UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Câmpus de São José do Rio Preto
Programa de Pós-graduação em Genética

FABRÍCIO RAMON LOPES

ELEMENTOS DE TRANSPOSIÇÃO COMO FONTE DE NOVIDADES GENÉTICAS EM NÍVEL TRANSCRICIONAL: UMA ABORDAGEM COMPUTACIONAL E MOLECULAR

Tese apresentada para obtenção do Título de Doutor em Genética.

São José do Rio Preto – SP

2011

UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Câmpus de São José do Rio Preto
Programa de Pós-graduação em Genética

FABRÍCIO RAMON LOPES

ELEMENTOS DE TRANSPOSIÇÃO COMO FONTE DE NOVIDADES GENÉTICAS EM NÍVEL TRANSCRICIONAL: UMA ABORDAGEM COMPUTACIONAL E MOLECULAR

Tese apresentada para obtenção do Título de Doutor em Genética.

Orientadora: Profa. Dra. Claudia Marcia Aparecida Carareto

São José do Rio Preto – SP

2011

Lopes, Fabrício Ramon.
Elementos de transposição como fonte de novidades genéticas em nível transcricional: uma abordagem computacional e molecular / Fabrício Ramon Lopes. – São José do Rio Preto : [s.n.], 2011.
151 f. : il. ; 30 cm.

Orientador: Claudia Marcia Aparecida Carareto
Tese (doutorado)-Universidade Estadual Paulista, Instituto de Biociências, Letras e Ciências Exatas

1. Genética. 2. Genômica. 3. Evolução molecular. 4. *Coffea spp*. I. Carareto, Claudia Marcia Aparecida. II. Universidade Estadual Paulista, Instituto de Biociências, Letras e Ciências Exatas. III. Título.

CDU – 575.852

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

FABRÍCIO RAMON LOPES

Elementos de transposição como fonte de novidades genéticas em nível transcricional: uma abordagem computacional e molecular

Tese apresentada 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.

BANCA EXAMINADORA

Prof[a]. Dr[a]. Claudia Marcia Aparecida Carareto
Professora Titular
UNESP – São José do Rio Preto
Orientadora

Prof[a]. Dr[a]. Marie-Anne Van Sluys
Professora Titular
Universidade de São Paulo

Prof. Dr. Elgion Lucio da Silva Loreto
Professor Associado
Universidade Federal de Santa Maria

Prof. Dr. Ivan de Godoy Maia
Professor Adjunto
UNESP – Botucatu

Prof[a]. Dr[a]. Maria Elisabete Jorge Amaral
Professora Adjunto
UNESP – São José do Rio Preto

São José do Rio Preto, 25 de Fevereiro de 2011.

O presente trabalho foi desenvolvido no Departamento de Biologia, da Universidade Estadual Paulista “Júlio de Mesquita Filho” – UNESP, de São José do Rio Preto, Estado de São Paulo, sob a orientação da Profa. Dra. Claudia Marcia Aparecida Carareto, com Bolsa de Estudo e Auxílio à Pesquisa da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP).

Dedico este trabalho científico a três pessoas...

À minha mãe, ***Liliana Lopes****,*

mulher guerreira,

a completa personificação do entusiasmo ao trabalho e à família!

Ao meu irmão, ***Fernando C. Lopes Filho****,*

por me apresentar ao mundo "Linux",

no início, um sistema desconhecido, agora, uma ferramenta poderosa!

À minha orientadora científica, ***Claudia M. A. Carareto****,*

"Por tantos anos de convivência, trabalhos realizados e formação".

Agradecimentos

Há poucos anos tive o prazer de acompanhar a defesa de inúmeros pós-graduandos e amigos. Agora, eu mesmo acabo de me apresentar para uma Defesa de Tese, uma oportunidade inesperada há alguns anos. Este trabalho científico é resultado da relação de co-dependência com diversos pesquisadores, além do bate-papo, do incentivo e da camaradagem de inúmeras pessoas. Diante disso, eu tentarei agradecer a todos nominalmente.

À minha orientadora Profa. Dra. **Claudia Marcia Aparecida Carareto** por tantos anos de convivência [agradável e profícua], pelos ensinamentos, experiências, oportunidades de formação e atuação. Você merece vida longa, felicidades e paz!

À minha mãe muito amada, **Liliana Lopes**. Por sua devoção aos descendentes, o meu amor do fundo do coração. Seus ensinamentos persistentes e repetitivos fazem me lembrar de maneira reveladora.

Aos colaboradores científicos, Dr. (s) **Alan Carvalho Andrade, Pierre Marracini** e **João Batista Teixeira** (EMBRAPA, Brasília - DF), **King I. Jordan** e **Daudi Jjingo** (Instituto de Tecnologia da Geórgia, Atlanta, EUA), **Carlos Augusto Colombo** (IAC, Campinas - SP), **Gonçalo A. G. Pereira** e **Marcelo Falsarella Carazzolle** (Univ. Estadual de Campinas, Campinas - SP), **André Luis Laforga Vanzela** e **Carlos R. Maximiano da Silva** (Univ. Estadual de Londrina, Londrina - PR), **Luiz Filipe P. Pereira** (IAPAR, Londrina - PR) pela oportunidade de desenvolver partes desta Tese.

À Profa. Dra. **Marie-Anne Van Sluys**, Prof. Dr. **Elgion Lucio da Silva Loreto**, Prof. Dr. **Ivan de Godoy Maia** e Profa. Dra. **Maria Elizabete Jorge Amaral** por fazerem parte da correção e avaliação desta Tese e pela oportunidade de compartilhar um momento tão significativo na minha vida. Muito obrigado!

À (s) Dra. (s) **Marlene Benchimol** (Univ. Santa Úrsula, Rio de Janeiro - RJ), **Joana C. Silva** (Instituto de Ciências Genômicas e Faculdade de Medicina, Univ. de Maryland, Maryland, EUA), **Cristina Vieira** e **Emmanuelle Lerat** (Univ. de Lyon I, Lyon, França), **André Luis Laforga Vanzela** (Univ. Estadual de Londrina, Londrina - PR) por mostrarem-me que a formação de um Doutor pode ir além da Tese.

À Comunidade da EMBRAPA Recursos Genéticos e Biotecnologia (Brasília - DF). À (s) Dra. (s) **Ana Cristina M. Brasileiro** e **Angela Mehta** pelo empréstimo de alíquotas de marcador radioativo e a Dra. **Leila M. G. Barros** e a Ms. **Norma S. Paes** por me acompanharem em diversas oportunidades no Laboratório de Radioatividade. Aos Dr. (s) **Alan C. Andrade**, **Pierre Marraccini**, **Carlos Bloch Junior**, **Francisco J. L. Aragão**, **Luciano Paulino da Silva** e **Hugo Bruno C. Molinari** pelos bate-papos altamente *lucrativos*. Ao analista **Mário A. P. Saraiva** pela construção dos macroarranjos e **Felipe Vinecky** pela manipulação das bibliotecas de cDNA de *Coffea spp.* Aos amigos, **Aline**, **André**, **Barbara**, **Betúlia**, **Daniel**, **Eder**, **Felipe**, **Flávio**, **Gabriel**, **Ingrid**, **Luciana**, **Marryelle**, **Maria Thereza**, **Mariana**, **Natália**, **Rebeca, Rodrigo** pela divertida convivência, pelos passeios [sensacionais] em Brasília...

Gostaria de render homenagem aos meus professores do Programa de Pós-graduação em Genética – UNESP/SJRP. Aos meus professores e amigos Dr. **Carlos Roberto Ceron** e Dra. **Paula Rahal**, que sempre falam bem ao meu respeito. À (s) Profa. (s) Dra. (s) **Ana Elisabete Silva, M. Elizabete J. Amaral** e **Sônia Maria Oliani**, outras grandes professoras e amigas. À Profa. Emérita **Hermione E. M. C. Bicudo** por sua vitalidade intelectual, dedicação e elegância. À (s) Profa. (s) Dra. (s) **Claudia R. Bonini Domingos, Lilian Madi Ravazzi** e **Maria Tercília V. A. Oliveira** pela atuação como Coordenadoras do Programa de Pós-graduação.

Aos queridos amigos de laboratório **Adriana Granzotto**, **Ana Luiza N. Rauber**, **Elaine S. Dias**, **Elias A. G. Carnelossi**, **Leliane S. Commar**, **Lilian Ricco Medeiros** e **Marcela A. Proença** (UNESP, SJRP - SP) e, aos que partiram **Julcimary Ricci, Luis Gustavo Galego** e **Nathalia de Setta Costa** pelos constantes bate-papos com e sem *fins lucrativos* e por compreenderem minha quietude.

Ao meu amigo Prof. Dr. **Luis Gustavo Galego** pela correção dos meus *Agradecimentos*, e homenageá-lo pela disposição e ânimo constantes.

Às minhas amigas **Cíntia Bittar** e **Ana Carolina Bonfim** (UNESP/SJRP - SP), **Agda Facincani** (UNESP/Jaboticabal - SP) e **Luciana Pereira Freire** (EMBRAPA/Brasília - DF) pelo seqüenciamento de partes dos clones de cDNA analisados nesta Tese e **Elias Alberto Gutierrez Carnelossi** pelo auxílio e atenção como bolsista de Treinamento Técnico da FAPESP. Muito obrigado!

Aos meus amigos de Pós-graduação, **Ana Carolina Bonini**, **Ana Carolina Jardim, Caroline Bonfim, Cíntia, Édis**, **Érica**, **Fernanda Manoel, Hector, Hederson, Jucimara**, **Larissa Venâncio, Larissa Fornitano, Lilian, Luciana, Márcia, Marília**, **Mariana, Marina**, **Mateus, Maurício, Melissa, Nedênia, Paola**, **Paula, Tadaiti**, **Sandra, Wanessa** e **Yvana** desejo felicidades e sucesso nessa longa jornada de formação, trabalho, emprego, família...

Ao meu irmão **Fernando C. Lopes Filho** por lutarmos juntos em nome de uma família melhor. Ao meu avô **Geraldo Lopes** por insistir que nossas metas e ações deveriam brotar de uma única "Motivação ELEvada"**.** À minha avó **Dolores de Freitas Lopes** por sempre me receber com um sorriso no rosto, um abraço apertado e um bolo pronto no forno.

À cidade de **São José do Rio Preto (SP)**, onde guardo amigos, onde fui criado e onde descobri meu papel.

À FAPESP pelo suporte.

A **DEUS** por me salvaguardar!

RESUMO

Elementos de transposição (TEs) são entidades genéticas que podem ter profundos impactos, estrutural, funcional, intra e interespecíficos na evolução dos genomas. A contribuição dos TEs para formação de novas seqüências codificadoras de proteínas é de particular interesse porque sua inserção em exons pode alterar a seqüência protéica influenciando diretamente o fenótipo. Além disso, o estudo das condições que provocam a ativação de TEs, e os mecanismos que os regulam, justifica o interesse de se identificar TEs expressos em tecidos ou condições específicas. Este estudo foca tais questões usando um modelo vegetal: *C. arabica*, única espécie híbrida e poliplóide do seu gênero, derivada de uma hibridização recente e natural entre *C. canephora* e *C. eugenioides*. As análises foram realizadas por meio de uma variedade de abordagens: 1) análises computacionais usando uma combinação de RepeatMasker, tBLASTx e diversas bibliotecas de TEs referência estocadas no Repbase; 2) análises de expressão baseadas em macroarranjos de DNA; e 3) avaliação do número de cópias e distribuição cromossomal de TEs ativos por Hibridização *in situ* fluorescente (*FISH*). Foram identificados 180 unigenes com fragmentos de TEs nas três espécies de café. Em uma primeira análise, com base em unigenes selecionados, foi possível sugerir 26 putativas proteínas com inserção de cassetes de TEs, demonstrando uma provável contribuição para a variabilidade do repertório protéico hospedeiro. Por outro lado, 327 ESTs similares a TEs expressos foram identificadas com uma possível abundância diferencial para duas famílias de *Ty3/Gypsy* (*dea1* e *Retrosat*) identificadas em apenas duas bibliotecas de *C. canephora* (sementes e pericarpo). Análises de expressão mostraram que muitos dos mRNAs quiméricos e TEs ativos apresentam baixa expressão. Por outro lado, vários desses transcritos tiveram sua expressão restabelecida em culturas de células tratadas com Cicloheximida, uma droga que permite o acúmulo de transcritos previamente silenciados por reverter uma mecanismo responsável por degradar mRNA contendo códon de terminação prematuro, chamado de "Decaimento de mRNA sem sentido" (NMD). Essa baixa expressão poderia ser explicada, para alguns casos pela atuação do mecanismo NMD, mas também não pode ser descartada a possibilidade de que alguns desses locos expressos tenham localização heterocromática. Análise de *FISH* com sondas de três TEs ativos com baixa expressão em *C. arabica* mostraram que muitas das inserções estão localizadas em blocos cromossômicos terminais e pericentrômeros, sugerindo que esta baixa expressão esteja relacionada com seu perfil de distribuição heterocromática. A aplicação de abordagens computacionais e moleculares permitiu avaliar a extensão com que os TEs contribuem e impactam a variabilidade transcricional em genomas de *Coffea spp*.

Palavras-chave: variabilidade transcricional; *Coffea spp*; elementos de transposição; macroarranjos; *FISH*.

ABSTRACT

Transposable elements (TEs) are genetic entities that can have profound structural, intra, interspecific impacts in evolution of the genomes. The contribution of the TEs in the protein coding region is of particular interest because insertions of TE into exons can alter the protein sequence influencing directly the phenotype. Moreover, the analysis of the conditions that cause the TE activation and the mechanisms that regulate them justify the interest of identifying expressed TEs in tissues or specific conditions. This study focus such subjects using the plant model: *C. arabica*, unique hybrid species and polyploidy of their genus, derived of a recent and natural hybridization between *C. canephora* and *C. eugenioides*. The analyses were performed by a variety of approaches: 1) computational analyses using a combination of RepeatMasker, tBLASTx e several libraries of reference TEs stored in Repbase; 2) expression analysis based in macroarrays; and 3) evaluation of copy number e chromosomal distribution of expressed TEs by Fluorescent *in situ* hybridization (*FISH*). 180 unigenes containing TE fragments were identified in the three *Coffea* species. Based in selected unigenes, it was possible to identify 26 putative proteins harboring TE-cassettes, demonstrating a probable contribution for the host protein repertory variability. In addition, 327 ESTs similar to expressed TEs were identified with a probable differential abundance for two families of *Ty3/Gypsy* (*dea1* and *Retrosat*) identified only in two cDNA libraries from *C. canephora* (seeds and pericarp). Our expression analyses showed that most of the chimerical mRNAs and expressed TEs has low or null expression. On the other hand, several transcripts had their expression reestablished in cell culture treated with Cycloheximide, a drug that permit the accumulation of transcripts previously silenced by reverting a mechanisms that degrade mRNAs containing premature terminate codons, named “nonsense mediated mRNA decay” (NMD). This low expression profile could be explained, for some cases, by action of the NMD mechanism, but also cannot to be ignored that some of these expressed locus have a heterochromatic localization. FISH analyses using probes of three expressed TEs with low expression profile in *C. arabica* showed that most of the hybridization signals are located in terminal chromosomal blocks and pericentromere, suggesting that this low expression is related with their heterochromatic distribution pattern. The application of computational and molecular approaches allowed evaluate the extent that the TEs contribute and impact the transcriptional variability in *Coffea spp* genomes.

Keywords: transcriptional variability; *Coffea spp*; transposable elements; macroarrays; *FISH*.

SUMÁRIO

I. INTRODUÇÃO

I. INTRODUÇÃO GERAL

Elementos de transposição (TEs) é um tipo de seqüência repetitiva que apresenta capacidade intrínseca de mobilização. Dados acumulados na literatura permitem elencar um amplo espectro de informações sobre a função desses elementos genéticos móveis no genoma, embora muito de seu papel reste a ser desvendado mesmo após a publicação de importantes genomas e transcriptomas. Crescente esforço tem sido dirigido ao estudo dessas seqüências, uma vez que os projetos de seqüenciamento em larga escala, de várias espécies de eucariotos, revelam que uma grande fração do genoma é constituída por TEs (KIDWELL 2002; SHAPIRO; VON STERNBERG, 2005). Por exemplo, 60 a 80% do genoma de milho (MEYERS et al., 2001), 45% do genoma humano (INTERNATIONAL HUMAN GENOME SEQUENCING CONSORTIUM, 2001), 39% do genoma de ratos (WATERSON et al., 2002) são constituídos por elementos de transposição.

Como conseqüência de sua prevalência e mobilidade, os TEs são mutagênicos e podem produzir profundas mudanças na organização do genoma e expressão e função de genes hospedeiros (KIDWELL; LISCH, 1997; JORDAN et al. 2003; VAN DE LAGEMAAT et al. 2003; THORNBURG et al., 2006); como por exemplo, podem levar a contração e expansão do genoma (FEDOROFF, 2000; BENNETZEN, 2002; CARLTON et al. 2007), transdução e amplificação de fragmentos gênicos hospedeiro (JIANG et al. 2004; KAPITONOV; JURKA, 2007; LOPES et al., 2008) e aumento da variabilidade do repertório transcricional e protéico (BROSIUS; GOULD, 1992; GERBER et al., 1997; MILLER et al. 1999; NEKRUTENKO; LI, 2001; HILGARD et al., 2002; HOENICKA et al., 2002; VOLFF, 2006; GOTEA; MAKALOWSKI, 2006; ALMEIDA et al., 2007; WU et al., 2007; LOPES et al., 2008). Dado esse enorme potencial como fonte de novidades genéticas, considerável esforço tem sido devotado para se caracterizar os diversos impactos dos TEs no genoma hospedeiro.

I.1 Impacto na função e expressão gênica hospedeira mediada por TEs

Regulação gênica é central para a relação genótipo-fenótipo em todos os organismos. As inserções de TEs em genes geram diferentes conseqüências funcionais e evolutivas para a espécie hospedeira. As sequências derivadas de TEs podem atuar como promotores quando inseridas na região 5' flanqueadora (BROSIUS, 1999), afetar a expressão gênica, levando ao seu aumento ou redução ou mudando os padrões de expressão temporal e/ou espacial ou, ainda, dar origem a novas combinações de genes gerados por mudanças no padrão de processamento do RNA mensageiro (VARAGONA et al., 1992; DAVIS et al., 1998; ZHENG et al., 2005; MEDSTRAND et al., 2005) e inserção de sítios de poliadenilação (MAGER et al., 1999). Adicionalmente, o uso de um sítio de *splicing* existente na sequência de um TE pode resultar na produção de uma nova proteína. Por exemplo, um alelo *waxy* mutado, em milho, mostrou uma expressão tecido-específica alterada resultando em uma atividade enzimática 30 vezes maior em pólen que no endosperma por um evento de *splicing* alternativo mediado pela inserção de um retrotransposon (VARAGONA et al., 1992; MARILLONNET; WESSLER, 1997). Além disso, elementos de transposição com baixo número de cópias provêem mecanismo para evolução de novos genes e fenótipos por adquirirem partes de outros genes e se transporem para novas localidades genômicas (JIN; BENNETZEN, 1994; ELROUBY; BUREAU, 2001). Por exemplo, a mutação *brown midrib*, em milho, é causada pela inserção de um retrotransposon na região codificante do gene COMT, que codifica para uma metiltransferase (VIGNOLS et al., 1995). Nos primatas antropóides, foi sugerido que um gene originado pela fusão de um gene *SET* histona metiltransferase com uma transposase *downstream*, nomeado *SETMAR*, forma novas redes regulatórias (CORDAUX et al., 2006).

Uma das discussões mais efervescentes trata-se dos efeitos da inserção de fragmentos ou cassetes de TEs dentro de seqüências de mRNAs gerados após a ativação de sítios crípticos de *splicing* em um TE residente em um íntron, ou *de novo* através de inserção dentro de exons (MITCHELL et al., 1991; MAKALOWSKI et al., 1994). A ocorrência e a abundância de transcritos hospedeiros, modificados pela inserção de

cassetes de TEs, têm sido descrita em diversos organismos, a saber: em humanos (NEKRUTENKO; LI, 2001), *Mus musculus* (DEBARRY et al., 2005), *Caenorhabditis elegans* (GANKO et al., 2003), *Drosophila melanogaster* (GANKO et al., 2006), *Bos taurus* (ALMEIDA et al., 2007), *Oryza sativa* (SAKAI et al., 2007; KROM et al., 2008), *Coffea spp* (LOPES et al., 2008).

A investigação de éxons contendo elementos *Alu* feita por Sorek et al. (2002) levou à interessante observação de que todos estes exons são alvo de eventos de *splicing* alternativo, dado pela presença de sítios doadores e aceptores de *splicing* nesses elementos, permitindo aos autores sugerirem um mecanismo para exonização e incorporação de TEs (LEV-MAOR et al., 2003). Mais recentemente, Wu e colaboradores (2007) revelaram por meio de análises filogenéticas que cassetes de TEs em transcritos humanos são mais divergentes em éxons constitutivos que em éxons alternativos, ao menos em média. Esta observação suporta a noção de exonização de TEs, pois fragmentos codificados por TEs são mais freqüentemente derivados da ativação de sítios crípticos de *splicing* localizados em uma seqüência de TE residente em íntrons que de inserções *de novo* em exons (WU et al., 2007). Entretanto, são ainda escassas as evidências que suportam a tradução destes cassetes em proteínas funcionais gerando polêmica quanto sua contribuição para o proteoma e, conseqüente impacto no fenótipo.

Por exemplo, um dos primeiros relatos sobre a contribuição de seqüências de TEs para o proteoma, o registro de Nekrutenko e Li (2001) de que 4% dos genes humanos analisados tinham CDSs (Seqüências Codificantes, do Inglês “Coding Sequences”) derivadas de TEs, foi revisto por Pavlicek e colaboradores (2002), que encontraram várias inconsistências. Primeiro, do conjunto de genes analisados, 30% foram anotados como hipotéticos e 63% foram anotados como preditos. Em outras palavras, não há evidências experimentais que suportam estes genes como sendo CDSs funcionais *bona fide*. Segundo, a maioria das CDSs humanas que Nekrutenko e Li encontraram com similaridade a TEs foram derivadas de elementos *Alu* (SINEs), que não têm capacidade codificante. Pavlicek et al. (2002) aplicaram uma abordagem mais conservadora para identificar CDSs derivadas de TEs: analisaram CDSs somente relacionadas a proteínas com estrutura tridimensional (3D) solucionadas. A estrutura 3D

representa a fonte mais acurada para a existência da proteína em consideração. Quando estas CDSs foram analisadas usando a mesma estratégia de detecção de Nekrutenko e Li, nenhuma evidência de seqüências derivadas de TEs foi encontrada. Uma técnica mais sensível revelou 28 casos de CDSs derivadas de TEs, mas daquelas que efetivamente codificam proteínas, nenhuma era formada por sequências *Alu* (revisado por Bowen and Jordan, 2007). Apesar da existência de algumas proteínas envolvidas no reconhecimento ou ligação molecular conterem seqüências de aminoácidos derivadas de elementos *Alu* (NEKRUTENKO; LI, 2001; HOENIKA et al., 2002), nenhuma delas parece ter claramente um papel funcional (PAVLICEK et al., 2002), com exceção de uma forma alternativa da proteína editase 1 específica a RNA de dupla fita (GERBER et al., 1997) e uma nova subunidade da proteína caseína quinase 2 (HILGARD et al., 2002). Entretanto, elementos *Alu* são altamente prevalentes em regiões ricas em genes humanos (INTERNATIONAL HUMAN GENOME SEQUENCING CONSORTIUM, 2001) e abrigam numerosos sítios potenciais de *splicing*, que podem facilitar a incorporação no mRNA (SOREK et al., 2002). Realmente, numerosos éxons humanos que apresentam processamento do RNA mensageiro (5%) contêm seqüências *Alu* (SOREK et al., 2002). Entretanto, o potencial codificante e a função biológica destas seqüências ainda são questões em aberto o que tem implicado na aparente ausência desses elementos em proteínas com estrutura 3D representativas (BOWEN; JORDAN, 2007).

Essa questão sobre a contribuição dos TEs para o proteoma foi também explorada por Gotea e Makalowski (2006), que analisaram proteínas não redundantes do Banco de dados de proteínas RCSB para fragmentos peptídicos derivados de TEs. Surpreendentemente, e corroborando as análises de Pavlicek et al. (2002), a proporção de TEs em nível protéico foi de apenas ~0,1% (3 proteínas não redundantes), muito menor que os 4% relatados por Nekrutenko e Li (2001). Essas proteínas – calpaína 1, granzima A e proteína tirosina fosfatase - com possíveis TE traduzidos caracterizam-se por apresentarem elementos antigos (degradados), indicando que a incorporação de elementos móveis em proteínas funcionais é baixa.

Além da ausência de capacidade codificante de muitos elementos incorporados no mRNA, outro importante fator parece contribuir fortemente para a raridade dos casos

de proteínas funcionais derivadas de cassetes de TEs. Como proposto por Hillman e colaboradores (2004), s*plicing* alternativo de éxons contendo TEs pode ser relacionado ao mecanismo NMD (da sigla em Inglês "*nonsense-mediated decay*") no qual o TE pode contribuir para o aparecimento de códons de terminação prematuros e, assim, ser sujeito a uma eficiente maquinaria de controle de qualidade resultando em um mecanismo adicional de controle da expressão gênica. Assim, tais variantes de transcritos "desfavoráveis" podem ser degradadas antes ou após a tradução, e mesmo se a proteína for produzida, ela pode ser imediatamente eliminada (JACOBSON; PELTZ, 1996; HILLEREN; PARKER, 1999; WAGNER; LYKKE-ANDERSEN, 2002; HILLMAN et al., 2004). A seguir, apresentamos uma fundamentação teórica e revisão bibliográfica atualizada para o mecanismo NMD em eucariotos.

I.2 Supervisão de mRNA em ação: manutenção da saúde vs severidade fenotípica

Um terço das desordens genéticas herdáveis, e muitas formas de câncer, são causadas por mudanças na matriz de leitura, e mutações sem sentido, que resultam em códons de terminação prematuros (PTC da sigla em Inglês *Premature Terminate Codon*). Atualmente tem sido demonstrado que, contrariamente ao esperado, a conseqüência predominante das mutações sem sentido não é a síntese de proteínas truncadas. A maioria dos transcritos sem sentido são reconhecidos e degradados, eficientemente pela maquinaria NMD (ver CONTI; IZAURRALDE, 2005). Esta via de supervisão de mRNA é ubíqua entre os eucariotos e se propõe ser uma das formas de proteção do organismo dos efeitos de ganho de função ou efeitos deletérios dominantes de proteínas truncadas resultantes de transcritos sem sentido, se esses fossem estáveis. De fato, a severidade fenotípica mediada por doenças causadas por mutações sem sentido pode ser predita pela intensa redução nos níveis de mRNA dos alelos mutantes (DIETZ et al., 1993; HALL et al.,1994). Assim, por meio da via NMD, os eucariotos identificam e degradam mRNAs aberrantes e asseguram a fidelidade da transcrição gênica.

Além do seu papel de controle da qualidade protéica, por degradar mRNAs PTC^+, recentes estudos têm revelado que a via NMD está também envolvida na regulação de várias atividades celulares. Por exemplo, em *Caenorhabitis elegans*, a expressão das proteínas ribossomais L3, L7a, L10a e L12 e as proteínas SR, SRp20 e SRp30b, é regulada pós-transcricionalmente via combinação dos processos de *splicing* alternativo e NMD (MITROVICH; ANDERSON, 2000; MORRISON et al., 1997). Em cada caso, mostrou-se que isoformas "produtivas" são produzidas *in vivo*, bem como aquelas "improdutivas" com um PTC. O *splicing* regulado gera isoformas "improdutivas" levando à redução da expressão da proteína, desde que essas isoformas sejam degradadas por NMD. Este sistema, chamado de "*Splicing* improdutivo regulado e tradução" (RUST da sigla em Inglês *Regulated unproductive splicing and translation*) ocorre no genoma humano. Por exemplo, mostrou-se que a proteína SC35 autoregula sua própria expressão usando RUST (SUREAU et al., 2001). Quando os níveis da proteína estão elevados, a SC35 se liga no próprio pré-mRNA induzindo a produção de mRNA *SC35* PTC^+, que é desestabilizado por NMD, resultando em níveis menores de proteína SC35.

Em eucariotos, a via NMD requer a tradução de mRNA ativos e fatores de ação *in trans* específicos a NMD (AMRANI et al., 2006; MAQUAT, 2004). Três fatores de ação *in trans* bem investigados em NMD são as proteínas codificadas pelos genes *UPF1*, *UPF2* e *UPF3*, que foram descobertas em buscas por supressão genética em *Saccharomyces cerevisiae*. Estes genes são evolutivamente conservados e sua deleção ou silenciamento previne o NMD em todos os eucariotos testados (CONTI; IZAURRALDE, 2005). Acredita-se que as três proteínas UPF formem um complexo trimérico que constitui o *core* da maquinaria NMD e se ligam ao códon de terminação prematuro desencadeando o processo que leva à degradação do mRNA (CONTI; IZAURRALDE, 2005). Fatores adicionais são provavelmente envolvidos em NMD. Em levedura, mutações em vários genes envolvidos na tradução e processamento do pré-mRNA também estabilizam seletivamente mRNA contendo PTC, ou mRNA PTC^+ (CUI et al., 1999; WELCH; JABOBSON, 1999). Em animais multicelulares, uma pletora de fatores adicionais tem sido relacionada ao NMD (CUI et al., 1999). As proteínas melhor caracterizadas são SMG1, SMG5, SMG6 e SMG7, primeiramente

identificadas como supressores de mutações em vários genes de *Caenorhabditis elegans* (HODGKIN et al., 1989; CALI et al., 1999). Uma busca genética inicial também identificou *smg-2*, *smg-3* e *smg-4*, os homólogos em *C. elegans* de *UPF1*, *UPF2* e *UPF3* (HODGKIN et al., 1989; PAGE et al., 1999). SMG1 é uma proteína quinase que pode fosforilar UPF1 (ou SMG2; GRIMSON et al., 2004); SMG5, SMG6 e SMG7 regulam o estado de fosforilação de UPF1, na qual a UPF1 fosforilada é necessária para a sua interação com fatores de decaimento do mRNA e a inibição de novos *rounds* de tradução (CONTI; IZAURRALDE, 2005; UNTERHOLZNER; IZAURRALDE, 2004).

A deleção de genes responsáveis pela codificação das proteínas do *core* NMD (UPF1, UPF2 e UPF3) não prejudica a viabilidade em *S. cerevisiae* ou *C. elegans* (CULBERTSON et al., 1980), mas alguns fatores são essenciais para outros organismos. Por exemplo, a *UPF1* (também conhecida como *Rent1* em camundongo) é essencial para o desenvolvimento embrionário (MEDGHALCHI et al., 2001); a depleção da *UPF2* em células hematopoiéticas causa a perda de todo o tecido hematopoiético e das populações progenitoras (WEISCHENFELDT et al., 2008) e, em *Arabidopsis*, a atividade de *UPF1* está envolvida no controle do tamanho da semente e no desenvolvimento da planta (YOINE et al., 2006). Ainda não é claro se NMD, ou alguma outra função essencial destas proteínas é necessária para a sobrevivência de *Arabidopsis* (AZZALIN; LINGNER, 2006; LUKE et al., 2007; AZZALIN et al., 2007; AJAMIAN et al., 2008).

I.2.1 Mecanismos envolvidos na identificação de PTCs em leveduras: o modelo "UTR 3' falso"

O fator chave em NMD consiste em como o processo distingue entre um códon de terminação normal e um PTC. Recentes estudos sugerem que a distância entre o PTC e cauda poli-A pode ser o determinante chave. Este mecanismo, conceituado como modelo *UTR 3' falso*, consideraria o ponto de terminação da tradução prematuro como intrinsecamente anormal porque estaria localizado a uma longa distância da extremidade

3', e esta distância, impediria a interação normal entre o ribossomo e as proteínas de ligação à cauda poli-A (PABP da sigla em Inglês *poly(A) binding protein*). Diante disso, os fatores NMD associar-se-iam ao ribossomo no ponto de terminação da tradução. Esse modelo é baseado na observação que os ribossomos localizados em um PTC e em um códon de terminação normal deixam diferentes marcas nos mRNAs, conhecidas como *toeprints*. Recentes observações - que o fator de terminação da tradução eRF3 e PABP interagem durante a terminação normal, e que superexpressão de PABP pode acentuar a terminação da tradução em alguns mutantes eRF3 – são consistentes com o modelo "UTR 3' falso" (detalhes ver BROGNA; WEN, 2009).

I.2.2 Reconhecimento de PTC em mamíferos: o modelo "EJC"

Uma característica diferencial da expressão gênica em eucariotos é que a transcrição e o processamento do RNA são realizados no núcleo, enquanto a tradução ocorre no citoplasma. Diante disso, acreditava-se que não havia ligação entre eventos nucleares (e.g. processamento diferencial do pré-mRNA) e eventos citoplasmáticos (e.g. tradução e destruição do mRNA). Esta visão foi desafiada em recentes estudos, primeiramente pela descoberta que a presença de íntrons em genes acentua a ação do mecanismo NMD. Vários trabalhos mostraram que PTCs em mRNAs de mamíferos causam NMD, quando eles estão localizados antes de um íntron e, que a inserção artificial de um íntron, após um códon de terminação selvagem, pode atuar similarmente a um PTC (BROGNA; WEN, 2009). Além disso, a descoberta que RNAms derivados de genes naturalmente sem íntrons são imunes ao NMD são também consistentes com estas observações. Diante disso, foi proposto que íntrons localizados após um PTC serviriam como um sinal ao NMD (BROGNA; WEN, 2009). Finalmente, também foi observado que muitos genes nos quais o PTC está muito próximo ao último íntron são imunes ao NMD, e isto levou a proposta de uma regra: para um PTC dirigir uma forte ativação NMD, este deveria estar localizado ao menos 50-55 bases antes da última junção exon-exon (e.g. NAGY; MAQUAT, 1998), de modo que um complexo

multiproteínas, com um *core* de quatro proteínas conhecido como “EJC”, depositado a 20-24 bases antes da junção exon-exon, possa servir como uma plataforma de ligação aos fatores NMD, sobretudo UPF2 e UPF3 (HIR et al., 2001). A versão corrente do modelo EJC é que na terminação da tradução, o “EJC” promove o recrutamento e ativação da UPF1 no ribossomo, iniciando a formação do “complexo de indução NMD” e isso promove a degradação acelerada do mRNA alvo (para detalhes, ver BROGNA; WEN, 2009).

I.2.3 Discriminação de PTC em plantas

Embora muito seja conhecido sobre o mecanismo NMD em leveduras e mamíferos, pouco se sabe sobre NMD em plantas. Em alguns cultivares de soja, uma mudança na matriz de leitura causa PTC, levando a níveis altamente atenuados do mRNA do inibidor de tripsina Kunitz (JOFUKU et al., 1989). Plantas de tabaco transgênico, expressando o gene da patatina, contendo PTC, têm níveis menores de mRNA que plantas que expressam o gene selvagem (VANCANNEYT et al., 1990). Em tabaco, o mRNA PTC^+ da ferredoxina-1 é altamente degradado por meio de NMD em tecidos expostos à luz (DICKEY et al., 1994; PETRACEK et al., 2000). Em alguns cultivares de cevada, mutações sem sentido no gene *waxy* reduzem o conteúdo de amilose do cloroplasto (DOMON et al., 2002).

Tem sido observada a degradação de transcritos contendo PTC derivados de genes contendo ou não íntrons (e.g. WU et al., 2007). Usando um sistema de transformação transiente, Kertesz et al. (2006) demonstrou que tanto UTRs 3´ longas quanto introns localizados na UTR 3´ poderiam atuar como elementos de ativação NMD em *cis* em plantas, bem como a distância do PTC até o próximo íntron é importante para ativar o NMD. Estudos recentes usando mutantes de *Arabidopsis* têm indicado o papel conservado dos genes *UPF* em plantas. Em mutantes UPF1 e UPF3 em *Arabidopsis*, espécies de mRNAs PTC^+ são mais abundantes que em plantas selvagens;

além disso, mudanças morfológicas nestes mutantes sugerem que NMD está envolvido no desenvolvimento (e.g. HORI; WATANABE, 2005; YOINE et al., 2006).

Em *Nicotiana attenuata*, um gene inibidor da proteína tripsina (*TPI* da sigla em Inglês "*Trypsin proteinase inhibitor*"), que tem um papel central na defesa da planta contra a herbivoria, tem servido com um modelo para a descoberta de importantes novidades sobre NMD em plantas (WU et al., 2007). Primeiro, a utilização de um sistema de silenciamento gênico induzido por vírus (VIGS da sigla em Inglês "*virus-induced gene silencing*"), para silenciar os níveis de transcritos dos três genes *UPF*, permitiu mostrar que os níveis de mRNA de TPI PTC^+ foram marcadamente aumentados, principalmente para UPF1 e UPF2-VIGS, sugerindo que os três genes *UPF* são requeridos para NMD em *N. attenuata*. Segundo, usando várias construções TPI com diferentes estruturas gênicas, isto é, construções que representam genes TPI com ou sem íntrons, e genes TPI abrigando PTC antes ou após o íntron, mostraram que tanto genes com íntrons, quanto sem, são alvos de NMD. Por outro lado, o mecanismo pode ser abolido quando o PTC está localizado muito próximo (4 pb antes) de um íntron. Terceiro, usando-se construções da região codificante do gene TPI, contendo PTC em diferentes posições, mostrou-se que PTCs localizados no meio da seqüência codificante ativam mais eficientemente NMD que PTC localizados no início ou fim da seqüência codificante. Este resultado é consistente com a idéia que a ação NMD correlaciona-se eficientemente com a posição do PTC em plantas, como em mamíferos e levedura. Por fim, usando calos embriogênicos tratados com o inibidor de biossíntese de proteínas Cicloheximida (CHX), mostrou-se que esse composto pode estabilizar espécies de mRNA contendo PTC, indicando a necessidade de que um *round* inicial de tradução de mRNA ocorra para genes que abrigam PTC sejam degradados por meio da via NMD, como ocorrem em mamíferos e levedura (e.g. CARTER et al., 1995). Dados em conjunto, estes resultados destacam a complexidade da ativação NMD em plantas.

I.3 Atividade transcricional de TEs em plantas

Devido à atividade mutagênica dos TEs, genomas de eucariotos têm desenvolvido eficientes mecanismos de silenciamento dessas sequências móveis (CASACUBERTA; SANTIAGO, 2003). Sabe-se que, em plantas, a atividade transcricional de retrotransposons é altamente controlada, embora alguns desses TEs sejam transcritos em cultura de células ou sob condições de estresse, tais como infecção por patógenos, injúrias físicas ou estresse abiótico (GRANDBASTIEN, 1998; TAKEDA et al., 1998). Dentre os TEs encontrados em plantas, sabe-se que os elementos *Mutator* são constitutivamente expressos em meristemas, órgãos imaturos e tecidos completamente diferenciados em milho e que sua atividade transposicional é restrita às últimas divisões celulares, durante o desenvolvimento de tecidos somático e germinal, e não é correlacionada com o nível de transcritos (RUDENKO; WALBOT, 2001; RUDENKO et al., 2003).

O trabalho de referência analisando abundância, distribuição e atividade transcricional de elementos repetitivos em larga escala foi realizado no genoma de milho (*Zea mays*) por Meyers et al. (2001). Os autores usaram a estratégia de sequenciamento genômico por amostras selecionadas randomicamente, combinado com análises por *macroarrays* de *BACs* (da sigla em Inglês, *Bacteria Artificial Chromosome*), para estudar a composição, estrutura e arranjo do genoma de *Zea mays*. Assim, foi possível sugerir que 80% deste genoma contêm sequências repetitivas. Além disso, buscas em um banco de dados de cerca de 407.000 *ESTs* permitiu identificar apenas 56 clones homólogos a TEs expressos. Essas buscas indicaram ainda que muitos dos retrotransposons encontrados nas bibliotecas de *ESTs* estão em baixo número de cópias nas seqüências genômicas, enquanto que elementos abundantes são infrequentemente encontrados entre as bibliotecas de *ESTs*.

Outra importante contribuição é destacada pelos trabalhos de Rossi et al. (2001) e Araújo et al. (2005) no transcriptoma de cana-de-açúcar (*Saccharum officinarum*). Em um primeiro momento, buscas de TEs expressos em *ESTs*, em condições estringentes (E= e^{-50}) para comparações de seqüência contra TEs de plantas completamente

caracterizados, permitiram identificar 267 clones (0,1%) similares a TEs expressos em um total de 260.781 *ESTs* (ROSSI et al., 2001). Posteriormente, uma caracterização transcricional de 64 TEs não redundantes em diferentes tecidos (calos, meristema apical, folhas e flor) por meio da combinação de *Northern* eletrônico, *macroarray* e transformação genética estável e transiente, indicou os calos como o tecido com o maior número de TEs expressos. Esse resultado revela que a cultura de tecidos induz drasticamente a expressão de diferentes elementos (ARAÚJO et al., 2005).

Jiao e Deng (2007) conduziram a análise de maior escala já realizada da atividade transcricional de genes relacionados a TEs. Os dados de atividades foram gerados usando experimentos de macroarranjo de 2.191 genes de arroz não redundantes relacionados a TEs associados a 15 estágios de desenvolvimento e condições de estresse. Cinco tecidos crescidos sob condições normais, em diferentes estágios de desenvolvimento, quatro culturas de tecidos e seis amostras sob condições de salinidade ou seca foram incluídos. Genes relacionados a TEs exibiram menor atividade transcricional que genes não relacionados (genes hospedeiros), embora transcritos de todas superfamílias de retrotransposons e transposons tinham sido detectados. Maior atividade transcricional foi detectada em elementos das superfamílias MULE e CACTA. Alguns elementos apresentaram expressão-específica a cultura de células.

I.3.1 Dinâmica de TEs e a regulação epigenética

Elementos de transposição são freqüentemente considerados como elementos parasitas devido à capacidade de aumentarem o número de cópias, o que pode ser negativamente correlacionado com o *fitness* do genoma hospedeiro. Embora muitos desses elementos estejam silenciados no genoma devido a mutações e deleções que suprimem a transposição, alguns elementos autônomos mantêm-se intactos, mas silenciados, na forma de “elementos crípticos”. Uma noção amplamente aceita é que mecanismos de defesa epigenética evoluíram como resposta aos efeitos potencialmente negativos de TEs ativos (SLOTKIN; MARTIENSSEN, 2007). O paradoxo da atividade

de TEs de plantas é dramaticamente ilustrado em milho, onde o genoma dobrou de tamanho por meio de retrotransposição nos últimos 3 milhões de anos (SANMIGUEL et al., 1998), mas somente um retrotransposon com LTR ativo tem sido identificado. Evidências recentes têm mostrado que a regulação dos TEs se dá por meio das três vias gerais de regulação epigenética, a remodelação da cromatina, a metilação do DNA e os pequenos RNAs não-codificantes (detalhes ver REBOLLO et al., 2010).

I.3.1.1 Modificações da cromatina

Um conjunto de modificações de cromatina suprime a transcrição de TEs, incluindo modificações na cauda de histonas, metilação de DNA e alterações no empacotamento e condensação da cromatina. Modificações da cauda amino (N)-terminal de histonas alteram a ligação de fatores protéicos e a aproximação de fatores de transcrição. Nucleossomos, que são associados a TEs, são enriquecidos na metilação da histona H3 na lisina 9 (H3K9), a qual é um sinal para cromatina inativa e transcricionalmente repressiva. Por fim, mutações em genes requeridos para a modificações na cauda de histonas repressivas levam a reativação de TEs (detalhes ver SLOTKIN; MARTIENSSEN, 2007).

I.3.1.2 Metilação de DNA

Metilação de DNA é considerada o primeiro nível de regulação gênica. Embora não seja presente em todos os eucariotos, a metilação do resíduo citosina é um importante sinal que reprime a transcrição em TEs. Em plantas e animais, o resíduo citosina pode ser metilado em um contexto simétrico (CpG), e essa metilação é copiada para uma nova fita de DNA no processo de replicação do DNA, provendo um mecanismo a herança do silenciamento de TEs. A metilação de DNA assimétrica também pode ter TEs como alvo, desde que o processo ocorra para elementos ativos em transposição e

não requeira substratos parcialmente metilados como guia (detalhes ver SLOTKIN; MARTIENSSEN, 2007; REBOLLO et al., 2010).

I.3.1.3 Modificações da cromatina mediada por siRNA

Em muitos eucariotos, componentes siRNA (do Inglês, *small interfering RNA*) são requeridos para as modificações da cromatina. Embora a manutenção de algumas modificações de cromatina repressiva não requeiram moléculas de siRNA, não se sabe se estas modificações são inicialmente estabelecidas por mecanismos baseados em siRNA (SLOTKIN; MARTIENSSEN, 2007). Dados acumulados na literatura mostram que a manutenção do silenciamento de TEs envolve funções redundantes de vias independentes ou não de siRNA, embora a compreensão de modificações da cromatina mediada por siRNA seja incompleta. Este tipo de modificação da cromatina tem sido extensivamente investigado em *A. thaliana*. Nessa espécie, as famílias gênicas codificadoras de siRNA participam na regulação pós-transcricional, bem como nas modificações de cromatina mediada por siRNA (XIE et al., 2004). Além disso, duas classes com tamanhos distintos de siRNA foram identificadas nesta espécie. Uma classe de siRNA menores, de 21-22 nts, compreende siRNAs envolvidos na regulação pós-transcricional (PTGS da sigla em Inglês *Post-transcriptional gene silencing*), enquanto a classe de siRNAs maiores, de 24-26 nts, compreende aqueles envolvidos na regulação transcricional (TGS da sigla em Inglês *Transcriptional gene silencing*) por meio da metilação do DNA dependente de RNA, e são derivados, predominantemente, de TEs e repetições em *tandem* (rasiRNA) (HAMILTON et al., 2002). O gene *dicer-like 3* (DCL3) gera a classe de siRNA de TEs maiores, e uma das 10 proteínas argonauta, AGO4, se liga a um subconjunto destes siRNA. O complexo AGO4+siRNA parace ter um papel catalítico e não catalítico no silenciamento de diferentes TEs por realizar a metilação assimétrica (QI et al., 2006).

I.3.1.4 Silenciamento Transcricional (TGS) e Pós-transcricional (PTGS) de TEs

Uma determinante chave para a ativação de TGS e PTGS é a ocorrência de múltiplas cópias idênticas ou quase idênticas de TEs no genoma. TGS de TEs pode ser ativado por interações diretas TE-TE. Além disso, somente TGS tem sido mostrado como capaz de regular TEs de plantas, para as quais é associada à metilação de promotores de TEs (VAUCHERET; FAGARD, 2001). Em PTGS, o silenciamento é causado pela degradação específica da seqüência de RNA no citoplasma. PTGS é ativado por RNAs de dupla fita (dsRNA) produzidos pelo pareamento de RNAs *sense* e *antisense* ou de RNAs que contenham um invertido terminal. Este mecanismo tem sido bem documentada em plantas como uma importante via de defesa contra replicação de vírus, como também por regular a atividade de TEs em animais em que a metilação de DNA é ausente (VOINNET, 2001). A formação de dsRNA pode também envolver polimerases de RNA dependentes de RNA, fato evidenciado pelo silenciamento de transgenes e vírus de plantas. Esses RNAs de dupla fita são degradados em pequenos RNAs de interferência (siRNAs) de 21-26 nts por um complexo enzimático que inclui as proteínas tipo DICER e helicases de RNA, os quais podem atuar como guias para a degradação por meio de RNAses. siRNAs também podem entrar no núcleo e mediar a metilação *de novo* de transgenes, possivelmente por meio de interações RNA-DNA. A metilação dirigida por RNA ocorre ao longo da região de DNA que são complementares com o guia de RNAs (detalhes ver FESCHOTTE et al., 2002).

I.4 Correlação entre hibridização, formação de genomas poliplóides e proliferação de TEs

O registro de que TEs se organizam de forma “aninhada” no genoma de milho, somado ao fato de que 80% deste genoma é constituído por TEs, levaram Voytas e Naylor (1998) a questionarem porque tantos elementos têm sido incorporados

rapidamente e porque eles são tolerados. Segundo os autores, estas duas perguntas são relacionadas, pois a tolerância pode promover a amplificação e a amplificação pode acelerar a tolerância. Os autores argumentam que a inserção de um retrotransposon em um sítio pode prover alvos múltiplos para inserções posteriores, e como a porcentagem de elementos aumenta no genoma, têm-se um aumento exponencial de sítios de integração. Isto pode explicar tanto a rápida amplificação como a organização "aninhada" de retrotransposons no genoma do milho.

Dados acumulados na literatura mostram que os genomas das plantas adaptam-se a perturbações em larga escala. Mudanças na ploidia, por exemplo, são comuns e cerca de 70% das Angiospermas sofreram poliploidização durante sua história evolutiva (HE et al., 2003). Um efeito óbvio da poliploidia é a duplicação de locos gênicos, que podem ou não conferir vantagens seletivas para os poliplóides que suportam extensivas mudanças em seus genomas nucleares em poucas gerações após sua formação, tais como deleções, translocações, inversões e duplicações resultando em um padrão alterado da expressão e função gênica (SONG et al., 1995; FELDMAN et al., 1997; HE et al., 2003). Embora os mecanismos dessas mudanças sejam pouco compreendidos, a transposição é claramente uma possibilidade: um elemento transponível quiescente em um diplóide pode ser ativado no novo ambiente genético do poliplóide. Alternativamente, a redundância genética no poliplóide pode mitigar os efeitos deletérios da transposição (NAYLOR, VOYTAS, 1998). Assim, TEs podem acumular e fixar em alopoliplóides, mesmo em regiões genômicas ricas em genes.

Uma terceira hipótese é que mecanismos de silenciamento hospedeiro tais como metilação, podem ser reduzidos em híbridos alopoliplóides, permitindo taxas de transposição se torne elevadas (MADLUNG et al., 2002, 2005). Por outro lado, o grande número de cópias por indivíduo pode reduzir a atuação de deriva, potencialmente aumentando a eficácia de seleção purificadora comparado com uma população diplóide de mesmo tamanho (HAZZOURI et al., 2008).

O cenário acima foi construindo segundo considerações sobre a dinâmica de TEs em diferentes organismos, mas poucos estudos focaram esta questão em experimentações refinadas. Evidências experimentais para a ativação de transposição em um poliplóide são encontradas de um estudo de Li et al. (2004) que realizaram

experimentos de *dot blot* com 17 famílias de TEs em *Triticum aestivum* diplóide e hexaplóide (AABBDD). As análises revelaram que o trigo poliplóide apresenta não apenas maior número de cópias de TEs que o trigo diplóide, mas também um número de cópias maior que a soma nos genomas de suas espécies parentais. Adicionalmente, Madlung et al. (2005), usando microarranjos genômicos de *Arabidopsis* para analisar uma região heterocromática do cromossomo 4, encontrou que o transposon *Em-Spm* estava ativo em um híbrido alopoliplóide, gerado experimentalmente, comparado às linhagens parentais autotetraplóides. Os autores mostraram ter ocorrido perda de metilação associada à alopoliploidia concomitante com ativação transcricional do elemento.

Enquanto estes estudos suportam a hipótese de redução do silenciamento de TEs associada com hibridização, há poucas evidências na literatura para uma conexão clara entre duplicação gênica e um relaxamento da seleção em TEs. Dois estudos em particular demonstraram que elementos de transposição são super-expressos em regiões duplicadas de genomas individuais, a saber: *A. thaliana* (HUGHES et al., 2003) e *Magnaporthe grisae* (THONG et al., 2006), embora outro estudo em levedura encontrou padrão oposto (HUGHES; FRIEDMAN, 2004). Enquanto os autores interpretaram seus resultados como evidência que TEs são importantes em causar duplicação, a explicação alternativa é que estas regiões duplicadas experimentam seleção relaxada contra TEs devido à redundância na função gênica. Por exemplo, *Brassica oleracea*, uma espécie hexaplóide antiga (ZIOLKOWSKI et al., 2006), mostra evidências para um forte acúmulo de muitas famílias de elementos de transposição relacionados aos de *A. thaliana* (ZHANG; WESSLER, 2004), que é também consistente com poliploidia permitindo a proliferação de TEs. Interessantemente, a ativação de certos TEs pode resultar no silenciamento de genes adjacentes *downstream* (KASHKUSH et al., 2003). Por exemplo, a ativação do transposon *Wis* causou silenciamento do gene *purB*, mostrando que a alopoliploidia pode induzir a atividade de elementos de transposição, assim criando potencial para mutagênese insercional, enquanto altera o padrão local de expressão gênica por transcrição "*readout*".

I.5 Variabilidade do repertório transcricional em Coffea spp: estratégias para avaliar o impacto dos TEs sobre o genoma hospedeiro e os mecanismos de controle do hospedeiro sobre os TEs

A genômica de plantas sofreu uma revolução com a publicação do genoma completo de *A. thaliana* (ARABIDOPSIS GENOME INITIATIVE, 2001), *O. sativa* (INTERNACIONAL RICE GENOME SEQUENCING PROJECT, 2005) e *Populus trichocarpa* (TUSKAN et al., 2006). A cultura cafeeira também está se beneficiando dessa revolução. Por iniciativa do Consórcio Brasileiro de Pesquisa e Desenvolvimento do Café (CBP&DC), em colaboração com a Fundação de Amparo a Pesquisa do Estado de São Paulo (FAPESP) e o Centro Nacional de Recursos Genéticos e Biotecnologia, teve início, em fevereiro de 2002, o "Projeto Genoma Café Brasileiro" (PGCB). Esse projeto foi criado para desenvolver ferramentas úteis para descoberta de genes e análises de genética funcional em café e espécies relacionadas e contribuir ao avanço do conhecimento da organização e evolução do genoma do café, este foco de nossos estudos.

A disponibilidade de seqüências de cDNA de bibliotecas de uma ampla variedade de tecidos, estágios de desenvolvimento e material vegetal submetido a estresse biótico e abiótico no banco de dados do Projeto Genoma Café Brasileiro resultou na identificação de mais de 30 mil unigenes (seqüências oriundas de um mesmo lócus expresso) diferentes (VIEIRA et al., 2006). Esta informação proporciona a oportunidade de contribuir em uma área da literatura completamente escassa e recente como o estudo da participação de TEs na composição e expressão de transcriptomas oriundos de genomas complexos, como *Coffea*.

Os dados gerados pelo PGCB foram armazenados em duas plataformas de bioinformática. Uma delas foi desenvolvida pelo Laboratório de Genômica e Expressão (LGE - http://www.lge.ibi.unicamp.br/cafe) e a outra criada pela Embrapa Recursos Genéticos e Biotecnologia (https://alanine.cenargen.embrapa.br/EST/). Apesar do conjunto de dados ser o mesmo em ambas as plataformas, os tratamentos empregados foram diferentes. Inicialmente, por exemplo, na plataforma do LGE, as *ESTs* foram

agrupadas separadamente por espécies, enquanto na base da Embrapa, as *ESTs* das três espécies (*C. arabica*, *C. canephora* e *C. racemosa*) foram tratadas conjuntamente.

Coffea arabica é uma espécie alotetraplóide (2n = 4x = 44) e autogâmica, enquanto *C. canephora* é diplóide (2n = 2x = 22) e exogâmica. Tem sido mostrado por evidências moleculares e citogenéticas que *C. arabica* foi derivada de uma hibridização espontânea entre duas espécies de café, *C. canephora* e *C. eugenoides*. Um recente estudo sobre o padrão e nível de divergência de seqüência de DNA entre *C. arabica* e *C. canephora* usando fragmentos genômicos ortólogos mostrou que a identidade de seqüência varia de 94-98%. Apesar do alto nível de similaridade, foram encontrados numerosos rearranjos cromossomais entre os dois conjuntos de fragmentos genômicos, incluindo inversões, deleções e inserções. Análise de 10 pares de genes ortólogos resultou na idade estimada de divergência de 0,36 a 1,01 milhão de anos, indicando uma origem recente da espécie alotetraplóide *C. arabica* (YU et al., 2008). A alta identidade entre *C. arabica* e um de seus parentais permite caracterizar a expressão de transcritos de interesse nessa espécie híbrida e um de seus parentais, *C. canephora*, apenas usando a seqüência da primeira. Essa estratégia não somente contribuirá na caracterização da atividade transcricional em duas diferentes espécies de *Coffea*, bem como no entendimento da atuação dos mecanismos de regulação gênica hospedeira sobre a tolerância ou controle da ativação relacionada aos processos de hibridização/poliploidização.

Um recente registro de Vidal *et al.* (2010) tem adicionado importantes informações sobre mudanças de expressão em genes homeólogos, que são os genes ortólogos das espécies ancestrais que compõe um poliplóide. Neste estudo foram utilizados dados de seqüencias expressas produzidas pelo PGCB (VIEIRA et al., 2006) e pelo Instituto de Pesquisa da Nestle (LIN et al., 2005). Combinando os dados de agrupamento (clusterização) de ESTs do alotetraplóide *C. arabica* e de um de seus parentais, *C. canephora*, com a ocorrência de SNPs (Polimorfismos de Nucleotídeos Únicos, da sigla em Inglês *Single Nucleotide Polymorphyms*), os autores analisaram a variação de expressão gênica de cada subgenoma de *C. arabica* (*C. canephora* e *C. eugenioides*). Dentre os resultados, destaca-se, primeiramente, que o transcriptoma de *C. arabica* é composto da contribuição aproximadamente igual de ESTs dos dois

ancestrais (48% das ESTs derivadas do subgenoma de *C. canephora* e 52% das ESTs de *C. eugenioides*), exceto, pela maior abundância de alguns transcritos oriundos de um dos dois subgenomas. Em segundo lugar, foi mostrado que os homeólogos mais expressos derivados de *C. eugenioides* apresentam uma função molecular particularmente relacionada à fotossíntese, metabolismo de carboidratos, respiração aeróbica e fosforilação; enquanto que aqueles derivados de *C. canephora* estão relacionados à resposta a estímulos hormonais, transdução de sinais GTP, tradução, biogênese de ribossomos e transporte mediado por vesículas. Em terceiro, para cinco homeólogos com expressão diferencial, que se apresentavam como pares de genes parálogos, em quatro deles, o parálogo mais expresso em *C. canephora* continuaram a ser mais expresso em *C. arabica*, mostrando uma conservação do padrão de expressão em relação a um de seus ancestrais.

Em face do exposto, com o objetivo de averiguar o impacto de fragmentos de TEs na expressão e função gênica, bem como as condições que provocam a ativação transcricional de TEs ativos e os mecanismos que os regulam em tecidos e condições específicas, pretende-se responder nesta Tese, usando como modelo elementos de transposição em espécies de *Coffea*, as seguintes indagações: Transcritos contendo cassetes de TEs são silenciados, comparados àqueles sem TEs? E sobre os TEs ativos? Hipotetizando que transcritos modificados por TEs sejam silenciados, o processo de hibridização/poliploidização confere algum relaxamento nos mecanismos de controle da expressão gênica hospedeira, possibilitando que as isoformas (quais) sejam mais expressas? Estresse hídrico e, sobretudo, cultura de células induzem a ativação de TEs?

II. OBJETIVOS

Com o intuito de testar a hipótese de que elementos de transposição exonizados afetam o padrão de expressão gênica, bem como avaliar a variação de expressão de TEs ativos em decorrência de sua distribuição cromossomal, este trabalho foi executado usando os transcriptomas das três espécies de café que compõem o banco de dados do PGCB (*C. arabica*, *C. canephora* e *C. racemosa*) e outros bancos de dados públicos com os seguintes objetivos específicos:

1) Avaliar a freqüência e função molecular de seqüências codificantes quiméricas, bem como, características estruturais dos fragmentos de TEs (tamanho, sentido de inserção e nível de conservação de seqüência) em *ESTs* e unigenes usando ferramentas computacionais;
2) Caracterizar o perfil de expressão de transcritos quiméricos ou não, ambas as seqüências altamente relacionadas por similaridade, por meio de macroarranjos;
3) Avaliar a abundância, e classificar por família, os TEs transcricionalmente ativos em *ESTs*;
4) Caracterizar o perfil de expressão e distribuição cromossomal de TEs expressos selecionados por meio de macroarranjos e Hibridização *in situ* Fluorescente (FISH), respectivamente.

Os resultados obtidos foram organizados e distribuídos no formato de artigos científicos. O capítulo 1 refere-se à conclusão do objetivo 1, que se encontra publicado no periódico *Molecular Genetics and Genomics*. O capítulo 2 refere-se ao objetivo 2, e o capítulo 3 refere-se aos objetivos 3 e 4. Ambos os capítulos serão submetidos à publicação.

III. RESULTADOS

Elementos transponíveis em transcritos de *Coffea* (Gentianales: Rubiacea) e seu papel na origem de diversidade de proteínas em Angiospermas

Capítulo 1

Mol Genet Genomics (2008) 279:385–401
DOI 10.1007/s00438-008-0319-4

ORIGINAL PAPER

Transposable elements in *Coffea* (Gentianales: Rubiacea) transcripts and their role in the origin of protein diversity in flowering plants

Fabrício Ramon Lopes · Marcelo Falsarella Carazzolle · Gonçalo Amarante Guimarães Pereira · Carlos Augusto Colombo · Claudia Marcia Aparecida Carareto

Received: 25 June 2007 / Accepted: 2 January 2008 / Published online: 30 January 2008

Abstract Transposable elements are major components of plant genomes and they influence their evolution, acting as recombination hot spots, acquiring specific cell functions or becoming part of protein-coding regions. The latter is the subject of the present analysis. This study is a report on the annotation of transposable elements (TEs) in expressed sequences of *Coffea arabica*, *Coffea canephora* and *Coffea racemosa*, showing the occurrence of 383 ESTs and 142 unigenes with TE fragments in these three *Coffea* species. Based on selected unigenes, it was possible to suggest 26 putative proteins with TE-cassette insertions, demonstrating a likely contribution to protein variability. The genes for two of those proteins, the fertility restorer (FR) and the pyrophosphate-dependent phosphofructokinase (PPi-PFKs) genes, were selected for evaluating the impact of TE-cassettes on host gene evolution of other plant genomes (*Arabidopsis thaliana*, *Oryza sativa* and *Populus trichocarpa*). This survey allowed identifying a FR gene in *O. sativa* harboring multiple insertions of LTR retrotransposons that originated new exons, which however does not necessarily mean a case of molecular domestication. A possible transduction event of a fragment of the PPi-PFK β-subunit gene mediated by Helitron ATREPX1 in *Arabidopsis thaliana* was also highlighted.

Keywords Transposable elements · *Coffea* genome · Protein diversity · Molecular domestication · Gene transduction

Communicated by M.-A. Grandbastien.

F. R. Lopes · C. M. A. Carareto (✉)
Laboratory of Molecular Evolution, Department of Biology,
UNESP, São Paulo State University,
15054-000 São José do Rio Preto, São Paulo, Brazil
e-mail: carareto@ibilce.unesp.br

M. F. Carazzolle · G. A. G. Pereira
Laboratory of Genomics and Expression,
Department of Genetics and Evolution, Institute of Biology,
UNICAMP, State University of Campinas,
13083-970 Campinas, São Paulo, Brazil

C. A. Colombo
IAC, Agronomic Institute of Campinas,
13001-970 Campinas, São Paulo, Brazil

Introduction

Transposable elements (TEs) are genetic units capable of moving within genomes and often making duplicate copies of themselves. As a consequence of this activity, they are mutagenic and can produce sequence changes in single genes, as well as large genome rearrangements (Zhang and Peterson 1999), both of which can alter the pattern of gene expression and function (Jordan et al. 2003). Moreover, TEs generate an enormous variability that can be used to create new genes or exons (reviewed in Kidwell and Lish 1997; Bennetzen 2000, 2005; Jordan et al. 2003; van de Lagemaat et al. 2003; Volff 2006) and new regulatory sequences (Jordan et al. 2003), besides being the source of transcription-regulating signals (Thornburg et al. 2006) and genome expansion or contraction (Fedoroff 2000; Bennetzen 2002). TEs are present in almost all organisms so far studied (Kidwell 2002; Shapiro and von Sternberg 2005), and in some genomes, like *Zea mays*, they can represent about 60–80% of the nuclear genome (Meyers et al. 2001).

The occurrence of TEs in intronic and intergenic regions has been widely reported (SanMiguel et al. 1996; Tikhonov et al. 1999; Bennetzen 2000). It was further demonstrated that these elements also contribute substantially to the

evolution of many genes at the transcriptional level through TE-cassettes (Makalowski et al. 1994; Makalowski 2000; International Human Genome Sequencing Consortium 2001; Nekrutenko and Li 2001; Sorek et al. 2002; Ganko et al. 2003; van de Lagemaat et al. 2003). TE-cassettes are fragments of TEs inserted into mRNA sequences. It has been proposed that a TE-cassette is generated after the activation of cryptic splice sites in an intron-residing TE sequence, or de novo through insertion into exons (Mitchell et al. 1991; Makalowski et al. 1994). Surprisingly, evidence also supports the translation of these cassettes, showing their contribution to the proteome (Gerber et al. 1997; Hilgard et al. 2002; Hoenicka et al. 2002). The presence of TEs in this region is of great interest, because they can change the function of the gene product. If this change is adaptive and conserved over evolutionary time, it is named exaptation (Brosius and Gould 1992), molecular domestication (Miller et al. 1999), or co-opted events (Sarkar et al. 2003).

Transposable elements have been extensively studied in human genomes, but less in plant genomes (e.g., Arabidopsis Genome Initiative 2000; Mao et al. 2000; Turcotte et al. 2001; Meyers et al. 2001; Sakai et al. 2007). However, to our knowledge, no information is available so far regarding TEs in the *Coffea arabica* genome or transcriptome. This fact enhances the relevance of studying the participation of TEs in the composition and expression of complex genomes such as that of *Coffea*. The availability of the Brazilian Coffee Genome Project database gave us the opportunity to analyze the contribution of sequences derived from TEs in ESTs and unigenes of *Coffea arabica*, *Coffea canephora* and *Coffea racemosa* (Family: Rubiaceae), and to estimate the frequency of TEs within protein-coding regions. Moreover, it was possible to evaluate the impact of TE-cassettes on the evolution of host genes in other plant genomes, comparing TE-cassettes of *Coffea* with gene models stored in public databases. This survey allowed identifying a fertility restorer (FR) gene in *Oryza sativa* harboring multiple insertions of LTR retrotransposons that originated new exons, which however does not necessarily imply a case of molecular domestication. A possible transduction event of a fragment of the PPi-PFK β-subunit gene from *Arabidopsis thaliana* mediated by the ATREPX1 element that belongs to the Helitron group was also highlighted.

Materials and methods

Coffea ESTs and unigenes database

The Brazilian Coffee Genome Project database (http://www.lge.ibi.unicamp.br/cafe) contains partial sequences of cDNA libraries of a wide range of tissues, developmental stages and plant material submitted to biotic and abiotic stressful conditions. Sequences accepted in the database had more than 250 bases with Phred quality >20; ribosomal sequences, of the vector and fragments of Poly-A^+ tail were removed (Vieira et al. 2006). Identification of TEs was carried out in 131,150 ESTs of *Coffea arabica* and in 10,566 ESTs of *Coffea racemosa*, clustered by the CAP3 program (Huang and Madan 1999) into 39,312 and 4,056 EST clusters, respectively, so-called unigenes. Initially, the analysis in *Coffea canephora* had been performed in 12,607 ESTs, however, aiming to increase the EST diversity in tissues not evaluated in the Brazilian Project, 46,914 ESTs produced by Lin et al. (2005), deposited in the SOL Genomics Network (http://www.sgn.cornell.edu/content/coffee.pl), were added, totaling 50,255 ESTs and 17,420 unigenes.

Occurrence of TE-cassettes at the transcriptional level

Aiming to identify TE-derived sequences in coding regions of the *Coffea* genomes, ESTs and unigenes were annotated de novo by the RepeatMasker (RM) version 3.1.5 (http://www.repeatmasker.org), with Cross_match version 0.990329 (Phrap/cross_match/swat package: http://www.phrap.org/phredphrapconsed.html) against a database of 1,010 reference TE sequences, such as *Arabidopsis thaliana* (487), *O. sativa* (361), *Hordeum vulgare* (67), *Z. mays* (63), *Triticum aestivum* (23) and *Sorghum bicolor* (9). For this analysis, the 10.06.2006 RepBase version was used, where a total of 1,304 plant reference sequences are available (Jurka et al. 2005—http://www.girinst.org/RepBase/update/index.html). To avoid spurious results, only scores of matches with RM >250, high sensitivity/low speed search conditions (condition "-s"), relative matrix based on GC level query ("-gcclac"), and where sequences of low complexity DNA had not been masked ("-nolow") were accepted. The RM cutoff scores >250 were chosen to eliminate false-positives. We considered all matches reported by RM for further analysis, without imposing additional scores or length thresholds. When ESTs or unigene sequences were annotated by the alignment with more than one reference TE from different plant TE databases, the matches between *Coffea* transcripts and reference TEs with the higher RM score was chosen. If the TE sequence occupied more than 70% of an EST or unigene, it was considered likely to represent a transcriptionally active element. Since this analysis focused TE-cassettes within coding regions, such transcribed TEs were discarded. The frequency of TE classes in ESTs, as well as sense and antisense TE-cassette orientation in unigenes, was evaluated and compared using the χ^2 test.

Transcripts diversity and TE-cassettes in the proteome

The functional annotation of all unigenes containing inserted TEs was obtained by BLASTX analysis against protein sequences stored at NCBI (National Center for Biotechnology Information) databases, particularly the NR (non-redundant) database (http://www.ncbi.nlm.nih.gov/blast/html/blastcgihelp.shtml#databases) [for details see Vieira et al. (2006)]. The BLASTX and RepeatMasker results were compared, and only the perfect overlapping between the localization of the TE-cassette and the protein sequence was accepted for inference of a putative TE-cassette in the proteome (Fig. 1a). This strategy allows eliminating all the TE-cassettes that are inserted into unigenes, but does not correspond to the protein sequence (Fig. 1b) or those that correspond to putative transcriptionally active TEs (Fig. 1c). Thus, overestimation of TE insertions in the proteome of the three *Coffea* species was avoided. The function of proteins with TEs in an ORF was assigned according to *the Gene Ontology* nomenclature (The Gene Ontology Consortium 2001) or to the LocusLink description (Pruitt and Maglott 2001).

Relationships between TE-cassettes and host genes

Gene models of the FR and pyrophosphate-dependent phosphofructokinase (PPi-PFK) genes of the *Arabidopsis thaliana* and *O. sativa* (LocusLink: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi.db=gene), *Ricinus communis* (stored in GenBank) and *Populus trichocarpa* (http://genome.jgi-psf.org/Poptr1/Poptr1.info.html) genomes were studied to evaluate the impact of TE-cassettes found in *Coffea* on host gene evolution. The gene models of *Arabidopsis thaliana* and *O. sativa* were identified by the name of the protein, and in *Populus trichocarpa* by the Gene Ontology identification number of the PPi-PFK gene (GO_ID: 0003872). Despite the non-availability of other genome projects in monocotyledons, the transcript sequences of PPi-PFK β-subunit genes of the four genomes under study were used as query against nucleotide and protein databases of the NCBI. RepeatMasker was used to analyze gene and transcript sequences with regard to occurrence of the similar TE fragments detected in the *Coffea* unigenes.

Evolutionary analysis

The multiple alignments of PPi-PFKs from plants and other homologous proteins were performed with CLUSTAL W (Thompson et al. 1994). The evolutionary relationship between PPi-PFK and ATP-PFK sequences was reconstructed using the maximum parsimony method (heuristic algorithm), as implemented in PAUP v.4.0b10 (Swofford 1998). The bootstrap analysis consisted of 1,000 replicates. The sequences used in the evolutionary analyses were obtained from the GenBank and SwissProt databases.

Results availability

The results of the annotations of TE-cassettes within ESTs and unigenes of the three *Coffea* species analyzed using the RepeatMasker are available online at http://www.lge.ibi.unicamp.br/cafe—Transposable elements icon.

Results

A set of 131,150 ESTs of *Coffea arabica*, 50,225 of *Coffea canephora*, and 10,566 of *Coffea racemosa* was compared to the RepBase TE database using the RepeatMasker. A total of 491 candidate sequences with TE-cassettes were identified in the three *Coffea* species. However, 108 sequences with TEs constituting more than 70% of an EST were not included in this analysis, because they were considered as putative transcriptionally active TEs rather than TE-cassettes in protein-coding regions of host genes. Thus, a set consisting of 383 sequences (0.18% of the total of 191,150 ESTs) with a TE-cassette within the EST was obtained. The screening of the 39,312 unigenes of *Coffea*

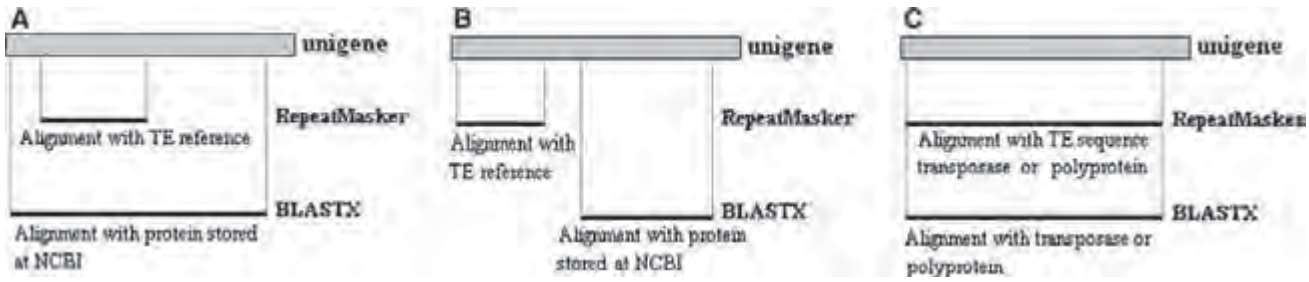


Fig. 1 Possible results of TE localization by RepeatMasker (*Coffea* unigenes database against RepBase plant TE bank) and the region of similarity of the transcript to the protein by BLASTX (*Coffea* unigene database against NR database). **a** The TE-cassette overlaps perfectly with the protein sequence stored at NCBI; **b** TE fragment aligns with the unigene outside the protein sequence; **c** TE sequence aligns with a transposase or polyprotein within protein sequence

Table 1 Basic statistics of TE-cassettes identified in three *Coffea* species

TEs' classification	Number of TE-containing ESTs (%)[a]	Average length of TE-cassettes (nt)	Minimum and maximum length of TE-cassettes (nt)	RM score average ± SE
Coffea arabica				
LTR	166 (71.24)	79.7 ± 1.2	36–122	303.7 ± 4.5
LINE	2 (0.85)	149 ± 91	58–240	288.5 ± 37.5
DNA	34 (14.60)	101 ± 9.3	42–299	307.1 ± 14.5
Other	31 (13.30)	99.5 ± 7.58	47–217	338.3 ± 25.8
Total	233 (100)	–	–	–
Coffea canephora				
LTR	72 (51.43)	96.5 ± 7.2	37–322	308.4 ± 7.1
LINEs	1 (0.71)	–	221	397
DNA	46 (32.86)	93.89 ± 10.5	44–212	319.8 ± 10.5
Other	21 (15)	84.2 ± 19.5	25–458	304.3 ± 21.8
Total	140 (100)	–	–	–
Coffea racemosa				
LTR	6 (60)	89.33 ± 4.87	76–110	291.8 ± 20.0
LINE	–	–	–	–
DNA	3 (30)	108.33 ± 7.67	93–116	454.3 ± 32.7
Other	1 (10)	–	89	356
Total	10 (100)	–	–	–

[a] Percentage of ESTs containing TEs

arabica, 17,420 of *Coffea canephora*, and 4,056 of *Coffea racemosa* resulted in 143 unigenes (0.23% of the total of 60,788) with TE-cassettes, after excluding the putative transcriptionally active TEs.

Transposable element-cassettes in the ESTs

From the 383 sequences harboring TE-cassettes, 233 ESTs were obtained in *Coffea arabica*, 140 in *Coffea canephora* and 10 in *Coffea racemosa* libraries. The analyses of copy number and minimum, mean and maximum length of the TE sequences within ESTs are presented in Table 1. A predominance of TE-containing ESTs classified as LTR-retrotransposons (63.7%) was observed in the three species ($P < 0.05$). Transposons comprised 21.4%, LINEs 0.78% and no SINEs were identified. Interestingly, 54 of the TE-containing ESTs (14%) were classified as Helitrons (ATREPX1 family) and MITEs (Stowaway family). The mean length of the TE fragments varied from 79.7 nt (LTR-retrotransposon in *Coffea arabica*) to 149 nt (LINE in *Coffea arabica*). The mean RM score was higher than 288 for any class of TEs of the three *Coffea* species.

Transposable element-cassette frequencies in the *Coffea* EST database were evaluated in 31 cDNA libraries of *Coffea arabica* corresponding to a wide range of tissues (e.g., seeds, embryogenic calli, roots, leaves, flowers), developmental stages and plant material submitted to biotic (e.g., stems infected with *Xylella spp* and nematodes) and abiotic (e.g., water deficit) stress conditions. Eight cDNA libraries of *Coffea canephora* were analyzed, with emphasis on seeds and fruits, while only two libraries of *Coffea racemosa* were evaluated, focusing fruits (Table 2).

The frequency of TE-cassettes in the ESTs varied, depending on the source tissue of the cDNA libraries (Table 2). No TE sequences were identified in ESTs from cDNA libraries of *Coffea arabica* derived from zygotic embryos (immature fruits, EB1 library), germinating seeds (EM1), plantlets treated with arachidonic acid (LP1), and root tissues (RT3).

Transposable element-cassettes in unigenes

Of the set of 143 unigenes harboring TE-encoded fragments, 72 were registered in *Coffea arabica*, 66 in *Coffea canephora* and five in *Coffea racemosa* libraries. The analyses of minimum, mean and maximum length of the TE sequences within unigenes, as well as the mean percentage occupied by a TE-cassette in the CDS, are presented in Fig. 2.

In the present study, TEs were inserted in a 5′–3′ orientation with regard to the host gene sequence, as well as in the opposite orientation (Table 3). Elements of the Copia/Ty1 superfamily were preferentially inserted in antisense orientation in the three species ($P < 0.05$). On the other hand, Gypsy/Ty3 elements were preferentially inserted in sense orientation in *Coffea arabica*, and the same was observed for transposons in *Coffea arabica* and *Coffea canephora* ($P < 0.05$), but the numbers of unigenes with inserted TEs are very small and therefore these differences should be considered with caution.

Table 2 Description of libraries and occurrence of different TE classes in 191,921 ESTs in the transcriptome from three *Coffea* species

Library code	Tissue/developmetal stage	Number of accepted reads	TE frequency (%)	TEs' classification			
				LTR	LINEs	DNA	Others
Coffea arabica							
AR1	Leaves treated with arachidonic acid	2,364	0.084	2	–	–	–
BP1	Suspension cells treated with acibenzolar-S-methyl	6,218	0.11	7	–	–	–
CA1	Non-embryogenic calli	4,278	0.18	7	–	1	1 (HE)
CB1	Suspension cells treated with acibenzolar-S-methyl and brassinosteroids	8,237	0.19	14	–	1	1 (HE)
CL2	Hypocotyls treated with acibenzolar-S-methyl	9,299	0.14	6	–	5	1 (HE) 1(ST)
CS1	Suspension cells treated with NaCl	7,804	0.24	14	1	1	2 (HE) 1(ST)
EA1	Embryogenic calli	4,847	0.22	10	–	1	–
EB1	Zygotic embryo (immature fruits)	138	0	–	–	–	–
EM1	Germinating seeds—zygotic embryos	139	0	–	–	–	–
FB1	Flower buds in stages 1 and 2—long	7,537	0.21	10	–	3	3 (HE)
FB2	Flower buds in stages 1 and 2—short	5,429	0.27	10	–	3	2 (HE)
FB4	Flower buds in stages 3 and 4—short	2,705	0.18	3	–	1	1 (HE)
FR1	Flower buds no 6, pinhead fruits no 1 and fruits (stages 1 and 2)—long	3,369	0.17	3	–	2	1 (HE)
FR2	Flower buds no 6, pinhead fruits no 1 and fruits (stages 1 and 2)—short	5,474	0.31	13	–	2	2 (HE)
IA2	Embryogenic calli with 2,4 D	2,052	0.09	2	–	–	–
IC1	Non-embryogenic calli without 2,4 D	2,227	0.22	5	–	–	–
LP1	Plantlets treated with arachidonic acid	2,301	0	–	–	–	–
LV4	Young leaves from orthotropic branch —long	4,742	0.12	3	–	1	2 (HE)
LV5	Young leaves from orthotropic branch—short	6,724	0.14	5	–	1	3 (HE) 1(ST)
LV8	Mature leaves from plagiotropic branches—long	7,320	0.15	8	1	1	1 (HE)
LV9	Mature leaves from plagiotropic branches—short	3,017	0.10	3	–	–	–
NS1	Roots infected with nematodes *Meloidogyne paranaensis*	321	0.31	1	–	–	–
PA1	Primary embryogenic calli	1,761	0.17	3	–	–	–
PC1	Non-embryogenic calli with 2,4 D	2,204	0.22	3	–	–	2 (HE)
RM1	Leaves infected with leaf miner (*Perileucoptera coffeella*) and coffee leaf rust	3,512	0.02	1	–	–	–
RT3	Roots	356	0	–	–	–	–
RT5	Roots with acibenzolar-S-methyl	1,507	0.06	1	–	–	–
RT8	Suspension cells stressed with aluminum	5,074	0.13	6	–	1	–
RX1	Stems infected with *Xyllela spp*	7,117	0.14	8	–	1	1 (HE)
SH2	Water deficit stresses field plants (pool of tissues)	5,185	0.23	7	–	4	1 (HE)
SI3	Germinating seeds—whole seeds	7,493	0.26	11	–	5	4 (HE)
Subtotal				166	2	34	31
Total		131,150	0.14 ± 0.01			233	
Coffea canephora							
EC1	Embryogenic calli	6,738	0.28	10	–	7	2 (HE)
LF1[a]	Leaf—young	6,448	0.24	10	–	4	2 (HE)
PP1[a]	Pericarp—all developmental stages	8,695	0.48	29	1	7	5 (HE)
SE1[a]	Whole cherries—22 week after pollination	1,001	0.19	2	–	–	–
SE2[a]	Seeds—18 week after pollination	6,390	0.21	3	–	10	1 (HE) 3 (ST)
SE3[a]	Endosperm and perisperm of seeds—30 week after pollination	9,097	0.20	9	–	7	3 (ST)
SE4[a]	Endosperm and perisperm of seeds—42 and 46 week after pollination	6,292	0.23	5	–	7	1 (HE) 2(ST)
SH3	Leaves from water deficit stresses plants	5,594	0.17	4	–	4	2 (HE)
Subtotal				72	01	46	21
Total		50,255	0.25 ± 0.03			140	

Table 2 continued

Library code	Tissue/developmetal stage	Number of accepted reads	TE frequency (%)	TEs' classification			
				LTR	LINEs	DNA	Others
Coffea racemosa							
FV2	Fruits, stages 1, 2 and 3	5,080	0.13	3	–	3	1 (HE)
FR4	Fruits	5,436	0.05	3	–	–	–
Subtotal				6	–	3	1
Total		10,516	0.09 ± 0.04			10	

Library code: library name; Tissues/developmental stages: tissues description, developmental stages or stress conditions of plant material analyzed; Number of accepted reads: number of ESTs studied in each library after elimination of low quality sequences; TE frequency: the percentage of ESTs with TE-cassettes insertion in each library; TEs' classification: occurrence of different classes of TEs in each library

HE Helitron, *ST* Stowaway

[a] Libraries produced by Lin et al. (2005)

The function of TE-cassette transcripts and their putative occurrence in the proteome

The functional annotation of 143 unigenes containing inserted TEs was obtained by BLASTX analysis against the NR (non-redundant) database. As a result, 89 unigenes presented a match to some proteins of the NR bank; 29 of these contained TE-cassettes within the region of similarity between the *Coffea* unigene and the protein (Fig. 1a; S1), and 60 unigenes had TE-cassettes inserted outside this region (Fig. 1b). Sixty-two of the 89 unigenes were similar to proteins with known molecular function and were assigned to functional groups according to *the Gene Ontology* nomenclature or to the LocusLink descriptions. The largest group of *Coffea* transcripts with TE-cassettes was represented by enzymes, 12 cases for *Coffea arabica* and *Coffea canephora* and one case for *Coffea racemosa*, and the enzymes were also the class of transcripts with a greater number of ESTs (100 for *Coffea arabica*, 16 for *Coffea canephora* and 1 for *Coffea racemosa*; Table 4). However, the mean number of ESTs contained in each unigene for each molecular function group was higher for chaperone proteins in *Coffea arabica* (30.5) and for transporter proteins in *Coffea arabica* and *Coffea canephora* (17 and 2.6, respectively; Table 4). The greater occurrence of enzyme transcripts containing TE-cassettes had already been reported (Lorenc and Makalowski 2003), but without any explanation for this relationship. Since enzymes act on several targets and TE-cassettes are sources of variability, we could hypothesize that they may represent a genomic plasticity essential to adaptation to an ever-changing environment.

Analysis of the length, fraction of the CDS occupied by TEs and RM score was performed, in order to assess differences between the unigenes containing TE-cassettes within (according to Fig. 1a) and outside (according to Fig. 1b) the perfect overlapping between the localization of the cassette in the unigene and the protein sequence from the NR database. The TE-cassettes within the region of similarity with a protein are about 23% longer (100.4 ± 11.4) than those from the outside region (77.6 ± 3.34), occupy a two times larger fraction of a CDS (17.3 ± 3.4 against 8.6 ± 0.5), and are more similar (327.1 ± 19.3 against 291.6 ± 7.6) to the reference TEs.

Exons origin from TE-cassettes and gene fragments intercepted by TEs

Three of the 29 identified proteins containing putative inserted TE-cassettes in the *Coffea* transcriptome did not represent true TE-cassettes, but rather a TE-mediated transduction of host gene sequences: the homeobox-leucine zipper protein in *Coffea arabica* and *Coffea canephora* (GenBank gi no 15232122) that matched to ATMU6N1 (Kapitonov and Jurka 2001a), the phosphoenolpyruvate translocator precursor protein (PPT) in *Coffea arabica* (GenBank gi no 7489171) that matched to ATDNA1T9A (Kapitonov and Jurka 2000), both transposons described in *Arabidopsis thaliana*, and the DNA helicase protein in *Coffea canephora* (GenBank gi no 102139878) that matched to Helitron4 of *O. sativa* (Kapitonov and Jurka 2005). Based on the structure of ATMU6N1, it can be inferred that the region of similarity between the mRNA of *Coffea arabica* and the TE is due to a small sequence of exon 2 of the homeobox-leucine zipper gene intercepted by the TE (Fig. 3a). Similarly, the match of *Coffea arabica* mRNA to ATDNA1T9A is associated with a portion of exon 5 of the PPT1 gene captured by this TE (Fig. 3b). In the third case, the alignment between the *Coffea canephora* DNA helicase mRNA and Helitron4 is due to a total sharing of the DNA helicase domain between the host gene and the TE (Fig. 3c).

Two additional proteins containing putative inserted TE-cassettes—FR and PPi-PFK—were investigated with

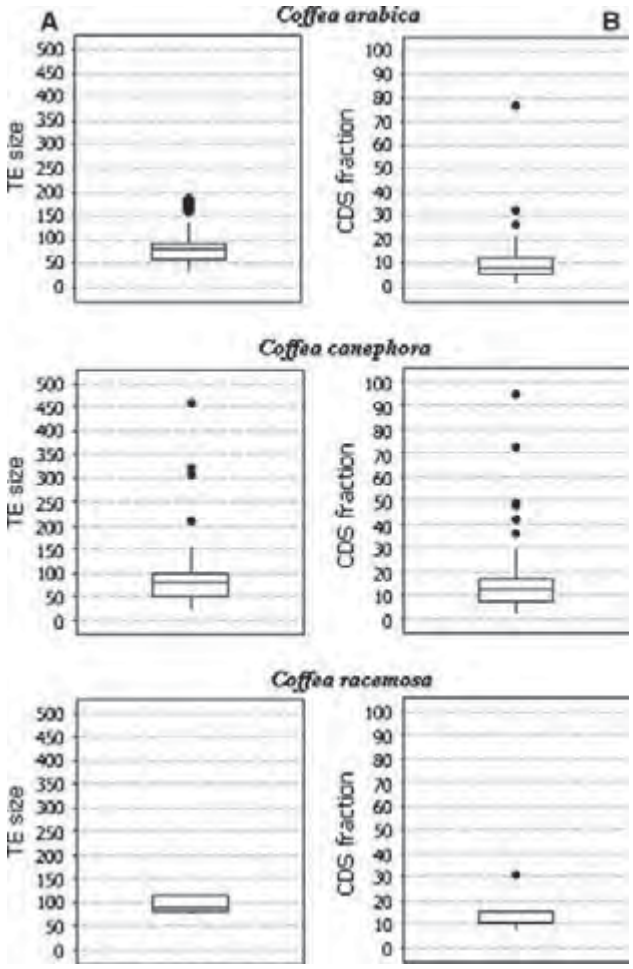


Fig. 2 *Box-plot* of TE-cassettes in unigenes of *Coffea arabica*, *Coffea canephora* and *Coffea racemosa*. **a** Distribution of TE-cassette sizes (in nucleotides) occupied in a CDS. The lengths of TE-cassettes were: *Coffea arabica*: mean = 84.56 ± 3.82 nt and median = 79.50 nt (minimum: 36 nt, maximum: 182 nt); *Coffea canephora*: mean = 96.06 ± 8.59 nt and median = 83 nt (minimum: 25 nt, maximum: 458 nt); *Coffea racemosa*: mean = 93.86 ± 5.79 nt and median = 90 nt (minimum: 76 nt, maximum: 116 nt). **b** Distribution of a CDS fraction covered by a TE-cassette: *Coffea arabica* = 10 ± 1.13%; *Coffea canephora* = 16 ± 1.90% and *Coffea racemosa* = 14 ± 3.03%. For each category, the *central box* depicts the middle half of the data between the 25th and 75th percentiles, and the *internal lines* indicate the median of each distribution. The *circles* that fall outside the main body of data in each distribution indicate extreme values

regard to the relationship between the TE and the host gene sequences. This analysis, using the gene models of *Arabidopsis thaliana*, *O. sativa*, *R. communis* and *Populus trichocarpa*, showed that the FR gene had exons similar to *copia*-like retrotransposons; on the contrary, the exons of PPi-PFK were putatively captured by an ATREPX1 Helitron.

Regarding the FR genes, 21 models were identified in the genome of *Arabidopsis thaliana*, 10 models in the genome of *O. sativa* and 2 in the genome of *Populus trichocarpa*. All these genes and their transcripts were annotated de novo by RepeatMasker and the results showed that only one gene of *O. sativa*, the one with the greatest length (11,618 bp; Gene ID: 4329334), presented the RIRE5 element composing exons 4 and 5 and intron 4, a *copia*-exon association, as described in the *Coffea* unigene (Fig. 4).

Regarding the PFK genes, all the models from *Arabidopsis thaliana*, *O. sativa*, *Populus trichocarpa* and *R. communis* (Table 5) were obtained and classified as full-length or truncated genes and PPi-PFK beta or alpha subunit coding genes, as well as genes similar to PFKs from the bacteria *Amycolatopsis methanolica*. Again, all these genes and their transcripts were annotated de novo by the RepeatMasker. A very interesting relationship between the full-length PPi-PFK β-subunit gene and the ATREPX1 Helitron was observed: the TE-exon association of exons 9–10 and intron 9 with ATREPX1, as described in the *Coffea* unigene (Fig. 5a and S1), being supported by a high-score RM (Fig. 4). Two possibilities could explain such an association: (1) a molecular domestication of ATREPX1 that might have originated these new exons; or (2) a transduction event of the host exon, mediated by the TE. In view of these two possibilities, several analyses were performed.

The first analysis was aimed at comparing the ATREPX1 sequence with PPi-PFK β-subunit genes in flowering plants. It showed similarities between the TE and the PPi-PFK β-subunit genes of *Arabidopsis thaliana* (0.86 and 0.68, respectively for GeneID: 837752 and 825716), *O. sativa* (0.68), *R. communis* (0.72), *Populus trichocarpa* (0.70), *Medicago truncatula* (0.70) and transcripts of *Solanum tuberosum* (0.69), *Citrus paradisi* (0.72), *Coffea arabica* (0.70) and *Z. mays* (0.66). The proteins encoded by these genes are highly conserved, even regarding exons 9 and 10, which present similarity to ATREPX1. This sequence conservation is illustrated by a highly supported polytomic clade (bootstrap 99%), obtained in a phylogenetic analysis by parsimony (Fig. 6) and neighbor-joining (tree not shown), and also by the alignment between in PPi-PFK β-subunit-related protein sequences (Fig. 7). An additional alignment was performed with the PPi-PFK protein of *Arabidopsis thaliana* against the translated ATREPX1 sequences (Fig. 5b), and two stop codons were observed in the element, which are absent in the host protein.

In the second analysis, all the PPi-PFK β-subunit genes and transcripts were compared against Helitrons from *Arabidopsis thaliana* and *O. sativa* (38) under low stringency conditions (BLASTN, $E = e - 02$, $V = 10{,}000$, $B = 10{,}000$), and matches were observed only for ATREPX1. Finally, an alignment was performed of these PPi-PFK proteins from plants, of other homologous proteins from bacteria (*Chlamydia trachomatis*, *Treponema pallidum* and *Borrelia burgdoferi*) which have been considered to have received a copy of this gene by horizontal transfer from plants (Stephens et al. 1998), and of two other bacteria (*Treponema*

Table 3 Characterization of the TE-cassettes according to their relative orientation to the host gene sequences

Superfamily	*Coffea arabica*		*Coffea canephora*		*Coffea racemosa*	
	Sense *N* (%)	Antisense *N* (%)	Sense *N* (%)	Antisense *N* (%)	Sense *N* (%)	Antisense *N* (%)
Copia/Ty1	12 (28)	31 (72)*	2 (13.3)	13 (86.6)*	1 (33)	2 (66)*
Gypsy/Ty3	3 (75)	1 (25)*	12 (57.1)	9 (42.9)	–	–
Transposons	14 (63.6)	8 (36.3)*	12 (63.1)	7 (36.9)*	–	1
Helitrons	2 (66.6)	1 (33.3)*	5 (55.5)	4 (44.5)	–	1
Stowaway	–	–	2	–	–	–
Subtotal	31 (43)	41 (57)	33 (50)	33 (50)	1 (20)	4 (80)*
Total	72 (100)		66 (100)		05 (100)	

χ^2 for test of homogeneity
*$P < 0.05$

Table 4 Categories of TE-containing transcripts in three *Coffea* species

Categories	*Coffea arabica*		*Coffea canephora*		*Coffea racemosa*		Total (%)[b]
	No. unigenes[a] (%)[b]	No. ESTs[c] (median)[d]	No. unigenes[a] (%)[b]	No. ESTs[c] (median)[d]	No. unigenes[a] (%)[b]	No. ESTs[c] (median)[d]	
Enzyme	12 (40.0)	100 (8.3)	12 (40)	16 (1.3)	1 (100)	1 (1)	25 (41.0)
Nucleic acid binding	5 (16.7)	17 (3.4)	6 (20)	15 (2.5)	–	–	11 (18.0)
Ligand binding	5 (16.7)	15 (3)	2 (6.7)	14 (7)	–	–	7 (11.4)
Chaperone	2 (6.7)	61 (30.5)	–	–	–	–	2 (3.2)
Transporter	2 (6.7)	34 (17)	3 (10)	8 (2.6)	–	–	5 (8.2)
Defense immunity protein	1 (3.3)	2 (2)	–	–	–	–	1 (1.6)
Regulatory enzyme	1 (3.3)	9 (9)	–	–	–	–	1 (1.6)
Apoptosis regulator	1 (3.3)	2 (2)	–	–	–	–	1 (1.6)
Structural protein	1 (3.3)	23 (23)	6 (20)	13 (2.1)	–	–	7 (11.4)
Pathogenesis	–	–	1 (3.3)	1 (1)	–	–	1 (1.6)
Function unknown	15	87 (5.8)	9	16 (1.7)	2	4 (2)	26
Total	45	100	39	100	3	100	87 (100)

[a] Number of TE-cassettes containing unigenes for each molecular function
[b] Percentages calculated excepting the transcripts with function unknown
[c] Number of ESTs contained in each unigene to each group of molecular function
[d] Mean number of ESTs for each unigenes

denticola and *Spirochaeta thermophila*) in which such a transfer has not been reported yet. This alignment showed high conservation of this region between bacteria and plants (Fig. 7).

Discussion

The annotation of TEs into the protein-coding region in *Coffea arabica, Coffea canephora* and *Coffea racemosa* performed in this study, using ESTs and unigenes, showed that a substantial number of *Coffea* transcripts harbor insertions of TEs, not only LTR-retrotransposons and transposons, but also MITEs and Helitrons, which contribute to mRNA in plants and possibly to protein diversity. Such a contribution had been described more than a decade ago by Makalowski et al. (1994), who characterized in detail the mechanism of an Alu element integration into part of the human mRNA. Later on, Nekrutenko and Li (2001) registered a large-scale annotation of TE-cassettes in all publicly available human unigenes, and Lorenc and Makalowski (2003) analyzed vertebrate protein diversity by TE insertions into ORFs. Sakai et al. (2007) registered the first annotation of TEs into the CDS in protein-coding genes from plants (*O. sativa*), using computationally predicted ORFs sequences as query.

Occurrence of TE-cassettes in ESTs and unigenes of *Coffea*

In this study, we firstly registered the annotation of TE-cassettes in expressed partial sequences from *Coffea arabica*, *Coffea canephora* and *Coffea racemosa*, showing the occurrence of 383 ESTs (0.18%) with TE fragments. Analysis of biased distribution of ESTs with TE-cassettes

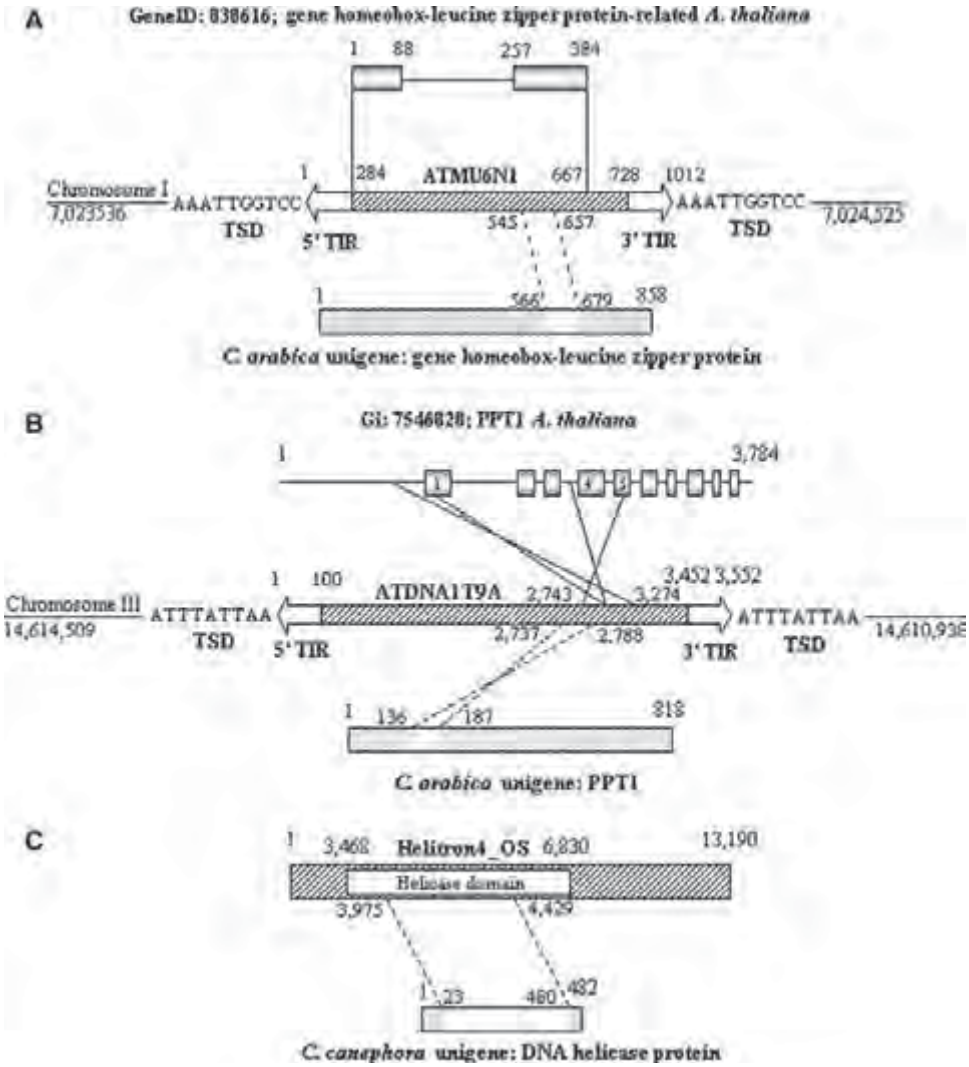


Fig. 3 Relationships between *Coffea* mRNA and genes captured by different TEs. **a** ATMU6N1 harbors the full sequences of exons 1–2 and intron 1 of the ortholog Athb-1 gene of *Arabidopsis thaliana* (position 284–657); **b** ATDNA1T9A harbors a partial sequence of the PPT1 host gene (position 2743–3274) corresponding to a partial region of exon 1, intron 3 and exon 5, and to a complete exon 4 and intron 4 of *Arabidopsis thaliana*; **c** DNA helicase gene transduced by a Helitron 4 of *O. sativa*. At the top of each drawing, *gray boxes* represent exon regions and lines correspond to introns. *Hatched boxes* and the *inverted arrows* represent internal region and ITR of the TE, respectively (*TSD* target site duplications). *Inclined continuous lines* correspond to regions of the gene captured by the TE and non-continuous lines to regions of similarity of the *Coffea* mRNA and the TE

amongst different libraries (varied tissues, developmental stages and response to biotic and abiotic stressful conditions) was not possible, due to the absence of a reference frequency value. The absence of TE sequences in ESTs of cDNA libraries from *Coffea arabica* derived from zygotic embryos (immature fruits), germinating seeds, plants treated with araquidonic acid and root tissues might reflect small sampling size.

Moreover, the 140 unigenes (0.23%) with TE fragments identified in a total of 60,788 unigenes of the three *Coffea* species suggest a low frequency of TE-cassettes in protein-coding regions when compared to the 533 UniGenes (3.87%) of *Homo sapiens* (Nekrutenko and Li 2001) and the 439 CDS (2.2%) of *O. sativa* (Sakai et al. 2007). However, a low rate of TE-derived sequences in exons has also been observed in *Mus musculus*: out of 186,823 exons analyzed, only 263 (0.14%) showed LTR insertions (DeBarry et al. 2005). Similarly low frequencies have been observed in *Caenorhabditis elegans*: out of 19,000 genes, 35 (0.18%) presented LTR insertions (Ganko et al. 2003); in *Drosophila melanogaster*, out of 13,300 genes, 25 (0.18%) showed LTR insertions (Ganko et al. 2006); and finally, in *Bos taurus*, out of 22,805 genes analyzed, only 630 exons (0.12%) presented TE insertions (Almeida et al. 2007).

The differences in frequencies of TE fragments in the protein-coding regions of the human and the genomes of other vertebrates, invertebrates, rice and *Coffea* can be explained by two reasons: (1) the frequency of cassettes could be species-specific; or (2) we and the above cited authors have applied more stringent criteria. In our study, the second option seems more likely, since we applied a strict RM score (>250), and the mRNA sequence coding for active TEs and the matches between *Coffea* mRNA and exons of host genes intercepted by the TEs were eliminated. Therefore, false-positives are unlikely. On the other hand, these criteria may have resulted in loosing some legitimate TE-cassettes. However, it is worth pointing out that all the above cited analyses were done in silico and might not represent the true contribution of TEs to proteomes.

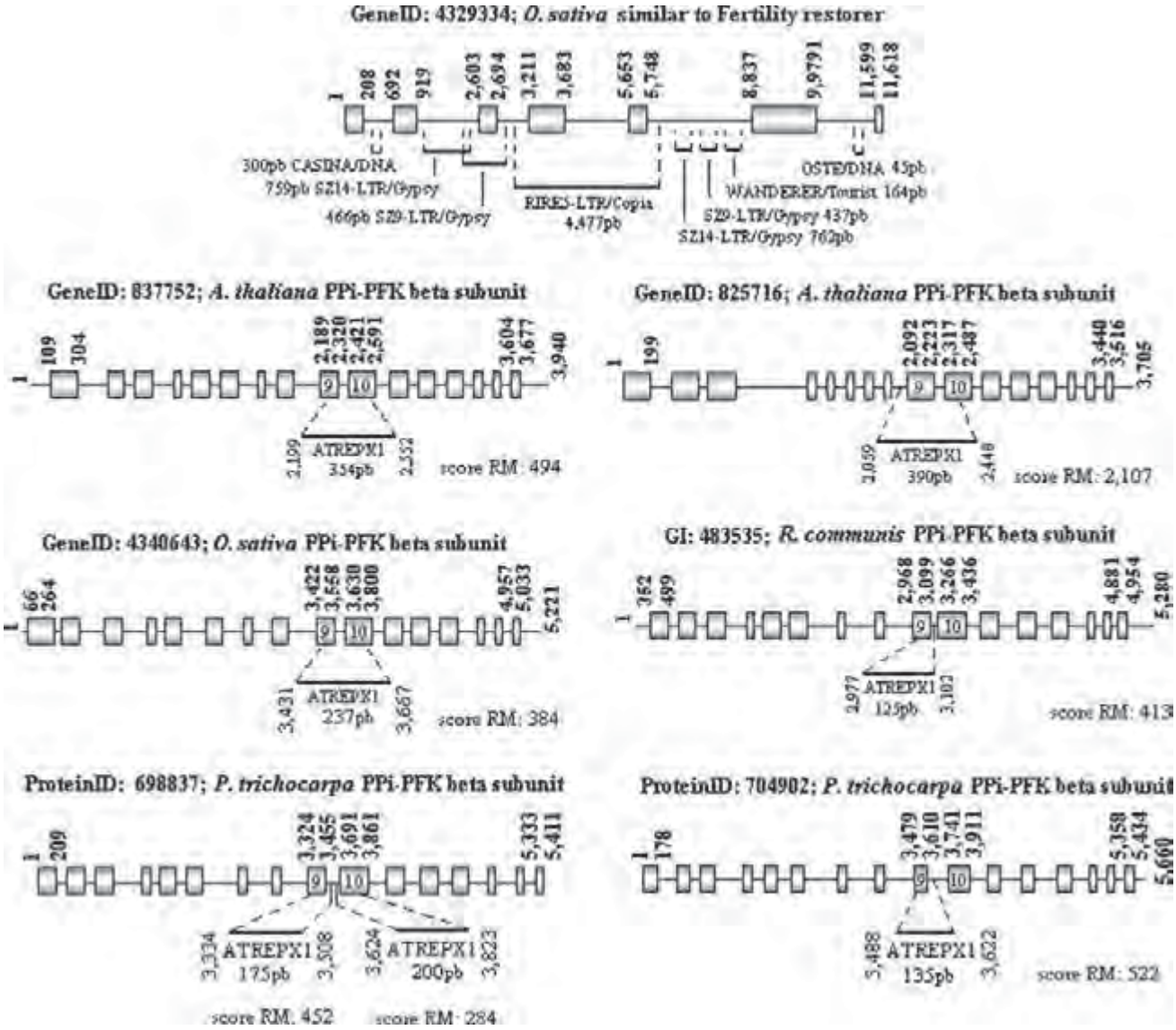


Fig. 4 Schematic representation of the fertility restorer (*FR*) and pyrophosphate-dependent phosphofructokinase (*PPi-PFK*) genes, with exon and intron sequences similar to TEs. The FR gene structure is similar to various TEs, and PPi-PFK presents regions spanning exons 9–10 similar to Helitrons. *Squares* represent exons and *lines* correspond to introns. *Numbers* show the beginning and the end of exons, the length of the TEs and the RM score

Putative presence of TE-encoded fragments in the proteome: discrepancy between the frequency of occurrence in transcripts and in proteins

The impact of TEs on protein-coding regions is of particular interest because they can directly influence the phenotype by altering the protein sequence. This aspect has been well documented only at the transcriptional level in the human genome (Li et al. 2001; Nekrutenko and Li 2001; Lorenc and Makalowski 2003). However, a few TE-encoded fragments have been confirmed at the protein level (Gerber et al. 1997; Hilgard et al. 2002; Hoenicka et al. 2002), possibly because insertion of TEs within coding protein regions is frequently associated to deleterious effects and consequently to some degree of loss of function (e.g., Deininger and Batzer 1999).

Putative *Coffea* proteins have been shown to harbor fewer TE-cassettes (~0.04%) than would be expected from the translation of TE-containing transcripts (0.23%). The cassettes within the region of similarity to the protein could mean new exaptation events. However, the fact that the TE-cassettes are longer, occupy a larger fraction of a CDS, and are more similar to the reference-TEs than those outside this region suggests that they are recent insertions, thus they could be tolerated as an alternatively spliced form of cognate mRNAs that does not encode functional proteins. Another possibility would be that those transcripts harboring large cassettes are encoded by some

Table 5 Identification of the gene models of PFKs in the genomes used in this study

Genome	Number of gene models of PFKs			Other PFK[e]
	Total	PPi-PFK beta-subunit[a]	PPi-PFK alpha-subunit[a]	
Arabidopsis thaliana	11	2 (825716, 837752)[b]	2 (838689, 843988)[b]	7
Oryza sativa	13	1 (4340643)[b]	3 (4330512, 4340909, 4345338)[b]	–
Populus trichocarpa	16	2 (698837, 704902)[c]	2 (550714, 73215)[c]	2
Ricinus communis	2	1 (483535)[d]	1 (483546)[d]	–

[a] Number of full-length gene models in relation to total

[b] GeneID in the LocusLink

[c] ProteinID in the *Populus trichocarpa* genome database

[d] Gene identifier in the GenBank

[e] Genes similar to PPi-PFK from bacteria *Amycolatopsis methanolica*

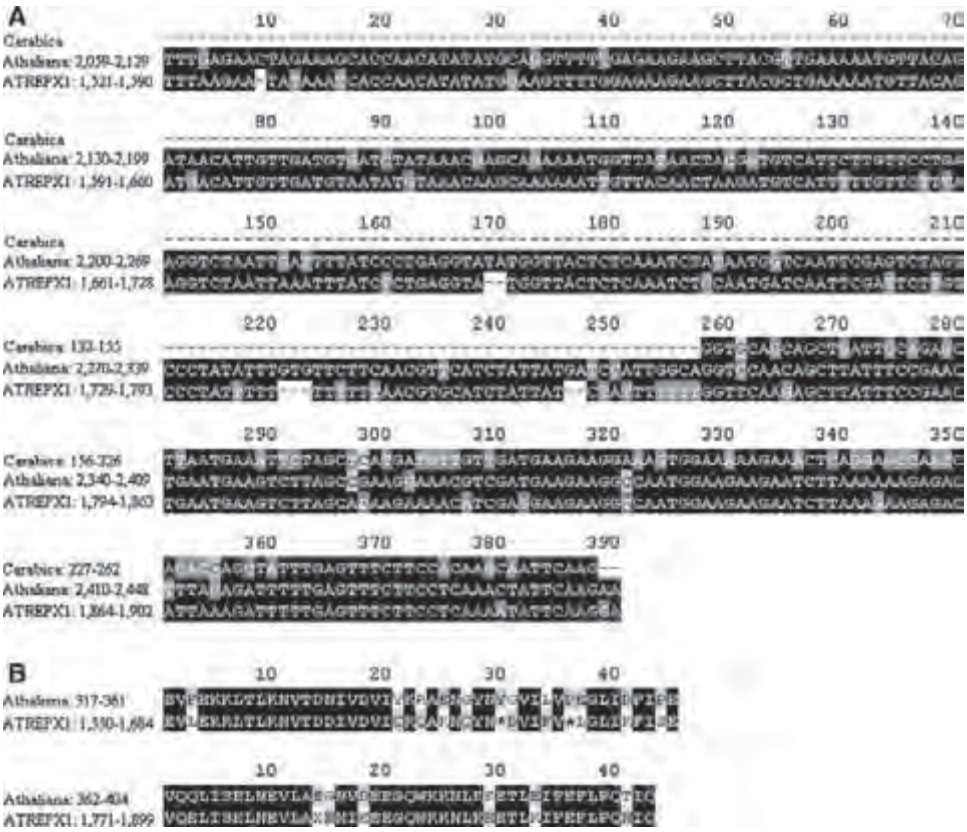


Fig. 5 Alignments of the PPi-PFK β-subunit sequences and the ATREPX1 element. **a** Nucleotide sequences of *Coffea arabica* unigene, of *Arabidopsis thaliana* gene (GeneID: 825716) and ATREPX1. The alignment encompasses 130 nt of *Coffea* transcript and 390 nt of *Arabidopsis thaliana* gene, and the identity to ATREPX1 is 70% for *Coffea* PPi-PFK and 86% for *Arabidopsis thaliana* PPi-PFK. **b** Partial protein of *Arabidopsis thaliana* and translated ATREPX1. The PFK sequences are shown in the *top line* of the alignment and the consensus sequence of ATREPX1 (ATREPX1#Other) stored in RepBase. These sequences and their coordinates were taken from the libraries provided by RepeatMasker or bl2seq programs, respectively

members of a certain gene family, and that these do not influence gene function, since the non-modified members would maintain the function. These multiple copies could provide the raw material for evolutionary "innovations", since some of them can be free of functional constraints and undergo significant changes, such as the exonization/exaptation of TE fragments and the origin of a new specific function (Ohno 1970; Gotea and Makalowski 2006). Hence, recently inserted large cassettes could acquire biological function if they had enough time to evolve (Hayashi et al. 2003) and the coding sequence be positively selected.

Impact of TE-cassettes in the host gene evolution

Transposable elements have been denigrated for a long time as poorly selfish sequences acting as parasites in the genome of living organisms (Doolittle and Sapienza 1980; Orgel and Crick 1980). However, this view has meanwhile been considerably challenged by way of new information demonstrating the importance of TEs in the evolution and function of genes and genomes (Brosius 2003). It has been reported that several functional genes harbor sequences that have been almost completely derived from mobile elements, such as in the genomes of

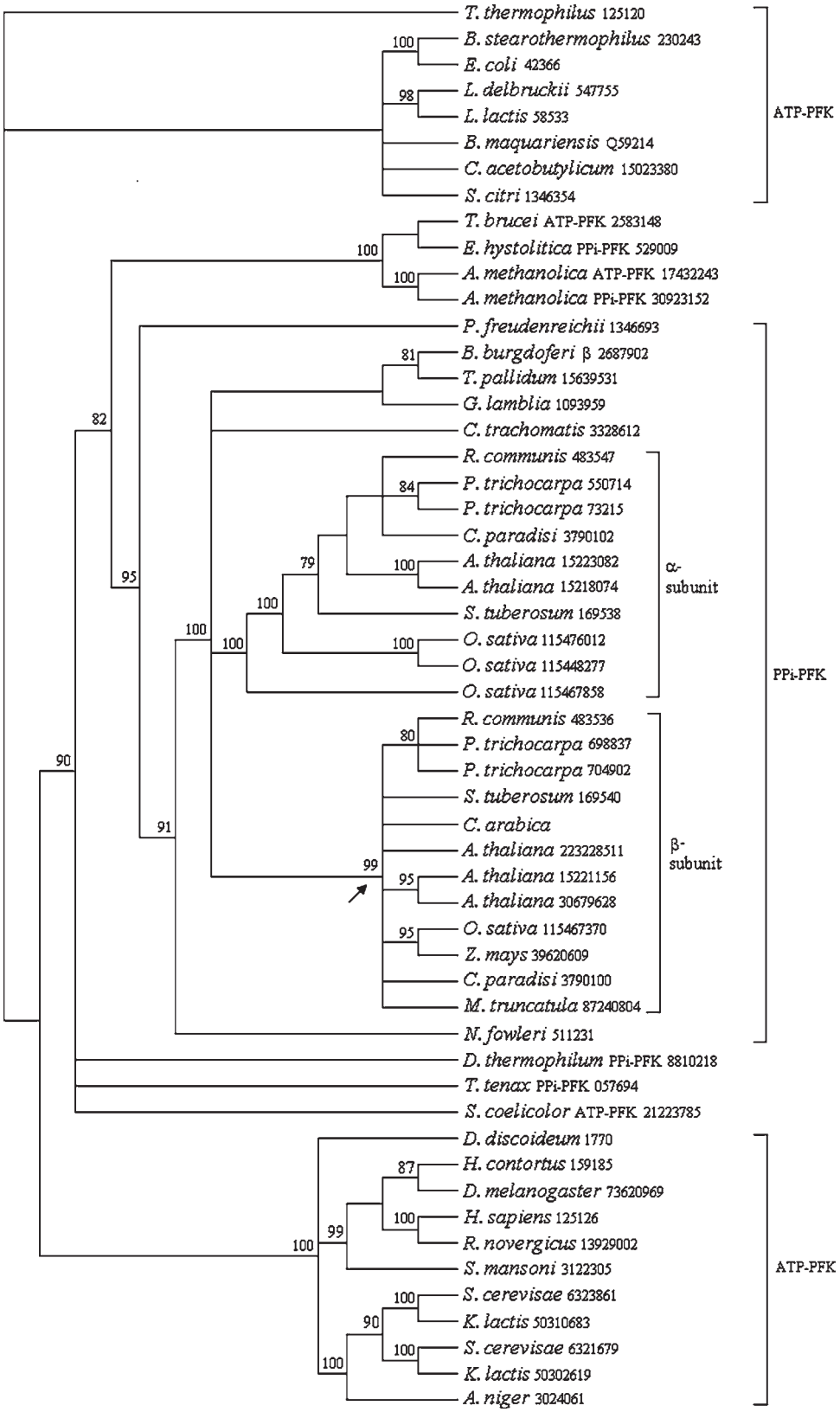


Fig. 6 Unrooted phylogenetic tree of PFK proteins. The cladogram was generated by parsimony analysis using the heuristic algorithm. The PPi- and ATP-PFK sequences were obtained from the GenBank and SwissProt databases and are identified by the host names and GenInfo Identifier (gi). Of 1,122 characters, 179 were uninformative and 845 were parsimony-informative. The consistency index was 0.67, and the retention index was 0.74. The *numbers* indicate the branch support calculated by bootstrap analysis that consisted of 1,000 replicates (over 70%). The *arrow* represents the clade that contains all PPi-PFK *β*-subunit-related sequences

H. sapiens (Britten 2004), *Bos taurus* (Almeida et al. 2007) and *O. sativa* (Sakai et al. 2007). All TE-cassettes detected so far, either as full-exons or those that originate multiple exons, are in agreement with the hypothesis of indirect recruitment of an intronic TE insertion (Nekrutenko and Li 2001).

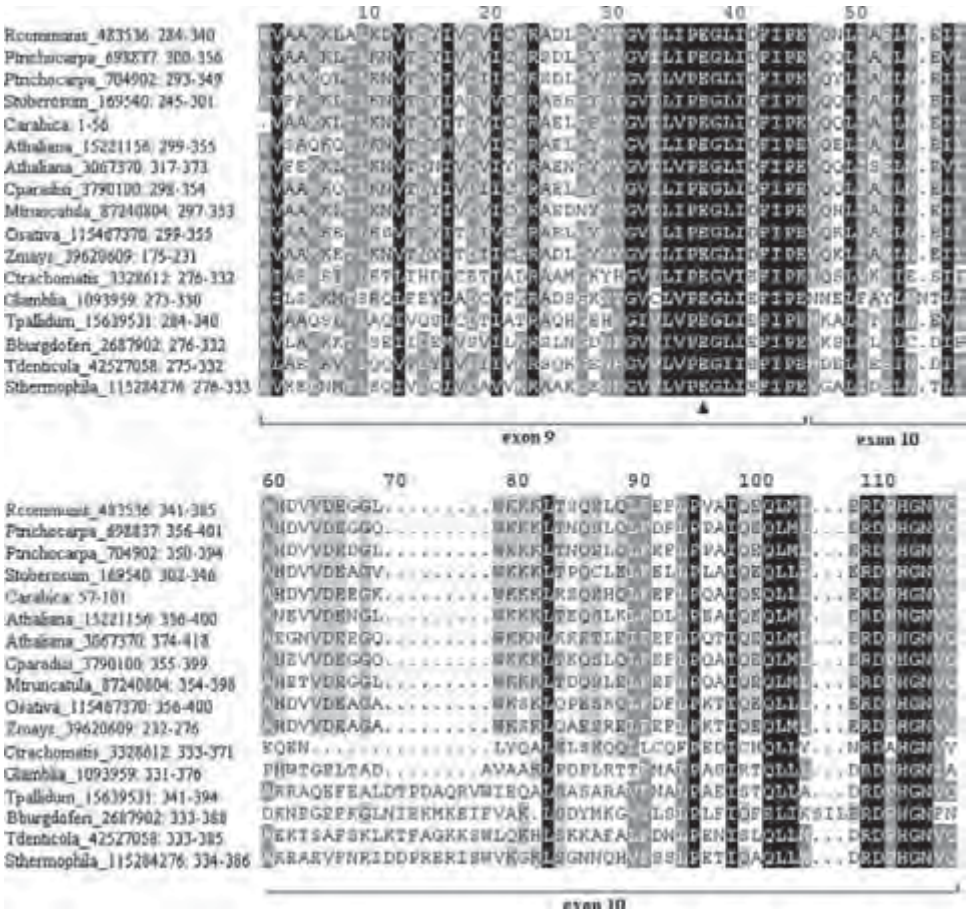


Fig. 7 Multiple alignment between sequences of PPi-PFK homologous proteins using exons 9 and 10. Species names, gi accession numbers and coordinates of residues included in alignment are indicated before every sequence. Identical and similar residues are highlighted in *black* and *gray*, respectively. The symbol (*dark filled triangles*) below the alignment corresponds to conserved active site residues (based on Moore et al. 2002)

One of the *Coffea* unigenes analyzed for evaluating the impact of TE-encoded fragments on the host gene evolution corresponds to genes from other plant genomes similar to the FR gene. This gene encodes a nuclear factor that suppresses the effect of a mitochondrial ORF that in turn encodes a hydrophobic protein capable in some way of altering mitochondrial function, leading to pollen abortion and consequently to infertility (reviewed in Andrés et al. 2007). This analysis showed that only one out of a total of seven genes of *O. sativa* presented TE-cassettes composing three entire exons. Likewise, this association was found in just one out of a total of seven unigenes identified in the Brazilian Coffee Genome Project database (S1). It is probable that this association does not represent a successful exaptation for a protein-coding region, because it was found in just one copy of the *O. sativa* FR genes and the *Coffea* transcripts (results not shown). This finding reinforces the idea that some gene copies can be modified by a TE insertion, without damaging the gene function of the ancestral non-modified sequences. Instead of inferring new impacts of TEs on the host gene structure without at least evaluating the other non-modified copies which possibly encode the functional protein, as has been done so far, we considered it essential to evaluate the occurrence of TEs in members of a gene family, for a robust indication of new domestication events.

Transposable elements-mediated transduction of host gene fragments

The analysis of TE-cassettes taking part in protein-coding regions of *Coffea* genomes was the focus of the present study. Additionally, we were also able to revisit three cases (ATMU6N1, ATDNA1T9A and Helitron4) of host gene fragments that had been captured by TEs and to propose a new one (ATREPX1). Of the revisited cases, transposon ATMU6N1 of the MuDR superfamily, that harbors the homeobox-leucine zipper gene, was highlighted because the capturing of gene fragments has been previously reported for *Mutator*-like elements such as Pack-MULEs in rice, which carry over 3,000 fragments of cellular genes (Jiang et al. 2004). A more obvious case of transduction is the Helitron4 of *O. sativa* and *Coffea* DNA helicase transcripts (S1), whose gene domains are shared by the TE and the host gene. It has been widely recognized that Helitrons or Rolling-circle transposons, which are abundant in the genomes of plants, cnidarians, invertebrates, worms, fish and mammals, harbor genes with enzymatic activities,

including the DNA helicase gene, encoded by autonomous copies (Kapitonov and Jurka 2001b; Kapitonov and Jurka 2003; Lal et al. 2003; Poulter et al. 2003; Arkhipova and Meselson 2005; Zhou et al. 2006; Pritham and Feschotte 2007). Gene transduction mediated by TEs is due to the intrinsic capacity of TE mobilization that can lead to deep changes in genome organization and has an important evolutionary impact on gene function (Eickler and Sankoff 2003; Messing et al. 2004; Biemont and Vieira 2006).

We present here an original description of TE-mediated transduction of the PPi-PFK gene, which presents similarity to an ATREPX1 sequence, as observed in the *Coffea arabica* PPi-PFK transcript. PFKs are a group of key regulatory enzymes in glycolysis, in which PPi-PFK catalyzes the reversible conversion of fructose 6-phosphate and pyrophosphate (PPi) to fructose 1,6-bisphosphate and inorganic phosphate (Pi), and ATP-PFK catalyzes the same pathway using ATP, but in an irreversible reaction. The PFKs can be divided into two large superfamilies, based on amino acid sequence (Wu et al. 1991; Siebers et al. 1998; Moore et al. 2002). The largest one is the archetypal PFKA family that includes ATP- and PPi-dependent PFKs, and the PFKB family includes ribokinase and other homologous sugar kinases that are distinct in sequence and structure from the PFKA family (Sigrell et al. 1998; Moore et al. 2002).

The PPi-PFKs are found in bacteria, protists and plants and can be divided into two subgroups, based on the size of the encoded polypeptides (Moore et al. 2002). The subgroup of the small subunits (monomers of ~320 amino acids) includes the subunits of *Amycolatopsis methanolica* (Alves et al. 1996), *Trichomonas vaginalis* (Mertens et al. 1998) and *Thermoproteus tenax* (Siebers et al. 1998), which are more closely related to the eubacterial ATP-PFKs. The subgroup of the large subunits (monomers of ~550 amino acids) includes the plant PPi-PFK subunits (Carlisle et al. 1990; Blakeley et al. 1992), some amitochondriate protists, such as *Amycolatopsis methanolica* (Alves et al. 1996) and *Entamoeba histolytica* (Mertens et al. 1998; Deng et al. 1998), and two bacterial groups, the spirochetes *Borrelia burgdorferi* (Deng et al. 1999) and *Treponema pallidum* (Roberson et al. 2000), and *Chlamydia* (Stephens et al. 1998). The large PPi-PFKs in plants are heterotetramers $(\alpha\beta)_2$, and the β-subunit of the above-cited bacteria and protists appears to be derived from plants by horizontal transfer (Stephens et al. 1998).

ATREPX1 is a non-autonomous 2,434-bp Helitron described in *Arabidopsis thaliana* that carries an ATHATN1 insertion (position 276–753), a HAT-like DNA transposon (Kapitonov and Jurka 2001b). The comparison of ATREPX1 with the *Arabidopsis thaliana* genome showed several dozens of copies of this TE and at least 14 of them harboring the complete exons 9–10 and intron 9 of the PPi-PFK gene (data no shown). The analysis also showed similarity between exons of the PPi-PFK β-subunit of *Arabidopsis thaliana*, *O. sativa*, *R. communis* and *Populus trichocarpa*, and sequences of ATREPX1. As pointed out before, this relationship could result from a molecular domestication event or from transduction of PPi-PFK into ATREPX1. The fact that the PPi-PFK of both monocot and dicot is similar to ATREPX1 and that all these sequences form a monophyletic clade could lead us to infer that domestication of ATREPX1 by the gene occurred before the diversification of the flowering plants. Since the alignment of the PPi-PFK protein of *Arabidopsis thaliana* against the translated ATREPX1 sequences showed two stop codons in the element that are absent in the host protein, and taking into consideration that the protein of *Arabidopsis thaliana* is functional (Mustroph et al. 2007) and that exon 9 encodes an active site conserved residue (Moore et al. 2002), it is more likely that ATREPX1 has captured the two exons of PPi-PFK and not the opposite. However, we cannot rule out the possibility that the mutations within ATREPX1 occurred after domestication. On the other hand, the transduction of exons 9–10 and intron 9 can also be suggested by the comparison of all the PPi-PFK β-subunit sequences with Helitrons from *Arabidopsis thaliana* and *O. sativa* which resulted in matches of PPi-PFKs only with ATREPX1. Finally, in order to assume TE domestication and considering the high conservation of these exons between bacteria of several genera and plants, we had to suppose that the exaptation had occurred in the dicot/monocot ancestor and that thereafter the gene with the domesticated TE was several times horizontally transferred to bacteria and protists. Although this hypothesis depends on multiple horizontal transfer events, it is not completely unrealistic, since several genes of plants have been transferred to bacteria (Möhlmann et al. 1998; Tjaden et al. 1998; Stephens et al. 1998), among them the PPi-PFK β-subunit, that was transferred from plants to spirochaetes and *Chlamydia* (Stephens et al. 1998). However, taking into account that Helitrons frequently capture host gene fragments, as in maize (Lal et al. 2003; Gupta et al. 2005; Morgante et al. 2005; Xu and Messing 2006), thale cress, rice and worms (Kapitonov and Jurka 2001b; Kapitonov and Jurka 2007) and bats (Pritham and Feschotte 2007), a more parsimonious scenario would consider that ATREPX1 contains a captured partial copy of the PPi-PFK β-subunit gene from *Arabidopsis thaliana*, based on the high RM score (2,107) and the nucleotide similarity to this element (0.86).

Conclusion and perspectives

In this report, we showed that several transcripts harbor insertions of TEs of different lengths into the protein-coding

regions and of gene fragments into TEs. These findings, for which the literature is relatively scarce and recent, revitalize the question whether these messenger RNAs containing TEs or gene fragments produce a functional protein. The scarcity of such reports may be temporary. The National Center for Biotechnology Information (NCBI) presents a wave of new public information, such as the registration of 92 Genome Projects from plants, three of which are already completed: *Arabidopsis thaliana* (Arabidopsis Genome Initiative 2000), *O. sativa* (International Rice Genome Sequencing Project 2005), and *Populus trichocarpa* (Tuskan et al. 2006). Moreover, the NCBI made available 24 transcriptomes from flowering plants based on UniGenes (18 from dicotyledonous and six from monocotyledonous plants), as well as dozens of expressed sequence databases based on ESTs. This enormous mass of information enables large-scale analyses about the contribution of TEs to gene structure, especially the occurrence of TE-encoded sequences in the protein-coding region. Moreover, a survey of transcriptomes can unveil the impact of those fragments on the structure and function of host genes and on the increase of variability of the protein repertories, turning TEs definitively into gold.

Acknowledgments We thank V.V. Kapitonov (GIRINST, Mountain View, USA) for valuable suggestions, L.M. Almeida (University of Alberta, Edmonton, Canada) for the drawing of Fig. 1 and three anonymous referees for critically reviewing the manuscript. This work was supported by grants provided by the Brazilian agencies FAPESP (fellowship 05/57212-3 to F.R.L) and CNPq (to C.M.A.C and G.A.G.P.).

References

Almeida LM, Silva IT, Silva WAS Jr, Castro JP, Riggs PK, Carareto CMA, Amaral MEJ (2007) The contribution of transposable elements to *Bos taurus* gene structure. Gene 390:180–189

Alves AMCR, Meijer WG, Vrijbloed JW, Dijkhuizen L (1996) Characterization and phylogeny of the *pfp* gene of *Amycolatopsis methanolica* encoding PPi-dependent phosphofructokinase. J Bacteriol 178:149–155

Andrés C, Lurin C, Small ID (2007) The multifarious roles of PPR proteins in plant mitochondrial gene expression. Physiol Plant 129:14–22

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

Arkhipova IR, Meselson M (2005) Diverse DNA transposons in rotifers of the class Bdelloidea. Proc Natl Acad Sci USA 102:11781–11786

Bennetzen JL (2000) Transposable element contributions to plant gene and genome evolution. Plant Mol Biol 42:251–269

Bennetzen JL (2002) Mechanisms and rates of genome expansion and contraction in flowering plants. Genetica 115:29–36

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

Biemont C, Vieira C (2006) Genetics: junk DNA as an evolutionary force. Nature 443:521–524

Blakeley SD, Crews L, Todd JF, Dennis DT (1992) Expression of the genes for the α-and β-subunits of pyrophosphate-dependent phosphofructokinase in germinating and developing seeds from *Ricinus communis*. Plant Physiol 99:1245–1250

Britten RJ (2004) Coding sequence of functioning human genes derived entirely from mobile elements sequences. Proc Natl Acad Sci USA 101:16825–16830

Brosius J (2003) The contribution of RNAs and retroposition to evolutionary novelties. Genetica 118:99–116

Brosius J, Gould SJ (1992) On 'genomenclature': a comprehensive (and respectful) taxonomy for pseudogenes and other 'junk DNA'. Proc Natl Acad Sci USA 89:10706–10710

Carlisle SM, Blakeley SD, Hemmingsen SM, Trevanion SJ, Hiyoshi T, Kruger NJ, Dennis DT (1990) Pyrophosphate-dependent phosphofructokinase. Conservation of protein sequence between the subunits and with the ATP-dependent phosphofructokinase. J Biol Chem 265:18366–18371

DeBarry JD, Ganko E, McDonald JF (2005) The contribution of LTR retrotransposon sequences to gene evolution in *Mus musculus*. Mol Biol Evol 23:479–481

Deng Z, Huang M, Singh K, Albach RA, Latshaw SP, Chang KP, Kemp RG (1998) Cloning and expression of the gene for the active PPi-dependent phosphofructokinase of *Entamoeba histolytica*. Biochem J 329:659–664

Deng Z, Roberts D, Wang X, Kemp RG (1999) Expression, characterization, and crystallization of the pyrophosphate dependent phosphofructo-1-kinase of *Borrelia burgdorferi*. Arch Biochem Biophys 371:326–331

Deininger PL, Batzer MA (1999) Alu repeats and human disease. Mol Genet Metab 67:183–193

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

Eickler EE, Sankoff D (2003) Structural dynamics and eukaryotic chromosomal evolution. Science 301:793–797

Fedoroff N (2000) Transposon and genome evolution in plants. Proc Natl Acad Sci USA 97:7002–7007

Ganko EW, Bhattacharjee V, Schliekelman P, McDonald JF (2003) Evidence for the contribution of LTR retrotransposon to *C. elegans* gene evolution. Mol Biol Evol 20:1925–1931

Ganko EW, Greene CS, Lewis JA, Bhattacharjee VM, McDonald JF (2006) LTR retrotransposon-gene associations in *Drosophila melanogaster*. J Mol Evol 62:111–120

Gerber A, O'Connell MA, Keller W (1997) Two forms of human double-stranded RNA-specific editase 1 (hRED1) generated by the insertion of an Alu cassette. RNA 3:453–463

Gotea V, Makalowski W (2006) Do transposable elements really contribute to proteomes? Trends Genet 22:260–267

Gupta S, Gallavotti A, Stryker GA, Schmidt RJ, Lal SK (2005) A novel class of Helitron-related transposable elements in maize contain portions of multiple pseudogenes. Plant Mol Biol 57:115–127

Hayashi Y, Sakata H, Makino Y, Urabe I, Yomo T (2003) Can an arbitrary sequence evolve towards acquiring a biological function? J Mol Evol 56:162–168

Hilgard P, Huang TM, Wolkoff AW, Stockert RJ (2002) Translated Alu sequence determines nuclear localization of a novel catalytic subunit of casein kinase 2. Am J Physiol Cell Physiol 283:C472–C483

Hoenicka J, Arrasate M, de Yebenes JG, Avila J (2002) A two-hybrid screening of human Tau protein: interactions with Alu-derived domain. Neuroreport 13:343–349

Huang X, Madan A (1999) CAP3: a DNA sequence assembly program. Genome Res 9:868–77

International Human Genome Sequencing Consortium (2001) A physical map of the human genome. Nature 409:934–941

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

Jordan IK, Rogozin IB, Glazko GV, Koonin EV (2003) Origin of a substantial fraction of human regulatory sequences from transposable elements. Trends Genet 19:68–72

Jiang N, Bao Z, Zhang X, Eddy SR, Wessler SR (2004) Pack-MULE transposable elements mediate gene evolution in plants. Nature 431:569–573

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

Kapitonov VV, Jurka J (2000) Non autonomous DNA transposon ATDNA1T9A—a consensus sequence. Repbase Update. http://www.girinst.org

Kapitonov VV, Jurka J (2001a) ATMU6N1 is a non-autonomous DNA transposon—a consensus sequence. Repbase Update. http://www.girinst.org

Kapitonov VV, Jurka J (2001b) Rolling-circle transposon in eukaryotes. Proc Natl Acad Sci USA 98:8714–8719

Kapitonov VV, Jurka J (2003) Molecular paleontology of the transposable elements in the *Drosophila melanogaster* genome. Proc Natl Acad Sci USA 100:6569–6574

Kapitonov VV, Jurka J (2005) Helitron4_OS: a new, possibly active helitron from rice. Repbase Rep 5:181–181

Kapitonov VV, Jurka J (2007) Helitrons on a roll: eukaryotic rolling-circle transposons. Trends Genet 23:521–529

Kidwell MG (2002) Transposable elements and the evolution of genome size in eukaryotes. Genetica 115:49–63

Kidwell MG, Lish D (1997) Transposable elements as sources of variation in animals and plants. Proc Natl Acad Sci USA 94:7704–7711

Lal SK, Giroux MJ, Brendel V, Vallejos CE, Hannah LC (2003) The maize genome contains a Helitron insertion. Plant Cell 15:381–391

Li WH, Gu ZL, Wang HD, Nekrutenko A (2001) Evolutionary analyses of the human genome. Nature 409:847–849

Lin C, Mueller LA, Carthy JM, Crouzillat D, Pétiard V, Tanksley SD (2005) Coffee and tomato share common gene repertories as revealed by deep sequencing of seed and cherry transcripts. Theor Appl Genet 112:114–130

Lorenc A, Makalowski W (2003) Transposable elements and vertebrate protein diversity. Genetica 118:183–191

Makalowski W (2000) Genomic scrap yard: how genomes utilize all that junk. Gene 259:61–67

Makalowski W, Mitchel GA, Labuda D (1994) Alu sequences in the coding regions of mRNA: a source of protein variability. Trends Genet 10:188–193

Mao L, Wood TC, Yu Y, Budiman MA, Tomkins J, Woo S, Sasinonowski M, Presting G, Frish D, Goff S, Dean RA, Wing RA (2000) Rice transposable elements: a survey of 73,000 sequence-tagged-connectors. Genome Res 10:982–990

Mertens E, Ladror US, Lee JA, Miretsky A, Morris A, Rozario C, Kemp RG, Muller M (1998) The pyrophosphate-dependent phosphofructokinase of the protist, *Trichomonas vaginalis*, and the evolutionary relationships of protist phosphofructokinases. J Mol Evol 47:739–750

Messing J, Bharti AK, Karlowski WM, Gundlach H, Kim HR, Yu Y, Wei FS, Fuks G, Soderlund CA, Mayer KFX, Wing RA (2004) Sequence composition and genome organization of maize. Proc Natl Acad Sci USA 101:14349–14354

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

Miller WJ, McDonald JF, Nouaud D, Anxolabéhère D (1999) Molecular domestication—more than a sporadic episode in evolution. Genetica 107:197–207

Mitchell GA et al (1991) Splice-mediated insertion of an *Alu* sequence inactivates ornithine delta-aminotransferase: a role for Alu elements in human mutation. Proc Natl Acad Sci USA 88:815–819

Möhlmann T, Tjaden J, Schwoppe C, Winkler HH, Kampfenkel K, Neuhaus HE (1998) Occurrence of two plastidic ATP/ADP transporters in *Arabidopsis thaliana* L—molecular characterisation and comparative structural analysis of similar ATP/ADP translocators from plastids and *Rickettsia prowazekii*. Eur J Biochem 252:353–359

Moore SA, Ronimus RS, Roberson RS, Morgan HW (2002) The structure of a pyrophosphate-dependent phosphofructokinase from the Lyme disease spirochete *Borrelia burgdoferi*. Structure 10:659–671

Morgante M, Brunner S, Pea G, Fengler K, Zuccolo A, Rafalski A (2005) Gene duplication and exon shuffling by Helitron-like transposons generate intraspecies diversity in maize. Nat Genet 37:997–1002

Mustroph A, Sonnewald U, Biemelt S (2007) Characterization of the ATP-dependent phosphofructokinase gene family from *Arabidopsis thaliana*. FEBS Lett 581:2401–2410

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

Ohno S (1970) Evolution by gene duplication. Springer, Berlin

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

Poulter RTM, Goodwin TJD, Butler MI (2003) Vertebrate helentrons and other novel Helitrons. Gene 313:201–212

Pritham EJ, Feschotte C (2007) Massive amplification of rolling-circle transposons in the lineage of the bat *Myotis lucifugus*. Proc Natl Acad Sci USA 104:1895–1900

Pruitt KD, Maglott DR (2001) RefSeq and LocusLink: NCBI gene-centered resources. Nucleic Acids Res 29:137–140

Roberson RS, Ronimus RS, Gephart S, Morgan HW (2000) Biochemical characterization of an active pyrophosphate-dependent phosphofructokinase from *Treponema pallidum*. FEMS Microbiol Lett 9735:1–4

Sakai H, Tanaka T, Itoh T (2007) Birth and death of genes promoted by transposable elements in *Oryza sativa*. Gene 392:59–63

SanMiguel P, Tihkonov A, Jin YK, Motchoulskaia N, Zkharov D, Melake-Berhan A, Springer PS, Edwards KJ, Lee M, Avramova Z, Bennetzen JL (1996) Nested retrotransposon in the intergenic regions of the maize genome. Science 274:765–768

Sarkar A, Sim C, Hong YS, Hogan JR, Fraser MJ, Robertson HM, Collins FH (2003) Molecular evolutionary analysis of widespread *piggy-back* transposon family and related domesticated sequences. Mol Genet Genomics 270:173–180

Shapiro JA, von Sternberg R (2005) Why repetitive DNA is essential to genome function. Biol Rev 80:227–250

Siebers B, Klenk HP, Hensel R (1998) PPi-dependent phophofructokinase from *Thermoproteus tenax*, an archaeal descendant of an ancient line in phosphofructokinase evolution. J Bacteriol 180:2137–2143

Sigrell JA, Cameron AD, Jones TA, Mowbray SL (1998) Structure of the *Escherichia coli* ribokinase in complex with ribose and dinucleotide determined to 1.8 Å resolution: insights into a new family of kinase structures. Structure 6:183–193

Sorek R, Ast G, Graur D (2002) Alu-containing exons are alternatively spliced. Genome Res 12:1060–1067

Stephens RS, Kalman S, Lammel C, Fan J, Marathe R, Arvind L, Mitchell W, Olinger L, Tatusov RL, Zhao Q (1998) Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. Science 282:754–759

Swofford DL (1998) PAUP*: phylogenetic analysis using parsimony (* and other methods). Sinauer Associates, Sunderland

The Gene Ontology Consortium (2001) Creating the gene ontology resource: design and implementation. Genome Res 11:1425–1433

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTALW: improving the sensitivity of progressive multiple sequence alignment

through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680

Thornburg BG, Gotea V, Makalowski W (2006) Transposable elements as source of transcription regulation signals. Gene 365:104–110

Tikhonov AP et al. (1999) Collinearity and its exceptions in orthologous *adh* regions of maize and sorghum. Proc Natl Acad Sci USA 96:7409–7414

Tjaden J, Schwoppe C, Mohlmann T, Quick PW, Neuhaus HE (1998) Expression of a plastidic ATP/ADP transporter gene in *Escherichia coli* leads to a functional adenine nucleotide transport system in the bacterial cytoplasmic membrane. J Biol Chem 273:9630–9636

Turcotte K, Srinivasan S, Bureau T (2001) Survey of transposable elements from rice genomic sequences. Plant J 25:169–179

Tuskan GA et al (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science 313:1596–1604

van de Lagemaat LN, Landry JR, Mager DL, Medstrand P (2003) Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. Trends Genet 19:530–536

Vieira LGE, Andrade AC, Colombo CA et al (2006) Brazilian coffee genome project: an EST-based genomic resource. Braz J Plant Physiol 18:95–108

Volff JN (2006) Turning junk into gold: domestication of transposable elements and the creation of new genes in eukariotes. BioEssays 28:913–922

Wu LF, Reizer A, Reizer J, Cai B, Tomich JM, Saier MH (1991) Nucleotide sequence of the *Rhodobacter capsulatus fruK* gene, which encodes fructose-1-phosphate kinase: superfamily including both phosphofructokinases of *E. coli*. J Bacteriol 173:3117–3127

Xu JH, Messing J (2006) Maize haplotyple with a Helitron-amplified cytidine deaminase gene copy. BMC Genetics 9:7–52

Zhang J, Peterson T (1999) Genome rearrangements by nonlinear transposons in maize. Genetics 153:1403–1410

Zhou Q et al (2006) Helitron transposons on the sex chromosomes of the platyfish *Xiphophorus maculatus* and their evolution in animal genomes. Zebrafish 3:39–52

Transcriptional activity of exonized transposable elements sequences in *Coffea* genomes

Capítulo 2

Transcriptional activity of exonized transposable element sequences in *Coffea* genomes

Fabrício R. Lopes[1], Alan C. Andrade[2], Pierre Marraccini[2, 3], Daudi Jjingo[4], I. King Jordan[4], João B. Teixeira[2], Marcelo F. Carazzolle[5], Gonçalo A. G. Pereira[5], Claudia M. A. Carareto*[1]

[1]UNESP – São Paulo State University, Department of Biology, 15054-000, São José do Rio Preto, São Paulo, Brazil,

[2]EMBRAPA – Brazilian Enterprise for Agropecuary Research, Genetic Resources and Biotechnology, 70770-900, Brasília, Distrito Federal, Brazil,

[3]CIRAD – Agricultural Research for Development, UMR PIA, F-34398, Montpellier Cedex 5, France,

[4]School of Biology, Georgia Institute of Technology, Atlanta, GA 30332, USA,

[5]UNICAMP – State University of Campinas, Department of Genetics and Evolution, Institute of Biology, 13083-970, Campinas, São Paulo, Brazil,

*Corresponding author (fax 55 17 3221 2390; e-mail carareto@ibilce.unesp.br).

Abstract

A wave of public information has enabled the report of innumerable examples where transposable elements (TEs) fragments are incorporated as exons into mRNAs, however, the proposition that such exonized TEs encode functional proteins is controversial. In order to characterize the gene expression pattern of TE-derived coding sequences identified in the *Coffea arabica* genome, we analyzed the macroarray profiling of transcripts with distinct molecular functions of sequences containing TE-cassettes or not. The evolutionary changes of the gene expression pattern mediated by exonized TE sequences using probes of tissues and stress conditions derived from *C. arabica* and *C. canephora* are presented and discussed.

Introduction

The availability of a large public database has allowed the identification of a wide plethora of evolutionary and functional roles of the transposable elements (TEs) in the genomes of prokaryotes and eukaryotes, namely by providing ready-to-use motifs for new transcriptional regulatory elements and polyadenilation signals, serving as recombination

hotspots (Lorenc and Makalowski, 2003), and contributing to both genome expansion and contraction (Fedoroff, 2000). Although a small fraction of the genome is constituted of protein-coding regions, these regions are also occupied by TE-cassette insertions, called exonized TE insertions, which provide an important, putative raw material for the host protein variability and versatility. While the host transcriptome is frequently impacted by novel chimerical transcripts (mature mRNA plus TE), as described in the human genome (Nekrutenko and Li, 2001), and other organisms, as in *Mus musculus* (Debarry *et al.*, 2006), *Caenorhabditis elegans* (Ganko *et al.*, 2003), *Drosophila melanogaster* (Ganko *et al.*, 2006), *Bos taurus* (Almeida *et al.*, 2007), *Oryza sativa* (Sakai *et al.*, 2007; Krom *et al.*, 2008), very few studies support the opportunity of repetitive fragments inserted into coding sequences to be translated and to have a functional impact on the organism fitness (Gotea and Makalowski, 2006). Adaptative changes caused by fragments of translated TEs are commonly called molecular domestication (MILLER *et al.*, 1999). In the plant genomes, a variety of cases are available, as for example, the DNA binding protein *Daysleeper* (Bundock and Hooykaas, 2005), and transcription factors *Fhy3* and *Far1* in *Arabidopsis* (Hudson *et al.*, 2003), the proteins with an unknown molecular function *Gary* in cereal grasses (Muehlbauer *et al.*, 2006) and *Mustang* in flowering plants (Cowan *et al.*, 2005).

The relatively low number of reports on translated TEs are due, firstly, because the presence of TE cassettes in transcripts do not guarantee their translation and, secondly, because eukaryotic cells contain several quality control mechanisms that can initiate the degradation of the transcript, even acting over the protein immediately after translation (Wagner *et al.*, 2002). For example, alternatively spliced exons that carry premature termination codons (PTCs) in the frame are used to regulate downwards the amount of protein products of certain genes by being targets of the nonsense mediate mRNA decay (NMD) pathway (Wagner *et al.*, 2002). Many TEs can provide alternatively spliced exons with PTCs, thus a potential role in regulating gene expression post-transcriptionally needs to be taken into consideration. Therefore, the mRNA surveillance onto TE-harboring isoforms needs to be evaluated.

In a previous study, we analyzed the contribution of TE-cassettes for the origin of new transcripts in economically important species belonging the genus *Coffea* (Lopes *et al.*,

2008). There, a combination of consensus TE banks stored in Repbase (Jurka *et al.*, 2005) and several cDNA libraries of distinct tissues, development stages and stress conditions (Vieira *et al.*, 2006; Lin *et al.*, 2005) allowed the identification of 143 unigenes (0.23% out of 60,788), namely 72 from *C. arabica*, 66 from *C. canephora* and five from *C. racemosa*, harboring TE-cassettes (Lopes *et al.*, 2008). Facing a scarcity of data about the changes in expression pattern mediated by exonized TE, we analyzed this already open question by macroarray profiling of transcripts with distinct molecular functions whether containing TE-cassettes or not (identified in Lopes *et al.*, 2008 and also in this one), both sequences highly related by sequence similarity. The expression analyses evidenced the low expression pattern of these chimerical transcripts. Searches by putative transcripts that could trigger the NMD mechanisms and to explain their low expression profile was performed and here discussed.

Material and Methods

Identification of novel cases of TEs incorporated into mature mRNAs from Coffea spp

A recent report by Piriyapongsa *et al.* (2007) showed that the choice of sequence similarity search methods to detect TE-derived sequences strongly influence the estimate of TE-cassettes that can be identified in the protein coding region. Thus, Hidden Markov model based searches followed by BLAST methods (tBLASTx $\rightarrow$ tBLASTn $\rightarrow$ BLASTx $\rightarrow$ BLASTp $\rightarrow$ BLASTn) and Repeat Masker are more sensitive/informative for identifying new exonized TEs. Facing this observation, in order to identify TE-derived sequences in novel protein coding sequences in the transcriptomes from *Coffea*, plus those identified by Repeat Masker (Lopes *et al.*, 2008), 131,150 ESTs from *C. arabica* and 10,566 from *C. racemosa* (Vieira *et al.*, 2006) were compared by tBLASTx (Altschul *et al.*, 1997) against 2,503 plant consensus TEs of the Repbase (Jurka *et al.*, 2005). To avoid spurious results, only the best $E \leq e^{-10}$ matches were accepted, without imposing additional scores or length thresholds. The new ESTs were compared to EST clusters of each species for the identification of the respective unigenes.

Sequence similarity searches between unigenes containing TE-cassettes or not

The unigenes containing TE-cassettes obtained in our previous study (72 for *C. arabica*, *C. canephora* and *C. racemosa*; (Lopes *et al.*, 2008) and here (only for *C. arabica* and *C. racemosa*) were compared by BLASTn (Altschul *et al.*, 1997) against all unigenes of each species. The examples of alignment between the two unigene sets are shown in Figure 1 (the two sequence sets are available under request).

Expression analyses

a) *Plant Material:* for the probe synthesis, total mRNA was extracted from *C. arabica* in the following tissue and treatment conditions: a) Drought stress: Cultivars tolerant (Iapar59) and sensitive (Rubi) to drought maintained in field conditions with (ψpd= -0.5 MPa) and without (ψpd= -1.7 MPa) water; b) Cell culture: embriogenic callus from *C. arabica* cv. Catuaí Vermelho maintained in a multiplication medium during ~4 months; c) Inhibition treatment: the same embriogenic callus was treated with the protein biosynthesis inhibitor Cycloheximide (CHX: 10 and 30 mg/mL in alcohol) added to the cell culture for a final concentration of 100 and 300 µg/mL, respectively. *C. canephora:* a) Drought stress: clones var. Conillon tolerant (14) and sensitive (22) to drought were selected by the INCAPER (Ferrão *et al.*, 2000) and grown in a greenhouse with (unstressed condition) or without (stress, ψpd leaves = -3.0 MPa) water.

b) *RNA isolation, DNAse treatment and Reverse transcription:* The 10 total RNA samples were extracted from cells using Concert™ reagent (Invitrogen, Carlsbad, CA, USA) according to the manufacturer´s protocol. A total RNA sample (10 µg) was incubated for 30 min at 37 ºC with 3 U of *RQ1 RNAse-Free DNAse* (Promega, Madison, WI, USA) in a final volume of ~10 µl. Each total RNA sample was mixed with 1.5 µl Oligo$(dT)_{12\text{-}18}$ (Invitrogen) and heated at 75 ºC for 10 min, then cooled for 5 min on ice. The reaction mixture for the reverse transcription contained 5 µl of 5 × first-strand buffer, 2.5 µl of 0.1 mol/l DTT, 40 U of RNAseOUT, 50 µCi [α-^{33}P]-dCTP and 2.5 µl of a 10 mM mixture of unlabeled dNTPs (dATP, dTTP and dGTP) was heated at 42 ºC for 5 min. The reverse transcription was performed at 42 ºC for 20 min with 300 U of *SuperScript III*

First-Strand Synthesis System for RT-PCR (Invitrogen). Then, 1.25 µl of unlabeled dCTP (10 mM) was added and maintained for 1 h, terminated by heating at 94 ºC for 5 min and cooled for 5 min on ice. The total volume (30.25 µl) was used in the hybridization experiments.

c) Amplification of target DNA: Each target cDNA (100 ng) was amplified by PCR in a volume of 25 µl with 1.25 U of *Platinum® Taq DNA Polymerase* (Invitrogen) in 10 × polymerase buffer, 2 mM $MgCl_2$, a 200 µM concentration of each dNTP and 10 µM of each universal primer, M13F (5´-GTAAAACGACGGCCAG-3´) and M13R (5´-CAGGAAACAGCTATGAC-3´). The solutions were heated to 94 °C for 2 min, and followed by 35 cycles of denaturation (94 °C for 30 sec), annealing (50 °C for 30 sec), and extension (72 °C for 4 min), followed by a final extension at 72 °C for 7 min. The target DNAs were used for making the array membranes.

d) Macroarray experiments: The expression analyses were carried out using transcripts with distinct molecular functions whether containing TE-cassettes (109) or not (83). Both sequences were strongly related by sequence similarity and were identified in both a previous study (Lopes et al., 2008) and in this one. The PCR products were denatured in DMSO (50% v/v) for 30 min at 37 ºC, arranged in a 384-well microtiter plate and then spotted twice in the same position onto *Performa II nylon filters* (Genetix Limited, Hampshire, UK) using the high-throughput robot system *Q-BOT* (Genetix Limited). To increase the signal homogeneity in the amount of PCR product among spots and filters, the set of 203 cDNAs were spotted in duplicate in a 2×2 array into two identical arrays arranged on the same nylon filters (222 × 222 mm). Additionally, eight spots containing cDNA of the reference gene ubiquitin were applied to delimit the two arrays. After sample deposition, the filters were dampened with a denaturant solution (0.13 M NaCl and 0.5 M NaOH) for 10 min, a neutralization solution (1.5 M NaCl and 1 M Tris) for 5 min, later fixed by UV light exposition (1,200 µj/cm^2) for 12 seconds and stored at -80 ºC. The filters were pre-hybridized for 2 h at 65 ºC in a *Modified Church and Gilbert Buffer* (0.5 M Phosphate Buffer pH 7.2, 7 % SDS, 10 mM EDTA) and hybridized

overnight with cDNA sample probes. Membranes were washed for 15 min three times with 0.1 % SDS/1 × SSC and three times with 0.1 % SDS/0.1 × SSC at 65 ºC. After washing, the filters were exposed on imaging plates *BAS-MS 2340* (Fujifilm, Tokyo, Japan) for 72 h into *BAS 2340 cassette* (Fujifilm) and scanned using a fluorescent image analyzer *FLA3000* (Fujifilm). The radioactive intensity of each spot was quantified by *Array Guage* software (Fujifilm), corrected by the background level, normalized to the average intensities of the reference gene ubiquitin (uni_CA_090) and for the differences of probe labeling using the average signals of all genes studied. The homogeneity of the spot replicates were evaluated and represented by an medium value using *limma* of the Bioconductor package (Gentleman *et al.*, 2004). The final expression values were analyzed by *heatmap* using R (http://www.r-project.org). The expression profiles obtained for each probe were considered as a group and compared aiming to identify outstanding expression differences related to treatment (callus *vs* callus treated with CHX), stress (I59_I *vs* I59_NI and Rubi_I *vs* Rubi_NI for *C. arabica*; 14_I *vs* 14_NI for *C. canephora*) and status to stress (tolerant *vs* sensitive only for *C. arabica*). To compute the average differences between each group of comparison, the average of one group was subtracted from that of the other group. To test if these differences were significant or not, a two-tailed paired student's T-test was performed between each pair of groups being tested on an Excell spreadsheet. This yielded a *P*-value that was used to determine significant differences ($p<0.05$).

Results and Discussion

Identification of novel cases of TE-derived CDSs using BLAST algorithm

A recent report showed that the type of TE-annotation methods strongly influence the estimates of proteins containing TEs (Piriyapongsa *et al.*, 2007). Searches based on patterns (Hidden Markov model) followed by BLAST methods (tBLASTx → tBLASTn → BLASTx → BLASTp → BLASTn) and Repeat Masker have been reported to be more ensitive/informative for identifying new exonized TEs. Facing this finding, a search on 131,150 ESTs from *C. arabica* and 10,566 from *C. racemosa* was undertaken using 2,503

consensus TEs that allowed an identification of 421 and 23 matches, respectively. Due to the high number of matches in the first species, only those with E<e-^{18} were analyzed, resulting in 72% of the total (303 matches); while in the second one, all the results were evaluated. Finally, a comparison of the ESTs retrieved against the database of unigenes (EST clusters) of the LGE platform allowed an identification of 27 TE-derived protein coding sequences in *C. arabica* (Supplementary Material: Table S1) and 10 new CDSs in *C. racemosa* (Supplementary Material: Table S3). Moreover, the length of TE-cassettes identified by tBLASTx are larger than those by Repeat Masker for both species (209.3 ± 20bp against 84.56 ± 3.82 bp in *C. arabica*; 207.2 ± 11 bp against 93.86 ± 5.79 bp in *C. racemosa*). Hypothesizing that more divergent TE-cassettes can be more easily tolerated by host genome or until acquiring a biological function (Hayashi *et al.*, 2003) and considering the findings of Piriyapongsa *et al.* (2007), showing that approaches based on comparisons in amino acid residues allow the identification of more divergent TE fragments, the use of a new search method in ESTs from *Coffea spp* showed a significantly distinct result from that initially obtained. Moreover, differently from the results obtained by Repeat Masker, all unigenes matched to proteins with a clearly defined function, namely: protein factors related to transcription and spliceosomal machinery, chaperones and alcohol dehydrogenase (see Supplementary Material).

Unigenes containing TE-cassettes or not: identification and similarity patterns

A comparison of the unigenes containing TE-cassettes from *C. arabica* (86), *C. canephora* (59) and *C. racemosa* (15) against their respective sets of unigenes allowed the identification of 111, 47 and zero unigenes respectively, related by their high similarity sequence (Supplementary Material: Tables S1-S3). The pair of aligned sequences was classified into four categories regarding the region of similarity and TE insertion site (Figures 1 A-D). Two sets of alignments were highlighted by a strong suggestion of putative alternative splicing events in *C. arabica*, such as three transcripts (uni_CA_046, uni_CA_125 and uni_CA_127) similar to a rust resistance Rp1-D-like protein (13310480) and two transcripts (uni_CA_055 and uni_CA_138) similar to a universal stress protein (USP) family protein (30681955). Moreover, several other transcripts presented a polymorphic pattern for the presence/absence of TE, as for example, those similar to SC35-like putative splicing factor and heat shock cognate 70 KDa protein. Moreover, 13

transcripts in *C. arabica* and five in *C. canephora*, showed non-conclusive patterns of similarity for alternative splicing since the similarity regions finished before initiation or began after the TE insertion site (see Supplementary Material).

The lack of evidence of alternative *splicing* in many of the transcripts analyzed does not exclude the possibility of such an event because although not identified, host genes probably produce wild transcripts besides those containing TEs. Moreover, this analysis is influenced by several factors, such as: 1) EST libraries were built based only on the representation of the 5' end of mRNAs preventing the identification of TEs in other portions of the transcript; 2) several libraries were not targeted for a great sequencing effort (saturation), mainly *C. canephora* and *C. racemosa*, which present few tissues represented and ESTs obtained; 3) the unavailability of genomic resources in coffee does not allowed an improved analysis of gene structure, TE insertion sites, as well as the identification of transcripts related to each expressed loci. In addition, the identification of novel unigenes containing TEs in the BLAST results was observed most probably for those cassettes presenting a RM score <250.

Finally, the TE-derived CDSs and CDSs without TEs from *C. arabica* and *C. canephora* were assigned to functional groups based on Gene Ontology (GO) classification data (Figure 2, Supplementary Material Tables S1-S3). As a result, the category enzymes (14% and 9%), nucleic acid binding (28% and 13%), ligand binding (6% and 21%) and chaperones (11% and 16%) were the most frequent for all the transcripts containing or not TEs in *C. arabica*, respectively, besides of number important transcripts that did not match proteins stored in GO, here identified as *no hits found* (20% and 21%, respectively). On the other hand, the TE-derived CDSs and CDSs without TEs identified in *C. canephora* were similar to proteins of a variety of functional categories. However, in this species, the matches to unknown proteins (12% and 13%, respectively) and mainly those considered as *no hits found* (44% and 61%, respectively) were the most frequent.

Expression profiles of TE-derived CDSs

A total of 109 (Figure 3) and 83 (Figure 4) transcripts containing TE-cassettes or not, for sharing high a similarity sequence were analyzed for their expression profile by

macroarrays in five samples in *C. arabica* (callus, callus treated with CHX; leaves cv. Iapar59 and Rubi, irrigated or not) and two samples in *C. canephora* (leaves clone 14, irrigated or not). With the exception of the great number of TE-derived CDSs with a low expression profile, five transcripts had a detectable expression in all samples: a probable kinesin heavy chain (Genbank accession # 7487566, uni_CA_008), known protein (uni_CA_022), Putative Cer1 (Genbank accession # 37535278, uni_CA_072), known protein (Genbank accession # 18417290, uni_CA_057) and EIL2 (Genbank accession # 30016896, uni_CA_157). Another two transcripts can be cited as having a high expression in the four samples of drought stress in *C. arabica*: heavy-metal-associated domain-containing protein (Genbank accession # 42572261, uni_CA_034) and trans-membrane MLO family protein (Genbank accession # 18398444, uni_CA_053).

Moreover, 19 transcripts had their expression profiles increased or reestablished in callus treated with the protein biosynthesis inhibitor Cycloheximide. A portion of the growing literature shows that inhibition of the protein biosynthesis machinery using CHX allows the accumulation of previously silenced transcripts by reversing a mechanism named nonsense-mediated mRNA decay (NMD). NMD is a conserved pathway of mRNA surveillance present in all eukaryotes by degrading mRNAs containing premature terminate codons (PTCs) that could synthesize truncated proteins and have a dominant deleterious effect (Conti and Izaurralde, 2005). In humans, ~30% of the hereditary genetic disorders are related to genes harboring frame shifts that result in PTCs (Holbrook *et al.*, 2004). PTCs, for example, can be generated during rearrangement events that occur in the origin of the variability of immunoglobulin and genes for the T-cell receptor. Most importantly, these defective transcripts can be generated by alternative splicing, that has a fundamental role in the generation of variability of the transcriptional repertory, generating a substantial number of mRNAs harboring PTC, commonly identified as mRNA PTC+ (Lejeune and Maquat, 2005).

Several transcripts that trigger the NDM mechanism have been described using CHX, as reported in plants (Lambein *et al.*, 2003; Wu *et al.*, 2007), humans (Noensie and Dietz, 2001; Lamba *et al.*, 2003; Hillman *et al.*, 2004) and mice (Lei *et al.*, 2001; Vignier-Schlossarek *et al.*, 2009). In addition, the test with a higher concentration of CHX (300

µg/mL) was chosen because it showed a total reduction of expression of the reference gene ubiquitin, a transcript that is not targeted to NMD, compared to the test using a lower concentration of CHX (100 µg/mL) (data not shown). This exploratory analysis allowed the identification of a target that would be able to trigger the NMD mechanism, but two complementary analyses should be performed for a conclusive scenario: 1) the full-length sequencing of the cDNAs could show the defective character of these elements and the confirmation of their differential expression pattern by RT-qPCR (callus and callus treated with alcohol *vs* callus treated with CHX); 2) analysis of the protein coding potential of both unigene sets focusing the identification of frame shifts and stop codons.

The expression analysis of 83 CDSs without TEs, but related by a high similarity sequence and molecular function, also presented a very low expression in the tested samples from *C. arabica* and *C. canephora* and, in addition, an expression profile of reestablishment or increase in callus treated with CHX (Figure 3). However, we cannot discard the possibility that those CDSs contained TEs in other regions not represented in the unigenes since the transcriptomes from *Coffea spp* were sampled only by the 5' end of mRNAs. Given together, the expression profiles of isophorms containing TEs or not, highlight the complexity of the evolutionary and physiologic impact of TE insertions into the protein coding region.

The expression profiles obtained for each probe were considered as a group and compared by statistical analyses aiming at evaluating significant expression differences. For the TE-derived CDSs, the comparison between the groups of callus treated with CHX *vs* callus showed a significant expression difference for a higher expression in the condition of treatment (average difference among the groups= 0.4448, *P*-value= 2.5E-03). On the contrary, for gene putatively without TE (average difference= -0.1486, *P*-value= 5.3E-01) no difference was observed, at least on average.

Both TE-derived CDSs and the gene without TE showed a significant expression difference in the other groups of comparison, namely: for the cultivar tolerant submitted to drought stress (I59_I *vs* I59_NI, average difference= -0.0667 and -0.1501, respectively; *P*-value= 1.1E-03 for both data set) showed higher expression in the irrigated condition; for

the cultivar sensitive Rubi (Rubi_I *vs* Rubi_NI, average difference= 0.1844 and 0.2102, *P*-value= 1.3E-05 and 7.9E-0.6, respectively) showed higher expression in the irrigation condition; status to stress (cultivar tolerant *vs* sensitive= I59_NI *vs* Rubi_NI, average difference= 0.1125 and 0.1353, *P*-value= 6.8E-07 and 2.7E-06, respectively) a higher expression in the cultivar tolerant from *C. arabica* was observed*;* and the drought stress in *C. canephora* (14_I *vs* 14_NI, average difference= -0.19 and -0.4413, *P*-value= 2.1E-08 and 6.3E-05, respectively), a higher expression in the not irrigated condition was observed. These results suggest that the drought stress influences the expression levels for the transcriptional activity of transcripts containing TE-cassettes or not.

Analyses of complete genomes have allowed the description of several examples of TE sequences incorporated into mature mRNAs assuming that such exonized TEs codify proteins. However, Piriyapongsa *et al.* (2007) reported that database and annotation methods used influence the frequency that such chimerical transcripts are identified. The different search methods were complementary since they resulted in a large number of non-overlapping hits and differed in the ability to identify known cases of TE-derived CDSs. Probabilistic analyses of exons derived from TEs indicated that these sequences have a low protein coding potential, at least on average. Non-autonomous TEs, those destitute of conserved ORFs, are frequently exonized, but are unlikely to encode functional proteins. Moreover, Piriyapongsa *et al.* (2007) suggest that many of the exonized TEs act as post-transcription regulators of the gene expression and would act on a variety of regulatory pathways related to double strand RNAs. As discussed above, we are preliminarily presenting a proposal of putative NMD-inducing factors since the insertion of TE fragments into CDS could contribute to frame-shift and PTC, so that such unfavorable isophorms would be degraded.

Perspective

A remarkable question about TE-derived CDSs is the extension to which such chimerical transcripts contribute to the functional protein repertory variability in *Coffea*. Piriyapongsa *et al.* (2007) comment that protein coding sequences present a specific periodicity of nucleotide frequencies across the first, second and third codon positions thus being a hallmark for gene prediction. Using genomic sequences related to TE-derived transcripts

and the distribution of observed nucleotide frequencies, it is possible to identify protein coding nucleotide sequences. The authors used the earliest gene predictor, GeneMark (http://exon.biology.gatech.edu/), for evaluating the protein coding potential of TE-derived CDSs in the human genome. To answer our question about capacity of TE-derived CDSs encoding for functional proteins, an important set of genomic sequences from *Coffea*, not yet publicly available, could be used for a self-training stage of GeneMark that would allow identifying the coding capacity of the portion of our transcripts containing TEs or not, as well as, frame-shifts and stop codons. Thus it could be strongly suggested that they would contribute for transcriptional variability but with small influence on the host capacity of encoding functional proteins.

Acknowledgments

We thank Mário A. P. Saraiva and Felipe Vinecky for both making the array membranes and for manipulating the BCGP's cDNA libraries, and Elias A. G. Carnelossi for technical support. This study was funded by grants provided by the Brazilian agencies FAPESP (2008/05894-1 to C.M.A.C. and fellowship 2007/55985-0 to F.R.L.) and CNPq (to C.M.A.C.).

References

ALMEIDA, L. M. et al. The contribution of transposable elements to *Bos taurus* gene structure. **Gene,** v. 390, n. 1-2, p. 180-189, Apr 2007.

ALTSCHUL, S. F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. **Nucleic Acids Research,** v. 25, n. 17, p. 3389-3402, Sep 1997.

BUNDOCK, P.; HOOYKAAS, P. An *Arabidopsis* hAT-like transposase is essential for plant development. **Nature**, v. 436, p. 282-284, 2005.

CONTI, E.; IZAURRALDE, E. Nonsense-mediated mRNA decay: molecular insights and mechanistic variations across species. **Current Opinion in Cell Biology,** v. 17, n. 3, p. 316-325, 2005.

COWAN, R. K. et al. *MUSTANG* is a novel family of domesticated transposase genes found in diverse angiosperms. **Molecular Biology and Evolution**, v. 22, p. 2084-2089, 2005.

DEBARRY, J. D. et al. The contribution of LTR retrotransposon sequences to gene evolution in Mus musculus. **Molecular Biology and Evolution,** v. 23, n. 3, p. 479-481, Mar 2006.

FEDOROFF, N. Transposons and genome evolution in plants. **Proceedings of the National Academy of Sciences of the United States of America,** v. 97, n. 13, p. 7002-7007, Jun 2000.

FERRÃO, R. G. et al. EMCAPA 8141 - Robustão Capixaba, variedade clonal de café conilon tolerante à seca, desenvolvida para o estado do Espírito Santo. **Revista Ceres,** v. 273, p. 555-560, 2000.

GANKO, E. W. et al. Evidence for the contribution of LTR retrotransposons to *C. elegans* gene evolution. **Molecular Biology and Evolution,** v. 20, n. 11, p. 1925-1931, Nov 2003.

_____. LTR retrotransposon-gene associations in *Drosophila melanogaster*. **Journal of Molecular Evolution,** v. 62, n. 1, p. 111-120, Jan 2006.

GENTLEMAN, R. C. et al. Bioconductor: open software development for computational biology and bioinformatics. **Genome Biology,** v. 5, n. 10, 2004.

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

HAYASHI, Y. et al. Can an arbitrary sequence evolve towards acquiring a biological function? **Journal of Molecular Evolution,** v. 56, n. 2, p. 162-168, Feb 2003.

HILLMAN, R. T.; GREEN, R. E.; BRENNER, S. E. An unappreciated role for RNA surveillance. **Genome Biology,** v. 5, n. 2, 2004.

HOLBROOK, J. A. et al. Nonsense-mediated decay approaches the clinic. **Nature Genetics,** v. 36, n. 8, p. 801-808, 2004.

HUDSON, M. E.; LISCH, D. R.; QUAIL, P. H. The FHY3 and FAR1 genes encode transposase-related proteins involved in regulation of gene expression by the phytochrome A-signaling pathway. **Plant J**., v. 34, p. 453-471, 2004.

JURKA, J. et al. Repbase update, a database of eukaryotic repetitive elements. **Cytogenetic and Genome Research,** v. 110, n. 1-4, p. 462-467, 2005.

KROM, N.; RECLA, J.; RAMAKRISHNA, W. Analysis of genes associated with retrotransposons in the rice genome. **Genetica,** v. 134, n. 3, p. 297-310, Nov 2008.

LAMBA, J. K. et al. Nonsense mediated decay downregulates conserved alternatively spliced ABCC4 transcripts bearing nonsense codons. **Human Molecular Genetics,** v. 12, n. 2, p. 99-109, Jan 2003.

LAMBEIN, I. et al. Decay kinetics of autogenously regulated CGS1 mRNA that codes for cystathionine gamma-synthase in Arabidopsis thaliana. **Plant and Cell Physiology,** v. 44, n. 9, p. 893-900, Sep 2003.

LEI, X. D.; CHAPMAN, B.; HANKINSON, O. Loss of CYP1A1 messenger RNA expression due to nonsense-mediated decay. **Molecular Pharmacology,** v. 60, n. 2, p. 388-393, Aug 2001.

LEJEUNE, F.; MAQUAT, L. E. Mechanistic links between nonsense-mediated mRNA decay and pre-mRNA splicing in mammalian cells. **Current Opinion in Cell Biology,** v. 17, n. 3, p. 309-315, 2005.

LOPES, F. R. 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, Apr 2008.

LORENC, A.; MAKALOWSKI, W. Transposable elements and vertebrate protein diversity. **Genetica,** v. 118, n. 2-3, p. 183-191, Jul 2003.

MUEHLBAUER, G. J. et al. A hAT superfamily transposase recruited by the cereal grass genome. **Molecular Genetics and Genomics**, v. 9, p. , 2006.

NEKRUTENKO, A.; LI, W. H. S. Transposable elements are found in a large number of human protein-coding genes. **Trends in Genetics,** v. 17, n. 11, p. 619-621, 2001.

NOENSIE, E. N.; DIETZ, H. C. A strategy for disease gene identification through nonsense-mediated mRNA decay inhibition. **Nature Biotechnology,** v. 19, n. 5, p. 434-439, May 2001.

PIRIYAPONGSA, J. et al. Evaluating the protein coding potential of exonized transposable element sequences. **Biology Direct,** v. 2, Nov 2007.

SAKAI, H.; TANAKA, T.; ITOH, T. Birth and death of genes promoted by transposable elements in Oryza sativa. **Gene,** v. 392, n. 1-2, p. 59-63, May 2007.

VIEIRA, L. G. E. E. A. Brazilian Coffee Genome Project: an EST-based genomic resource. **Brazilian Journal of Plant Physiology**. 18**:** 95-108 p. 2006.

VIGNIER, N. et al. Nonsense-Mediated mRNA Decay and Ubiquitin-Proteasome System Regulate Cardiac Myosin-Binding Protein C Mutant Levels in Cardiomyopathic Mice. **Circulation Research,** v. 105, n. 3, p. 239-U89, Jul 2009.

WAGNER, E.; LYKKE-ANDERSEN, J. mRNA surveillance: the perfect persist. **Journal of Cell Sciences**, v. 115, p. 3033-3038, 2002

WU, J. Q. et al. Nonsense-mediated mRNA decay (NMD) silences the accumulation of aberrant trypsin proteinase inhibitor mRNA in *Nicotiana attenuata*. **Plant Journal,** v. 51, n. 4, p. 693-706, Aug 2007.

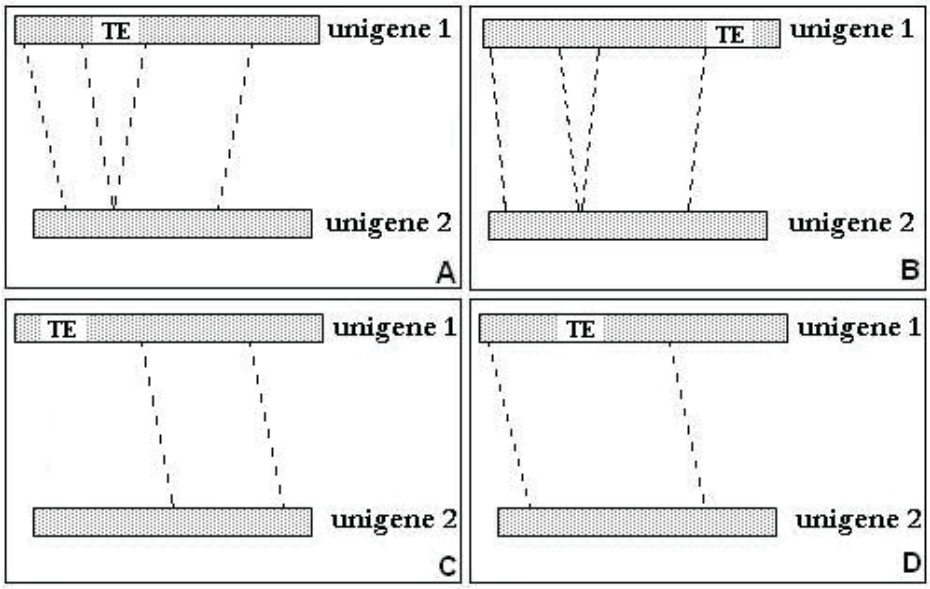


Figure 1. Examples of alignments between unigenes containing TEs and the highly related unigenes using sequence similarity searches by BLASTn. **A.** putative alternative *splicing* event. Non-conclusive *splicing* evidences: **B.** the unigenes present one or more similarity regions, however, these regions finish before the start or begin after the TE insertion; **C.** the similarity regions are distant from the TE insertion site. **D.** TE-cassette is within the similarity region between the unigenes.

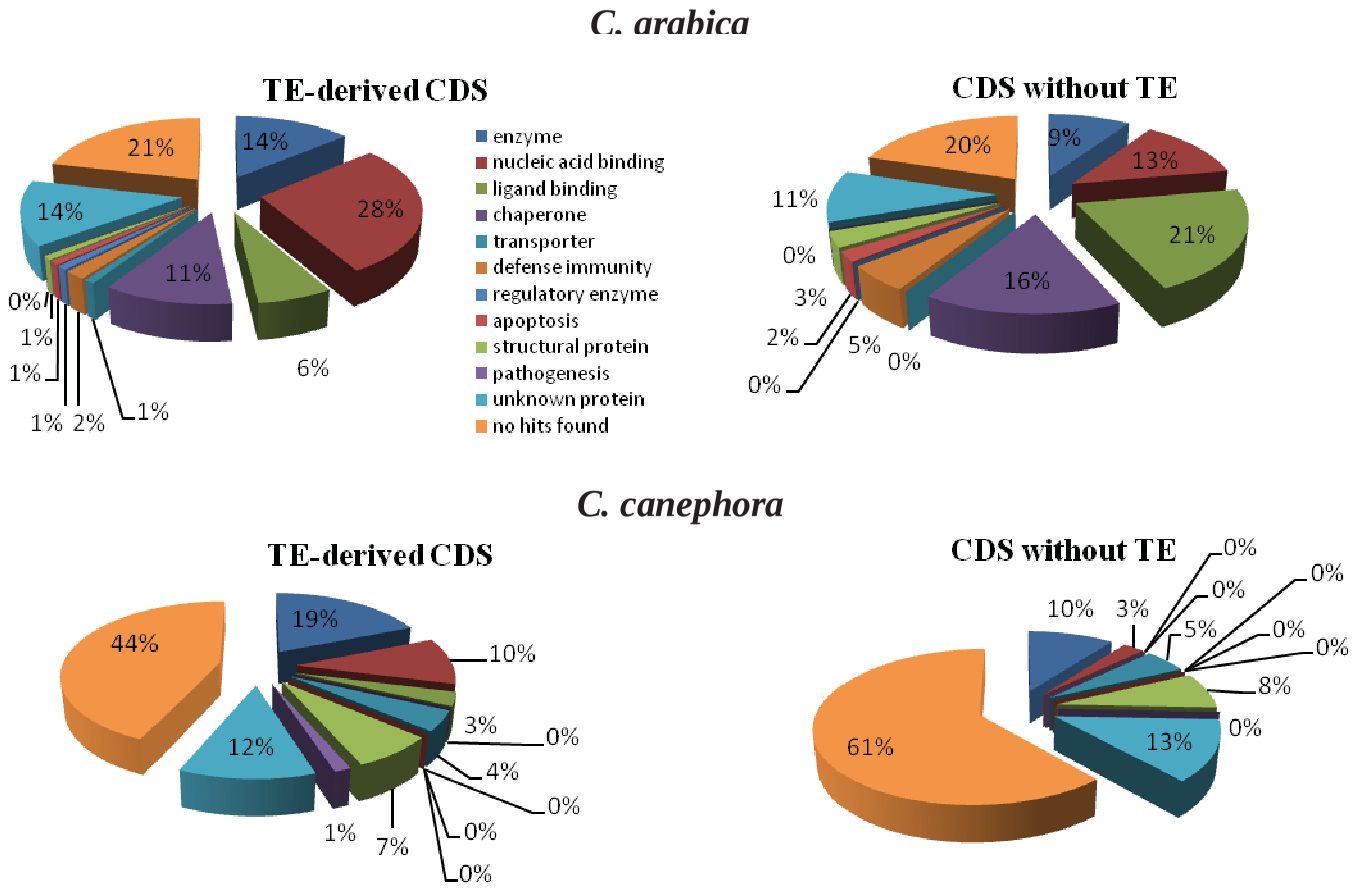


Figure 2. Frequency of TE-derived CDSs and CDSs without TEs, closely related by sequence similarity for distinct functional categories (data set available in Supplementary Material Tables S1-S3).

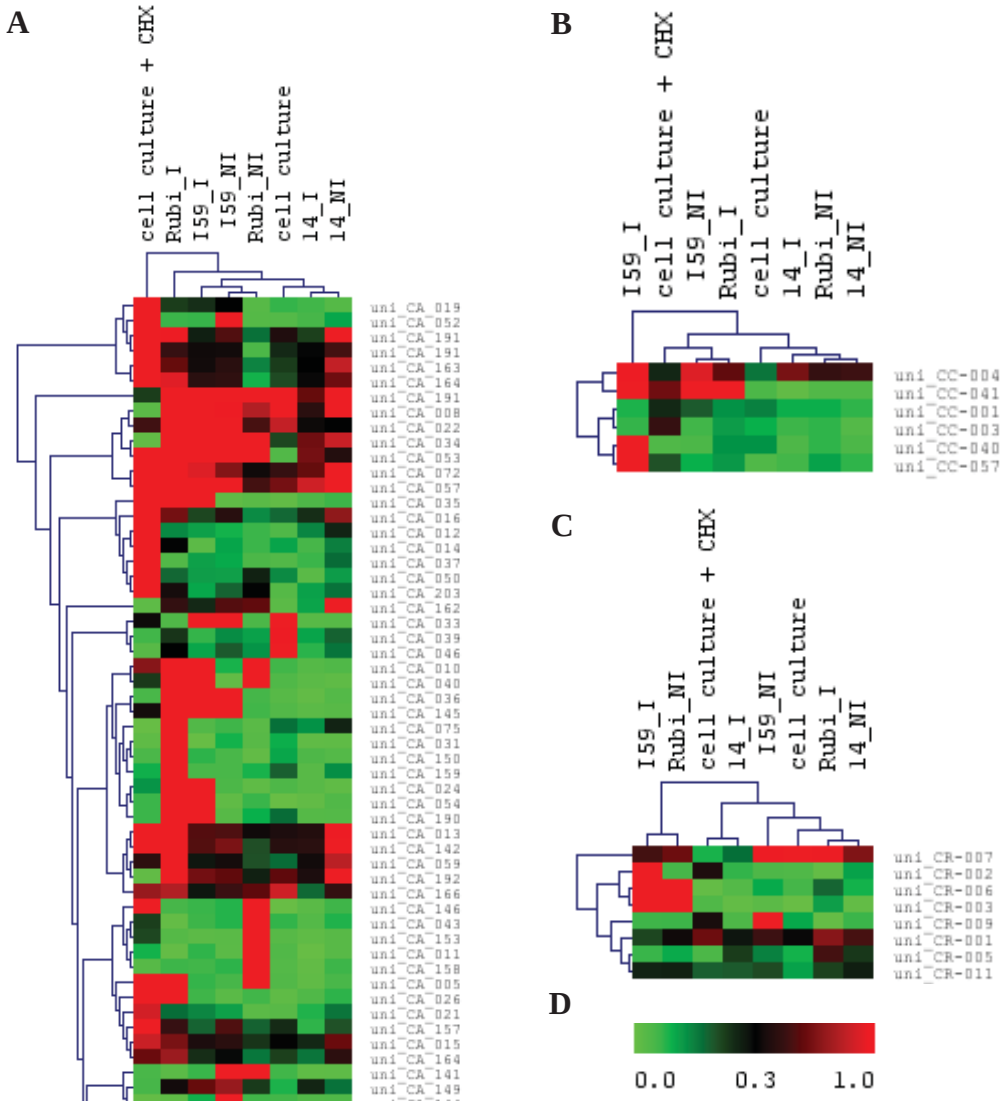


Figure 3. Expression profiles of 109 TE-derived CDSs. This analysis includes 99 transcripts from *C. arabica* **(A)**, six from *C. canephora* **(B)** and eight from *C. racemosa* **(C)** using *heatmap*. **(D)** Caption. Left: results obtained by macroarrays. Lines= cDNA clones (arbitrary identification plus protein name), column: distinct samples from *C. arabica* (embriogenic callus + 300 µg/mL of CHX; callus; I59_I and I59_NI: cultivar tolerant (Iapar59) irrigated or not irrigated, respectively; Rubi_I and Rubi_NI: cultivar sensitive (Rubi) irrigated or not irrigated, respectively) and *C. canephora* (14_I and 14_NI: clone var. Conillon tolerant (14) irrigated or not irrigated, respectively). Right: Digital *northern*. Lines= clones in the same ordination of that presented in the macroarrays. Column: cDNA libraries identified as in Lopes *et al.* (2008) of the three *Coffea* species studied here.

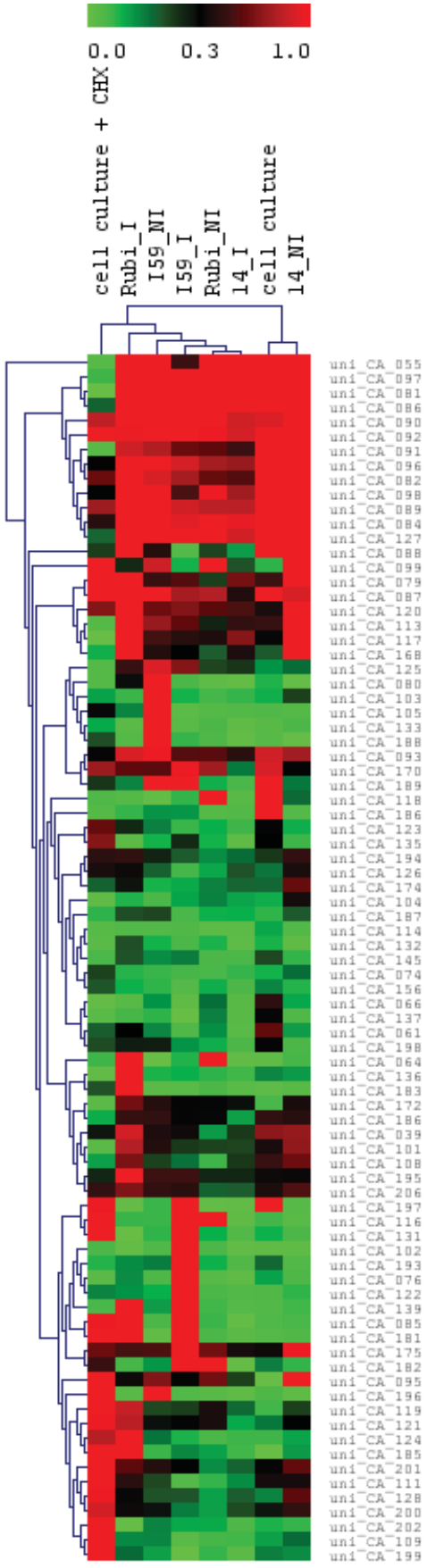


Figure 4. Expression profiles of 83 CDSs without TE-cassettes identified in the transcriptome from *C. arabica*. Caption as presented in Figure 3.

Supplementary Material

Legends: Tables S1-S3

unigenes containing TE-cassette insertions ("query"): arbitrary identification of the unigenes identified by RepeatMasker (Lopes et al., 2008) and tBLASTx (this study), length followed by localization of the TE-cassette insertion site into unigene and results of the BLASTx searches of the unigenes against nr databank.

unigenes related those containing TE by BLASTn comparisons ("subject"): arbitrary identification of the unigene, length followed by regions of similarity between "query"/"subject" (between brackets); results of the BLASTx searches of the unigenes against nr databank (same= means unigenes with same molecular function detected for the "query"). Unigenes marked in green means putative alternative *splicing* events (Figure 1A, see main text); unigenes marked in blue means that the unigenes present one or more similarity regions, however, these regions finish before of the initial or begin after the TE insertion (Figure 1B) and, finally, unigene marked in red means *EST* cluster that also harbor TE-cassette as the *query* (Figure 1D).

Table S1. List of unigenes containing or not TE-cassette insertions in *C. arabica*.

unigenes containing TE-cassette insertions ("query")			unigenes related those containing TE by BLASTn comparisons ("subject")		
Identification	unigene length (TE insertion)	First protein hit in BLASTx searches (Genbank accession #)	Identification	unigene lenght (bp) [regions of similarity between "query"/"subject"]	First protein hit in BLASTx searches (Genbank accession #)
unigenes identified by RepeatMasker					
uni_CA_001	898 (585-641)	unknown protein (15230124)	uni_CA_032	1,039 [351-898/13-561]	unknown (15230124)
uni_CA_002	689 (502-623)	no hits found	uni_CA_060	724 [1-126/445-571]	no hits found
uni_CA_003	679 (192-280)	no hits found	uni_CA_061	671 [410-652/431-671]	no hits found
			uni_CA_062	718 [1-184/273-89]	unknown
			uni_CA_063	657 [425-574/219-369]	no hits found
			uni_CA_064	776 [1-216/166-381 ; 231-617/378-761]	unknown (21536582)
			uni_CA_065	705 [410-640/453-682]	no hits found
			uni_CA_066	593 [425-679/322-574]	no hits found
			uni_CA_067	915 [254-506/536-790]	no hits found
uni_CA_004	843 (386-461)	Calreticulin 1 precursor (11131631)	uni_CA_039	1,655 [1-329/696-1024 ; 331-586/1122-1380 ; 615-736/1534-1655]	Calreticulin 1 precursor (11131904)
			uni_CA_020	832 [26-326/6-308 ; 326-586/394-658]	Calreticulin 1 precursor (31249766)
			uni_CA_068	865 [123-202/2-81 ; 201-260/160-219 ; 258-329/304-375]	Calreticulin 1 precursor (11131631)
uni_CA_005	690 (62-243)	DRL1 (deformed roots and leaves 1) (23504265)	uni_CA_069	416 [1-162/35-198]	DRL1 (deformed roots and leaves 1) (15222972)
uni_CA_006	649 (14-103)	no hits found	uni_CA_049	1,014 [1-171/293-463 ; 173-584/603-1014]	no hits found
			uni_CA_070	1,087[1-171/366-736 ; 194-593/685-1080]	no hits found
Uni_CA_007	649 (164-219)	no hits found	uni_CA_049	1,014 [1-171/293-463 ; 173-584/603-1014]	no hits found
			uni_CA_070	1,087[1-171/366-736 ; 194-593/685-1080]	no hits found
uni_CA_008	763 (40-75)	probable kinesin heavy chain (7487566)	no hits found		
uni_CA_009	625 (141-231)	no hits found	no hits found		
uni_CA_010	454 (184-250)	no hits found	no hits found		
uni_CA_011	528 (64-234)	no hits found	uni_CA_013	618 [38-201/67-231 ; 113-495/229-613]	no hits found
			uni_CA_071	854[1-374/407-773]	Putative Cer1 (10716610)
			uni_CA_072	998 [1-202/796-998]	Putative Cer1 (37535278)
			uni_CA_073	2,116 [1-83/2034-2116]	Putative Cer1 (30678265)
uni_CA_012	822 (458-560)	unknown (15230463)	uni_CA_074	875 [67-609/78-621]	unknown (15230463)
uni_CA_013	618 (223-352)	no hits found	uni_CA_011	618 [67-231/38-201 ; 229-613/113-495]	no hits found
			uni_CA_071	854 [66-231/443-608 ; 229-492/519-773]	Putative Cer1 (10716610)
			uni_CA_072	998 [66-231/832-997 ; 229-319/908-998]	Putative Cer1 (37535278)
uni_CA_014	861 (458-560)	unknown (18401997)	no hits found		
uni_CA_015	651 (575-646)	unknown (15235843)	no hits found		

Table S1. Continuation

uni_CA_016	852 (379-483)	unknown (30678837)	no hits found		
uni_CA_017	751 (272-310)	protein kinase family - Ser/Thr protein kinase (15220275)	no hits found		
uni_CA_018	864 (624-669)	no hits found	no hits found		
uni_CA_019	685 (321-402)	no hits found	uni_CA_058	997 [18-647/369-997]	no hits found
			uni_CA_075	216 [449-2/663-216]	no hits found
uni_CA_020	832 (454-529)	putative calreticulin precursor (31249766)	uni_CA_039	1,655 [6-314/721-1027 ; 397-693/1120-1415]	calreticulin precursor (11131904)
			uni_CA_004	865 [6-308/26-326 ; 394-658/326-586]	calreticulin precursor (11131631)
			uni_CA_068	843 [102-184/1-81 ; 183-242/160-219 ; 240-397/304-461]	calreticulin precursor (11131631)
			uni_CA_076	316 [590-765/3-178]	no hits found
uni_CA_021	792 (603-644)	aldo/keto reductase family (15219786)	no hits found		
uni_CA_022	580 (81-159)	no hits found	uni_CA_023	721 [37-577/153-693]	no hits found
			uni_CA_077	808 [37-338/497-808]	no hits found
uni_CA_023	721 (197-275)	no hits found	uni_CA_022	580 [153- 693/37-577]	no hits found
			uni_CA_077	808 [1-454/345-808]	no hits found
uni_CA_024	836 (759-821)	unknown (15221754)	no hits found		
uni_CA_025	786 (603-770)	no hits found	no hits found		
uni_CA_026	871 (97-186)	no hits found	no hits found		
uni_CA_027	885(403-505)	no hits found	no hits found		
uni_CA_028	1,989 (809-892)	Probable MAP kinase phosphatase (13540262)	no hits found		
uni_CA_029	1,292 (504-542)	no hits found	uni_CA_078	891 [15-867/33-891]	no hits found
			uni_CA_079	654 [15-366/25-369 ; 836-1087/382-639]	no hits found
uni_CA_030	934 (29-85)	polyubiquitin 5 (70645)	uni_CA_080	1,544 [77-738/644-1305]	ubiquitin (902586)
			uni_CA_081	902 [66-771/38-742]	polyubiquitin (1332579)
			uni_CA_082	873 [51-692/47-688]	pentameric ubiquitin (602076)
			uni_CA_083	1,729 [66-692/57-683]	polyubiquitin (100934)
			uni_CA_084	793 [69-738/1-670]	ubiquitin (902586)
			uni_CA_085	1,342 [79-774/532-1227]	ubiquitin (70644)
			uni_CA_086	857 [127-774/1-648]	hexameric polyub. (170352)
			uni_CA_087	886 [69-681/1-615]	polyubiquitin (30680052)
			uni_CA_088	1,097 [127-774/335-981]	hexameric polyubiquitin (170352)
			uni_CA_089	1,540 [79-762/579-1262]	ubiquitin (4115337)
			uni_CA_090	695 [77-440/275-640]	polyubiquitin (1332579)
			uni_CA_091	1,315 [77-740/530-1193]	polyubiquitin (3126967)
			uni_CA_092	960 [75-737/86-748]	polyubiquitin (1332579)
			uni_CA_093	1,575 [75-737/846-1507]	polyubiquitin (1332579)
			uni_CA_094	793 [148-771/603-1226]	polyubiquitin (1332579)
			uni_CA_095	1,410 [75-680/62-668]	polyubiquitin (70645)

Table S1. Continuation

			uni_CA_096	1,482 [75-737/96-758]	polyubiquitin (481477)
			uni_CA_097	762 [66-738/46-717]	polyubiquitin (421929)
			uni_CA_098	822 [136-755/123-740]	polyubiquitin (1332579)
uni_CA_031	762 (379-480)	fertility restorer (22128587)	No hits found		
uni_CA_032	1,039 (247-303)	6b-interacting protein 1 (15230124)	uni_CA_001	898 [13-561/351-898]	unknown (15230124)
uni_CA_033	1,118 (997-1106)	transfactor-like (30699418)	No hits found		
uni_CA_034	1,135 (146-233)	heavy-metal-associated domain-containing protein (42572261)	uni_CA_099	501 [702-1135/1-434]	same (30679432)
			uni_CA_100	339 [1023-1135/52-164]	no hits found
uni_CA_035	1.093	unknown (42569311)	uni_CA_101	859 [1-713/148-859]	unknown (26452742)
			uni_CA_102	896 [1-650/58-706]	unknown (30678837)
uni_CA_036	1,990 (1462-1590)	rab GDP dissociation inhibitor (7489133)	uni_CA_103	307 [58-665/50-656]	No hits found
uni_CA_037	1,316 (185-244)	PSTVd RNA-binding protein Virp1a (10179602)	uni_CA_104	845 [298-1080/1-784]	no hits found
			uni_CA_051	2,655 [53-264/1740-1951 ; 360-503/1951-2094 ; 590-1084/2093-2588]	Same
			uni_CA_105	367 [946-1312/1-367]	no hits found
			uni_CA_106	813 [53-184/448-579]	viroid RNA-binding protein (19171209)
uni_CA_038	2,054 (2-61)	glyceraldehyde-3-phosphate dehydrogenase, cytosolic (18072805)	uni_CA_107	1,298 [46-364/359-680 ; 373-978/690-1298]	same
			uni_CA_108	1,139 [1013-1158/1-146 ; 1155-1436/356-637 ; 1467-1937/660-1132]	same
			uni_CA_109	721 [1314-1436/17-139 ; 1432-1554/286-406]	same
			uni_CA_110	650 [46-128/558-650]	no hits found
uni_CA_039	1,655 (1177-1252)	calreticulin precursor (11131904)	uni_CA_111	876 [146-711/304-865]	calreticulin 2 (15217459)
			uni_CA_020	832 [721-1027/6-314]	calreticulin percusor (31249466)
			uni_CA_004	843 [696-1024/1-329 ; 1122-1380/331-586 ; 1534-1655/615-736]	calreticulin percusor (11131631)
			uni_CA_076	379 [1311-1461/1-149]	No hits found
			uni_CA_112	823 [1460-658-809]	molecular chaperone Hsp90-1 (38154489)
			uni_CA_068	865 [817-897/1-81 ; 953-1027/304-378 ; 1021-1120/458-557]	calreticulin percusor (11131631)
			uni_CA_113	162 [1528-1632/54-158]	no hits found
uni_CA_040	2,286 (750-817)	multidomain cyclophilin type peptidyl-prolyl cis-trans isomerase - CYP63 (15229425)	uni_CA_114	860 [1446-1695/62-311]	same
uni_CA_041	961 (87-142)	SRG1 (senescence-related gene 1), oxidoreductase, 2OG-Fe(II) oxygenase family (15219988)	uni_CA_115	936 [1-897/44-936]	same
			uni_CA_116	2,265 [139-961/1-824]	same
			uni_CA_117	900 [162-934/1-773]	same
			uni_CA_118	773 [380-961/1-583]	Same
			uni_CA_119	691 [641-961/1-322]	SRG1 homolog (25285681)
			uni_CA_120	677 [737-961/1-224]	SRG1 like protein (26451337)
			uni_CA_121	482 [870-961/3-95]	SRG1 like protein (26451337)
uni_CA_042	1,102 (46-118)	ribosomal protein L7 (445613)	uni_CA_122	765 [102-803/60-765]	same
			uni_CA_123	877 [102-280/107-285 ; 278-809/347-877]	same
			uni_CA_124	1,143 [101-838/219-956]	same

Table S1. Continuation

uni_CA_043	1,998 (349-447)	ubiquitinating enzyme (49328017)	No hits found		
uni_CA_044	656 (382-445)	no hits found	No hits found		
uni_CA_045	909 (373-451)	no hits found	No hits found		
uni_CA_046	861 (415-495)	rust resistance Rp1-D-like protein (13310480)	uni_CA_125	911 [307-422/1-117 ; 503-856/135-488]	similar to NBS-LRR type resistance protein (28564572)
			uni_CA_126	701 [503-861/59-416]	putative NBS-LRR type resistance protein (13872974)
			uni_CA_127	820 [112-416/91-396 ; 497-861/411-774]	putative disease resistant protein rga4 (32470648)
			uni_CA_128	874 [122-352/295-526]	Vrga1 (6606266)
uni_CA_047	1,075 (433-512)	pre-mRNA splicing factor cwc15/Cwc15 cell cycle control (21592359)	No hits found		
uni_CA_048	714 (480-527)	CONSTANS-like protein (18390719)	uni_CA_129	890 [27-714/3-681]	same
uni_CA_049	1,014 (282-366)	no hits found	uni_CA_070	1,081 [1-535/74-608 ; 531-625/616-710 ; 624-685/1014-1071]	no hits found
			uni_CA_007	649 [293-463/1-171 ; 603-173/1014-584]	no hits found
uni_CA_050	1,729 (1323-1372)	sucrose synthase (29289943)	uni_CA_130	660 [731-1391/4-660]	same
			uni_CA_131	259 [1489-1729/23-252]	no hits found
			uni_CA_132	286 [1585-1729/135-274]	no hits found
uni_CA_051	2,655 (1872-1931)	PSTVd RNA-binding protein Virp1a (10179602)	uni_CA_133	881 [113-419/9-315 ; 1043-1604/320-881]	same
			uni_CA_104	845 [1951-2094/63-206 ; 2093-2635/293-834]	no hits found
			uni_CA_037	1,316 [2093-2588/590-1084]	same
			uni_CA_106	813 [1651-1871/358-579]	same
			uni_CA_134	872 [570-1149/55-635]	same
			uni_CA_105	367 [2449-2588/1-139]	no hits found
uni_CA_052	759 (520-590)	no hits found	uni_CA_135	870 [13-244/591-818]	no hits found
uni_CA_053	1,274 (39-109)	transmembrane MLO family protein (18398444)	uni_CA_136	764 [920-1234/14-354]	Same
			uni_CA_137	872 [928-1234/48-354]	transmembrane MLO protein (201398)
uni_CA_054	1,017 (606-685)	unknown (30681632)	No hits found		
uni_CA_055	1,208 (185-244)	universal stress protein (USP) family protein (30681955)	uni_CA_138	1,166 [38-101/5-68 ; 277-607/63-393 ; 606-701/477-572 ; 702-1140/663-1105]	Same
uni_CA_056	750 (469-550)	Unknown protein (18413696)	No hits found		
uni_CA_057	439 (119-201)	Unknown protein (18417290)	No hits found		
uni_CA_058	997 (672-754)	no hits found	uni_CA_019	685 [369-997/18-647]	no hits found
			uni_CA_075	216 [797-994/2-197]	no hits found
uni_CA_059	1,185 (798-876)	nucleosome assembly protein (NAP) (15221298)	uni_CA_139	676 [10-375/33-398 ; 561-839/398-676]	nucleosome/chromatin assembly factor A (46805241)

Table S1. Continuation

unigenes identified by tBLASTx					
uni_CA_140	786 (527-694)	galactokinase GHMP kinase-like (2326372)	No hits found		
uni_CA_141	829 (649-798)	GHMP kinase-like protein (42409310)	No hits found		
uni_CA_142	856 (429-575)	SC35-like putative splicing factor (31249706)	uni_CA_145	798 (337-427/1-91 ; 427-705/205-483 ; 714-856/480-623)	Same
uni_CA_143	781 (485-631)	SC35-like putative splicing factor (31249706)	uni_CA_146	1613 (1-192/163-354 ; 404-778/350-721)	SC35-like splicing factor (9843661)
			uni_CA_147	1177 (1-192/119-310 ; 404-778/306-677)	Same
			uni_CA_142	856 (418-479/441-502)	Same
			uni_CA_145	798 (418-479/219-280)	Same
uni_CA_144	662 (254-409)	SC35-like putative splicing factor (31249706)	uni_CA_167	760 (124-497/188-566)	SC35-like splicing factor (30687014)
			uni_CA_148	649 (192-269/546-623)	putative pre-mRNA splicing factor (47497118)
uni_CA_145	798 (201-353)	SC35-like putative splicing factor (31249706)	uni_CA_142	856 (1-91/337-427 ; 205-483/427-705 ; 480-623/714-856)	Same
			uni_CA_143	781 (219-280/418-479)	Same
			uni_CA_146	1613 (219-280/364-425)	SC35-like splicing factor (9843661)
			uni_CA_147	1177 (219-286/320-387)	Same
uni_CA_146	1,613 (352-498)	SC35-like putative splicing factor	uni_CA_147	1177 (59-957/15-913 ; 954-1201/929-1177 ; 1170-1292/1005-1129)	Same
			uni_CA_143	781 (163-354/1-192 ; 350-721/404-778)	Same
			uni_CA_142	856 (364-425/441-502)	Same
			uni_CA_145	798 (364-425/219-280)	Same
uni_CA_147	1,177 (308-454)	SC35-like putative splicing factor (31249706)	uni_CA_146	1613 (15-913/59-957 ; 929-1177/954-1201 ; 1005-1129/1170-1292)	SC35-like splicing factor (9843661)
			uni_CA_143	781 (306-677/404-778 ; 119-310/1-192)	Same
			uni_CA_142	856 (320-387/441-508)	Same
			uni_CA_145	798 (320-387/219-286)	Same
uni_CA_148	1,195 (608-763)	SC35-like putative splicing factor (47497118)	uni_CA_168	768 (13-692/86-795)	unknown (12002297) - SC35
			uni_CA_169	756 (13-263/86-336 ; 235-361/341-467)	no hits found
			uni_CA_144	662 (546-623/192-269)	SC35-like splicing factor (30687014)
uni_CA_149	847 (414-824)	protein F21D18.16 (25405918)	No hits found		
uni_CA_150	951 (303-500)	heat shock cognate 70 kd protein (123650)	uni_CA_153	2475 (40-951/141-1053)	heat shock cognate protein 70 (26985223)
			uni_CA_154	2209 (87-951/51-916)	heat shock cognate protein 70 (26985221)
			uni_CA_152	901 (87-817/151-877)	heat shock cognate protein 70 (26985221)
			uni_CA_151	2340 (87-951/122-987)	dnaK-type molecular chaperone hsp70 (1076746)
			uni_CA_170	768 (86-280/342-536)	heat shock cognate protein 70 (26985219)
			uni_CA_171	751 (189-316/247-374 ; 363-490/421-548)	heat shock protein (6969976)
			uni_CA_172	768 (184-280/199-295)	molecular chaperone HSP71.2 (1076529)

Table S1. Continuation

uni_CA_151	2,340 (338-535)	dnaK-type molecular chaperone hsp70 (1076746)	uni_CA_153	2475 (122-1974/188-2040)	heat shock cognate protein 70 (26985223)
			uni_CA_154	2209 (114-1825/43-1754 ; 1907-1977/1836-1906)	heat shock cognate protein 70 (26985221)
			uni_CA_173	722 (1353-1452/1-100 ; 1415-1439/122-98 ; 1429-1993/116-683)	Same
			uni_CA_174	402 (1892-2147/1-248 ; 2173-2306/269-402)	heat shock protein (6969976)
			uni_CA_150	951 (122-987/87-951)	heat shock cognate 70 kd protein (123650)
			uni_CA_152	901 (114-816/143-843)	heat shock cognate protein 70 (26985221)
			uni_CA_170	768 (122-449/343-670)	Same
			uni_CA_175	804 (880-1284/12-415)	molecular chaperone Nthsp70 (100335)
			uni_CA_176	900 (1289-1506/81-297 ; 1602-1857/340-598)	heat shock protein hsp70 (15230534)
			uni_CA_177	804 (947-1026/386-465 ; 1112-1259/549-696)	heat shock protein (6969976)
			uni_CA_178	751 (572-678/595-701)	heat shock cognate 70 kd protein (123650)
			uni_CA_179	1,888 (546-639/56-149)	Luminal binding protein 5 precursor (729623)
			uni_CA_171	969 (546-639/650-743)	Luminal binding protein 5 precursor (729623)
			uni_CA_172	768 (221-300/201-280 ; 554-619/534-599)	heat shock protein hsp70 (15230534)
			uni_CA_180	824 (569-625/220-276)	Luminal binding protein precursor (1346172)
			uni_CA_181	2,591 (569-625/825-881)	heat shock protein (6969976)
				1,238 (575-639/10-74)	Luminal binding protein 5 precursor (729623)
uni_CA_152	901 (367-564)	non-cell-autonomous heat shock cognate protein 70 (26985221)	uni_CA_154	2,209 (101-901/1-806)	Same
			uni_CA_153	2475 (151-878/188-918)	heat shock cognate protein 70 (26985223)
					Same
			uni_CA_150	951 (151-877/87-817)	dnaK-type molecular chaperone hsp70 (1076746)
			uni_CA_151	2,340 (143-843/114-816)	heat shock protein (6969976)
			uni_CA_177	804 (217-380/409-572)	
uni_CA_153	2,475 (404-601)	cell-autonomous heat shock cognate protein 70 (26985223)	uni_CA_150	951 (143-1053/42-951)	heat shock cognate protein 70 (26985221)
			uni_CA_154	2,209 (188-1861/51-1724 ; 1899-2045/1762-1908)	heat shock cognate protein 70 (26985221)
			uni_CA_151	2,340 (188-2040/122-1974)	dnaK-type molecular chaperone hsp70 (1076746)
			uni_CA_152	901 (188-918/151-878)	heat shock cognate protein 70 (26985221)
			uni_CA_173	722 (1419-1518/1-100 ; 1508-1852/129-477)	same
			uni_CA_175	804 (983-1335/48-400)	molecular chaperone Nthsp70 (100335)
			uni_CA_176	900 (1358-1584/84-309 ; 1670-1861/342-533 ; 1943-2022/615-694)	heat shock protein hsp70 (15230534)
			uni_CA_183	711 (1821-1853/69-101 ; 1868-2095/116-344)	heat shock protein 70 (21327033)
			uni_CA_177	804 (939-1059/311-432 ; 1178-1325/549-696)	heat shock protein (6969976)
			uni_CA_171	721 (290-417/247-374 ; 464-567/421-524)	heat shock cognate 70 kd protein (123650)
			uni_CA_170	768 (187-510/342-665)	molecular chaperone Nthsp70 (100335)
			uni_CA_174	402 (1958-2040/1-83)	heat shock cognate 70 kd protein (123650)

Table S1. Continuation

uni_CA_154	2,209 (267-464)	non-cell-autonomous heat shock cognate protein 70 (26985221)	uni_CA_152	901 (1-806/101-901)	same
			uni_CA_153	2,475 (51-1724/188-1861 ; 1762-1908/1899-2045)	heat shock cognate protein 70 (26985223)
			uni_CA_151	2,340 (43-1754/114-1825 ; 1836-1906/1907-1977)	dnaK-type molecular chaperone hsp70 (1076746)
			uni_CA_176	900 (1217-1472/80-334 ; 1523-1975/332-790)	heat shock protein hsp70 (15230534) heat shock protein (6969976)
			uni_CA_177	804 (555-1269/64-777)	heat shock cognate 70 kd protein (123650)
			uni_CA_150	951 (51-916/87-956)	No hits found
			uni_CA_184	277 (1976-2140/110-274)	molecular chaperone Nthsp70 (100335) dnaK-type molecular chaperone hsp70
			uni_CA_175	804 (846-1208/48-410 ; 1281-1381/483-583)	(1076746)
			uni_CA_173	722 (138-271/189-322)	molecular chaperone Nthsp70 (100335) chaperone HSP71.2 (1076529)
			uni_CA_170	768 (117-301/409-593)	heat shock cognate 70 kd protein (123650)
			uni_CA_172	768 (138-271/189-322)	heat shock cognate 70 kd protein (123650)
			uni_CA_171	751 (133-271/227-365)	Luminal binding protein 5 precursor (729623)
			uni_CA_174	402 (1836-1906/16-86)	Luminal binding protein 5 precursor (729623)
			uni_CA_178	1,888 (475-559/56-140)	
			uni_CA_179	969 (475-650/559-734)	
uni_CA_155	884 (12-482)	EIL3 (14280044)	uni_CA_185	840 (40-190/674-820)	EIN3-like protein (15425735)
			uni_CA_157	2,580 (63-370/875-1182)	EIL2 (30016896)
uni_CA_156	1123	EIL3 (30016898)	uni_CA_157	2,580 (298-423/13-142 ; 421-1123/205-907)	EIL2 (30016896)
			uni_CA_186	1,297 (735-926/12-203)	EIL2 (30016896)
			uni_CA_187	753 (650-840/96-294)	EIL2 (14280042)
			uni_CA_188	806 (644-704/97-157 ; 791-918/247-374)	EIN3-like protein (15425735)
			uni_CA_185	840 (644-704/247-307 ; 791-918/397-524)	EIN3-like protein (15425735)
uni_CA_157	2,580 (827-1294)	EIL2 (30016896)	uni_CA_186	1,297 (519-710/12-203 ; 1405-2481/195-1270)	EIL2 (30016896)
			uni_CA_156	1,123 (13-142/298-423 ; 205-907/421-1123)	EIL3 (30016898)
			uni_CA_155	884 (875-1182/63-370)	EIL3 (14280044)
			uni_CA_188	806 (428-488/97-157 ; 575-702/247-374)	EIN3-like protein (15425735)
			uni_CA_185	840 (428-488/247-307 ; 575-702/397-524)	EIN3-like protein (15425735)
			uni_CA_187	753 (434-632/96-294)	EIL2 (14280042)
uni_CA_158	766 (2-739)	expressed protein (15219020)	uni_CA_189	657 (605-713/1-109)	expressed protein (15219020)
uni_CA_159	612 (44-307)	histone H3.2 (15236103)	uni_CA_162	734 (42-441/102-501)	Same
			uni_CA_160	945 (44-417/127-500)	Same
			uni_CA_161	726 (71-426/132-487)	histone H3 (15232146)
			uni_CA_190	723 (43-300/71-333 ; 329-426/362-459)	histone H3 (15232146)
			uni_CA_191	667 (43-300/67-324 ; 329-426/353-450)	histone H3(15232146)
			uni_CA_192	755 (43-285/59-301 ; 329-426/345-442)	histone H3(15232146)

Table S1. Continuation

uni_CA_160	667 (127-390)	histone H3 (15232146)	uni_CA_190	723 (11-667/20-678)	Same
			uni_CA_161	726 (52-477/89-514)	Same
			uni_CA_192	755 (64-477/56-469)	Same
			uni_CA_162	734 (95-450/131-486)	histone H3.2 (15236103)
			uni_CA_160	945 (227-465/286-524)	histone H3.2 (15236103)
			uni_CA_163	798 (67-450/41-424)	histone H3.2 (15236103)
			uni_CA_159	612 (67-324/43-300 ; 353-450/329-426)	histone H3.2 (15236103)
uni_CA_161	726 (105-365)	ENSANGP00000016056 histone (31210957)	uni_CA_192	755 (88-512/43-467)	histone H3 (15232146)
			uni_CA_190	723 (89-512/61-484)	histone H3 (15232146)
			uni_CA_191	667 (89-512/52-475)	histone H3 (15232146)
			uni_CA_160	945 (105-178/127-200 ; 255-502/277-524)	histone H3.2 (15236103)
			uni_CA_163	798 (132-487/69-424)	histone H3.2 (15236103)
			uni_CA_162	734 (104-487/103-486)	histone H3.2 (15236103)
			uni_CA_159	612 (132-487/71-426)	histone H3.2 (15236103)
uni_CA_162	734 (104-367)	histone H3.2 (15236103)	uni_CA_163	798 (102-501/40-439)	Same
			uni_CA_159	612 (102-501/42-441)	Same
			uni_CA_160	945 (104-515/127-538)	Same
			uni_CA_166	1,780 (104-515/82-493)	rubisco small subunit (24940138)
			uni_CA_190	723 (131-486/104-459)	histone H3 (15232146)
			uni_CA_191	667 (131-486/95-450)	histone H3 (15232146)
			uni_CA_192	755 (103-486/59-442)	histone H3 (15232146)
			uni_CA_161	726 (103-486/104-487)	histone H3 (15232146)
			uni_CA_164	1,599 (102-171/499-568)	putative Ruv DNA-helicase (7208771)
			uni_CA_165	1,837 (102-171/454-523)	putative Ruv DNA-helicase (7208771)
uni_CA_163	798 (96-359)	histone H3.2 (15236103)	uni_CA_159	612 (1-607/3-605)	Same
			uni_CA_162	734 (40-439/102-501)	Same
			uni_CA_160	945 (42-415/127-500)	Same
			uni_CA_161	723 (69-424/132-487)	histone H3 (15232146)
			uni_CA_191	726 (41-424/67-450)	histone H3 (15232146)
			uni_CA_192	667 (41-283/59-301 ; 327-424/345-442)	histone H3 (15232146)
uni_CA_164	1,599 (571-500)	putative Ruv DNA-helicase (7208771)	uni_CA_165	1,837 (45-573/1-528 ; 617-903/534-820 ; 952-1333/869-1250 ; 1333-1598/1374-1642)	Same
			uni_CA_193	1,356 (1333-1599/299-568 ; 1276-1333/13-70)	no hits found
			uni_CA_194	920 (607-715/944-1052 ; 712-771/1149-1208)	Same
			uni_CA_195	798 (657-835/1-179 ; 1162-1599/180-620)	Same
			uni_CA_196	787 (607-715/247-355 ; 712-771/452-511)	Same
			uni_CA_197	777 (1333-1403/295-365)	no hits found

Table S1. Continuation

uni_CA_165	1,837 (526-455)	putative Ruv DNA-helicase (7208771)	uni_CA_193	920 (1193-1374/13-194 ; 1373-1837/298-761)	no hits found
			uni_CA_164	1,599 (1-67/45-111 ; 120-528/165-573 ; 534-805/617-888 ; 870-1250/953-1333 ; 1374-1642/1333-1598)	Same
			uni_CA_195	798 (574-752/1-179 ; 1079-1250/180-351 ; 1374-1813/351-788)	Same
					no hits found
			uni_CA_197	777 (1371-1444/292-365)	Same
			uni_CA_196	787 (534-632/257-355 ; 629-688/452-511)	Same
			uni_CA_194	1,356 (534-632/954-1052 ; 629-688/1149-1208)	same
uni_CA_166	1,780 (82-345)	histone H3 (296083584)	uni_CA_162	734 (82-493/104-515)	Same

Table S2. List of unigenes containing or not TE-cassette insertions in *C. canephora*.

unigenes containing cassette-TE insertions ("query")			unigenes related those containing TE by BLASTn comparisons ("subject")		
Identification	unigene length in bp (TE insertion)	First protein hit in BLASTx searches (Genbank accession #)	Identification	unigene lenght (bp) [regions of similarity between "query"/"subject"]	First protein hit in BLASTx searches (Genbank accession #)
unigenes identified by RepeatMasker					
uni_CC_001	792 (171-252)	RNA Binding Protein 47 (9663769)	uni_CC_060	1,546 [16-785/88-869]	DNA binding protein ACBF-like (82621158)
			uni_CA_198	2,158 [37-160/113-236 e 214-785/290-867]	unknown protein (15226402)
uni_CC_002	811 (574-672)	no hits found	no hits found		
uni_CC_003	792 (12-99)	TTN9 (TITAN9) (22331190)	no hits found		
uni_CC_004	775 (316-425)	Ribosomal protein L21 family protein, putative 50S ribosomal protein L21 (115465385)	uni_CC_061	482 [126-402/9-276 ; 413-621/263-471]	same
uni_CC_005	815 (96-195)	Unknown (21536685)	no hits found		
uni_CC_006	707 (102-149)	adenylyl cyclase-like (57899387)	no hits found		
uni_CC_007	1,126 (492-567)	Cwf15/Cwc15 cell cycle control protein (92872088)	uni_CC_062	987 [114-212/8-106 ; 319-930/106-720]	Same
			uni_CC_063	908 [319-520/293-494 ; 114-203/195-284]	Same
			uni_CA_199	1087 [1-60/54-113 e 111-791/285-961]	Unknown protein (18379014)
uni_CC_008	807 (58-103)	Unknown (18379014)	no hits found		
uni_CC_009	631 (363-408)	Unknown (21554182)	uni_CC_019	486 [103-583/12-486]	hypothetical protein MA4_112I10.14 (102140015)
uni_CC_010	867 (479-566)	major latex protein homolog (3064039)	no hits found		
uni_CC_011	1,045 (59-105)	no hits found	no hits found		
uni_CC_012	102 (6-79)	no hits found	no hits found		
uni_CC_013	429 (271-427)	no hits found	no hits found		
uni_CC_014	487 (364-405)	no hits found	no hits found		
uni_CC_015	477 (383-476)	no hits found	no hits found		
uni_CC_016	506 (32-243)	no hits found	no hits found		
uni_CC_017	467 (28-85)	no hits found	no hits found		
uni_CC_018	486 (132-226)	no hits found	uni_CC_064	472(11-129/391-271)	no hits found
			uni_CC_045	995(115-174/158-217)	Capip2 (47558819)
			uni_CC_048	1,034 (116-174314-256)	13-lipoxygenase (1495804)
uni_CC_019	486 (272-317)	hypothetical protein MA4_112I10.14 (102140015)	uni_CC_009	631 [12-486/103-583]	unknown (21554182)
uni_CC_020	200 (8-58)	hypothetical protein (115640798)	uni_CC_065	200 [1-200/1-200]	same
uni_CC_021	511 (31-114)	no hits found	no hits found		
uni_CC_022	707 (132-168)	no hits found	uni_CC_066	526 [264-611/109-452]	no hits found

Table S2. Continuation

uni_CC_023	777 (521-608)	no hits found	uni_CC_067	669 [9-152/16-159 ; 638-757/159-278]	no hits found
uni_CC_024	842 (271-335)	no hits found	uni_CC_068	723 [285-627/11-354]	no hits found
			uni_CC_069	874 [258-426/71-239]	no hits found
uni_CC_025	883 (580-733)	14-3-3 homologues mediates signal transduction by binding to phosphoserine-containing proteins (92891732)	No hits found		
uni_CC_026	886 (17-69)	no hits found	No hits found		
uni_CC_027	793 (676-770)	structural constituent of ribosome (18401997)	No hits found		
uni_CC_028	331 (14-102)	no hits found	No hits found		
uni_CC_029	529 (46-161)	no hits found	uni_CC_070	480 [326-515/480-291]	same
			uni_CC_071	968 [248-524/655-380]	same
			uni_CC_072	701 [107-231/576-451 ; 334-524/441-252]	same
			uni_CC_073	457 [107-231/304-428]	same
			uni_CC_074	579 [107-231/244-368]	same
			uni_CC_075	897 [163-225/8-70 ; 248-414/76-243]	same
			uni_CC_076	502 [144-231/390-477]	same
			uni_CC_077	518 [278-386/275-165]	same
uni_CC_030	441 (10-221)	no hits found	No hits found		
uni_CC_031	469 (60-112)	no hits found	No hits found		
uni_CC_032	886 (628-764)	unknown protein (14334448)	No hits found		
uni_CC_033	341 (211-300)	putative proline-rich protein APG isolog (10638955)	No hits found		
uni_CC_034	1,086 (162-218)	no hits found	No hits found		
uni_CC_035	1,075 (482-539)	unknown protein (116830383)	No hits found		
uni_CC_036	1,034 (10-316)	no hits found	No hits found		
uni_CC_037	1,056 (876-926)	protein binding / transporter, Protein transport protein Sec24-like , putative Sec24-like COPII protein (30680129)	uni_CC_078	1,152 [33-715/57-741]	Same
			uni_CC_079	919 [529-719/534-725]	same
uni_CC_038	1,247 (105-174)	26S proteasome regulatory complex component, contains PCI domain (90265119)	uni_CC_080	508 [52-127/20-95 ; 266-678/96-508]	26S proteasome regulatory particle non-ATPase subunit7 (17297977)
uni_CC_039	827 (446-497)	hexokinase 6 (45387415)	No hits found		
uni_CC_040	830 (347-471)	putative CEO protein (alternative splicing products) (29126336)	uni_CC_081	1,290 [561-771/1072-1289]	unknown protein (158110143)
			uni_CA_200	833 [561-830/298-567]	CEO protein (29126336)
			uni_CA_201	1,725 [12-346/1017-1351 e 468-686/1507-1725]	No hits found
			uni_CA_202	2,737 [561-830/2082-2352]	CEO protein (11044957)
uni_CC_041	589 (257-332)	CID11; RNA binding (79319100)	No hits found		

Table S2. Continuation

uni_CC_042	655 (311-632)	echinoderm microtubule associated protein like 5 (68355594)	No hits found		
uni_CC_043	367 (47-107)	ribosomal protein L7 (445613)	uni_CC_082	1,028 [16-69/12-65 ; 75-227/43-198]	60S ribosomal protein L7 (77999275)
uni_CC_044	944 (820-898)	DNA binding, phospatase 2A inhibitor (15221298)	No hits found		
uni_CC_045	995 (175-250)	Capip2 (47558819)	uni_CC_083	854 [9-204/655-850]	No hits found
			uni_CC_048	1,034 [11-251/459-222]	13-lipoxygenase (1495804)
			uni_CC_084	523 [10-172/299-459]	No hits found
			uni_CC_085	297 [10-172/97-259]	No hits found
			uni_CC_018	486 [158-217/115-174]	No hits found
uni_CC_046	1,012 (42-78)	no hits found	uni_CC_086	487 [830-1012/136-328]	No hits found
			uni_CC_087	298 [892-1012/14-136]	No hits found
			uni_CC_088	879 088 [836-1007/586-761]	No hits found
			uni_CC_089	469 [610-829/285-63]	No hits found
			uni_CC_090	527 [727-829/103-205]	No hits found
			uni_CC_091	434 [830-995/105-275]	No hits found
			uni_CC_092	480 [830-987/210-371]	No hits found
uni_CC_047	1,413 (567-681)	cell wall-plasma membrane linker protein (11994733)	No hits found		
			uni_CA_203	1,537 [9-437/160-588 e 414-645/598-829 e 755-1280/906-1429]	same
uni_CC_048	1,034 (165-294)	13-lipoxygenase (1495804)	uni_CC_093	958 [614-1034/252-673]	Lipoxygenase (32454714)
			uni_CC_094	1,140 [614-988/402-777]	Lipoxygenase (14589309)
			uni_CC_083	854 [269-614/850-501]	No hits found
			uni_CC_095	893 [298-605/54-363]	No hits found
			uni_CC_096	685 [347-611/525-259]	No hits found
			uni_CC_045	995 [222-469/251-11]	Capip2 (47558819)
			uni_CC_097	221 [415-614/221-21]	No hits found
			uni_CC_084	523 [319-596/441-162]	No hits found
			uni_CC_085	297 [319-538/240-19]	No hits found
			uni_CC_098	342 [9-149/165-25]	Retrosat (28558781)
			uni_CC_099	751 [13-226/229-16]	Retrosat (28558781)
			uni_CC_100	475 [411-520/118-9]	No hits found
			uni_CC_101	609 [410-538/263-135]	membrane-associated salt-inducible (3249103)
			uni_CC_102	562 [397-538/545-404]	No hits found
			uni_CC_018	486 [256-314/174-116]	No hits found
uni_CC_049	716 (117-210)	Nascent polypeptide-associated alpha subunit like protein complex NAC; UBA-like (87241023)	uni_CC_103	804 [8-716/12-722]	Same
			uni_CA_204	852 [6-118/1-113 e 158-716/147-710]	nascent polypeptide associated complex alpha chain (34907258)
			uni_CA_205	903 [6-118/28-140 e 159-716/178-740]	nascent polypeptide associated complex alpha chain (34907258)
uni_CC_050	500 (394-477)	No hits found	uni_CC_104	575 [352-500/9-157]	No hits found
uni_CC_051	504 (205-287)	Proteína unknown (18417290)	uni_CC_105	593 [85-375/146-436]	Same
			uni_CC_106	475 [389-174/391-176]	No hits found
uni_CC_052	575 (51-134)	No hits found	uni_CC_050	500 [155-9/498-352]	No hits found

Table S2. Continuation

uni_CC_053	910 (79-218)	unknown protein (4734014)	No hits found
uni_CC_054	1,049 (460-512)	ARRZ-1A; RNA binding / nucleotide binding, RNA-binding protein (15231557)	No hits found
uni_CC_055	967 (536-626)	No hits found	No hits found
uni_CC_056	368 (154-214)	Yippee-like protein (77555722)	No hits found
uni_CC_057	1,966 (957-1009)	hydrolase-like protein (53828585)	No hits found
uni_CC_058	1,129 (161-284)	putative SCO1 protein (18398306)	No hits found
uni_CC_059	1,169 (440-524)	No hits found	No hits found

Table S3. List of unigenes containing or not TE-cassette insertions in *C. racemosa*.

unigenes containing cassette-TE insertions ("query")			unigenes related those containing TE by BLASTn comparisons ("subject")		
Identification	unigene length in bp (TE insertion)	First protein hit in BLASTx searches (Genbank accession #)	Identification	unigene lenght (bp) [regions of similarity between query/subject]	First protein hit in BLASTx searches (Genbank accession #)
unigenes identified by RepeatMasker					
uni_CR_001	641 (230-322)	No hits found	No hits found		
uni_CR_002	833 (670-758)	copper chaperone (CCH)-related (62321736)	No hits found		
uni_CR_003	791 (345-424)	unknown protein (26451901)	No hits found		
uni_CR_004	486 (383-458)	No hits found	No hits found		
uni_CR_005	369 (320-435)	granule-bound starch synthase (3688123)	No hits found		
unigenes identified by tBLASTx					
uni_CR_006	766 (156-344)	putative cinnamyl alcohol dehydrogenase (2981475)	uni_CR_013	1304 [71-517/44-490]	same
uni_CR_007	748 (244-441)	histone H3 (1568559)	No hits found		
uni_CR_008	713 (82-342)	histone H3 (62701879)	No hits found		
uni_CR_009	778 (319-486)	putative galactosyltransferase (41469412)	No hits found		
uni_CR_010	626 (38-316)	histone protein Hist2h3c1(34858260)	uni_CR_008 uni_CR_015	713 [86-351/112-377] 567 [113-355/12-254]	histone H3 (62701879) histone H3.2 precursor (488573)
uni_CR_011	812 (55-201)	SC35-like splicing factor (9843661)	No hits found		
uni_CR_012	779 (497-676)	putative pathogenesis-related protein (7269429)	No hits found		
uni_CR_013	1304 (129-317)	putative cinnamyl alcohol dehydrogenase (2981475)	uni_CR_006	766 [44-692/81-729]	same
uni_CR_014	1091 (779-582)	heat shock protein 70 (20557)	No hits found		
uni_CR_015	567 (3-218)	histone H3.2 precursor (488573)	uni_CR_008 uni_CR_010	713 [6-337/133-464] 626 [12-254/113-355]	histone H3 (62701879) histone protein (34858260)

Transcriptional activity and chromosomal distribution of transposable elements in *Coffea* genomes

Capítulo 3

Transcriptional activity and chromosomal distribution of transposable elements in *Coffea* genomes

Fabrício R. Lopes[1], Alan C. Andrade[2], Pierre Marraccini[2, 3], Carlos R. M. da Silva[4], André L. L. Vanzela[4], Daudi Jjingo[5], I. King Jordan[5], João B. Teixeira[2], Carlos A. Colombo[6], Marcelo F. Carazzolle[7], Gonçalo A. G. Pereira[7], Luiz Filipe P. Pereira[8,9], Claudia M. A. Carareto*[1]

[1]UNESP – São Paulo State University, Department of Biology, 15054-000, São José do Rio Preto, São Paulo, Brazil,

[2]EMBRAPA – Brazilian Enterprise for Agropecuary Research, Genetic Resources and Biotechnology, 70770-900, Brasília, Distrito Federal, Brazil,

[3]CIRAD – Agricultural Research for Development, UMR PIA, F-34398, Montpellier Cedex 5, France,

[4]UEL – State University of Londrina, Department of General Biology, 86051-990, Londrina, Paraná, Brazil,

[5]School of Biology, Georgia Institute of Technology, Atlanta, GA 30332, USA,

[6]IAC – Agronomic Institute of Campinas, Centre for Genetics, Molecular Biology and Phytochemicals, 13012-970, Campinas, São Paulo, Brazil,

[7]UNICAMP – State University of Campinas, Department of Genetics and Evolution, Institute of Biology, 13083-970, Campinas, São Paulo, Brazil,

[8]EMBRAPA – Brazilian Enterprise for Agropecuary Research, Coffee, 70770-901, Brasília, Distrito Federal, Brazil, and

[9]IAPAR – Agronomic Institute of the State of Paraná, Laboratory of Biotechnology, 86001-970, Londrina, Paraná, Brazil.

*Corresponding author (fax 55 17 3221 2390; e-mail carareto@ibilce.unesp.br).

Keywords: *Coffea* genomes, transposable elements, macroarray profiling, FISH.

Abstract

This study reports on the identification of expressed transposable elements in 191,971 *ESTs* derived from cDNA libraries of several tissues, development stages and stress conditions in three *Coffea* species, namely *Coffea arabica*, *C. canephora* and *C. racemosa*. Using a combination of BLAST algorithm and TE libraries we identified 327 expressed TEs (0.01% of the total *ESTs*). Macroarray profiling for 64 of them using probes of cell culture and water deficit stress plants obtained from *C. arabica* and from one of its parental genome, *C. canephora*, showed a variable expression pattern. Embriogenic callus seems not to induce activation of TEs. Interestingly, the analysis of chromosomal distribution of three selected TEs by presenting low expression levels showed a higher frequency of hybridization signals in terminal chromosomal segments. This finding strongly suggests that the expression pattern of these TEs be related to their heterochromatic distribution.

Introduction

Transposable elements are genetic entities with an intrinsic capacity of mobilization. As a result of this characteristic, they are responsible for considerable levels of genomic instability mediating chromosome rearrangements (Zhang and Peterson, 1999), altering both gene expression and function (Jordan *et al.*, 2003), and creating novel genes and exons (Kidwell and Lisch, 1997), regulatory sequences (Jordan *et al.*, 2003) or transcription regulatory signals (Thornburg *et al.*, 2006). Such mobilization can also result in the host genome contraction and expansion (Fedoroff, 2000; Bennetzen, 2002). In a recent publication, Wicker *et al.* (2007) proposed a unified classification system for eukaryotic transposable elements reconciling an old criteria of classification (Finnegan, 1989), as well as a categorization that allows including the novel TEs groups. TEs can be grouped into two classes according to their transposition mode - the Class I elements, which use the Reverse Transcriptase (RTase) to transpose via RNA intermediate for a new genome insertion site, and the Class II elements, which are transposed directly via DNA molecule using a Transposase (Tpase) enzyme. Class I is divided into subclasses for distinguishing elements that copy themselves or leave the donor site to reintegrate elsewhere. The elements of this Class are divided into five orders (LTR, DIRS, PLE, LINEs, SINEs), following the subdivision of each into superfamilies (LTR: *Copia*,

Gypsy, *Bel-Pao*, *Retrovirus*, *ERV*; DIRS: *DIRS*, *Ngaro*, *VIPER*; PLE: *Penelope*; LINEs: *R2*, *RTE*, *Jockey*, *L1* and *I*; SINEs: *tRNA*, *7SL*, *5S*). Class II, which harbors the DNA transposons, is also divided into two subclasses: in subclass I the elements are grouped into superfamilies that harbor TIR (*Tc1-Mariner*, *hAT*, *Mutator*, *Merlin*, *Transib*, *P*, *PiggyBac*, *PIF-Harbinger* and *Cacta*) or not (*Crypton*), while in subclass II the elements are grouped into the Helitron and Maverick superfamilies.

There are increasing evidences that the plant genomes are massively invaded by repetitive DNA, mainly LTR retrotransposons. Many of them are located near genes, that could impact the expression of host neighboring genes mediated by a variety of mechanisms, such as RNAi (Zaratiegui *et al.*, 2007), methilation (Kinoshita *et al.*, 2007) and *readout* transcription (Kashkush *et al.*, 2003). But transcriptional activity of TEs seems to be under the "arrest" of the host genome (Biemont, 2009), given, for example, the involvement of transcriptional gene silencing mechanisms that could prevent the access of the host transcriptional machinery (Okamoto e Hirochika, 2001). However, reports also shown that TEs can be activated under certain stressful conditions, such as pathogen infection, physical injuries and abiotic stress (Grandbastien, 1998) or cell culture (Pouteau *et al.*, 1991; Hirochika, 1993; De Araujo, Rossi *et al.*, 2005; Vicient, 2010).

Three species of the *Coffea* genus (*Coffea arabica*, *C. canephora* and *C. racemosa*) were chosen to analyze the transcriptional activity of TEs. Apart from about one hundred of the species that composes the *Coffea* genus, which are cultivated throughout tropical regions of the world, commercial production relies on only two species, *Coffea arabica* and *C. canephora*, which represent ~70 and 30% of the global coffee production respectively. *C. arabica* is the unique polyploidy species of the genus (2n= 4x= 44 chromosomes), derived from a recent (1 million years ago) and natural hybridization between *C. canephora* and *C. eugenioides* (Lashermes *et al.*, 1999). *C. canephora* is a diploid (2n= 2x= 22 chromosomes) and auto-incompatible species that grows in humid and lowland habitats. It is more resistant to diseases and hydric stress and has a higher productivity and caffeine content. However, the quality of the beverage is regarded inferior to that produced by *C. arabica*. Finally, *C. racemosa*, another diploid species, is characterized by also presenting high resistance to

disease (e.g. to leaf-minner) and drought and is of particular importance for breeding programs by the introgression of traits to *C. arabica* (Guerreiro *et al.*, 1991).

A detailed analysis of the abundance and activity of transcriptionally active TEs in coffee genomes and the interplay with its chromosomal distribution has not to our knowledge been previously studied. To target this open question, we previously searched for TE-like coding sequences (CDSs) in dozens of libraries of expressed sequence tags (*ESTs*) derived from the three *Coffea* genomes (Lopes *et al.*, 2008) which allowed us to identify a wide spectrum of active elements. A set of TEs from *C. arabica* was used for evaluating their expression levels by macroarray profiling in cell cultures and plants under water stress obtained both from *C. arabica* and from one of its parental species, *C. canephora*. After an analysis of the transcriptional activity of these elements, the chromosomal distribution by FISH was investigated for transposons and retrotransposon sequences, which, interestingly, allowed us to observe that most of the genomic insertion sites are localized in terminal chromosome blocks. This finding strongly suggests that their low expression levels are related to their heterochromatic distribution pattern.

Material and Methods

Identification of expressed TEs: The identification of transcriptionally active TEs was carried out on 131,150 *ESTs* from *C. arabica,* 10,566 from *C. racemosa* and 50,255 from *C. canephora* (12,607 *ESTs* obtained by BCGP plus 46,914 *ESTs* produced by Lin *et al.*, 2005). *EST* clusters (or unigenes) similar to expressed TEs were not considered because they can represent the mRNA assembly of distinct insertions. The transcriptionally active TE searches were performed using keywords, such as: “transposon”, “transposase”, “polyprotein”, “retrotransposon”, “retroposon”, “MITEs”, “LINEs”, “SINEs” and family names (e.g. “hAT”, “MuDR”, “En/Spm”). Many BLAST *hits* were obtained, however, for eliminating spurious and low reliability results, a stringent *cut-off* ($E=1e^{-30}$) was applied between the comparisons of the coffee *ESTs* similar to TEs and the protein sequences stored at NCBI database, particularly the NR (non-redundant) database using BLASTx (Altschul *et al.*, 1997). These transcripts were classified into families according to the best alignment by BLASTx or tBLASTx (Altschul *et al.*, 1997) against a completely characterized library of 96 reference TEs (see Supplementary Material: Table S1), as well as against 840 consensus TEs of

phylogenetic species close to *Coffea* (e.g. *Vitis vinifera*, *Populus trichocarpa*, *Solanum lycopersicum*, *Solanum tuberosum*, etc) stored in the dicotyledonous plant library of the Repbase (www.girinst.com). The frequencies of Retrotransposons and Transposons, and expressed TEs of the *Copia/Ty1* and *Gypsy/Ty3* superfamilies were compared using χ^2.

cDNA sequencing and in silico analyses: The 64 clones homologous to TEs, cloned into pSPORT1 vector, were sequenced using the *BigDye® Terminator v3.1 Cycle Sequencing kit* (Applied Biosystems, Foster City, CA, USA) and run on an *3730xl DNA Sequencer* (Applied Biosystems). The sequences were clustered using CodonCode Aligner v.3.5.6 (www.codoncode.com) and bases had a Phred quality ≥ 20. The identification of these active TEs followed as: "Ca_" (for *Coffea arabica*), "TE-" (for transposable elements) plus "three numerical digits". Descriptions of the coffee *EST* libraries considered in our analyses can be found in Lopes *et al.* (2008).

Plant Material: for the probe synthesis, total mRNA was extracted from the following samples of *C. arabica:* a) Drought stress: Cultivars tolerant (Iapar59) and sensitive (Rubi) to drought maintained in field conditions with (ψpd= -0.5 MPa) and without (ψpd= -1.7 MPa) water; b) Cell culture: embriogenic callus from *C. arabica* cv. Catuaí Vermelho maintained in a multiplication medium for ~4 months; c) Inhibition treatment: the same embriogenic callus was treated with the protein biosynthesis inhibitor Cycloheximide (CHX: 10 and 30 mg/mL in alcohol) added to cell culture for a final concentration of 100 and 300 µg/mL during 4h, respectively. *C. canephora:* a) Drought stress: clones var. Conillon tolerant (14) to drought were selected by the INCAPER (Ferrão *et al.*, 2000) and grown in a greenhouse with (unstressed condition) or without (stress, ψpd leaves = -3.0 MPa) water.

RNA isolation, DNAse treatment and Reverse transcription: The 10 total RNA samples were extracted from cells using Concert™ reagent (Invitrogen, Carlsbad, CA, USA) according to the manufacturer´s protocol. A total RNA sample (10 µg) was incubated for 30 min at 37 ºC with 3 U of *RQ1 RNAse-Free DNAse* (Promega, Madison, WI, USA) in a final volume of ~10 µl. Each total RNA sample was mixed with 1.5 µl Oligo$(dT)_{12\text{-}18}$ (Invitrogen) and heated at 75 ºC for 10 min, then cooled for 5 min on ice. The reaction mixture for the reverse transcription

contained 5 µl of 5 × first-strand buffer, 2.5 µl of 0.1 mol/l DTT, 40 U of RNAseOUT, 50 µCi [α-^{33}P]-dCTP and 2.5 µl of a 10 mM mixture of unlabeled dNTPs (dATP, dTTP and dGTP) was heated at 42 ºC for 5 min. The reverse transcription was performed at 42 ºC for 20 min with 300 U of *SuperScript III First-Strand Synthesis System for RT-PCR* (Invitrogen). Then, 1.25 µl of unlabeled dCTP (10 mM) was added and maintained for 1 h, terminated by heating at 94 ºC for 5 min and cooled for 5 min on ice. The total volume (30.25 µl) was used in the hybridization experiments.

Amplification of target DNA: Each target cDNA (100 ng) was amplified by PCR in a volume of 25 µl with 1.25 U of *Platinum® Taq DNA Polymerase* (Invitrogen) in 10 × polymerase buffer, 2 mM $MgCl_2$, a 200 µM concentration of each dNTP and 10 µM of each universal primer, M13F (5´-GTAAAACGACGGCCAG-3´) and M13R (5´-CAGGAAACAGCTA TGAC-3´). The solutions were heated to 94 °C for 2 min, and followed by 35 cycles of denaturation (94 °C for 30 sec), annealing (50 °C for 30 sec), and extension (72 °C for 4 min), followed by a final extension at 72 °C for 7 min. The target DNAs were used for making the array membranes.

Macroarray experiments: The PCR products of the 64 clones were denatured in DMSO (50% v/v) for 30 min at 37 ºC arranged in 384-well microtiter plate and then spotted twice in the same position onto *Performa II nylon filters* (Genetix Limited, Hampshire, UK) using the high-throughput robot system *Q-BOT* (Genetix Limited). To increase the signal homogeneity among spots and filters, the set of 64 cDNAs was spotted in duplicate (2 × 2 array) into two identical arrays arranged in the same nylon filters (222 × 222 mm). Additionally, 16 spots containing cDNA of the reference gene ubiquitin were applied to delimit the two arrays. After sample deposition, the filters were dampened with a denaturant solution (NaCl 0.13 M and 0.5 M NaOH) for 10 min, a neutralization solution (NaCl 1.5 M and Tris 1 M) for 5 min, then fixed by UV light exposition (1,200 µj/cm^2) for 12 sec and stored at -80 ºC. The filters were pre-hybridized for 2 h at 65 ºC in *Modified Church and Gilbert Buffer* (0.5 M Phosphate Buffer pH 7.2, 7 % SDS, 10 mM EDTA) and hybridized overnight with cDNA sample probes. Membranes were washed for 15 min three times with 0.1 % SDS/1 × SSC and three times with 0.1 % SDS/0.1 × SSC at 65 ºC. After washing, the filters were exposed on imaging plates *BAS-MS 2340* (Fujifilm, Tokyo, Japan) for 72 h onto a *BAS 2340 cassette* (Fujifilm)

and scanned using a fluorescent image analyzer *FLA3000* (Fujifilm). The radioactive intensity of each spot was quantified by *Array Guage* software (Fujifilm), corrected by the level of the local background, normalized to the average intensities of the reference gene ubiquitin (except for callus treated with CHX, in which the reference gene expression was completely suppressed) and for the differences of probe labeling using the average signals of all genes studied. The homogeneity of the spot replicates were evaluated and represented by average values using *limma* of the Bioconductor package (Gentleman *et al.*, 2004). The final expression values were analyzed by *heatmap* using the statistical environment R (http://www.r-project.org). The expression profiles were grouped and compared aiming at identifying significant expression differences related to treatment (callus vs callus treated with CHX), stress (I59_I vs I59_NI and Rubi_I vs Rubi_NI for *C. arabica*; 14_I vs 14_NI for *C. canephora*) and status to stress (tolerant vs sensitive only for *C. arabica*). To compute the average differences between each group of comparison, the average of one group was subtracted from that of the other group. To test if these differences were significant or not, a two-tailed paired student's T-test was performed between each pair of groups being tested on an excel spreadsheet. This yielded a *P*-value that was used to determine significant differences ($p<0.05$).

Fluorescent in situ hybridization: FISH was performed as described by Heslop-Harrison and Schwarzacher, 1991) with modifications. Three expressed TE cDNA clones (Ca_TE-009, Ca_TE-050 and Ca_TE-079 similar to *MuDR*, *dea1* and *Tip100*, respectively) were used for synthesizing the probes with biotin-14-dATP by nick translation. The reaction mixture (total volume 33 µl) contained 15 µl of 100 % formamide, 6 µl of polyethylene glycol, 3 µl of 20 × SSC, 1 µl of calf thymus DNA (100 ng), 4 µl of water and 4 µl of each probe (200 ng). The material was denatured at 70 ºC for 10 min and hybridized at 37 ºC overnight in a humidified chamber. The washes were carried out in 6 × SSC and 4 × SSC/0.2 % Tween 20 at room temperature. The probes were detected with avidin-FITC, followed by post-detection washes in 4 × SSC/0.2 % Tween 20 at room temperature. Slides were mounted with 25 µl of antifade, composed of glycerol 90 %, 1,4-diaza-bicyclo(2,2,2)-octane(2.3 %), 20 mM TrisHCl pH 8.0 (2 %), water and 1 µl of 2 µg/ml 4,6-diamidino-2-phenylindole (DAPI). The images were acquired by *Leica DM4500 B Microscope* (Leica Microsystems, Wetzlar, Germany), equipped with a *DFC 300FX Digital Color Camera* (Leica Microsystems), and overlapped

with red color for DAPI using the *Leica IM50 4.0* image management software (Leica Microsystems).

Results

Frequency and Classification of expressed TEs in the transcriptome from Coffea spp

A set of 191,971 *ESTs* derived from 31 cDNA libraries from *C. arabica*, eight from *C. canephora* and two from *C. racemosa* were used for analyzing the classes, families and number of expressed TEs. The similarity sequence searches based on the BLASTx results using keywords allowed the identification of 327 clones homologous to TEs in the three *Coffea* species (Table 1; Supplementary Material: Tables S2-S4). For *C. arabica* and *C. racemosa*, the proportion of homolog sequences to Transposons (51% and 50%, respectively) and Retrotransposons (49% and 50%, respectively) were similar ($P<0.05$ for both species). Likewise, the amount of expressed TEs of *Ty1/Copia* was also similar to *Ty3/Gypsy* for both species ($P<0.05$). Finally, in *C. canephora*, the proportion of Transposons (13%) and Retrotransposons (87%) was very different, as well as the difference between *Ty3/Gypsy* and *Ty1/Copia* (both $P>0.05$). These differences reflect the higher abundance/expression of *Ty3/Gypsy* elements in the libraries SE2 (Seed – 18 weeks after pollination) and PP1 (Pericarp – all development stages) (Figure 1).

The *ESTs* similar to expressed TEs were classified into families regarding the best alignment against a completely characterized TE's library (reference TEs) using BLAST algorithm, which classified the 101 clones from *C. arabica* into 25 families, 220 clones from *C. canephora* into 19 families and six clones from *C. racemosa* into four families (Figure 2). For example, the most expressed families in *C. arabica* were *Retrosat1-2* (12 ESTs: 24% of the total of expressed TEs identified), *Melmoth* (8 ESTs: 12%), *Cin4* (5 ESTs: 12%), for retrotransposons; and, *MuDR* (12 ESTs: 23%), *Activator* (11 ESTs: 21%) and *Soymar* (11 ESTs: 21%), for transposons. On the other hand, in *C. canephora* the most expressed elements were *dea1* (94 ESTs: 48%), *Retrosat* (56 ESTs: 31%) and *Del1-46* (18 ESTs: 9%) for retrotransposons; and, *MuDR* (7 ESTs: 24%), *AtMu* (6 ESTs: 21%), *Activator* (6 ESTs: 21%), *Jittery* (5 ESTs: 17%), *Tag2* (4 ESTs: 14%), *Soymar* (1 ESTs: 3%) for transposon.

Characterization of 28 sequenced cDNA clones similar to TEs from C. arabica

After the classification of the 101 clones identified in *C. arabica* into families, the redundant clones (identical ESTs and identified in the same library) were excluded, 64 clones remained, from which 28 had already been fully sequenced (Table 2). Again, this set of sequences was compared against two TE banks: the reference elements cited above and the collection of consensus TEs stored in the Repbase. The results evidenced that these cDNAs present higher similarity to elements from Repbase described in species closely related to genus *Coffea*, such as *Vitis vinifera* and *Solanum tuberosum*, for example. The occurrence of frame shifts and stop codons was distinct for some comparisons due the differential choice of frame in the sequence translation. Putatively complete transposase or polyprotein searches were performed for evaluating CDSs that spanned at least 60% of the reference or consensus TE and absence of frame shifts and stop codons, as applied by Araujo *et al.* (2005). Up to now, two clones, similar to *MuDR* (Ca_TE-008) and *Ty1* (Ca_TE-098) were identified. Perhaps, most importantly, the sequencing reactions are in process for obtaining the 64 clones.

Expression levels of 64 TEs

Macroarray profiling of 64 TEs (31 DNA transposon and 33 retrotransposon) analyzed in five samples from the allotetraploid *C. arabica* (callus, leaves cv. Iapar59 and Rubi, irrigated or not for both) showed that many TEs present low expression (Figure 3). Expression was detected in the five groups of analysis for three families of retrotransposons (*Cin4*, *Del1* and *Melmoth*) and *Retrosat2* only in the callus. On the other hand, expression of the transposons similar to *AtMu1* and *Tip100* was detected in the four groups of analysis related to hydric stress and *Tag2* only in the callus. Moreover, the three transcriptionally active retrotransposons in the five *C. arabica* groups were also in two samples from *C. canephora* (leaves clone 14, irrigated or not) suggesting that the main features for its activity were maintained in these genomes with an outstanding difference in ploidy. Importantly, Real Time PCR analyses should be performed to confirm the profile obtained.

Finally, eight transcripts (*MuDR*: Ca_TE-003, *Tag2*: Ca_TE-038, *Tip100*: Ca_TE-049 and 050, *Del1*: Ca_TE-076, *Retrosat2*: Ca_TE-061 and 062, *Tnt1*: Ca_TE-093) had their transcriptional activity reestablished in a cell culture from *C. arabica* treated with 300 µg/ml Cycloheximide (CHX). A portion of the growing literature shows that inhibition of the protein

biosynthesis machinery using CHX allows the accumulation of previously silenced transcripts by reversing a mechanism named nonsense-mediated mRNA decay (NMD). By this regulatory pathway, the host genome detects and degrades aberrant mRNAs containing premature terminate codons derived from mutations, errors during transcription or RNA processing (Brogna and Wen, 2009). Several transcripts that trigger the NDM mechanism has been described using CHX, as reported in plants (Lambein *et al.*, 2003; Wu *et al.*, 2007), human (Noensie and Dietz, 2001; Lamba *et al.*, 2003; Hillman *et al.*, 2004) and mice (Lei *et al.*, 2001; Vignier-Schlossarek *et al.*, 2009). In fact, the test with a higher concentration of CHX (300 µg/mL) was chosen because it showed a total reduction of expression of the reference gene ubiquitin, a transcript that is not target to NMD, compared to the test using a lower concentration of CHX (100 µg/mL) (data not shown). Finally, this relative expression increase suggests that these TEs could act as a putative NMD-inducing feature, but the full-length sequencing of the cDNAs could show the defective character of these elements and the confirmation of their differential expression pattern by RT-qPCR (callus and callus treated with alcohol *vs* callus treated with CHX) is essential for a conclusive indication.

Statistical analyses were carried out for evaluating significant expression differences between groups of comparison. As a result, the comparison of the groups of callus *vs* callus treated with CHX (average difference among the groups= 0.3370, *P*-value= 5.5E-02) did not show significant expression differences, at least on average, suggesting that the expression differences are restricted to specific transcripts. On the other hand, the comparison for the other groups showed significant expression differences, namely the cultivar tolerant submitted to drought stress (I59_I *vs* I59_NI, average difference= -0.0769, *P*-value= 1.1E-02) showed a higher expression, for the sensitive Rubi (Rubi_I *vs* Rubi_NI, average difference= 0.1917, *P*-value= 3.4E-06) irrigation condition showed a higher expression; status to stress (cultivar tolerant *vs* sensitive= I59_NI *vs* Rubi_NI, average difference= 0.1039, *P*-value= 7.8E-05) showed a higher expression in the cultivar tolerant from *C. arabica;* and the drought stress in *C. canephora* (14_I *vs* 14_NI, average difference= -0.1769, *P*-value= 4.2E-04) showed a non irrigated higher expression Together, the significant expression differences observed suggest that the drought stress influences the expression levels for TE activity.

Differential amplification of TEs and genome size variation

The intrinsic mobilization capacity of the TEs and the differential amplification of certain families allow them to contribute to the genome size variation, as previously shown in *Zea mays* (Meyers *et al.*, 2001) and genus *Oryza* (Zuccolo *et al.*, 2007) genomes. Three TE-like CDSs (two DNA transposons and one LTR retrotransposon) with low expression levels were selected for an evaluation of their chromosomal distribution by FISH. This approach allowed us to identify the distribution of these repeats in the allotetraploid *C. arabica* and in their progenitors (*C. canephora* and *C. eugenioides*) for determining if the genome size variation among the species is associated to TE proliferation. As a result, the hybridizations of the transposons, *MuDR* (Ca_TE-009, Figures 4 A, D and G) and *Tip100* (Ca_TE-050, Figures 4 C, F and I), as well as for the LTR retrotransposon *dea1* (Ca_TE-079, Figures 4 B, E and H) are presented for the three species. On the other hand, the chromosomal distribution pattern will be obtained for the retrotransposons, *Del1-46* (Ca_TE-077) and *Retrosat2* (Ca_TE-062).

Most of the hybridization signals in *C. arabica* are located in terminal chromosomal blocks for the three TEs (Figures 4 A, B and C), except for one signal in the pericentromeric region of the *dea1* element (Figure 4 B). For *C. canephora*, the pattern of TE distribution in terminal chromosomal segments is also observed, in addition to signals located in the pericentromeric region of *MuDR* (Figure 4D) and interstitial ones for both *dea1* (Figure 4 E) and *Tip100* (Figure 4 F). Taken together, the three analyzed TEs presented a higher frequency of hybridization in terminal chromosomal segments. This finding strongly suggests that low expression levels of the selected TEs are related to their heterochromatic distribution, a region where expression and transposition of TEs is mostly suppressed mediated by epigenetic regulation mechanisms (e.g. DNA methylation, histone modification and RNAi) (Slotkin and Martienssen, 2007). Additionally, the TEs tend to be maintained in the pericentromeric heterochromatin, a region where the selection against TEs is relaxed as well as being submitted to positive selection, due to their role in heterochromatin formation and cell division (Biemont, 2009).

Discussion

This study allows a first glimpse of TE dynamics in the allotetraploid and complex genome of *C. arabica*. The findings here presented show that the species can be consolidated as an interesting organism to be studied because most DNA transposons and retrotransposons seem to be submitted to a fine transcriptional control. Also, contrary to commonly reported in literature, cell culture seems not to induce the TE activation in *Coffea*.

Other plants present some differences in their functional and evolutionary scenario when compared to *C. arabica*. For example, gene silencing and activation, cytogenetic abnormalities (e.g. translocations, inversions and insertion/deletion), DNA sequence changes and methylation instability are frequent events in the process of tissue culture or also called tissue culture-induced variation (Kaeppler *et al.,* 2000). This phenomenon has been described in a variety of species, such as sugarcane (Irvine *et al.*, 1991), barley (Bregitzer *et al.*, 1998), oats (Johnson *et al.*, 1987), maize (Lee and Phillips, 1987; Kaeppler and Phillips, 1993), carrots (Loschiavo *et al.*, 1989), including clonally propagated coffee plants derived from embryogenic cell suspension (Etienne and Bertrand, 2001). Interestingly, the genome stress caused by DNA changes has been shown to activate TEs. For example, the retrotransposon *Tnt1* from tobacco (Courtial *et al.*, 2001), *Tos10, Tos17 and Tos19* from rice (Hirochika *et al.*, 1996), several DNA transposons and retrotransposons from sugarcane (De Araujo, Rossi *et al.*, 2005) and mainly LTR retrotransposons from rice (Vicient, 2010) had increased transcription rates in cell culture. On the other hand, our computational and expression analyses showed that callus in *C. arabica* does not induce activation of TEs. Therefore, these data suggest that both phenotype and DNA variation in coffee cannot be significantly caused by the activity of TEs.

Apart from the differences, similarities with other plant genomes are also evidenced. Another example, ploidy changes due to polyploidization are a recurrent phenomenon in flowering plants (~70% of the species) and an obvious effect is the duplication of gene loci. Extensive changes in polyploidy genomes have been reported, such as insertions/deletions, inversions and translocations, as well as an alteration in the gene expression pattern (e.g. He *et al.*, 2003). Although the mechanisms of these changes are poorly understood, an increasing transposition activity of TEs is a possibility because a quiescent transposable element in a

diploid genome can be activated in the new polyploidy genetic environment. Alternatively, the genetic redundancy in a polyploidy genome can mitigate/minimize the deleterious effects of the transposition (Voytas and Naylor, 1998). Therefore, TEs can proliferate in allotetraploids, including the insertions within gene-rich chromosomal regions. Another possibility is the relaxation of host silencing mechanisms (e.g. methylation) in allotetraploid that would allow increased transposition rates (Madlung *et al.*, 2005). The three *Coffea* species here studied presented a low TE-like mRNA abundance, in which only 0.17% expressed TEs which were identified (327 of 191,971 ESTs). This low abundance is also observed in other plant genomes. For example, Meyers *et al.* (2001) showed that 60% of the *Z. mays* genome is constituted by retroelements, but only 0.014% of them (56 out of 407,000 ESTs) were identified as expressed. But, a recent systematic search carried out by Vicient (2010) in the maize transcriptome showed that 1.5% (25,282 ESTs out of more than 2 million) were similar to 56 well characterized TE families. In *Saccharum officinarum*, out of 260,781 ESTs, 276 (0.1%) were considered as expressed (Rossi, Araujo *et al.*, 2001); and finally, in *Eucalyptus grandis*, out of 123,889 ESTs, 124 (0.1%) identified as transcriptionally active TEs (BACCI Jr. *et al.*, 2005). Our data together with that described above reinforce the fact that TEs are poorly represented in the transcriptome although the plant genomes are enriched for those repetitive sequences. This paradox reflects a strong host transcriptional control action in plants, as illustrated by the heterochromatic distribution of three TE sequences here analyzed.

Acknowledgments

We thank Mário A. P. Saraiva and Felipe Vinecky for making the array membranes and for manipulating the BCGP's cDNA libraries, respectively and Elias A. G. Carnelossi for technical support. This study was funded by grants provided by the Brazilian agencies FAPESP (2008/05894-1 to C.M.A.C. and fellowship 2007/55985-0 to F.R.L.) and CNPq (to C.M.A.C.).

References

ALTSCHUL, S. F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. **Nucleic Acids Research,** v. 25, n. 17, p. 3389-3402, Sep 1997.

BACCI Jr, M. et al. Identification and frequency of transposable elements in *Eucalyptus*. **Genetics and Molecular Biology**, v. 28, n. 3, p. 634-639, 2005.

BENNETZEN, J. L. Mechanisms and rates of genome expansion and contraction in flowering plants. **Genetica,** v. 115, n. 1, p. 29-36, May 2002.

BIEMONT, C. Are transposable elements simply silenced or are they under house arrest? **Trends in Genetics,** v. 25, n. 8, p. 333-334, Aug 2009.

BREGITZER, P.; HALBERT, S. E.; LEMAUX, P. G. Somaclonal variation in the progeny of transgenic barley. **Theoretical and Applied Genetics,** v. 96, n. 3-4, p. 421-425, 1998.

BROGNA, S.; WEN, J. K. Nonsense-mediated mRNA decay (NMD) mechanisms. **Nature Structural & Molecular Biology,** v. 16, n. 2, p. 107-113, 2009.

COURTIAL, B. et al. *Tnt1* transposition events are induced by in vitro transformation of *Arabidopsis thaliana*, and transposed copies integrate into genes. **Molecular Genetics and Genomics,** v. 265, n. 1, p. 32-42, 2001.

DE ARAUJO, P. G. et al. Transcriptionally active transposable elements in recent hybrid sugarcane. **Plant Journal,** v. 44, n. 5, p. 707-717, 2005.

ETIENNE, H.; BERTRAND, B. Trueness-to-type and agronomic characteristics of *Coffea arabica* trees micropropagated by the embryogenic cell suspension technique. **Tree Physiology,** v. 21, n. 14, p. 1031-1038, 2001.

FEDOROFF, N. Transposons and genome evolution in plants. **Proceedings of the National Academy of Sciences of the United States of America,** v. 97, n. 13, p. 7002-7007, Jun 2000.

FERRÃO, R. G. et al. EMCAPA 8141 - Robustão Capixaba, variedade clonal de café conilon tolerante à seca, desenvolvida para o estado do Espírito Santo. **Revista Ceres,** v. 273, p. 555-560, 2000.

FINNEGAN, D. J. Eukaryotic transposable elements and genome evolution. **Trends Genetics**, v. 5, p. 103-107, 1989.

GENTLEMAN, R. C. et al. Bioconductor: open software development for computational biology and bioinformatics. **Genome Biology,** v. 5, n. 10, 2004.

GRANDBASTIEN, M. A. Activation of plant retrotransposons under stress conditions. **Trends in Plant Science,** v. 3, n. 5, p. 181-187, May 1998.

HESLOP-HARRISON, J. S.; SCHWARZACHER, T. In situ hybridization with automated chromosome denaturation. **Technique,** v. 3, p. 109-116, 1991.

HILLMAN, R. T.; GREEN, R. E.; BRENNER, S. E. An unappreciated role for RNA surveillance. **Genome Biology,** v. 5, n. 2, 2004.

HIROCHIKA, H. Activation of tobacco retrotransposons during tissue-culture. **Embo Journal,** v. 12, n. 6, p. 2521-2528, 1993.

HIROCHIKA, H. et al. Retrotransposons of rice involved in mutations induced by tissue culture. **Proceedings of the National Academy of Sciences of the United States of America,** v. 93, n. 15, p. 7783-7788, Jul 1996.

IRVINE, J. E. et al. The frequency of marker changes in sugarcane plants regenerated from callus-culture. 2. evidence for vegetative and genetic transmission, epigenetic effects and chimeral disruption. **Plant Cell Tissue and Organ Culture,** v. 26, n. 2, p. 115-125, 1991.

JOHNSON, S. S.; PHILLIPS, R. L.; RINES, H. W. Possible role of heterochromatin in chromosome breakage induced by tissue-culture in oats (*Avena sativa* l). **Genome,** v. 29, n. 3, p. 439-446, 1987.

JORDAN, I. K. et al. Origin of a substantial fraction of human regulatory sequences from transposable elements. **Trends in Genetics,** v. 19, n. 2, p. 68-72, Feb 2003.

KAEPPLER, S. M.; KAEPPLER, H. F.; RHEE, Y. Epigenetic aspects of somaclonal variation in plants. **Plant Molecular Biology,** v. 43, n. 2-3, p. 179-188, 2000.

KAEPPLER, S. M.; PHILLIPS, R. L. Tissue culture-induced DNA methylation variation in maize. **Proceedings of the National Academy of Sciences of the United States of America,** v. 90, n. 19, p. 8773-8776, 1993.

KASHKUSH, K.; FELDMAN, M.; LEVY, A. A. Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat. **Nature Genetics,** v. 33, n. 1, p. 102-106, Jan 2003.

KIDWELL, M. G.; LISCH, D. Transposable elements as sources of variation in animals and plants. **Proceedings of the National Academy of Sciences of the United States of America,** v. 94, n. 15, p. 7704-7711, Jul 1997.

KINOSHITA, Y. et al. Control of FWA gene silencing in *Arabidopsis thaliana* by SINE-related direct repeats. **Plant Journal,** v. 49, n. 1, p. 38-45, Jan 2007.

LAMBA, J. K. et al. Nonsense mediated decay downregulates conserved alternatively spliced ABCC4 transcripts bearing nonsense codons. **Human Molecular Genetics,** v. 12, n. 2, p. 99-109, Jan 2003.

LAMBEIN, I. et al. Decay kinetics of autogenously regulated CGS1 mRNA that codes for cystathionine gamma-synthase in Arabidopsis thaliana. **Plant and Cell Physiology,** v. 44, n. 9, p. 893-900, Sep 2003.

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.

LEE, M.; PHILLIPS, R. L. Genomic rearrangements in maize induced by tissue-culture. **Genome,** v. 29, n. 1, p. 122-128, 1987.

LEI, X. D.; CHAPMAN, B.; HANKINSON, O. Loss of CYP1A1 messenger RNA expression due to nonsense-mediated decay. **Molecular Pharmacology,** v. 60, n. 2, p. 388-393, Aug 2001.

LIN, C. W. et al. Coffee and tomato share common gene repertoires as revealed by deep sequencing of seed and cherry transcripts. **Theoretical and Applied Genetics,** v. 112, n. 1, p. 114-130, Dec 2005.

LOPES, F. R. 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, Apr 2008.

LOSCHIAVO, F. et al. DNA methylation of embryogenic carrot cell-cultures and its variations as caused by mutation, differentiation, hormones and hypomethylating drugs. **Theoretical and Applied Genetics,** v. 77, n. 3, p. 325-331, 1989.

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

MEYERS, B. C.; TINGLEY, S. V.; MORGANTE, M. Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. **Genome Research,** v. 11, n. 10, p. 1660-1676, 2001.

NOENSIE, E. N.; DIETZ, H. C. A strategy for disease gene identification through nonsense-mediated mRNA decay inhibition. **Nature Biotechnology,** v. 19, n. 5, p. 434-439, May 2001.

GUERREIRO, F. O.; MEDINA FILHO, H. P.; CARVALHO, A.. Fontes de resistência ao bicho-mineiro. *Perileucoptera coffeella* em *Coffea* sp. Bragantia. 50**:** 45-55 p. 1991.

OKAMOTO, H.; HIROCHIKA, H. Silencing of transposable elements in plants. **Trends in Plant Science,** v. 6, n. 11, p. 527-534, Nov 2001.

POUTEAU, S. et al. Specific expression of the tobacco *Tnt1* retrotransposon in protoplasts. **Embo Journal,** v. 10, n. 7, p. 1911-1918, Jul 1991.

ROSSI, M.; ARAUJO, P. G.; VAN SLUYS, M. A. Survey of transposable elements in sugarcane expressed sequence tags (ESTs). **Genetics and Molecular Biology,** v. 24, n. 1-4, p. 147-154, 2001.

SLOTKIN, R. K.; MARTIENSSEN, R. Transposable elements and the epigenetic regulation of the genome. **Nature Reviews Genetics,** v. 8, n. 4, p. 272-285, Apr 2007.

THORNBURG, B. G.; GOTEA, V.; MAKALOWSKI, W. Transposable elements as a significant source of transcription regulating signals. **Gene,** v. 365, p. 104-110, Jan 2006.

VICIENT, C. M. Transcriptional activity of transposable elements in maize. **Bmc Genomics,** v. 11, 2010.

VIGNIER, N. et al. Nonsense-Mediated mRNA Decay and Ubiquitin-Proteasome System Regulate Cardiac Myosin-Binding Protein C Mutant Levels in Cardiomyopathic Mice. **Circulation Research,** v. 105, n. 3, p. 239-U89, Jul 2009.

VOYTAS, D. F.; NAYLOR, G. J. P. Rapid flux in plant genomes. **Nature Genetics,** v. 20, n. 1, p. 6-7, 1998.

WICKER, T. et al. A unified classification system for eukaryotic transposable elements. **Nature Reviews Genetics,** v. 8, n. 12, p. 973-982, Dec 2007.

WU, J. Q. et al. Nonsense-mediated mRNA decay (NMD) silences the accumulation of aberrant trypsin proteinase inhibitor mRNA in *Nicotiana attenuata*. **Plant Journal,** v. 51, n. 4, p. 693-706, Aug 2007.

ZARATIEGUI, M.; IRVINE, D. V.; MARTIENSSEN, R. A. Noncoding RNAs and gene silencing. **Cell,** v. 128, n. 4, p. 763-776, Feb 2007.

ZHANG, J.; PETERSON, T. Genome rearrangements by nonlinear transposons in maize. **Genetics,** v. 153, n. 3, p. 1403-1410, 1999.

ZUCCOLO, A. et al. Transposable element distribution, abundance and role in genome size variation in the genus *Oryza*. **Bmc Evolutionary Biology,** v. 7, Aug 2007.

Tables

Table 1. Numbers and proportions of transposable elements of distinct classes of 191,921 *ESTs* obtained from three *Coffea* species.

Classes	Superfamily	*EST* number					
		C. arabica		*C. canephora*		*C. racemosa*	
		N	%	N	%	N	%
I	*Ty1/Copia*	12	11.7	8	3.6	1	16.7
	Ty3/Gypsy	22	21.6	177	80.4	2	33.3
	LTR[1]	10	9.8	1	0.4	-	-
	Retroposon	6	5.9	5	2.3	-	-
II	Transposon	51	51.0	29	13.2	3	50.0
TOTAL		**101**	**100**	**220**	**100**	**6**	**100**

[1] Not classified.

Table 2. Similarity pattern of 28 cDNA clones from *Coffea arabica* with other plant TEs.

TE-derived cDNA			Comparison with Reference Elements (RE)				Comparison with Consensus Elements (CE)			
Name[1]	Library[2]	Size (bp)[3]	Description[4]	Stop/ frame shift[5]	Similarity with RE (aa) (%)[6]	Aligned region (nt)[7]	Description[8]	Stop/ frame shift[5]	Similarity with CE (aa) (%)[6]	Aligned region (nt)[7]
DNA Transposons					48.5 ± 3.8				53 ± 3.2	
Ca_TE-009	FB1	1,481	*MuDRA* AAA21566 *Zm* (823 aa)	-/-	44	239/310-710/69	*MuDR4 Vv* (815 aa)	-/-	50	239/411-811/12
Ca_TE-008	LV5	1,799	“	-/-	44	239/178-710/6	“	-/-	51	239/300-811/6
Ca_TE-046	CB1	740	*AtMu1* AAG52094 *At* (761 aa)	-/-	41	254/556-699/114	*MuDR9 Vv* (679 aa)	-/-	44	269/514-673/15
Ca_TE-043	SH2	1,431	“	-/-	42	198/266-694/3	*MuDR4 Vv* (815 aa)	-/-	49	198/412-811/3
Ca_TE-044	LV4	869	“	-/-	35	171/451-693/8	“	-/-	42	159/593-814/17
Ca_TE-019	FB1	2,130	*Jittery* AAF66982 *Zm* (709 aa)	-/-	55	1,187/267-580/13	*MuDR-21 Vv* (769 aa)	-/-	55	1187/298-606/13
Ca_TE-035	CB1	1,227	*Tag2* AAD24567 *At* (577 aa)	-/-	49	257/252-542/55	*hAT-6 Vv* (652 aa)	-/-	53	245/329-648/7
Ca_TE-042	FR2	1,075	“	-/-	52	257/300-541/19	“	-/-	58	260/395-642/25
Ca_TE-050	SH2	1,402	*Tip100* BAA36225 *Ip (808* aa*)*	+/-	75	379/380-560/68	*hAT-4 Pt* (564 aa)	-/-	75	791/135-315/68
LTR Retrotransposons					62.9 ± 3.4				72.1 ± 2.3	
Ca_TE-061	SH2	773	*Retrosat2* AAD27547 *Os* (1,521 aa)	+/-	47	214/1344-1521/49	*Gypsy3 Pt* (1,407 aa)	+/-	68	256/1230-1397/34
Ca_TE-062	RT5	818	“	-/-	51	482/1351-1466/3	*Cop18-I Mt* (1,312 aa)	-/-	80	5/1175-1290/9
Ca_TE-063	FR2	1,078	*Cin4* Y00086 *Zm* (RTase 259 aa)	-/-	55	114/2-160/484	*Shaline14 Mt* (1,387 aa)	+/-	56	18/392-693/167
Ca_TE-064	RT8	1,582	“	-/-	58	1,072/2-162/24	“	-/-	55	127/179-659/30
Ca_TE-074	PA1	1,389	*Del1* 1510387A *Lh* (1443 aa)	-/-	73	1/5-175/875	*Gypsy2 Vv* (1,515 aa)	-/-	71	331/332-752/637
Ca_TE-080	RT5	363	*Dea1* CAA73042 *Ac* (871 aa)	-/-	68	113/613-692/10	*Gypsy3 Pt* (1,407 aa)	-/-	82	113/1068-1147/10
Ca_TE-072	CL2	552	*Tst1* CAA36615 *St* (675 aa)	+/+	60	5/58-203/292	*Tvv1 Vv* (1,382 aa)	-/+	75	2/319-556/46
Ca_TE-082	LV5	1,160	*Tst1* CAA36616 *St* (390 aa)	-/-	75	324/100-355/62	*Copia23 Vv* (318 aa)[9]	-/-	83	192/16-317/62
Ca_TE-085	CA1	941	*Tst1* CAA36614 *St* (334 aa)	-/-	49	293/52-260/15	*Copia35 Vv* (523 aa)[9]	-/+	68	1/88-314/360
Ca_TE-071	BP1	1,445	*Retrofit* AAB82754 *Ol* (1,445 aa)	-/-	37	65-513/11	*Copia32-I Vv* (1,803 aa)	-/-	74	1194-1674/5
Ca_TE-089	CL2	564	“	+/-	60	315/1348-1429/3	*Copia47 Pt* (564 aa)	+/-	84	249/468-569/3
Ca_TE-087	SH2	797	“	-/-	70	197/1233-1430/6	*Copia23 Vv* (318 aa)[9]	-/+	85	144/103-317/9
Ca_TE-088	CB1	747	*Hopscotch* AAA57005 *Zm* (1,439 aa)	-/-	70	307/1276-1416/17	*Copia23 Vv* (318 aa)[9]	-/-	74	247/159-317/23
Ca_TE-093	FR1	718	*Tnt1* CAA32025 *Nt* (1328 aa)	-/-	74	485/1277-1326/83	*Cop18-I Mt* (1,312 aa)	-/-	78	485/1261-1310/83
Ca_TE-094	CL2	362	*Ta1-1* CAA37917 *At* (587 aa)	-/-	76	8/447-563/3	*Gypsy-3 Pt* (1,407 aa)	+/-	68	5/923-1377/12
Ca_TE-095	CB1	1,284	*Osr1* BAB03249 *Os* (1,268 aa)	+/-	51	22/46-382/263	*Copia22 Pt* (316 aa)[9]	+/+	57	394/79-311/220
Ca_TE-096	LV4	854	*Athila1* CAA57397 *At* (935 aa)	-/-	59	249/52-248/14	*Gypsy1 Pt* (1,581 aa)	-/-	68	81/12-254/11
Ca_TE-098	CA1	1,067	*Ty1* Q07163 *Sc* (pol 1,756 aa)	-/+	100	2/494-846/5	-	-	-	-

cDNA similar to TEs from *C. arabica*: [1]Arbitrary identification (details see Material and Methods); [2]Tissue, developmental stage or stress condition in which the clone was obtained: BP1 - Suspension cells treated with acibenzolar-S-methyl, CA1 - Non-embryogenic callus, CB1 - Suspension cells treated with acibenzolar-S-methyl and brassinosteroids, CL2 - Hypocotyls treated with acibenzolar-S-methyl, FB1 - Flower buds in stages 1 and 2—long, FR1 - Flower buds no 6, pinhead fruits no 1 and fruits (stages 1 and 2)—long, FR2 - Flower buds no 6, pinhead fruits no 1 and fruits (stages 1 and 2)—short, LV4 - Young leaves from orthotropic branch — long, LV5 - Young leaves from orthotropic branch—short, PA1 - Primary embryogenic callus, RT5 - Roots with acibenzolar-S-methyl, RT8 - Suspension cells stressed with aluminum, SH2 - Water deficit stresses plants (pool of tissues); [3]Size of the cDNA

clone after full-length sequencing cloned into pSPORT1. Comparison with Reference Elements (completely characterized plant TEs): [4]TE name, Genbank accession number, species (At – *Arabidopsis thaliana*, Zm – *Zea mays*, Ac – *Ananas comosus*, Ol – *Oryza longistaminata*, Ip – *Ipoema purpurea*, Lh – *Lilium henryi*, Nt – *Nicotiana tabacum*, St – *Solanum tuberosum*, Sc – *Saccharomyces cerevisae*) and protein length; [5]Presence (+) or absence (-) of premature stop codon and frame shift in the TE´s sequence; [6]Percent similarity between putative *C. arabica* TE-encoded protein and reference element in BLASTx alignment; [7]Reference element amino acid region similar to our cDNA clone (central numbers), as well as nucleotides of the cDNA clone from *C. arabica* not included in the alignment in the both 5´ and 3´ ends (external numbers) as applied by Araujo *et al.* (2005). Comparison with Consensus Elements (stored in Repbase): [8]TE's family, species (Gm – *Glycine max*, Mt – *Medicago truncatula*, Pt – *Populus trichocarpa*, Vv – *Vitis vinifera*) and protein length, [9]Incomplete polyprotein.

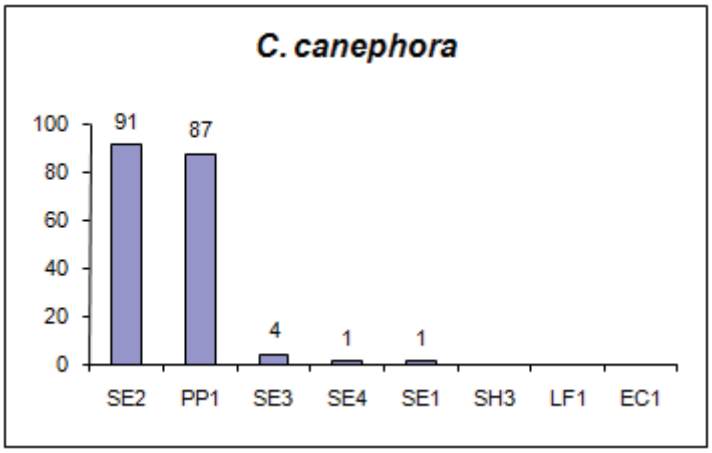


Figure 1. Frequency of *TEs* of the *Ty3/Gypsy* superfamily in the eight cDNA libraries from *C. canephora*. Libraries produced by Nestle Research Center, namely SE2: whole cherries – 18 weeks after pollination, PP1: Pericarp – all development stages, SE3: endosperm and perisperm of seeds – 30 weeks after pollination, SE4: endosperm and perisperm of seeds – 42 and 46 weeks after pollination, LF1: leaf – young; and those of the Brazilian Coffee Genome Project, such as SH3: leaves from water deficit stresses plants and EC1: embryogenic callus.

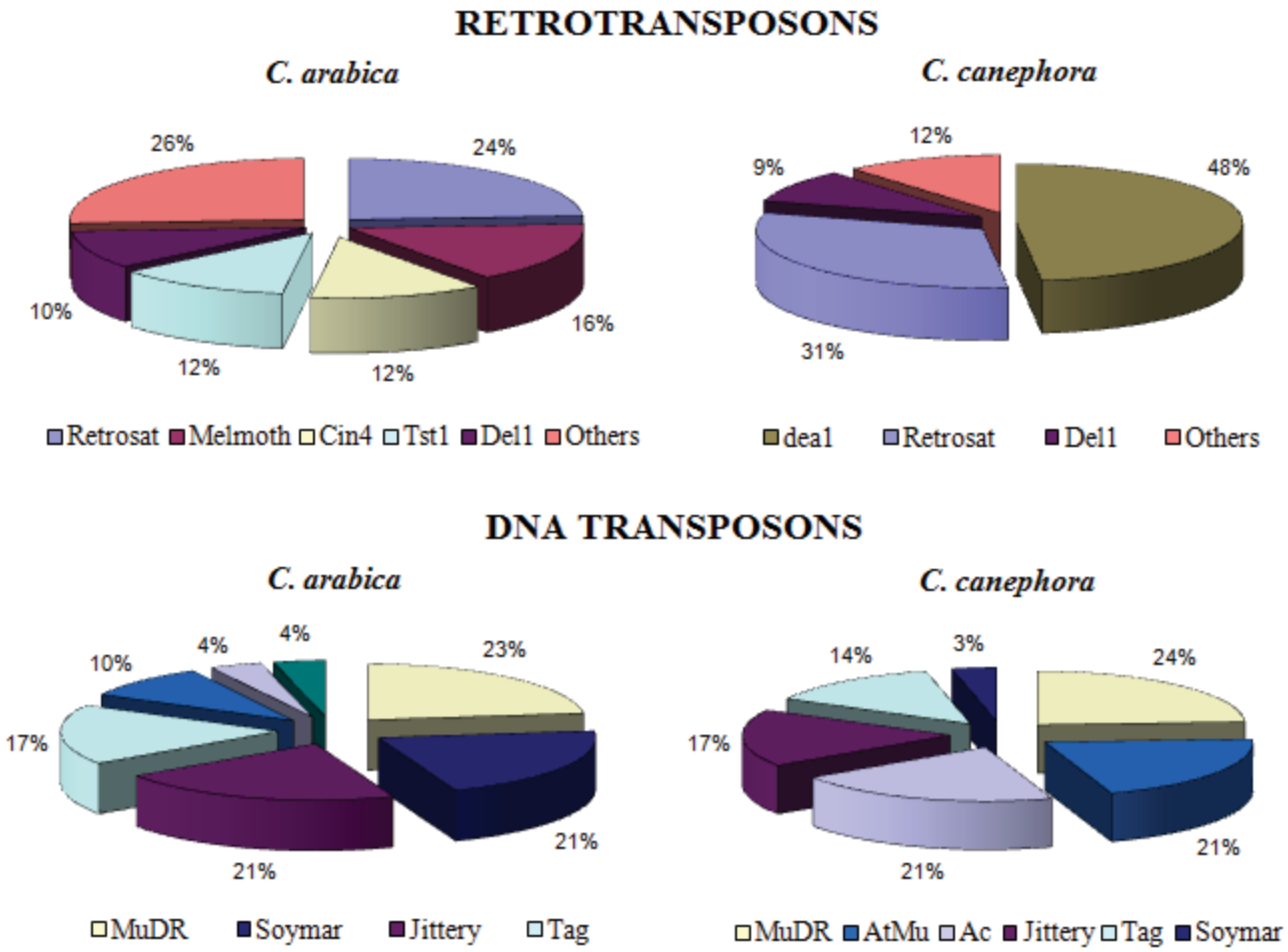


Figure 2. Comparative analysis and frequency of the distinct TE families based on *EST* searches identified in the transcriptome from *C. arabica* and *C. canephora*.

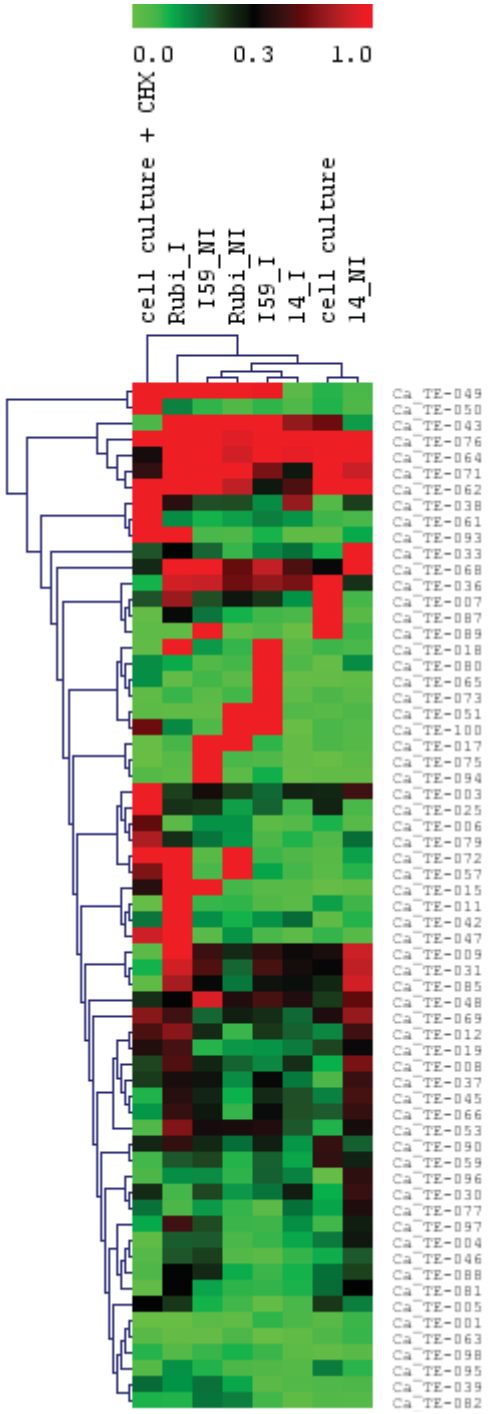
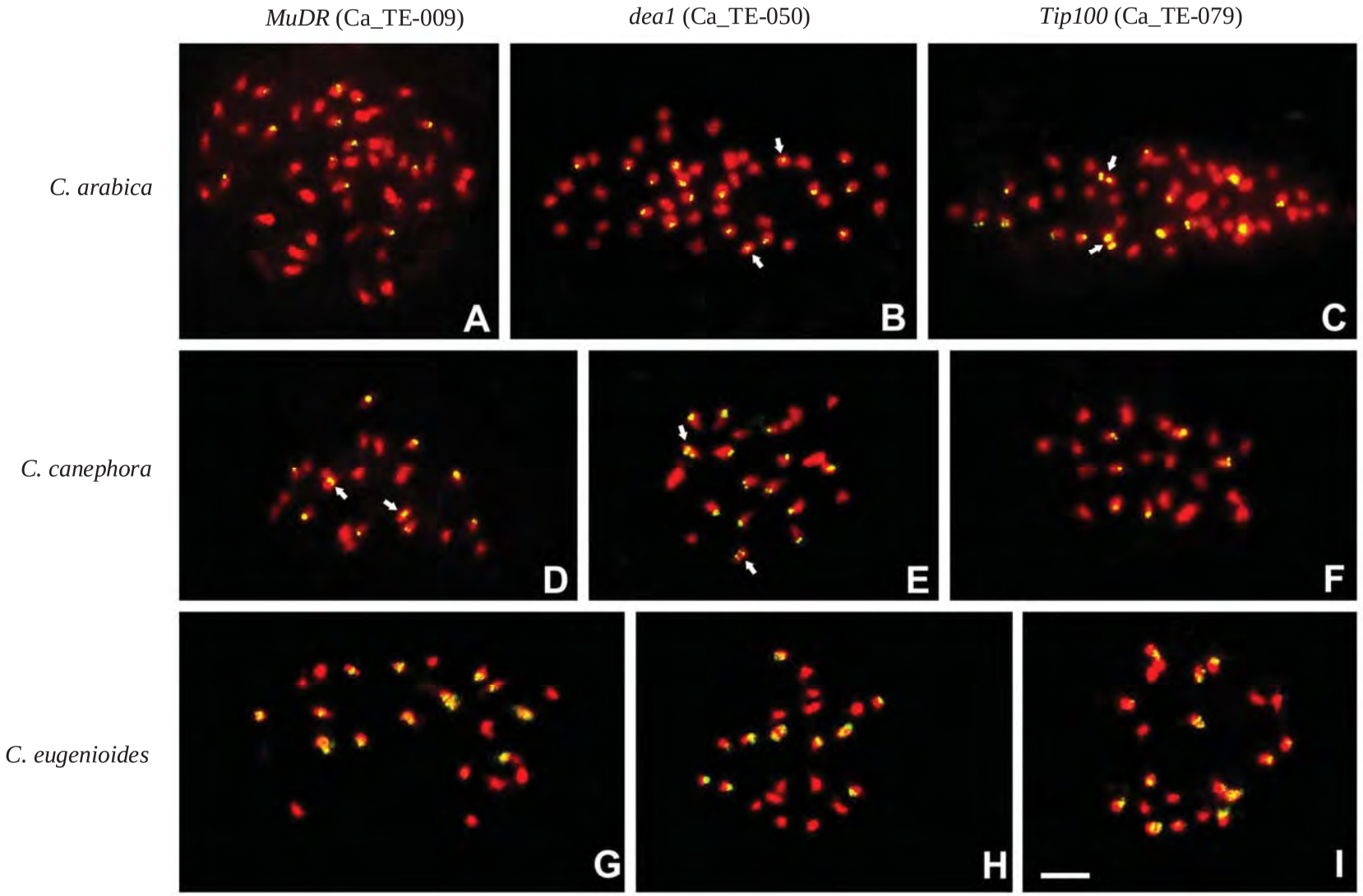


Figure 3. Gene expression profiles of 64 TEs. Left: results obtained by macroarrays. Lines= cDNA clones (arbitrary identification), column: distinct samples from *C. arabica* (embriogenic callus + 300 µg/mL of CHX; callus; I59_I and I59_NI: cultivar tolerant (Iapar59) irrigated or not irrigated, respectively; Rubi_I and Rubi_NI: cultivar sensitive (Rubi) irrigated or not irrigated, respectively) and *C. canephora* (14_I and 14_NI: clone tolerant (14) irrigated or not irrigated, respectively).



Captions on the next page.

Figure 4. *FISH* with three TE probes selected from the *C. arabica* transcriptome and hybridized against *C. arabica* var. *typica* and their progenitors, *C. canephora* and *C. eugenioides*. Probes derived from the cDNA clones Ca_TE-009, Ca_TE-050 and Ca_TE-079 similar to *MuDR*, *dea1* and *Tip100*, respectively. **A.** Signals located in terminal blocks of 12 chromosomes. **B.** Signals also located in terminal blocks of 14 chromosomes and in the pericentromere in one pair of the little chromosome (arrowed). **C.** Signal detected in terminal blocks of 10 chromosomes and one chromosome exhibited two signals, one large terminal signal and another a little proximal signal (arrowed). **D.** Evident signals in 10 chromosomes, two of them in pericentromere regions (arrowed) and another exhibited terminal signals, four of them being more intense (arrow-head). **E.** Evident signals in 14 chromosomes, only one pair (arrowed) exhibited two interstitial signals; the others showed signals in terminal regions. **F.** Terminal signals in four chromosomes and interstitial signals proximal to terminal region in two chromosomes. **G.** Signal located in terminal blocks of 10 chromosomes and two others in the pericentromere. **H.** Signals detected in terminal blocks of 10 chromosomes, one of them presented two signals. **I.** Terminal signals in 13 chromosomes, in which one of them showed two signals: one interstitial and the other a terminal one. Bar represents a scale of 5 μM.

Supplementary Material

Table S1. Completely characterized transposable elements library used for the classification into families of the expressed TEs identified in the *Coffea* transcriptome. The description of this library includes both species and TE names, Genbank accession number (nt and aa) and References for the elements categorized in six groups (*Ty1/Copia*, *Ty3/Gypsy*, Non classified LTRs, *LINEs*, DNA Transposons, *MITEs*). Cells marked with "-" means non available sequences.

Species	TE name	Genbank accession number		Reference
		nucleotides	amino acids	
		Ty1/Copia		
Arabidopsis thaliana	Ta1-1	X53973	RTase: CAA37917	Konieczny et al., 1991
	Ta1-2	X53976	ORF1: CAA37923 ORF2: CAA37924 ORF3: CAA37925	Konieczny et al., 1991
	Ta1-3	X13291	CAA31653	Voytas and Ausubel, 1988
	AtRE1	AB021265	BAA78425	Kuwahara et al., 2000
	AtRE2	AB021264	BAA78424	Kuwahara et al., 2000
Hordeum vulgare	BARE-1	Z17327	Z17327	Manninen and Schulman, 1993
Lycopersicon peruvianum	Retrolyc1	AF228701	-	Costa et al., 1999
Nicotiana tabacum	Tnt1	X13777	CAA32025	Grandbastien et al., 1989
	Tto1	D83003	BAA11674	Hirochika et al., 1996
Oryza sativa	Osr1	-	BAB03249	Jwa, 2000
	Riccopia	M94492	AAA33902	Voytas et al., 1992.
Oryza longistaminata	Retrofit		AAB82754	Song et al., 1995
Saccharomyces cerevisiae	Ty1	-	Q07163	Boecke et al., 1998
Solanum tuberosum	Tst1	X52387	ORF1: CAA36613 ORF2: CAA36614 ORF3: CAA36615 ORF4: CAA36616	Camirand and Brisson, 1990
Solanum lycopersicon	TORTL1	U68072	-	Daraselia et al., 1996
Sorghum bicolor	pHind12	AF078902	-	Miller et al., 1998
Zea mays	PREM-2	U41000	AAB04689	Turcich et al., 1996
	Hopscotch	U12626	AAA57005	White et al., 1994
	Opie-2	U68408	gag: AAC49501 pol: AAC49502	SanMiguel et al., 1996
	Stonor4	AF082134	AAD12998	Marillonnet and Wessler, 1998
	Fourf	U68401 (LTR 5')	-	San Miguel et al., 1996.
	Victim	U68410 (LTR 5')	-	San Miguel et al., 1996.
	Ji-3	U68405 (LTR 5')	-	San Miguel et al., 1996.

		Ty3/Gypsy		
Zea mays	Tekay	AF050455	-	SanMiguel and Bennetzen, 1998
	Cinful1	AF049110	pol: AAD11615	SanMiguel et al., 1996
	Cinful2	AF049111	-	SanMiguel et al., 1996
	Magellan	AF015269	-	Purugganan and Wessler, 1994
	Reina	U69258	-	Wing and Bennetzen, 1996
	Zeon1	U11059	gag: AAA93147	Hu et al., 1995
	Rle	AF057037	-	Wing and Bennetzen, 1996
Arabidopsis thaliana	Athila1	X81801	ORF1: CAA57397	Pelissier et al., 1995
Magnaporthe grisea	Maggy	L35053	gag: AAA33419 pol: AAA33420	Farman et al., 1996
	Pyret	AB062507	AB062507	Nakayashiki et al., 2001
Oryza sativa	RIRE1	D85597	BAA22288	Noma et al., 1987
	RIRE2	AB030283.1	gag: BAA84457 pol: BAA84458	Ohtsubo et al., 1999.
	RIRE3	AB014738	AB014738	Kumekawa et al., 1999
	Retrosat1	AF111709	gag/pol: AAD27548	Llaca et al., non published
	Retrosat2	AF111709	gag-pol: AAD27547	Llaca et al., non published
Sorghum bicolor	Retrosor1	AF098806	AAD19359	Llaca et al., non published
Lilium henryi	Del1-46	X13886	1510387A	Smyth et al., 1989
Ananas comosus	dea1	Y12432	CAA73042	Thomson et al., 1998
		Non classified LTR		
Arabidopsis thaliana	Endovir1-1	AY016208	AAG52950 gag/pol: AAG52949	Peterson-Burch et al., non published
	Frodo	AY923749	Não encontrado	Kalendar and Schulman, non published
Brassica oleracea	Melmoth	Y12321	ORF1: CAA72989 ORF2: CAA72990	Pastuglia et al., non published
Lycopersicon esculentum	Jinling	DQ445619	Não encontrado	Wang et al. 2006.
	LARD1	DQ445621	Não encontrado	Wang et al. 2006.
	LARD2	DQ445622	Não encontrado	Wang et al. 2006.
Lycopersicon chilense	TLC1-1	AF279585	gag/pol: AAK29467	Tapia et al., 2005.
Solanum chacoenses	Sch2	U91993	-	Oosumi and Belknap, 1997
Solanum tuberosum	Potten1	U91987	-	Oosumi and Belknap, 1997
Zea mays	BS1	M25397	ORF1: AAA66269 ORF2: AAA66271 ORF3: AAA66270	Jin and Bennetzen, 1989
		LINEs		
Arabidopsis thaliana	ATLINE1	AB016128	-	Noma et al., 1999.
	ATLINE2	AB016129	-	Noma et al., 1999.

	ATLINE3	AB016130	-	Noma et al., 1999.
	ATLINE4	AB016131	-	Noma et al., 1999.
Lilium speciosum	Del2	Z17425	-	Leeton and Smyth, 1993
Solanum tuberosum	POLN1	AB016140 (endonuclease)	-	Noma et al., 1999.
	POLN2	AB016141 (endonuclease)	-	Noma et al., 1999.
Solanum melongena	EGLN1	AB016142 (endonuclease)	-	Noma et al., 1999.
	EGLN2	AB016143 (endonuclease)	-	Noma et al., 1999.
	EGLN3	AB016144 (endonuclease)	-	Noma et al., 1999.
Zea mays	Cin4	Y00086	-	Schwarz-Sommer et al., 1987
		DNA Transposons		
Arabidopsis thaliana	AtMu1	AC012680.3	AAG52094	Singer et al., 2001
	AtMu2	AB023031.1	BAB 09991	Singer et al., 2001
	TAG1	L12220	AAC25101	Tsay et al., 1993
	TAG2	AF120335	AAD24567	Henk et al., 1999
	Lemi1	DQ888711	ABI75344	Loot et al., 2006.
	Limpet1	U76697	-	Klimyuk and Jones, 1997
	Harbinger	Repbase	-	Kapitonov and Jurka, 2005
	Sadhu	DQ385059	-	Ranghala et al., 2006.
Antirrhinum majus	Tam1	-	CAA40555	Nacken et al., 1991
	Tam2	X06266	-	Krebbers et al., 1987
	Tam3	X55078	CAA38906	Hehl et al., 1991
Glycine max	Soymar	AF078934	AAC28384	Jarvick and Lark, 1998
Ipomoea purpurea	Tip100	AB004906	BAA36225	Habu et al., 1998
Oryza sativa	Tnr3	D63711	-	Motohashi et al., 1996
Sorghum bicolor	Candystripe	AF206660	-	Chopra et al. 1999
Silene latifolia	Thelma13	-	AAP59878	Pritham et al., 2003
Zea mays	Activator	X01380	ORF1: CAA25635 ORF2: CAA25636 ORF3: CAA25637	Kunze and Starlinger, 1989
	BG	X56877	-	Hartings et al., non published
	Doppia4	AF187822	AAG17043	Bercury et al., 2001
	En-1	M25427	AAA66266	Pereira et al., 1986
	Jittery	AF247646	AAF66982	Xu et al., 2004
	MuDR	M76978	MuDRA: AAA21566 MuDRB: AAA21567	Hersberger et al., 1991
	Shooter	AF136220	AAD55677	Panavas et al., 1999
		MITEs		
Arabidopsis thaliana	Emigrant	Z97337	Non-coding	Casacuberta et al., 1998
Oryza sativa	L-K	AF172282	Non-coding	Tarchini et al., 2000
	Wanderer	U34601	Non-coding	Ronald et al., 1992

	Explorer	U70541	Non-coding	Chen and Bennetzen, 1996
Zea mays	Tourist	S48688	Non-coding	Bureau and Wessler, 1992
	Heartbreaker	AF205927	Non-coding	Zhang et al., 2000
	Ditto	AC079830	Non-coding	Buel et al., non published
	Castaway	AC079830	Non-coding	Buel et al., non published
	Gaigin	AC079830	Non-coding	Buel et al., non published
	Stowaway	AC079830	Non-coding	Buel et al., non published
	Crackle	AC079830	Non-coding	Buel et al., non published

Table S2. List of CDSs similar to expressed TE families identified in the transcriptome from *C. arabica*.

Query id	Library	Subject id	% identity	alignment length	mis-matches	gap openings	query start	query end	subject start	subject end	e-value	bit score
DNA Transposons												
Ca_TE-001	RM1	MuDRA	27.00	100	72	1	394	690	118	217	2,00E-06	42.4
Ca_TE-002	SI3	MuDRA	29.90	194	132	2	190	759	133	326	2,00E-23	99.4
Ca_TE-003	IC1	MuDRA	39.71	68	40	1	433	633	136	203	3,00E-10	55.1
Ca_TE-004	IA2	MuDRA	28.91	128	90	1	398	778	136	263	2,00E-16	75.5
Ca_TE-005	EA1	MuDRA	30.36	247	169	2	137	868	136	382	5,00E-36	140
Ca_TE-006	CS1	MuDRA	29.38	194	132	4	89	655	146	339	7,00E-22	94.0
Ca_TE-007	SH2	MuDRA	33.33	264	173	3	99	881	165	424	6,00E-43	164
Ca_TE-008	LV5	MuDRA	29.92	254	173	4	32	778	178	431	4,00E-31	124
Ca_TE-009	FB1	MuDRA	29.50	200	140	3	114	710	310	497	2,00E-22	95.9
Ca_TE-010	FB1	MuDRA	23.43	239	181	5	69	779	337	563	2,00E-17	79.3
Ca_TE-011	CA1	MuDRA	26.54	260	186	4	59	823	343	590	2,00E-23	99.4
Ca_TE-012	FR1	MuDRA	26.11	180	132	3	70	606	343	510	8,00E-20	70.9
Ca_TE-013	FR1	Jittery	28.83	111	67	2	8	304	40	148	5,00E-10	54.7
Ca_TE-014	SI3	Jittery	28.40	257	150	6	96	764	52	304	4,00E-28	107
Ca_TE-015	RT8	Jittery	29.53	298	181	5	77	883	92	388	7,00E-38	147
Ca_TE-016	LV8	Jittery	38.25	251	155	3	59	811	142	389	2,00E-45	171
Ca_TE-017	LV8	Jittery	37.97	266	161	3	76	861	142	403	3,00E-50	188
Ca_TE-018	EA1	Jittery	36.70	188	119	0	300	863	194	381	2,00E-38	149
Ca_TE-019	FB1	Jittery	35.09	285	181	2	53	895	267	550	6,00E-50	187
Ca_TE-020	FB1	Jittery	32.71	321	199	4	52	963	267	572	3,00E-49	185
Ca_TE-021	LV8	Jittery	35.92	142	90	2	77	499	298	431	3,00E-28	85.9
Ca_TE-022	SI3	Jittery	34.88	86	55	1	208	462	538	623	1,00E-08	50.1
Ca_TE-023	SI3	Jittery	34.88	86	55	1	206	460	538	623	1,00E-08	50.1
Ca_TE-024	SI3	Soymar	45.83	168	91	2	279	782	22	184	1,00E-40	156
Ca_TE-025	SI3	Soymar	47.49	179	94	2	357	893	22	195	6,00E-44	167
Ca_TE-026	SI3	Soymar	55.94	261	115	3	105	887	46	300	1,00E-80	288
Ca_TE-027	SI3	Soymar	56.28	183	80	2	96	644	50	227	1,00E-66	206
Ca_TE-028	SI3	Soymar	57.00	200	86	1	109	708	79	276	8,00E-74	238

Ca_TE-029	SH2	Soymar	58.15	270	112	3	90	896	106	372	2,00E-86	308
Ca_TE-030	LV4	Soymar	57.71	175	74	1	52	576	153	326	1,00E-55	202
Ca_TE-031	LV4	Soymar	58.11	265	111	1	58	852	153	416	3,00E-87	311
Ca_TE-032	LV4	Soymar	58.70	247	102	1	54	794	153	398	3,00E-81	291
Ca_TE-033	CS1	Soymar	60.30	199	79	1	63	659	180	377	9,00E-65	236
Ca_TE-034	SI3	Soymar	59.82	112	45	1	407	742	266	376	3,00E-35	132
Ca_TE-035	CB1	TAG2	39.02	205	122	5	108	713	65	264	2,00E-32	129
Ca_TE-036	RX1	TAG2	46.72	244	129	2	824	96	94	335	4,00E-61	224
Ca_TE-037	LV4	TAG2	32.68	205	136	2	80	688	110	311	4,00E-34	127
Ca_TE-038	RT8	TAG2	44.02	209	117	0	50	676	194	402	4,00E-49	184
Ca_TE-039	LV5	TAG2	31.36	287	188	8	94	927	252	516	8,00E-26	107
Ca_TE-040	LV5	TAG2	40.96	83	49	0	528	280	253	335	1,00E-33	77.0
Ca_TE-041	FR2	TAG2	25.13	195	144	3	63	641	300	479	9,00E-23	72.0
Ca_TE-042	FR2	TAG2	29.80	255	177	3	72	830	300	539	3,00E-32	128
Ca_TE-043	SH2	AtMu1	29.48	173	118	3	92	598	266	436	3,00E-22	95.1
Ca_TE-044	LV4	AtMu1	24.18	244	171	3	97	786	451	693	3,00E-15	72.0
Ca_TE-045	IC1	AtMu1	28.46	123	87	1	97	462	491	613	7,00E-09	50.8
Ca_TE-046	CB1	AtMu1	28.47	144	83	3	166	537	556	699	2,00E-11	59.3
Ca_TE-047	PA1	Activator_orf1	32.03	153	103	2	277	732	54	204	3,00E-28	81.3
Ca_TE-048	LV4	Activator_orf2	25.74	202	136	5	278	841	12	188	3,00E-14	55.8
Ca_TE-049	FB2	Tip100	52.06	194	90	2	309	881	380	569	2,00E-48	182
Ca_TE-050	SH2	Tip100	59.55	178	72	0	689	156	383	560	6,00E-55	203
Ca_TE-051	SI3	TAG1	23.53	238	156	8	137	772	22	252	6,00E-15	70.9
Retrotransposons												
Ca_TE-052	LV8	Retrosat2	53.95	215	99	0	23	667	631	845	2,E-71	251
Ca_TE-053	LV8	Retrosat2	53.11	241	113	0	57	779	633	873	3,E-75	271
Ca_TE-054	LV8	Retrosat2	50.62	243	120	0	57	785	633	875	4,E-70	253
Ca_TE-055	LV8	Retrosat2	52.70	241	114	0	56	778	633	873	1,E-76	270

Ca_TE-056	FB1	Retrosat2	35.55	211	135	2	57	686	813	995	3,E-31	125
Ca_TE-057	BP1	Retrosat2	28.28	244	150	4	227	883	891	1131	4,E-35	98.6
Ca_TE-058	BP1	Retrosat2	40.00	105	63	0	263	577	891	995	5,E-30	84.0
Ca_TE-059	FB1	Retrosat2	36.40	250	149	6	762	43	1246	1487	1,E-36	142
Ca_TE-060	FB1	Retrosat2	29.84	248	164	6	769	56	1248	1487	1,E-22	96.3
Ca_TE-061	SH2	Retrosat2	34.08	179	109	5	135	644	1344	1521	9,E-17	77.0
Ca_TE-062	RT5	Retrosat2	31.37	102	67	2	74	370	1365	1466	7,E-09	50.8
Ca_TE-063	FR2	Cin4	34.74	190	123	3	276	842	456	636	2,E-26	109
Ca_TE-064	RT8	Cin4	35.98	164	101	3	566	87	528	685	7,E-27	91.3
Ca_TE-065	PC1	Cin4	35.56	180	113	2	101	631	729	907	1,E-19	86.7
Ca_TE-066	FR1	Cin4	31.03	87	60	0	69	329	1030	1116	5,E-08	48.1
Ca_TE-067	FR1	Cin4	31.03	87	60	0	69	329	1030	1116	5,E-08	48.1
Ca_TE-068	FB2	Melmoth_orf1	36.32	234	149	3	56	757	13	237	6,E-41	156
Ca_TE-069	FR1	Melmoth_orf1	25.97	181	130	2	333	863	21	197	2,E-19	85.9
Ca_TE-070	FR1	Melmoth_orf1	25.71	175	126	2	339	851	21	191	2,E-18	82.4
Ca_TE-071	BP1	Melmoth_orf1	25.33	225	157	5	61	702	55	277	2,E-16	76.3
Ca_TE-072	CL2	Melmoth_orf1	44.16	77	43	1	358	588	543	618	6,E-18	80.9
Ca_TE-073	LV5	Del1	32.75	171	114	2	930	421	66	235	1,E-20	90.1
Ca_TE-074	PA1	Del1	32.35	102	69	0	422	117	134	235	5,E-12	61.2
Ca_TE-075	RM1	Del1	20.42	191	132	6	120	632	255	432	4,E-06	41.6
Ca_TE-076	IC1	Del1	51.02	49	24	1	514	660	1189	1236	7,E-08	47.4
Ca_TE-077	FR1	Del1	42.86	175	100	1	777	253	1268	1441	4,E-42	150
Ca_TE-078	LV8	dea1	67.03	182	60	0	40	585	225	406	1,E-68	249
Ca_TE-079	LV8	dea1	60.70	257	97	2	54	812	225	481	5,E-83	296
Ca_TE-080	RT5	dea1	48.75	80	41	0	312	73	613	692	2,E-20	89.0
Ca_TE-081	FR1	Tst1_orf4	72.60	146	40	0	233	670	1	146	2,E-64	227
Ca_TE-082	LV5	Tst1_orf4	62.02	208	79	2	84	707	100	305	6,E-75	264
Ca_TE-083	CA1	Tst1_orf3	38.34	193	114	3	920	357	202	392	8,E-32	127
Ca_TE-084	SH2	Tst1_orf3	48.57	35	18	0	156	260	358	392	2,E-06	42.7
Ca_TE-085	CA1	Tst1_orf2	33.91	230	133	7	56	688	52	260	2,E-24	102

Ca_TE-086	SH2	Tst1_orf2	40.91	66	37	1	662	853	263	328	1,E-09	53.1
Ca_TE-087	SH2	Melmoth_orf2	53.37	193	90	0	97	675	40	232	4,E-59	217
Ca_TE-088	CB1	Melmoth_orf2	53.50	157	73	0	38	508	88	244	8,E-45	170
Ca_TE-089	CL2	Melmoth_orf2	44.83	87	48	0	76	336	160	246	1,E-16	76.6
Ca_TE-090	PA1	Endovir1-1	38.60	57	35	0	572	742	950	1006	2,E-15	48.5
Ca_TE-091	EA1	Endovir1-1	48.72	156	80	0	579	112	1301	1456	5,E-44	149
Ca_TE-092	PA1	Retrosat1	41.46	41	24	0	500	378	788	828	1,E-09	43.9
Ca_TE-093	FR1	Tnt1	48.00	50	26	0	127	276	1277	1326	7,E-11	57.4
Ca_TE-094	CL2	Ta1_1_rt	61.86	118	45	0	39	392	449	566	9,E-39	150
Ca_TE-095	CB1	Osr1	24.12	199	139	3	309	869	46	239	3,E-19	85.1
Ca_TE-096	LV4	Athila1_orf1	41.62	197	115	0	647	57	52	248	4,E-44	167
Ca_TE-097	EA1	Hopscotch	54.03	124	57	0	354	725	922	1045	3,E-39	140
Ca_TE-098	CA1	Opie2_pol	22.87	188	142	6	94	648	61	232	2,E-08	40.4
Ca_TE-099	CA1	Tto1	22.93	205	157	7	68	679	334	517	6,E-10	54.3
Ca_TE-100	FB1	Maggy_pol	24.27	206	123	5	62	580	564	769	1,E-16	55.5

Query id: arbitrary identification; Library: tissue, developmental stage or stress condition in which the clone was obtained (details see Lopes et al., 2008); **Subject id:** identification of the reference TEs (full description see Additional file 1). Gray cells means cDNA clones in process of sequencing. Legend for the Tables S1-3.

Table S3. List of CDSs similar to expressed TE families identified in the transcriptome from *C. canephora*.

Query id	Library	Subject id	% identity	alignment length	mis-matches	gap openings	q. start	q. end	s. start	s. end	e-value	bit score
DNA Transposons												
CC_TE-001	SE2	MuDRA	30.04	243	165	5	4	717	138	379	1,E-26	110
CC_TE-002	SE4	MuDRA	28.33	180	104	7	505	41	171	347	7,E-15	70.9
CC_TE-003	SE4	MuDRA	27.08	192	105	8	475	5	171	359	2,E-14	69.3
CC_TE-004	SE4	MuDRA	28.33	180	104	7	469	5	171	347	9,E-15	70.9
CC_TE-005	SE4	MuDRA	28.33	180	104	7	468	4	171	347	1,E-14	70.9
CC_TE-006	EC1	MuDRA	29.05	241	168	3	54	767	171	407	2,E-30	122
CC_TE-007	SE4	MuDRA	26.80	291	209	7	9	869	265	522	3,E-22	95.5
CC_TE-008	SE2	Activator_orf1	32.97	182	120	2	11	550	26	194	1,E-22	97.1
CC_TE-009	SE2	Activator_orf1	32.42	182	121	2	6	545	26	194	2,E-22	96.7
CC_TE-010	SE2	Activator_orf1	32.97	182	120	2	6	545	26	194	1,E-22	97.1
CC_TE-011	SE4	Activator_orf1	34.69	147	95	2	43	480	50	194	3,E-21	92.4
CC_TE-012	SH3	Activator_orf1	43.56	101	55	1	232	528	94	194	4,E-20	88.2
CC_TE-013	SH3	Activator_orf1	43.56	101	55	1	231	527	94	194	3,E-20	88.2
CC_TE-014	SE4	AtMu1	24.62	264	194	5	5	781	241	478	4,E-17	78.6
CC_TE-015	SE4	AtMu1	25.00	184	138	3	184	735	302	459	3,E-12	55.8
CC_TE-016	PP1	AtMu1	32.58	89	60	0	112	378	531	619	4,E-11	57.4
CC_TE-017	SE3	AtMu1	25.16	159	104	1	112	543	536	694	3,E-11	58.9
CC_TE-018	SE3	AtMu1	33.33	51	34	0	69	221	563	613	7,E-07	42.7
CC_TE-019	SE3	AtMu1	33.33	51	34	0	30	182	563	613	2,E-06	42.7
CC_TE-020	SE4	Jittery	34.76	233	151	2	70	765	164	394	1,E-39	152
CC_TE-021	LF1	Jittery	41.18	119	70	0	193	549	226	344	3,E-27	100
CC_TE-022	SE4	Jittery	45.37	205	111	1	17	628	227	431	2,E-57	188
CC_TE-023	SE4	Jittery	43.24	259	146	1	20	793	227	485	4,E-61	224
CC_TE-024	SE4	Jittery	37.60	242	150	2	16	738	262	495	3,E-40	155
CC_TE-025	LF1	TAG2	31.23	269	180	6	75	866	275	522	1,E-27	107
CC_TE-026	SE3	TAG2	35.03	197	127	2	18	605	218	413	2,E-33	132
CC_TE-027	SE2	TAG1	32.14	224	151	4	951	283	492	705	9,E-36	140

CC_TE-028	SE2	TAG1	30.16	252	175	4	1034	282	464	705	6,E-35	137
CC_TE-029	SE4	Soymar	36.82	258	162	3	1253	483	167	422	2,E-37	145
Retrotransposons												
CC_TE-030	SE2	Del1	26.09	184	134	5	841	296	63	243	5,E-10	55.1
CC_TE-031	SE2	Del1	50.00	154	75	4	5	460	983	1133	3,E-58	164
CC_TE-032	PP1	Del1	26.79	168	122	4	602	102	60	222	5,E-10	54.7
CC_TE-033	PP1	Del1	42.11	228	122	7	759	106	360	574	6,E-24	100
CC_TE-034	PP1	Del1	38.75	160	97	5	578	102	376	525	2,E-22	94.7
CC_TE-035	SE3	Del1	42.50	280	153	5	868	53	482	756	3,E-39	152
CC_TE-036	SE3	Del1	42.05	264	145	4	828	61	494	756	5,E-42	160
CC_TE-037	SE2	Del1	49.68	157	77	4	58	522	980	1133	5,E-68	165
CC_TE-038	PP1	Del1	54.91	224	99	3	819	154	1201	1420	1,E-67	246
CC_TE-039	PP1	Del1	50.92	163	78	3	632	150	1261	1420	3,E-58	164
CC_TE-040	PP1	Del1	46.55	174	93	1	152	673	1268	1440	3,E-46	175
CC_TE-041	PP1	Del1	55.86	145	62	2	584	156	1277	1420	2,E-42	161
CC_TE-042	PP1	Del1	55.56	144	62	2	579	154	1278	1420	2,E-41	158
CC_TE-043	PP1	Del1	55.63	142	61	2	574	155	1280	1420	3,E-41	157
CC_TE-044	PP1	Del1	55.63	142	61	2	573	154	1280	1420	3,E-41	157
CC_TE-045	PP1	Del1	56.03	141	60	2	571	155	1281	1420	4,E-41	157
CC_TE-046	PP1	Del1	54.68	139	61	2	554	144	1283	1420	4,E-39	150
CC_TE-047	PP1	Del1	55.97	134	57	2	549	154	1288	1420	6,E-39	149
CC_TE-048	PP1	Dea1	75.32	158	39	0	131	604	2	159	5,E-72	248
CC_TE-049	PP1	Dea1	56.50	177	75	3	129	653	2	176	3,E-45	171
CC_TE-050	PP1	Dea1	75.24	105	26	0	102	416	23	127	1,E-37	145
CC_TE-051	PP1	Dea1	67.16	67	22	0	246	446	61	127	3,E-37	102
CC_TE-052	PP1	Dea1	37.04	270	168	2	906	103	70	339	1,E-43	166
CC_TE-053	PP1	Dea1	64.20	162	58	0	616	131	74	235	1,E-59	218
CC_TE-054	PP1	Dea1	70.37	108	32	1	429	106	94	200	9,E-41	156
CC_TE-055	PP1	Dea1	43.02	172	95	1	619	113	103	274	3,E-37	144
CC_TE-056	PP1	Dea1	44.16	154	83	1	560	108	121	274	5,E-33	130
CC_TE-057	PP1	Dea1	44.37	151	81	1	551	108	124	274	6,E-33	129
CC_TE-058	PP1	Dea1	58.52	229	95	0	784	98	131	359	2,E-73	265
CC_TE-059	PP1	Dea1	55.97	134	59	1	175	576	137	265	2,E-37	146
CC_TE-060	PP1	Dea1	41.57	166	97	0	605	108	139	304	4,E-36	140

CC_TE-061	PP1	Dea1	69.80	245	74	0	845	111	158	402	1,E-100	355
CC_TE-062	PP1	Dea1	59.21	152	62	0	561	106	169	320	3,E-49	184
CC_TE-063	PP1	Dea1	51.70	176	84	1	880	356	169	344	1,E-79	178
CC_TE-064	PP1	Dea1	58.39	149	62	0	547	101	172	320	2,E-47	177
CC_TE-065	PP1	Dea1	58.42	190	79	0	786	217	183	372	9,E-66	230
CC_TE-066	PP1	Dea1	63.69	157	57	0	107	577	259	415	3,E-56	207
CC_TE-067	PP1	Dea1	67.65	170	53	1	548	45	280	449	2,E-64	234
CC_TE-068	PP1	Dea1	55.06	158	67	2	573	112	339	496	2,E-42	161
CC_TE-069	PP1	Dea1	59.39	229	93	0	102	788	476	704	1,E-85	305
CC_TE-070	PP1	Dea1	42.60	223	116	5	974	342	477	684	1,E-45	172
CC_TE-071	PP1	Dea1	68.92	74	23	0	374	153	564	637	2,E-42	110
CC_TE-072	PP1	Dea1	58.82	170	70	0	614	105	582	751	1,E-57	212
CC_TE-073	PP1	Dea1	55.49	173	71	1	673	173	595	767	1,E-51	192
CC_TE-074	PP1	Dea1	58.97	156	64	0	577	110	596	751	6,E-54	199
CC_TE-075	PP1	Dea1	58.55	152	63	0	565	110	600	751	1,E-51	191
CC_TE-076	PP1	Dea1	58.55	152	63	0	566	111	600	751	1,E-51	191
CC_TE-077	PP1	Dea1	57.89	76	26	1	229	438	629	704	9,E-39	93.2
CC_TE-078	PP1	Dea1	57.89	76	26	1	233	442	629	704	2,E-38	93.2
CC_TE-079	PP1	Dea1	56.94	72	31	0	127	342	642	713	5,E-22	87.8
CC_TE-080	PP1	Dea1	64.85	165	57	2	626	135	691	852	6,E-62	226
CC_TE-081	PP1	Dea1	57.58	132	55	2	499	107	708	836	7,E-43	162
CC_TE-082	PP1	Dea1	61.49	148	56	2	577	137	708	852	5,E-52	193
CC_TE-083	PP1	Dea1	57.58	132	55	2	490	98	708	836	6,E-43	162
CC_TE-084	PP1	Dea1	61.74	149	56	2	539	96	710	855	4,E-51	190
CC_TE-085	PP1	Dea1	60.96	146	56	2	572	138	710	852	5,E-51	189
CC_TE-086	SE2	Dea1	75.90	166	40	0	645	148	2	167	4,E-66	241
CC_TE-087	SE2	Dea1	72.84	162	44	0	640	155	3	164	2,E-59	219
CC_TE-088	SE2	Dea1	76.15	218	52	0	186	839	18	235	5,E-93	330
CC_TE-089	SE2	Dea1	58.01	362	151	5	50	1132	66	421	1,E-114	402
CC_TE-090	SE2	Dea1	60.67	328	129	1	53	1036	66	392	2,E-115	405
CC_TE-091	SE2	Dea1	64.60	274	97	1	54	875	66	338	1,E-103	365
CC_TE-092	SE2	Dea1	70.83	96	28	0	97	384	74	169	8,E-59	145
CC_TE-093	SE2	Dea1	68.35	297	94	0	136	1026	92	388	1,E-122	422
CC_TE-094	SE2	Dea1	67.54	191	61	1	252	821	124	314	1,E-80	266
CC_TE-095	SE2	Dea1	67.91	134	43	0	554	153	267	400	2,E-49	186
CC_TE-096	SE2	Dea1	45.24	42	23	1	1105	980	267	307	5,E-108	33.1
CC_TE-097	SE2	Dea1	67.91	134	43	0	560	159	267	400	2,E-49	186

CC_TE-098	SE2	Dea1	67.91	134	43	0	557	156	267	400	2,E-49	186
CC_TE-099	SE2	Dea1	67.91	134	43	0	554	153	267	400	2,E-49	186
CC_TE-100	SE2	Dea1	57.79	353	141	3	1097	63	270	618	2,E-115	405
CC_TE-101	SE2	Dea1	54.67	353	151	4	1087	56	270	618	3,E-103	364
CC_TE-102	SE2	Dea1	58.36	353	139	3	1096	62	270	618	7,E-117	410
CC_TE-103	SE2	Dea1	44.02	343	170	8	1097	135	271	594	1,E-59	220
CC_TE-104	SE2	Dea1	56.90	348	142	4	1077	58	276	618	2,E-110	388
CC_TE-105	SE2	Dea1	57.10	345	138	6	1064	60	280	618	7,E-106	373
CC_TE-106	SE2	Dea1	56.56	343	141	3	1062	58	280	618	3,E-110	387
CC_TE-107	SE2	Dea1	56.80	338	138	3	1050	61	285	618	1,E-107	379
CC_TE-108	SE2	Dea1	58.63	336	131	3	1045	62	287	618	1,E-111	392
CC_TE-109	SE2	Dea1	54.87	277	117	3	1037	231	288	560	1,E-103	290
CC_TE-110	SE2	Dea1	57.53	332	133	3	1031	60	291	618	5,E-107	377
CC_TE-111	SE2	Dea1	57.83	332	132	3	1028	57	291	618	9,E-111	380
CC_TE-112	SE2	Dea1	58.13	332	130	4	1021	53	291	618	2,E-106	375
CC_TE-113	SE2	Dea1	56.36	330	136	4	1023	58	294	618	7,E-103	363
CC_TE-114	SE2	Dea1	58.66	329	128	3	1022	60	294	618	7,E-109	383
CC_TE-115	SE2	Dea1	47.19	231	112	5	1006	344	298	524	2,E-70	163
CC_TE-116	SE2	Dea1	58.81	318	123	3	987	58	305	618	8,E-108	370
CC_TE-117	SE2	Dea1	58.54	316	123	3	982	59	307	618	3,E-104	368
CC_TE-118	SE2	Dea1	58.86	316	122	3	981	58	307	618	5,E-108	370
CC_TE-119	SE2	Dea1	58.58	309	120	3	963	61	314	618	3,E-101	358
CC_TE-120	SE2	Dea1	59.15	306	117	3	949	56	317	618	1,E-101	359
CC_TE-121	SE2	Dea1	59.67	305	115	3	949	59	318	618	1,E-104	361
CC_TE-122	SE2	Dea1	34.98	243	153	4	937	224	323	564	1,E-35	140
CC_TE-123	SE2	Dea1	60.00	300	112	3	932	57	323	618	5,E-103	358
CC_TE-124	SE2	Dea1	60.00	300	112	3	932	57	323	618	2,E-101	358
CC_TE-125	SE2	Dea1	47.41	251	122	5	325	1047	323	568	1,E-53	200
CC_TE-126	SE2	Dea1	77.36	106	24	0	330	647	323	428	3,E-52	174
CC_TE-127	SE2	Dea1	49.82	273	133	2	319	1125	323	595	6,E-72	260
CC_TE-128	SE2	Dea1	59.33	300	114	3	934	59	323	618	2,E-99	352
CC_TE-129	SE2	Dea1	59.67	300	113	3	935	60	323	618	5,E-101	357
CC_TE-130	SE2	Dea1	58.66	283	109	3	888	64	340	618	8,E-97	332
CC_TE-131	SE2	Dea1	58.16	282	110	3	877	56	341	618	1,E-99	328
CC_TE-132	SE2	Dea1	49.17	240	114	3	756	61	382	617	1,E-66	222
CC_TE-133	SE2	Dea1	53.22	171	77	1	81	584	476	646	8,E-70	191
CC_TE-134	SE2	Dea1	59.66	119	48	0	411	55	500	618	3,E-87	167

CC_TE-135	SE2	Dea1	54.86	319	144	1	1	957	556	871	9,E-100	353
CC_TE-136	SE2	Dea1	56.95	302	130	1	50	955	573	871	2,E-99	352
CC_TE-137	SE2	Dea1	57.09	296	127	1	50	937	573	865	2,E-96	342
CC_TE-138	SE2	Dea1	50.48	210	104	2	209	838	613	818	2,E-57	212
CC_TE-139	SE2	Dea1	54.88	82	37	1	860	1105	724	803	2,E-31	89.0
CC_TE-140	SE2	Dea1	60.90	312	116	3	137	1054	1	306	1,E-95	339
CC_TE-141	SE1	Dea1	38.31	154	95	0	643	182	649	802	8,E-27	109
CC_TE-142	PP1	Retrosat2	26.57	143	103	3	557	135	96	234	5,E-10	53.9
CC_TE-143	PP1	Retrosat2	28.04	107	75	3	120	434	132	234	7,E-06	40.0
CC_TE-144	PP1	Retrosat2	43.11	225	123	3	169	828	468	676	8,E-38	147
CC_TE-145	PP1	Retrosat2	52.72	239	112	3	112	825	468	694	2,E-61	225
CC_TE-146	PP1	Retrosat2	65.81	117	40	0	452	102	545	661	2,E-39	151
CC_TE-147	PP1	Retrosat2	64.91	114	40	0	449	108	545	658	4,E-37	144
CC_TE-148	PP1	Retrosat2	69.04	197	61	0	695	105	545	741	2,E-74	268
CC_TE-149	PP1	Retrosat2	55.70	158	70	0	94	567	658	815	1,E-51	192
CC_TE-150	PP1	Retrosat2	60.16	123	49	0	150	518	728	850	4,E-43	164
CC_TE-151	PP1	Retrosat2	60.00	85	34	0	413	159	757	841	2,E-30	122
CC_TE-152	PP1	Retrosat2	59.38	128	52	0	598	215	812	939	2,E-44	159
CC_TE-153	PP1	Retrosat2	42.11	114	66	1	494	153	853	962	6,E-20	86.7
CC_TE-154	PP1	Retrosat2	44.63	121	67	1	101	463	894	1013	2,E-22	95.1
CC_TE-155	PP1	Retrosat2	66.32	95	32	0	392	108	1103	1197	7,E-38	147
CC_TE-156	PP1	Retrosat2	44.88	254	138	4	905	150	1241	1488	4,E-56	207
CC_TE-157	PP1	Retrosat2	36.11	108	69	1	492	169	1243	1344	3,E-15	70.9
CC_TE-158	PP1	Retrosat2	54.90	153	68	1	549	94	1336	1488	5,E-43	163
CC_TE-159	PP1	Retrosat2	54.90	153	68	1	610	155	1336	1488	6,E-43	163
CC_TE-160	PP1	Retrosat2	54.90	153	68	1	605	150	1336	1488	6,E-43	163
CC_TE-161	PP1	Retrosat2	53.03	132	61	1	544	152	1357	1488	3,E-40	136
CC_TE-162	SE2	Retrosat2	66.54	269	90	0	65	871	545	813	2,E-103	365
CC_TE-163	SE2	Retrosat2	55.18	299	134	5	67	963	545	831	1,E-82	296
CC_TE-164	SE2	Retrosat2	65.16	221	77	0	716	54	546	766	8,E-81	290
CC_TE-165	SE2	Retrosat2	64.25	221	79	0	714	52	546	766	5,E-79	284
CC_TE-166	SE2	Retrosat2	65.16	221	77	0	716	54	546	766	7,E-81	290
CC_TE-167	SE2	Retrosat2	65.16	221	77	0	719	57	546	766	7,E-81	290
CC_TE-168	SE2	Retrosat2	65.16	221	77	0	715	53	546	766	7,E-81	290
CC_TE-169	SE2	Retrosat2	49.44	180	88	2	318	848	888	1065	5,E-50	173
CC_TE-170	SE2	Retrosat2	46.95	213	110	2	315	944	888	1098	4,E-62	191

CC_TE-171	SE2	Retrosat2	48.62	181	93	0	32	574	1034	1214	2,E-68	185
CC_TE-172	SE2	Retrosat2	50.00	176	88	0	47	574	1039	1214	8,E-87	186
CC_TE-173	SE2	Retrosat2	50.57	174	86	0	55	576	1041	1214	2,E-89	186
CC_TE-174	SE2	Retrosat2	50.57	174	86	0	52	573	1041	1214	1,E-89	186
CC_TE-175	SE2	Retrosat2	50.57	174	86	0	52	573	1041	1214	1,E-90	186
CC_TE-176	SE2	Retrosat2	50.57	174	86	0	43	564	1041	1214	2,E-88	186
CC_TE-177	SE2	Retrosat2	50.57	174	86	0	54	575	1041	1214	2,E-72	186
CC_TE-178	SE2	Retrosat2	50.57	174	86	0	52	573	1041	1214	3,E-87	186
CC_TE-179	SE2	Retrosat2	50.57	174	86	0	53	574	1041	1214	2,E-68	186
CC_TE-180	SE2	Retrosat2	50.57	174	86	0	54	575	1041	1214	2,E-89	186
CC_TE-181	SE2	Retrosat2	37.24	341	203	6	50	1039	1041	1366	9,E-60	220
CC_TE-182	SE2	Retrosat2	50.88	171	84	0	62	574	1044	1214	2,E-88	185
CC_TE-183	SE2	Retrosat2	50.88	171	84	0	59	571	1044	1214	2,E-60	185
CC_TE-184	SE2	Retrosat2	50.88	171	84	0	64	576	1044	1214	3,E-85	185
CC_TE-185	SE2	Retrosat2	50.88	171	84	0	63	575	1044	1214	8,E-90	185
CC_TE-186	SE2	Retrosat2	50.88	171	84	0	65	577	1044	1214	6,E-94	185
CC_TE-187	SE2	Retrosat2	51.19	168	82	0	84	587	1047	1214	3,E-71	190
CC_TE-188	SE2	Retrosat2	50.90	167	82	0	74	574	1048	1214	8,E-90	182
CC_TE-189	SE2	Retrosat2	49.13	173	87	1	80	595	1103	1275	1,E-44	169
CC_TE-190	SE2	Retrosat2	49.13	173	87	1	84	599	1103	1275	2,E-44	169
CC_TE-191	SE2	Retrosat2	49.13	173	87	1	81	596	1103	1275	2,E-44	169
CC_TE-192	SE2	Retrosat2	49.13	173	87	1	80	595	1103	1275	2,E-44	169
CC_TE-193	SE2	Retrosat2	49.69	163	81	1	111	596	1113	1275	4,E-43	165
CC_TE-194	SE2	Retrosat2	49.35	154	77	1	139	597	1122	1275	5,E-40	154
CC_TE-195	SE2	Retrosat2	49.13	173	87	1	83	598	1103	1275	1,E-44	169
CC_TE-196	SE3	Retrosat2	39.85	133	79	1	493	98	686	818	2,E-31	88.6
CC_TE-197	SE3	Retrosat2	54.81	104	46	1	764	456	852	955	6,E-48	119
CC_TE-198	SE4	Cinful1_pol	48.15	135	70	0	459	55	271	405	2,E-34	135
CC_TE-199	PP1	Cinful1_pol	39.83	118	71	0	452	99	282	399	4,E-25	104
CC_TE-200	PP1	Cinful1_pol	37.82	119	74	0	456	100	282	400	3,E-24	102
CC_TE-201	PP1	Cinful1_pol	35.27	224	141	2	764	105	780	1002	1,E-36	142
CC_TE-202	PP1	Cin4	38.06	134	82	2	511	113	512	642	1,E-19	86.7
CC_TE-203	PP1	Cin4	38.06	134	82	2	511	113	512	642	1,E-19	86.7
CC_TE-204	PP1	Cin4	38.06	134	82	2	510	112	512	642	1,E-19	86.7
CC_TE-205	PP1	Cin4	38.06	134	82	2	509	111	512	642	1,E-19	86.7

CC_TE-206	PP1	Cin4	33.94	109	65	3	523	218	803	907	2,E-14	69.7
CC_TE-207	PP1	Opie2_pol	59.62	104	42	0	90	401	658	761	1,E-33	128
CC_TE-208	PP1	Opie2_pol	57.71	201	85	0	709	107	799	999	5,E-63	230
CC_TE-209	PP1	Opie2_pol	63.83	94	34	0	394	113	888	981	1,E-32	128
CC_TE-210	PP1	Maggy_pol	41.53	183	104	1	644	105	319	501	6,E-47	150
CC_TE-211	PP1	Maggy_pol	41.53	183	104	1	650	111	319	501	2,E-46	150
CC_TE-212	PP1	Maggy_pol	41.53	183	104	1	650	111	319	501	5,E-47	150
CC_TE-213	PP1	Osr1	41.38	232	133	4	109	795	290	519	1,E-47	173
CC_TE-214	PP1	Osr1	42.41	224	126	4	111	773	290	511	4,E-45	171
CC_TE-215	SE2	Osr1	58.77	114	47	0	534	193	1000	1113	1,E-53	139
CC_TE-216	PP1	Retrosor1	45.19	104	57	0	414	103	929	1032	9,E-22	92.8
CC_TE-217	SE2	Retrosor1	30.16	189	130	2	87	647	1595	1781	5,E-25	105
CC_TE-218	SE4	AtRE2	46.32	285	152	1	60	911	885	1169	9,E-72	252
CC_TE-219	SE4	AtRE2	46.32	285	152	1	62	913	885	1169	3,E-78	252
CC_TE-220	PP1	TLC1_1	52.21	136	64	2	200	604	4	138	3,E-40	154

Table S4. List of CDSs similar to expressed TE families identified in the transcriptome from *C. racemosa*.

Query id	Library	Subject id	% identity	alignment length	mis-matches	gap openings	q. start	q. end	s. start	s. end	e-value	bit score
DNA Transposons												
CR_TE-001	FV2	TAG1	61.26	111	42	1	178	507	577	687	2E-39	152
CR_TE-002	FV2	TAG1	54.14	133	60	3	133	528	587	714	1E-34	136
CR_TE-003	FV2	TAG1	64.65	99	35	0	103	399	577	675	5E-39	150
Retrotransposons												
CR_TE-004	FR4	Tnt1	58.67	271	112	2	53	865	935	1201	6E-84	300
CR_TE-005	FV2	dea1	64.52	155	55	1	564	100	617	770	2E-60	222
CR_TE-006	FV2	RIRE1	46.30	54	29	0	449	288	421	474	8E-24	60.8

IV. DISCUSSÃO GERAL

IV. DISCUSSÃO GERAL

Genomas de eucariotos são entidades muito complexas e dinâmicas. Somente uma fração desses genomas é ocupada por exons codificantes de proteínas, enquanto a maioria das seqüências não-exônicas consistem de elementos repetitivos, principalmente elementos de transposição (TEs). Embora os TEs fossem, inicialmente, considerados como DNA lixo ou parasita (DOOLITTLE, SAPENZA, 1980; ORGEL, CRICK, 1980), devido a sua aparente falta de função genética, passaram a ser gradualmente reconhecidos por ter um amplo impacto evolutivo. Dada sua prevalência e mobilidade, esses elementos móveis apresentam atividade mutagênica por produzirem mudanças em um único gene, bem como grandes rearranjos genômicos (ZHANG, PETERSON, 1999), que podem alterar o padrão da expressão e função gênica. Além disso, TEs podem gerar enorme variabilidade que pode ser utilizada para criar novos genes ou exons dentro de genes existentes (revisado por KIDWELL, LISH, 1997; BENNETZEN, 2000, 2005; VOLFF, 2006).

Vários estudos recentes focando o genoma humano têm sugerido que eventos de *splicing* alternativo contribuem substancialmente para a criação de éxons dentro de genes existentes, o que, em geral, podem relaxar a seleção negativa contra vários tipos de mudanças evolutivas em um gene funcional (MODREK, LEE, 2003; BOUE et al., 2003; LAREAU et al., 2004). A contribuição dos TEs para a evolução de muitos genes de eucariotos em nível transcricional ocorre por meio de cassetes ou fragmentos inseridos dentro de seqüências de mRNA (MAKALOWSKI et al., 1994; MAKALOWSKI, 2000; NEKRUTENKO, LI, 2001; SOREK et al., 2002; GANKO et al., 2003; VAN DE LAGEMAAT et al., 2003; LOPES et al., 2008).

Há mais de uma década, Makalowski e colaboradores (1994) caracterizaram o mecanismo de integração de um elemento *Alu* em um mRNA humano. Posteriormente, Nekrutenko e Li (2001) anotaram em larga escala a ocorrência de cassetes de TEs em todos os UniGenes humanos disponíveis publicamente, encontrando cerca de 4% dessas seqüências na região codificadora de proteínas (principalmente *Alus*). Lorenc e Makalowski (2003) mostraram que 0,38% das proteínas de vertebrados apresentavam inserções de fragmentos de TEs.

A contribuição dos TEs para formação de novas seqüências codificadoras de proteínas é, desta forma, de particular interesse porque a inserção de TEs em éxons pode alterar a seqüência protéica influenciando diretamente o fenótipo. Diante desse cenário, na primeira fase de nossos estudos, como apresentado no Capítulo 1, realizamos a análise da ocorrência de TEs em ESTs de *Coffea* usando a ferramenta de anotação RepeatMasker o que permitiu identificar 72 unigenes de *C. arabica*, 66 em *C. canephora* e cinco em *C. racemosa* contendo fragmentos derivados de TEs (LOPES et al., 2008). Recentemente, uma parcela da literatura em crescimento sugere duas formas com as quais TEs podem ser incorporados em seqüências expressas do genoma hospedeiro. Na primeira, genes contendo cassetes de TEs expressam diversos transcritos, contendo ou não seqüências derivadas de TEs, gerados por mudanças no padrão de processamento do RNA mensageiro (e.g. SOREK et al., 2003; LEV-MAOR et al., 2003). Na segunda, transcritos contendo cassetes de TEs são codificados por alguns membros de certa família gênica modificados pela inserção, que, entretanto não influenciam a função gênica, desde que os membros não modificados mantenham a informação original (e.g. GOTEA; MAKALOWSKI, 2006).

Entre os unigenes contendo TEs identificados no transcriptoma de *C. arabica*, dois deles, Restaurador da Fertilidade (FR) e Fosfofrutoquinase dependente de pirofosfato (PPi-PFKs) foram selecionados para avaliar os sítios de inserção do TE na estrutura de genes relacionados em *Arabidopsis thaliana*, *Oryza sativa* e *Populus trichocarpa*. Esta análise permitiu identificar um gene FR no genoma do arroz que continha múltiplas inserções de retrotransposons com LTR que originaram novos exons. Esta cópia modificada não necessariamente significa um caso de domesticação molecular devido à ocorrência de outros genes destituídos de inserções de TEs que devem manter a expressão de transcritos selvagens. Além disso, realizamos um levantamento das múltiplas cópias de fosfofrutoquinases (subunidades alfa e beta) nos três genomas, o que permitiu evidenciar uma alta similaridade de seqüência entre o Helitron ATREPX1 e os exons 9 e 10 e intron 9 para as PPi-PFKs subunidade beta. Levando em consideração a freqüência com que Helitrons capturam fragmentos gênicos hospedeiros e a maior similaridade do TE com o gene de *A. thaliana* pudemos destacar

um possível evento de transdução de um fragmento de PPi-PFK subunidade beta daquela espécie mediado pelo Helitron ATREPX1.

O recente registro de Piriyapongsa et al. (2007) de que a escolha do método de busca de TEs influenciaria as estimativas de ocorrência de cassetes de TEs na região codificadora de proteínas nos motivou a reanalisar a freqüência com que estes transcritos são identificados nos transcriptomas de *Coffea spp*, como apresentado no Capítulo 2. Usando o algoritmo tBLASTx e as bibliotecas de TEs consenso do Repbase contra as ESTs do PGCB obtivemos um novo conjunto de unigenes contendo TEs (27 em *C. arabica* e 10 em *C. racemosa*) não redundante àqueles identificados por meio de RepeatMasker, incluindo transcritos com função molecular conhecida e fragmentos de TEs de maior comprimento.

Além disso, procuramos analisar as relações de similaridade entre estas seqüências codantes contendo cassetes de TEs e aquelas sem o cassete, com o objetivo de identificar possíveis eventos de *splicing* alternativo e levantar a diversidade de transcritos putativamente oriundos de genes de cópia única, famílias gênicas e alelos múltiplos (no caso do genoma poliplóide de *C. arabica*). Como resultado, apresentados no Capítulo 2, poucos pares de seqüências alinhadas evidenciaram possíveis casos de processamento diferencial do RNAm, entretanto, alguns fatores parecem limitar esta abundância, como a estratégia de representar os clones de cDNA apenas pela extremidade 5’ e número reduzido de clones em algumas bibliotecas. Por outro lado, estes dois conjuntos de seqüências foram alvos de análises de perfis de expressão por macroarranjos usando cultura de células tratadas ou não com o inibidor de síntese de proteínas Cicloheximida e plantas submetidas a estresse hídrico de *C. arabica* e *C. canephora*. Assim, foi possível selecionar alvos com expressão a ser confirmada por RT-qPCR para os tecidos avaliados, bem como possíveis transcritos que poderiam iniciar a ação do mecanismo de supervisão de qualidade para seqüências de RNAm contendo códon de terminação prematuro, denominado Decaimento de RNA sem sentido.

Além da contribuição de fragmentos de TEs, inseridos na região codificadora de proteína, para a variabilidade do repertório transcricional hospedeiro, outro foco deste estudo foi a análise da expressão de transcritos unicamente relacionados à TEs, ou

comumente chamados de TEs expressos. O estudo das condições que provocam a ativação de TEs, e os mecanismos que os regulam, justificam o interesse de se identificar TEs expressos em diferentes tecidos ou condições específicas. Inicialmente, no Capítulo 3, avaliamos a freqüência de ESTs similares a TEs expressos nas diferentes bibliotecas de cDNA das três espécies de *Coffea* analisadas. Essa busca permitiu identificar 327 ESTs similares a famílias de TEs completamente caracterizados e uma maior abundância/expressão de elementos da superfamília *Ty3/Gypsy* em duas bibliotecas de *C. canephora* (sementes e pericarpo) relacionadas aos elementos *dea1* e *Retrosat* foi também destacada. Uma análise do perfil de expressão para 64 deles mostrou um espectro variável de expressão. Esforço de seqüenciamento tem sido empregado para obter a informação nucleotídica completa desses clones de cDNA, o que permitirá avaliar a sua capacidade em codificar proteínas funcionais. Três TEs ativos que apresentaram baixo perfil de expressão foram selecionados para análise de *FISH*. Como um resultado, os sinais de hibridização evidenciaram que muitas das inserções estão localizadas em blocos cromossômicos terminais e pericentrômeros, sugerindo que a baixa expressão esteja relacionada com seu perfil de distribuição heterocromática. Este achado não é inesperado porque TEs são pobremente representados no transcriptoma embora os genomas vegetais sejam enriquecidos por tais seqüências repetitivas que podem ocorrer dispersas ou em blocos (GAETA et al., 2010), mas tendem a concentrar em centrômeros (JIANG et al., 2004) e em regiões intergênicas mais que gênicas (SANMIGUEL et al., 1998); e assim têm contribuído para expansão ou contração do genoma (BENNETZEN, 2002) e para a diversidade intergênica entre e dentro de espécies (MORGANTE et al., 2007).

O estudo da atividade transcricional de TEs e de genes que abrigam cassetes de TEs no genoma de *Coffea*, usando abordagens computacionais e moleculares, permitiu obter informações relevantes, tanto do ponto de vista evolutivo, permitindo uma descrição preliminar da dinâmica dessas seqüências abundantes e ubíquas nos genomas; como também, para o entendimento da atuação dos mecanismos de regulação gênica hospedeira que poderiam atuar no controle da ativação de TEs e minimizar o impacto dos TEs na expressão e função gênica. Algumas perspectivas de análises foram apresentadas e outras surgirão com a disponibilidade em larga escala de dados

genômicos provenientes da Rede Internacional do Genoma Café, que propiciará traçar um cenário amplo da contribuição dos TEs para a estrutura gênica e sua distribuição cromossomal em espécies do gênero *Coffea*.

V. CONCLUSÕES

O presente trabalho permitiu estabelecer as seguintes conclusões:

1) O transcriptoma das três espécies de *Coffea* analisadas (*C. arabica*, *C. canephora* e *C. racemosa*) apresenta seqüências codificantes modificadas pela inserção de fragmentos de elementos de transposição, e aqueles categorizados como enzimas são os mais freqüentes;
2) Transcritos quiméricos e TEs ativos apresentam um perfil geral de baixa expressão, o qual pode ser associado a um baixo potencial codificante, que deve ser analisado por meio de ferramentas computacionais, e de expressão; como também, alguns desses transcritos podem portar códons de terminação prematuros e serem alvos do mecanismo denominado “Decaimento de RNA sem sentido” (NMD);
3) Cultura de células parece não induzir a ativação de TEs em *C. arabica*.
4) Análises de Hibridização *in situ* Fluorescente para três TEs ativos com baixa expressão em *C. arabica* indicam uma distribuição preferencialmente heterocromática.

VI. REFERÊNCIAS

VI. Referências

AJAMIAN, L. et al. Unexpected roles for UPF1 in HIV-1 RNA metabolism and translation. *RNA,* v. 14, p. 914-927, 2008.

ALMEIDA, L. M. et al. The contribution of transposable elements to *Bos taurus* gene structure. *Gene*, v. 390, p. 180-189, 2007.

AMRANI, N.; SACHS, M. S.; JACOBSON, A. Early nonsense: mRNA decay solves a translational problem. *Nat. Rev. Mol. Cell Biol.*,v. 7, p. 415–425, 2006.

ARABIDOPSIS GENOME INITIATIVE. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, v. 408, p. 796-815, 2000.

ARAÚJO, P. G. et al. Transcriptionally active transposable elements in recent hybrid sugarcane. *The Plant J.*, v. 44, p. 707-717, 2005.

AZZALIN, C. M. et al. Telomeric repeat containing RNA and RNA surveillance factors at mammalian chromosome ends. *Science,* v. 318, p. 798-801, 2007.

AZZALIN, C. M.; LINGNER, J. The human RNA surveillance factor UPF1 is required for S phase progression and genome stability. *Curr. Biol.*, v. 16, p. 433-439, 2006.

BENNETZEN, J. L. Mechanisms and rates of genome expansion and contraction in flowering plants. *Genetica*, v. 115, p. 29-36, 2002.

BENNETZEN, J. L. Transposable elements, gene creation and genome rearrangement in flowering plants. *Curr. Opin. Genet. Dev.*, v. 15, p.621-627, 2005.

BOWEN, N. J.; JORDAN, I. K. Exaptation of protein coding sequences from transposable elements. *Genome Dynamics*, v. 3, p. 131-146, 2007.

BOUE, S.; LETUNIC, I.; BORK, P. Alternative splicing and evolution. *Bioessays*, v. 25, p. 1031-1034, 2003.

BROGNA, S.; WEN, J. Nonsense-mediated mRNA decay (NMD) mechanisms. *Nat. Struct. Mol. Biol.*, v. 16, p. 107-113, 2009.

BROSIUS, J.; GOULD, S. J. On 'genomenclature': a comprehensive (and respectful) taxonomy for pseudogenes and other 'junk DNA'. *Proc. Natl. Acad. Scie. USA*, v. 89, p. 10706-10710, 1992.

CALI, B. M. et al. *smg-7* is required for mRNA surveillance in *Caenorhabditis elegans*. *Genetics*, v. 151, p. 605-616, 1999.

CARLTON, J. M. et al. Draft genome sequence of the sexually transmitted pathogen *Trichomonas vaginalis*. *Nature*, v. 315, p. 207-212, 2007.

CARTER, M. S. et al. A regulatory mechanism that detects premature nonsense codons in T-cell receptor transcripts *in vivo* is reversed by protein synthesis inhibitors *in vitro*. *J. Biol. Chem.*, v. 270, p. 28995-29003, 1995.

CASACUBERTA, L. M.; SANTIAGO, N. Plant LTR-retrotransposons and MITEs: control of transposition and impact on the evolution of plant genes and genomes. *Gene*, v. 311, p. 1-11, 2003.

CONTI, E.; IZAURRALDE, E. Nonsense-mediated mRNA decay: molecular insights and mechanistic variations across species. *Curr. Opin. Cell Biol.,v.* 17, p. 316-325, 2005.

CORDAUX, R. et al. Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. *Proc. Natl. Acad. Scie. USA*, v. 103, p. 8101-8106, 2006.

CUI, Y. et. al. Mutations in the MOF2/SUI1 gene affect both translation and nonsense-mediated mRNA decay. *RNA* v. 5, p. 794-804, 1999.

CULBERTSON, M. R.; UNDERBRINK, K. M.; FINK, G. R. Frameshift suppression in *Saccharomyces cerevisiae* II. Genetic properties of group II suppressors. *Genetics*, v. 95, p. 833-853, 1980.

DAVIS, M. B. et al. Transposable element insertions respecify alternative exon splicing in three *Drosophila* myosin heavy chain mutants. *Genetics*, v. 150, p. 1105-1114, 1998.

DEBARRY, J. D.; GANKO, E.; MCDONALD, J. F. The contribution of LTR retrotransposon sequences to gene evolution in *Mus musculus*. *Mol. Biol. Evol.*, v. 23, p. 479-481, 2005.

DICKEY, L. F. et al. Light modulation of ferredoxin mRNA abundance requires an open reading frame. *Plant Cell*, v. 6, p. 1171-1176, 1994.

DIETZ, H. C. et al. Four novel FBN1 mutations: significance for mutant transcript level and EGF-like domain calcium binding in the pathogenesis of Marphan syndrome. *Genomics*, v. 17, p. 468-475, 1993.

DOOLITTLE W. F.; SAPIENZA, C. Selfish genes, the phenotype paradigm and genome evolution. *Nature*, v. 284, p. 601-603, 1980.

DOMON, E.; SAITO, A; TAKEDA, K. Comparison of the *waxy* locus sequence from a non-*waxy* strain and two *waxy* mutants of spontaneous and artificial origins in barley. *Genes Genet. Syst*., v. 77, p. 351-359, 2002.

ELROUBY, N.; BUREAU, T. E. A novel hybrid open reading frame formed by multiple cellular gene transductions by a plant long terminal repeat retroelement. *J. Biol. Chem.*, v. 276, p. 41963-41968, 2001.

FEDOROFF, N. Transposon and genome evolution in plants. *Proc. Natl. Acad. Scie. USA*, v. 97, p. 7002-7007, 2000.

FELDMAN, M. et al. Rapid elimination of low-copy DNA sequences in polyploidy wheat: A possible mechanism for differentiation of homologous chromosomes. *Genetics*, v. 147, p. 1381-1387, 1997.

FESCHOTTE, C.; JIANG, N.; WESSLER, S. R. Plant transposable elements: where genetics meets genomics. *Nat. Rev. Genet.*, v. 3, p. 329-341, 2002.

GANKO, E. W. et al. Evidence for the contribution of LTR retrotransposon to *C. elegans* gene evolution. *Mol. Biol. Evol.*, v. 20, p. 1925-1931, 2003.

GANKO, E. W. et al. LTR retrotransposon-gene associations in *Drosophila melanogaster*. *J. Mol. Evol.*, v. 62, p. 111-120, 2006.

GAETA, M. L. et al. Occurrence and chromosome distribution of retroelements and NUPT sequences in *Copaifera langsdorffii* Desf. (Caesalpinioideae). *Chromosome Res.*, v. 18, p. 515-524, 2010.

GERBER, A.; O'CONNELL, M. A.; KELLER, W. Two forms of human double-stranded RNA-specific editase 1 (hRED1) generated by the insertion of an *Alu* cassette. *RNA*, v. 3, p. 453-463, 1997.

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

GRANDBASTIEN, M. A. Activation of plant retrotransposons under stress conditions. *Trends in Plant Sciences*, v. 3, p. 181-187, 1998.

GRIMSON, A. et al. SMG-1 is a phosphatidylinositol kinase-related protein kinase required for nonsense-mediated mRNA decay in *Caenorhabditis elegans. Mol. Cell. Biol.* v. 24, p. 7483-7490, 2004.

HAMILTON, A. et al. Two classes of short interfering RNA in RNA silencing. *EMBO J.*, v. 21, p. 4671-4679, 2002.

HAZZOURI, K. M. et al. Contrasting patterns of transposable element insertion polymorphism and nucleotide diversity in autotetraploid and allotetraploid *Arabidopsis* species. *Genetics*, v. 179, p. 581-592, 2008.

HE, P. et al. Allopolyploidy alters gene expression in the highly stable hexaploid wheat. *Plant Mol. Biol.*, v. 52, p. 401-414, 2003.

HILGARD, P. et al. Translated *Alu* sequence determines nuclear localization of a novel catalytic subunit of casein kinase 2. *Am. J. Physiol. Cell. Physiol.*, v. 283, p. C472-C483, 2002.

HILLEREN, P.; PARKER, R. Mechanisms of mRNA surveillance in eukaryotes. *Annu. Rev. Genet.*, v. 33, p. 229-260, 1999.

HILLMAN, R. T.; GREEN, R.E.; BRENNER, S.E. An unappreciated role for RNA surveillance. *Genome Biol.*, v. 5, R8, 2004.

HIR, H. L. et al. The exon-exon junction complex provides a binding platform for factors involved in mRNA export and nonsense-mediated mRNA decay. *EMBO J.*, v. 20, p. 4987-4997, 2001.

HODGKIN, J. et al. A new kind of informational suppression in the nematode *Caenorhabditis elegans. Genetics,* v. 123, p. 301-313, 1989.

HOENICKA, J. et al. A two-hybrid screening of human Tau protein: interactions with *Alu*-derived domain. *Neuroreport*, v. 13, p. 343-349, 2002.

HUGHES, A. L.; FRIEDMAN, R. Transposable element distribution in the yeast genome reflects a role in repeated genomic rearrangement events on an evolutionary time scale. *Genetica*, v. 121, p. 181-185, 2004.

HUGHES, A. L. et al. Nonrandom association of transposable elements with duplicated genomic blocks in *Arabidopsis thaliana*. *Mol. Phyl. Evol.*, v. 29, p. 410–416, 2003.

INTERNATIONAL HUMAN GENOME SEQUENCING CONSORTIUM. A physical map of the human genome. *Nature*, v. 409, p. 934-941, 2001.

INTERNATIONAL RICE GENOME SEQUENCING PROJECT. The map-based sequence of the rice genome. *Nature*, v. 436, p. 793-800, 2005.

JACOBSON, A.; PELTZ, S. W. Interrelationships of the pathways of mRNA decay and translation in eukaryotic cells. *Annu. Rev. Biochem.*, v. 65, p. 693-739, 1996.

JIANG, N. et al. Pack-MULE transposable elements mediate gene evolution in plants. *Nature*, v. 431, p. 569-573, 2004.

JIAO, Y. L.; DENG, X. W. A genome-wide transcriptional activity survey of rice transposable element-related genes. *Genome Biology*, v. 8, R28, 2007.

JIN, Y. K.; BENNETZEN, J. L. Integration and nonrandom mutation of a plasma membrane proton ATPase gene fragment within the *Bs1* retroelement of maize. *Plant Cell*, v. 6, p.1177-1186, 1994.

JOFUKU, K. D; SCHIPPER, R. D.; GOLDBERG, R. B. A frameshift mutation prevents Kunitz trypsin inhibitor mRNA accumulation in soybean embryos. *Plant Cell*, v. 1, p. 567, 1989.

JORDAN, I. K. et al. Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends Genet.*, v. 19, p. 68-72, 2003.

KAPITONOV, V. V.; JURKA, J. Helitrons on a roll: eukaryotic rolling-circle transposons. *Trends Genet.*, v. 23, p. 521-529, 2007.

KASHKUSH, K.; FELDMAN M.; LEVY, A. A. Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat. *Nature Genet.*, v. 33, p. 102-106, 2003.

KERTESZ, S. et al. Both introns and long 3′-UTRs operate as cis-acting elements to trigger nonsense-mediated decay in plants. *Nucleic Acids Res.*, v. 34, p. 6147-6157, 2006.

KIDWELL, M. G. Transposable elements and the evolution of genome size in eukaryotes. *Genetica*, v. 115, p. 49-63, 2002.

KIDWELL, M. G.; LISH, D. Transposable elements as sources of variation in animals and plants. *Proc. Natl. Acad. Scie. USA*, v. 94, p. 7704-7711, 1997.

KROM, N.; RECLA, J.; RAMAKRISHNA, W. Analysis of genes associated with retrotransposons in the rice genome. *Genetica*, v. 134, p. 297-310, 2008.

LEV-MAOR, G. et al. The birth of an alternatively splice exon: 3’ splice-site selection in *Alu* exons. *Science*, v. 300, p. 1288-1291, 2003.

LI, W. et al. Sequence composition, organization, and evolution of the core Triticeae genome. *Plant J.*, v. 40, p. 500-511, 2004.

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

LORENC, A., MAKALOWSKI, W. Transposable elements and vertebrate protein diversity. *Genetica*, v. 118, p. 183-191, 2003.

LUKE, B. et al. *Saccharomyces cerevisiae* Ebs1p is a putative ortholog of human Smg7 and promotes nonsense-mediated mRNA decay. *Nucleic Acids Res.*, v. 35, p. 7688-7697, 2007.

MADLUNG, A. et al. Remodeling of DNA methylation and phenotypic and transcriptional changes in synthetic *Arabidopsis* allotetraploids. *Plant Physiol.*, v. 129, p. 733-746, 2002.

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

MAGER, D. L. et al. Endogenuous retroviruses provide the primary polyadenization signal for two human genes (HHLA2 and HHLA3). *Genomics*, v. 59, p. 255-263, 1999.

MAKALOWSKI ,W. Genomic scrap yard: how genomes utilize all that junk. *Gene*, v. 259, p. 61-67, 2000.

MAKALOWSKI, W.; MITCHEL, G. A.; LABUDA, D. *Alu* sequences in the coding regions of mRNA: a source of protein variability. *Trends Genet.*, v. 10, p. 188-193, 1994.

MAQUAT, L. E. Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics. *Nat. Rev. Mol. Cell Biol.*, v. 5, p. 89-99, 2004.

MARILLONNET, S.; WESSLER, S. R. Retrotransposon insertion into the maize *waxy* gene results in tissue-specific RNA processing. *Plant Cell*, v. 9, p. 967–978, 1997.

MEDGHALCHI, S. M. et al. Rent1, a *trans*-effector of nonsense-mediated mRNA decay, is essential for mammalian embryonic viability. *Hum. Mol. Genet.*, v. 10, p. 99-105, 2001.

MEDSTRAND, P. et al. Impact of transposable elements on the evolution of mammalian gene regulation. *Cytogenet. Genome Res.*, v. 110, p. 342–352, 2005.

MEYERS, B. C.; TINGEY, S. V.; MORGANTE, M. Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. *Genome Res.*, v. 11, p. 1660-1676, 2001.

MITCHELL GA et al. Splice-mediated insertion of an *Alu* sequence inactivates ornithine delta-aminotransferase: a role for *Alu* elements in human mutation. *Proc. Natl. Acad. Scie. USA*, v. 88, p. 815-819, 1991.

MITROVICH, Q. M.; ANDERSON, P. Unproductively spliced ribosomal protein mRNAs are natural targets of mRNA surveillance in *C. elegans*. *Genes Dev.*, v. 14, p. 2173-2184, 2000.

MODREK, B., LEE, C. A genomic view of alternative splicing. *Nature Genet.*, v. 30, p. 13-19, 2003.

MORGANTE, M.; DE PAOLI, E.; RADOVIC, S. Transposable elements and the plant pan-genomes. *Curr. Opin. Plant. Biol.*, v. 10, p. 149-155, 2007.

MORRISON, M.; HARRIS, K. S.; ROTH, M. B. *smg* mutants affect the expression of alternatively spliced SR protein mRNAs in *Caenorhabditis elegans*. *Proc. Natl. Acad. Sci. USA*, v. 94, p. 9782-9785, 1997.

NAGY, E.; MAQUAT, L. E. A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance. *Trends Biochem. Sci.,* v. 23, p. 198-199, 1998.

NEKRUTENKO, A.; LI, W. H. Transposable elements are found in a large number of human protein-coding genes. *Trends Genet.*, v. 17, p. 619-621, 2001.

PAGE, M. F. et al. SMG-2 is a phosphorylated protein required for mRNA surveillance in *Caenorhabditis elegans* and related to Upf1p of yeast. *Mol. Cell. Biol.*, v. 19, p. 5943-5951, 1999.

PAVLICEK, A.; CLAY, O.; BERNARDI, G. Transposable elements enconding functional proteins: pitfalls in unprocessed genomic data? *FEBS Letters*, v. 523, p. 252-253, 2002.

PETRACEK, M. E. et. al. Premature termination codons destabilize ferredoxin-1 mRNA when ferredoxin-1 is translated. *Plant J.*, v. 21, p. 563-569, 2000.

QI, Y. et al. Distinct catalytic and non-catalytic roles of ARGONAUTE4 in RNA-directed DNA methylation. *Nature,* v. 443, p. 1008-1012, 2006.

REBOLLO, R. et al. Jumping genes and epigenetics: Towards new species. *Gene*, v. 454, p. 1-7, 2010.

ROSSI, M.; ARAUJO, P. G.; VAN SLUYS, M. A. Survey of transposable elements in sugarcane expressed sequence tags (ESTs). *Genet. Mol. Biol.*, v. 24, p. 147-154, 2001.

RUDENKO, G. N.; WALBOT, V. Expression and post-transcriptional regulation of maize transposable elements *MuDR* and its derivates. *Plant Cell*, v. 13, p. 553-570, 2001.

RUDENKO, G. N.; ONO, A.; WALBOT, V. Initiation of silencing of maize *MuDR/Mu* transposable elements. *The Plant J.*, v. 33, p. 1013-1025, 2003.

SAKAI, H.; TANAKA, T.; ITOH, T. Birth and death of genes promoted by transposable elements in *Oryza sativa*. *Gene*, v. 392, p. 59-63, 2007.

SANMIGUEL, P. et al. The paleontology of intergene retrotransposon of maize. Nat. Genet., v. 20, p. 44-45, 1998.

SHAPIRO, J. A.; VON STERNBERG, R. Why repetitive DNA is essential to genome function. *Biol. Rev.*, v. 80, p. 227-250, 2005.

SLOTKIN, R. K.; MARTIENSSEN, R. Transposable elements and the epigenetic regulation of the genome. *Nat. Rev. Genet.*, v. 8, p. 272-285, 2007.

SONG, K. et al. Rapid genome change in synthetic polyploids of *Brassica* and its implications for genome evolution. *Proc. Natl. Acad. Scie. USA*, v. 92, p. 7719-7723, 1995.

SOREK, R.; AST, G.; GRAUR, D. *Alu*-containing exons are alternatively spliced. *Genome Res.*, v. 12, p. 1060-1067, 2002.

SUREAU, A. et al. SC35 autoregulates its expression by promoting splicing events that destabilize its mRNAs. *EMBO J.*, v. 20, p. 1785-1796, 2001.

TAKEDA, S. et al. Transcriptional activation of the tobacco retrotransposon *Tto1* by wounding and methyl jasmonate. *Plant Mol. Biol.*, v. 36, p. 365-376, 1998.

THONG, M. R. et al. The role of transposable element clusters in genome evolution and loss of synteny in the rice blast fungus *Magnaporthe oryzae*. *Genome Biol.*, v. 7, p. R16, 2006.

THORNBURG, B. G.; GOTEA, V.; MAKALOWSKI, W. Transposable elements as source of transcription regulation signals. *Gene*, v. 365, p. 104-110, 2006.

TUSKAN, G. A. et al. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science*, v. 313, p. 1596-1604, 2006.

UNTERHOLZNER, L.; IZAURRALDE, E. SMG7 acts as a molecular link between mRNA surveillance and mRNA decay. *Mol. Cell*, v. 16, p. 587-596, 2004.

VAN DE LAGEMAAT, L. N. et al. Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. *Trends Genet.*, v. 19, p. 530-536, 2003.

VANCANNEYT, G.; ROSAHL, S.; WILLMITZER, L. Translatability of a plant-mRNA strongly influences its accumulation in transgenic plants. *Nucleic Acids Res.*, v. 18, p. 2917-2921, 1990.

VARAGONA, M. J.; PURUGGANAN, M.; WESSLER, S. R. Alternative splicing induced by insertion of retrotransposon into the maize *waxy* gene. *Plant Cell*, v. 4, p. 811-820, 1992.

VAUCHERET, H; FAGARD, M. Transcriptional gene silencing in plants: targets, inducers and regulators. *Trends Genet.*, v. 17, p. 29-35, 2001.

VICIENT, C. M. Transcriptional activity of transposable elements in maize. *BMC Genomics*, v. 11, no. 601, 2010.

VIDAL, R. O. et al. A high-throughput data mining of single nucleotide polymorphism in *Coffea* species Expressed Sequence Tags suggests differential homeologous gene

expression in the allotetraploid *Coffea arabica*. *Plant Physiology*, v. 154, p. 1053-1066, 2010.

VIEIRA, L. G. E.; ANDRADE, A. C.; COLOMBO, C. A. et al. Brazilian coffee genome project: an EST-based genomic resource. *Braz. J. Plant. Physiol.*, v. 18, p. 95-108, 2006.

VIGNOLS, F. et al. The *brown midrib3* (*bm3*) mutation in maize occurs in the gene encoding caffeic acid *O*-methyltransferase. *Plant Cell*, v. 7, p. 407-416, 1995.

VOLFF, J. N. Turning junk into gold: domestication of transposable elements and the creation of new genes in eukariotes. *BioEssays*, v. 28, p. 913-922, 2006.

VOINNET, O. RNA silencing as a plant immune system against viruses. *Trends Genet.*, v. 17, p. 449-459, 2001.

VOYTAS, D. F.; NAYLOR, G. J. P. Rapid flux in plant genomes. *Nature Genet.*, v. 20, p. 6-7, 1998.

XIE, Z. et al. Genetic and functional diversification of small RNA pathways in plants. *PLoS Biol.* **2**, e104, 2004.

WAGNER, E.; LYKKE-ANDERSEN, J. mRNA surveillance: the perfect persist. *J. Cell Scie.*, v. 115, p. 3033-3038, 2002.

WEISCHENFELDT, J. et al. NMD is essential for hematopoietic stem and progenitor cells and for eliminating by-products of programmed DNA rearrangements. *Genes Dev.*, v. 22, p. 1381-1396, 2008.

WELCH, E. M.; JACOBSON, A. An internal open reading frame triggers nonsense-mediated decay of the yeast SPT10 mRNA. *EMBO J.*, v. 18, p. 6134-6145, 1999.

WU, J. et al. Nonsense-mediated mRNA decay (NMD) silences the accumulation of aberrant trypsin proteinase inhibitor mRNA in *Nicotiana attenuate*. *Plant J.*, v. 51, p. 693-706, 2007.

YOINE, M.; NISHII, T.; NAKAMURA, K. *Arabidopsis* UPF1 RNA helicase for nonsense mediated mRNA decay is involved in seed size control and is essential for growth. *Plant Cell Physiol.*, v. 47, p. 572-580, 2006.

YU, Q. et al. Recent speciation event of allotetraploid *Coffea arabica* revelead by orthologous BAC sequence comparison. *The 5th Solanaceae Genome Workshop 2008*, Cologne: Alemanha, pp. 69, 2008.

ZHANG, J.; PETERSON, T. Genome rearrangements by nonlinear transposons in maize. *Genetics*, v. 153, p. 1403-1410, 1999.

ZHANG, X.; WESSLER, S. Genome-wide comparative analysis of the transposable elements in related species *Arabidopsis thaliana* and *Brassica oleracea*. *Proc. Natl. Acad. Scie. USA*, v. 101, p. 5589-5594, 2004.

ZHENG, C. L.; FU, X. D.; GRIBSKOV, M. Characteristics and regulatory elements defining constitutive splicing and different modes of alternative splicing in human and mouse. *RNA*, v. 11, p. 1777-1787, 2005.

ZIOLKOWSKI, P. A. et al. Genome evolution in *Arabidopsis/Brassica*: conservation and divergence of ancient rearranged segments and their breakpoints. *The Plant J.*, v. 47, p. 63-74, 2006.

Autorizo a reprodução xerográfica para fins de pesquisa.

São José do Rio Preto, / / .

_____

Assinatura