C.canephora		98	98	96	96	97	97	96	98	94	94	95	93	92	92	89	91	92	91	90	92	92	89	89	92	89	90	89	91	88	89	89	89	89	87	86
2_C._eugenioides_c1	98		97	98	97	96	95	96	97	94	94	96	93	92	93	91	92	92	92	90	93	93	90	90	92	90	89	88	90	91	91	90	90	90	89	87
3_C._humilis_c3	98	97		96	96	97	97	96	98	93	93	94	93	92	93	89	91	91	91	90	91	92	90	90	92	90	90	89	91	90	90	90	89	89	88	86
4_C._eugenioides_c3	96	98	96		98	95	94	94	96	93	93	96	93	93	92	90	91	92	91	91	93	93	90	90	92	90	90	89	91	90	90	91	90	90	89	87
5_C._eugenioides_c2	96	97	96	98		94	94	94	96	94	94	96	93	93	92	91	91	92	91	91	93	92	90	90	92	90	90	89	90	90	90	90	89	90	88	86
6_C._canephora_c1	97	96	97	95	94		96	96	97	92	92	93	91	90	91	88	89	89	90	89	90	91	88	88	91	88	88	88	91	89	89	88	88	88	88	86
7_C._stenophylla_c1	97	95	97	94	94	96		100	97	95	94	95	93	93	91	87	90	90	91	91	90	90	88	88	91	89	89	88	90	88	89	88	88	87	87	85
8_C._stenophylla_c2	96	96	96	94	94	96	100		97	94	94	96	93	93	91	87	90	90	91	90	90	91	88	88	90	89	89	88	90	88	89	89	88	88	87	85
9_20_C.canephora	98	97	98	96	96	97	97	97		94	94	95	93	92	92	89	90	91	91	91	92	92	89	89	92	89	89	88	90	89	89	89	88	89	87	86
10_C._costatifructa_c1	94	94	93	93	94	92	95	94	94		99	97	93	93	90	89	90	90	90	91	89	89	88	88	90	87	88	87	90	88	89	88	88	87	86	85
11_C._costatifructa_c2	94	94	93	93	94	92	94	94	94	99		96	92	92	90	89	90	90	90	92	89	89	88	88	90	87	88	87	90	88	88	88	87	87	86	85
12_C._ebracteolatus_c3	95	96	94	96	96	93	95	96	95	97	96		94	94	90	89	91	91	91	92	91	91	88	88	91	89	90	88	90	89	90	89	89	89	88	86
13_C._resinosa_c3	93	93	93	93	93	91	93	93	93	93	92	94		93	90	89	89	89	91	91	88	88	88	88	90	90	89	88	89	89	89	89	88	88	87	86
14_C._tetragona_c1	92	92	92	93	93	90	93	93	92	93	92	94	93		89	89	89	90	89	91	88	89	88	88	89	88	87	87	88	87	87	87	87	86	86	84
15_Tricalysia_congesta_c1	92	93	93	92	92	91	91	91	92	90	90	90	90	89		94	90	89	89	87	87	87	90	90	90	86	86	85	85	90	89	89	88	89	89	85
16_Tricalysia_congesta_c2	89	91	89	90	91	88	87	87	89	89	89	89	89	89	94		90	89	89	87	86	86	89	89	89	86	86	84	87	91	88	88	87	88	89	85
17_Polysphaeria_parvifolia_c1	91	92	91	91	91	89	90	90	90	90	90	91	89	89	90	90		97	92	87	87	87	91	91	92	90	91	88	87	91	90	90	89	89	87	85
18_Polysphaeria_parvifolia_c2	92	92	91	92	92	89	90	90	91	90	90	91	89	90	89	89	97		91	87	87	88	90	90	91	89	90	87	88	89	88	88	88	87	86	84
19_Coptosperma_spp	91	92	91	91	91	90	91	91	91	90	90	91	91	89	89	89	92	91		88	87	87	90	90	92	89	90	89	89	90	90	90	89	89	88	86
20_C._vianneyi_c1	90	90	90	91	91	89	91	90	91	91	92	92	91	91	87	87	87	87	88		86	86	86	86	88	86	88	86	87	86	86	86	86	86	85	83
21_C._arabica_Typica_c3	92	93	91	93	93	90	90	90	92	89	89	91	88	88	87	86	87	87	87	86		89	85	85	87	86	86	84	87	85	85	85	84	85	84	82
22_C._pseudozanguebariae_c2	92	93	92	93	92	91	90	91	92	89	89	91	88	89	87	86	87	88	87	86	89		85	85	87	86	87	85	88	86	87	86	86	85	86	84
23_Bertiera_iturensis_c1	89	90	90	90	90	88	88	88	89	88	88	88	88	88	90	89	91	90	90	86	85	85		100	91	87	87	85	85	88	89	89	88	88	88	85
24_Bertiera_iturensis_c2	89	90	90	90	90	88	88	88	89	88	88	88	88	88	90	89	91	90	90	86	85	85	100		91	87	87	85	85	88	89	89	88	88	88	85
25_Oxyanthus_formosus_c2	92	92	92	92	92	91	91	90	92	90	90	91	90	89	90	89	92	91	92	88	87	87	91	91		88	89	87	87	90	90	90	90	90	88	86
26_C._dolichophylla_c1	89	90	90	90	90	88	89	89	89	87	87	89	90	88	86	86	90	89	89	86	86	86	87	87	88		92	90	89	87	88	87	86	87	86	83
27_C._resinosa_c2	90	89	90	90	90	88	89	89	89	88	88	90	89	87	86	86	91	90	90	88	86	87	87	87	89	92		90	90	87	89	88	87	87	86	84
28_C._perrieri_c4	89	88	89	89	89	88	88	88	88	87	87	88	88	87	85	84	88	87	89	86	84	85	85	85	87	90	90		89	86	87	86	86	85	85	82
29_C._dolichophylla_c2	91	90	91	91	90	91	90	90	90	90	90	90	89	88	85	87	87	88	89	87	87	88	85	85	87	89	90	89		86	87	86	86	86	87	84
30_19_C.canephora	88	91	90	90	90	89	88	88	89	88	88	89	89	87	90	91	91	89	90	86	85	86	88	88	90	87	87	86	86		89	89	88	88	87	84
31_45_M.acuminata	**89**	**91**	**90**	**90**	**90**	**89**	**89**	**89**	**89**	**89**	**88**	**90**	**89**	**87**	**89**	**88**	**90**	**88**	**90**	**86**	**85**	**87**	**89**	**89**	**90**	**88**	**89**	**87**	**87**	**89**		**98**	**97**	**97**	**95**	**93**
32_46_M.acuminata	**89**	**90**	**90**	**91**	**90**	**88**	**88**	**89**	**89**	**88**	**88**	**89**	**89**	**87**	**89**	**88**	**90**	**88**	**90**	**86**	**85**	**86**	**89**	**89**	**90**	**87**	**88**	**86**	**86**	**89**	**98**		**98**	**97**	**96**	**93**
33_49_M.acuminata	**89**	**90**	**89**	**90**	**89**	**88**	**88**	**88**	**88**	**88**	**87**	**89**	**88**	**87**	**88**	**87**	**89**	**88**	**89**	**86**	**84**	**86**	**88**	**88**	**90**	**86**	**87**	**86**	**86**	**88**	**97**	**98**		**97**	**94**	**93**
34_48_M.acuminata	**89**	**90**	**89**	**90**	**90**	**88**	**87**	**88**	**89**	**87**	**87**	**89**	**88**	**86**	**89**	**88**	**89**	**87**	**89**	**86**	**85**	**85**	**88**	**88**	**90**	**87**	**87**	**85**	**86**	**88**	**97**	**97**	**97**		**95**	**92**
35_50_M.balbisiana	**87**	**89**	**88**	**89**	**88**	**88**	**87**	**87**	**87**	**86**	**86**	**88**	**87**	**86**	**89**	**89**	**87**	**86**	**88**	**85**	**84**	**86**	**88**	**88**	**88**	**86**	**86**	**85**	**87**	**87**	**95**	**96**	**94**	**95**		**92**
36_47_M.acuminata	**86**	**87**	**86**	**87**	**86**	**86**	**85**	**85**	**86**	**85**	**85**	**86**	**86**	**84**	**85**	**85**	**85**	**84**	**86**	**83**	**82**	**84**	**85**	**85**	**86**	**83**	**84**	**82**	**84**	**84**	**93**	**93**	**93**	**92**	**92**	
37_Ixora_coccinea_c1	89	90	89	90	90	89	88	88	89	88	88	89	87	86	87	85	88	87	88	86	85	86	86	86	89	85	86	86	86	87	92	93	92	92	91	89
38_Ixora_finlaysoniana_c3	90	91	90	91	91	89	88	88	90	89	88	90	87	88	89	88	90	89	89	86	86	86	89	89	91	87	88	86	86	88	93	93	92	93	92	90
39_Ixora_coccinea_c2	90	91	90	91	91	88	88	89	90	89	89	90	88	88	90	89	91	90	90	86	86	86	89	89	91	87	88	86	86	88	92	92	92	92	91	89
40_Ixora_sp_c1	91	92	91	92	92	90	89	89	91	90	89	91	89	89	90	89	90	90	90	88	87	88	88	88	90	86	88	86	87	88	93	93	92	92	91	90
41_Ixora_sp_c2	90	90	89	90	90	88	88	88	89	88	87	89	87	87	89	88	90	89	88	86	85	86	86	86	89	85	87	85	86	88	91	91	91	91	89	88
42_Ixora_foliicalyx_c1	89	90	89	90	90	87	87	88	89	87	87	89	86	86	90	88	89	88	87	85	85	85	87	87	89	85	85	84	84	86	91	91	90	91	90	88
43_Ixora_coccinea_c3	90	91	91	92	91	90	89	89	90	89	88	90	89	87	90	89	90	89	90	87	86	86	88	88	90	88	89	87	87	88	92	93	92	92	91	90
44_Ixora_foliicalyx_c2	88	89	88	89	88	86	86	86	87	87	86	87	85	86	88	87	87	87	87	84	84	85	86	86	87	84	86	84	84	84	90	91	90	90	90	88
45_C._humilis_c1	86	84	85	84	85	85	84	84	85	85	85	85	84	83	80	80	83	83	83	81	80	81	80	80	81	81	82	81	85	81	81	80	80	79	79	78
46_51_N.benthamiana	84	85	84	84	84	83	85	85	84	86	86	85	84	84	83	81	83	82	84	82	79	82	82	82	82	79	81	80	82	80	83	83	82	81	84	81
47_53_N.benthamiana	87	87	86	85	86	85	87	87	86	88	87	86	85	86	84	82	84	83	85	83	80	83	84	84	84	81	82	81	83	82	84	84	83	83	85	82
48_55_N.tabacum	88	88	88	87	87	87	89	89	88	89	89	89	87	87	86	83	86	86	87	86	83	85	84	84	86	83	85	83	85	84	86	87	86	85	86	84
49_56_N.tabacum	87	87	86	86	87	85	87	87	86	88	88	87	86	86	84	81	85	84	86	85	81	83	83	83	84	81	84	81	84	81	84	85	85	84	85	83
50_57_N.tabacum	88	88	87	87	87	86	89	89	88	89	89	89	87	87	86	83	86	85	87	85	83	85	84	84	86	83	84	83	84	83	86	86	86	85	86	86
51_85_S.tuberosum	90	89	90	89	89	88	91	91	90	90	90	90	89	89	87	85	87	86	89	87	84	87	86	86	87	84	86	84	86	85	87	87	87	86	87	85
52_86_S.tuberosum	90	89	90	89	89	88	91	91	90	90	90	90	89	89	87	85	87	86	89	87	84	87	86	86	87	84	86	84	86	85	87	87	87	86	87	85
53_88_S.tuberosum	88	88	89	87	87	87	90	90	88	88	88	88	87	87	86	84	85	85	87	85	83	85	85	85	85	83	84	83	84	84	86	86	85	85	86	83
54_87_S.tuberosum	86	86	87	85	85	85	87	87	86	87	87	86	85	85	85	82	84	83	86	84	80	84	83	83	84	81	83	82	83	82	85	85	84	84	85	82
55_C._vianneyi_c2	90	90	91	91	90	89	89	89	90	89	89	90	90	89	87	88	89	89	89	88	85	87	85	85	89	89	89	89	90	87	88	88	87	87	87	84
56_C._tetragona_c2	88	88	89	89	90	87	89	89	88	89	89	91	90	92	85	86	86	86	86	89	84	86	85	85	86	86	86	84	85	86	86	85	84	84	83	81
57_Craterispermum_schwenfurthii_c2	86	86	87	86	87	85	86	86	86	85	84	85	84	84	85	84	88	87	86	83	82	84	86	86	87	83	84	83	82	84	86	86	86	85	85	83
58_54_N.benthamiana	73	73	72	72	73	72	73	73	73	74	74	74	73	73	71	69	71	70	72	71	68	71	71	71	70	68	69	69	70	69	71	71	71	71	72	70
59_52_N.benthamiana	82	83	81	81	82	81	82	82	82	83	83	83	82	81	80	78	79	79	81	80	77	79	79	79	80	77	79	80	79	78	80	80	81	80	81	79
60_73_R.communis	79	79	79	79	79	77	79	79	79	78	78	79	78	78	77	76	79	79	79	77	75	77	77	77	78	78	77	75	77	76	79	80	80	79	78	78
61_21_E.guineensis	75	76	75	76	77	75	75	75	76	75	75	75	75	74	75	75	74	75	74	74	72	76	75	75	76	73	74	72	74	75	74	74	74	73	75	73
62_22_E.guineensis	75	76	75	76	77	75	76	76	76	76	75	76	74	75	75	75	75	75	74	73	72	76	74	74	75	72	73	72	74	74	74	74	74	74	77	74
63_Pyrostria_sp_c2	82	83	83	83	84	81	81	82	82	81	81	81	81	80	81	80	80	80	81	79	78	79	80	80	82	80	79	79	78	80	81	80	80	80	79	78
64_Pyrostria_sp_c3	82	83	83	83	84	82	81	82	82	81	80	81	81	80	81	80	80	80	82	79	78	79	80	80	82	79	79	79	78	80	80	80	79	79	79	78
65_Pyrostria_sp_c1	85	85	86	85	86	84	84	84	85	85	84	84	83	82	82	80	84	84	84	82	82	82	84	84	84	82	83	79	83	83	83	82	81	82	80	79

Table S9 (continued)

	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65
1_18_C.canephora	89	90	90	91	90	89	90	88	86	84	87	88	87	88	90	90	88	86	90	88	86	73	82	79	75	75	82	82	85
2_C._eugenioides_c1	90	91	91	92	90	90	91	89	84	85	87	88	87	88	89	89	88	86	90	88	86	73	83	79	76	76	83	83	85
3_C._humilis_c3	89	90	90	91	89	89	91	88	85	84	86	88	86	87	90	90	89	87	91	89	87	72	81	79	75	75	83	83	86
4_C._eugenioides_c3	90	91	91	92	90	90	92	89	84	84	85	87	86	87	89	89	87	85	91	89	86	72	81	79	76	76	83	83	85
5_C._eugenioides_c2	90	91	91	92	90	90	91	88	85	84	86	87	87	87	89	89	87	85	90	90	87	73	82	79	77	77	84	84	86
6_C._canephora_c1	89	89	88	90	88	87	90	86	85	83	85	87	85	86	88	88	87	85	89	87	85	72	81	77	75	75	81	82	84
7_C._stenophylla_c1	88	88	88	89	88	87	89	86	84	85	87	89	87	89	91	91	90	87	89	89	86	73	82	79	75	76	81	81	84
8_C._stenophylla_c2	88	88	89	89	88	88	89	86	84	85	87	89	87	89	91	91	90	87	89	89	86	73	82	79	75	76	82	82	84
9_20_C.canephora	89	90	90	91	89	89	90	87	85	84	86	88	86	88	90	90	88	86	90	88	86	73	82	79	76	76	82	82	85
10_C._costatifructa_c1	88	89	89	90	88	87	89	87	85	86	88	89	88	89	90	90	88	87	89	89	85	74	83	78	75	76	81	81	85
11_C._costatifructa_c2	88	88	89	89	87	87	88	86	85	86	87	89	88	89	90	90	88	87	89	89	84	74	83	78	75	75	81	80	84
12_C._ebracteolatus_c3	89	90	90	91	89	89	90	87	85	85	86	89	87	89	90	90	88	86	90	91	85	74	83	79	75	76	81	81	84
13_C._resinosa_c3	87	87	88	89	87	86	89	85	84	84	85	87	86	87	89	89	87	85	90	90	84	73	82	78	75	74	81	81	83
14_C._tetragona_c1	86	88	88	89	87	86	87	86	83	84	86	87	86	87	89	89	87	85	89	92	84	73	81	78	74	75	80	80	82
15_Tricalysia_congesta_c1	87	89	90	90	89	90	90	88	80	83	84	86	84	86	87	87	86	85	87	85	85	71	80	77	75	75	81	81	82
16_Tricalysia_congesta_c2	85	88	89	89	88	88	89	87	80	81	82	83	81	83	85	85	84	82	88	86	84	69	78	76	75	75	80	80	80
17_Polysphaeria_parvifolia_c1	88	90	91	90	90	89	90	87	83	83	84	86	85	86	87	87	85	84	89	86	88	71	79	79	74	75	80	80	84
18_Polysphaeria_parvifolia_c2	87	89	90	90	89	88	89	87	83	82	83	86	84	85	86	86	85	83	89	86	87	70	79	79	75	75	80	80	84
19_Coptosperma_spp	88	89	90	90	88	87	90	87	83	84	85	87	86	87	89	89	87	86	89	86	86	72	81	79	74	74	81	82	84
20_C._vianneyi_c1	86	86	86	88	86	85	87	84	81	82	83	86	85	85	87	87	85	84	88	89	83	71	80	77	74	73	79	79	82
21_C._arabica_Typica_c3	85	86	86	87	85	85	86	84	80	79	80	83	81	83	84	84	83	80	85	84	82	68	77	75	72	72	78	78	82
22_C._pseudozanguebariae_c2	86	86	86	88	86	85	86	85	81	82	83	85	83	85	87	87	85	84	87	86	84	71	79	77	76	76	79	79	82
23_Bertiera_iturensis_c1	86	89	89	88	86	87	88	86	80	82	84	84	83	84	86	86	85	83	85	85	86	71	79	77	75	74	80	80	84
24_Bertiera_iturensis_c2	86	89	89	88	86	87	88	86	80	82	84	84	83	84	86	86	85	83	85	85	86	71	79	77	75	74	80	80	84
25_Oxyanthus_formosus_c2	89	91	91	90	89	89	90	87	81	82	84	86	84	86	87	87	85	84	89	86	87	70	80	78	76	75	82	82	84
26_C._dolichophylla_c1	85	87	87	86	85	85	88	84	81	79	81	83	81	83	84	84	83	81	89	86	83	68	77	78	73	72	80	79	82
27_C._resinosa_c2	86	88	88	88	87	85	89	86	82	81	82	85	84	84	86	86	84	83	89	86	84	69	79	77	74	73	79	79	83
28_C._perrieri_c4	86	86	86	86	85	84	87	84	81	80	81	83	81	83	84	84	83	82	89	84	83	69	80	75	72	72	79	79	79
29_C._dolichophylla_c2	86	86	86	87	86	84	87	84	85	82	83	85	84	84	86	86	84	83	90	85	82	70	79	77	74	74	78	78	83
30_19_C.canephora	87	88	88	88	88	86	88	84	81	80	82	84	81	83	85	85	84	82	87	86	84	69	78	76	75	74	80	80	83
31_45_M.acuminata	92	93	92	93	91	91	92	90	81	83	84	86	84	86	87	87	86	85	88	86	86	71	80	79	74	74	81	80	83
32_46_M.acuminata	93	93	92	93	91	91	93	91	80	83	84	87	85	86	87	87	86	85	88	85	86	71	80	80	74	74	80	80	82
33_49_M.acuminata	92	92	92	92	91	90	92	90	80	82	83	86	85	86	87	87	85	84	87	84	86	71	81	80	74	74	80	79	81
34_48_M.acuminata	92	93	92	92	91	91	92	90	79	81	83	85	84	85	86	86	85	84	87	84	85	71	80	79	73	74	80	79	82
35_50_M.balbisiana	91	92	91	91	89	90	91	90	79	84	85	86	85	86	87	87	86	85	87	83	85	72	81	78	75	77	79	79	80
36_47_M.acuminata	89	90	89	90	88	88	90	88	78	81	82	84	83	86	85	85	83	82	84	81	83	70	79	78	73	74	78	78	79
37_Ixora_coccinea_c1		93	93	93	92	92	94	90	81	81	84	86	85	86	87	87	85	84	87	83	86	71	82	77	73	73	79	79	83
38_Ixora_finlaysoniana_c3	93		97	94	94	95	95	92	80	83	85	86	85	86	87	87	85	84	86	85	86	72	81	78	73	74	80	80	82
39_Ixora_coccinea_c2	93	97		96	95	95	95	93	80	83	85	86	85	85	87	87	86	84	86	85	86	71	80	79	73	73	80	80	82
40_Ixora_sp_c1	93	94	96		97	96	94	95	83	85	86	87	86	87	89	89	87	86	87	85	86	73	83	79	74	74	81	81	82
41_Ixora_sp_c2	92	94	95	97		95	93	93	83	83	85	86	85	85	88	88	86	84	86	84	86	72	81	79	73	74	79	79	81
42_Ixora_foliicalyx_c1	92	95	95	96	95		93	94	80	82	84	85	84	85	87	87	85	84	84	83	85	71	80	79	74	74	79	79	80
43_Ixora_coccinea_c3	94	95	95	94	93	93		91	80	81	83	85	84	84	86	86	85	83	87	86	86	70	79	78	72	72	81	81	83
44_Ixora_foliicalyx_c2	90	92	93	95	93	94	91		80	83	83	85	85	86	87	87	85	85	84	82	85	71	80	76	72	73	79	78	81
45_C._humilis_c1	81	80	80	83	83	80	80	80		77	78	79	78	79	81	81	79	78	83	80	77	65	74	72	69	69	74	73	77
46_51_N.benthamiana	81	83	83	85	83	82	81	83	77		90	92	91	92	91	91	90	89	82	80	81	78	86	74	73	76	75	75	77
47_53_N.benthamiana	84	85	85	86	85	84	83	83	78	90		92	91	92	93	93	91	90	82	81	82	80	88	75	73	76	74	75	78
48_55_N.tabacum	86	86	86	87	86	85	85	85	79	92	92		95	95	95	95	94	93	85	83	84	78	89	78	74	77	77	77	80
49_56_N.tabacum	85	85	85	86	85	84	84	85	78	91	91	95		95	94	94	93	91	83	82	83	78	89	76	72	75	76	75	78
50_57_N.tabacum	86	86	85	87	85	85	84	86	79	92	92	95	95		95	95	93	92	84	83	84	79	88	77	75	78	78	77	79
51_85_S.tuberosum	87	87	87	89	88	87	86	87	81	91	93	95	94	95		100	98	96	86	84	86	78	89	79	78	80	79	79	80
52_86_S.tuberosum	87	87	87	89	88	87	86	87	81	91	93	95	94	95	100		98	96	86	84	86	78	89	79	78	80	79	79	80
53_88_S.tuberosum	85	85	86	87	86	85	85	85	79	90	91	94	93	93	98	98		96	84	83	84	77	87	78	76	78	78	78	79
54_87_S.tuberosum	84	84	84	86	84	84	83	85	78	89	90	93	91	92	96	96	96		84	82	83	76	87	77	75	78	78	77	79
55_C._vianneyi_c2	87	86	86	87	86	84	87	84	83	82	82	85	83	84	86	86	84	84		86	83	70	80	77	74	74	80	80	80
56_C._tetragona_c2	83	85	85	85	84	83	86	82	80	80	81	83	82	83	84	84	83	82	86		81	68	77	76	72	71	78	77	81
57_Craterispermum_schwenfurthii_c2	86	86	86	86	86	85	86	85	77	81	82	84	83	84	86	86	84	83	83	81		69	77	77	73	73	77	77	81
58_54_N.benthamiana	71	72	71	73	72	71	70	71	65	78	80	78	78	79	78	78	77	76	70	68	69		75	64	62	64	64	65	65
59_52_N.benthamiana	82	81	80	83	81	80	79	80	74	86	88	89	89	88	89	89	87	87	80	77	77	75		72	73	74	73	73	74
60_73_R.communis	77	78	79	79	79	79	78	76	72	74	75	78	76	77	79	79	78	77	77	76	77	64	72		73	73	69	69	72
61_21_E.guineensis	73	73	73	74	73	74	72	72	69	73	73	74	72	75	78	78	76	75	74	72	73	62	73	73		92	68	68	70
62_22_E.guineensis	73	74	73	74	74	74	72	73	69	76	76	77	75	78	80	80	78	78	74	71	73	64	74	73	92		68	68	72
63_Pyrostria_sp_c2	79	80	80	81	79	79	81	79	74	75	74	77	76	78	79	79	78	78	80	78	77	64	73	69	68	68		99	79
64_Pyrostria_sp_c3	79	80	80	81	79	79	81	78	73	75	75	77	75	77	79	79	78	77	80	77	77	65	73	69	68	68	99		79
65_Pyrostria_sp_c1	83	82	82	82	81	80	83	81	77	77	78	80	78	79	80	80	79	79	80	81	81	65	74	72	70	72	79	79	

Table S10 Distance values of the pair-wise comparison between the sequences similar to *Copia25* in the plant genomes analyzed. The shaded values correspond to the sequences homologous to *Copia25*, supported by the clade in the phylogeny and the distance values. All positions containing gaps and missing data were eliminated. There were a total of 602 nucleotide sites in the final dataset, and a total of 98 nucleotide sequences. Dark gray: species with less than 0.2 of distance; Light gray: species between 0.2 and 0.28 of distance.

Sequence	p-distance			Tamura 3-parameter model (+G 1.2)		
	Copia25 Sub 1[1]	*Copia25* Sub 2[2]	*Copia25*[3]	*Copia25* Sub 1[1]	*Copia25* Sub 2[2]	*Copia25*[3]
20_C.canephora	0.022	0.118	-	0.022	0.146	-
46_M.acuminata	0.116	0.120	0.110	0.143	0.148	0.134
85_S.tuberosum	0.116	0.154	0.120	0.140	0.203	0.145
86_S.tuberosum	0.116	0.154	0.120	0.140	0.203	0.145
50_M.balbisiana	0.121	0.131	0.121	0.150	0.165	0.151
45_M.acuminata	0.121	0.126	0.115	0.151	0.158	0.142
48_M.acuminata	0.123	0.125	0.116	0.152	0.154	0.143
49_M.acuminata	0.128	0.126	0.121	0.161	0.157	0.152
19_C.canephora	0.128	-	-	0.161	-	-
57_N.tabacum	0.140	0.189	0.141	0.176	0.271	0.179
88_S.tuberosum	0.143	0.178	0.145	0.183	0.248	0.186
47_M.acuminata	0.148	0.159	0.141	0.194	0.216	0.184
87_S.tuberosum	0.148	0.184	0.151	0.189	0.258	0.195
55_N.tabacum	0.148	0.188	0.146	0.187	0.263	0.185
56_N.tabacum	0.158	0.201	0.159	0.208	0.297	0.211
53_N.benthamiana	0.164	0.199	0.171	0.217	0.287	0.230
54_N.benthamiana	0.169	0.199	0.171	0.226	0.287	0.230
52_N.benthamiana	0.173	0.204	0.176	0.234	0.304	0.242
51_N.benthamiana	0.183	0.214	0.186	0.249	0.316	0.256
73_R.communis	0.228	0.233	0.223	0.342	0.349	0.331
22_E.guineensis	0.271	0.277	0.264	0.452	0.464	0.432
21_E.guineensis	0.271	0.277	0.267	0.459	0.472	0.449
1_A.trichopoda	0.281	0.264	0.286	0.480	0.429	0.494
33_H.vulgare	0.286	0.281	0.291	0.470	0.455	0.484
94_T.aestivum	0.286	0.282	0.287	0.474	0.464	0.478
24_E.grandis	0.287	0.279	0.277	0.481	0.452	0.453
11_C.cajan	0.287	0.294	0.291	0.475	0.498	0.486
12_C.cajan	0.287	0.294	0.291	0.475	0.498	0.486
13_C.cajan	0.287	0.294	0.291	0.475	0.498	0.486
23_E.grandis	0.289	0.281	0.279	0.489	0.459	0.460
96_T.aestivum	0.289	0.286	0.291	0.484	0.473	0.487
95_T.aestivum	0.289	0.287	0.294	0.481	0.476	0.495
98_V.vinifera	0.291	0.282	0.284	0.484	0.454	0.464
70_P.vulgaris	0.292	0.287	0.299	0.485	0.472	0.505
97_V.vinifera	0.296	0.307	0.297	0.511	0.558	0.515
60_O.sativa	0.297	0.294	0.301	0.507	0.490	0.516
68_O.sativa	0.297	0.294	0.301	0.508	0.491	0.517
64_O.sativa	0.297	0.294	0.301	0.509	0.492	0.518
32_G.raimondii	0.299	0.291	0.291	0.510	0.476	0.481
71_P.vulgaris	0.302	0.292	0.309	0.522	0.491	0.543
92_T.cacao	0.302	0.294	0.299	0.520	0.490	0.509
91_T.cacao	0.302	0.297	0.299	0.522	0.502	0.512
59_O.sativa	0.302	0.297	0.306	0.524	0.501	0.533
62_O.sativa	0.302	0.297	0.306	0.524	0.501	0.533
63_O.sativa	0.302	0.299	0.306	0.528	0.510	0.538
58_O.glaberrima	0.302	0.306	0.306	0.524	0.530	0.534
67_O.sativa	0.302	0.319	0.309	0.513	0.568	0.533
65_O.sativa	0.304	0.294	0.307	0.533	0.493	0.543
66_O.sativa	0.304	0.297	0.307	0.530	0.501	0.540
93_T.cacao	0.306	0.296	0.302	0.531	0.495	0.520
90_T.cacao	0.306	0.296	0.302	0.532	0.498	0.525
35_L.japonicus	0.306	0.299	0.299	0.533	0.512	0.510
31_G.raimondii	0.306	0.302	0.299	0.536	0.513	0.510
27_G.max	0.307	0.307	0.309	0.535	0.534	0.541
69_P.vulgaris	0.309	0.297	0.319	0.543	0.505	0.578
28_G.hirsutum	0.309	0.304	0.302	0.542	0.514	0.517
39_M.truncatula	0.312	0.306	0.309	0.555	0.540	0.544

41_M.truncatula	0.312	0.306	0.309	0.555	0.540	0.544
7_B.distachyon	0.312	0.317	0.319	0.551	0.563	0.573
79_S.lycopersicum	0.312	0.332	0.309	0.535	0.603	0.522
82_S.lycopersicum	0.312	0.332	0.309	0.535	0.603	0.522
83_S.lycopersicum	0.312	0.332	0.309	0.535	0.603	0.522
40_M.truncatula	0.314	0.299	0.309	0.562	0.517	0.544
80_S.lycopersicum	0.314	0.331	0.311	0.541	0.598	0.528
84_S.lycopersicum	0.314	0.331	0.311	0.541	0.598	0.528
81_S.lycopersicum	0.316	0.332	0.312	0.545	0.603	0.533
6_B.distachyon	0.316	0.322	0.322	0.562	0.584	0.585
61_O.sativa	0.316	0.329	0.322	0.552	0.598	0.574
30_G.hirsutum	0.317	0.299	0.309	0.577	0.505	0.543
77_S.italica	0.332	0.332	0.336	0.623	0.615	0.634
74_S.italica	0.332	0.332	0.337	0.617	0.613	0.635
3_A.thaliana	0.332	0.355	0.334	0.594	0.688	0.601
72_P.trichocarpa	0.332	0.336	0.332	0.596	0.608	0.596
76_S.italica	0.334	0.334	0.337	0.629	0.621	0.641
89_S.bicolor	0.336	0.334	0.339	0.633	0.622	0.644
37_M.domestica	0.341	0.347	0.337	0.636	0.660	0.624
75_S.italica	0.341	0.336	0.341	0.655	0.629	0.656
78_S.italica	0.344	0.341	0.347	0.668	0.649	0.685
17_C.sinensis	0.346	0.365	0.347	0.640	0.723	0.645
14_C.sinensis	0.346	0.341	0.346	0.661	0.633	0.658
16_C.sinensis	0.347	0.369	0.350	0.653	0.748	0.665
25_F.ananassa	0.352	0.337	0.349	0.651	0.599	0.638
15_C.sinensis	0.354	0.370	0.355	0.685	0.760	0.690
34_L.japonicus	0.357	0.347	0.352	0.679	0.647	0.656
29_G.hirsutum	0.357	0.347	0.355	0.685	0.647	0.679
8_B.rapa	0.359	0.352	0.359	0.670	0.651	0.671
26_G.max	0.360	0.369	0.354	0.697	0.731	0.669
9_B.rapa	0.362	0.357	0.362	0.680	0.666	0.680
44_M.guttatus	0.365	0.355	0.360	0.694	0.661	0.674
5_A.thaliana	0.365	0.350	0.362	0.722	0.659	0.707
4_A.thaliana	0.365	0.367	0.367	0.706	0.710	0.714
2_A.thaliana	0.369	0.367	0.370	0.719	0.708	0.727
38_M.truncatula	0.370	0.349	0.367	0.781	0.665	0.758
43_M.truncatula	0.370	0.349	0.367	0.781	0.665	0.758
36_L.japonicus	0.374	0.384	0.367	0.735	0.783	0.707
42_M.truncatula	0.380	0.390	0.382	0.766	0.826	0.780
10_B.rapa	0.390	0.382	0.385	0.797	0.764	0.777

[1]*Copia25* Sub 1 = 18_*C.canephora*, reference sequence of Subfamily 1; [2]*Copia25* Sub 2 = 19_*C.canephora*, reference sequence of Subfamily 2; [3]*Copia25* = 20_*C.canephora*, sequence trimmed.

Table S11 The number of base substitutions per site, Tamura 3-parameter model (below) and p-distance (above), between homologous sequences of *Copia25*. All positions containing gaps and missing data were eliminated. There were a total of 686 nucleotides in the final dataset, and a total of 98 nucleotide sequences.

	18	20	45	46	48	49	50	47	19	51	53	55	56	57	85	86	87	88	52	54	73	21	22
18_C.canephora		0.022	0.121	0.116	0.123	0.128	0.121	0.148	0.128	0.183	0.164	0.148	0.158	0.140	0.116	0.116	0.148	0.143	0.173	0.169	0.228	0.271	0.271
20_C.canephora	0.022		0.115	0.110	0.116	0.121	0.121	0.141	0.118	0.186	0.171	0.146	0.159	0.141	0.120	0.120	0.151	0.145	0.176	0.171	0.223	0.267	0.264
45_M.acuminata	0.136	0.128		0.027	0.035	0.042	0.053	0.080	0.126	0.201	0.191	0.164	0.176	0.159	0.140	0.140	0.164	0.161	0.193	0.191	0.219	0.282	0.281
46_M.acuminata	0.130	0.122	0.027		0.028	0.030	0.043	0.071	0.120	0.196	0.186	0.156	0.168	0.154	0.131	0.131	0.159	0.153	0.188	0.184	0.211	0.284	0.271
48_M.acuminata	0.138	0.130	0.036	0.029		0.030	0.050	0.085	0.125	0.203	0.186	0.166	0.174	0.161	0.138	0.138	0.166	0.163	0.188	0.184	0.214	0.291	0.282
49_M.acuminata	0.145	0.137	0.043	0.031	0.031		0.055	0.080	0.126	0.206	0.191	0.169	0.174	0.161	0.145	0.145	0.171	0.166	0.188	0.191	0.214	0.296	0.279
50_M.balbisiana	0.136	0.136	0.056	0.045	0.052	0.058		0.085	0.131	0.194	0.178	0.164	0.169	0.156	0.130	0.130	0.153	0.151	0.186	0.179	0.226	0.284	0.271
47_M.acuminata	0.171	0.163	0.086	0.077	0.092	0.086	0.092		0.159	0.214	0.206	0.188	0.189	0.166	0.163	0.163	0.193	0.188	0.206	0.199	0.234	0.291	0.284
19_C.canephora	0.145	0.132	0.142	0.134	0.139	0.142	0.148	0.187		0.214	0.199	0.188	0.201	0.189	0.154	0.154	0.184	0.178	0.204	0.199	0.233	0.277	0.277
51_N.benthamiana	0.215	0.220	0.243	0.235	0.245	0.250	0.232	0.262	0.263		0.103	0.105	0.101	0.095	0.106	0.103	0.125	0.125	0.125	0.108	0.287	0.307	0.294
53_N.benthamiana	0.190	0.200	0.230	0.223	0.222	0.229	0.209	0.252	0.241	0.112		0.096	0.095	0.091	0.091	0.091	0.118	0.113	0.106	0.080	0.267	0.306	0.292
55_N.tabacum	0.168	0.166	0.191	0.179	0.193	0.197	0.190	0.223	0.224	0.114	0.105		0.058	0.048	0.063	0.063	0.086	0.078	0.103	0.100	0.248	0.297	0.282
56_N.tabacum	0.183	0.185	0.209	0.197	0.206	0.206	0.198	0.228	0.247	0.111	0.104	0.061		0.048	0.068	0.068	0.100	0.083	0.098	0.090	0.249	0.299	0.284
57_N.tabacum	0.158	0.160	0.185	0.179	0.187	0.187	0.180	0.194	0.228	0.103	0.100	0.050	0.051		0.063	0.063	0.088	0.081	0.098	0.085	0.249	0.281	0.262
85_S.tuberosum	0.128	0.133	0.158	0.148	0.156	0.165	0.145	0.189	0.178	0.117	0.099	0.066	0.073	0.067		0.003	0.040	0.027	0.100	0.093	0.233	0.277	0.259
86_S.tuberosum	0.128	0.133	0.158	0.148	0.156	0.165	0.145	0.189	0.178	0.112	0.099	0.066	0.073	0.067	0.003		0.037	0.027	0.100	0.093	0.231	0.276	0.257
87_S.tuberosum	0.168	0.173	0.191	0.184	0.193	0.200	0.175	0.232	0.220	0.139	0.131	0.093	0.109	0.095	0.041	0.038		0.053	0.126	0.123	0.252	0.297	0.271
88_S.tuberosum	0.163	0.165	0.188	0.177	0.190	0.195	0.174	0.226	0.212	0.140	0.126	0.083	0.090	0.088	0.027	0.027	0.056		0.118	0.115	0.243	0.294	0.279
52_N.benthamiana	0.203	0.208	0.233	0.226	0.225	0.225	0.222	0.253	0.251	0.140	0.117	0.114	0.108	0.108	0.110	0.110	0.143	0.133		0.101	0.272	0.292	0.284
54_N.benthamiana	0.197	0.200	0.230	0.220	0.219	0.229	0.211	0.241	0.241	0.118	0.085	0.109	0.098	0.092	0.101	0.101	0.137	0.128	0.112		0.267	0.301	0.287
73_R.communis	0.282	0.274	0.269	0.256	0.260	0.261	0.280	0.293	0.288	0.382	0.347	0.312	0.317	0.317	0.289	0.286	0.321	0.305	0.360	0.347		0.301	0.301
21_E.guineensis	0.356	0.350	0.375	0.379	0.390	0.402	0.377	0.390	0.365	0.417	0.418	0.400	0.407	0.370	0.365	0.362	0.404	0.398	0.394	0.406	0.423		0.103
22_E.guineensis	0.353	0.341	0.369	0.351	0.371	0.365	0.349	0.374	0.362	0.388	0.388	0.368	0.374	0.334	0.329	0.326	0.349	0.365	0.376	0.377	0.419	0.115	

Table S12 Inter-specific nucleotide sequence identities (Id), *Ks*, Ka/Ks and divergence time from the ancestral sequence for *Copia25* polyprotein genes, and the seven best-conserved COSII genes between banana, *Solanaceae* and coffee. The best-aligned region of each COS gene were used for computation. Details of COSII sequences and their alignment are in Table S10. (n. a. = not applicable).

Gene/COS	C.can/N.ben				C.can/S.tub				C.can/M.acu				M.acu/N.ben				M.acu/S.tub				N.ben/S.tub			
	Id (%)	*Ks*	Ka/Ks	Time (Mya)	Id (%)	*Ks*	Ka/Ks	Time (Mya)	Id (%)	*Ks*	Ka/Ks	Time (Mya)	Id (%)	*Ks*	Ka/Ks	Time (Mya)	Id (%)	*Ks*	Ka/Ks	Time (Mya)	Id (%)	*Ks*	Ka/Ks	Time (Mya)
Copia25-Subfamily1	78.3	0.644	0.271	49.5	78.5	0.676	0.246	52.0	**84.5**	0.462	0.233	35.5	77.2	0.717	0.253	55.15	77.4	0.722	0.247	55.6	86.6	0.348	0.287	26.7
Copia25-Subfamily2	75.8	0.773	0.257	59.5	76.4	0.767	0.246	59.0	**85.5**	0.412	0.250	31.7	-	-	-	-	-	-	-	-	-	-	-	-
Semialdehyde dehydrogenase	77.8	1.593	0.058	122.5	78.9	1.270	0.072	97.7	73.3	2.363	0.057	181.8	72.8	n.a.	-	-	72.9	n.a.	-	-	91.1	0.41	0.045	31.5
Biotinsynthase	80.2	1.479	0.046	113.8	80.6	1.192	0.066	91.7	77.0	n.a.	-	-	76.3	n.a.	-	-	75.2	n.a.	-	-	93.5	0.177	0.215	13.6
Copperamineoxidase1-like	81.0	1.073	0.070	82.6	80.9	1.201	0.056	92.4	77.2	1.932	0.047	148.6	75.7	2.421	0.043	186.2	75.8	2.61	0.038	200.8	92.9	0.262	0.096	20.1
Deoxycytidylatedeaminase	75.0	0.978	0.179	75.2	74.0	1.075	0.168	82.7	69.8	1.626	0.132	125	69.5	1.538	0.145	118.3	71.0	1.7	0.111	130.8	87.7	0.417	0.15	32.1
Dynein light chain type 1	76.0	2.201	0.060	169.3	76.0	1.924	0.071	148	67.8	n.a.	-	-	69.0	n.a.	-	-	66.6	n.a.	-	-	92.3	0.413	0.036	31.7
Glucose6phosphateisomerase1	85.1	0.860	0.050	66.2	83.8	0.941	0.057	72.4	80.2	1.229	0.067	94.5	80.0	1.352	0.059	104	79.2	1.523	0.054	117.1	95.3	0.163	0.103	12.5
Alpha-mannosidase 3	78.6	0.987	0.124	75.9	78.9	0.980	0.124	75.4	77.4	1.328	0.09	102.1	73.6	2.917	0.045	224.4	72.9	2.465	0.061	189.6	92.2	0.195	0.26	15

Supplementary Material Figures

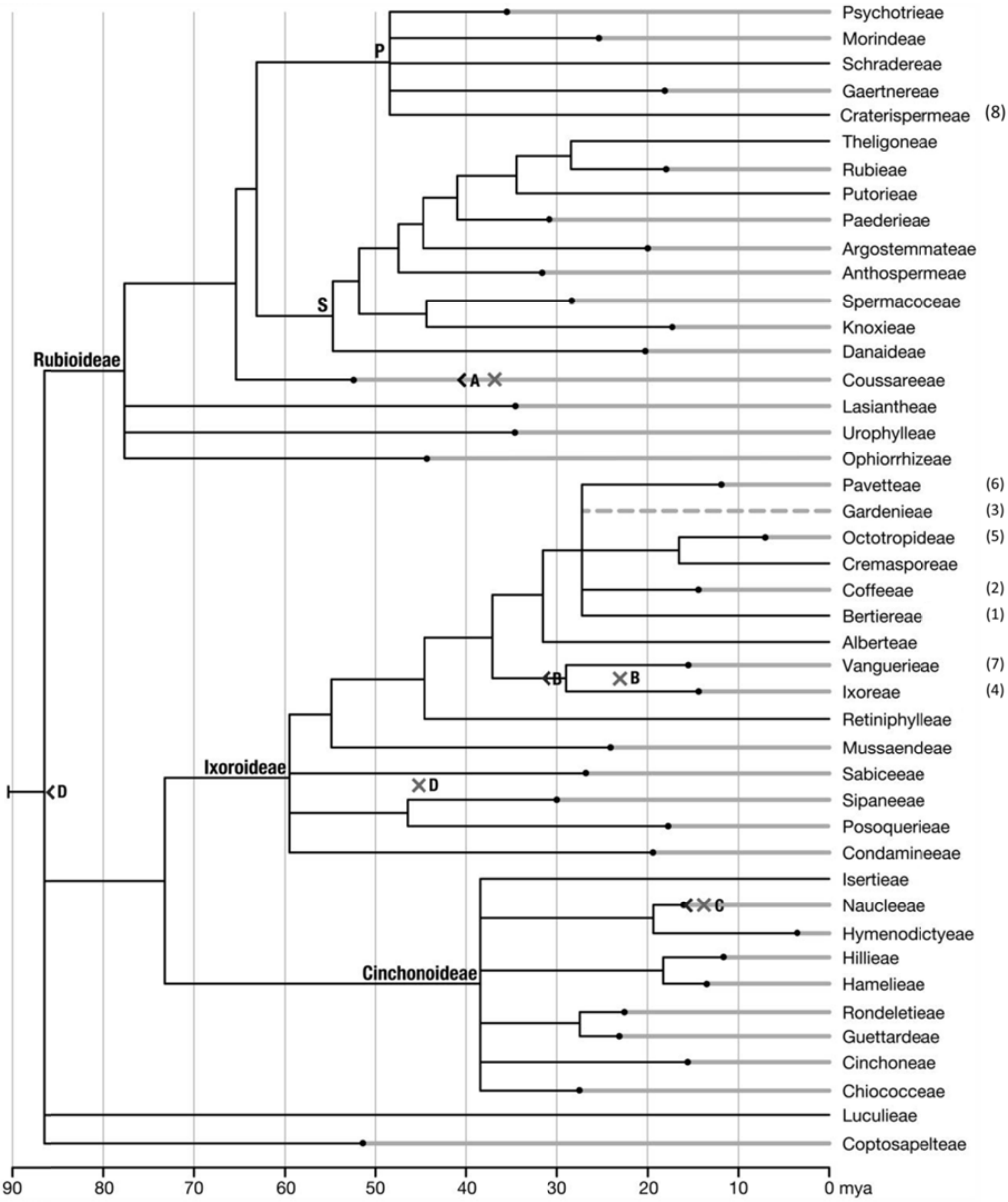


Fig. S1 Phylogeny of the Rubiaceae species: the numbers in parentheses refer to the following tribes which have species used in this work. (1) *B. iturensis*; (2) *C. arabica, C. canephora, C. eugenioides, C. humilis, C. stenophylla, C. millotii (ex-dolichophylla), C. perrieri, C. resinosa, C. tetragona, C. vianneyi, C. costatifructa, C. pseudozanguebariae, C. ebracteolatus* (ex *Psilanthus*) and *T. congesta;* (3) *O. formosus;* (4) *I. coccinea, I. finlaysoniana, I. foliicalyx* and *Ixora. sp;* (5) *P. parvifolia*; (6) *Coptosperma* sp; (7) *Pyrostria* sp; (8) *C. schwenfurthii*. Figure modified from Bremer and Erickson, 2009.

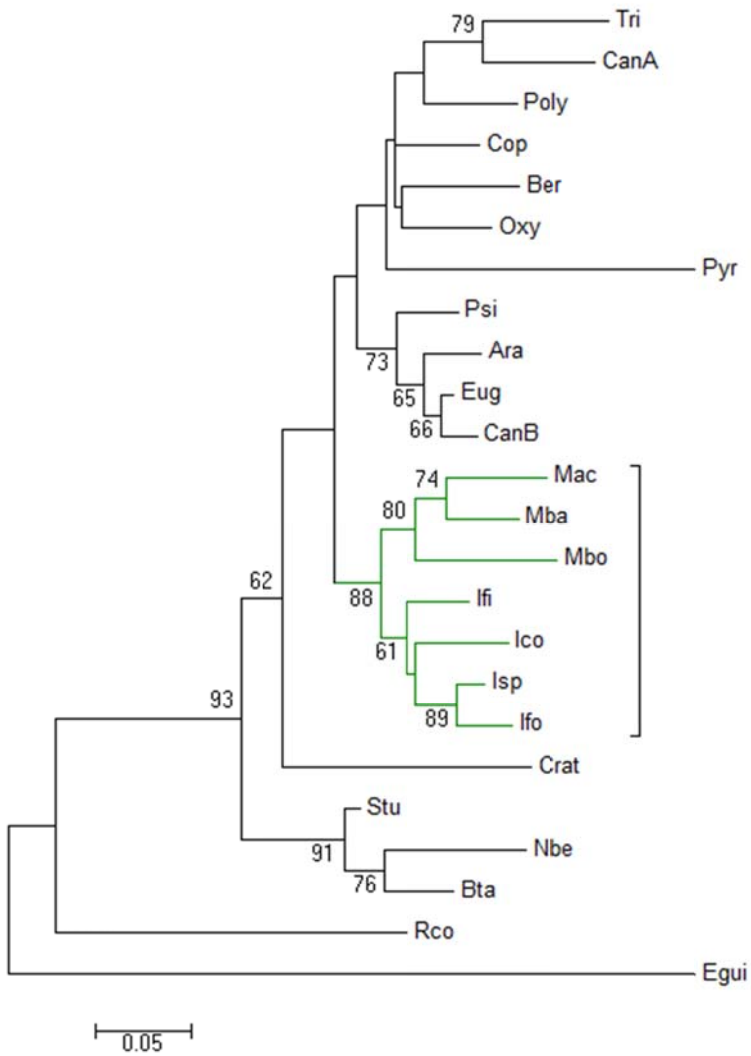


Fig. S2 Simplified phylogenetic tree reconstructed with RT *Copia25* used for the likelihood ratio tests (LRTs). The evolutionary history was inferred by using the Maximum Likelihood method based on the Tamura 3-parameter model [1]. The tree with the highest log likelihood (-2484.8487) is shown. A discrete Gamma distribution was used to model evolutionary rate differences among sites (2 categories (+G, parameter = 0.5721)). The tree is drawn to scale, with branch lengths measured by the number of substitutions per site. The analysis involved 24 nucleotide sequences: Ber = *Bertiera iturensis*, Clone 1; Ara = *C. arabica*, Clone 3; Eug = *C. eugenioides*, Clone 1; CanB = *C. canephora*, Subfamily 2; Psi = *Psilanthus ebracteolatus*, Clone 3; Poly = *Polysphaeria parvifolia*, Clone 1; Cop = *Coptosperma spp*; Oxy = *Oxyanthus formosus*, Clone 2; Tri = *Tricalysia congesta*, Clone 2; CanA = *C. canephora*, Subfamily 1; Ico = *Ixora coccinea*, Clone 1; Ifi = *Ixora finlaysoniana*, Clone 3; Isp = *Ixora sp*, Clone 1; Ifo = *Ixora foliicalyx*, Clone 1; Mac = *M. acuminata*, Clone 2; Mbo = *M. boman*, Clone 2; Mba = *M. balbisiana*, AC186755; Nbe = *N. benthamiana*, Niben044Scf00037679; Bta = *N. tabacum*, scaffold212962; Stu = *S. tuberosum*, chr02; Crat = *Craterispermum schwenfurthii*, Clone 2; Pyr = *Pyrostria sp*, Clone 1; Rco = *R. communis*, scaffold29815; Egui = *E. guineensis*, scaffold530537133. All positions containing gaps and missing data were eliminated. There were a total of 315 positions in the final dataset. Only the bootstrap values over 50% are shown. In green, the clade clustering *Ixora* and *Musa* sequences.

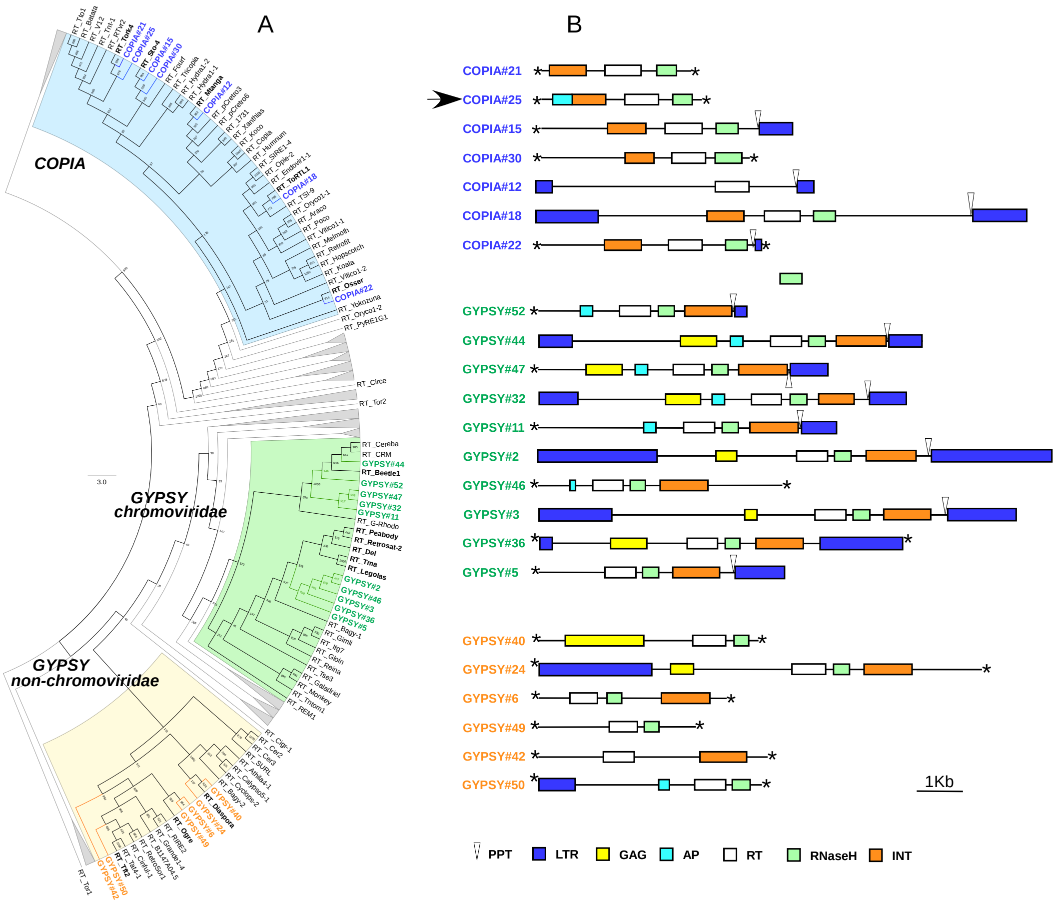


Fig. S3 Analysis of AAARF contigs showing similarities to RT_LTR proteins. A. Unrooted neighbor-joining tree of AAARF contigs based on their Reverse Transcriptase domains**.** RT domains were extracted from contigs and used for a phylogeny with all the RT domains stored at the GyDB (Loorens et al. 2011. In blue the plant *Ty1-Copia* superfamily, in green the plant *Ty3-Gypsy* superfamily, branch 1 (*Chromoviridae*) and in orange the plant *Ty3-Gypsy* family, branch 2. In grey are represented clusters of RT domains from other genus than plants. B. Schematic representation of the structure and domains of the 23 AAARF contigs an identified RT domain. The * indicates interrupted contigs, in blue: Long Terminal Repeat, in yellow: GAG domain, in light blue: Protease domain (AP), in white: Reverse Transcriptase domain, in green: RNase H domain and in orange Integrase domain (INT).

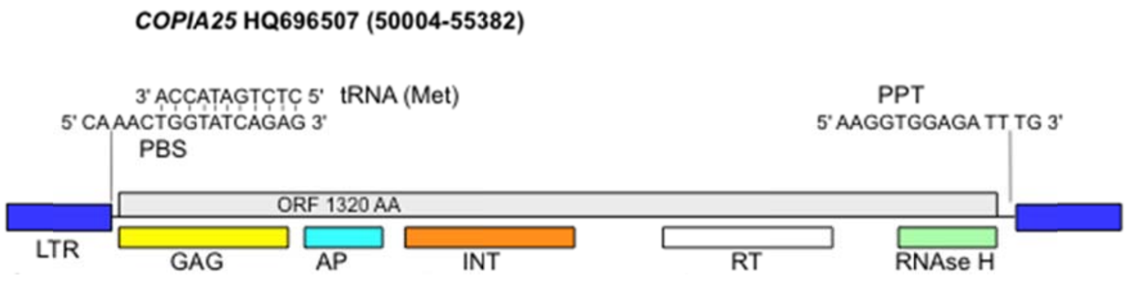
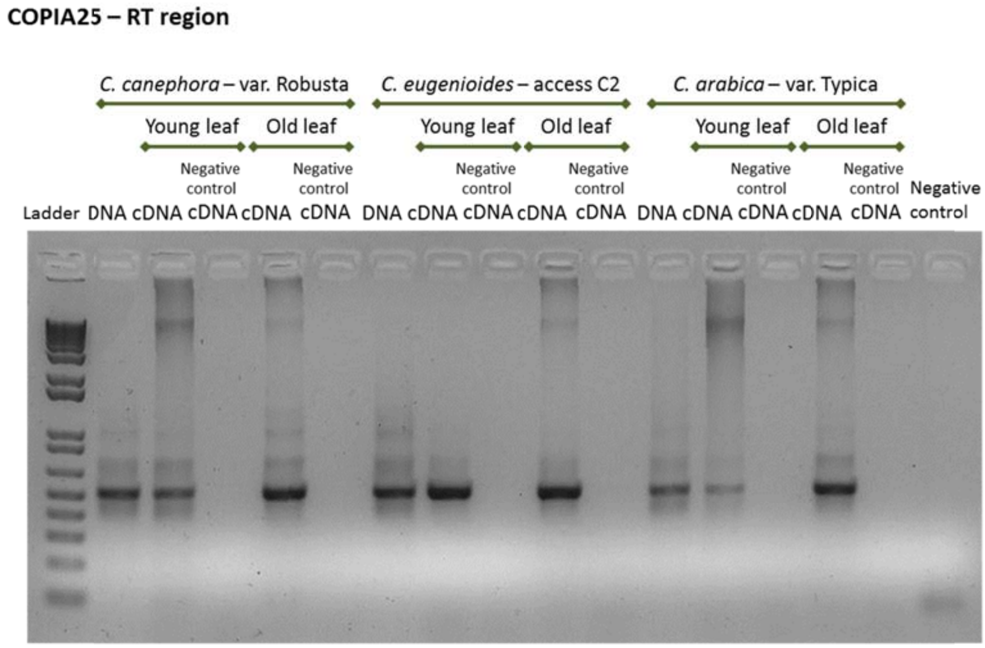


Fig. S4 The structure of the *Copia25* element found in the HQ696507 *C. canephora* BAC clone.



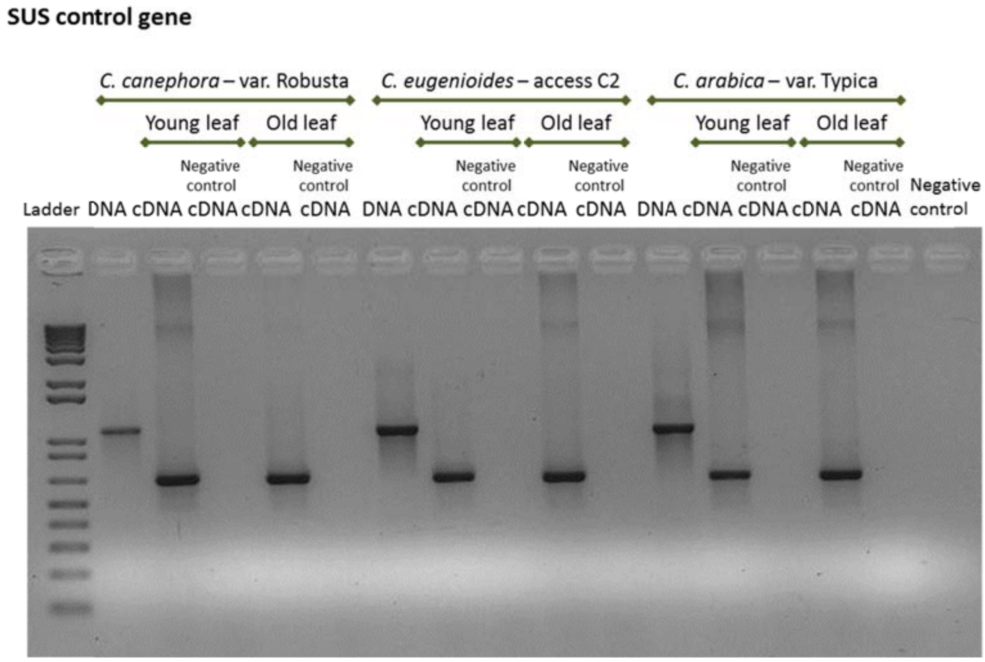


Fig. S5 Electrophoresis gel image of the RT-PCR of the RT *Copia25* region and the control gene *SUS*.

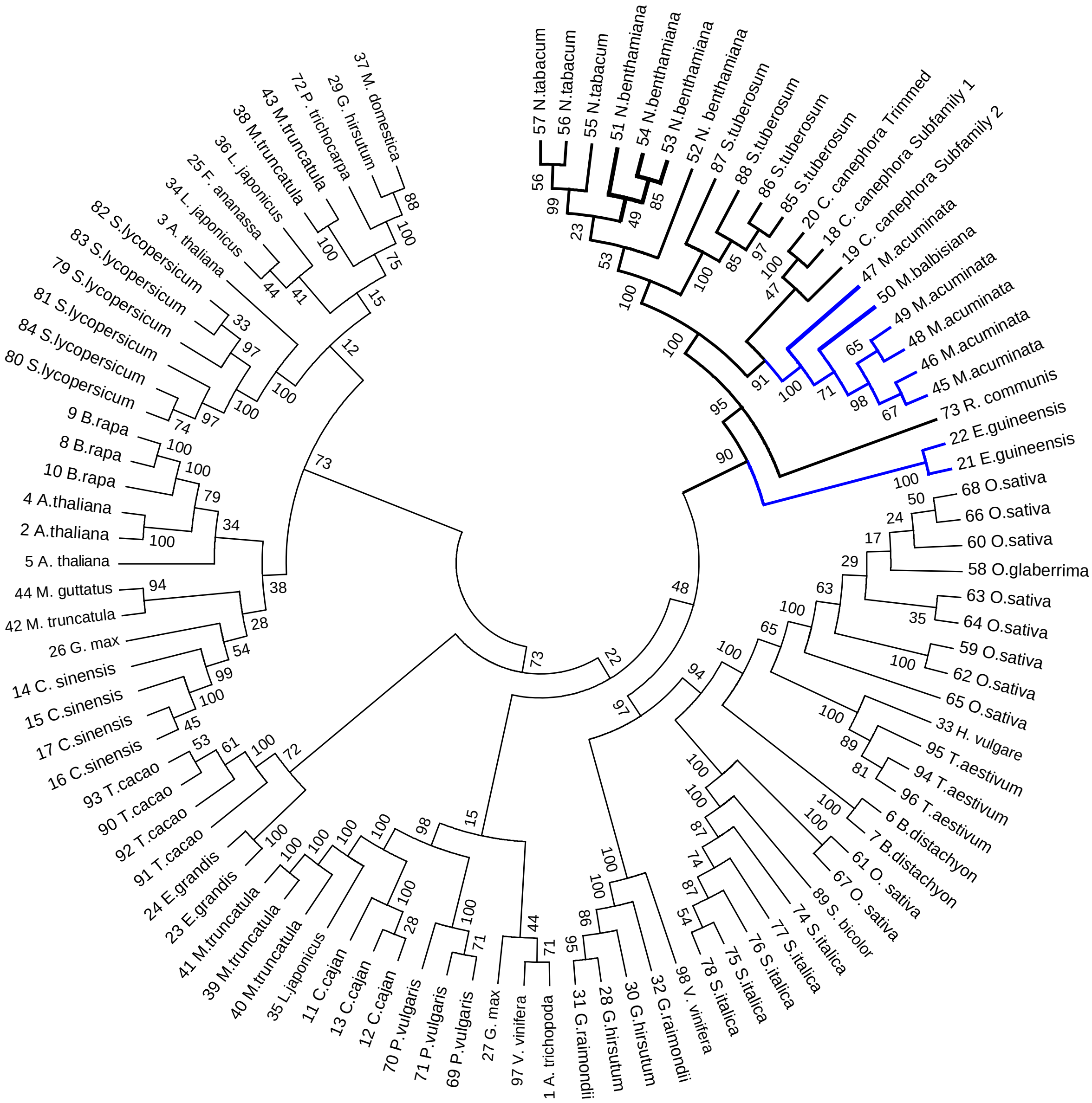
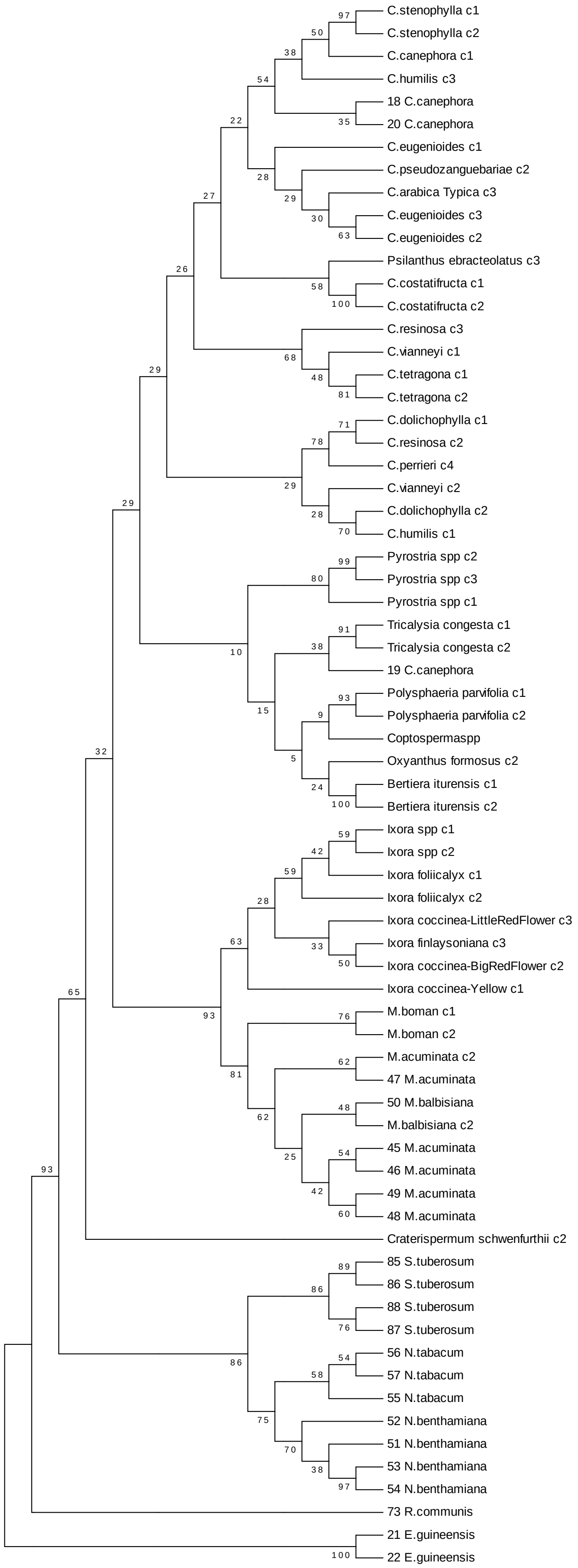


Fig. S6 Phylogeny of the RT domains from sequences similar to the *Copia25* elements in the 29 plant genomes analyzed – Not collapsed. The phylogeny was reconstructed using Maximum Likelihood, with the distance corrected by Tamura 3-parameter, and 1000 replicates; the bootstrap consensus tree inferred from the 1000 replicates has been taken to represent the evolutionary history of the taxa analyzed. All positions containing gaps and missing data were eliminated. There were a total of 602 nucleotide sites in the final dataset, and a total of 98 nucleotide sequences. A discrete Gamma distribution was used to model evolutionary rate differences among sites (2 categories (+G, parameter = 1.7864)). The rate variation model allowed for some sites to be evolutionarily invariable ([+I], 10.1863% of sites). The highlighted clade corresponds to the *Copia25* family; in blue, the monocot species in *Copia25* clade. Species abbreviations are given in Figure 3 and 4.


C.stenophylla c1
C.stenophylla c2
C.canephora c1
C.humilis c3
18 C.canephora
20 C.canephora
C.eugenioides c1
C.pseudozanguebariae c2
C.arabica Typica c3
C.eugenioides c3
C.eugenioides c2
Psilanthus ebracteolatus c3
C.costatifructa c1
C.costatifructa c2
C.resinosa c3
C.vianneyi c1
C.tetragona c1
C.tetragona c2
C.dolichophylla c1
C.resinosa c2
C.perrieri c4
C.vianneyi c2
C.dolichophylla c2
C.humilis c1
Pyrostria spp c2
Pyrostria spp c3
Pyrostria spp c1
Tricalysia congesta c1
Tricalysia congesta c2
19 C.canephora
Polysphaeria parvifolia c1
Polysphaeria parvifolia c2
Coptospermaspp
Oxyanthus formosus c2
Bertiera iturensis c1
Bertiera iturensis c2
Ixora spp c1
Ixora spp c2
Ixora foliicalyx c1
Ixora foliicalyx c2
Ixora coccinea-LittleRedFlower c3
Ixora finlaysoniana c3
Ixora coccinea-BigRedFlower c2
Ixora coccinea-Yellow c1
M.boman c1
M.boman c2
M.acuminata c2
47 M.acuminata
50 M.balbisiana
M.balbisiana c2
45 M.acuminata
46 M.acuminata
49 M.acuminata
48 M.acuminata
Craterispermum schwenfurthii c2
85 S.tuberosum
86 S.tuberosum
88 S.tuberosum
87 S.tuberosum
56 N.tabacum
57 N.tabacum
55 N.tabacum
52 N.benthamiana
51 N.benthamiana
53 N.benthamiana
54 N.benthamiana
73 R.communis
21 E.guineensis
22 E.guineensis

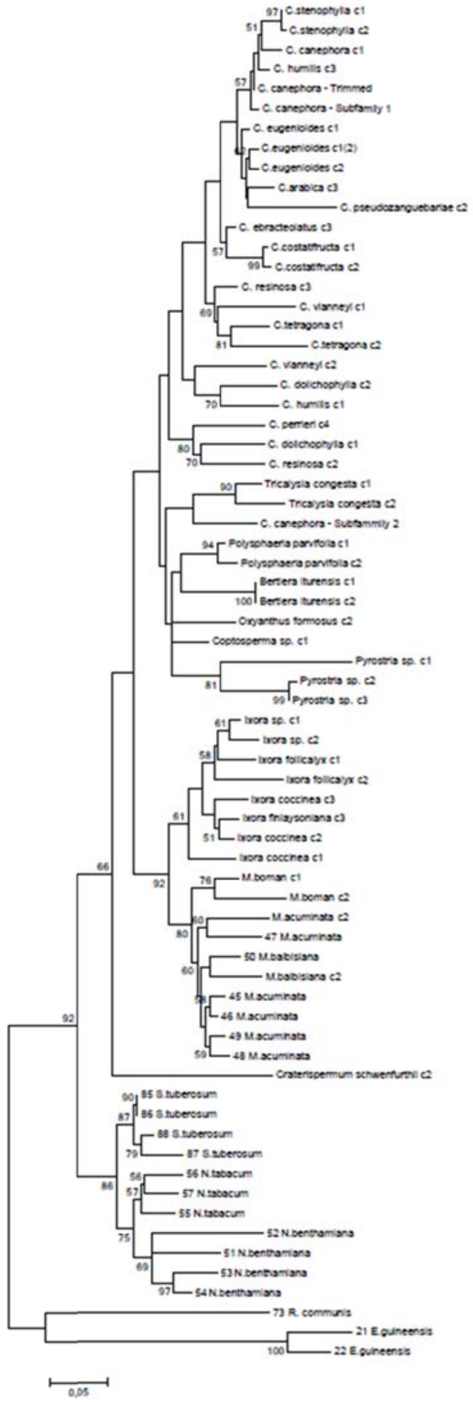


Fig. S7 Phylogenetic tree reconstructed with the RT *Copia25* homologs and the sequences amplified from the Rubiaceae species – Consensus Bootstrap tree. The phylogeny was reconstructed using Maximum Likelihood, with the distance corrected by Tamura 3- parameter, and 1000 replicates; the tree with the highest log likelihood (-4739.5265) is shown. A discrete Gamma distribution was used to model evolutionary rate differences among sites (2 categories (+G, parameter = 1.1187)). The tree is drawn to scale, with branch lengths measured by the number of substitutions per site. The analysis involved 69 nucleotide sequences. All positions containing gaps and missing data were eliminated. There were a total of 313 positions in the final dataset. Only the bootstrap values over 50% have been shown. In parentheses = the number of sequences collapsed in the tree; “c” indicates the clone sequences collapsed in the tree.

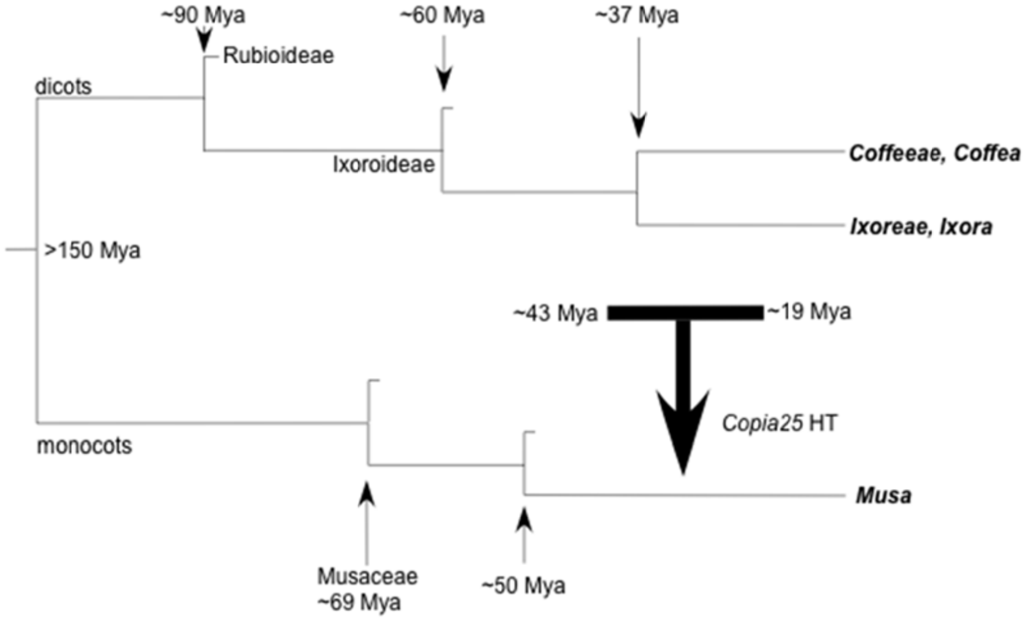


Fig. S8 Schematic representation of the phylogenetic relationships between *Coffea, Ixora* and *Musa*. The time scale of divergence is indicated as published in Bremer et al., 2009 and Christelová et al., 2011. The thick line indicates the putative time scale of *Copia25* horizontal transfer from *Ixora* to *Musa*.

4 CAPÍTULO 2

4 Capítulo II - Evolutionary dynamics of LTR-Retrotransposons in the allotetraploid *Coffea arabica*

Authors and Affiliations

Elaine Silva Dias[1,3] (esdias.bio@gmail.com)

Serge Hamon[3] (serge.hamon@ird.fr)

Perla Hamon[3] (perla.hamon@ird.fr)

Alan Carvalho Andrade[2] (alan.andrade@embrapa.br)

Pierre Marraccini[4] (pierre.marraccini@cirad.fr)

Romain Guyot[3] (romain.guyot@ird.fr)

Alexandre de Kochko[3] (alexandre.dekochko@ird.fr)

Claudia Marcia Aparecida Carareto[1*] (carareto@ibilce.unesp.br)

[1]UNESP – Univ. Estadual Paulista, Department of Biology, São José do Rio Preto, SP, Brazil.

[2]EMBRAPA Recursos Genéticos e Biotecnologia, Brasília, DF, Brazil.

[3]IRD UMR DIADE, EVODYN, BP 64501, 34394 Montpellier Cedex 5, France.

[4]UMR AGAP CIRAD, Montpellier, France

*Author for Correspondence: Claudia Marcia Aparecida Carareto
Email: carareto@ibilce.unesp.br
[1]Department of Biology, UNESP – Univ. Estadual Paulista, Department of Biology, São José do Rio Preto, SP, Brazil.

Abstract

Allopolyploidization can be followed by genomic reorganization and epigenetic changes associated to the repetitive portion of genomes. In plants, retrotransposons with LTR (LTR-RTs) comprise large amount of the genomes and have been involved with events of auto and allopolyploidization occurred throughout their evolutionary history. Retrotransposons sum about 42% of the *Coffea canephora* genome. *Coffea arabica* is an allotetraploid originated from a cross between *C. canephora*, the paternal species, and *C. eugenioides*, the maternal one, less than 1 Mya. Here, 10 LTR-RTs were annotated in the *C. canephora* genome and had their insertional profile obtained using IRAP and REMAP methods. The number of full-length copies in the genome and the transcriptional activity, evaluated in five genotypes, varied among the LTR-RTs, indicating that they are under different host control or stage of evolution. Losses of insertion sites were observed for all LTR-RTs, being the loss statistically significant for five of them. The reorganization in the allotetraploid seems to be correlated to the subgenomes, which the maternal subgenome more associated to losses or rearrangements than the paternal one. The LTR-RTs insertion sites in the genotypes of progenitor species form homogeneous populations suggesting that they share ancient copies remained from the ancestral. The *C. arabica* genotypes used, cultivated or derivate from them, have passed through several steps of artificial selection, hence population size shrink and bottleneck could be also related to the losses and reorganization observed.

Key-words: transposable elements, angiosperms, molecular marker, Rubiaceae.

4. 1 Introduction

Interspecific hybridization followed by the genome duplication - allopolyploidization - originates new species. The gene doubling in the polyploids confers genetic diversity and material to evolve, due to the increase of gene combination (Comai *et al.* 2005), which reflects in the high heterozygosity making the hybrids potentially adapted to different conditions those of their progenitors (Soltis 2013) and well succeeded in a long-term perspective (Mayrose *et al.* 2011). The evolutionary history of the flowering plants is remarkable for its rapid and extensive diversification, with the angiosperms comprising about 350,000 species, representing the majority of the land plants. Whole genome duplication have occurred in an ancient and a recent time at the evolutionary history of angiosperms and associated with their diversification (Jiao *et al.* 2011). Among 30 and 70% of angiosperm species have undergone polyploidization (Masterson, 1994), and about 30% of them are recent polyploids. Moreover, 15% of speciation events in angiosperms are directly associated with genome duplication (Wood *et al.* 2009).

Transposable elements (TEs) are ubiquitous in eukaryotic genomes and make up to 90% of plant genomes (Bennetzen and Kellogg 1997; Feschotte *et al.* 2002). They can promote punctual variations, at a locus context, and large changes, at the genome level. At a locus context, the insertion of TEs may interrupt genes, change distances between coding and regulatory sequences, “exonizes” and shuffles sequences (Gilbert 1978; Bureau *et al.* 1994), whereas their presence may supply *cis*-element, such as transcription factor binding sites, enhancers or insulators (Wang *et al.* 2006). At the genome level, TE induces gross and small genome reorganization in time, such as duplications, deletions, inversions and translocations, resulting in structural genomic changes (Lönnig and Saedler 2002). Drastic genetic and epigenetic changes are considered to take part in the rapid stabilization of hybrid genome after allopolyploidization events (Levy and Feldman 2002; Chang *et al.* 2010; Parisod and Senerchia 2012), and the TEs playing the decisive role in that reorganization (Parisod *et al.* 2009).

Transposable elements are classified into two major classes based on their mode of mobility, Class I elements, or retrotransposons, move through an RNA intermediate by a "copy and paste" mechanism, whereas Class II elements, or DNA transposons, move directly through a DNA molecule by a "cut and paste" mechanism. According to Wicker *et al.* (2007), the two classes are also subdivided; Class I is subdivided into five orders (LTR-retrotransposons, DIRS, Penelope, LINE and SINE) and Class II into two subclasses. These subclasses are themselves divided in two orders: orders TIR and Crypton for subclass I, and orders Helitron and Maverick for subclass II. Each order is also divided into one or several superfamilies, resulting in 29 superfamilies in total. In this study we focused on Class I TEs with LTRs, which includes the main superfamilies *Ty1/Copia* and *Ty3/Gypsy* (Wicker *et al.* 2007), which are differentiated on the basis of their internal coding region organization. Within each superfamily, families of TEs are determined based on the sequence conservation. Elements, which sequences sharing at least 80% of identity on at least 80% of their length are considered to belong to the same family (Wicker *et al.* 2007). Within each TE family, the age and the number of copies can drastically vary. The classification in subfamilies depends on TEs segregate in clades in phylogenetic analyses.

Coffea arabica (Rubiaceae) derived from a recent (less than 1 Mya) and natural hybridization between *C. canephora*, paternal progenitor, and *C. eugenioides*, maternal progenitor (Lashermes 1999; Tesfaye *et al.* 2007; Hamon *et al.* 2009; Yu *et al.* 2011). *Coffea arabica* is the unique tetraploid (2n = 4x = 44) species of the *Coffea* genus (Bouharmont 1959). Its recent origin and the strong bottleneck resulting from its cultivation resulted in a low genetic variability in the cultivated forms conducting to phenotype differences but very similar genetics varieties. Their phenotypic differences are mainly due to single point mutations. The self-fertile character of the species maintains this low diversity, which reduces the general variability, unlike the vast majority of the other *Coffea* species, which are auto-incompatible. But, the allopolyploidization that originated *C. arabica* could have contributed to genomic changes that can follow such events as genome react to the stress (McClintock 1984).

Hybridization events is one of several conditions that may trigger genome response, which could include activation of TEs that are carried in a silent state by the genome, and restructuration of the genome involving small or large segments of chromosomes, making with the allopolyploid may undergo rapid genetic and epigenetic changes – 'genome shock' hypothesis proposed by McClintock (1984). Both, extensive variations and discrete were already observed in allopolyploids, and this trend seems to be related to divergence of the parental species. Allopolyploidy between close related species could induce chromosomal and genomic rearrangements, while between distant related species few events of this nature were reported (Jackson and Chen 2010). The recent divergence of the progenitors of *C. arabica* – common ancestor between *C. canephora* and *C. eugenioides* lived about 4.2 Mya (Yu *et al.* 2011) –, and the high proportion of TEs in the parental *C. canephora,* amounting ~ 50% of its genome, 85% of which being LTR-RTs (Denoeud *et al.* 2014) suggest the potential role of these TEs to promote changes after the hybridization.

Accumulation and loss of the TE portion might occur in allotetraploids. TEs silenced in the progenitors can be reactivated in the hybrid, due to epigenetic changes, and results in bursts of transposition (Madlung *et al.* 2005). Allotetraploid also could accumulate TEs by relaxing the purifying selection against the TE insertions since coding regions are duplicated. And, the reduced effective population size, due to hybridization, leading to genetic drift can cause fixation of TEs (Charlesworth and Charlesworth 1983; Langley *et al.* 1983; Brookfield and Badge 1997; Le Rouzick and Capy 2005). By contrast, besides their natural propensity to chromosomal rearrangement due to their repetitive feature, the reorganization that occur after some allopolyploidization events led to losing of DNA sequences, usually the repetitive component (Vicient *et al.* 1999; Leitch and Bennett 2004; Bento *et al.* 2013) and result in decrease of the TE fraction. Although amass TEs is high likely and have being reported for some allopolyploids (Madlung *et al.* 2005; Piednoël *et al.* 2013), a similar or even greater number of studies have reported losses of sequences due to unequal or illegitimate recombination, and activation of few families of TEs (see Parisod and Senerchia 2012). Hence, both accumulation and losses of TE sequences are open scenarios in the allotetraploid *C. arabica* genome evolution.

Here, we investigated the above exposed possibilities of TEs action in changing the hybrid genome analyzing 10 LTR-RTs by IRAP and REMAP techniques in 21 genotypes of *C. arabica*, 18 of *C. canephora* and five of *C. eugenioides*, its parental species. Exclusive insertion sites observed in *C. arabica* suggested that these LTR-RTs could have mobilized recently, but there was no signal of bursts of transposition. The different patterns of the bands distribution of the progenitor species to the LTR-RTs composition suggest that *C. arabica* could have undergone genomic structural changes, involving loss of copies and overall rearrangements. These rearrangements could be directional, with bands of maternal progenitor being more often involved with the genome alteration.

4. 2 Material and Methods

Selection, classification and annotation of putatively active LTR-Retrotransposons in *Coffea canephora* genome

The draft genome sequence of the *C. canephora* accession DH 200-94 was used to annotate the TEs (Denoeud *et al.* 2014). Using the largest scaffolds from the genome, a manual annotation of TEs was performed and an initial database of 948 TEs was produced (Guyot R, unpublished data). Ten LTR-Retrotransposons (LTR-RTs) were selected based on their conserved structure: high conservation of their LTRs, presence of ORFs (using ORFinder) encoding essential protein domains for LTR-RTs to accomplish their transpositional cycle using BLASTp (Altschul *et al.* 1990) and, Artemis (Carver *et al.* 2005). The reverse transcriptase domain was used to classify the LTR-RTs by comparing with the reverse transcriptase domains available in the Gypsy Database2.0 (http://gydb.org/). The selected elements were characterized and their full-length copies were annotated in the *C. canephora* genome. Homologous sequences to each defined retrotransposon (using BLASTn - Altschul *et al.* 1990) were extracted from the genome and had their domains identified. Only copies with both LTRs and without uncertainties in their sequence were used. Afterward, copies of each element were classified into families based on the 80-80 Wicker’s criterion, (Wicker *et al.* 2007). For further analyses, the reverse transcriptase domain regions alone were used to classify each sequence at the family and subfamily levels according to the phylogenetic clustering.

Evolutionary sequence analyses

Phylogenetic analyses were performed with MEGA 6 (Tamura *et al.* 2013) using the alignment method by MAFFT (Katoh and Stanley 2013). The age of insertions of the LTR-RTs within *C. canephora* genome was estimated using the molecular clock equation, where *k* was the distance Kimura 2-parameter between both LTRs of the same copy, and *r* equals 1.3 x 10^{-8} base substitutions per site per year proposed for rice (Ma and Bennetzen 2004).

RNA Extraction and RT-PCR

RNA were extracted of leaves of five genotypes – *C. canephora* Robusta, *C. eugenioides* C1, and *C. arabica* Typica, Mundo Novo, and Iapar59 – using RNeasy Plant mini kit (QIAGEN). For the RT-PCR, 1 µg of total RNA was treated with RQ1 RNase-Free DNase (Promega) and reverse-transcribed using ImProm-II™ Reverse Transcription System (Promega) using oligo (dT) and random primers. The synthesized cDNAs were used as templates for the RT-PCR analyses. PCRs using oligonucleotide primers that anneal in the reverse transcriptase region of each retrotransposon selected (Table 1) were carried out as follows: 25 ng of cDNA, 5.0 µL of SYBR® Green master mix, and 600 nM of each primer, for a final volume of 10 µL; using an initial denaturation (95 °C, 10 min); followed by 40 cycles of denaturation (95 °C, 15 s) and annealing (60 °C, 1 min) (equipment StepOne Applied Biosystems). The constitutive gene of ubiquitin (BUBI) was used as a control and checking DNA contamination (Marracini *et al.* 2011). The results were analyzed using StepOne Software v2.3 (Applied Biosystems), and visualized by electrophoresis in a 1% agarose gel, 90V by 20 minutes.

Table 1 Nucleotide primers used in the RT-PCR.

Code	Sequence (5' - 3')	Amplicon (bp)
310_RT_F	GRTATTGAGGTAGYTCGGTCTA	116
310_RT_R	TCCATAGGAGTATCCACAGGTC	
645_RT_F	CCTTCTTCTTGTGCATATCCTTTAG	110
645_RT_R	CAACCTATCATTGGCACAAAGT	
763_RT_F	GAGACCTCATTTCCCATCCTAAC	92
763_RT_R	GGTGCATCCTACCATGTTACTC	
1070_RT_F	CCAAACCTTCTTGTTGCTTGTAT	127
1070_RT_R	CTTCCTCCTGGTTGTAGACCTTTAG	
1173_RT_F	ATGCYRCTATTGCAAGYAATATCC	99
1173_RT_R	CCAAAGGCTACAAGCAGAAAG	
1611_RT_F	TGGATACRGTYAGAGTACTGWTA	112
1611_RT_R	ATGTARACYTCCTCCTCMAGAT	
1054_RT_F	CCAYCCTGTYTTYCATGTRTC	210
1054_RT_R	TCCCAKGTRGCYTCTGCAGG	
1351_RT_F	YGAGTGGTTAGTTATRCCATTTGR	83
1351_RT_R	GRAATKGRCGAAGYACATGG	
1587_RT_F	GCTTGGTCTGCTTCTGTACTC	118
1587_RT_R	TCCTGGGATTGACAGGTTACTA	
1692_RT_F	CTGTTCGGGAGTTCCTTACTTC	193
1692_RT_R	TACCCGTCRAAGAGCCAGTCC	

DNA Extraction

DNA was extracted from 45 genotypes belonging to 3 *Coffea* species (Table 2) followed the protocol described in Deshmukh et al. (2007) with modifications (complete protocol available under request). Information on the origin of the genotypes is given in Supporting material (Table S1).

Table 2 Genotypes of the *Coffea* species used in this study.

Species	Genotype and/or botanical group	Source
C. canephora	DH200-94	IRD
	BA58 – Guianese	IRD
	BB56 – Congolese	IRD
	BC56 – Congolese	IRD
	BD64 – Congolese	IRD
	Apoatã IAC 3597 Col 1	IAC
	IAC 784 (C12)	IAC
	Guarini IAC 1598-11-3 Col 2	IAC
	Kouilou IAC 67-4	IAC
	Robusta IAC 1564 (C5)	IAC
	BuB2	IRD
	BuE1	IRD
	BuF4	IRD
	Bu10	IRD
	BuH1	IRD
	BuH3	IRD
	Zo03	IRD
	Zo05	IRD
C. eugenioides	IAC 1140-24 (C1)	IAC
	IAC 1098-7 (C2)	IAC
	DA	IRD
	DA56	IRD
	Iapar	Iapar
C. arabica	Typica IAC 537	IAC
	Acaiá IAC 1474-19	IAC
	Bourbon Amarelo IAC J19	IAC
	Bourbon Vermelho IAC	IAC
	Catuaí Amarelo IAC 62	IAC
	Catuaí Vermelho	IAPAR
	Caturra Vermelho IAC 477	IAC
	Ibairi IAC 4761	IAC
	Laurina IAC870	IAC
	Mundo Novo IAC 379-19	IAC
	IAC Ouro Verde H 5010-5	IAC
	Catiguá MG3	EPAMIG
	Obatã Vermelho IAC 1669-20	IAC
	Tupi IAC 1669-33	IAC
	Icatu Amarelo IAC 2944	IAC
	IcatuVermelho IAC 4041	IAC
	IPR 100	IAPAR
	IPR 102	IAPAR
	IPR 103	IAPAR
	IAPAR 59	EPAMIG
	Rubi	EPAMIG

Analyses of LTR-RT insertion sites polymorphism - IRAP and REMAP

IRAP - Inter-retrotransposon amplification polymorphism (Kalendar and Schulman 2006) - is a methodology that allows checking the distribution of given LTR-RTs closely inserted in the genome. Oligonucleotide primers designed to anneal in the LTRs region allow amplifying the in between copies region. On the other hand, REMAP - Retrotransposon-microsatellite amplified polymorphism (Kalendar and Schulman 2006) - allows the analysis of given LTR-RTs distribution proximal to microsatellites. In this technique, one of the primers is anchored in the LTR region, and the other in a SSR motif. For both IRAP and REMAP, two oligonucleotide primers were designed for each element. They were localized at the extremities of the LTRs, facing outward, and were designed based on the alignment of the copies annotated in the *C. canephora* genome (Table 2). For the IRAP analyses, both primers were used, so they were able to anneal and amplify regions between two close copies (up to 2 Kb) repeated in tandem or palindrome. For the REMAP analyses, only the reverse primer that anneals to the 5' end of the LTR, facing outward, was used combined with two microsatellite primers; ISSR1 $(AC)_8G$ and ISSR8 $(CT)_8G$ (Poncet *et al.* 2006). The Guanine nucleotide at the 3' end of the ISSR primer avoids the detection of variation in the number of repeats within the motif. Not all the combinations were used; ISSR1 was first associated to the reverse LTR primer of each element (10 RTs), the elements that did not show polymorphism were then tested with the ISSR8 (2 RTs) (Table 3).

Table 3 Primers used for the REMAP and IRAP analyses of insertion site polymorphism in the *Coffea* species.

Code	Primer sequence
ISSR1	ACACACACACACACACG
ISSR8	CTCTCTCTCTCTCTCTG
310_RMP_R	CACYRTATTATGAGTGACCAACT
310_RMP_F	TTTCCTYCCAARATACCWRTT
645_RMP_R	GRAWAGYTTGGTAGACAMTHGG
645_RMP_F	GACRTGAGTGYCCTGTTYTT
763_RMP_R	CTGTAYYGGTACCMMTTTYCCAAC
763_RMP_F	DYTGGCCCACAAATGGRCTR
1070_RMP_R	AAYCTCGAGCACTCYACCTGKA
1070_RMP_F	GKACCAYATTCATGTACATACAYG
1173_RMP_R	TYTAGGGTGATYCYTGRRTRG
1173_RMP_F	GTTTTAGGGCAWACATTCCAAC
1611_RMP_R	TCCTCTCACCTATCAACCACTAACAG
1611_RMP_F	GGCATCGGTACCATCATTAGTCTG
1054_RMP_R	GCGCCTTCCAGTGTATCTTCTT
1054_RMP_F	TCTTCCATTCCCAGCTGGATCA
1351_RMP_R	CYTTGGAGTTCCTYCCTCATCTT
1351_RMP_F	TTGGTRTCCTCTTGTCGATCCT
1587_RMP_R	CCTGAGCAAGGRAGTWKTAGGW
1587_RMP_F	GGATGCCAGCTTGGCACAACCG
1692_RMP_R	GRRYCAGCTCTYCCATCARCAC
1692_RMP_F	TTCAGGAGCCCTAGCAACGA

For both methods, the PCR reactions were performed as follows: 1.25 unit of Taq polymerase (Platinum Invitrogen, Brazil), 10 ng genomic DNA, 1 mM of $MgCl_2$, 1 X buffer, 5X TBT-PAR buffer (Samarakoon *et al.* 2013), 2.5 mM of dNTPs and 0.4 mM of each primer (the LTR-R labeled with 6-FAM, and the ISSR, in the REMAP, or the LTR-F, in the IRAP) were used in a final volume of 12.5 µL with ultrapure water QS. Amplification conditions were as follows: initial denaturation (94 °C, 120 s); followed by 45 cycles of denaturation (94 °C, 40 s), annealing (55 °C, 40 s) and extension (72 °C, 120 s), and a final extension of 420 s. The PCR products were diluted (1:20), mixed with size standard (GeneScan™-1200LIZ®, Applied Biosystems), and resuspended in HI-DI Formamide (Life Technologies®) before electrokinetic injection on capillary electrophoresis systems (ABI3730 DNA Analyzer, service performed in CEHG-CEL, São Paulo University, São Paulo, Brazil). Each fluorescent peak obtained from capillary electrophoresis was treated as a unit character (allele) for its respective locus and the amplicon size determined by GeneMarker V2.6.3 (Softgenetics®). The profiles were independently scored manually, and classified by the number one (presence) or zero (absence) among different genotypes. In order to avoid mis-typing of the bands, the band recording was done, at least, twice, and the faintly ambiguous bands were excluded.

Data analyses

The raw data, matrix 0/1, were used to estimate the frequency based analysis, and the distance-based analysis. In frequency analysis, the total number of insertion sites per species, percentage of polymorphic sites, number of private insertion sites, Shannon's information index [$I = -1^* (p * Ln (p) + q * Ln(q))$], and genetic diversity index [$h = 1 - (p^2 + q^2)$] were estimated using GenAlEx v.6.501 (Peakall and Smouse 2012), where for haploid binary data, p is the band frequency and q = 1 - p. Nei's distances and identities were also calculated.

The genetic structure of the populations also was analyzed by Analysis of Molecular Variance (AMOVA). Both, Nei's estimates and AMOVA were done using GenAlEx v.6.501. A dissimilarity matrix was built using the Bray–Curtis index in the vegan package in R software ver. 3.1.1 (R Core Team, 2014). The Bray–Curtis index for binary data corresponds to the Sorensen-Dice index, this index take into account the presence for the estimation. The difference among and between the dissimilarity index of the populations were tested using Kruskal-Wallis non-parametric test. The same vegan package was used to perform the principal coordinate analysis (PCoA) using the dissimilarity matrices transformed (square root), assessing the general arrangement of genetic variation. Besides being used in the frequency and the distance analyses, the raw data were used as input for the network analysis using the median joining (MJ) algorithm in the Phylogenetic Network Software (Bandelt *et al.* 1999).

4. 3 Results

Classification and Annotation of the LTR-RTs selected in the C. canephora *genome, and their recent activity*

Out of the ten selected LTR-RTs, six belong to the *Ty1/Copia* superfamily and the other 4 to the *Ty3/Gypsy* superfamily (Table 4 and Figure S1). The selected elements share an overall sequence identity with retrotransposons described in other plants deposited in the RepBase database. The identity shared ranging from 29% to 64%, being the lesser value between the *CcGypsy1692* and the Ogre-Sd1 described in *Solanum* sp. and the highest between the *CcCopia1611* and the *Copia-55* identified in *Vitis vinifera*. It is worth noting that the first species belongs to the Asterid, the same clade to what belongs the *Coffea* genus, and the last to the Rosid clade. Other elements showed a close relationship with retrotransposons described in *V. vinifera*, a species quit distantly related. Although the identity shared between the LTR-RTs does not support the hypothesis of them belong to the same family, it may be residue of a close and complex evolutionary history that have occurred in these groups of species, *Coffea* and *Vitis*, which have more syntenic regions than *Solanum* and *Coffea*, taxa closely related (Guyot *et al.* 2012).

In the *C. canephora* genome, the number of full-length copies of each LTR-RT family annotated varied from 2 to 80 (Table S2), totaling 258 full-length copies (206 of *Ty1/Copia* and 52 of *Ty3/Gypsy*). The mean nucleotide identity of the reverse transcriptase domain between the copies of each family was 91% (± 1.7%), ranging from 83% to 99%, supporting their individual classification as belonging to a given family, according the 80's rule (Wicker *et al.* 2007). Besides that, copies of the elements *CcCopia645, CcCopia1070*, *CcGypsy1054*, *CcGypsy1351* and *CcGypsy1692* branches in, at least, two clades in their individual phylogenies using a sequence of the *reverse transcriptase* domain (Figure S2). This segregation suggests the presence of subfamilies of these RTs in the *C. canephora* genome.

To evaluate the recent activity of these RT families in the coffee genome, the age of insertion of each copy in its genomic site was estimated using the divergence between both LTRs of the each copy, and the rate of substitution per site per year of 1.3×10^{-8} proposed to rice (Ma and Bennetzen 2004) in the molecular clock equation. Figure 1 shows the ages estimated for all LTR-RTs analyzed. The mean age of the insertion was 2.9 My (± 4.7, median = 1.65), being the oldest age observed for *CcGypsy1351* and *CcGypsy1692*, more than 25 My ago. Furthermore, all RTs presented copies inserted very recently in the *C. canephora* genome less than 1 My ago (red line in the Figure 1), probably after have produced the *C. arabica* hybrid. Although these time of divergence were estimated using a substitution rate estimated for rice, which can be at some extent different of that in *Coffea*, these estimates and the occurrence of full-length sequences suggest that these elements were active or have been active in a very recent time in this species. *Coffea canephora* shares a common ancestor about 4.2 Mya with *C. eugenioides* and, as mentioned above, from an intercross they originated *C. arabica* about 1 Mya in Northeast Africa (Yu *et al.* 2011; Lashermes *et al.* 1999).

The transcriptional activity of the LTR-RTs was evaluated by RT-PCR in five genotypes, one of each parental species and three of *C. arabica* (Table S3). The results indicated abundant presence of transcripts for *CcCopia310*, *CcCopia645* and *CcCopia763*, whereas *CcCopia1070*, *CcCopia1173*, *CcCopia1611*, and *CcGypsy1587* presented moderate amount of transcripts. The other LTR-RTs presented, in general, minimal amount of transcripts or no conclusive results. Particularly, *CcCopia1611* showed absence or minimal amount of transcripts in the progenitors, but abundant expression in the allotetraploid samples. Although these results should be treated cautiously, due to the absence of biological replicates, they suggest that most of the LTR-RTs are transcriptionally active, and could thus generate new copies in the genomes. Hence the genome and expression analyses showed that these LTR-RTs are, or have been, transpositionally active in the parental species and in the hybrid, they were used to investigate their occurrence, heritage and, at last, their impact in the hybrid species comparing their polymorphism for insertion sites with the progenitors.

Table 4 Characteristics of LTR-RTs concerning their classification and full-length copy number in the *C. canephora* genome.

LTR-RT	Classification (Superfamily, clade)	Reference copy genome localization	Size (bp)	Full-length copy genome	Mean Identity (%)*	Related retrotransposon RepBase	Identity (%)
CcCopia310	*Ty1/Copia*, Retrofit	chr11: 183768 - 187213	3,446	80	90.4 ± 0.0004	Copia23-VV_ICopiaVitis	48
CcCopia645	*Ty1/Copia*, SIRE	chr2: 13619256 - 13630213	10,958	62	88.7 ± 0.002	SZ-55_ICopiaOryza	41
CcCopia763	*Ty1/Copia*, Tork	chr0: 32330598 - 32336022	5,425	48	85.3 ± 0.003	Copia-4_PD-ICopiaPhoenix	41
CcCopia1070	*Ty1/Copia*, Tork	chr10: 20173933 - 20179816	5,884	10	92.8 ± 0.005	Copia15-VV_ICopiaVitis	42
CcCopia1173	*Ty1/Copia*, Tork	chr3: 24291196 - 24300739	9,544	4	99.9 ± 0.015	Copia-20_TC-ICopiaTheobroma	42
CcCopia1611	*Ty1/Copia*, Retrofit	chr6: 19221962 - 19226978	5,017	2	83.4	Copia-55_VV-ICopiaVitis	64
CcGypsy1054	*Ty3/Gypsy*, Reina	chr8: 15934846 - 15940597	5,752	6	81.6 ± 0.007	Gypsy-113_SB-IGypsySorghum	47
CcGypsy1351	*Ty3/Gypsy*, CRM	chr0: 88897664 - 88905609	7,946	18	96.1 ± 0.003	Gypsy-29_PTr-IGypsyPopulus	53
CcGypsy1587	*Ty3/Gypsy*, Reina	chr2: 34246937 - 34252179	5,243	4	90.5 ± 0.012	Gypsy-109_ZM-IGypsyZea	42
CcGypsy1692	*Ty3/Gypsy*, TAT	chr1: 29459929 - 29469576	9,648	24	86.8 ± 0.003	Ogre-SD1_IGypsySolanum	29

*Reverse transcriptase domain region, according to reference copy.

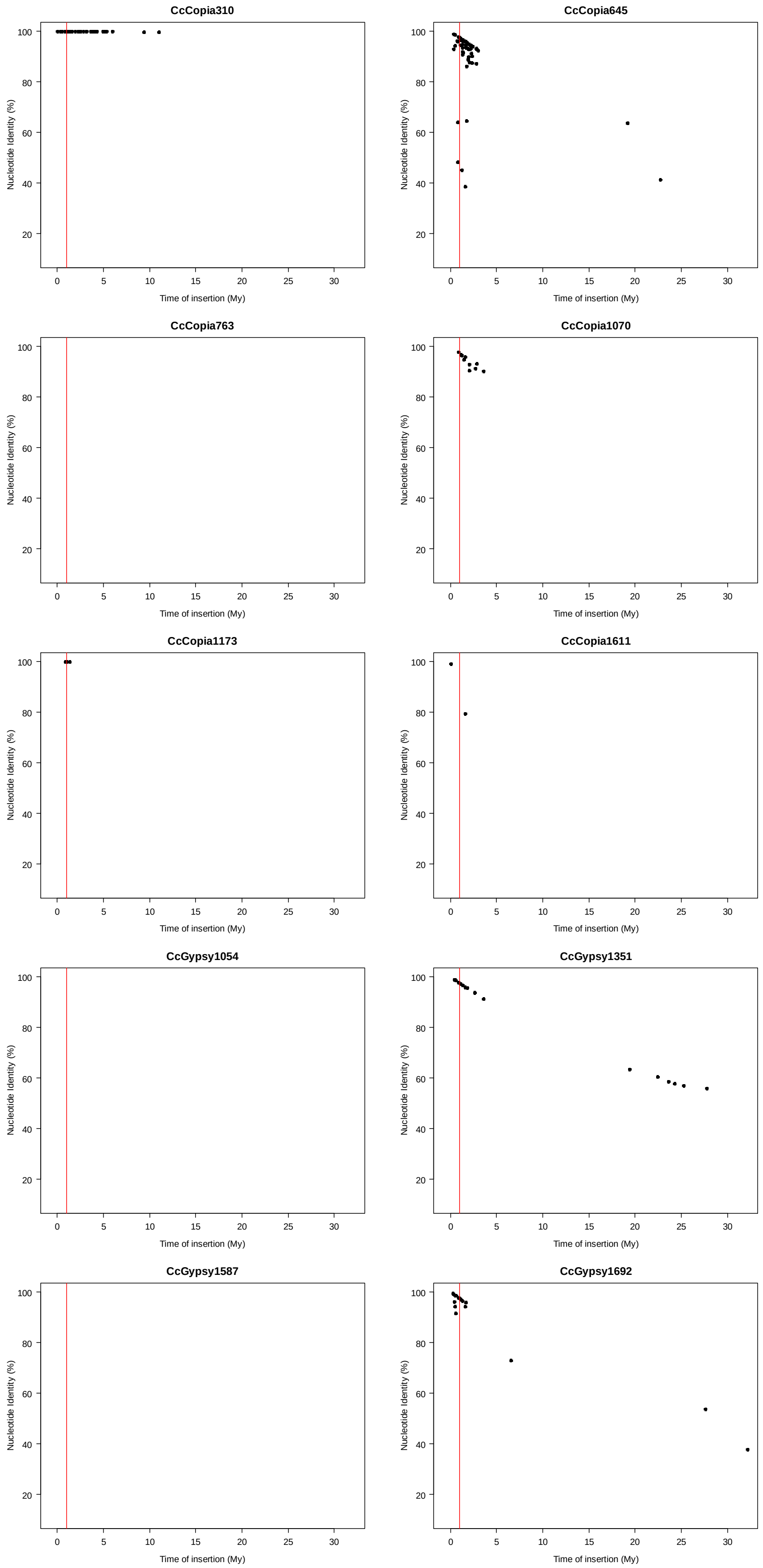


Figure 1 Distribution of the ages of insertion (in Millions of years) of the full-length copies of the ten LTR-RTs analyzed in *C. canephora* genome. Red line indicates 1 My.

TE Insertion Site Polymorphism Analyses

Twenty-one *C. arabica* genotypes, 18 *C. canephora*, and 5 *C. eugenioides* were analyzed with 10 IRAP and 10 REMAP primer combinations (Table 3, and Tables S4 and S5). IRAP produced a total of 374 bands, with a mean of 46 (± 6.55) bands, whereas REMAP produced 470 insertion sites with a mean of 47 (± 10.24) bands by primer combination. The elements *CcCopia763* and *CcCopia1611* produced no bands in the IRAP reactions. The highest number of polymorphic bands (frequency over 5% in the species) was found for the *CcGypsy1351* (80) using IRAP, and for *CcCopia763*/SSR1 (97) using REMAP. The lowest number of bands was observed for *CcGypsy1054* (24) and *CcCopia645*/SSR8 (4) in both types of analyses, respectively. The retrotransposon *CcCopia645* with SSR8 produced no bands in *C. eugenioides* and *C. arabica*, which means that their copies could be not associated with the SSRs motifs used. All LTR-RTs presented private bands in the three species. *CcGypsy1692* showed the highest number of private bands (61), followed by *CcCopia310* (53) and *CcGypsy1351* (46). Lowest copy numbers where found for *CcCopia1611*, followed by *CcCopia1070* and *CcCopia645*. Except for *CcCopia763* and *CcCopia1611*, *C. canephora* presented the highest number of private bands (199). *CcCopia763* presents the highest number of bands in *C. arabica*, and *CcCopia1611*, with a general low number of bands, presents only one exclusive band in *C. canephora* and in *C. arabica*.

In total, 90% of the bands produced by IRAP and 92% by REMAP were found to be over 5% in the analyzed species. *Coffea canephora*, again, stands out for presenting the highest percentages of polymorphic loci for almost all the LTR-RTs analyzed, with the exception of *CcCopia763* and *CcCopia1611*, for which *C. arabica* showed the highest values (Figure 2). This species had the second largest percentages of polymorphic loci, except for *CcCopia310* and *CcGypsy1351*, which are more polymorphic in *C. eugenioides*. *C. arabica* being polymorphic for higher number of LTR-RTs than *C. eugenioides* not only point out the considered variability observed in the first species for these RTs, but also evidences the low diversity in *C. eugenioides*, one of its progenitor, that has about 4.2 My (Yu *et al.* 2011).

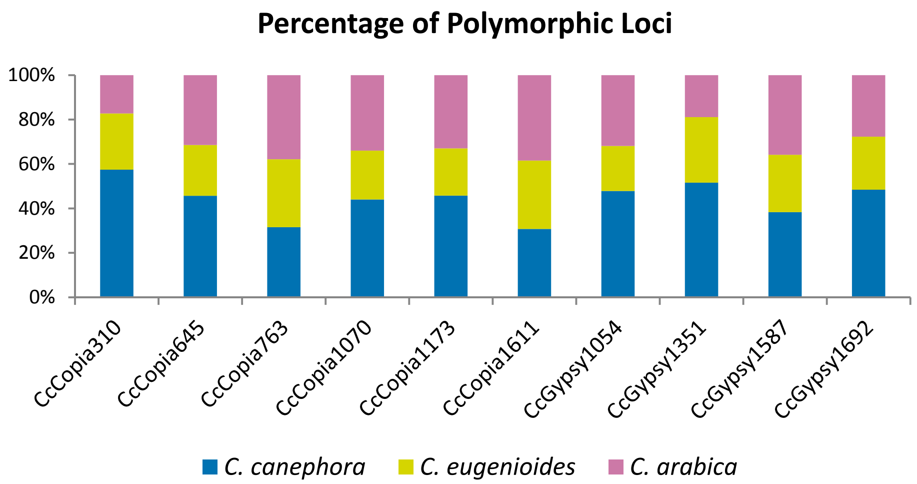


Figure 2 Distribution of the percentages of polymorphic bands produced by 10 LTR-RTs detected using the IRAP and REMAP methods in three *Coffea* species.

Bands distribution in the diploid and the allotetraploid species

The number of REMAP and IRAP bands private or shared between the three species was variable for each LTR-RTs (Figure 3 and Table 5). In total, 36% of the bands are shared by two or more species; a similar percentage (34%) of bands private to *C. canephora*, whereas *C. arabica* and *C. eugenioides* showed about half of this percentage, 17% and 13%, respectively. For six LTR-RTs, *C. arabica* shared with *C. canephora* the highest values among the percentages of the pair of species comparisons. This situation is particularly observed for the *Gypsy* RTs, as the *CcGypsy1054* for which 21.7% of the total bands are shared between this two species. Three elements

(*CcCopia645*, *CcCopia763*, and *CcCopia1611*) showed a higher number of shared bands by *C. arabica* and *C. eugenioides*.

Although the amount of bands has varied when the progenitor species were compared, the difference is not significant. Only *CcGypsy1692* presented an amount of bands significantly higher in *C. canephora* than in *C. eugenioides* (X^2 = 14.2045, p = 1.64e-04). *Coffea arabica* presented in average 50% fewer bands than its progenitors. Individually, this difference was significant for five of the ten LTR-RTs, *CcCopia310*, *CcCopia1173*, *CcGypsy1054*, *CcGypsy1351*, and *CcGypsy1692*. The difference is also observed once the expected amount is compared to the observed of each superfamily (*Copia*: X^2 = 41.823, *p-value* = 9.991e-11, *Gypsy*: X^2 = 59.2081, *p-value* = 1.418e-14).

From the total of bands detected in *C. arabica* (97), 11.3% (11) are present at a frequency higher than 0.8 in the genotypes analyzed. Six for *CcCopia310* bands, 2 for *CcGypsy1351*, and 1 for *CcCopia645*, *CcCopia1070* and *CcCopia1173*. This data suggest that very few new retrotransposition and/or insertions were relocated, result of ectopic recombination, had occurred in the ancestor lineage of the genotypes evaluated.

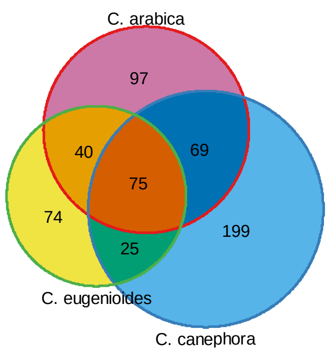


Figure 3 Venndiagram showing the number of bands shared by the three *Coffea* species for 10 LTR-RTs detected by IRAP and REMAP methods.

Table 5 Distribution of IRAP and REMAP bands from ten LTR-RTs within and among the progenitor *C. canephora* and *C. eugenioides* and the allotetraploid *C. arabica*.

LTR-RT	Total	*C. canephora**	*C. eugenioides**	*C. arabica**	C + E*	C + A*	A + E*	C + E + A*	C + E *vs.* *C. arabica***	Expected***
CcCopia310	76	38 (50.0)	13 (17.1)	11 (14.5)	4 (5.3)	4 (5.3)	2 (2.6)	4 (5.3)	CE > A[1] ****	< 67%
CcCopia645	32	13 (40.6)	4 (12.5)	9 (28.1)	0 (0.0)	2 (6.3)	3 (9.4)	1 (3.1)	NS	< 31%
CcCopia763	63	13 (20.6)	10 (15.9)	17 (27.0)	2 (3.2)	4 (6.3)	6 (9.5)	11 (17.5)	NS	< 19%
CcCopia1070	37	13 (35.1)	3 (8.1)	5 (13.5)	0 (0.0)	8 (21.6)	7 (18.9)	1 (2.7)	NS	< 52%
CcCopia1173	62	20 (32.3)	7 (11.3)	10 (16.1)	3 (4.8)	10 (16.1)	2 (3.2)	10 (16.1)	CE > A[2] ****	< 50%
CcCopia1611	8	1 (12.5)	0	1 (12.5)	1 (12.5)	2 (25.0)	3 (37.5)	0	-	-
CcGypsy1054	60	20 (33.3)	6 (10.0)	7 (11.7)	3 (5.0)	13 (21.7)	2 (3.3)	9 (15.0)	CE > A[3] ****	< 61%
CcGypsy1351	72	26 (36.1)	13 (18.1)	7 (9.7)	7 (9.7)	9 (12.5)	3 (4.2)	7 (9.7)	CE > A[4] ****	< 74%
CcGypsy1587	75	20 (26.7)	9 (12.0)	13 (17.3)	1 (1.3)	11 (14.7)	7 (9.3)	14 (18.7)	NS	< 40%
CcGypsy1692	94	35 (37.2)	9 (9.6)	17 (18.1)	4 (4.3)	6 (6.4)	5 (5.3)	18 (19.1)	CE > A[5] ****	< 48%
Total	579	199 (34)	74 (13)	97 (17)	25 (4)	69 (12)	40 (7)	75 (13)	CE > A[6] ****	< 50%

*Absolute number (percentage). **Yates one-sided chi-square testes. ***Observed related to the expected. ****Chi-Square: [1]28.0152, *p-value* = 1.20 e^{-07}; [2]9.025, *p-value* = 2.66e^{-03}; [3]12.25, *p-value* = 4.65e-04; [4]27.2453, 1.79e-07; [5]13.8462, *p-value* = 1.98e-04; [6]101.2658, *p-value* < 2.2e-16.

Genetic diversity in the diploid and the allotetraploid species

Using the raw data, the genetic diversity indexes were estimated in order to obtain the diversity variation within and among species. For that, each LTR-RT was considerate as a species, and the *C. canephora*, *C. eugenioides*, and *C. arabica*, as subpopulations of it. This strategy was considered an appropriate analysis based on the knowledge that a certain family coalesce to its ancient copy. Figure 4 shows the genetic diversity and distribution of the LTR-RT insertions. It is striking that the number of bands is not related to the genetic diversity estimated for each LTR-RT. The LTR-RTs *CcGypsy1351*, *CcGypsy1587* and *CcGypsy1692* presented the greatest numbers of bands, however they did not show discrepant Nei's gene diversity values. At the contrary, their values are similar or smaller than LTR-RTs with the lower number of bands for *CcCopia1611* and *CcCopia645*. This is understandable if we consider that the bands are shared by many genotypes in a population resulting low gene diversity.

The data on genetic distance and heterogeneity indicated that the variance among species is generally higher than within (Table 6; Table S6-S8). Only for the LTR-RTs *CcCopia310*, *CcCopia645* and *CcCopia1070* the variation is higher within than among species, for all the others the contrary is observed. Regarding the relationship between the species, in general, *C. arabica* shares with *C. canephora* and *C. eugenioides* the highest distance values, while the smallest are observed between *C. canephora* and *C. eugenioides*. The molecular variance analyses reinforce such observation, frequently segregating the LTR-RTs population from *C. arabica*.

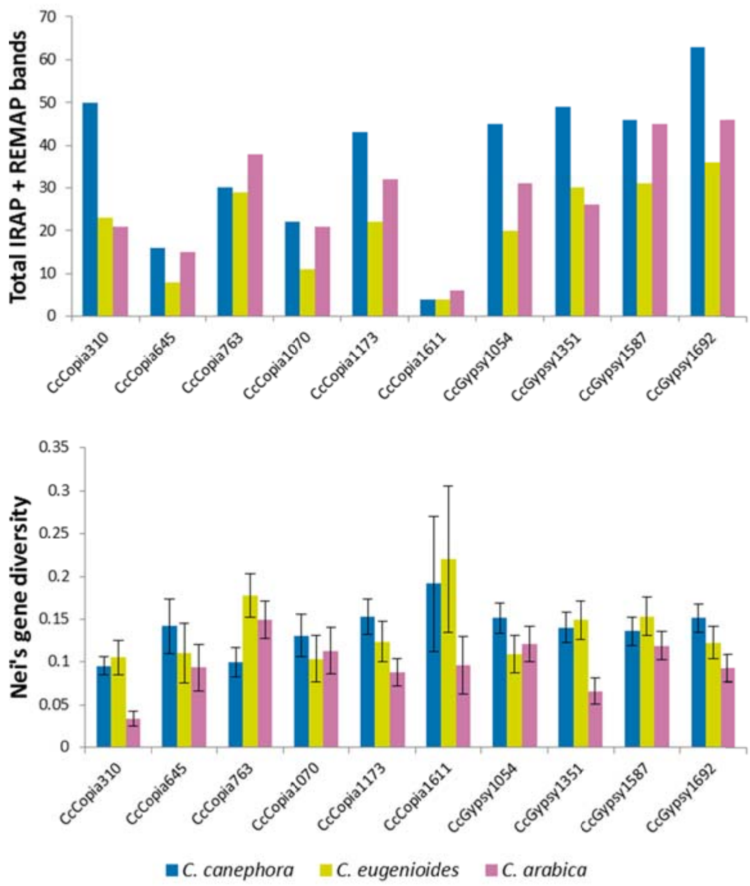


Figure 4 Genetic diversity and distribution of the LTR-RT insertions in the three *Coffea* species. (a) Number of IRAP and REMAP bands and (b) Nei's gene diversity index for each LTR-RT.

The Sorensen-Dice coefficient that measures the dissimilarity of sample pairs regarding the insertion site polymorphism of the LRT-RTs were also estimated (Table 6, Figure 5). As the dissimilarity data did not present a normal distribution within each species, the difference among and between each pair was tested using the non-parametric Kruskal-Wallis test (Table S7). The average dissimilarity of the species were significantly different for almost of the LTR-RTs; agreeing with previous analyses, *C. arabica* genotypes are more closely related between them showing low dissimilarity, whereas each progenitor species showed highest values. The dissimilarity data (previously transformed by square root) were used to obtain general arrangement of the variation by PCoAs checking the isolation of the species (Figure 5). In general, the first component explains the majority of the variation (mean 18.93 ± 1.32), with the second and third they were responsible on average for 27.45% (± 1.67) of the variation, ranging from 19.98% to 34.85%. Seven LTR-RTs presented a similar distribution of the insertion sites for *C. canephora* and *C. eugenioides* forming a group, and *C. arabica* another, significantly different. Conversely, for *CcCopia310* all the species are significantly different, agreeing for the spatial distribution observed. Finally, for *CcCopia645* and *CcCopia763*, *C. arabica* is different from *C. canephora*, but *C. eugenioides* does not differentiate itself from both. The results showed that, for almost all LTR-RTs, the distribution of the retrotransposons is different in the hybrid genome when compared with its progenitors.

Table 6 Diversity indexes among and between the three *Coffea* species.

LTR-RTs	**AMOVA***		**Nei's Genetic Distance**			**Nei's Genetic Identity**			**φ_{PT} values**			**Sorensen-Dice's dissimilarity coefficient**		
	Within	Among	Ara X Cane	Ara X Eug	Cane X Eug	Ara X Cane	Ara X Eug	Cane X Eug	Ara X Cane	Ara X Eug	Cane X Eug	*C. arabica*	*C. canephora*	*C. eugenioides*
CcCopia310	**58%**	42%	0.129	0.148	0.055	0.879	0.863	0.947	0.627	**0.704**	0.242	**0.129 (0.0756)** ^{a}	**0.948 (0.1126)** $^{a, b}$	0.600 (0.2583)$^{a, b}$
CcCopia645	**61%**	39%	**0.307**	0.215	0.165	0.736	0.807	0.848	**0.650**	0.605	0.436	0.180 (0.0945) ^{a}	**0.497 (0.2356)** ^{a}	0.480 (0.2310)
CcCopia763	26%	**74%**	**0.062**	0.117	0.060	0.940	0.890	0.941	**0.256**	0.300	0.214	**0.451 (0.1031)** ^{a}	0.536 (0.1959) ^{a}	0.643 (0.1865)
CcCopia1070	**58%**	42%	0.258	0.209	0.191	0.772	0.811	0.827	0.601	0.559	**0.498**	0.170 (0.0758)^{a,b}	0.525 (0.2199) ^{a}	0.551 (0.3997) ^{b}
CcCopia1173	44%	**56%**	0.134	0.180	0.089	0.874	0.835	0.915	0.457	0.566	0.248	0.242 (0.2096)^{a,b}	0.682 (0.2115) ^{a}	0.574 (0.1199) ^{b}
CcCopia1611	41%	**59%**	0.101	**0.320**	**0.253**	0.904	0.726	0.776	0.340	0.624	0.409	0.233 (0.2009)^{a,b}	0.561 (0.3564) ^{a}	0.597 (0.3004) ^{b}
CcGypsy1054	24%	**76%**	0.072	**0.077**	0.036	0.931	0.925	0.965	0.269	**0.280**	0.066**	0.440 (0.1329)^{a,b}	0.802 (0.1534) ^{a}	**0.719 (0.1151)** ^{b}
CcGypsy1351	48%	**52%**	0.140	0.192	0.061	0.869	0.825	0.941	0.516	0.623	**0.165**	0.151 (0.0878)^{a,b}	0.672 (0.1861) ^{a}	0.587 (0.1566) ^{b}
CcGypsy1587	34%	**66%**	0.110	0.086	0.073	0.896	0.918	0.929	0.387	0.286	0.211	0.312 (0.1389)^{a,b}	0.708 (0.1721) ^{a}	0.572 (0.2966) ^{b}
CcGypsy1692	37%	**63%**	0.117	0.108	**0.035**	0.890	0.898	0.965	0.417	0.421	0.057**	0.209 (0.0538)^{a,b}	0.594 (0.1626) ^{a}	**0.448 (0.1543)** ^{b}

Mean (Standard deviation); ** Supplementary information about AMOVA analyses see Supplementary Material Table S6. **p-value* not significant. Values with similar letters are significantly different. The highest and smallest values are highlighted.

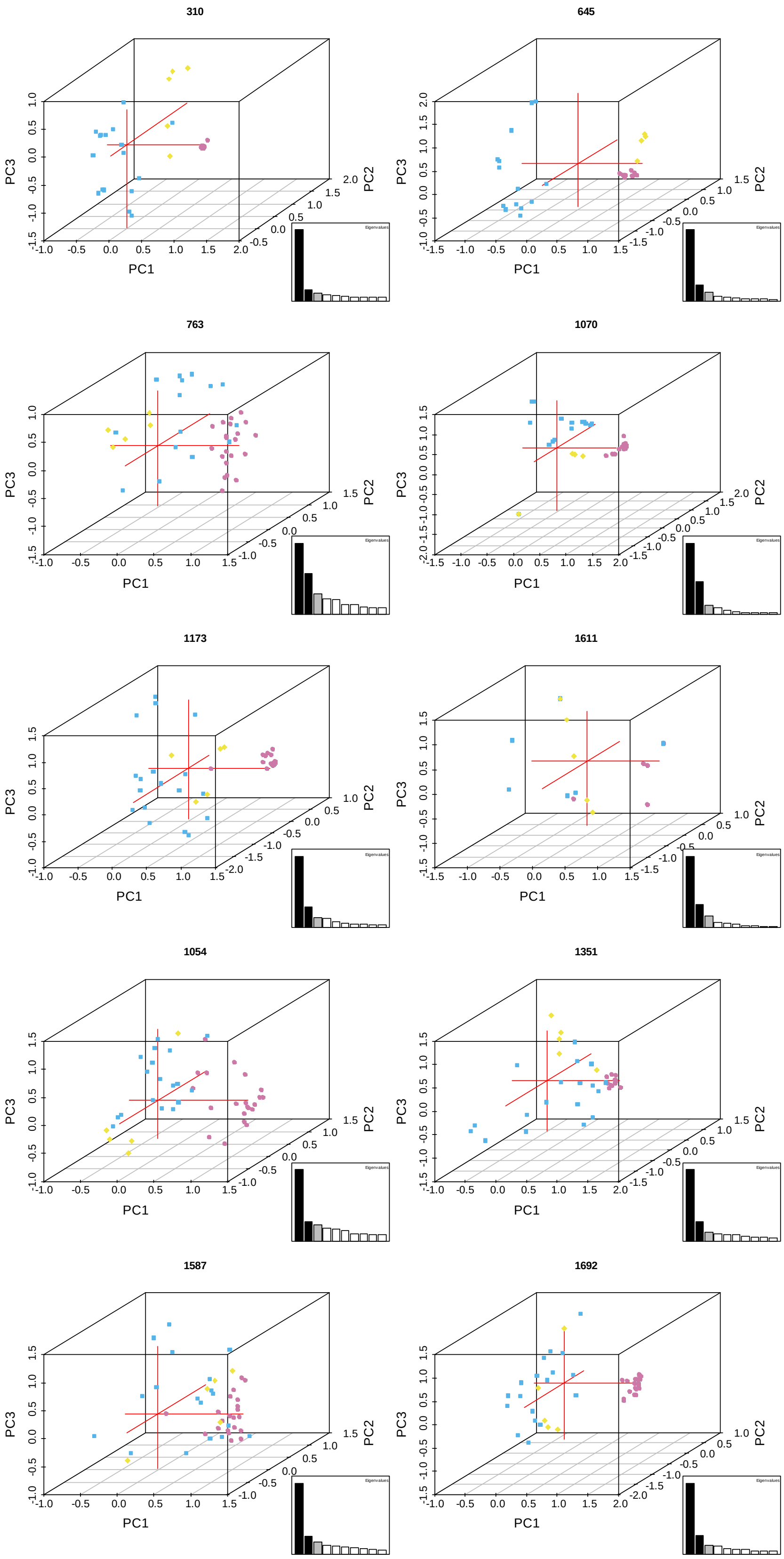


Figure 5 PCoAs of *Coffea* genotypes (blue: *C. canephora*, yellow: *C. eugenioides*, pink: *C. arabica*) based on the dissimilarity matrices, estimated by Sorensen-Dice index (Bray-Curtis for binary data, in vegan package in R). The PCoAs in three-dimensional factorial plan, the Eigen values are plotted in the right side.

Evolutionary history of LTR-RTs resulting from the hybridization

The network analysis is a valuable tool to address the evolutionary history of the LTR-RTs resulting from the *C. arabica* allotetraploidization (Figure 6). In a network the branch size is, in general, proportional to the differences segregating two nodes, and the small dark circles are ancestral state useful to explain the relationship between the genotypes. The differences among taxa cluster the genotypes joining one species with another by branches. The nodes in a central position are considered to be ancestor related to those positioned in the extremities, the derivate genotypes. When so many ancient relationships are present, intra-specific or inter-specific, the taxa involved are connected by reticulations. Although the results showed clear segregate the species on the insertion site polymorphism of the RTs, they did not illustrate a typical hybridization event, where a sum of the progenitors sites in the hybrid is expected, or it may occupy and intermediate position between the progenitors. In general, the genotypes of the same species cluster together, showing similar insertion sites pattern, in very few cases some of them cluster out of its main clade. These occurrences involve genotypes with unexpected low total insertion sites compared with others from the same species. This might result from technical artifacts, probably due to inhibitors in the sample DNA that could decrease the number of bands. A typical star topology is found only for *CcGypsy1587*, with *C. arabica* clade connecting in a central position with the branch of clade of *C. canephora* and *C. eugenioides* also well-defined. The species distribution on the *CcCopia1070* network also might support an intermediate relationship of the hybrid with the progenitor, but *C. arabica* does not connect with the other by a branch, and its genotypes are in a central position with some of the genotypes closer to *C. canephora* and others to *C. eugenioides*. Five LTR-RTs share high proportions (> 10%) of ancestral insertion sites by all species improved the reticulation (homoplasy signal) in the network, e.g. *CcCopia763*, *CcCopia1173*, *CcGypsy1054*, *CcGypsy1351*, and *CcGypsy1692*. Despite of the reticulation, the strength relationship with *C. canephora* is observed for these elements, which genotypes are localized next to central region. In the extreme case, *CcCopia763*, almost all the connections are mediated by polygons, which difficult to interpret although a close relationship with *C. canephora* can be seen. A closer relationship of *C. arabica* with *C. eugenioides* was observed only for *CcCopia645*, whose genotypes are in a central position with the branch of the hybrid at an extremity and *C. canephora* at the other. A very particular topology was found for *CcCopia1611*, a circle. The low loci number could not be enough for seeing the segregation.

The genotypes of both progenitors form a homogeneous population, involving intricate reticulations, for most of the LTR-RTs, the exceptions are *CcCopia645*, *CcCopia1070*, and *CcGypsy1587*. With rare exception, as *CcGypsy1587*, *C. arabica* set join to the progenitors through a branch connected to ancestral states, and not direct to a specific genotype. This junction occurs, in general, in the central part of topology of the progenitors set, where the wild genotypes of *C. canephora* (dark blue circles in the Figure 6) tend to be localized. Thus, the *C. arabica* genotypes seem to have a closer relationship with the wild *C. canephora* genotypes than with those derivate ones.

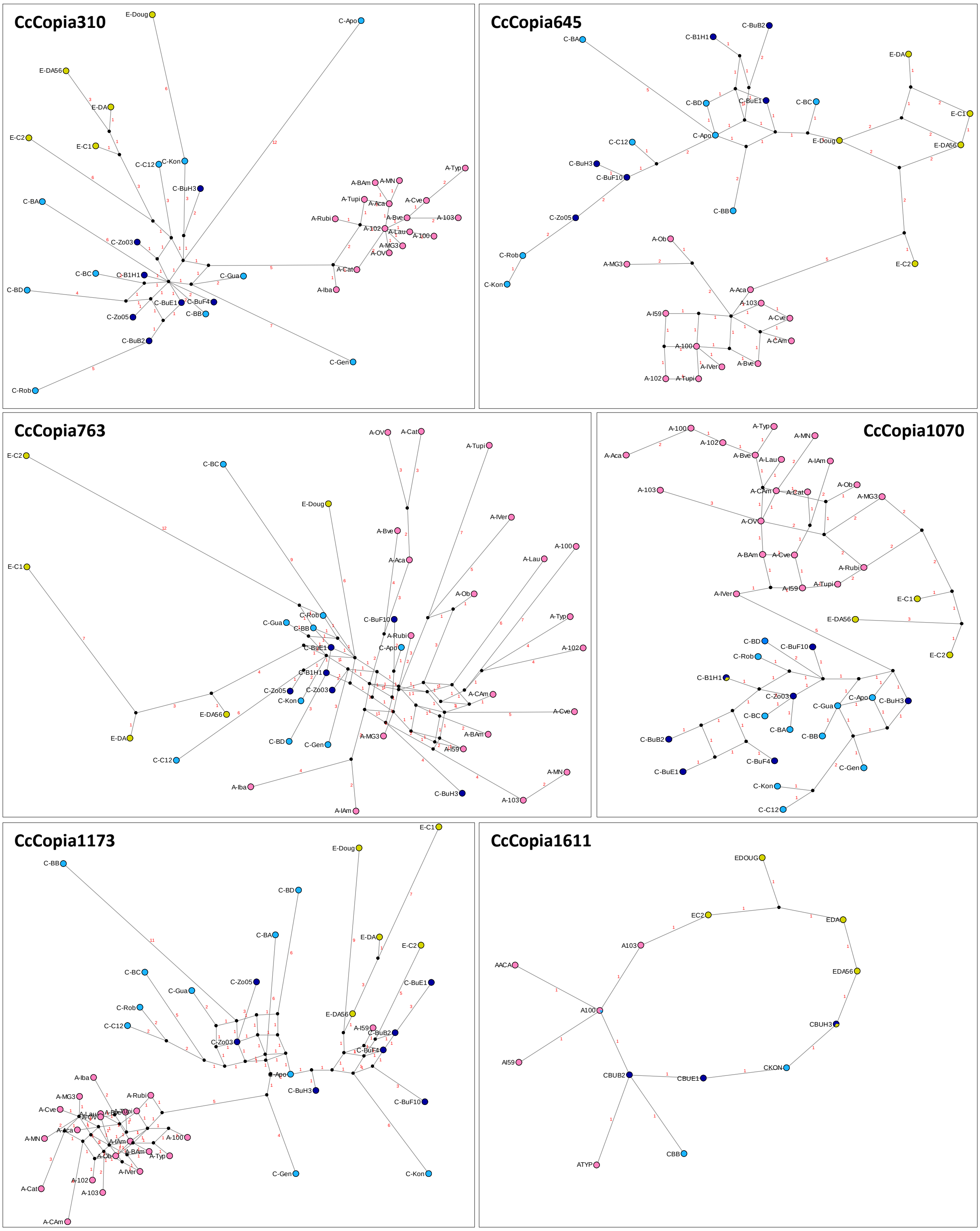


Figure 6 continues in the next page.

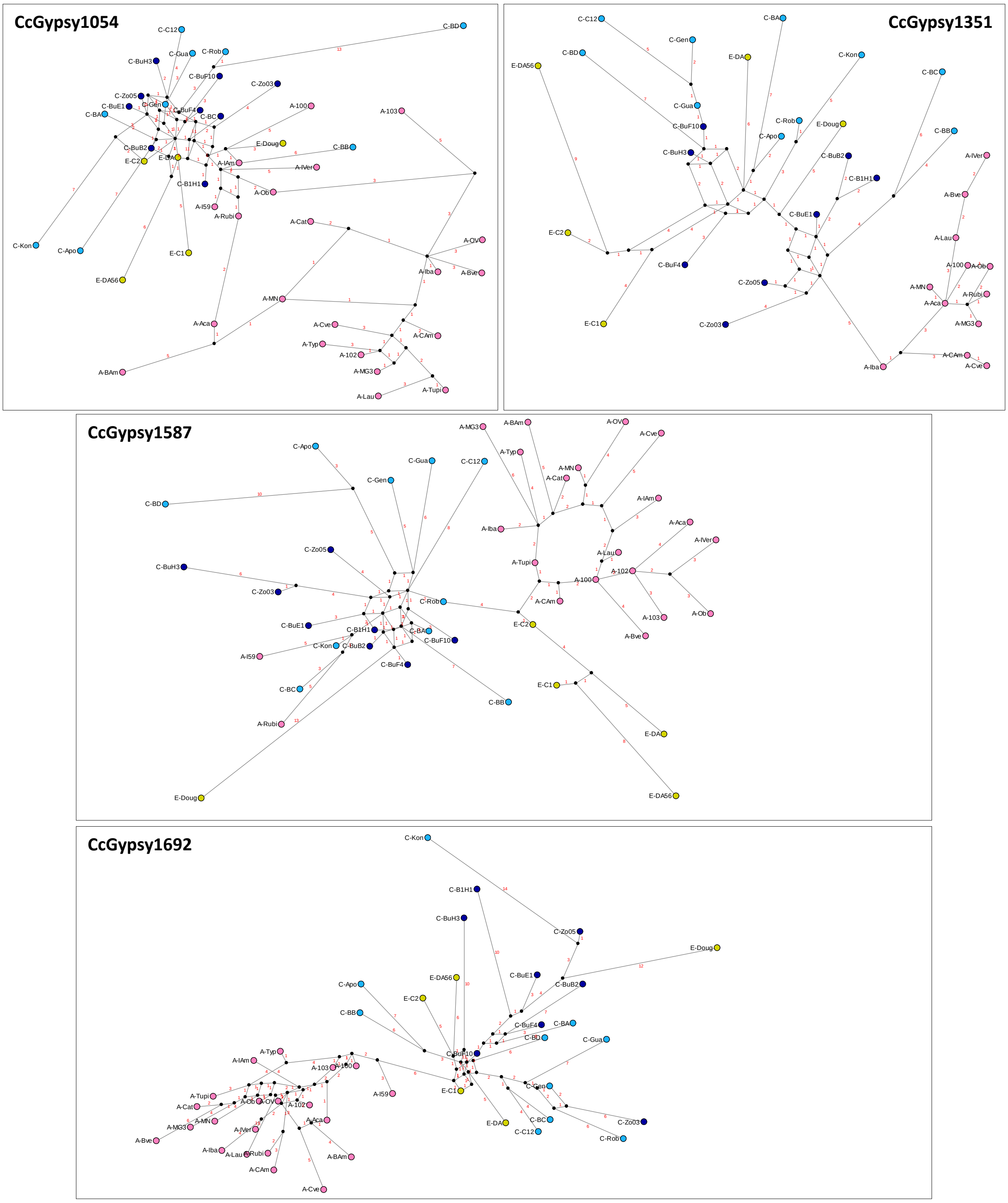


Figure 6 Network generated from the raw data matrices. The genotypes are indicated in the color nodes (dark blue: wild *C. canephora*, and light blue the other *C. canephora* genotypes; yellow: *C. eugenioides*; and, pink: *C. arabica*); the ancestor vectors, small black circle; the size of branches are proportional to differences between two nodes.

4. 4 Discussion

Transpositional activity and use of LTR-RTs as molecular markers

The capacity of integration, persistence and dispersion of the LTR-RTs in the genomes, and the occurrence of conserved motifs in their sequences, make them fit the requirements to be used as molecular markers (Kalendar *et al.* 2014). In this study, we show that none insertion site are fixed in all genotypes of the three *Coffea* species. The total genetic diversity values although relatively low, present a reduced variation (showed by the standard deviation), with only the LTR-RTs *CcGypsy1587* and *CcGypsy1692* presenting loci with wide variation, i. e. bands present in few as well as in almost all genotypes. But, absence of bands cannot mean straightforward a real absence (due to technical disturbs as inhibitors, troubles of pipetting, etc.), fixed sites might be present in the species. Regarding to *C. arabica*, only three sites are exclusive and fixed in all genotypes, two of *CcCopia310* and one of *CcGypsy1351*, all of them detected by REMAP. Although the results do not support these LTR-RTs as markers for tracking breeding of lineages, they foster their use as markers of diversity in the available resources for breeding selection. Incidentally, the presence of transcripts and of polymorphic insertion sites suggest that the LTR-RTs families here analyzed could be transpositionally active in the three *Coffea* species, though signals of transposition bursts have not been identified.

LTR-RTs conservation and genome distribution

LTR-RTs analyzed showed wide variation in the number of full-length copies in the *C. canephora* genome suggesting that these families are under different host controls or were originated at different period of time in this species. Differences in the transcriptional activity reinforce this proposition. Occurrence of bands for all 10 LTR-RTs in almost all the genotypes of the three species shows the recognizing of the primers designed in *C. canephora* in the other species. Such primer homology suggests that the LTR-RTs identified in *C. canephora* are present, with a good level of homology, in *C. eugenioides*, and, as expected due its hybrid origin, in *C. arabica*. We observed the sharing of the SSR motifs, in the REMAP results, among the three species, as previously shown (Poncet *et al.* 2004; Cubry *et al.* 2007). This conservation is expected hence the recent divergence time of the Coffeeae tribe (~ 15 Mya), to which the species belong (Bremer and Eriksson 2009), and the divergence of both parental species only about 4.2 Mya (Yu *et al.* 2011).

Both techniques, IRAP and REMAP, can help us to understand the distribution and the organization of the LTR-RTs in the host genomes. The total absence of bands of *CcCopia763*, and only a unique large band of *CcCopia1611*, in the IRAP analyses, suggest that these elements can be present in low copy number and/or dispersed in the genome, not presenting copies closely located one to each other (less than 2 Kbp). The presence of bands generated by the other elements, as well as bands with low molecular weight, amplified by the IRAP, indicates that these eight LTR-RTs contain copies close to each other in the genome. Similarly, the occurrence of bands for all LTR-RTs analyzed by REMAP indicates their proximity with microsatellites. SSRs present a random distribution in eukaryotes (Gupta and Varshney 2000), indicating a possible dispersion of the LTR-RTs studied throughout the host genomes. Although not the most frequent, the microsatellites here analyzed (di-nucleotides) show an intermediate abundance in the ESTs of *C. canephora*, the SSR8 $(CT)_n$, in special, is the most frequent among the di-nucleotides analyzed in coding sequences (Poncet *et al.* 2006), which could suggest the association of these LTR-RTs with genes. But, the low number of SSR motifs used and the small region amplified (< 2Kpb) could limit the application of these two last suggestions for all insertions of each family.

Genetic diversity in progenitor and hybrid species

Genetic diversity, expressed by the number of bands present and polymorphic in *C. arabica* is lower than in *C. canephora*, but higher than in *C. eugenioides*. Of the total bands (863) produced by the two analyses (IRAP + REMAP), about 42% (368) were observed in *C. canephora* genotypes, while only 25% (214) in *C. eugenioides*, and 33% (281) in *C. arabica*. A similar pattern was found for the average polymorphism, which was 61% and 35% for *C. canephora* and *C. eugenioides*, respectively, and 44% for *C. arabica*. These differences do not reflect the intra-specific diversity of *C. arabica* genotypes because they share more insertion sites than the parental species do. Although variable among the LTR-RTs (average of 0.252 and standard deviation (SD) of 0.117), the dissimilarity estimates for *C. arabica* is smaller, than those found for the progenitors (*C. canephora* with average of 0.652 and SD of 0.201, and *C. eugenioides* with average of 0.577 and SD 0.222). The results mean that though *C. arabica* genotypes had an intermediate number of total polymorphic bands, the bands are largely shared among the individual genotypes, decreasing the genetic diversity of the species. The reduction of the genetic diversity in wild and cultivated accessions of *C. arabica* has being reported (Carvalho *et al.* 1991; Anthony *et al.* 2002). Using AFLP markers, it was observed that the polymorphism among the wild genotypes, or among locally cultivated forms directly derived from wild ones, is much higher than among the widely cultivated varieties of *C. arabica* (Anthony *et al.* 2002).The genotypes used here are basically cultivated, their narrow genetic basis resulted from the artificial selection (Carvalho *et al.* 1991; Anthony *et al.* 2001; Anthony *et al.* 2002) could explain part of the reduced *C. arabica* intraspecific polymorphism observed here. If on the one hand, populations of an allotetraploid are expected to have higher levels of heterozygosity (Soltis and Soltis 2000), and redundant material to evolve, i.e. potential to accumulate TEs; on the other hand, genomic rearrangements following the allopolyploidization, and narrow genetic basis of its origin – and in the *C. arabica* case in the artificial selection in the genotypes origin – tend to reduce the diversity. The low diversity related to LTR-RTs insertion sites as shown in the hybrid could result from both, founder effect during its origin and dissemination, and genomic rearrangements following allopolyploidization.

An important point that should take into account is about the specific progenitor genotypes of the allotetraploid. *Coffea arabica* could be originated by a unique hybridization event between ancestors of the two closely related diploid coffee species, *C. eugenioides* and *C. canephora* (Lashermes *et al.* 1999). The 'modern' genotypes here analyzed, albeit some of *C. canephora* be wild, shall not reflect the real progenitor genotypes that originated *C. arabica*, and could input a bias in the data obtained. The populations of *C. canephora* form, at least, five well differentiated groups corresponding to geographical patterning in the individuals correlated with the natural distribution in Africa (Gomez *et al.* 2009). Among the samples here used there is at least one genotype from four of these groups, probably more since the other samples never were properly investigated under this aspect. However, the TE populations show to be homogeneous in *C. canephora* and *C. eugenioides* and not segregate one each other for most of them in PCoAs and network analyses. The connection of *C. arabica* set of elements to the central topology of *C. canephora* plus *C. eugenioides* set, together with the tendency to the wild *C. canephora* genotypes to be there localized suggest that *C. arabica* should be closer related to the wild than the derivate ones.

Evolution of the RTs inheritance in the allotetraploid hybrid

Individual variations were observed in the three species analyzed, being responsible not only for the highest genetic diversity indexes observed in *C. canephora*, but also by the moderate variation observed for *C. arabica*. When we take into account that this hybrid originated less than 1 Mya (Yu *et al.* 2011) and that 17% of the insertion sites are exclusive of it, and that *C. eugenioides,* which is much older presented 13% of specific insertions, a significant contribution to the hybrid species diversity by LTR-RTs can be pointed out; however, the data from the last species should be considered with caution due to the reduced number of genotypes used. This diversity can be a result not only of new insertions but also of rearrangements, outcome of illegitimate or unequal recombination, involving TE sequences, resulting in the increase of the intra-genotypes diversity. Globally, genomic rearrangements are improved by TEs, resulting in genomic large scale variation (Lönnig and Saedler 2002; Syvanen 1984).

Aside from TE transposition in the genomes, recombination and rearrangements of repetitive DNA are typical in the early phases of the allopolyploidization and often reported (Chang *et al.* 2010; Parisod *et al.* 2010; Koukalova *et al.* 2010; Parisod and Senerchia 2012). Although *C. arabica* shares 32% of the bands at least with its ancestors, the lower number of insertion sites in the *C. arabica* genotypes, when compared to what expected from the joining of the two progenitor genomes, marks a trend to a loss of these kinds of TE sequences. For all LTR-RTs, was observed a decrease of the total of bands in the hybrid when compared with the expected additive pattern regarding the progenitor profile, for five of ten the difference was significant. This occurrence suggests that the allopolyploid *C. arabica* could have passed by changes leading to rearrangements and losses of these types of TEs. The changes could be due to illegitimate recombination between short homologous regions, or unequal recombination between homologous or homeologous chromosomes, the occurrence of insertions close one each other (showed by IRAP) aim at the last one, but more specific analyses have to be done. Using SNPs in transcriptomic data, homoeolog losses were reported for *C. arabica*, where at least 5% of genes were inferred to display genomic changes

(Lashermes *et al.* 2014). Such losses seemed to be physically clustered in the genome sequences, including genomic fragments sizing up to several hundred kbp, suggesting that unequal crossover exchanges could be responsible for their origin (Lashermes *et al.* 2014).

Loss of diversity in allopolyploid hybrids, showed by the loss of TEs, was observed in *Spartina anglica* (Poaceae) (Baumel *et al.* 2002). This is a recent species (150 years old) originated by genome duplication of *Spartina x townsendii*, a species originated from the hybridization of the indigenous *Spartina maritima* and the introduced *Spartina alterniflora*. In *S. anglica*, few new insertions were observed, a non-exclusive band was identified, and 90.5% of its bands were shared with at least one of the parental species. TE losses also have been shown in *Nicotiana* allopolyploids, however TE genomic fractions restructure depends on the TE family, the lineage and the divergence time (Parisod *et al.* 2012). The results showed amplification and losses that varies in short-term and long-term of the *Nicotiana* allopolyploids evolution. The hybridization events that originated *Spartina* hybrid occurred recently, while those that originated the *Nicotiana* allotetraploid species had occurred earlier, 4.5 Mya. Revolutionary changes, i. e. drastic genome reorganization and bursts of transposition, is a common response to hybridization in a short-term. Evolutionary changes such as point mutations, indels and, in some cases, amplifications and local duplications occur on the long-term (Parisod and Senerchia 2012). The *C. arabica* allopolyploidization is a recent event, and there are three estimates for that: ~ 1 million years (Lashermes *et al.* 1999), ~ 665 thousand years (Yu *et al.* 2011), or 10,000–150,000 years (Cenci *et al.* 2012). Regardless the time of the origin, our results suggest that *C. arabica* could have undergone genomic reorganization at least in its TE fraction.

Our study also showed that *C. arabica* contains a considered amount of LTR-RT insertion sites inherited from *C. canephora* (12% exclusive of *C. canephora*, and 13% shared by *C. canephora* and *C. eugenioides*, and only 7% shared with *C. eugenioides*). Besides the copies shared with each progenitor, it is possible to point out a trend where *C. arabica* shares more exclusive bands with its paternal progenitor, *C. canephora*, than with its maternal progenitor, *C. eugenioides*; this is true for seven LTR-RTs studied, although only for *CcGypsy1054* the difference is significant (X^2 = 6.67, p-value = 0.0098). The closer relationship between the hybrid and its paternal progenitor becomes evident in the PCoAs and networks reconstructed analyses, in which the *C. arabica* genotypes are positioned closer to the *C. canephora* than to *C. eugenioides*. This closer relationship could be to preferential genome alteration as well as to sampling artifacts due to the reduced number of *C. eugenioides* genotypes. Besides the last possibility cannot be discarded, replacement of TE insertions in the maternal subgenome has already been observed in *C. arabica*, as shown in an analysis of 50 kbp region (Cenci *et al.* 2011). This preference was observed also for *Spartina* hybrids, where TE insertions from the maternal genome were more severely rearranged in *S. x townsendii* than in *S. x neyrautii*; while looking at the random loci this preferential behavior was not observed, with both hybrids exhibiting similar levels of rearrangements (Parisod *et al.* 2009).

Despite of some limitations of the proposals here presented, our study could be a useful framework for the understanding the LTR-RT dynamic in this allopolyploid. The data allow us to suggest that the LTR-RTs are a source of genetic diversity in the hybrid. This contribution could have occurred in two ways: a relative discreet increase in copy number, by new insertions, and moderate genomic changes, by unequal or illegitimate recombination, indicated by the private sites found and reduced insertion sites in the hybrid. Additionally, it seems to have a differential genome alteration that could result in more losses in the maternal subgenome, *C. eugenioides*, than the parental one, *C. canephora*. Further studies using wild *C. arabica* and *C. eugenioides* genotypes are needed to establish whether our findings are typical of artificial selective process to which *C. arabica* recently undergone or are a mirror of changes occurred just after the allopolyploidization that originated *C. arabica*.

References

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, 1990. Basic Local Alignment Search Tool. Journal of Molecular Biology 215: 403-410.

Anthony, F., B. Bertrand, O. Quiros, and A. Wilches, 2001 Genetic diversity of wild coffee (Coffea arabica L.) using molecular markers. Euphytica, 1 18: 53-65.

Anthony, F., C. Combes, C. Astorga, B. Bertrand, G. Graziosi, *et al.*, 2002 The origin of cultivated Coffea arabica L. varieties revealed by AFLP and SSR markers. Theor. Appl. Genet. 104: 894–900.

Bandelt, H. J., P. Forster, and A. Röhl, 1999 Median-joining networks for inferring intraspecific phylogenies. MolBiolEvol, 16:37-48.

Baumel, A., M. Ainouche, and R. Kalendar, 2002 Retrotransposons and genomic stability in populations of the young allopolyploid species Spartina anglica CE Hubbard (Poaceae). MolBiolEvol, 19:1218-27.

Bennetzen J. L., and Kellogg E. A., 1997 Do Plants Have a One-Way Ticket to Genomic Obesity? Plant Cell 9: 1509–1514.

Bento, M., D. Tomás, W. Viegas, and M. Silva, 2013 Retrotransposons Represent the Most Labile Fraction for Genomic Rearrangements in Polyploid Plant Species. Cytogenetic and Genome Research 140: 286–294.

Bouharmont, J., 1959 Recherche sur les affinités chromosomiques dans le genre CotTea. Publication I.N.E.A.C. Série scientifique n. 77, 94 p.

Bremer, B. and T. Eriksson, 2009Time Tree of Rubiaceae: Phylogeny and Dating the Family, Subfamilies, and Tribes. International Journal of Plant Sciences 170: 766793.

Brookfield, J., and R. M. Badge, 1997 Population genetics models of transposable elements. In Evolution and Impact of Transposable Elements (pp. 281-294). Springer Netherlands.

Bureau T. E., White S. E., and Wessler S. R., 1994 Transduction of a cellular gene by a plant retroelement. Cell 77: 479–480.

Carvalho, A., H. P. Medina Filho, L. C Fazuoli, O. Guerreiro Filho O, and M. M. A. Lima, 1991 Aspectos genéticos do cafeeiro. Rev Bras Genet l4:135–183

Carver, T., K. Rutherford, M. Berriman, M.-A. Rajandream, B. Barrell, *et al.*, 2005ACT: the Artemis comparison tool. Bioinformatics 21: 3422–3423.

Cenci, A., M.-C.Combes, and P. Lashermes, 2011Genome evolution in diploid and tetraploid Coffea species as revealed by comparative analysis of orthologous genome segments. Plant Mol. Biol. 78: 135–145.

Cenci, A., M.-C. C. Combes, and P. Lashermes, 2013 Differences in evolution rates among eudicotyledon species observed by analysis of protein divergence. J. Hered. 104: 459–64.

Chang, P. L., B. P. Dilkes, M. McMahon, L. Comai, and S. V. Nuzhdin, 2010 Homoeolog-specific retention and use in allotetraploid Arabidopsis suecica depends on parent of origin and network partners. Genome Biol. 11: R125.

Charlesworth, B., and D. Charlesworth, 1983 The population dynamics of transposable elements. Genetical Research, 42: 1–27.

Comai L., 2005 The advantages and disadvantages of being polyploid. Nat. Rev. Genet. 6: 836–46.

Cubry, P., P. Musoli, H. Legnaté, D. Pot, F. de Bellis, *et al.*, 2008 Diversity in coffee assessed with SSR markers: structure of the genus Coffea and perspectives for breeding. Genome 51: 50–63

Denoeud, F., L. Carretero-Paulet, A. Dereeper, G. Droc, R. Guyot, *et al.*, 2014 The genome of coffee displays ancestral asterid gene order and provides insight into caffeine, aroma and flavor evolution. Science 345: 1181–4.

Deshmukh, V. P. et al. A simple method for isolation of genomic DNA from fresh and dry leaves of Terminalia arjuna (Roxb.) Wight and Arnot. **Electronic Journal of Biotechnology**, v. 10, n. 3, p. 468–472, 2007.

Dietrich, W. F., J. C. Miller, R. G. Steen, M. Merchant, D. Damron, *et al.* 1994 A genetic map of the mouse with 4,006 simple sequence length polymorphisms. Nat Genet. 7: 220-45.

Feschotte C., Jiang N., and Wessler S., 2002 Plant transposable elements: where genetics meets genomics. Nature Reviews Genetics 3: 329–341

Gilbert W., 1978 Why genes in pieces? Nature 271: 501–501.

Gupta, P.K., and R. K. Varshney, 2000 The development and use of microsatellite markers for genetic analysis and plant breeding with emphasis on bread wheat. Euphytica 113: 163–185.

Guyot, R., F. Lefebvre-Pautigny, C. Tranchant-Dubreuil, M. Rigoreau, P. Hamon, *et al.* 2012 Ancestral synteny shared between distantly-related plant species from the asterid (Coffea canephora and Solanum Sp.) and rosid (Vitis vinifera) clades. BMC Genomics 13: 103.

Hamon, P., S. Siljak-Yakovlev, S. Srisuwan, O. Robin, V. Poncet, *et al.*, 2009 Physical mapping of rDNA and heterochromatin in chromosomes of 16 Coffea species: a revised view of species differentiation. Chromosome Res., 17, 291–304.

Jackson, S, and Z. J. Chen, 2010 Genomic and expression plasticity of polyploidy. CurrOpin Plant Biol, 13: 153-159

Jiao, Y., N. J. Wickett, S. Ayyampalayam, A. S. Chanderbali, L. Landherr, *et al.*, 2011 Ancestral polyploidy in seed plants and angiosperms. Nature 473: 97–100.
Kalendar, R. and A. H. Schulman, 2006. IRAP and REMAP for retrotransposon-based genotyping and fingerprinting. Nat Protoc., 1: 2478-84.

Kalendar, R. and A. H. Schulman, 2014 Transposon-based tagging: IRAP, REMAP, and iPBS. Methods Mol. Biol. 1115: 233–55.

Katoh, K. and D. M. Standley, 2013 MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular biology and evolution 30: 772–780.

Koukalova, B., A. P. Moraes, S. Renny-Byfield, R. Matyasek, A. R. Leitch, *et al.*, 2010 Fall and rise of satellite repeats in allopolyploids of Nicotiana over c. 5 million years. New Phytol. 186: 148–60.

Langley, C. H., J. Brookfield, and N. Kaplan, 1983 Transposable elements in Mendelian populations. I. A theory. Genetics, 104: 457–471.

Lashermes, P., M. C. Combes, J. Robert, P. Trouslot, A. D'Hont, *et al.*, 1999 Molecular characterisation and origin of the Coffea arabica L. genome. Mol. Gen. Genet. 261: 259–66.

Lashermes, P., M.-C. C. Combes, Y. Hueber, D. Severac, and A. Dereeper, 2014 Genome rearrangements derived from homoeologous recombination following allopolyploidy speciation in coffee. Plant J. 78: 674–85.

Leitch, I. J. and M. D. Bennett, 2004 Genome downsizing in polyploid plants. Biological Journal of the Linnean Society, 82: 651-663.

Feldman, M., and A. A. Levy, 2009 Genome evolution in allopolyploid wheat - a revolutionary reprogramming followed by gradual changes. J Genet Genomics 36: 511–8.

Lönnig, W-E., and H. Saedler, 2002 Chromosome rearrangements and transposable elements.AnnuRevGenet 36: 389–410.

Ma, J., and J. Bennetzen, 2004 Rapid recent growth and divergence of rice nuclear genomes.Proceedings of the National Academy of Sciences of the United States of America. 101: 12404–10.

Madlung, A., A. P. Tyagi, B. Watson, H. Jiang, T. Kagochi, *et al.*, 2005 Genomic changes in synthetic Arabidopsis polyploids. Plant J. 41: 221–30.

Marraccini, P., L. P. Freire, G. S. Alves, N. G. Vieira, F. Vinecky, *et al.*, 2011 RBCS1 expression in coffee: Coffea orthologs, Coffea arabica homeologs, and expression variability between genotypes and under drought stress. BMC Plant Biol. 11: 85.

Masterson, J., 1994 Stomatal size in fossil plants – evidence for polyploidy in majority of angiosperms. Science 264:1759-1763.

Mayrose, I., S. H. Zhan, C. J. Rothfels, and K. Magnuson-Ford, 2011 Recently formed polyploid plants diversify at lower rates. Science.333:1257.

McClintock, B., 1984 The significance of responses of the genome to challenge. Science 226: 792-801.

Parisod, C., and N. Senerchia, 2012. Responses of transposable elements to polyploidy. In: Plant Transposable Elements (Eds. Grandbastien&Casacuberta), Topics in Current Genetics, Springer 24: 147-168.

Parisod, C., A. Salmon, T. Zerjal, M. Tenaillon, M.-A. A. Grandbastien *et al.*, 2009 Rapid structural and epigenetic reorganization near transposable elements in hybrid and allopolyploid genomes in Spartina. New Phytol. 184: 1003–15.

Parisod, C., C. Mhiri, K. Y. Lim, J. J. Clarkson, M. W. Chase, *et al.* (2012) Differential Dynamics of Transposable Elements during Long-Term Diploidization of Nicotiana Section Repandae (Solanaceae) Allopolyploid Genomes. PLoS ONE 7: e50352.

Parisod, C., K. Alix, J. Just, M. Petit, V. Sarilar, *et al.* 2010 Impact of transposable elements on the organization and function of allopolyploid genomes. The Newphytologist. 186:37–45.

Parisod, C., R. Holderegger, and C. Brochmann, 2010 Evolutionary consequences of autopolyploidy.New Phytol. 186:5-17.

Peakall, R., and P. E. Smouse, 2012 GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research--an update. Bioinformatics 28: 2537–9.

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

Poncet, V., M. Rondeau, C. Tranchant, A. Cayrel, S. Hamon, *et al.* 2006. SSR mining in coffee tree EST databases: potential use of EST-SSRs as markers for the Coffea genus. Molecular Genetics and Genomics 276: 436-449.

R Core Team, 2014 R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.URL http://www.R-project.org/.

Rouzic, A. Le and P. Capy, 2005The first steps of transposable elements invasion: parasitic strategy vs. genetic drift. Genetics 169: 1033–43.

Samarakoon, T., S. Y. Wang, and M. H. Alford, 2013 Enhancing PCR amplification of DNA from recalcitrant plant specimens using a trehalose-based additive. Applications in Plant Sciences 1: 1200236.

Soltis, P., 2013 Hybridization, speciation and novelty. Journal of Evolutionary Biology 26: 291–293.

Soltis, P. and D. Soltis, 2000 The role of genetic and genomic attributes in the success of polyploids. Proceedings of the National Academy of Sciences 97: 7051-7057

Syvanen, M., 1984 The evolutionary implications of mobile genetic elements. Annual Review of Genetics, 18: 271-293.

Tamura, K., G. Stecher, D. Peterson, A. Filipski, and S. Kumar 2013 MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Molecular Biology and Evolution: 30 2725-2729.

Tesfaye, K, T. Borsch, K. Govers, and E. Bekele 2007. Characterization of Coffea chloroplast microsatellites and evidence for the recent divergence of *C. arabica* and *C. eugenioides* chloroplast genomes. Genome, 50: 1112–1129.

Vicient, C. M., and A. Suoniemi, 1999 Retrotransposon BARE-1 and its role in genome evolution in the genus Hordeum. Plant Cell,11:1769-1784.

Wang, W., H. Zheng, C. Fan, J. Li, J. Shi, *et al.*, 2006 High rate of chimeric gene origination by retroposition in plant genomes. The Plant Cell 18: 1791–1802.

Wicker, T., F. Sabot, A. Hua-Van, J. Bennetzen, P. Capy *et al.*, 2007 A unified classification system for eukaryotic transposable elements. Nature Reviews Genetics 8: 973–982.

Wood, T. E., N. Takebayashi, M. Barker, I. Mayrose, P. B. Greenspoon, *et al.*, 2009 The frequency of polyploid speciation in vascular plants. Proceedings of the National Academy of Sciences 106: 13875–13879.

Yu, Q., R. Guyot, A. de Kochko, A. Byers, R. Navajas-Perez, *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.

Supplementary Material

Table S1 Description of the genotype pedigree or botanical group information of the genotypes used in this work.

	Species	Genotype and/or botanical group	Characteristics of interest
C. canephora			
1	DH200-94	Double haploid, Congolese diversity group. Genome sequenced.	
2	BA58	Guianese diversity group.	
3	BB56	Congolese diversity group.	
4	BC56	Congolese diversity group.	
5	BD64	Congolese diversity group.	
6	Apoatã IAC 3597	Moderately resistant to rust and susceptible to root-knot nematode. Apoatã IAC 2258 (rootstock)	
7	IAC 784 (C12)	-	
8	Guarini IAC 1598-11-3		Moderately resistant to rust and to *Meloidogyne exígua*, *M. incognita* and *M. paranaensis*.
9	Kouilou IAC 67-4	It came from Indonesia by Dr. Edmundo Navarro de Andrade.	Resistant to rust and Highly resistant to *M. exígua*, and/or tolerant to M. incognita and M. paranaensis
10	Robusta IAC 1564 (C5)		Moderately resistant to rust and to *M. exígua*, *M. incognita* and *M. paranaensis*.
11	BuB2	Budango Forest, Biiso block, Compartment B3	
12	BuE1	Budango Forest, Nyafungo block, Compartment N1 & 2	
13	BuF4	Budango Forest, Nyafungo block, Compartment N3	
14	BuF10	Budango Forest, Nyafungo block, Compartment N3	
15	BuH1	Budango Forest, Siba block, Compartment S1	Wild
16	BuH3	Budango Forest, Siba block, Compartment S1	
17	Zo03	Zoka Forest North Uganda	
18	Zo05	Zoka Forest North Uganda	
C. eugenioides			-
1	IAC 1140-24 (C1)		-
2	IAC 1098-7 (C2)		
3			
4			
5			
C. arabica			
1	Typica IAC 537	An ancient genotype, probably ancestral of Bourbon Vermelho.	
2	Acaiá IAC 1474-19	Individual selection of Mundo Novo plants, probably from the Sumatra genotype, which is involved in the Mundo Novo origin. It is a hybrid between Sumatra and Bourbon Vermelho.	Susceptible to rust and to root-knot nematode.
3	Bourbon Amarelo IAC J19	Originated from spontaneous mutation of Bourbon, or by the natural cross between Bourbon Vermelho and Amarelo de Botucatu (Xanthocarpa).	
4	Bourbon Vermelho IAC	It came from Island Reunion.	
5	CatuaíAmarelo IAC 62	AS. It is a hybrid of Mundo Novo (IAC 374-19) and Caturra Amarelo (IAC 476-11).	
6	CatuaíVermelho	AS. It is a hybrid of Mundo Novo (IAC 374-19) and	

		Caturra Amarelo (IAC 476-11).	
7	CaturraVermelho IAC 477	It is originated from one or two mutations of the Bourbon Vermelho genotype.	
8	Ibairi IAC 4761	It is a hybrid of Mokka and Bourbon Vermelho.	
9	Laurina IAC870	It came from the Island Reunion; it was considered an interspecific hybrid between *C. arabica* and *C. mauritiana*, but it might also be considered as originated from mutations in Bourbon Vermelho.	
10	Mundo Novo IAC 379-19	It is a hybrid between Bourbon Vermelho and Typica/Sumatra.	
11	IAC Ouro Verde H 5010-5	It is a hybrid between CatuaíAmarelo IAC H 2077-2-12-70 and Mundo Novo IAC 515-20.	
12	Catiguá MG3	AS. It is a hybrid between CatuaíAmarelo IAC 86 and Híbrido de Timor* (UFV 440-10).	Resistant to rust and to root-knot nematode.
13	ObatãVermelho IAC 1669-20	It is a hybrid between Villa Sarchi and Híbrido de Timor (CIFC 832/2), progenies of the F2 crossed with CatuaíVermelho.	Resistant to rust and susceptible to root-knot nematode.
14	TupiVermelho IAC 1669-33	It is a hybrid between Villa Sarchi and Híbrido de Timor (CIFC 832/2).	
15	IcatuAmarelo IAC 2944	It is a hybrid between IcatuVermelho and Bourbon Amarelo or Mundo Novo Amarelo.	Resistant or moderately resistant to rust; and
16	IcatuVermelho IAC 4041	It is a interspecific hybrid between *C. canephora* (a tetraploid plant) and *C. arabica* (Bourbon Vermelho)	susceptible to root-knot nematode
17	IPR 100 “Catindú”	Catuaí SH2,SH , originated from the cross between Catuaí with a hydric before originated between Catuaí and H7314-4 of the serie BA-10, carrying genes of *C. liberica*.	Resistant to root-knot nematode.
18	IPR 102 “Catucaí”	It is a hybrid between Catuaí and Icatu.	Moderately resistant to rust.
19	IPR 103 “Catucaí”	It is a hybrid between Catuaí and Icatu.	
20	IAPAR 59*	It is a hybrid between Villa Sarchi (CIFC 971/10) and Híbrido de Timor (CIFC 832/2).	Resistant to rust and to root-knot nematode (*M. exigua*). Tolerant to deficit hydric.
21	Rubi*- MG 1192	It is a hybrid between Catuaí and Mundo Novo.	Susceptible to rust and to root-knot nematode.

CIFC – Centro de Investigação das Ferrugens do Cafeeiro, Portugal.
*Originated from a natural cross between *C. arabica* and a gamete not reduced of *C. canephora*.

Table S2 Information about the sequences of the 10 LTR-RTs in *C. canephora* genome. * Indicate the reference copy.

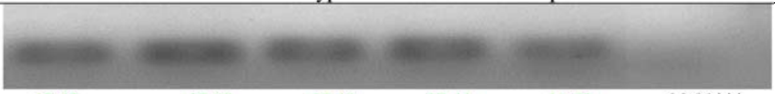
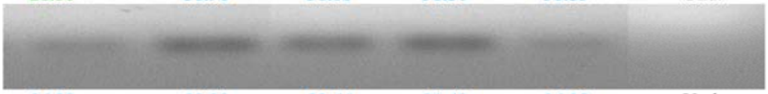
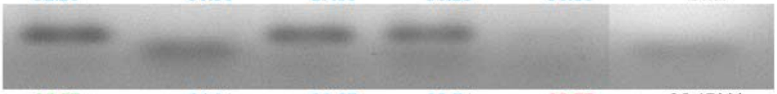
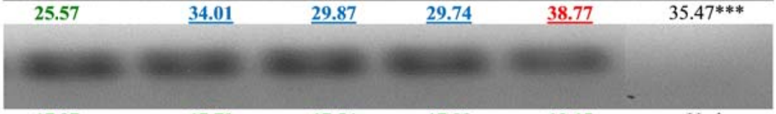
LTR-RT	Sequence	Localization	Length	Start	End
CcCopia310	1	chr0	3731	174401773	174398043
	2	chr0	3892	119296511	119300402
	3	chr0	3839	58229930	58226092
	4	chr0	4179	8469839	8465661
	5	chr0	4755	111689904	111685150
	6	chr0	4775	142876542	142871768
	7	chr0	4615	8374102	8378716
	8	chr0	3816	81954712	81958527
	9	chr0	3837	62653025	62656861
	10	chr0	3789	147348731	147344943
	11	chr0	3978	44125451	44121474
	12	chr0	4199	154053651	154057849
	13	chr0	4006	113256985	113260990
	14	chr0	3913	110341352	110345264
	15	chr0	3236	15564049	15560814
	16	chr0	3839	106975104	106971266
	17	chr0	4047	56047779	56043733
	18	chr0	4600	17709349	17713948
	19	chr0	3940	7042344	7038405
	20	chr0	3157	109146871	109150027
	21	chr0	3066	66272433	66275498
	22	chr1	3847	17520881	17524727
	23	chr1	3786	34907495	34903710
	24	chr1	4685	13215791	13211107
	25	chr1	3741	156526	152786
	26	chr1	4727	19912718	19907992
	27	chr2	3762	2129002	2125241
	28	chr2	3862	49629968	49626107
	29	chr2	3919	50312519	50316437
	30	chr2	3861	48291566	48287706
	31	chr2	4746	42737163	42732418
	32	chr2	3873	4856375	4852503
	33	chr2	3859	38276817	38272959
	34	chr2	3857	23666191	23670047
	35	chr2	4349	3465699	3461351
	36	chr3	4672	21118235	21113564
	37	chr3	3729	18277332	18281060
	38	chr3	5565	21397850	21392286
	39	chr4	3799	24482727	24478929
	40	chr4	3898	18139457	18135560
	41	chr4	3796	18835459	18831664
	42	chr4	3837	13362569	13358733
	43	chr4	4679	28039192	28034514
	44	chr4	4700	22493945	22489246
	45	chr4	2782	5772122	5769341
	46	chr5	4670	2060929	2056260
	47	chr5	3678	30161	26484
	48	chr5	4655	11306543	11311197
	49	chr6	3829	19534468	19538296
	50	chr6	3803	12475493	12479295
	51	chr6	4356	8349209	8353564
	52	chr6	5472	16075549	16081020
	53	chr6	3666	20669264	20672929
	54	chr7	3850	26275639	26271790
	55	chr7	3793	24841502	24837710
	56	chr7	4675	29825854	29830528
	57	chr7	5441	6006697	6001257
	58	chr7	3798	20494719	20490922
	59	chr7	4841	3828115	3832955
	60	chr8	3570	17686733	17683164
	61	chr8	3780	17084228	17088007
	62	chr9	3825	17647384	17643560

	63	chr9	3778	19967985	19964208
	64	chr9	4657	19779415	19784071
	65	chr10	3831	8396232	8392402
	66	chr10	3812	15159979	15156168
	67	chr10	3848	22275125	22278972
	68	chr10	3862	13244995	13248856
	69	chr10	4655	27235268	27239922
	70	chr10	4042	19234222	19230181
	71	chr10	4024	21428174	21432197
	72*	chr11	3446	183768	187213
	73	chr11	3805	162037	158233
	74	chr11	3430	21465495	21462066
	75	chr11	4628	14330333	14325706
	76	chr11	4589	215316	219904
	77	chr11	4692	31465819	31461128
	78	chr11	4952	2501292	2506243
	79	chr11	3740	6826290	6822551
	80	chr11	4642	2974916	2979557
CcCopia645	1	chr0	12128	3928288	3940415
	2	chr0	8812	77363339	77372150
	3	chr0	9587	164380775	164371189
	4	chr0	9630	64021658	64012029
	5	chr0	9615	151058809	151049195
	6	chr0	10074	14716744	14706671
	7	chr0	9534	57371158	57361625
	8	chr0	9338	111574495	111565158
	9	chr0	9524	111098754	111108277
	10	chr0	9529	118246339	118236811
	11	chr0	9660	92438123	92447782
	12	chr0	8654	91065530	91056877
	13	chr0	9523	2114356	2104834
	14	chr0	9518	131502873	131512390
	15	chr0	8876	21291727	21300602
	16	chr0	9493	168144588	168154080
	17	chr0	8401	129224159	129232559
	18	chr0	9560	24924468	24934027
	19	chr0	8576	83879055	83887630
	20	chr0	9885	20839711	20829827
	21	chr0	10160	109186488	109196647
	22	chr0	9321	28499775	28509095
	23	chr0	8897	51155493	51146597
	24	chr0	10000	122441312	122451311
	25	chr0	10407	88065578	88055172
	26	chr0	9895	154083952	154093846
	27	chr1	10724	14447203	14436480
	28	chr1	9463	14571307	14561845
	29*	chr2	10958	13619256	13630213
	30	chr2	9761	42685391	42675631
	31	chr2	9638	31583590	31573953
	32	chr2	8886	30174128	30183013
	33	chr2	8878	28153445	28162322
	34	chr2	8985	36671130	36680114
	35	chr2	9590	13608751	13618340
	36	chr3	10591	10945733	10935143
	37	chr3	9601	17772733	17763133
	38	chr3	9345	16825395	16834739
	39	chr4	9451	16584007	16574557
	40	chr4	10086	9481559	9471474
	41	chr5	9493	10432122	10441614
	42	chr5	9996	10574185	10564190
	43	chr5	10032	12920902	12910871
	44	chr6	9485	23587226	23577742
	45	chr6	9541	30339998	30349538
	46	chr6	9606	24265426	24255821
	47	chr6	9861	31231635	31221775
	48	chr6	10008	27210763	27200756
	49	chr7	9556	24397108	24387553

	50	chr8	9526	11400289	11390764
	51	chr8	10002	12183974	12173973
	52	chr9	9892	15162853	15152962
	53	chr9	9519	15361405	15351887
	54	chr10	10568	19865765	19855198
	55	chr10	9623	13257677	13267299
	56	chr10	9550	12899597	12909146
	57	chr11	9523	4386801	4377279
	58	chr11	9622	2241785	2251406
	59	chr11	8775	13121296	13112522
	60	chr11	8941	18504689	18513629
	61	chr11	9540	15892658	15883119
	62	chr11	8135	3669459	3661325
CcCopia763	1*	chr0	5425	32330598	32336022
	2	chr0	5443	175214065	175208623
	3	chr0	6715	98178689	98185403
	4	chr0	5441	49607979	49602539
	5	chr0	5426	88829973	88824548
	6	chr0	5681	124201118	124206798
	7	chr0	5445	31587053	31592497
	8	chr0	5394	9815060	9820453
	9	chr0	5429	132514255	132508827
	10	chr0	5426	31036151	31030726
	11	chr0	10727	22347013	22336287
	12	chr0	5245	31305328	31300084
	13	chr1	5428	27965674	27960247
	14	chr1	5404	23622140	23616737
	15	chr1	5450	19641296	19635847
	16	chr1	5316	30379469	30384784
	17	chr1	5573	10357787	10363359
	18	chr1	5644	4898203	4903846
	19	chr2	5353	36191120	36196472
	20	chr2	5370	25986888	25981519
	21	chr2	8367	14652617	14644251
	22	chr2	5423	41882470	41887892
	23	chr4	5380	22802479	22797100
	24	chr4	5402	22336130	22330729
	25	chr5	5423	17580408	17585830
	26	chr5	5601	11285903	11280303
	27	chr5	5379	7286352	7280974
	28	chr6	5387	19954421	19949035
	29	chr6	5418	32442534	32437117
	30	chr6	5402	27213275	27218676
	31	chr6	9020	14666664	14675683
	32	chr6	5326	13415389	13410064
	33	chr6	5245	16723278	16728522
	34	chr6	5528	12107646	12102119
	35	chr7	5460	21797468	21802927
	36	chr7	6484	28665984	28659501
	37	chr8	5403	3471953	3466551
	38	chr9	5395	4379355	4373961
	39	chr10	5390	11036051	11041440
	40	chr10	5431	24815546	24810116
	41	chr11	5425	18552203	18557627
	42	chr11	5376	15683484	15688859
	43	chr11	5394	10030955	10036348
	44	chr11	10483	17690760	17680278
	45	chr11	5418	16514721	16520138
	46	chr11	5539	16553146	16558684
	47	chr11	5626	25363046	25368671
	48	chr11	5367	24964519	24959153
CcCopia1070	1	chr1	5648	2636884	2631237
	2	chr2	5603	36045900	36040298
	3	chr4	5645	14161275	14155631
	4	chr5	5697	891651	897347
	5	chr7	5656	9745176	9739521
	6	chr7	5133	16020600	16025732

	7	chr9	5673	8159365	8165037
	8*	chr10	5717	20173933	20179649
	9	chr10	5671	27597120	27602790
	10	chr10	5309	5823450	5818142
CcCopia1173	1	chr0	14041	125359470	125373510
	2*	chr3	9544	24300739	24291196
	3	chr4	11042	12834830	12845871
	4	chr9	9162	15109090	15118251
CcCopia1611	1*	chr6	5017	19221962	19226978
	2	chr6	4967	32369764	32364798
CcGypsy1054	1	chr0	5797	129011255	129017051
	2	chr0	5741	112459132	112453392
	3	chr0	5820	68830152	68824333
	4	chr2	6174	43675488	43669315
	5*	chr8	5752	15934846	15940597
	6	chr8	5784	20367937	20373720
CcGypsy1351	1*	chr0	7946	88897664	88905609
	2	chr0	5298	125118514	125113217
	3	chr0	7777	46964049	46956273
	4	chr0	8072	147687009	147678938
	5	chr0	7795	16119709	16127503
	6	chr0	6280	88935308	88929029
	7	chr0	7357	93350349	93357705
	8	chr0	7785	125232135	125239919
	9	chr0	8769	109023358	109014590
	10	chr0	8122	44405461	44397340
	11	chr0	8177	59374223	59366047
	12	chr0	7337	59366824	59359488
	13	chr1	7317	15483963	15476647
	14	chr3	7376	23888948	23881573
	15	chr5	7525	12248768	12256292
	16	chr5	7777	12241727	12249503
	17	chr7	10250	16883590	16893839
	18	chr9	8601	12477472	12486072
CcGypsy1587	1	chr0	4280	84941729	84937450
	2*	chr2	5243	34246937	34252179
	3	chr2	5183	14584739	14579557
	4	chr10	3936	9591554	9587619
CcGypsy1692	1	chr0	10737	29688423	29677687
	2	chr0	9989	71644271	71634283
	3	chr0	10399	35307067	35317465
	4	chr0	10723	146884354	146873632
	5	chr0	10421	84118345	84128765
	6*	chr1	9648	29459929	29469576
	7	chr1	10626	5581718	5571093
	8	chr1	10447	26730765	26741211
	9	chr3	10535	23716660	23706126
	10	chr4	10711	2782972	2772262
	11	chr6	10508	14254111	14243604
	12	chr6	10660	32920995	32931654
	13	chr6	10236	28136698	28126463
	14	chr6	5563	25658555	25652993
	15	chr7	10374	4577349	4587722
	16	chr7	10250	11588942	11599191
	17	chr8	10152	17960233	17950082
	18	chr8	10283	14045190	14034908
	19	chr8	10327	13302027	13291701
	20	chr9	10439	17785518	17775080
	21	chr9	10485	299053	309537
	22	chr10	10446	12630866	12620421
	23	chr10	10597	3854443	3865039
	24	chr11	9587	31365584	31375170

Table S3 Electrophoresis gel images and Ct values of the RT-PCR of the ten LTR-RTs and the constitutive gene ubiquitin (BUBI) in five samples, one of each progenitor and tree of the allotetraploid.

LTR-RT*	*C. canephora*	*C. eugenioides*	*C. arabica*			negative
	Robusta	C1	Typica	Mundo Novo	Iapar 59	
CcCopia310						
	17.45	**17.83**	**17.19**	**17.41**	**18.59**	38.90***
CcCopia 645						
	26.38	**20.59**	**24.27**	**24.14**	**26.02**	39.40***
CcCopia 763						
	27.28	**25.07**	**28.14**	**28.82**	**32.66**	Und.
CcCopia 1070						
	28.56	**31.96**	**31.81**	**31.50**	**31.53**	Und.
CcCopia 1173						
	36.20	**30.30**	**32.66**	**31.49**	**36.38**	Und.
CcCopia 1611						
	Und.	**35.99**	**28.48**	**27.41**	**31.17**	Und.
CcGypsy1054						
	33.80	**28.73**	**32.99**	**33.37**	**34.15**	36.77***
CcGypsy 1351						
	Und.	Und.	Und.	Und.	Und.	Und.
CcGypsy 1587						
	32.26	**30.51**	**29.81**	**30.28**	**31.81**	Und.
CcGypsy 1692						
	25.57	**34.01**	**29.87**	**29.74**	**38.77**	35.47***
BUBI						
	17.97	**17.70**	**17.54**	**17.99**	**19.15**	Und.
BUBI (RNA - rt)**						
	Und.	Und.	Und.	Und.	Und.	Und.

* The values above the images correspond to the Ct (cycle threshold), defined as the number of cycles required for the fluorescent signal to cross the threshold, which were determined for each primer combination in the exponential phase (logarithm graph). Und. = Undetermined, is value as negative. In **green**: Cts< 29 are strong positive reactions indicative of abundant target nucleic acid in the sample; **blue**: Cts of 30-37 are positive reactions indicative of moderate amounts of target nucleic acid; **red**: Cts of 38-40 are weak reactions indicative of minimal amounts of target nucleic acid which couldrepresent an infection state or environmental contamination. Reference: Real Time PCR Ct Values, Retrieved June 07, 2015, from http://www.wvdl.wisc.edu/wp-content/uploads/2013/01/WVDL.Info_.PCR_Ct_Values1.pdf.

** Checking DNA contamination. Results of the negative cDNA, i.e., RNA minus transcriptase reverse enzyme.

*** The values underlined can't be consider because the difference between the Ct of the negative and the sample is less than 10, indicating that the value of the sample can be background of the reaction. The values without underline can be considerate.

Table S4 Summary of genetic parameters for 18 genotypes of *C. canephora*, 5 of *C. eugenioides* and 21 of *C. arabica* obtained from IRAP analysis.

LTR-RT	Species	Insertion sites	Polymorphism (Freq.>= 5%)	Private sites	I ± SE	h ± SE	uh ± SE
CcCopia310	***C. arabica***	1	1	0	0.00 ± 0.000	0.00 ± 0.000	0.00 ± 0.000
	C. canephora	30	30	27	0.19 ± 0.022	0.10 ± 0.014	0.11 ± 0.014
	C. eugenioides	11	11	8	0.14 ± 0.039	0.09 ± 0.026	0.12 ± 0.033
	Total	**42**	**42**	**35**	**0.11 ± 0.016**	**0.06 ± 0.011**	**0.07 ± 0.013**
CcCopia645	***C. arabica***	15	14	9	0.14 ± 0.038	0.09 ± 0.025	0.09 ± 0.027
	C. canephora	12	12	9	0.14 ± 0.038	0.09 ± 0.026	0.09 ± 0.028
	C. eugenioides	8	8	4	0.14 ± 0.046	0.10 ± 0.032	0.13 ± 0.040
	Total	**35**	**34**	**22**	**0.14 ± 0.023**	**0.09 ± 0.016**	**0.10 ± 0.018**
CcCopia1070	***C. arabica***	11	9	2	0.18 ± 0.050	0.11 ± 0.033	0.11 ± 0.035
	C. canephora	13	13	7	0.26 ± 0.054	0.16 ± 0.038	0.17 ± 0.040
	C. eugenioides	4	4	0	0.11 ± 0.053	0.08 ± 0.035	0.09 ± 0.044
	Total	**28**	**26**	**9**	**0.18 ± 0.031**	**0.11 ± 0.021**	**0.13 ± 0.023**
CcCopia1173	***C. arabica***	18	8	9	0.12 ± 0.026	0.07 ± 0.017	0.07 ± 0.017
	C. canephora	24	24	15	0.20 ± 0.035	0.12 ± 0.024	0.13 ± 0.025
	C. eugenioides	12	12	6	0.15 ± 0.039	0.10 ± 0.026	0.12 ± 0.033
	Total	**54**	**44**	**30**	**0.15 ± 0.019**	**0.09 ± 0.013**	**0.11 ± 0.015**
CcGypsy1054	***C. arabica***	8	4	4	0.12 ± 0.043	0.07 ± 0.028	0.07 ± 0.030
	C. canephora	14	14	10	0.23 ± 0.041	0.13 ± 0.027	0.14 ± 0.029
	C. eugenioides	2	2	1	0.04 ± 0.035	0.03 ± 0.025	0.03 ± 0.032
	Total	**24**	**20**	**15**	**0.13 ± 0.025**	**0.08 ± 0.016**	**0.08 ± 0.018**
CcGypsy1351	***C. arabica***	19	19	5	0.11 ± 0.029	0.07 ± 0.019	0.08 ± 0.020
	C. canephora	38	38	19	0.25 ± 0.029	0.15 ± 0.020	0.16 ± 0.021
	C. eugenioides	23	23	8	0.23 ± 0.040	0.16 ± 0.028	0.20 ± 0.035
	Total	**80**	**80**	**32**	**0.20 ± 0.020**	**0.13 ± 0.013**	**0.14 ± 0.016**
CcGypsy1587	***C. arabica***	22	12	6	0.26 ± 0.038	0.15 ± 0.027	0.16 ± 0.029
	C. canephora	15	15	4	0.18 ± 0.040	0.11 ± 0.027	0.11 ± 0.029
	C. eugenioides	13	13	2	0.26 ± 0.056	0.18 ± 0.038	0.22 ± 0.048
	Total	**50**	**40**	**12**	**0.23 ± 0.026**	**0.15 ± 0.018**	**0.17 ± 0.021**
CcGypsy1692	***C. arabica***	18	9	12	0.10 ± 0.025	0.06 ± 0.016	0.06 ± 0.017
	C. canephora	30	30	25	0.16 ± 0.022	0.08 ± 0.014	0.09 ± 0.014
	C. eugenioides	13	13	5	0.13 ± 0.033	0.08 ± 0.022	0.11 ± 0.028
	Total	**61**	**52**	**42**	**0.13 ± 0.016**	**0.08 ± 0.010**	**0.09 ± 0.012**

*N: number of genotypes per species, I: Shannon's information index, h: genetic diversity index, uh: Unbiased Diversity = (N / (N-1)) * h, where for Haploid Binary data, p = Band Freq. and q = 1 – p; SE: standard error.

Table S5 Summary of genetic parameters for 18 genotypes of *C. canephora*, 5 of *C. eugenioides* and 21 of *C. arabica* obtained from REMAP analysis.

LTR-RT	Species	Insertion sites	Polimorphism (Freq.>= 5%)	Private sites	I ± SE	h ± SE	uh ± SE
CcCopia310	***C. arabica***	18	12	10	0.16 ± 0.033	0.09 ± 0.020	0.10 ± 0.021
	C. canephora	17	17	8	0.19 ± 0.033	0.10 ± 0.020	0.11 ± 0.021
	C. eugenioides	6	6	0	0.11 ± 0.042	0.07 ± 0.028	0.09 ± 0.035
	Total	**41**	**35**	**18**	**0.15 ± 0.021**	**0.09 ± 0.013**	**0.10 ± 0.015**
CcCopia645	***C. arabica***	0	0	0	0.00 ± 0.000	0.00 ± 0.000	0.00 ± 0.000
	C. canephora	4	4	4	0.55 ± 0.075	0.37 ± 0.065	0.39 ± 0.069
	C. eugenioides	0	0	0	0.00 ± 0.000	0.00 ± 0.000	0.00 ± 0.000
	Total	**4**	**4**	**4**	**0.18 ± 0.081**	**0.12 ± 0.056**	**0.13 ± 0.060**
CcCopia763	***C. arabica***	38	27	17	0.24 ± 0.032	0.15 ± 0.022	0.16 ± 0.023
	C. canephora	30	30	13	0.17 ± 0.026	0.10 ± 0.017	0.11 ± 0.018
	C. eugenioides	29	29	10	0.26 ± 0.037	0.18 ± 0.025	0.22 ± 0.032
	Total	**97**	**86**	**40**	**0.22 ± 0.019**	**0.14 ± 0.013**	**0.16 ± 0.015**
CcCopia1070	***C. arabica***	7	6	2	0.22 ± 0.081	0.15 ± 0.057	0.16 ± 0.060
	C. canephora	8	8	5	0.20 ± 0.054	0.11 ± 0.035	0.12 ± 0.037
	C. eugenioides	3	3	1	0.12 ± 0.061	0.07 ± 0.039	0.09 ± 0.049
	Total	**18**	**17**	**8**	**0.18 ± 0.038**	**0.11 ± 0.026**	**0.12 ± 0.028**
CcCopia1173	***C. arabica***	14	13	1	0.22 ± 0.050	0.13 ± 0.034	0.14 ± 0.036
	C. canephora	19	19	5	0.35 ± 0.044	0.22 ± 0.033	0.23 ± 0.035
	C. eugenioides	10	10	1	0.26 ± 0.067	0.18 ± 0.047	0.22 ± 0.058
	Total	**43**	**42**	**7**	**0.28 ± 0.032**	**0.18 ± 0.022**	**0.20 ± 0.026**
CcCopia1611	***C. arabica***	6	4	1	0.18 ± 0.058	0.10 ± 0.033	0.10 ± 0.035
	C. canephora	4	4	1	0.28 ± 0.113	0.19 ± 0.079	0.20 ± 0.084
	C. eugenioides	4	4	0	0.31 ± 0.121	0.22 ± 0.085	0.28 ± 0.106
	Total	**14**	**12**	**2**	**0.26 ± 0.057**	**0.17 ± 0.040**	**0.19 ± 0.047**
CcGypsy1054	***C. arabica***	23	18	3	0.23 ± 0.039	0.15 ± 0.027	0.15 ± 0.028
	C. canephora	31	31	10	0.27 ± 0.032	0.16 ± 0.022	0.17 ± 0.024
	C. eugenioides	18	18	5	0.23 ± 0.041	0.15 ± 0.027	0.19 ± 0.034
	Total	**72**	**67**	**18**	**0.24 ± 0.022**	**0.15 ± 0.015**	**0.17 ± 0.017**
CcGypsy1351	***C. arabica***	7	4	2	0.08 ± 0.037	0.05 ± 0.023	0.05 ± 0.024
	C. canephora	11	11	7	0.20 ± 0.052	0.12 ± 0.036	0.13 ± 0.038
	C. eugenioides	7	7	5	0.19 ± 0.060	0.13 ± 0.040	0.16 ± 0.050
	Total	**25**	**22**	**14**	**0.16 ± 0.029**	**0.10 ± 0.020**	**0.11 ± 0.023**
CcGypsy1587	***C. arabica***	23	19	7	0.16 ± 0.031	0.10 ± 0.020	0.10 ± 0.021
	C. canephora	31	31	16	0.25 ± 0.032	0.15 ± 0.021	0.16 ± 0.023
	C. eugenioides	18	18	7	0.21 ± 0.040	0.14 ± 0.027	0.17 ± 0.034
	Total	**72**	**68**	**30**	**0.21 ± 0.020**	**0.13 ± 0.013**	**0.15 ± 0.015**
CcGypsy1692	***C. arabica***	28	25	5	0.21 ± 0.040	0.14 ± 0.028	0.14 ± 0.029
	C. canephora	33	33	10	0.36 ± 0.039	0.23 ± 0.028	0.25 ± 0.029
	C. eugenioides	23	23	4	0.25 ± 0.044	0.17 ± 0.030	0.21 ± 0.038
	Total	**84**	**81**	**19**	**0.27 ± 0.024**	**0.18 ± 0.017**	**0.20 ± 0.019**

*N: number of genotypes per species, I: Shannon's information index, h: genetic diversity index, uh: Unbiased Diversity = (N / (N-1)) * h, where for Haploid Binary data, p = Band Freq. and q = 1 – p; SE: standard error.

Table S6 Gene diversity parameters in the three *Coffea* species.

LTR-RTs	Total	*C. arabica*			*C. canephora*			*C. eugenioides*		
	H_T	H_S	D_{ST}	G_{ST}	H_S	D_{ST}	G_{ST}	H_S	D_{ST}	G_{ST}
CcCopia310	**0.130 (0.1463)**	**0.034 (0.0786)**	0.096	**0.741**	**0.096 (0.0915)**	0.034	0.263	0.105 (0.1704)	0.025	0.189
CcCopia645	**0.238 (0.1786)**	0.094 (0.1541)	**0.144**	0.606	0.142 (0.1822)	0.096	0.402	0.110 (0.1967)	**0.128**	0.537
CcCopia763	0.168 (0.1442)	**0.149 (0.1726)**	**0.019**	**0.111**	0.100 (0.1367)	0.068	0.404	0.178 (0.2014)	**-0.010**	**-0.061**
CcCopia1070	0.230 (0.1856)	0.114 (0.1643)	0.117	0.507	0.131 (0.1503)	**0.099**	**0.430**	**0.104 (0.1653)**	0.126	**0.549**
CcCopia1173	0.185 (0.1675)	0.088 (0.1281)	0.097	0.522	0.153 (0.1588)	0.032	0.174	0.124 (0.1865)	0.061	0.330
CcCopia1611	0.224 (0.1376)	0.096 (0.0938)	0.127	0.569	**0.191 (0.2247)**	0.032	0.144	**0.220 (0.241)**	0.004	0.017
CcGypsy1054	0.166 (0.1459)	0.121 (0.1618)	0.045	0.269	0.152 (0.1354)	**0.014**	**0.085**	0.109 (0.1653)	0.056	0.340
CcGypsy1351	0.174 (0.1644)	0.066 (0.1303)	0.108	0.621	0.140 (0.1489)	0.033	0.191	0.149 (0.1943)	0.025	0.142
CcGypsy1587	0.179 (0.1638)	0.119 (0.1426)	0.060	0.335	0.136 (0.1453)	0.043	0.239	0.154 (0.1905)	0.025	0.141
CcGypsy1692	0.171 (0.1674)	0.093 (0.1536)	0.078	0.455	0.152 (0.1596)	0.020	0.114	0.123 (0.1809)	0.049	0.284

Mean (Standard deviation); H_T is the total diversity; H_S diversity within the population; $D_{ST}= H_S - H_{T,}$ is differentiation among the populations; $G_{ST} = D_{ST} / H_S$, apportionment of diversity among the populations (Nei, 1973). Highlighted in black the highest values, and in red the lowest.

Table S7 Results of the Kruskal-Wallis test using the dissimilarity matrices regarding to the insertion site polymorphism in the three *Coffea* species. In red p-values not significant. ω = 0.001.

LTR-RT	Hypotheses H_0	Chi-squared	df		*p-value*
CcCopia310	E = C = A	287.4118	2	<	2.20E-16
	E = C	34.145	1	=	5.12E-09
	E = A	25.6892	1	=	4.01E-07
	C = A	275.2715	1	<	2.20E-16
CcCopia645	E = C = A	156.5712	2	<	2.20E-16
	E = C	6.00E-04	1	=	0.9797
	E = A	9.2365	1	=	0.002372
	C = A	153.6084	1	<	2.20E-16
CcCopia763	E = C = A	26.1313	2	=	2.12E-06
	E = C	2.6754	1	=	0.1019
	E = A	10.2012	1	=	0.001403
	C = A	19.0543	1	=	1.27E-05
CcCopia1070	E = C = A	205.3203	2	<	2.20E-16
	E = C	0.0432	1	=	0.8354
	E = A	14.2415	1	=	0.000161
	C = A	202.4212	1	<	2.20E-16
CcCopia1173	E = C = A	192.3539	2	<	2.20E-16
	E = C	3.1378	1	=	0.0765
	E = A	18.7263	1	=	1.51E-05
	C = A	185.5596	1	<	2.20E-16
CcCopia1611	E = C = A	65.3715	2	=	6.38E-15
	E = C	0.2092	1	=	0.6474
	E = A	14.2704	1	=	0.000158
	C = A	58.2281	1	=	2.33E-14
CcGypsy1054	E = C = A	232.7078	2	<	2.20E-16
	E = C	3.0919	1	=	0.07868
	E = A	22.5782	1	=	2.02E-06
	C = A	224.6799	1	<	2.20E-16
CcGypsy1351	E = C = A	270.9707	2	<	2.20E-16
	E = C	1.9074	1	=	0.1673
	E = A	28.4787	1	=	9.47E-08
	C = A	260.8222	1	<	2.20E-16
CcGypsy1587	E = C = A	219.9218	2	<	2.20E-16
	E = C	2.7179	1	=	0.09923
	E = A	12.1809	1	=	0.000483
	C = A	217.374	1	<	2.20E-16
CcGypsy1692	E = C = A	271.923	2	<	2.20E-16
	E = C	6.9625	1	=	0.008323
	E = A	26.4451	1	=	2.71E-07
	C = A	261.7624	1	<	2.20E-16

Table S8 Analysis of Molecular Variance (AMOVA) of *Coffea* species (df = Degrees of freedom; SS = Sum of square; % Percentage of variation). Permutation of 9,999.

	Source	df	SS	MS	Est. Var.	%	P(rand>= data)
CcCopia310	**Among Pops**	2	104.971	52.486	3.820	**58%**	0.001
	Within Pops	41	112.302	2.739	2.739	42%	0.001
	Total	43	217.273		6.559	100%	
CcCopia645	**Among Pops**	2	86.133	43.066	3.155	**61%**	0.001
	Within Pops	41	81.117	1.978	1.978	39%	0.001
	Total	43	167.250		5.134	100%	
CcCopia763	**Among Pops**	2	49.012	24.506	1.539	26%	0.001
	Within Pops	41	183.238	4.469	4.469	**74%**	0.001
	Total	43	232.250		6.008	100%	
CcCopia1070	**Among Pops**	2	89.944	44.972	3.271	**58%**	0.001
	Within Pops	41	97.306	2.373	2.373	42%	0.001
	Total	43	187.250		5.644	100%	
CcCopia1173	**Among Pops**	2	90.306	45.153	3.164	44%	0.001
	Within Pops	41	162.057	3.953	3.953	**56%**	0.001
	Total	43	252.364		7.116	100%	
CcCopia1611	**Among Pops**	2	13.091	6.545	0.453	41%	0.001
	Within Pops	41	26.273	0.641	0.641	**59%**	0.001
	Total	43	39.364		1.094	100%	
CcGypsy1054	**Among Pops**	2	44.057	22.029	1.364	24%	0.001
	Within Pops	41	174.670	4.260	4.260	**76%**	0.001
	Total	43	218.727		5.625	100%	
CcGypsy1351	**Among Pops**	2	107.463	53.731	3.812	48%	0.001
	Within Pops	41	167.514	4.086	4.086	**52%**	0.001
	Total	43	274.977		7.898	100%	
CcGypsy1587	**Among Pops**	2	80.758	40.379	2.699	34%	0.001
	Within Pops	41	214.197	5.224	5.224	**66%**	0.001
	Total	43	294.955		7.924	100%	
CcGypsy1692	**Among Pops**	2	104.771	52.386	3.556	37%	0.001
	Within Pops	41	249.229	6.079	6.079	**63%**	0.001
	Total	43	354.000		9.635	100%	

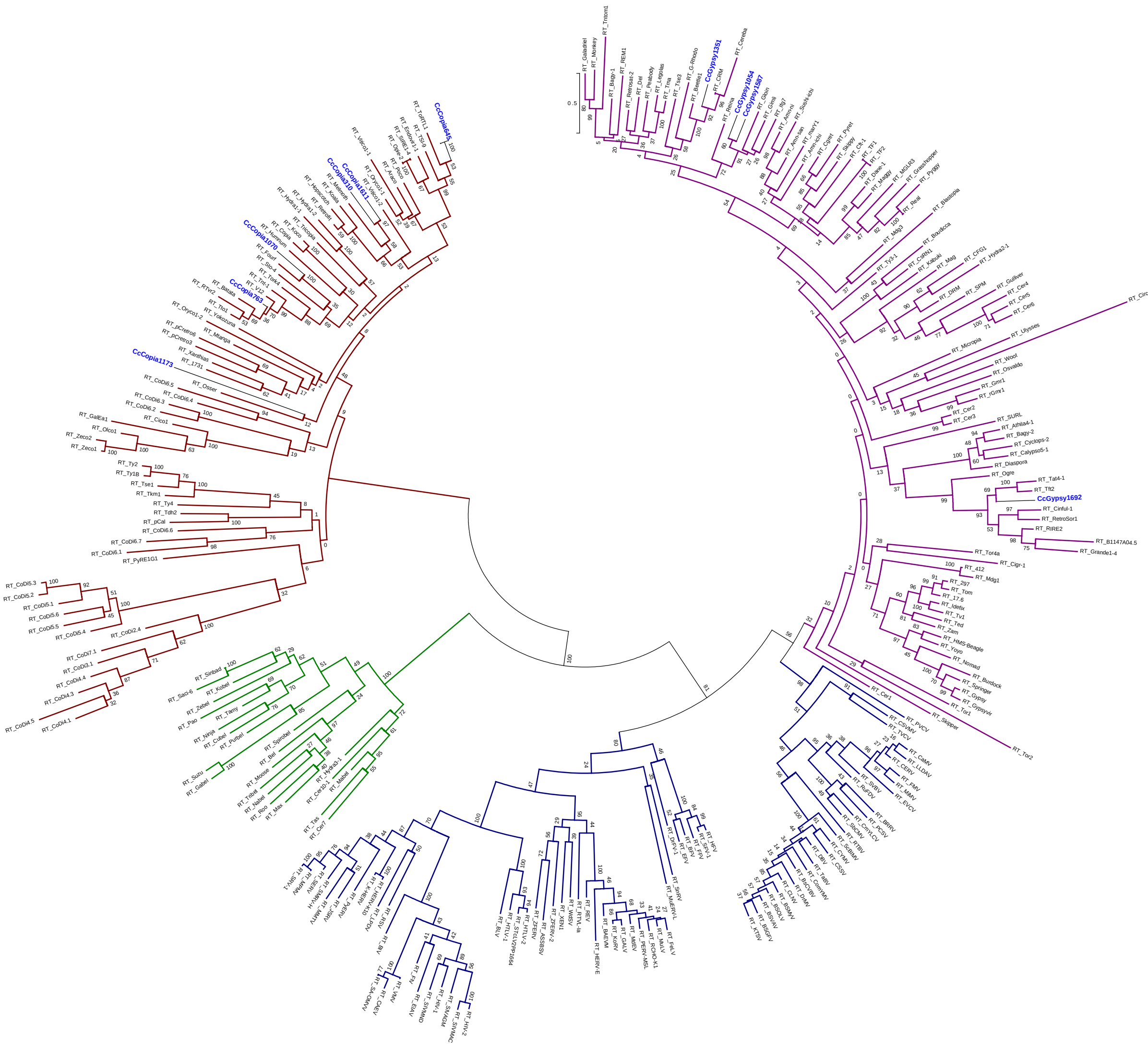


Figure S1 Phylogeny of the *reverse transcriptase* domain of 10 LTR retrotransposons identified in *C. canephora* genome and of LTR-RT *reverse transcriptase* domain deposited in GyDB. In red, the *Ty1/Copia* clade; pink: *Ty3/Gypsy*; green: Bel/Pao; blue: retrovirus clades. In blue, the elements characterized in this study; and in black, other *reverse transcriptase* sequences of retrotransposons from GyDB. The phylogenetic relationships were inferred by using the Maximum Likelihood method based on the JTT matrix-based model, with 100 replicates of bootstrap. The tree with the highest log likelihood (-64946.5978) is shown. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using a JTT model, and then selecting the topology with superior log likelihood value. A discrete Gamma distribution was used to model evolutionary rate differences among sites (2 categories (+G, parameter = 3.1189)). The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. The analysis involved 279 amino acid sequences. All positions with less than 80% site coverage were eliminated. That is, fewer than 20% alignment gaps, missing data, and ambiguous bases were allowed at any position. There were a total of 174 positions in the final dataset. Evolutionary analyses were conducted in MEGA6 (Tamura et al. 2013).

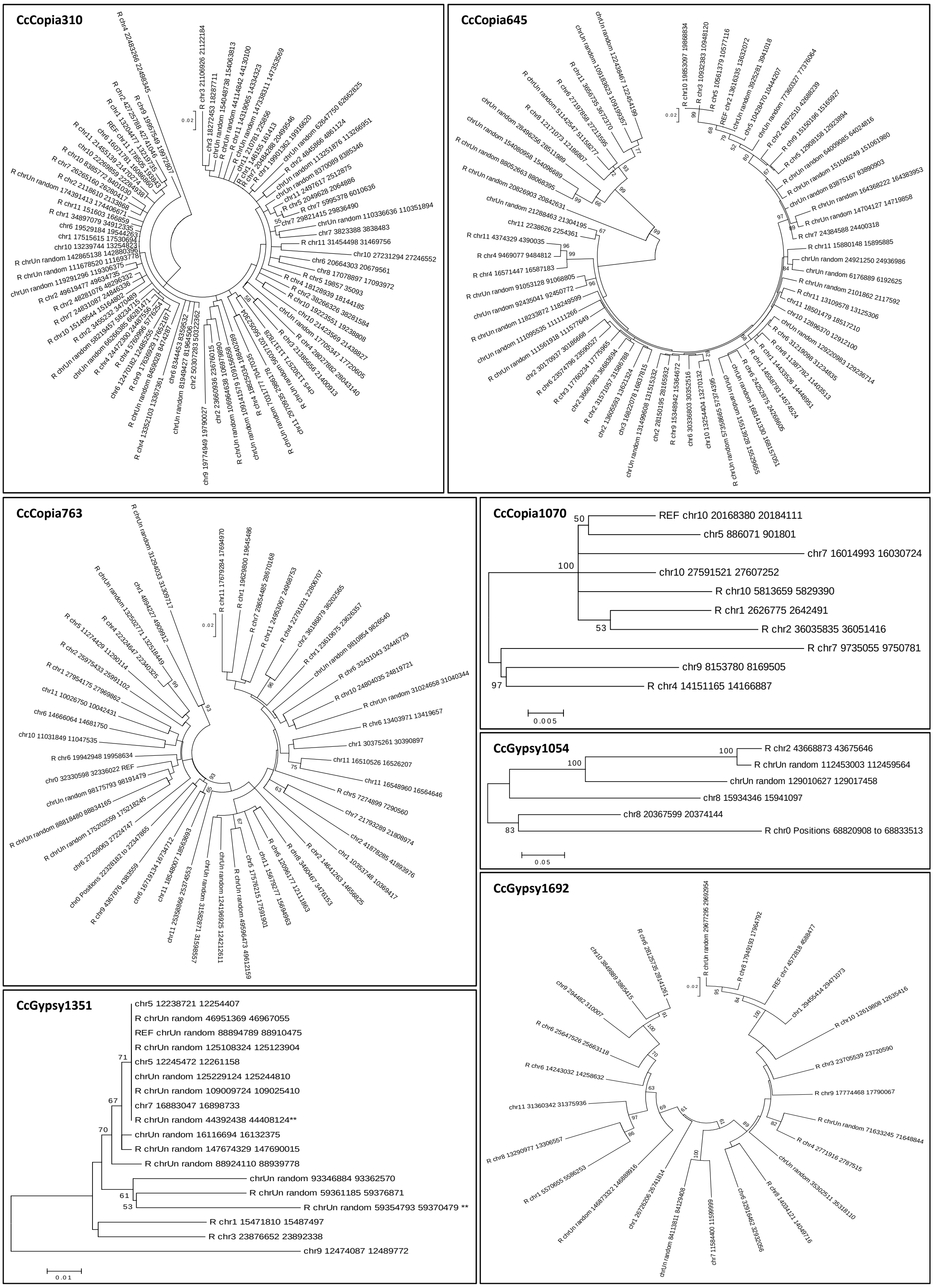


Figure S2 Phylogenies of the LTR-RTs reconstructed using the reverse transcriptase domain, showing the presence or not of subfamilies in the *C. canephora* genome. Only the bootstraps upper 70 was considered, and upper 50% were showed. The phylogenies were reconstructed using Tamura-3 parameters, with 1000 replicates of bootstrap.

5 DISCUSSÃO GERAL

5 DISCUSSÃO GERAL

Os TEs tem um papel importante na evolução dos genomas. Em geral, todas as famílias de DNA repetitivo podem potencialmente ter um impacto sobre a organização genômica como geradores de instabilidade – uma vez que sequências com múltiplas cópias são substratos poderosos para recombinação (HEDGES; DEININGER, 2007) –, ou como componentes de domínios cromossômicos essenciais, tais como centrômeros e telômeros (WONG; CHOO, 2004; LAMB et al., 2007). Nesse contexto, ressalta-se a ocorrência dos LTR-RTs nos genomas de plantas. O modo de transposição replicativa, característico desses elementos, pode resultar em um grande acúmulo no número de cópias, como observado em *Vicia faba*, em que somente os elementos da superfamília *Ty1/Copia* somam cerca de um milhão de cópias (PEARCE et al., 1996), ou na composição do genoma de algumas espécies de plantas, como no milho, em que os LTR-RTs constituem mais de 75% das sequências nucleotídicas (SANMIGUEL et al., 1998; TENAILLON et al.,2011). Essa presença conspícua e constante no genoma de plantas e a complexidade genômica relacionada à sua presença ressaltam a importância do entendimento da origem e da expansão dos TEs para que, dessa forma, possa-se entender a estrutura e a evolução desses genomas.

A transmissão dos TEs em geral ocorre verticalmente e muitas de suas histórias evolutivas são marcadas por perdas nas espécies divergentes. Em casos eventuais, essas sequências são transmitidas horizontalmente e colonizam novos genomas. Essa transferência, no caso dos retrotransposons, é propiciada principalmente por dois fatores relacionados ao seu ciclo de vida: a ocorrência de uma forma intermediária estável – o RNA transcrito reversamente – e de uma fase citoplasmática. A transferência horizontal (HT) constitui uma oportunidade para a colonização de novos genomas e, cada vez mais, estudos mostram que foi, e continua sendo, uma estratégia importante na evolução dos TEs, garantindo sua permanência e sobrevivência ao longo do tempo evolutivo. Eventos dessa natureza envolvendo TEs têm sido inferidos há um longo tempo para animais (revisão em LORETO et al., 2008, SCHAACK et al., 2010; WALLAU et al., 2012) e apenas mais recentemente para plantas (DIAO et al., 2006; FORTUNE et al., 2008; ROULIN et al., 2008; CHENG et al., 2009; EL BAIDOURI et al., 2014). Nossos resultados com o retrotransposon *Copia25* acrescem esses achados e descrevem um evento de HT e de uma inesperada alta similaridade de sequências envolvendo espécies de dicotiledôneas e monocotiledôneas que divergiram há 150 milhões de anos. Sequências de *Copia25* identificadas em espécies do gênero *Musa* (monocotiledônea, subclasse Zingiberidae) se agrupam com alto suporte filogenético com sequências do gênero *Ixora* (dicotiledônea, subclasse Asteridae) dentro do clado da família Rubiaceae. Hipóteses alternativas à HT foram ponderadas e refutadas. *Copia25* apresenta uma distribuição heterogênea entre os táxons – estando presente em sete dos 41 genomas avaliados –, uma alta conservação de sequências entre espécies distantemente relacionadas – 85% de identidade nucleotídica na *pol* entre *Copia25* de *C. canephora* e *M. acuminada*, dicotiledônea e monocotiledônea, respectivamente, que divergiram há 150 milhões de anos –, bem como incongruência filogenética envolvendo esses táxons – sequências *Copia25* de *Musa* se agrupam no clado de Rubiaceae, como grupo irmão de sequência de *Ixora* com alto suporte – juntas essas ocorrências suportam a inferência de HT proposta. Adicionalmente, as linhagens ancestrais dos gêneros envolvidos, *Musa* e *Ixora*, compartilharam a região geográfica do sudeste asiático entre 30 e 50 milhões de anos (LIU et al., 2010; CHRISTELOVÁ et al., 2011; LORENCE et al., 2007; TOSH et al., 2013), o que reforça a proposição. Uma ocorrência semelhante foi reportada recentemente envolvendo a transferência de um LTR-RT entre uma espécie de palmeira e uma de uva em uma análise ampla de 40 genomas de plantas (EL BALDOURI et al., 2014). Nesse estudo, 32 eventos de HT foram propostos, um envolvendo espécies de classes distintas, monocotiledôneas e dicotiledôneas, mas a maioria envolveu espécies de categorias taxonômicas inferiores, oito entre diferentes ordens e 22 entre gêneros da mesma família. O retrotransposon *Rider*, que apresenta similaridade com o *Copia25*, embora não pertença a mesma família, teria-se originado no genoma de *Solanum lycopersycum* advindo de *Arabidopsis thaliana* (CHENG et al., 2009) via HT.

Além do evento de HT do LTR-RT *Copia25* sugerido entre *Ixora* e *Musa*, esse elemento mostrou alta identidade nucleotídica com espécies distantemente relacionadas, *Elais guinensis*, 75%, uma monocotiledônea, e *Ricinus communis*, 79%, uma dicotiledônea pertencente à subclasse Rosidae. O posicionamento basal na filogenia da família *Copia25* descarta a hipótese de HT, permanecendo a inesperada identidade compartilhada entre espécies que divergiram há cerca de 150 milhões de anos (dicotiledôneas e monocotiledôneas) e 95 milhões de anos (Asteridae e Rosidae) (WANG et al., 2009). Outro retrotransposon, *Tvv1*, também apresentou alta similaridade envolvendo espécies dessas duas subclasses. *Tvv1* foi identificado em *Vitis vinífera* (Rosidae) e apresentou alta identidade com sequências oriundas do gênero *Solanum* (Asteridae) (MOISY et al., 2014). Ambas as ocorrências, HT e identidade nucleotídica alta, denotam a história evolutiva complexa da família *Copia25* e ressaltam sua origem e permanência em genomas de plantas.

Os elementos de transposição são os mais importantes agentes de remodelação genômica devido a sua transposição e por causarem significantes rearranjos cromossômicos, como translocações, inversões, deleções e duplicações (ZHANG et al., 2009; HEDGES; DEININGER, 2007). Além da contribuição para o entendimento da evolução dos TEs nos genomas de plantas, os resultados obtidos a partir da análise dos LTR-RTs fornecem informações acerca dessas sequências em um genoma novo. As reorganizações genômicas decorrentes da porção repetitiva do DNA, particularmente dos TEs, constituem fenômenos importantes em eventos de alopoliploidização. Essa reorganização influenciaria os genomas na promoção de rearranjos, levando a translocações, perdas e duplicações, que, por sua vez, poderiam alterar o contexto epigenético do genoma hospedeiro (TEIXEIRA et al., 2009; PARISOD; SENERCHIA, 2012). Amplificação e perdas de famílias de TEs têm sido observadas durante essa reorganização pós alopoliploidização (PARISOD; SENERCHIA, 2012). Os resultados encontrados na análise de 10 LTR-RTs nos parentais e no híbrido *C. arabica* sugerem a ocorrência de reorganização e perda de TEs no alotetraploide. Análises semelhantes em outros alopoliploides mostram uma amplificação limitada, restrita a algumas famílias de TEs, prevalecendo perdas dessas sequências repetitivas (BAUMEL et al., 2002; PETIT et al., 2010; PARISOD et al., 2009; 2012). Nenhuma das famílias, aqui, analisadas, apresentou amplificação no híbrido, sugerindo que as alterações epigenéticas, decorridas da incompatibilidade dos parentais e da reorganização genômica, não teriam reativado essas famílias. Em geral, o controle dos TEs no híbrido ocorre por eventos epigenético tanto pre-transcricionalmente, por alteração de estados de metilação, quanto pós transcricionalmente, por RNAi (GRANDBASTIEN et al., 2012). Diversos casos reportam a ocorrência de alteração no estado de metilação, que, de modo geral, torna-se demetilado, no alopoliploide nas primeiras gerações (BEAULIEU et al., 2009; PARISOD et al., 2009; KASHKUSH et al., 2002), embora, em alguns organismos, esse estado mude rapidamente para um estado hipermetilado (KRAITSHTEIN et al., 2010) e, em nem todos, observa-se a ocorrência de bursts propriamente de transposição, e sim apenas a presença do transcrito (MADLUNG et al., 2005).

Ao contrário de aumento do número de cópias, para os 10 RTs analisados, os resultados mostram uma perda de inserções no alotetraploide, sendo mais significativa em cinco das dez famílias investigadas, e sugerem, ainda, a ocorrência de alterações direcionais nos subgenomas, sendo que inserções do subgenoma materno, oriundo de *C. eugenioides*, estariam mais frequentemente envolvidas nessa reorganização do que as do subgenoma paterno, de *C. canephora*. A alopoliploidização pode ser caracterizada por eventos que ocorrem antes e após a hibridização, onde a partir de espécies divergentes, tem-se a hibridização e a duplicação do genoma total, seguidas por duas fases distintas. As gerações iniciais caracterizam-se por grande instabilidade genômica que envolve reorganização estrutural e epigenética, enquanto que nas gerações subsequentes, o genoma evolui restaurando a diploidização genética, caracterizam-se pela ocorrência mutações de ponto e rearranjos genômicos típicos de genomas diploides (GRANDBASTIEN et al., 2012; CHANG et al., 2010). A reorganização das gerações iniciais, também referida como mudanças revolucionárias, estaria relacionada à distância evolutiva das espécies que se hibridizaram, sendo geralmente observadas em eventos envolvendo espécies proximamente aparentadas, como é o caso de *C. eugenioides* e *C. canephora* (LASHERMES et al., 1999).

Afora a reorganização genômica sofrida, os resultados obtidos permitem inferências no que concerne a evolução dessas das 10 famílias de LTR-RTs nas três espécies de *Coffea*. Populações de *C. canephora* do continente africano, seu local de origem, formam grupos geneticamente distintos que se segregam em análises filogenéticas utilizando dados de marcadores SSRs e RFLPs associada a sua distribuição geográfica (GOMEZ et al., 2009). Contudo, os dados aqui obtidos, a partir de sítios de inserções das famílias LTR-RTs analisadas, por PCoAs e network, formam, de modo geral, uma população homogênea envolvendo genótipos de ambos os progenitores, *C. canephora* e *C. eugenioides*, com os genótipos de *C. arabica* formando um grupo separado, para a maioria das famílias de TEs. Salvo casos de HT e de perdas estocásticas, TEs são transmitidos verticalmente sendo herdados do ancestral para as espécies derivadas, e permanecem ou não em um genoma devido a uma gama de fatores relacionados ao TE, ao hospedeiro e a interação TE-hospedeiro (LE ROUZIC et al., 2007). A tribo Coffeae as espécies estudadas pertencem divergiu em um tempo evolutivo recente, 15 milhões de anos (BREMER; ERICKSON, 2009), e as espécies progenitoras há apenas 4.2 milhões de anos (YU et al., 2011). Famílias de TEs ativas são mantidas seletivamente, e produzem em sua mobilização cópias muito similares a cópia mãe, formando uma metapopulação, cujos indivíduos são passiveis de serem reconhecidos mesmo após um longo tempo evolutivo, como as sequências de *Copia25* aqui estudado. A população homogênea formada entre os genótipos de *C. canephora* e *C. eugenioides* para a maioria das 10 famílias de LTR-RTs poderia ser resultado da recente divergência dessas espécies, cujas cópias remanesceriam do ancestral. A identificação de cópias com idade de inserção no genoma superior a divergência das espécies progenitoras, bem como intrincadas reticulações no network (sinal de relações antigas marcadas por homoplasia) reforçam essa sugestão. Essas populações de LTR-RTs, devido a recente divergência, não teriam coalescido formando populações isoladas.

Juntos, os resultados aqui obtidos contribuem para o entendimento da evolução dos LTR-RTs no genoma, da colonização de novos genomas por esses elementos, bem como, da sua dinâmica evolutiva em um genoma recém-originado. Uma vez que a conjugação de dois genomas diferentes, no mesmo organismo, como é o caso dos alopoliplóides, pode envolver adaptações significativas de todos os mecanismos de regulação, incluindo regulação epigenética dos TEs, o estudo da regulação epigenética desses 10 TEs nas espécies parentais, *C. canephora* e *C. eugenioides*, e no híbrido, *C. arabica*, constitui uma próxima etapa indispensável para compreender como os genomas híbridos evoluem.

6 CONCLUSÕES

6 **CONCLUSÕES**

O presente estudo permitiu estabelecer as seguintes conclusões:

1. O LTR-RT, *Copia25*, amplamente distribuído na família Rubiaceae (Asteridae), e presente em outras dicotiledôneas das famílias Solanaceae e Euphorbiaceae, em monocotiledôneas das famílias Arecaceae e Musaceae, apresenta uma história evolutiva complexa, caracterizada por transferência horizontal entre espécies distantemente relacionadas (dicotiledônea *Ixora* e monocotiledônea *Musa*), conservação de sequências e perdas estocásticas.
2. As análises de 10 LTR-RTs no alotetraploide *C. arabica* sugerem que esses elementos de transposição mediaram alterações genômicas ocorridas após o evento de hibridização, resultando, principalmente, em perdas dessas sequências no genoma híbrido.
3. Uma reorganização envolvendo preferencialmente o subgenoma materno também pôde ser sugerida devido ao menor compartilhamento de sítios insercionais entre *C. arabica* e *C. eugenioides*.
4. Nos progenitores, a população dos LTR-RTs forma uma população homogênea resultado da presença de inserções do ancestral compartilhada por ambas as espécies, e a não coalescência devido ao recente tempo de divergência. No híbrido, forma uma subpopulação isolada, resultado de reorganização genômica e perdas.

REFERÊNCIAS BIBLIOGRÁFICAS

REFERÊNCIAS BIBLIOGRÁFICAS

BAACK, E. J.; RIESEBERG, L. H. A genomic view of introgression and hybrid speciation. **Current Opinion in Genetics & Development**, v. 17, n. 6, p. 513-518, 2007.

BAUMEL, A.; AINOUCHE, M.; KALENDAR, R. Retrotransposons and genomic stability in populations of the young allopolyploid species Spartina anglica CE Hubbard (Poaceae). **Molecular Biology and Evolution**, v. 19, n. 8, p. 1218-1227, 2002.

BEAULIEU, J.; JEAN, M.; BELZILE, F. The allotetraploid Arabidopsis thaliana-Arabidopsis lyrata subsp. petraea as an alternative model system for the study of polyploidy in plants. **Molecular Genetics and Genomics**, v. 281, n. 4, p. 421-435, 2009.

BRANDES, A. et al. Comparative analysis of the chromosomal and genomic organization of Ty1-copia-like retrotransposons in pteridophytes, gymnosperms and angiosperms. **Plant molecular biology**, v. 33, n. 1, p. 11-21, 1997.

BREMER, B.; ERIKSSON, T. Time tree of Rubiaceae: Phylogeny and dating the family, subfamily, and tribes. **International Journal of Plant Sciences**, v. 170, n. 6, p.766-793, 2009.

BROOKFIELD, J. F.; BADGE, R. M. Population genetics models of transposable elements. In **Evolution and Impact of Transposable Elements,** p. 281-294, Springer Netherlands, 1997.

CAPY, P.; ANXOLABÉHÈRE, D.; LANGIN, T. The strange phylogenies of transposable elements: are horizontal transfers the only explantation? **Trends in Genetics**, v. 10, n. 1, p. 7-12, 1994.

CENCI, A.; COMBES, M. C.; LASHERMES, P. Genome evolution in diploid and tetraploid Coffea species as revealed by comparative analysis of orthologous genome segments. **Plant Molecular Biology**, v. 78, n. 1-2, p. 135-45, 2012.

CHANG, P. L. et al. Homoeolog-specific retention and use in allotetraploid Arabidopsis suecica depends on parent of origin and network partners. **Genome Biology**, v. 11, n. 12, p. R125, 2010.

CHENG, X. et al. A New Family of Ty1-copia-Like Retrotransposons Originated in the Tomato Genome by a Recent Horizontal Transfer Event. **Genetics**, v. 181, n. 4, p. 1183-1193, 2009.

CHRISTELOVA, P. et al. A multi gene sequence-based phylogeny of the Musaceae (banana) family. **Bmc Evolutionary Biology**, v. 11, p. 103, 2011.

COMAI, L. Genetic and epigenetic interactions in allopolyploid plants. **Plant Molecular Biology**. v. 43, n. 2-3, p. 387-399, 2000.

COMAI, L. The advantages and disadvantages of being polyploid. **Nature Reviews Genetics**, v. 6, n. 11, p. 836-46, 2005.

CROS, J. et al. Phylogenetic analysis of chloroplast DNA variation in Coffea L. **Molecular Phylogenetics and Evolution**, v. 9, n. 1, p.109-117, 1998.

CUMMINGS, M. P. Transmission patterns of eukaryotic transposable elements: arguments for and against horizontal transfer. **Trends in Ecology and Evolution**, v. 9, n. 4, p. 141-5, 1994.

CUOMO, C. A. et al. The Fusarium graminearum genome reveals a link between localized polymorphism and pathogen specialization. **Science**, v. 317, p.1400-2, 2007.

DENOEUD, F. et al. The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. **Science**, v. 345, n. 6201, p. 1181-4, 2014.

DEVOS, K. M.; BROWN, J. K. M.; BENNETZEN, J. L. Genome Size Reduction through Illegitimate Recombination Counteracts Genome Expansion in Arabidopsis. **Genome Research**, v. 12, n. 7, p.1075-9, 2002.

DIAO, X. M.; FREELING, M.; LISCH, D. Horizontal transfer of a plant transposon. **Plos Biology**, v. 4, n. 1, p. 119-128, 2006.

EL BAIDOURI, M. et al. Widespread and frequent horizontal transfers of transposable elements in plants. **Genome Research**, v. 24, n. 5, p. 831-8, 2014.

FELDMAN, M.; LEVY, A. A. Genome evolution in allopolyploid wheat - a revolutionary reprogramming followed by gradual changes. **Journal of Genetics and Genomics**, v. 36, n. 9, p. 511-8, 2009.

FESCHOTTE, C.; ZHANG, X.; WESSLER, R. Miniature inverted-repeat transposable elements and their relationships to established DNA transposons. In: NL, C.;R, C., et al (Ed.). **Mobile DNA II**. Washington, DC: ASM Press, 2002.p.1093-1110.

FORTUNE, P. M.; ROULIN, A.; PANAUD, O. Horizontal transfer of transposable elements in plants. **Communicative & Integrative Biology**, v. 1, n. 1, p. 74-7, 2008.

GAETA, R. T. et al. Genomic changes in resynthesized Brassica napus and their effect on gene expression and phenotype. **Plant Cell**, v. 19, n. 11, p. 3403-3417, 2007.

GOMEZ, C. et al. 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 Evolutionary Biology**, v. 9, p. 167, 2009.

GOTEA, V.; MAKALOWSKI, W. Do transposable elements really contribute to proteomes? **Trends in Genetics**, v. 22, n. 5, p. 260-267, 2006.

HANSKI, I. Metapopulation dynamics. **Nature**, v. 396, p. 41-49, 1998.

HEDGES, D.J.; DEININGER, P.L. Inviting instability: Transposable elements, double-strand breaks, and the maintenance of genome integrity. **Mutation Research**, v. 616, n. 1-2, p. 46-59, 2007.

HODSON, M. J.; BRYANT, J. A. **Functional Biology of Plants.** By. Chichester, UK: Wiley-Blackwell, pp. 326, 2012. ISBN 978-0-470-69939-3.

JACKSON, S.; CHEN, Z. J. Genomic and expression plasticity of polyploidy. **Current opinion in plant biology**, v. 13, n. 2, p. 153-9, 2010.

JIAO, Y. et al. Ancestral polyploidy in seed plants and angiosperms. **Nature**, v. 473, n. 7345, p. 97-100, 2011.

JIN, Y-K.; BENNETZEN, J. L. Structure and coding properties of Bs1, a maize retrovirus-like transposon. **Proc Natl Acad Sci U S A,** v. 86, p. 6235-39, 1989.

JORDAN, I. K.; MCDONALD, J. F. Evolution of the copia retrotransposon in the Drosophila melanogaster species subgroup. **Molecular Biology and Evolution**, v. 15, n. 9, p. 1160-1171, 1998.

JOSEFSSON, C.; DILKES, B.; COMAI, L. Parent-dependent loss of gene silencing during interspecies hybridization. **Current Biology**, v. 16, n. 13, p. 1322-8, 2006.

KALENDAR, R. et al. Cassandra retrotransposons carry independently transcribed 5S RNA. **Proc Natl Acad Sci U S A,** v. 105, n. 15, p. 5833-8, 2008.

KALENDAR, R. et al. Large retrotransposon derivatives: abundant, conserved but non autonomous retroelements of barley and related genomes. **Genetics**, v. 166, n. 3, p. 1437-50, 2004.

KASHKUSH, K.; FELDMAN, M.; LEVY, A. A. Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat. **Nature genetics**, v. 22, p. 102-106, 2003.

KIDWELL, M. G.; LISCH, D. R. Perspective: Transposable elements, parasitic DNA, and genome evolution. **Evolution**, v. 55, n. 1, p. 1-24, 2001.

KIDWELL, M. G.; LISCH, D. R. Transposable elements and host genome evolution. **Trends in Ecology & Evolution**, v. 15, n. 3, p. 95-99, 2000.

KRAITSHTEIN, Z. et al. Genetic and epigenetic dynamics of a retrotransposon after allopolyploidization of wheat. **Genetics**, v. 186, p. 801-12, 2010.

KRAMEROV, D. A.; VASSETZKY, N. S. Short retroposons in eukaryotic genomes. **International Review of Cytology**, v. 247, p. 165-221, 2005.

KUMAR, A. The adventures of the Ty1- copia group of retrotransposons in plants. **Trends In Genetics**, v. 12, n. 2, p. 41-43, 1996.

KUMAR, A.; BENNETZEN, J. PLANT RETROTRANSPOSONS. **Genetics**, v. 33, p. 479-532, 2003.

KUMAR, A. et al. The Ty1-copia group of retrotransposons in plants: genomic organisation, evolution, and use as molecular markers. **Genetica**, v. 100, n. 1-3, p. 205-217, 1997.

KUMEKAWA, N. et al. Identification and characterization of novel retrotransposons of the gypsy type in rice. **Molecular and General Genetics**, v. 260, p. 593-602, 1999.

LAMB, J.C. et al. Plant chromosomes from end to end: telomeres, heterochromatin and centromeres. **Current Opinion in Plant Biology**, v. 10, n. 2, p. 116-22, 2007.

HODSON, M. J.; BRYANT, J. A. **Functional Biology of Plants**. Wiley-Blackwell, 2012. 334 ISBN 9780470699393.

LASHERMES, P.et al. Molecular characterisation and origin of the Coffea arabica L. genome. **Molecular and General Genetics**, v. 261, n. 2, p. 259-266, 1999.

LE ROUZIC, A.; BOUTIN, T. S.; CAPY, P. Long-term evolution of transposable elements. **Proc. Natl. Acad. Sci. U.S.A.**, v. 104, p. 19375-80, 2007.

LIU, A.; KRESS, W.; LI, D. Phylogenetic analyses of the banana family (Musaceae) based on nuclear ribosomal (ITS) and chloroplast (trnL-F) evidence. **Taxon**, v. 59, n. 1, p. 20-28, 2010.

LIU, B., et al. Polyploid formation in cotton is not accompanied by rapid genomic changes. **Genome**, v. 44, n. 3, p. 321-30, 2001.

LOHE, A. R. et al. Horizontal transmission, vertical inactivation, and stochastic loss of mariner-like transposable elements. **Molecular Biology and Evolution**, v. 12, n. 1, p. 62-72, 1995.

LOPES, F. et al. Transposable elements in Coffea (Gentianales: Rubiacea) transcripts and their role in the origin of protein diversity in flowering plants. **Molecular Genetics and Genomics**, v. 279, n. 4, p. 385-401, 2008.

LOPES, F. R. et al. Transcriptional activity, chromosomal distribution and expression effects of transposable elements in Coffea genomes. **PLoS One**, v. 8, n. 11, p. e78931, 2013.

LORENCE, D. et al. Revision of Ixora (Rubiaceae) in the Marquesas Islands (French Polynesia). **Botanical Journal of The Linnean Society**, v. 155, n. 4, p. 581-597, 2007.

LORETO, E. L. S.; CARARETO, C. M. A.; CAPY, P. Revisiting horizontal transfer of transposable elements in Drosophila. **Heredity**, v. 100, n. 6, p. 545-554, 2008.

MADLUNG, A. et al. Genomic changes in synthetic Arabidopsis polyploids. **Plant Journal**, v. 41, n. 2, p. 221-230, 2005.

MAKALOWSKI, W.; TODA, Y. Modulation of host genes by mammalian transposable elements. **Genome dynamics**, v. 3, p. 163-74, 2007.

MARUYAMA, K.; HARTL, D. L. Evolution of the transposable element mariner in Drosophila species. **Genetics**, v. 128, n. 3, p. 19-329, 1991.

MASTERSON, J. Stomatal size in fossil plants: evidence for polyploidy in majority of angiosperms. **Science**, v. 264, n. 5157, p. 421-4, 1994.

MAYROSE, I. et al. Recently formed polyploid plants diversify at lower rates. **Science**, v. 2, n. 333, p.1257, 2011.

MCCLINTOCK, B. The significance of responses of the genome to challenge. **Science**, v. 226, n. 4676, p. 792-801, 1984.

MOISY, C. et al. The Tvv1 retrotransposon family is conserved between plant genomes separated by over 100 million years. **Theoretical and Applied Genetics**, v. 127, n. 5, p. 1223-35, 2014.

OLIVER, K. R.; MCCOMB, J. A.; GREENE, W. K. Transposable elements: powerful contributors to angiosperm evolution and diversity. **Genome Biology and Evolution**, v. 5, n. 10, p. 1886-901, 2013.

PARISOD, C. et al. Rapid structural and epigenetic reorganization near transposable elements in hybrid and allopolyploid genomes in Spartina. **New Phytologist**, v. 184, p. 1003-15, 2009.

PARISOD, C.; SENERCHIA, N. Responses of transposable elements to polyploidy. In: **Plant Transposable Elements** (Eds. Grandbastien & Casacuberta), Topics in Current Genetics, Springer, v. 24, p. 147-168, 2012.

PEARCE, S. et al. The Ty1-copia group retrotransposons in *Vicia* species: Copy number, sequence heterogeneity and chromosomal localization**. Molecular and General Genetics**, v. 250, n. 3, p. 305-315, 1996.

PETIT, M. et al. Mobilization of retrotransposons in synthetic allotetraploid tobacco. **New Phytologist**, v. 186, n. 1, p.135-47, 2010.

PINSKER, W. et al. The evolutionary life history of P transposons: From horizontal invaders to domesticated neogenes. **Chromosoma**, v. 110, n. 3, p. 148-158, 2001.

ROULIN, A. et al. Evidence of multiple horizontal transfers of the long terminal repeat retrotransposon RIRE1 within the genus Oryza. **Plant Journal**, v. 53, n. 6, p. 950-959, 2008.

SABOT, F.; SCHULMAN, A. H. Parasitism and the retrotransposon life cycle in plants: a hitchhiker's guide to the genome. **Heredity**, v. 97, n. 6, p. 381-8, 2006.

SANMIGUEL, P. et al. The paleontology of intergene retrotransposons of maize. **Nature Genetics**, v. 20, n. 1, p. 43-45, 1998.

SANMIGUEL, P.; BENNETZEN, J. L. Evidence that a recent increase in maize genome size was caused by the massive amplification of intergene retrotransposons. **Annals of Botany**, v. 82, p. 37-44, 1998.

SCHAACK, S.; GILBERT, C.; FESCHOTTE, C. Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. **Trends in Ecology & Evolution**, v. 25, n. 9, p. 537-546, 2010.

SHAKED, H. et al. Sequence elimination and cytosine methylation are rapid and reproducible responses of the genome to wide hybridization and allopolyploidy in wheat. **Plant Cell**, v. 13, n. 8, p. 1749-1759, 2001.

SOLTIS, P. S.; SOLTIS, D. E. The role of genetic and genomic attributes in the success of polyploids. **Proc Natl Acad Sci U S A,** v. 97, n. 13, p. 7051-7, 2000.

SOLTIS, P. S.; SOLTIS, D. E. The role of hybridization in plant speciation. **Annual Review of Plant Biology**, v. 60, p. 561-88, 2009.

STEBBIS, G. L. Types of polyploids; their classification and significance. *Advances in* **Genetics**, v. 1, p. 403-29, 1947.

TATE, J. A. et al. Evolution and expression of homeologous loci in Tragopogon miscellus (Asteraceae), a recent and reciprocally formed allopolyploid. **Genetics**, v. 173, n. 3, p. 1599-1611, 2006.

TEIXEIRA, F. K. et al. A role for RNAi in the selective correction of DNA methylation defects. **Science**, v. 323, n. 5921, p. 1600-4, 2009.

TENAILLON, M. I. et al. Genome size and transposable element content as determined by high-throughput sequencing in maize and Zea luxurians. **Genome Biology and Evolution**, v. 3, p. 219-29, 2011.

TESFAYE, K. et al. Characterization of Coffea chloroplast microsatellites and evidence for the recent divergence of C. arabica and C. eugenioides chloroplast genomes. **Genome**, v. 50, n. 12, p. 1112-29, 2007.

TOSH, J. et al. Evolutionary history of the Afro-Madagascan Ixora species (Rubiaceae): species diversification and distribution of key morphological traits inferred from dated molecular phylogenetic trees. **Annals of Botany**, v. 112, n. 9, p. 1723-1742, 2013.

TOUCHON, M.; ROCHA, E. P. Causes of insertion sequences abundance in prokaryotic genomes. **Molecular Biology and Evolution**, v. 24, p. 969-81, 2007.

WALLAU, G. L.; ORTIZ, M. F.; LORETO, E. L. Horizontal transposon transfer in eukarya: detection, bias, and perspectives**. Genome Biology and Evolution**, v. 4, n. 8, p. 689-99, 2012.

WANG, H. et al. Rosid radiation and the rapid rise of angiosperm-dominated forests. **Proc Natl Acad Sci U S A,** v. 106, n. 10, p. 3853-8, 2009.

WENDEL, J. F. Genome evolution in polyploids. **Plant Molecular Biology**, v. 42, n. 1, p. 225-49, 2000.

WICKER, T. et al. A unified classification system for eukaryotic transposable elements. **Nature Reviews Genetics**, v. 8, p. 973-982, 2007.

WICKER, T. So Many Repeats and So Little Time: How to Classify Transposable Elements. In **Plant Transposable Elements** - Impact on Genome Structure and Function. Editors:

Grandbastien, Marie-Angèle, Casacuberta, Josep (Eds.), Topics in Current Genetics 2012.

WIKSTRÖM, N.; SAVOLAINEN, V.; CHASE, M.W. Evolution of the angiosperms: calibrating the family tree. **Proceedings of the Royal Society**, v. 268, n. 1482, p. 2211-2220, 2001.

WITTE, C. P. et al. Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes. **Proc Natl Acad Sci U S A,** v. 98, n. 24, p. 13778-83, 2001.

WONG, L.H.; CHOO, K. H. Evolutionary dynamics of transposable elements at the centromere. **Trends In Genetics**, v. 20, n. 12, p. 611-6, 2004.

WOOD, T. E. et al. The frequency of polyploid speciation in vascular plants. **Proc Natl Acad Sci U S A,** v. 106, n. 33, p. 13875-9, 2009.

YU, Q. et al. Micro-collinearity and genome evolution in the vicinity of an ethylene receptor gene of cultivated diploid and allotetraploid coffee species (Coffea). **Plant Journal**, v. 67, n. 2, p. 305-17, 2011.

ZHANG, J. et al. Alternative Ac/Ds transposition induces major chromosomal rearrangements in maize. **Genes Development**, v. 23, n. 6, p. 755-765, 2009.