

Report of a chimeric origin of transposable elements in a bovine-coding gene

L.M. Almeida^{1*}, M.E.J. Amaral^{1*}, I.T. Silva², W.A. Silva Jr.²,
P.K. Riggs³ and C.M. Carareto¹

¹Departamento de Biologia, IBILCE, Universidade do Estado de São Paulo,
São José do Rio Preto, SP, Brasil

²Departamento de Genética, Faculdade de Medicina de Ribeirão Preto,
Universidade de São Paulo, Ribeirão Preto, SP, Brasil

³Department of Animal Science, Texas A&M University,
College Station, TX, USA

*These authors contributed equally to this study.

Corresponding author: L.M. Almeida

E-mail: carreto@ibilce.unesp.br

Genet. Mol. Res. 7 (1): 107-116 (2008)

Received October 24, 2007

Accepted January 2, 2008

Published February 1, 2008

ABSTRACT. Despite the wide distribution of transposable elements (TEs) in mammalian genomes, part of their evolutionary significance remains to be discovered. Today there is a substantial amount of evidence showing that TEs are involved in the generation of new exons in different species. In the present study, we searched 22,805 genes and reported the occurrence of TE-cassettes in coding sequences of 542 cow genes using the RepeatMasker program. Despite the significant number (542) of genes with TE insertions in

exons only 14 (2.6%) of them were translated into protein, which we characterized as chimeric genes. From these chimeric genes, only the FAST kinase domains 3 (FASTKD3) gene, present on chromosome BTA 20, is a functional gene and showed evidence of the exaptation event. The genome sequence analysis showed that the last exon coding sequence of bovine FASTKD3 is ~85% similar to the ART2A retrotransposon sequence. In addition, comparison among FASTKD3 proteins shows that the last exon is very divergent from those of *Homo sapiens*, *Pan troglodytes* and *Canis familiares*. We suggest that the gene structure of bovine FASTKD3 gene could have originated by several ectopic recombinations between TE copies. Additionally, the absence of TE sequences in all other species analyzed suggests that the TE insertion is clade-specific, mainly in the ruminant lineage.

Key words: Cow; Genome; Exaptation; Domestication; Transposon; FASTKD3

INTRODUCTION

To date, the large number of transposable elements (TEs) in almost all genomes sequenced is an interesting phenomenon that has challenged evolutionists and has generated extensive research to examine whether these sequences could play a role in the host genomes.

TEs can be captured by host genome mechanisms and acquire novel functions. Miller et al. (1992, 1995) described the first example of the conversion of a TE-derived DNA into a stable integrated host gene in *Drosophila*. They called this phenomenon molecular domestication, although the same event is also reported in the literature under different names such as exaptation (Brandt et al., 2005) or co-opted events (Sarkar et al., 2003). Additional cases of neofunctionalization have been described in a variety of organisms. In mice, the *Fv1* gene, which inhibits murine leukemia virus infection, is derived from the gag region of an endogenous retrovirus (Best et al., 1996); in vertebrates, RAG proteins, which catalyze the V(D)J recombination of immunoglobulin, might have evolved from a transposase (Agrawal et al., 1998; Hiom et al., 1998); in *Drosophila*, an enzyme similar to telomerase may be derived from the reverse transcriptase from a retrotransposon (Eickbush, 1997; Pardue and DeBaryshe, 2003). Recently, molecular and computational analyses have shown sequences of functional genes derived entirely from mobile element sequences (Britten, 2004; Mallet et al., 2004; Brandt et al., 2005; Iwashita et al., 2006; Cordaux et al., 2006; Gotea and Makalowski, 2006; Volff, 2006; Sakai et al., 2006; Almeida et al., 2007). Iwashita et al. (2003, 2006) reported that bucentaur gene (*bcnt⁹⁷*), which encodes the craniofacial developmental protein1, recruited an AP-END of BovB/RTE as a coding exon, and Almeida et al. (2007) observed that zinc finger 452 gene, involved in various cellular functions including cell proliferation, differentiation and apoptosis, recruited a *Charlie 10* DNA transposon entirely as an exon. In this paper, we report a TE fragment as part of an exon of the bovine FAST kinase domains

3 (FASTKD3) gene, reinforcing that TE-derived sequences have been acquired or gene expression and function.

MATERIAL AND METHODS

Data collection and detection of transposable elements

As in a previous study (Almeida et al., 2007), we retrieved the *Bos taurus* draft assembly (based on Btau_3.1) entries from GenBank and converted them to FASTA-formatted sequence files using the BioPerl toolkit (Stajich et al., 2002). These multifasta files contained exonic regions from 22,805 nuclear genes. The localization and identities of all TE sequences were determined using the RepeatMasker program (<http://www.repeat-masker.org>) which uses the RepBase library of repeat sequences. In addition, our analysis considered only those TE sequences that show a match greater than 100 nucleotides and identity higher than 80% against the sequences deposited in the RepBase database. These stringent parameters were set to avoid spurious results. The RepeatMasker output was parsed with an in-house prepared parser. The most relevant RepeatMasker output values were stored in an MySQL database for more advanced data-mining. Additionally, the NetGene2 program was applied to identify the occurrence of acceptor and donor splicing sites on both last intron and exon of FASTKD3 gene (<http://www.cbs.dtu.dk/services/NetGene2>). The BLAT program was used to give a complete view of TE insertion in FASTKD3 in the bovine genome (<http://genome.ucsc.edu/cgi-bin/hgBlat>).

Comparison between cattle, human, chimp, and dog proteins

The multiple alignments of FASTKD3 protein and other homologous proteins were performed with Clustal W (Thompson et al., 1994). The sequences used in the analyses were obtained from GenBank sequences from FASTKD3 of *Bos taurus* (NP_001019699.1), *Homo sapiens* (NP_076996.2), *Canis familiaris* (XP_545176.2), and *Pan troglodytes* (XP_517625.2).

RESULTS AND DISCUSSION

Evolutionary significance of transposable elements in exons

The initial idea about the role of TEs in host genome is that insertional events of TEs in exon regions should be deleterious because they frequently disrupt the protein functions of genes (Nekrutenko and Li, 2001; Sorek et al., 2002). However, nowadays, several segments derived from TEs have been found in coding regions of host genes (Nekrutenko and Li, 2001; Ganko et al., 2003; Britten, 2004; Lipatov et al., 2005; Iwashita et al., 2003, 2006; Cordaux et al., 2006; Gotea and Makalowski, 2006; Sakai et al., 2007; Almeida et al., 2007). Although the most common event probably is the elimination of the TE sequence before translation by several post-transcriptional mechanisms, these sequences are in some cases required for proper expression and functionality of proteins (Gotea and Makalowski, 2006). Our results confirm such hypothesis. From 542 cow genes

previously identified as containing TE-derived sequences in exons, only 14 (2.6%) are translated into proteins (Almeida et al., 2007). Among these, six TE insertions corresponded to an entire exon. In the current study, we analyzed the remaining 536 genes and identified eight additional possible events of domestication. The chimeric genes identified were: FASTKD3 (NM_001024528, chromosome 20); similar to zinc finger protein 237 - zinc finger protein 198-like 1 (LOC536389, chromosome 12); similar to olfactory receptor MOR227-1 (LOC514818, chromosome 15); similar to olfactory receptor Olfr1418 (LOC527077, unplaced); similar to fetal Alzheimer antigen (LOC532986, chromosome 19); similar to putative 4 repeat voltage-gated ion channel (LOC536103, chromosome 8); hypothetical LOC523303 (LOC523303, chromosome 15), and hypothetical LOC618384 (LOC618384, chromosome 4). Table 1 and Figure 1 describe these cases.

The occurrence of two events could explain the presence of a TE-cassette in an open reading frame. The first possibility would be the direct TE insertion caused by transposition into an exon, and the second would be the indirect insertion by recruiting an intronic TE (Lorenc and Makalowski, 2003). The fact that all TE-cassettes detected in exons in *Bos taurus* begin in an intronic region and extend to the exons is in agreement with mechanism of indirect recruitment of an intronic TE insertion as proposed by Nekrutenko and Li (2001). Curiously, most of the events detected in this study are associated with the last exon (Figure 1). Therefore, another possibility is that insertion of a TE adjacent into a 3' gene boundary can provide new splice donor or acceptor sites, affecting the processing of the primary RNA transcript. Regardless the mechanism used to create a gene novelty, a certain TE insertion will only persist if it is beneficial or at least neutral to the host.

Among the eight chimeric genes cited above, the largest exon composed by a TE-cassette was detected in the protein similar to zinc finger 237 - zinc finger protein 198-like 1 gene

Table 1. Chimeric genes which have a coding sequence derived from transposable element insertions in bovine genome.

<i>Bos taurus</i>	Gene name	Coding sequence (nt)	% TE	TE identifier	Match divergence	GenBank accession number
20	FAST kinase domains 3	1680	19.6	ART2A/SINE	15.3	NM_001024528
12	Similar to zinc finger protein 237	2022	35.9	<i>Zaphod 2</i> /DNA Tip100	16.5	XM_616517.3
15	Similar to olfactory receptor MOR227-1	957	5.9	L1_BT/Line	14.0	XM_592724.3
Un	Similar to olfactory receptor Olfr1418	987	5.9	ART2A/SINE	6.8	XM_605465.2
19	Similar to fetal Alzheimer antigen	8700	0.9	ART2A/SINE	10.4	XM_61229.3
8	Similar to putative 4 repeat voltage-gated ion channel	5200	1.8	BovB/Line RTE	3.8	XM_616223.3
15	Hypothetical LOC523303	1209	16	BovTA/SINEBovA	14.5	XM_601599.3
4	Hypothetical LOC618384	1146	11.2	LTR16A/LTR ERV	26.0	XM_883349.2

This result was obtained using the RepeatMasker program. TE = transposable elements; Un = unplaced.

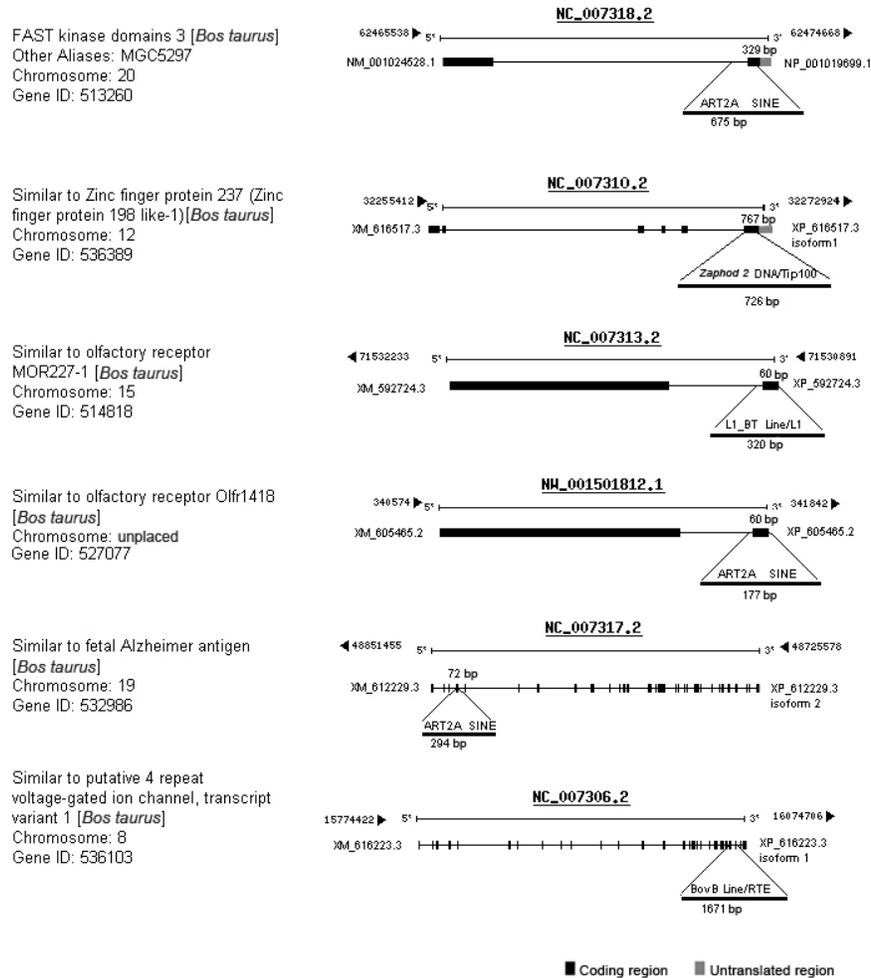


Figure 1. Schematic representation of genes with exon sequences similar to transposable elements **A.** Comparison between FAST kinase domains 3 gene structure and ART2A SINE. **B.** Comparison between similar zinc finger 237 gene structure and *Zaphod 2* DNA transposon. **C.** Comparison between similar to olfactory receptor MOR227 gene structure and *L1BT* retrotransposons. **D.** Comparison between similar to olfactory receptor Olfr1418 and ART2A SINE. **E.** Comparison between similar to fetal Alzheimer antigen and ART2A SINE retrotransposons. **F.** Similar to putative 4 repeat voltage-gated ion channel, transcript variant 1. Squares represent exons and lines introns. The numbers represent the beginning and the end of the exons, as well as the length of the transposable element insertions.

(LOC536389, chromosome 12). The sixth exon of this gene (start point: 16,107 bp, end point: 16,874 bp) is composed of a TE-cassette of 767 bp that shows 83.5% identity with *Zaphod 2* DNA transposon. Almeida et al. (2007) have described a similar case, where the last exon of the mammalian zinc finger 452 gene is originated by *Charlie 10* DNA transposon. There are two extra examples of TE-cassettes inserted into zinc finger genes: LOC617220 (similar to zinc finger protein 193), LOC615117 (similar to zinc finger protein 496). In addition, the olfactory receptor family genes also have TE-cassettes in their structure: similar to olfactory

MOR227-1 (LOC514818); similar to olfactory receptor Olfr1418 (LOC527077), and similar to olfactory receptor Olfr1169 (LOC514883; Almeida et al., 2007). According to Ohno (1999) and Gotea and Makalowski (2006) the molecular domestication or exaptation events should occur more frequently in duplicated genes because one of the duplicates can retain its original function, while the other accumulates molecular changes such that, in time, it can perform a rather different function. Thus, new duplicated copies that receive a TE insertion could evolve free of functional constraints.

The central question of our analysis of exaptation is whether the 8 chimeric genes are actually functional. In fact, there is no evidence of whether or not the majority of such mRNAs result in functional proteins. Among these genes, the FASTKD3 is the only functional candidate gene, showing protein inferred from electronic annotation which provides evidence of being a result of exaptation event in bovine.

FAST kinase domains 3

The FAST kinase gene family represents a conserved region of eukaryotic FAS-activated serine/threonine (FAST) kinases that contain several conserved leucine residues. FAST kinase is rapidly activated during FAS-mediated apoptosis, when it phosphorylates TIA-1, a nuclear RNA binding protein that has been implicated as an effector of apoptosis. The gene extension is 9132 bp and the mRNA encoded has 2021 bp. The last exon has 638 bp, but only 329 bp are translated into protein. All such coding sequence regions are composed of a TE-cassette, which extends more 346 bp within the intron (Figure 1). RepeatMasker analysis showed that the last exon coding sequence matches the regions 40 to 373nt of ART2A (Figure 2) and that this TE-cassette is only 15.3% divergent of ART2A SINE retrotransposon.

FASTKD3	1351	CTTCTCCTCTGCCCCTCA&ATCTTTCCC&CCATCA&G&TCTTTTC&AGTGA	1400
ART2A#3 INE	373	CTTCTCCTCTGCCCCTCA&ATCTTTCCC&CCATCAGGG-CTTTTC&A&TGA	325
FASTKD3	1401	GCCAGTTCTTCG&ATC&CTGCTGCC&A&MT&TTGG>TTC-----AGCA	1444
ART2A#3 INE	324	CTC&GCTCTTCG&ATG&GCTGCC&A&MT&CTGG>TTC&GCTTT&AGCA	275
FASTKD3	1445	TC>CCTTCCC&TG&A&T&TTC&AGG&TG&AT&TTCTTTC&G&G&ATGG&ACTG	1494
ART2A#3 INE	274	TC&ATTCCTTCCC&TG&A&T&TTC&AGG&CTG&A-TCTCCTTT&G&A&TGG&ACTG	226
FASTKD3	1495	GTTGG&ATCTCCTTGC>CC&A&GG&ACTCTCA&G&AGC-CTTCTCCA&AC&C	1543
ART2A#3 INE	225	GTTGG&ATCTCCTTGC>CC&A&GG&ACTCTCA&G&AGCCTTCTCCA&AC&C	176
FASTKD3	1544	C&C&ATTTCA&A&GC&ATCA&TTCTTC>GCTC&AGCTCTCTT&AT&AGTCCA	1593
ART2A#3 INE	175	C&C&AGTTC-CA&A&GC&ATCA&TTCTTC>GCTC&AGCTCTCTT&C&C&AGTCCA	127
FASTKD3	1594	ACTGTC&CATCC&G&C&ATG&ATT&CTGG&A&A&ACC&C&GCTTTG&AC>&T	1643
ART2A#3 INE	126	ACTCTCAC&ATCC&AT&CATG&ACC&AC&GG&A&A&A&ACC&AT&CTCCTTG&ACT&AG&C	77
FASTKD3	1644	GG&ACCTTTGTC&GC&A&G&CTG&ATG&CTT&TG&CTTTT&A&	1680
ART2A#3 INE	76	GG&ACCTTTGTTGG&CA&A>&A&T>CTT&TG&CTTTT&G&A&	40

Figure 2. Alignment between FASTKD3 (FAST kinase domains 3) gene and ART2/SINE using the Clustal W program.

The presence of a TE within the gene is apparently beneficial or neutral to cattle genome since an additional FASTKD3 without the TE-cassette was not detected suggesting the TE-cassette fixation. There are in the literature several examples of positive selection driving

the gene to fixation, diversifying function, and/or changing the function of the gene (Long and Langley, 1993; Nurminsky et al., 1998; Yi and Charlesworth, 2000; Iwashita et al., 2006).

The first step to understanding the evolutionary history of a particular gene is to examine its occurrence, functional products and phylogenetic relationships among species. In order to get some insight into the evolutionary history of FASTKD3, we searched for homologous genes in Amniota species such as *Homo sapiens* (NM_024091.2), *Canis familiaris* (XM_545176.2), *Pan troglodytes* (XM_517625.2), *Mus musculus* (NM_027123.2), and *Rattus norvegicus* (XM_214418.4). An initial analysis showed that the gene structure of *Bos taurus* is very different from other species (Figure 3A). The *Bos taurus* FASTKD3 gene has only two exons; in contrast, FASTKD3 genes of the other species have 6 exons. Only the first exon of *Bos taurus* FASTKD3 is conserved when compared with the other species. A BLAT search showed a high number of TE fragment in the intronic region of *Bos taurus* FASTKD3, which could suggest that the TE insertions into the first intron could have been responsible for chromosomal rearrangements due to ectopic recombination between elements inserted at different genomic sites or even between homologous chromosomes (Figure 3B). In addition, the most probable hypothesis to explain the origin of the new exon is that ART2A element inserted in the intron provided acceptor and donor splicing sites that changed the transcription pattern. Using the NetGene2 program, a site acceptor/donor of splicing was detected in the beginning of the last exon (tctttcccag/aatcaagatc).

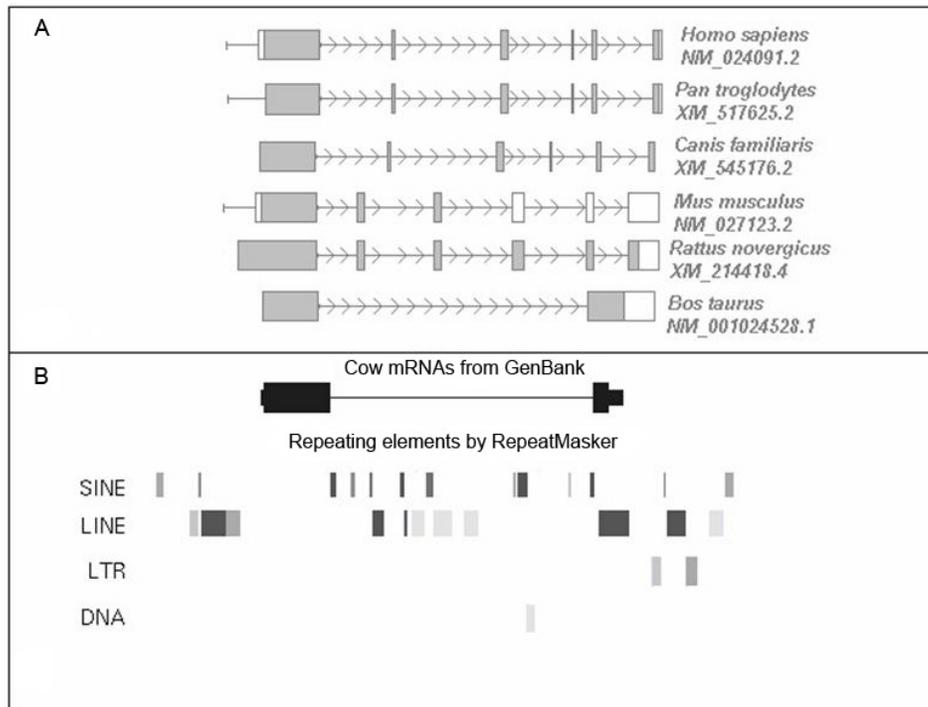


Figure 3. A. FAST kinase domains 3 gene structure among species. The gray squares represent coding sequences and white squares no coding sequences. **B.** The mRNA produced by FAST kinase domains 3 in bovine and the occurrence of transposable elements along the gene (BLAT search).

The functional products of each gene were aligned using Clustal W (Figure 4). The analysis shows that the last exon of *Bos taurus*, which is highly similar to ART2A retrotransposon fragment, is very divergent from the protein encoded by *Homo sapiens*, *Pan troglodytes* and *Canis familiares*. The arrow indicates the point of the TE insertion in the gene and consequently the initial mismatch between the cow and the other sequences. Consequently, the absence of the TE-cassette in all the other species analyzed suggests that the TE insertion was clade-specific, probably in the ruminant ancestor lineage.

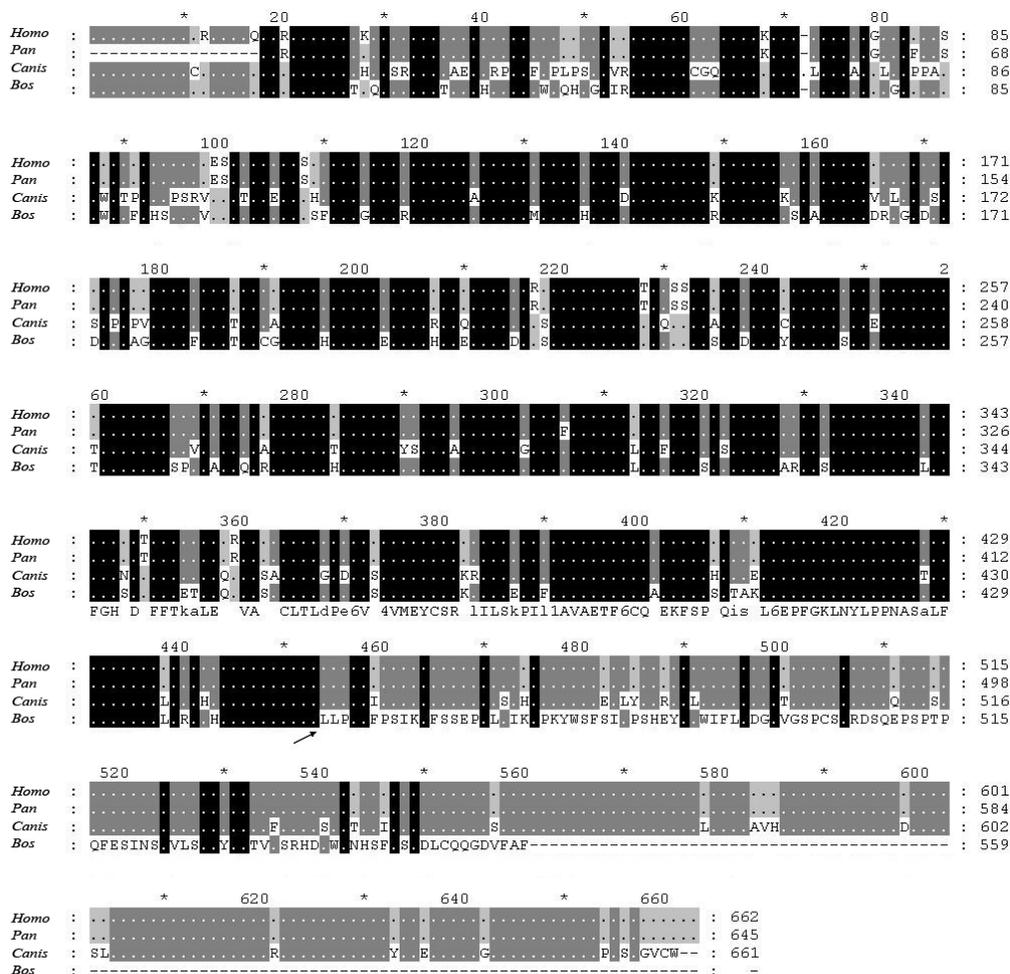


Figure 4. Alignment between FASTKD3 protein from *Bos taurus* (NP_001019699.1), *Homo sapiens* (NP_076996.2), *Canis familiaris* (XP_545176.2), and *Pan troglodytes* (XP_517625.2). The arrow represents the TE-cassette insertion.

CONCLUSIONS

The identification of a TE-cassette in an exon sequence itself is not indicative of adaptive significance because it may only represent a polymorphism. However, there is substantial evidence showing that TEs can insert in gene sequences providing novel exons to resident genes which can evolve with new cellular functions. In this study, we identified eight potential exaptation events responsible for the origination of neogenes. This report shows, together with findings of Almeida et al. (2007) regarding zinc finger 452 and with findings of Iwashita et al. (2006) regarding *bcnt⁹⁷*, that TE gene association seems to be common in the bovine genome and that the novel genes with domesticated TEs, such as *FASTKD3*, have been maintained in the host genome over very long periods of evolution.

ACKNOWLEDGMENTS

Research supported by FAPESP (Grants 04/10148-6 to M.E.J Amaral and a post-doctoral fellowship to L.M. Almeida, 004/00905-4).

REFERENCES

- Agrawal A, Eastman QM and Schatz DG (1998). Transposition mediated by RAG1 and RAG2 and its implications for the evolution of the immune system. *Nature* 94: 744-751.
- Almeida LM, Silva IT, Silva WA Jr, Castro JP, et al. (2007). The contribution of transposable elements to *Bos taurus* gene structure. *Gene* 390: 180-189.
- Best S, Le TP, Towers G and Stoye JP (1996). Positional cloning of the mouse retrovirus restriction gene *Fv1*. *Nature* 382: 826-829.
- Brandt J, Schrauth S, Veith AM, Froschauer A, et al. (2005). Transposable elements as a source of genetic innovation: expression and evolution of a family of retrotransposon-derived neogenes in mammals. *Gene* 345: 101-111.
- Britten RJ (2004). Coding sequences of functioning human genes derived entirely from mobile element sequences. *Proc. Natl. Acad. Sci. U. S. A.* 101: 16825-16830.
- Cordaux R, Udit S, Batzer MA and Feschotte C (2006). Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. *Proc. Natl. Acad. Sci. U. S. A.* 103: 8101-8106.
- Eickbush TH (1997). Telomerase and retrotransposons: which came first? *Science* 277: 911-912.
- Ganko EW, Bhattacharjee V, Schliekelman P and McDonald JF (2003). Evidence for the contribution of LTR retrotransposons to *C. elegans* gene evolution. *Mol. Biol. Evol.* 20: 1925-1931.
- Gotea V and Makalowski W (2006). Do transposable elements really contribute to proteomes? *Trends Genet.* 22: 260-267.
- Hiom K, Melek M and Gellert M (1998). DNA transposition by the RAG1 and RAG2 proteins: a possible source of oncogenic translocations. *Cell* 94: 463-470.
- Iwashita S, Osada N, Itoh T, Sezaki M, et al. (2003). A transposable element-mediated gene divergence that directly produces a novel type bovine Bcnt protein including the endonuclease domain of RTE-1. *Mol. Biol. Evol.* 20: 1556-1563.
- Iwashita S, Ueno S, Nakashima K, Song SY, et al. (2006). A tandem gene duplication followed by recruitment of a retrotransposon created the paralogous bucentaur gene (*bcntp97*) in the ancestral ruminant. *Mol. Biol. Evol.* 23: 798-806.
- Lipatov M, Lenkov K, Petrov DA and Bergman CM (2005). Paucity of chimeric gene-transposable element transcripts in the *Drosophila melanogaster* genome. *BMC Biol.* 3: 24.
- Long M and Langley CH (1993). Natural selection and the origin of jingwei, a chimeric processed functional gene in *Drosophila*. *Science* 260: 91-95.
- Lorenc A and Makalowski W (2003). Transposable elements and vertebrate protein diversity. *Genetica* 118: 183-191.
- Mallet F, Bouton O, Prudhomme S, Cheynet V, et al. (2004). The endogenous retroviral locus ERVWE1 is a bona fide gene involved in hominoid placental physiology. *Proc. Natl. Acad. Sci. U. S. A.* 101: 1731-1736.

- Miller WJ, Hagemann S, Reiter E and Pinsker W (1992). P-element homologous sequences are tandemly repeated in the genome of *Drosophila guanche*. *Proc. Natl. Acad. Sci U. S. A.* 89: 4018-4022.
- Miller WJ, Paricio N, Hagemann S, Martinez-Sebastian MJ, et al. (1995). Structure and expression of clustered P element homologues in *Drosophila subobscura* and *Drosophila guanche*. *Gene* 156: 167-174.
- Nekrutenko A and Li WH (2001). Transposable elements are found in a large number of human protein-coding genes. *Trends Genet.* 17: 619-621.
- Nurminsky DI, Nurminskaya MV, De Aguiar D and Hartl DL (1998). Selective sweep of a newly evolved sperm-specific gene in *Drosophila*. *Nature* 396: 572-575.
- Ohno S (1999). The one-to-four rule and paralogues of sex-determining genes. *Cell Mol. Life Sci.* 55: 824-830.
- Pardue ML and DeBaryshe PG (2003). Retrotransposons provide an evolutionarily robust non-telomerase mechanism to maintain telomeres. *Annu. Rev. Genet.* 37: 485-511.
- Sakai H, Tanaka T and Itoh T (2007). Birth and death of genes promoted by transposable elements in *Oryza sativa*. *Gene* 392: 59-63.
- Sarkar A, Sim C, Hong YS, Hogan JR, et al. (2003). Molecular evolutionary analysis of the widespread *piggyBac* transposon family and related "domesticated" sequences. *Mol. Genet. Genomics* 270: 173-180.
- Sorek R, Ast G and Graur D (2002). Alu-containing exons are alternatively spliced. *Genome Res.* 12: 1060-1067.
- Stajich JE, Block D, Boulez K, Brenner SE, et al. (2002). The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.* 12: 1611-1618.
- Thompson JD, Higgins DG and Gibson TJ (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22: 4673-4680.
- Volff JN (2006). Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. *Bioessays* 28: 913-922.
- Yi S and Charlesworth B (2000). A selective sweep associated with a recent gene transposition in *Drosophila miranda*. *Genetics* 156: 1753-1763.