Available online at www.sciencedirect.com
ScienceDirect
Gene 390 (2007) 180–189
GENE
SECTION
EVOLUTIONARY GENOMICS
www.elsevier.com/locate/gene

The contribution of transposable elements to *Bos taurus* gene structure

Luciane M. Almeida [a,1], Israel T. Silva [b], Wilson A. Silva Jr. [b], Juliana P. Castro [a,1], Penny K. Riggs [c], Claudia M. Carareto [a,1], M. Elisabete J. Amaral [a,*]

[a] *Department of Biology, UNESP-São Paulo State University, IBILCE, Rua Cristovao Colombo, 2265, CEP: 15054-000, São José Rio Preto, SP, Brazil*
[b] *Department of Genetics, School of Medicine of Ribeirão Preto, SP, Brazil*
[c] *Texas A&M University, Department of Animal Science, College Station, TX, USA*

Received 5 June 2006; received in revised form 28 September 2006; accepted 14 October 2006
Available online 28 October 2006
Received by I. King Jordan

Abstract

In an effort to identify the contribution of TEs to bovine genome evolution, the abundance, distribution and insertional orientation of TEs were examined in all bovine nuclear genes identified in sequence build 2.1 (released October 11, 2005). Exons, introns and promoter segments (3 kb upstream the transcription initiation sites) were screened with the RepeatMasker program. Most of the genes analyzed contained TE insertions, with an average of 18 insertions/gene. The majority of TE insertions identified were classified as retrotransposons and the remainder classified as DNA transposons. TEs were inserted into exons and promoter segments infrequently, while insertion into intron sequences was strikingly more abundant. The contribution of TEs to exon sequence is of great interest because TE insertions can directly influence the phenotype by altering protein sequences. We report six cases where the entire exon sequences of bovine genes are apparently derived from TEs and one of them, the insertion of *Charlie* into a bovine transcript similar to the zinc finger 452 gene is analyzed in detail. The great similarity of the TE-cassette sequence to the ZNF452 protein and phylogenetic relationship strongly suggests the occurrence of *Charlie 10* DNA exaptation in the mammalian zinc finger 452 gene.

Keywords: Bovine; Exaptation; Domestication; *Charlie* transposon

1. Introduction

Significant interest exists in studies of the *Bos taurus* genome. This species is an economically important animal, and represents an alternative mammalian model for obesity, infectious diseases and female health (Larkin et al., 2003; Wilson et al., 2005). Since completion of the *B. taurus* draft assembly, this genome sequence has been utilized in studies of non-primate and non-rodent genomes as well as in comparative genomics (Barendse et al., 1994). Similar to human, dog and mouse (3000; 2400 and 2500 Mb respectively) bovine genome size is 3000 Mb with 30 haploid chromosomes. To date, 22,818 genes (22,805 nuclear and 13 mitochondrial genes) have been identified and characterized in the bovine genome. The next step is to determine the location, structure, function and expression of genes affecting health, reproduction, production and product quality in cattle.

Over the last 10 years, an abundance of experimental evidence has accumulated that directly points to the contribution of transposable elements (TEs) to host gene structure, function and expression (Britten, 1996a,b, 1997, 2004; Nekrutenko and Li, 2001; Landry et al., 2001, 2002; Landry and Mager, 2003; Sorek et al., 2002; Jordan et al., 2003; Lorenc and Makalowski, 2003; Van de Lagemaat et al., 2003; Jordan et al., 2003; Han and Boeke, 2004, 2005; Bacci et al., 2005; Dunn et al., 2005, 2006; DeBarry et al.,

Abbreviations: Bov-A, Bov-A2; Bov-tA, Bovidae Short Insterspersed Nuclear Elements; Bov-B, Bovidae Long Insterspersed Nuclear Elements; bp, base pair; CR1, chicken repeat; ERV, endogenous retroviruses; kb, kilo bases; L1, Line-1; L2, Line-2; LINEs, long interspersed sequences; LTRs, long terminal repeat; Mb, mega bases; MER, medium reiterated repeat; *MIR*, mammalian-wid interspersed repeat; nt, nucleotide; ORF, open reading frame; RTE, retrotransposable elements; SCAN domain, domain-swapped homologue of hiv capsid c-terminal domain; SINEs, short interspersed sequences; TEs, transposable elements; ZNF452, zinc finger 452 protein; hAT domain, hobo, activator and Tam3 element domain.

* Corresponding author. Tel.: +55 17 32212407; fax: +55 17 32212390.
E-mail address: eamaral@ibilce.unesp.br (M.E.J. Amaral).
[1] Tel.: +55 17 32212407; fax: +55 17 32212390.

0378-1119/$ - see front matter © 2006 Published by Elsevier B.V.
doi:10.1016/j.gene.2006.10.012

2005; Ganko et al., 2003; Ganko, 2006). These studies also have shown that the evolutionary consequences of TE insertions in the host genome are diverse when they occur within genes.

One of the most extensive literature surveys of TE contribution to host gene regulation identified approximately 80 cases where regulatory elements of vertebrate genes are derived from TEs (Brosius, 1999). Since TEs contain several *cis*-regulatory components including promoter and enhancer sequences, they can influence not only their own activity but also the expression of adjacent genes (Landry et al., 2001, 2002; Medstrand et al., 2001; Jordan et al., 2003; Dunn et al., 2003, 2005, 2006). In addition to acting as promoters and enhancers of nearby genes, TE insertions have also been shown to influence gene expression by providing alternative splicing sites (Varagona et al., 1992; Davis et al., 1998) and polyadenylation sites (Sugiura et al., 1992; Mager et al., 1999) when inserted into intronic regions. For example, an *Alu* element has several cryptic splicing sites embedded within its sequence and can create alternative splicing sites in host genes (Makalowski, 2000; Sorek et al., 2002; Lev-Maor et al., 2003). Besides these regulatory effects, TEs may also contribute to the evolution of coding regions, implicating TE exaptation or domestication as a mechanism for the origin of genetic novelties (Nekrutenko and Li, 2001; Lorenc and Makalowski, 2003; Ganko et al., 2003; Britten, 2004). Several examples of neofunctionalization have been described, in which the TEs have been recruited as neogenes with new cellular functions and have concomitantly lost their ability to transpose (Agrawal et al., 1998; Miller et al., 1999a,b; Donnelly et al., 1999; Pardue and DeBaryshe, 2003; Mallet, 2004; Brandt et al., 2005).

Specifically within the bovine genome, TEs have not yet been well-characterized with regard to mobilization activity, number of insertions within genes and genomes, mode of transmission and impact of their presence on host evolution (Lenstra et al., 1993; Smit, 1996; Okada and Hamada, 1997; Malik and Eickbush, 1998; Kordis and Gubensek, 1999). One method to assess the contributions of TEs to host gene evolution is identification of their presence within genes and promoter sequence regions proximal to host genes. Thus, in an effort to verify the evolutionary influence of TEs, the bovine genome was searched to identify the abundance, distribution and insertional orientation of TEs in all known bovine nuclear genes. Exons, introns and sequences 3 kb upstream transcription initiation sites were screened. The aim of this study was to provide a complete overview of the diversity of TEs present in bovine genes, and evaluate whether TEs can contribute to gene structure with their DNA sequences. The cases described in this survey, in conjunction with those previously described (Lorenc and Makalowski, 2003; Ganko et al., 2003; Britten, 2004) demonstrate the methods by which TEs have contributed to mammalian evolution.

2. Material and methods

2.1. Data collection

We retrieved the *B. taurus* draft assembly build 2.1 (October 11, 2005) entries from GenBank and converted them to FASTA-formatted sequence files using the BioPerl toolkit (Stajich et al., 2002). These multifasta files contained 3 kb sequence upstream of each gene and the intronic and exonic regions from 22,805 nuclear genes. These sequence files were screened with RepeatMasker (http://www.repeatmasker.org) to identify TE insertions. Data containing locus identification, coding sequence, and exon and intron coordinates were stored in a MySQL database where relationships between sequences and repeats were defined.

2.2. Sequence analysis

To identify possible TE insertions in *B. taurus* genes, bovine genomic sequence was screened with RepeatMasker 3.1.2; (http://www.repeatmasker.org). This program identifies copies of TEs by pairwise sequence comparisons with a library of known TEs (RepBase 10.11; http://www.girinst.org/Repbase_Update.html). The following parameters were used for this search: "cross_match" as the search engine; "slow" to obtain a search 0–5% more sensitive than default; "nolow" to not mask low complexity DNA or simple repeats; "norna" to no mask small RNA (pseudo) genes; "species cow" to specify the species or clade of the input sequence; "alignment" to generate a output file showing the alignment. In addition to the parameters selected from the program, our analysis identified a TE insertion as a sequence of at 100 nucleotides that possessed at least 80% identity to a TE sequence in the Repbase database. A TE insertion was designated as a TE-cassette when a fragment of a TE was inserted into an mRNA coding sequence (Gotea and Makalowski, 2006). 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 a MySQL database for more advanced data-mining. TE insertions were classified into three categories according to the gene region where it was identified. Insertions residing completely within an intron were classified as intronic and those completely inserted within a region up to 3 kb from the transcription initiation site were classified as promoter insertions. The window size of the upstream region was chosen based on previous studies in mammalian genome regarding estimates of potential regulatory region size (Jordan et al., 2003; Van de Lagemaat et al., 2003; Landry and Mager, 2003; Thornburg et al., 2005). To classify a TE insertion as exonic, two possibilities were considered: TE insertions completely contained within an exon, as well as those extending from the boundary regions (upstream region/exon or exon/intron), with at least with five nucleotides remaining inside the promoter or exon region.

2.3. Analysis of Repeat Masker results

A PERL script algorithm was developed in-house to classify the relatively large amount of TE data generated by RepeatMasker. Various TE parameters were sorted, including the number of genes that contain TE sequences; the mean number of insertion by gene; the class frequency (SINEs, LINEs, LTR and DNA percentages), the TE diversity found in cow genes; the rate of TE insertions in the same and in the opposite orientation in relation to the gene; the frequency of insertions in promoter

segments, introns and exons; the mean length of TE insertions in each region (promoter segments, introns and exons).

Comparisons of the TE length among different gene regions (promoter segment, introns and exons) were performed using Kruskal–Wallis Test. The frequency of sense and antisense TE insertions was compared using χ^2 test.

2.4. Splicing sites

NetGene2 program has been applied to the prediction of splicing site locations in gene sequences. The last intron and exon of zinc finger 452 gene of *Homo sapiens* (NC_000006.10), *Pan troglodytes* (NW_120489.1), *Canis familiaris* (NC_006617) and *B. taurus* (NW_983609) were submitted to NetGene2 to identify the occurrence of acceptor and donor splicing sites in these regions.

2.5. Evolutionary analysis

The multiple alignments of Zinc finger 452 protein and other homologous proteins were performed with CLUSTAL W (Thompson et al., 1994). The evolutionary relationships among Zinc finger 452 sequences were assessed using the maximum parsimony method (branch and bound algorithm), as implemented in PAUP v.4.0b10 (Swofford, 2000). The sequence used in evolutionary analyses were obtained from GenBank sequences from Zinc finger 452 protein of *B. taurus* (XP_618238); *H. sapiens* (AAI11743; AAI11742; NP_443155; AAI10835; AAS01734); *C. familiaris* (XP_545451); *P. troglodytes* (XP_527300); *Mus musculus* (XP_923559) and protein homologies of *H. sapiens* (BAB67818; BAA92591; BAA24856); *P. troglodytes* (XP_513301); *B. taurus* (XP_590215; XP_592089); *Rattus norvegicus* (XP_342922); *M. musculus* (AAH34119; AAF18453; NP_803413) and *Tetraodon nigroviridis* (CAF95678).

3. Results and discussion

3.1. Many transposable elements are associated with genes in bovine

Of the 22,805 *B. taurus* nuclear genes analyzed in the present study, 20,366 (89.30%) contained TE insertions with an average of 18.41 insertions/gene. The greatest number of insertions (378) was contained within KCNB2 (NM_001024563.1), a 490 kb bovine gene. In contrast, 2079 other genes contained only one TE insertion in their sequence. Fig. 1 shows the number of genes analyzed by chromosome and the number of genes with TE insertions. The frequency of genes with TE insertions ranged from 84.46% for genes located on chromosome 4 to 97.15% for those on chromosome 8.

In this study 375,011 TE insertions were detected, and 359,173 (95.77%) were classified as non-LTR retrotransposons. Of that group, 246,871 (65.83%) were identified as short interspersed sequences (SINEs), 112,302 (29.95%) as long interspersed sequences (LINEs) and 5166 (1.38%) as long terminal repeat retrotransposons (LTRs). The remaining insertions were classified as DNA transposon 10,672 (2.84%). This result shows that mobile elements from all categories contribute to bovine gene variability. However, the TEs are distributed primarily as retrotransposons (98.62%) rather than transposons (1.38%). DNA transposons are not as well documented in mammalian genomes, but have been extensively studied outside mammals (Berg and Howe, 1989; Capy et al., 1998; Craig, 2002). Current researches indicate that DNA transposons in mammals are mostly "fossilized" elements representing ancient, long inactive sequences (Robertson, 1996; Smit and Riggs, 1996).

In general, TE frequency in the bovine genome is consistent with those observed in human genome, where 75% of TE sequences are SINEs and LINEs, 19% are LTR sequences and only 6% are DNA transposons (The Human Genome Sequencing Consortium, 2001). Outside mammals, other genomes, such as maize and lily, show high abundance of retrotransposons in which 50% and 90%, respectively, of their genome is represented by retrotransposons (SanMiguel et al., 1996; Flavell, 1986). On the other hand, in species with smaller genomes, such as yeast, nematodes and fruit flies, the percentage of retrotransposons in the genomes is much lower, typically ranging from 1% to 10% (Cherry et al., 1997; Kaminker et al., 2002; Kidwell, 2002). The reasons for the difference in transposable element content among species are not completed understood, but it is assumed that it can reflect the TE dynamism, capacity for self-regulation of copy number, frequency of vertical and horizontal transfers (Silva et al., 2004), as well as host ability in repressing transposition.

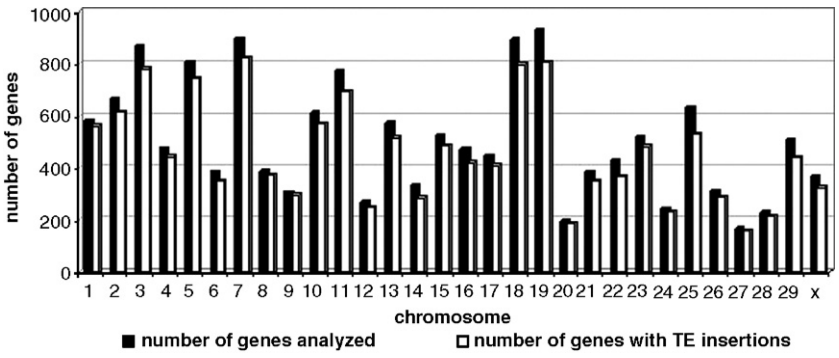

Fig. 1. Distribution of TE-gene association by chromosome. Number of bovine genes analyzed versus number of bovine genes with TE insertions.

3.2. Transposable element diversity in cow genes

In the human genome, most of the copies are retroelements belonging to a few families of LINEs and SINEs, while the remaining TEs represent numerous families with relatively low copy number (The Human Genome Sequencing Consortium, 2001; Ostertag and Kazazian, 2001; Hua-Van et al., 2005). In order to evaluate whether this pattern is also observed in the bovine genome, we analyzed TE family diversity (number of different TE sequences) versus its respective abundance (number of insertion in genes). As result, it was observed that the most abundant class, non-LTR retrotransposons, presented the lowest diversity. All the 246,871 sequences classified as SINE can be divided into only three family types: *Bov-A*, *t-RNA-Glu* and *MIR*. Similarly, LINE sequences, which also have high abundance (112,302 sequences) showed low diversity (*RTE*, *L1*, *L2* and *CR1*). On the other hand, DNA transposon, which presented a low number of copies in cow genes (10,672 sequences), showed the greatest diversity (*Tip 100*, *Achobo*, *MER1_type*, *MER2-type*, *PiggyBac*, *Tc2* and *Mariner*). Finally, the LTR sequences showed the smallest abundance (5166 insertions) and only three groups of sequences *ERVL*, *ERV1* and *MaLR* (Fig. 2). Most of the families identified in this study are widely distributed in mammalian lineages, for instance *L1*, *L2*, *CR1* and *MIR*. Some ancient DNA transposons, such as *Tigger*, are distributed outside mammals (Robertson, 1996). However, *Bov-B*, *Bov-tA* and *Bov-A2* were found to be specific for the suborder Ruminantia.

3.3. TE insertions orientation in relation to the cow genes

TEs can be inserted in the 5′–3′ orientation of the host gene sequences as well as in the opposite orientation. If TE insertional is a random event, it would be expected that sense and antisense insertions occur in the same frequency. However, some studies have shown that some TE sequences are preferentially inserted into the opposite orientation in relation to host gene (Makalowski et al., 1994; Smit, 1999; Medstrand et al., 2002, Lorenc and Makalowski, 2003; Van de Lagemaat et al., 2003; Singer et al., 2004). In order to evaluate whether there is any bias in the rate of TE orientation insertion in the bovine genome, the frequencies of sense and antisense insertions were analyzed. As general result, a statistically similar distribution of sense and antisense insertions (46.30% and 53.70%, respectively) was observed. However, analyzing each class, LINEs and LTR classes were preferentially inserted into the opposite orientation, 61% and 69.75% respectively (both $P<0.05$). Accordingly, studies have shown that L1 (LINEs) and LTRs are more likely to be oriented in the antisense transcriptional direction in the human genome (Smit, 1999; Medstrand et al., 2002; Van de Lagemaat et al., 2003). It is thought that regulatory motifs such as polyadenylation signals within these elements are more likely to be detrimental by, for instance, leading to truncated proteins when oriented in the same direction as the gene. In relation to the DNA and SINE classes any bias in insertion direction was observed (48% and 49.9%, respectively of antisense insertions).

Considering only exons, the frequencies of sense and antisense insertions were similar for DNA transposons (52.72% sense and 47.27% antisense), LINEs (48.19 % sense and 51.81% antisense), LTRs (59% sense and 41% antisense) and SINEs (53.66% sense and 46.33% antisense). Singer et al. (2004) showed that 85% of *Alu* sequences contained in exons of human genes are in the antisense orientation. Hence, our results for the bovine genome differ from those observed in the human genome, where *Alu* sequences (SINE subclass) are preferentially inserted into the opposite orientation in relation to the host gene (Makalowski et al., 1994; Lorenc and Makalowski, 2003; Singer et al., 2004).

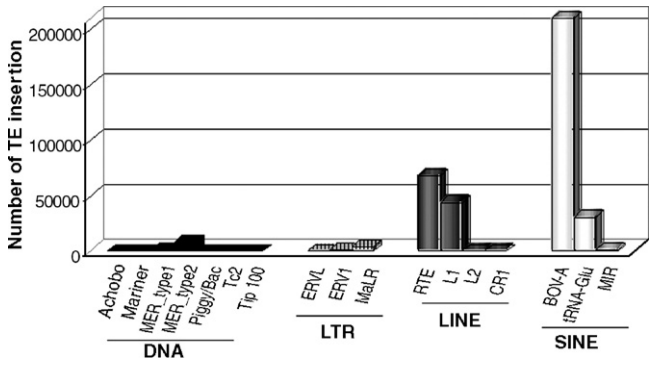

Fig. 2. Transposable elements diversity (number of different families of transposable elements) versus their occurrence in *Bos taurus* genes (number of insertion).

3.4. TE insertions in promoter segments, exons and introns

The availability of complete genome sequences is providing an unprecedented opportunity to assess the contribution of TEs to gene structure and function. The genomic approach has typically begun with the identification of TE-gene associations (i.e. the occurrence of TE sequences near or within genes) in a sequenced genome (Maside et al., 2003; Petrov et al., 2003; Franchini et al., 2004; DeBarry et al., 2005; Ganko, 2006). In order to identify this kind of association, bovine nuclear genes were screened. The occurrence of TEs was evaluated in exons, introns and promoter segments individually.

From all genes analyzed, 19,577 (85.84%) showed TE-derived sequences in introns. A similar percentage of TE insertions were detected in human genes, where 75% of genes contain at least one L1 sequences usually as part of introns (Han et al., 2004). A recent survey that analyzed 846 functionally characterized *cis*-regulatory elements from 288 genes showed that 21 of those elements (2.5%) from 13 genes (4.5%) reside in TE-derived sequences (Jordan et al., 2003). The same study showed that 24% of TE-derived sequences are present in promoter regions; defined as 500 bp located 5′ of functionally characterized transcription initiation sites. This significant number of insertions in the 5′ promoter region suggests that TE can play a role in regulation of gene expression. Finally, our analysis showed that 542 (2.37%) cow genes contain TE-derived sequences within exons. This number is smaller than that of Nekrutenko and Li (2001) who estimated that 4% of human protein-coding genes contain retrotransposons sequences, but shows that TEs are

contained within bovine genes in a significant proportion which are consistent with earlier studies of *H. sapiens* (Nekrutenko and Li, 2001; Jordan et al., 2003; Han et al., 2004), *M. musculus* (DeBarry et al., 2005), *Drosophila* (Ganko, 2006) and *C. elegans* (Ganko et al., 2003) genomes.

Insertion frequencies in exons and promoter segment were very low, 630 (0.12%) and 3349 (0.89%) respectively, while insertions in introns were strikingly more abundant 371,032 (98.94%). The low rate of TE-derived sequences in exons were observed in other mammals, in mouse, for instance, from the 186,823 exons analyzed, 263 (0.14%) showed LTR insertions (DeBarry et al., 2005). Additionally, a wide study has shown that 0.38% (751 of 196,937) of vertebrate proteins have TE sequences in their structure (Lorenc and Makalowski, 2003). Our data are consistent with the above-cited reports, which show a smaller frequency of TE insertions in human exons than Nekrutenko and Li (2001). In this study, 630 insertions were identified in exons, including 44 LTRs cassettes, 64 DNA cassettes, 157 LINEs cassettes and 365 SINEs cassettes. Since the TE maintenance is tolerated if the effect of its insertion is neutral or beneficial to gene host function, a hypothesis to explain the maintenance of relatively high number of TEs associated with exons is the production of one or more novel alternative transcripts while the native transcript maintains the original gene function (DeBarry et al., 2005). Over evolutionary time, novel transcripts containing TEs may evolve to encode a beneficial function and thus can be selectively maintained with or in place of the original transcript. However, the mere identification of a TE in exons is not, in itself, indicative of adaptive significance because it may only represent an insertional mutant unique to the sequenced strain. However, the strategy of searching for TE-cassettes within genes allows selecting the most relevant TE-gene associations to be investigated in further studies.

In addition to the evolutionary consequences of the exon insertions, TE-cassettes in promoter segments can also modify the normal transcription of gene. According to Van de Lagemaat et al. (2003), TEs in promoter regions affect the expression of many genes through the donation of transcriptional regulatory signals. Several studies have demonstrated a role for TEs in human gene transcription in individual cases (Murane and Morales, 1995; Brosius, 1999; Hamdi et al., 2000; Van de Lagemaat et al., 2003). From the 3349 TE-derived sequences found in promoter segments in the cow genome, 78 were DNA elements; 112 LTR elements, 1039 LINEs and 2120 SINEs.

When inserted into introns, TEs can influence gene expression by providing alternative splicing sites (Varagona et al., 1992; Davis et al., 1998) and polyadenylation sites (Sugiura et al., 1992; Mager et al., 1999). From 371,932 TE sequences found in introns, 5087 were LTR elements; 10,539 DNA; 111,099 LINEs and 244,307 SINEs.

3.5. Insertion lengths in exons, introns and promoter segment

Since exons and promoter are under stronger selection pressures than intron regions, it could be expected that TE-derived sequences in exons are smaller than those found in other gene locations. Statistical comparisons among TE lengths in each region were performed using the Kruskal–Wallis test. The median (minimum and maximum) value of each TE class is shown in Table 1. This result shows that there is a significant variation of length between TE classes and between gene regions. The lengths were statistically different between the exons, introns and promoter regions of DNA transposons, LINEs and SINEs, which is in agreement with the expected higher functional constraints in exons. No difference was detected in LTR class. In all classes, the largest TE insertions were found in intronic regions (DNA=2680 bp; LTR=660 bp; LINE=4722 bp and SINE=301 bp).

3.6. Full-exon origins from TE-cassettes

The contribution of TEs to coding regions is of particular interest, because they can influence the phenotype by changing protein sequences. There are several indications that transposable elements can be recruited for normal function in host organism (Miller et al., 1999a,b; Donnelly et al., 1999; Pardue and DeBaryshe, 2003; Britten, 2004; Mallet, 2004; Brandt et al., 2005). A classical example of retrotransposons playing a functional role for the host organism is the *HetA* and *TART* elements that have telomerase activity in *Drosophila* (Pardue and DeBaryshe, 2003). Britten (2004) described several examples of functional genes whose sequences have been almost completely derived from mobile elements in human genome. We identified 10 genes in the bovine genome that had an entire exon similar to a TE fragment. Therefore, the presence of TE-cassettes in transcripts does not guarantee their translation considering that eukaryotic organisms contain several post-transcriptional mechanisms that can eliminate the TE sequence before translation (Gotea and Makalowski, 2006). We evaluated the presence of these exon associations in coding sequences. As result, it was observed that the TE-exon associations of the LOC521822, LOC614028, LOC617472 and LOC505293 genes are not translated into protein. However, six genes with TE-exon associations are translated into proteins. These genes were LOC538046 (similar to zinc finger protein 452), LOC616159 (similar to

Table 1
Length of TE insertions (median, minimum and maximum) in each gene region

Median length/gene	DNA (minimum–maximum)	LTR (minimum–maximum)	LINE (minimum–maximum)	SINE (minimum–maximum)
Exons	199 (106–484)	296 (104–606)	212 (111–714)	175.50 (104–285)
Promoter segments	200 (118–1225)	246 (112–910)	392 (101–913)	196.00 (101–399)
Introns	255 (101–2680)	236 (101–1733)	320.5 (100–4722)	185.50 (114–301)
P	0.018 *	0.076	0.004 *	0.000 *

* $P<0.05$.

prefoldin 4), LOC514883 (similar to olfactory receptor Olr1169), LOC617220 (similar to zinc finger protein 193), LOC510331 (similar to Cullin-5, Vasopressin-activated calcium-mobilizing receptor), LOC615117 (similar to zinc finger protein 496).

Two possible mechanisms exist for insertion of transposable elements into an ORF: directly by transposition into the exon or indirectly by recruiting an intronic TE (Lorenc and Makalowski, 2003). The fact that all TE-cassettes detected as full-exons begin in intron regions and extend to the exons is in agreement with the hypothesis of indirect recruitment of an intronic TE insertion (Nekrutenko and Li, 2001). Curiously, all events detected in this study are associated with the final exon (Fig. 3). The biggest full-exon composed of TE-cassette was detected in LOC538046 gene (similar to zinc finger protein 452). The fourth exon of this gene (start point: 17,006 bp, end point: 18,919 bp) shares 89.3% identity with *Charlie 10* DNA transposon (Fig. 3A). There are other proteins with *Charlie* insertion described in the human genome such as GTF2IRD2, LOC58486, LOC285550 and DkFZp727G1 (Smit, 1999; Britten, 2004). Another transposon-cassette detected in this study as a full-exon gene (LOC615117: similar to zinc finger protein 496) was homologous to *Tigger 1*. The last exon (start point: 1506 bp, end point: 1614 bp) shares 81.25% identity with *Tigger 1* DNA transposon (Fig. 3B). There is a growing body of evidence that DNA transposons have been frequent source of protein domains for the assembly of new genes during evolution (Nouaud and Anxolabéhère, 1997; Sarkar et al., 2003; Hammer et al., 2005). Retrotransposon insertions were also identified (Fig. 3C to F), for example 182 bp segment of the element CHR2/SINE was detected in the last intron and complete exon of LOC510331 gene (similar to Cullin-5), 483 bp segment of L1MC/LINE was identified in the same region of LOC616159 gene (similar to prefoldin 4), also 307 bp segment of L1BT/LINE was detected in LOC514883 gene (similar to olfactory receptor Olr1169) and finally 185 bp

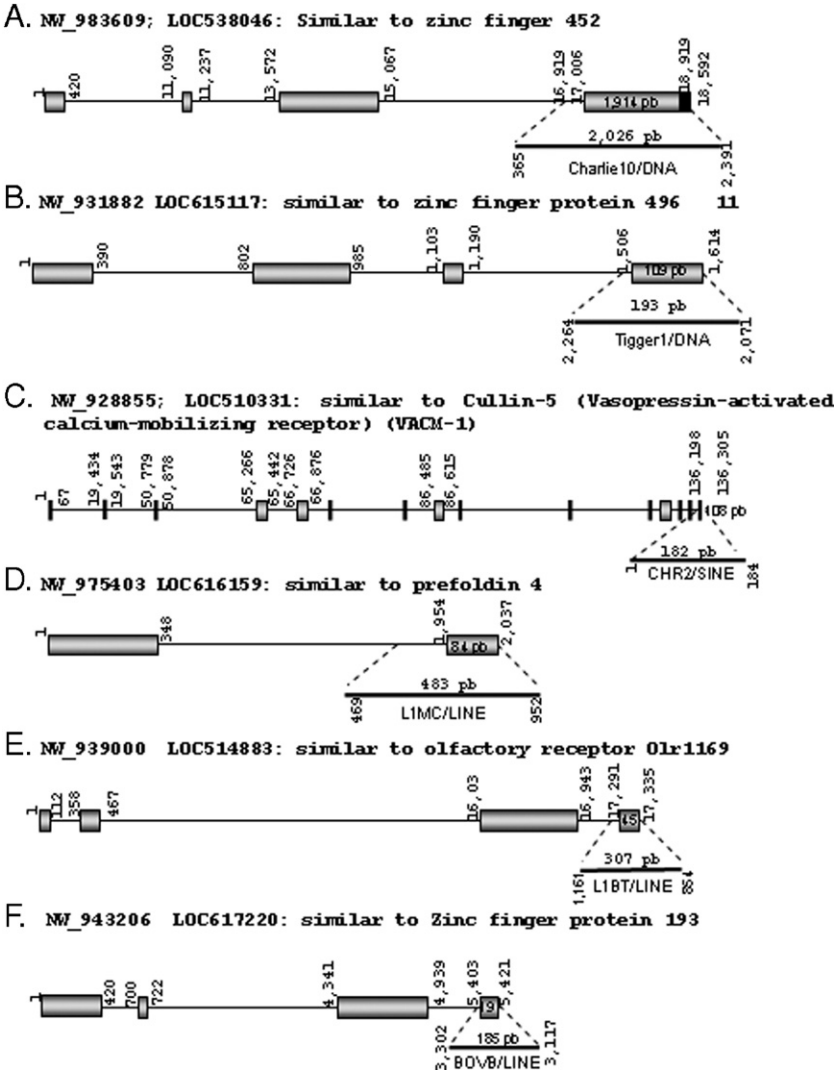

Fig. 3. Schematic representation of genes with exon sequences similar to TEs. (A) Comparison between similar zinc finger 452 gene structure and *Charlie 10* DNA transposon. (B) Comparison between similar zinc finger 496 gene structure and *Tigger 1* DNA transposon. (C) Comparison between similar to Cullin-5 gene structure and *CHR2* retrotransposon. (D) Comparison between similar to prefoldin gene structure and *L1MC* retrotransposons. (E) Comparison between similar to olfactory receptor Olr1169 gene structure and *L1BT* retrotransposons. (F) Comparison between zinc finger 193 gene structure and *BOVB* retrotransposons. Squares represent exons regions and lines introns. Numbers show the beginning and the end of exons and the length of TE insertion.

segment of BOVB/LINE was identified in LOC617220 (similar to zinc finger protein 193 (PRD51).

Three of these probable exaptation events are associated with the zinc finger family of proteins, which is one of the largest protein families in human genome. This protein functions diversely in regulation of transcription playing important roles in various cellular functions including cell proliferation, differentiation and apoptosis (Luo et al., 2006). Particularly interesting for us, the protein Zinc finger 452 has 1369 amino acids and a molecular weight of 156.5 kDa. It contains one transcriptional regulator SCAN domain, one integrase, catalytic region domain, one hAT dimerisation domain. This dimerisation domain is 50 amino acids located at the C terminus of the transposases of elements belonging to the Activator superfamily (*hobo* of *D. melanogaster*; *Ac* of maize and *Tam3* elements of *A. majus*) (Rubin et al., 2001).

According to Ohno (1970) and Gotea and Makalowski (2006) the molecular domestication or exaptation events should occur more frequently in duplicated genes because new duplicated copies are free of functional constraint and can undergo significant changes until they acquire new specific functions. In addition, Gotea and Makalowski (2006) suggest that the exaptation scenario should be favored when support from phylogeny exists, because the probability of having both random sequence similarity (TE versus protein) and phylogenetics support for the same protein fragment is lower than a random match between the protein and TE sequence alone. Based on this information we evaluated the possibility of exaptation in the Zinc finger protein 452 gene using phylogenetic relationship and sequence similarity. As result, we present three arguments supporting the hypothesis of exaptation. The first is the high similarity between *Charlie 10* transposon element and Zinc finger 452 amino acid sequences of *H. sapiens* (0.78), *P. troglodytes* (0.78), *B. taurus* (0.67) and *C. familiaris* (0.71). Fig. 4A shows the alignment between last exon of Zinc finger 452 from *B. taurus* (XP_618238),

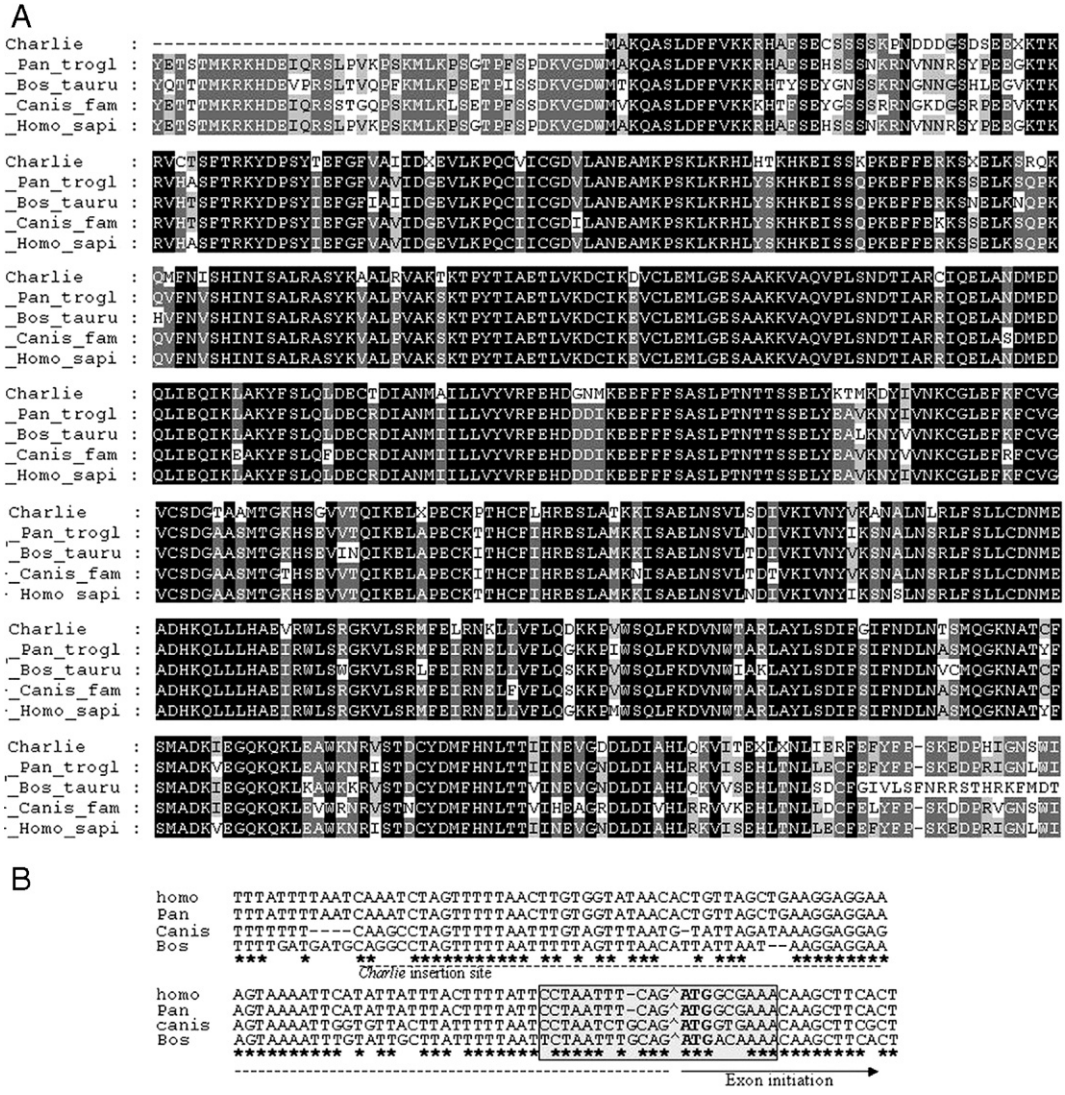

Fig. 4. A. Alignment between last exon of zinc finger 452 from *Bos_tauru* (*Bos taurus*: XP_618238), *Canis_fam* (*Canis familiaris*: XP_54541), *Homo_sapi* (*Homo sapiens*: NP_443155) and *Pan_trogl* (*Pan troglodytes*: XP_527300) and the *Charlie 10* DNA transposon. B. Multiple alignment of acceptor splicing site of last intron of zinc finger 452 is showed as square. The last intron and exon of zinc finger 452 gene of *Homo sapiens* (NC_000006.10), *Pan troglodytes* (NW_120489.1), *Canis familiaris* (NC_006617) and *Bos taurus* (NW_983609). Sites predicted by NetGene2 program.

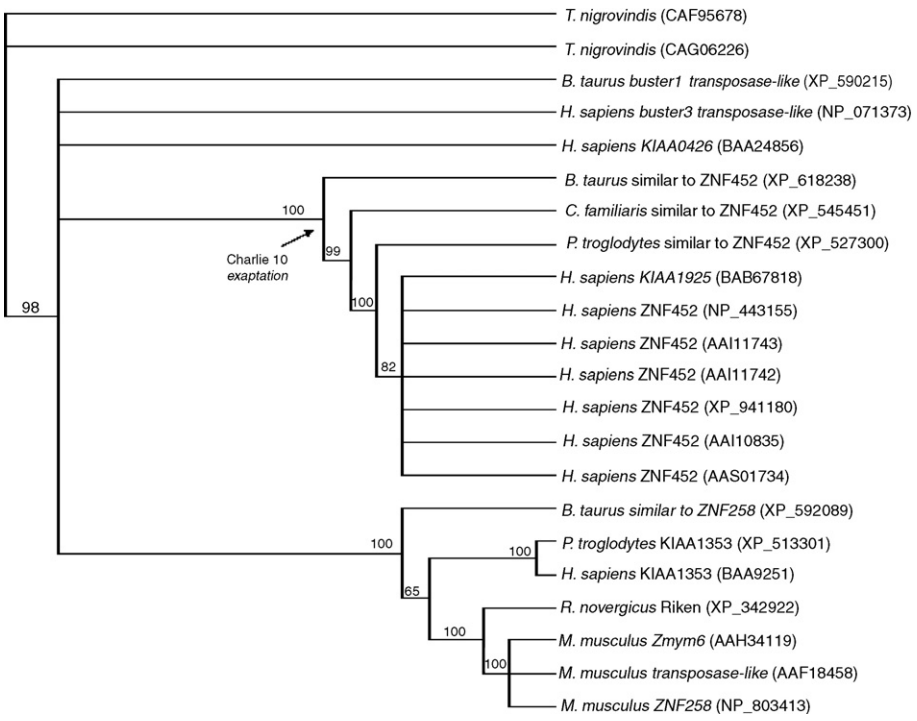

Fig. 5. Phylogenetic relationships between zinc finger 452 protein and other homologous proteins. The arrow represents the introduction of *Charlie 10* transposons. The cladogram was generated by parsimony analysis using the branch and bound algorithm (Swofford, 2000). The numbers indicated the branch support calculated by bootstrap analysis consisting of 1000 replicates.

C. familiaris (XP_54541), *H. sapiens* (NP_443155) and *P. troglodytes* (XP_527300) and the *Charlie 10* DNA transposon. In addition to such a high level of similarity, the sequence analysis suggests that insertion of *Charlie 10* near a zinc finger ancestral gene generated a change in the splicing site given rise to the current zinc finger 452 gene structures. Fig. 4B shows a motif matching the consensus sequence for acceptor splicing site (predicted by NETGENE with a confidence level of 46% in human).

The second is based on the fact that zinc finger is a large gene family that diversified by a series of duplication events during evolution. In such case new duplicated gene copy with a TE insertion, could pass through the sieve of selection evolve and acquire new function. Third, this protein in mammalian group (Fig. 5) show the phylogenetic relationships that *Charlie 10* insertion in the last functional exon of Zinc finger 452 protein is shared by *H. sapiens*, *P. troglodytes*, *B. taurus* and *C. familiaris* suggesting that this exaptation event occurred before the diversification of these mammals. The evidence of this event is shown by the highly supported clade that contain this sequences (bootstrap 100) pointed by an arrow. This finding suggests a new function of *Charlie 10* DNA in the mammalian genome and reinforces that TE fragments are utilized by genome and can contribute to host gene expression.

4. Conclusion

The origin of genetic novelties is of great interest in evolutionary and functional biology. Evidence from the human, mouse, worm and fruit fly genomes have indicated that, in addition to providing the source of genetic variability, transposable elements can also provide template DNA for novel genes or regulatory sequences. In this report we identified TE-gene associations in bovine genomes and showed that a great number of genes harbor insertions of transposable elements of different lengths and a small fraction of them are inserted into the coding sequences. The presence of TEs in this small fraction is of great interest because they can change the function of the gene product. When this change is adaptive and conserved over the evolutionary time it is named of molecular domestication (Miller et al., 1997), exaptation (Brandt et al., 2005) or co-opted events (Sarkar et al., 2003). In this study six potential exaptation events were identified within the bovine genome. From these findings, new questions arise, such as how the insertions of TEs into a gene can affect the host protein properties and/or function. Can any TE insertions be associated with different patterns of *B. taurus* gene expression? At the moment there are no answers for such questions, but there is no doubt that TE fragments modify genomes and can contribute to host gene expression. Further studies, including molecular and biochemical analyses are required to fully understand this phenomenon.

Acknowledgment

This work was supported by FAPESP (Grant 04/10148-6 to MEJA; post-doctoral fellowship to LMA 004/00905-4).

References

Agrawal, A., Eastman, Q.M., Schatz, D.G., 1998. Implications of transposition mediated by V(D)J-recombination proteins RAG1 and RAG2 for origins of antigen-specific immunity. Nature 394, 744–751.

Bacci Jr., M., et al., 2005. Identification and frequency of transposable elements in *Eucalyptus*. Genet. Mol. Biol. 28, 634–639.

Barendse, W., et al., 1994. A genetic linkage map of the bovine genome. Nat. Genet. 6, 227–235.

Berg, D.E., Howe, M.M., 1989. Mobile DNA. American Society for Microbiology, Washington DC. pp. xii, 972.

Brandt, J., et al., 2005. Transposable elements as a source innovation: expression and evolution of a family of retrotransposon-derived neogenes in mammals. Gene 345, 101–111.

Briten, R.J., 1996a. DNA sequence insertion and evolutionary variation in gene regulation. Proc. Natl. Acad. Sci. U. S. A. 93, 9374–9377.

Britten, R.J., 1996b. Cases of ancient mobile element DNA insertions that now affect gene regulation. Mol. Phylogenet. Evol. 5, 13–17.

Britten, R.J., 1997. Mobile elements inserted in the distant past have taken on important functions. Gene 205, 177–182.

Britten, R.J., 2004. Coding sequences of functioning human genes derived entirely from mobile element sequences. PNAS 101, 16825–16830.

Brosius, J., 1999. Genomes were forged by massive bombardments with retroelements and retrosequences. Genetica 107, 209–238.

Capy, P., Bazin, C., Higuet, D., Langin, T., 1998. Dynamics and Evolution of transposable elements. 1° France: Landes Bioscience, Springer, Heidelberg. 197p.

Cherry, J.M., Ball, C., Weng, S., 1997. Genetic and physical maps of *Saccharomyces cerevisiae*. Nature 387, 67–73.

Craig, N.L., 2002. Mobile DNA II. ASM Press, Washington DC. pp.xviii, 1204.

Davis, M.B., Dietz, J., Standiford, D.M., Emerson Jr., C.P., 1998. Transposable element insertions respecify alternative exon splicing in three *Drosophila* myosin heavy chain mutants. Genetics 150, 1105–1114.

DeBarry, J.D., Ganko, E., McDonald, J.F., 2005. The contribution of LTR retrotransposon sequences to gene evolution in *Mus musculus*. Mol. Biol. Evol. 23, 479–481.

Donnelly, S.R., Hawkins, T.E., Moss, S., 1999. A conserved nuclear element with a role in mammalian gene regulation. Hum. Mol. Genet. 1723–1728.

Dunn, C.A., Medstrand, P., Mager, D.L., 2003. An endogenous retroviral long terminal repeat is the dominant promoter for human beta1,3-galactosyltransferase 5 in the colon. Proc. Natl. Acad. Sci. U. S. A. 100, 12841–12846.

Dunn, C.A., van Lagemaat, L.N., Baillie, G.J., Mager, D.L., 2005. Endogenous retroviruses long terminal repeats as ready-to-use mobile promoters: the case of primate β3GAL-T5. Gene 364, 2–12.

Dunn, C.A., Romanish, M., Gutierrez, L.E., van Lagemaat, L.N., Mager, D.L., 2006. Transcription of two genes from a bidirectional endogenous retrovirus promoter. Gene 366, 335–342.

Flavell, R.B., 1986. Repetitive DNA and chromosome evolution in plants. PhIlos. Trans. R. Soc. Lond., B Biol. Sci. 312, 227–242.

Franchini, L.F., Ganko, E.W., McDonald, J.F., 2004. Retrotransposon-gene associations are widespread among *D. melanogaster* populations. Mol. Biol. Evol. 21, 1323–1331.

Ganko, E.W., 2006. LTR retrotransposon-gene associations in *Drosophila melanogaster*. J. Mol. Evol. 62, 111–120.

Ganko, E.W., Bhattacharjee, V., Schliekelman, P., McDonald, J.F., 2003. Evidence for the contribution of LTR retrotransposon to *C. elegans* gene evolution. Mol. Biol. Evol. 20, 1925–1931.

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

Hamdi, H.K., Nishio, H., Tavis, J., Zielinski, R., Dugaiczyk, A., 2000. Alu-mediated phylogenetic novelties in gene regulation and development. J. Mol. Biol. 299, 931–939.

Hammer, S.E., Strehe, S., Hagemann, S., 2005. Homologs of *Drosophila P* transposon were mobile in *Zebrafish* but have been domesticated in common ancestor of chicken and human. Mol. Biol. Evol. 22, 833–844.

Han, J.S., Boeke, J.D., 2004. A highly active synthetic mammalian retrotransposons. Nature 429, 314–318.

Han, J.S., Boeke, J.D., 2005. LINE-1 retrotransposons: modulators of quantity and quality of mammalian gene expression? BioEssays 27, 775–784.

Han, J.S., Szak, S.T., Boeke, J.D., 2004. Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. Nature 429, 268–274.

Hua-Van, A., Rouzic, A., Maisonhaute, C., Capy, C., 2005. Abundance, distribution and dynamics of retrotransposable elements and transposons: similarities and differences. Cytogenet. Genome Res. 110, 426–440.

Jordan, I.K., Rogozin, I.B., Glazko, G.V., Koonin, E.V., 2003. Origin of a substantial fraction of human regulatory sequences from transposable elements. Trends Genet. 19, 68–72.

Kaminker, J.S., Bergman, C.M., Kronmiller, B., 2002. The transposable elements of the *Drosophila melanogaster* euchromatin: a genomics perspective. Genome Biol. 3 (RESERCH0084).

Kidwell, M.G., 2002. Transposable elements and the evolution of genome size in eukaryotes. Genetica 115, 49–63.

Kordis, D., Gubensek, F., 1999. Horizontal transfer of non-LTR retrotransposons in vertebrates. Genetica 107, 121–128.

Landry, J.R., Mager, D.L., 2003. Functional analysis of the endogenous retroviral promoter of the human endothelin B receptor gene. J. Virol. 77, 7459–7466.

Landry, J.R., Rouhi, A., Medstrand, P., Mager, D.L., 2001. Repetitive elements in the 5′untranslated region of a human zinc-finger modulate transcription and translation efficiency. Genomics 76, 110–116.

Landry, J.R., Rouhi, A., Medstrand, P., Mager, D.L., 2002. The Opitz syndrome gene *M1D1* IS transcribed from a human endogenous retroviral promoter. Mol. Biol. Evol. 19, 1934–1942.

Larkin, D.M., et al., 2003. A cattle–human comparative map built with cattle BAC ends and human genome sequence. Genome Res. 13, 1966–1972.

Lenstra, J.A., Van Boxtel, J.A.F., Zwaagstra, K.A., Schwerin, M., 1993. Short interspersed nuclear element (SINE) sequences of the Bovidae. Anim. Genet. 24, 33–39.

Lev-Maor, G., Sorek, R., Shomron, N., Ast, G., 2003. The birth of an alternatively spliced exon: 3′ splice-site selection in Alu exons. Science 300, 1288–1291.

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

Luo, K., et al., 2006. Activation of transcriptional activities of AP1 and SRE by a novel zinc finger protein ZNF445. Gene 33, 51–57.

Mager, D.L., Hunter, D.G., Schertzer, M., Freeman, J.D., 1999. Endogenuous retroviruses provide the primary polyadenization signal for two human genes (HHLA2 and HHLA3). Genomics 59, 255–263.

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

Makalowski, W., Mitchell, G.A., Labuda, D., 1994. Alu sequences in the coding regions of mRNA: a source of protein variability. Trends Genet. 10, 188–193.

Malik, H.S., Eickbush, T.H., 1998. The RTE class of non-LTR retrotransposons is widely distributed in animals and is the origin of many SINEs. Mol. Biol. Evol. 15, 1123–1134.

Mallet, F., 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.

Maside, X., Bartolome, C., Charlesworth, B., 2003. Inferences on the evolutionary history of the S-element family of *Drosophila melanogaster*. Mol. Biol. Evol. 20, 1183–1187.

Medstrand, P., Landry, J.R., Mager, D.L., 2001. Long terminal repeats are used as alternative promoters for the endothelin B receptor and apolipoprotein C-I genes in humans. J. Biol. Chem. 276, 1896–1903.

Medstrand, P., van Lagemaat, L.N., Mager, D.L., 2002. Retroelement distribution in the human genome: variations associated with age and proximity to genes. Genome Res. 12, 1483–1495.

Miller, W.J., McDonald, J.F., Pinsker, W., 1997. Molecular domestication of mobile elements. Genetica 100, 261–270.

Miller, K., Lynch, C., Martin, J., Herniou, E., Tristem, M., 1999a. Identification of multiple Gypsy LTR-retrotransposon lineages in vertebrate genomes. J. Mol. Evol. 49, 358–366.

Miller, W.J., McDonald, J.F., Nouaud, D., Anxolabéhère, D., 1999b. Molecular domestication — more than a sporadic episode in evolution. Genetica 107, 197–207.

Murane, J.P., Morales, J.F., 1995. Use of mammalian interspersed repetitive (MIR) element in coding and processing sequences of mammalian genes. Nucleic Acids Res. 23, 2837–2839.

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

Nouaud, D., Anxolabéhère, D., 1997. *P* element domestication: a stationary truncated *P* element may encode a 66 kDa repressor-like protein in the *Drosophila mountium* species subgroup. Mol. Biol. Evol. 14, 1132–1144.

Ohno, S., 1970. Evolution by Gene Duplication. Springer-Verlag.

Okada, N., Hamada, M., 1997. The 3′ ends of tRNA-derived SINEs originated from 3′ end LINEs: a new example from the bovine genome. J. Mol. Evol. 44, 52–56.

Ostertag, E.M., Kazazian Jr., H.H., 2001. Biology of mammalian L1 retrotransposons. Annu. Rev. Genet. 35, 501–538.

Pardue, M.L., DeBaryshe, P.G., 2003. Retrotransposons provide an evolutionary robust non-telomerase mechanism to maintain telomeroes. Annu. Rev. Genet. 37, 485–511.

Petrov, D., Aminetzach, Y.T., Davis, J.C., Bensasson, D., Hirsh, A.E., 2003. Size matters: non-LTR retrotransposable elements and ectopic recombination in *Drosophila*. Mol. Biol. Evol. 20, 880–892.

Robertson, H.M., 1996. Members of *pogo* subfamily of DNA mediated transposon in the human genome. Mol. Gen. Genet. 252, 761–766.

Rubin, E., Lithwick, G., Levy, A.A., 2001. Structure and evolution of the hAT transposon superfamily. Genetics 158, 949–957.

SanMiguel, P., et al., 1996. Nested retrotransposons in the intergenic regions of the maize genome. Science 274, 765–768.

Sarkar, A., et al., 2003. Molecular evolutionary analysis of widespread *piggyback* transposon family and related domesticated sequences. Mol. Genet. Genomics 270, 173–180.

Silva, J.C., Loreto, E.L., Clark, J.B., 2004. Factors that affect the horizontal transfer of transposable elements. Curr. Issues Mol. Biol. 6, 57–71.

Singer, S.S., Mànnel, D.N., Hehlgans, T., Brosius, J., Schmitz, J., 2004. From "junk" to gene: curriculum vitae of a primate receptor isoform gene. J. Mol. Biol. 341, 883–886.

Smit, A.F.A., 1996. The origin of interspersed repeats in the human genome. Curr. Opin. Genet. Dev. 6, 743–748.

Smit, A.F.A., 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr. Opin. Genet. Dev. 9, 657–663.

Smit, A.F.A., Riggs, A.D., 1996. Tiggers and DNA transposons fossils in the human genome. PNAS 93, 1443–1448.

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

Stajich, J.E., et al., 2002. The Bioperl toolkit: Perl modules for the life sciences. Genome Res. 12, 1611–1618.

Sugiura, N., Hagiwara, H., Hirose, S., 1992. Molecular cloning of porcine soluble angiotensin-binding protein. J. Biol. Chem. 267, 18067–18072.

Swofford, D., 2000. PAUP: Phylogenetic Analysis Using Parsimony (and other Methods), Version 4.0.b10. Sinauer, Sunderland, Massachusetts.

The Human Genome Sequencing Consortium, 2001. Initial sequence analysis of human genome. Nature 409, 860–921.

Thompson, J.D., Higgins, D.G., Gibson, T.J., 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.

Thornburg, B.G., Gotea, V., Makalowski, W., 2005. Transposable elements as a significant source of transcription regulating signals. Gene 365, 104–110.

Van de Lagemaat, L.N., Landry, J.R., Mager, D.L., Medstrand, P., 2003. Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. Trends Genet. 19, 530–536.

Varagona, M.J., Purugganan, M., Wessler, S.R., 1992. Alternative splicing induced by insertion of retrotransposon into the maize waxy gene. Plant Cell 4, 811–820.

Wilson, H.L., et al., 2005. Molecular analyses of disease phatogenesis: application of bovine microarrays. Vet. Immunol. Immunopathol. 105, 277–287.