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*

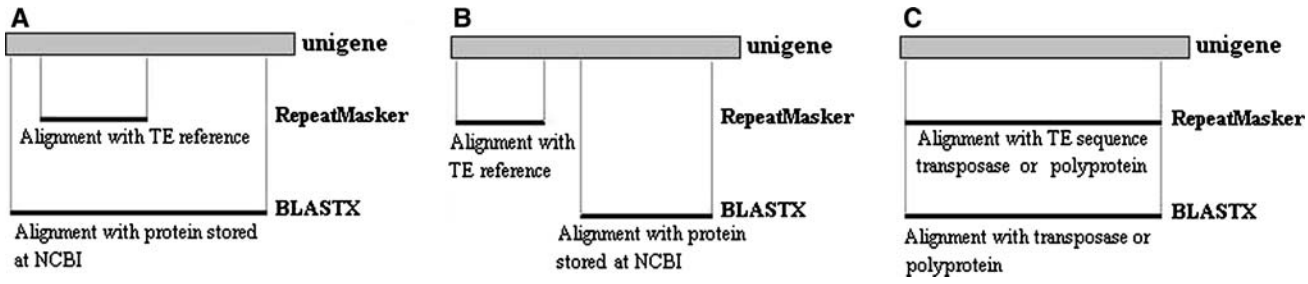

Fig. 1 Possible results of TE localization by RepeatMasker (*Coffea* unigenes database against RepBase plant TE bank) and the region of similarity of the transcript to the protein by BLASTX (*Coffea* unigene database against NR database). **a** The TE-cassette overlaps perfectly with the protein sequence 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

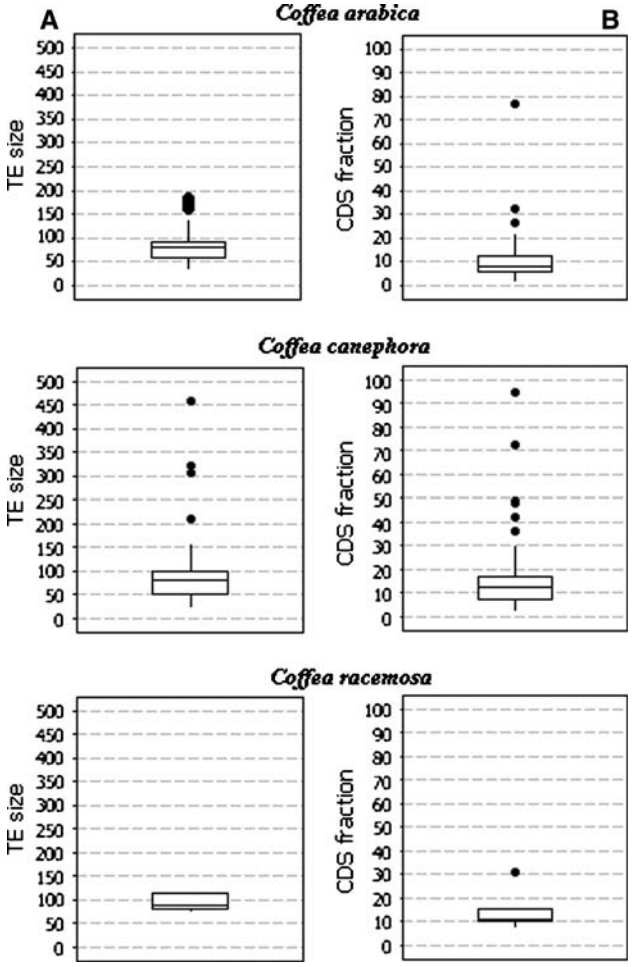

Fig. 2 *Box-plot* of TE-cassettes in unigenes of *Coffea arabica*, *Coffea canephora* and *Coffea racemosa*. **a** Distribution of TE-cassette sizes (in nucleotides) occupied in a CDS. The lengths of TE-cassettes were: *Coffea arabica*: mean = 84.56 ± 3.82 nt and median = 79.50 nt (minimum: 36 nt, maximum: 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

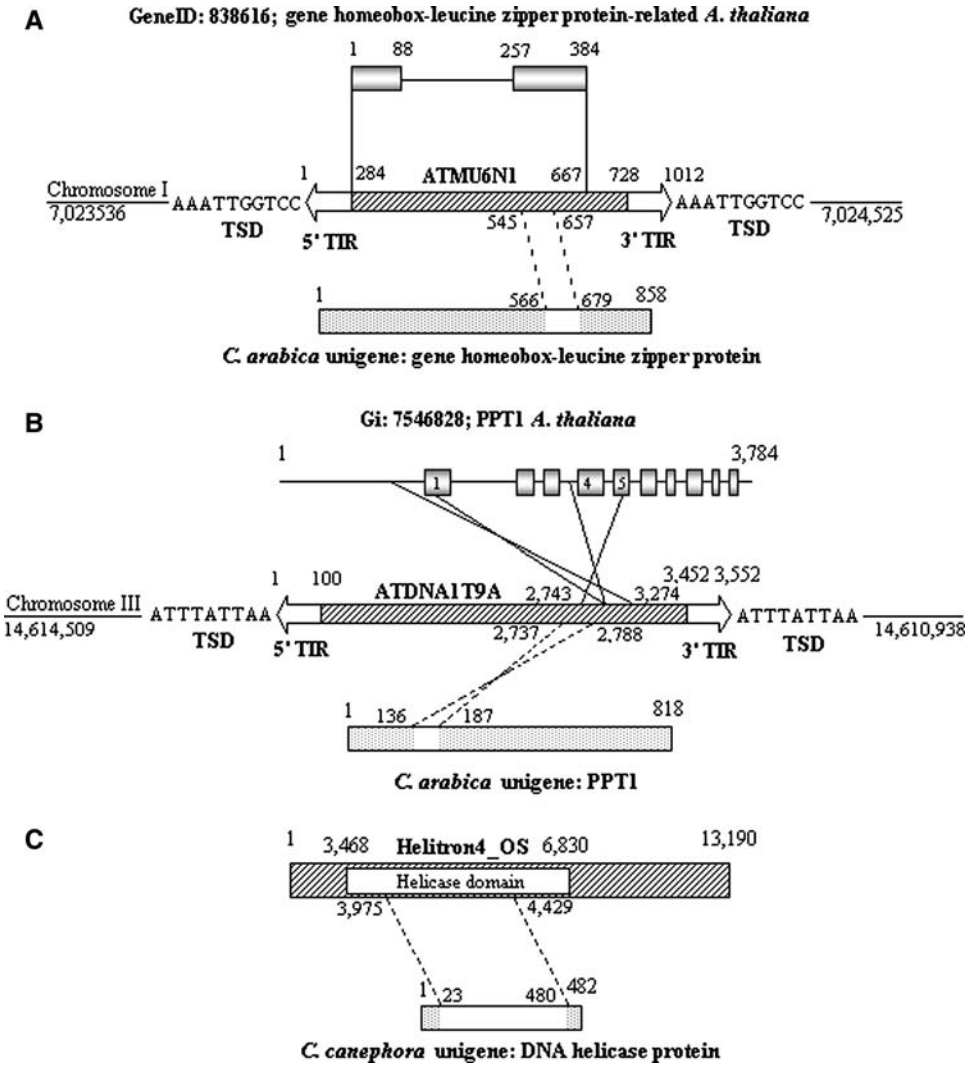

Fig. 3 Relationships between *Coffea* mRNA and genes captured by different TEs. **a** ATMU6N1 harbors the full sequences of exons 1–2 and intron 1 of the ortholog Athb-1 gene of *Arabidopsis thaliana* (position 284–657); **b** ATDNA1T9A harbors a partial sequence of the PPT1 host gene (position 2743–3274) 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.

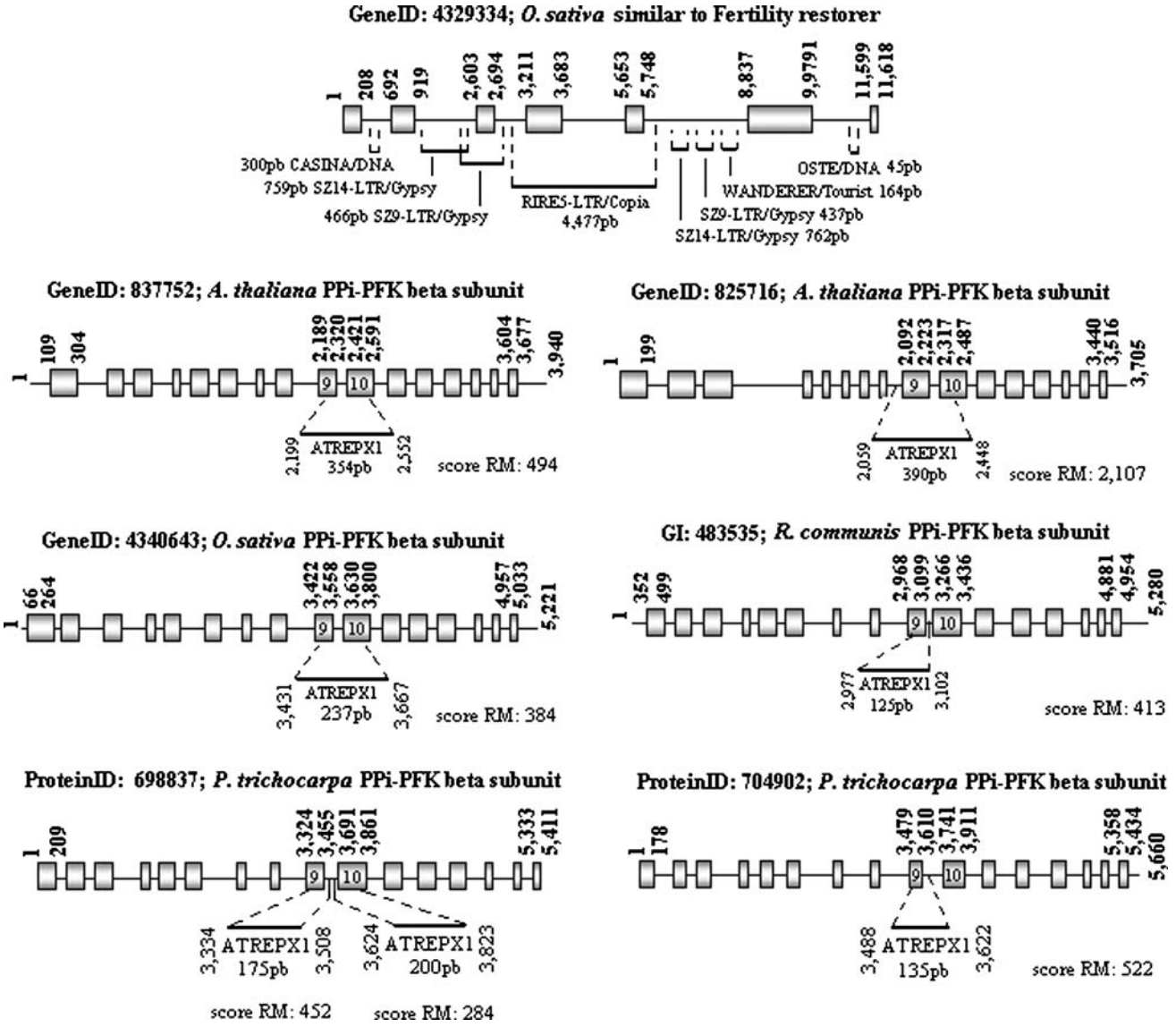

Fig. 4 Schematic representation of the fertility restorer (*FR*) and pyrophosphate-dependent phosphofructokinase (*PPi-PFK*) genes, with exon and intron sequences similar to TEs. The FR gene structure is similar to various TEs, and PPi-PFK presents regions spanning exons 9–10 similar to Helitrons. *Squares* 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*

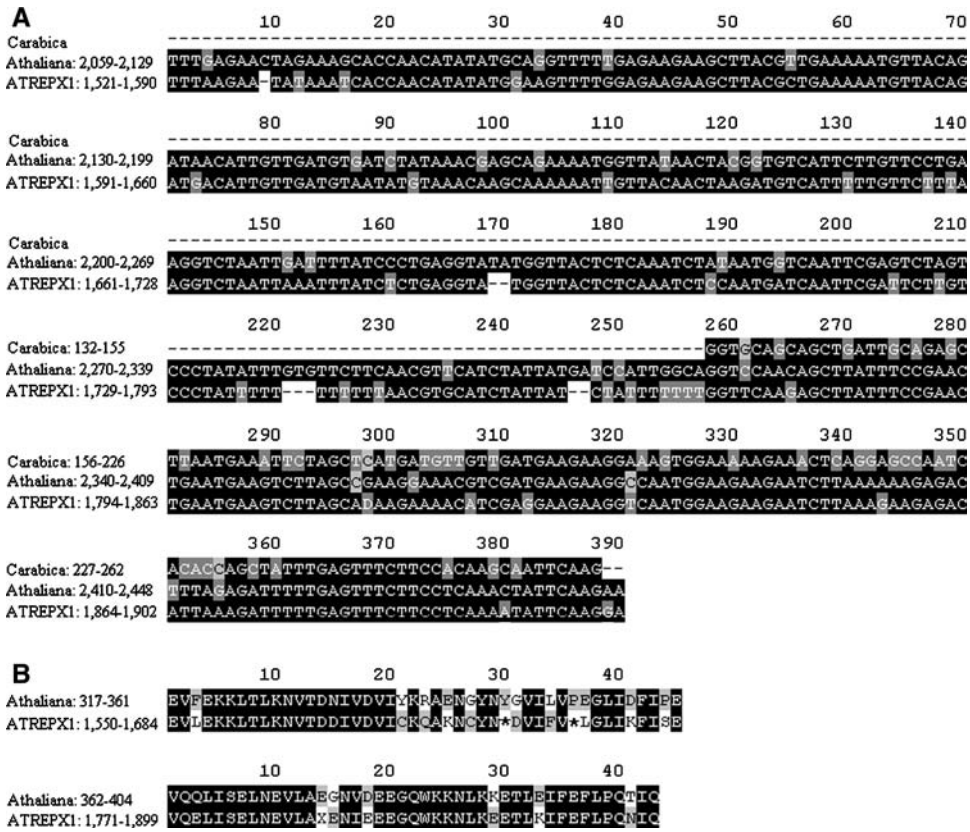

Fig. 5 Alignments of the PPi-PFK *β*-subunit sequences and the ATREPX1 element. **a** Nucleotide sequences of *Coffea arabica* unigene, of *Arabidopsis thaliana* gene (GeneID: 825716) and ATREPX1. The alignment encompasses 130 nt of *Coffea* transcript and 390 nt of *Arabidopsis thaliana* gene, and the 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

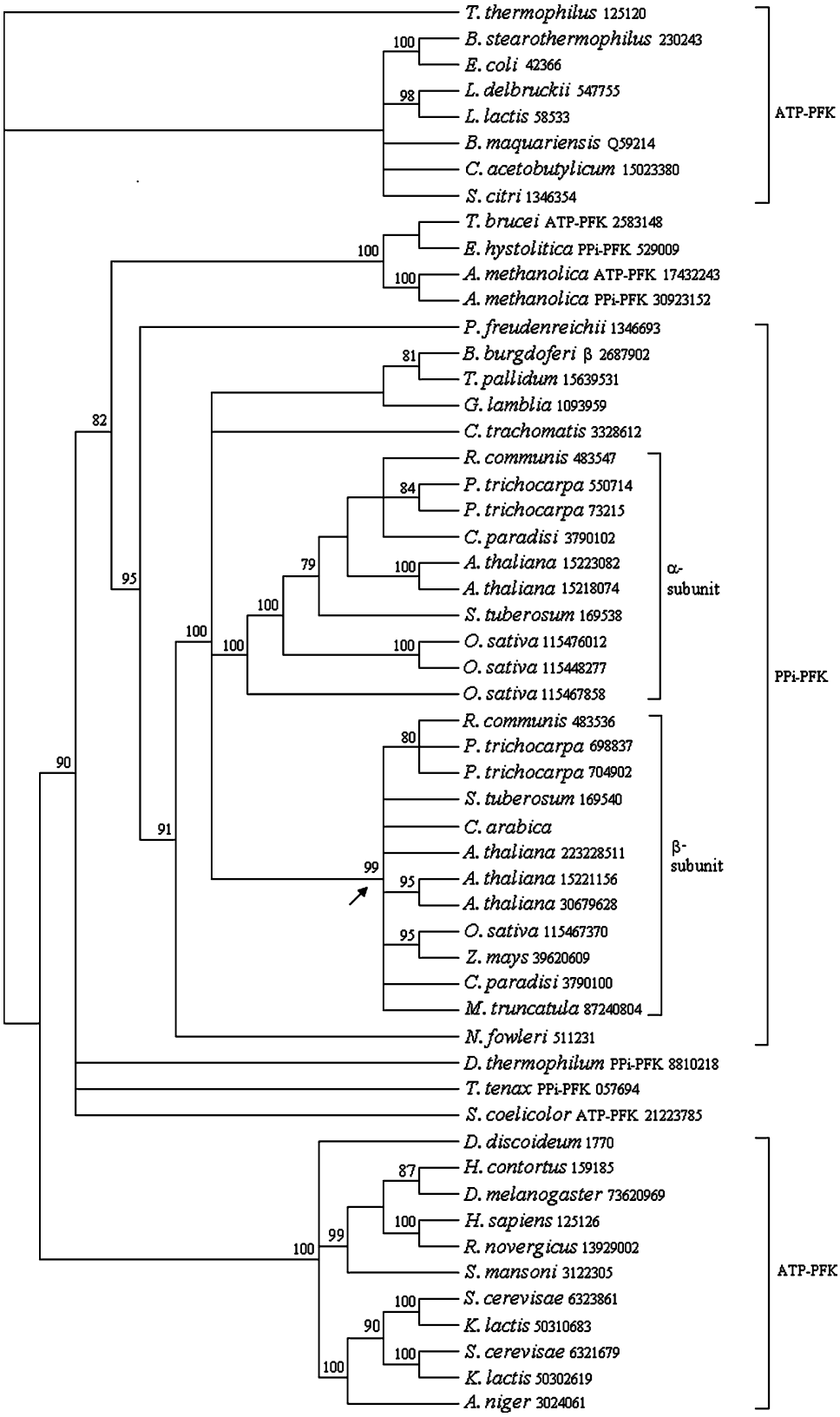

Fig. 6 Unrooted phylogenetic tree of PFK proteins. The cladogram was generated by parsimony analysis using the heuristic algorithm. The PPi- and ATP-PFK sequences were obtained from the GenBank and SwissProt databases and are identified by the host names and GenInfo Identifier (gi). Of 1,122 characters, 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).

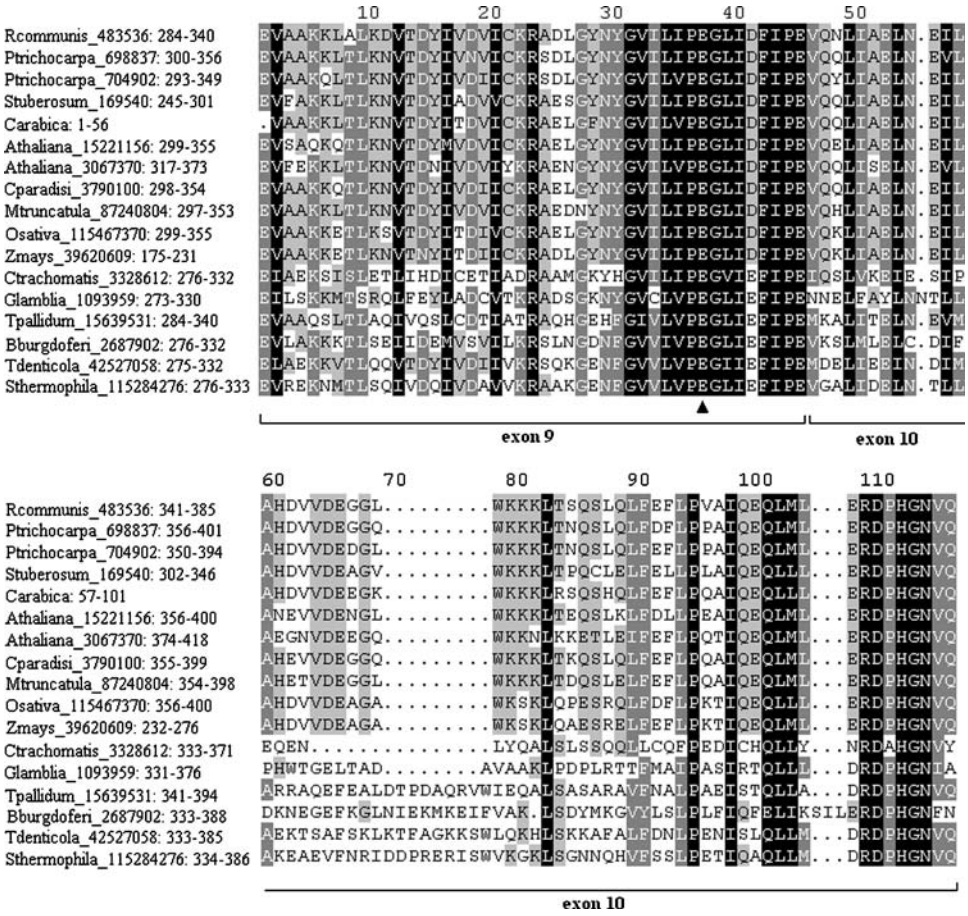

Fig. 7 Multiple alignment between sequences of PPi-PFK homologous proteins using exons 9 and 10. Species names, gi accession numbers and coordinates of residues included in alignment are indicated before every sequence. Identical and similar residues are highlighted in *black* and *gray*, respectively. 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, Sasinowski 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