Mol. Cells *30*, 77-87, July 31, 2010
DOI/10.1007/s10059-010-0091-2

Communication

Molecules and Cells

Functional Impact of Transposable Elements Using Bioinformatic Analysis and a Comparative Genomic Approach

Dae-Soo Kim[1,4], Jae-Won Huh[1,4], Young-Hyun Kim[1,2,4], Sang-Je Park[1,3], and Kyu-Tae Chang[1,*]

A dual coding event, which is the translation of different isoforms from a single gene, is one of the special patterns among the alternative splicing events. This is an important mechanism for the regulation of protein diversity in human and mouse genomes. Although the regulation for dual coding events has been characterized in a few genes, the individual mechanism remains unclear. Numerous studies have described the exonization of transposable elements, which is the splicing mediated insertion of transposable element sequence fragments into mature mRNAs. Therefore, in this study, we investigated the number of transposable element (TE)-derived dual coding genes in human, chimpanzee and mouse genomes. TE fusion exons appeared in the dual coding regions of 309 human genes. Functional protein domain alterations by TE-derived dual coding events were observed in 129 human genes. Comparative TE-derived dual coding events were also analyzed in chimpanzee and mouse orthologs. Seventy chimpanzee orthologs had TE-derived dual coding events, but mouse orthologs did not have any TE-derived dual coding events. Taken together, our analyses listed the number of TE-derived dual coding genes which could be investigated by experimental analysis and suggested that TE-derived dual coding events were major sources for the functional diversity of human genes, but not mouse genes.

INTRODUCTION

Alternative splicing has been considered an important mechanism that contributes to the functional diversification of a single gene (Lopez, 1998). It can change how a gene acts in different tissues and developmental stages by generating distinct mRNA variations composed of different selections of exons (Modrek et al., 2002). The completion of large genomic sequencing projects reveals that various organisms abundantly use alternative splicing. Indeed, alternative splicing affects a majority of genes, and plays an extremely important role in expanding protein diversity (Licatalosi et al., 2010). Alternative splicing events may therefore explain the apparent discrepancy between gene number and organism complexity (Szathmary et al., 2001).

Dual coding events are one of the specialized cases of alternative splicing. Because alternative splicing can concei-vably give birth to some dual coding regions in which the same exon sequence can result in different transcripts which code for amino acids in different reading frames. Recently, several specific examples of dual-coding transcripts have been reported in human and mouse genomes (Liang et al., 2005, Scorilas et al., 2001; Zhao et al., 2004), even though many such candidate genes have been predicted by bioinformatics analysis (Heng et al., 2010). Until now, there has been no large scale study of dual coding regions in alternative splicing genes. Therefore, they provide a special opportunity to trace the origin of unique splicing patterns and to understand the evolutionary constraints on coding sequences in open reading frames.

Recently, it has been reported that alternative splicing mediated the insertion of parts of transposable element sequences into mature mRNAs (Sorek et al., 2002). This is possible because sequences of transposable elements (TEs) contain motifs that resemble consensus splice sites (Sorek et al., 2002). Also, a recent study reported that TEs are found in the protein coding region of approximately 4% of human genes, and that Alu elements account for about one third of these insertions (Nekrutenko et al., 2001). The vast majority of TE-sequence insertions into mature mRNAs are splicing mediated (Makalowski et al., 1994; Nekrutenko and Li, 2001). This is possible because both strands of TE-sequences contain motifs that resemble consensus splice sites (Makalowski et al., 1994). Mutations within intronic TE-sequences may yield active splice sites, that is, a part of the intronic TE-sequence will be exonized. Furthermore, species-specific alternative splicing often is due to the presence of an Alu insertion in a human gene versus, for example, its mouse homolog (Sorek et al., 2002).

The insertion of TEs into mature mRNA may cause a genetic disease or contribute to protein variability in the genome (Ahn et al., 2009; Christensen, 2005; Halling et al., 1999; Makalowski et al., 1994). As the sequences are mostly intronic, TEs have been assumed to be spliced out and not affecting target gene

[1]National Primate Research Center, Korea Research Institute of Bioscience and Biotechnology, Ochang 363-883, Korea, [2]Functional Genomics, University of Science and Technology, Daejeon 305-333, Korea, [3]Division of Biological Sciences, College of Natural Sciences, Pusan National University, Busan 609-735, Korea, [4]These authors contributed equally to this work.
*Correspondence: changkt@kribb.re.kr

Received January 11, 2010; revised March 19, 2010; accepted March 25, 2010; published online July 14,2010

Keywords: bioinformatic analysis, dual coding gene, TE-derived dual coding genes, transposable element

expression. However, recent studies have been shown that TEs could affect protein sequences, splicing patterns, and expression of human genes (Sin et al., 2006). In most cases, TEs are deleterious to the host as a consequence of their insertion into coding regions that lead to alterations of gene expression (Makalowski, 2000). But some TEs have more favorable effects on the expression of host genes (Gotea et al., 2006). For example, PTPN1 is a 435 amino acid protein that belongs to the large family of protein tyrosine phosphatases (PTPs), which catalyze protein dephosphorylation. The L3 fragment found in the PTPN1 mRNA corresponds to part of this RT domain (Andersen et al., 2004; Gotea et al., 2006). The TEs-fusion event is one of the most important evolutionary mechanisms for the creation of new functions (Yi et al., 2006). Therefore, TE insertions might be one important reason for the high frequency of alternative splicing in human protein-coding genes (Nekrutenko et al., 2001). In addition, the fact that the majority of TEs found in protein-coding regions are fragmented also indicates that the exon recruitment mechanism is the most common (Yi et al., 2006).

In the present study, we systematically identified TEs within dual coding genes from human sequences and discovered some unique features. Also, evolutionary pathways of the TEs were investigated using chimpanzee and mouse genomes and bioinformatics tools.

MATERIALS AND METHODS

Data source

The RefSeq Database reflects our current knowledge of genome biology and is generally accepted as a standard for genome annotation (Pruitt et al., 2005). It also harbors all valid transcripts and proteins which could be used for various genomic analyses. Therefore, we used high quality and fully annotated mRNAs from the NCBI Reference Sequence (RefSeq) Database in our analysis. The RefSeq mRNA (human; Build 35.1, chimpanzee; Build 1.1, mouse; Build 35.1) and genome sequences were downloaded from the NCBI database (Pruitt et al., 2005). These sequences include the coding region as well as the 5′ and 3′ untranslated regions of each gene. The HomoloGene database (http://www.ncbi.nlm.nih.gov/homologene) was downloaded from the National Center for Biotechnology Information. NCBI accessions for human-mouse and human-chimpanzee orthologous pairs that were described as either reciprocal best hits or manually curated were extracted from the databases. The HomoloGene database is based on similarity searches of genes.

Identification of dual coding genes

We used the publicly available human, chimpanzee, and mouse genome sequences from the NCBI database. Dual coding genes were collected from the database described above by selecting the gene sequence that contained two amino acid sequences. A dual coding gene contains two reading frames read in the same gene. The Entrez gene database was used to identify the genomic location of reference sequences in RefSeq. Genes were located to the genomic contig according to the GenBank database, to ensure that muliti-copy genes and pseudogenes would not interfere during the SIM4 (Florea et al., 1998) alignment process. Only those that aligned precisely, without a gap or overlap, were retained. Redundant entries (with the same protein sequence) were removed from the data set. If a gene had more than one RefSeq sequence, their corresponding genomic regions were required to not overlap. Various intron/exon structures of single genes were obtained by alignment of RefSeq mRNAs. Exon-intron structures of genes containing TE sequences were checked using the SIM4 program for spliced alignment (Florea et al., 1998). An intron was defined as a vertex containing only the genomic sequence, and a true intron as an intron abiding by the GT/AG, GC/AG, or AT/AC rules. Finally, repeat elements integrated in dual coding gene transcripts and the genomic sequences were detected using the RepeatMasker program (http://www.repeatmasker.org/). Descriptions of specific repetitive elements were obtained from Repbase Update (Jurka, 2000).

To identify orthologous genes between human, chimpanzee and mouse, pairs of genes reported as reciprocal best matches in the HomoloGene database and were mapped successfully onto the genomic sequence were used. For the comparative analysis of the dual coding effect of TEs, orthologous sequences were extracted from the Entrez gene database using GenBank information. To reduce noise, only well defined gene data were included in the analysis. TEs were then screened by RepeatMasker (http://repeatmasker.org/) using consensus sequences of various repeat elements from Repbase Update. Next, further analysis of the TE fusion gene was conducted to check their influence on the transcripts they were inserted into. Protein domains associated with TE elements were identified in version 21.0 of the SMART and Pfam database. Sequence analysis of all transposable elements inserted in human coding regions was performed by the InterProScan program (Mulder et al., 2002; Quevillon et al., 2005). Protein domains were identified by characteristic signatures in SMART (Schultz et al., 1998) and Pfam (Bateman et al., 2004).

RESULTS AND DISCUSSION

Identification and analysis of TE-derived dual coding genes

Before describing our analyses, we would like to define the terms used in this paper. A dual coding gene is a gene which makes different isoforms (different coding regions) in a single gene. By using different exon combinations, some genes can encode protein in multiple reading frames in different transcripts. Here, we performed a systematic search through a set of high quality human, chimpanzee, and mouse transcripts that encode amino acids in more than one reading frame due to alternative splicing. The genome locus for RefSeq transcript variation forms of each gene was identified according to the annotation in the NCBI database. We searched 28171 genes listed in the NCBI database for only two RefSeq transcript variation forms in each gene and analyzed the structural patterns and expression of TE elements located in the gene, producing 13,584 genes that had their mRNA sequences affected by alternative splicing events. In our data set, 18,082 transcripts were included, representing 13,584 distinct genes, with redundant entries removed from the data set. Finally, we identified 2585 dual coding genes, from which we analyzed TEs in the dual coding regions.

Human RefSeq mRNA sequences were aligned with the genome using the SIM4 program (Florea et al., 1998). Only alignments with greater than 97% sequence identity were used in subsequent stages. The position information of the exons and genome sequences were extracted to be matched. Next, RepeatMasker (http://repeatmasker.org/) was used to search for TE sequences in the genome. Based on this information, the location of the TEs and exons were calculated. Alignments having >97% sequence identity and TEs with a minimum exonization length of 50 bp were used in this study. Dual coded RefSeq mRNA sequences were extracted in relation with TEs, and 309 TE-derived dual coding genes were found (486 RefSeq mRNA sequences in total). This result supports the obser-

Table 1. Distribution of transposable elements within human dual coding genes

Transposable elements		TEs within dual coding region	Percent (%)	Gene number	RefSeq mRNA
Family	Subfamily				
SINE	AluJ	28	5.07	144 (35.87 %)	189
	AluSc	52	9.42		
	AluY	6	1.09		
	FAM	2	0.36		
	FLAM	10	1.81		
	FRAM	1	0.18		
	MIR	93	16.85		
	THER1	6	1.09		
LINE	HAL1	2	0.36	129 (42.6%)	226
	L1M	49	8.88		
	L1P	8	1.45		
	L2	73	13.22		
	L3	110	19.93		
	L4	13	2.36		
LTR	LTR/ERVK	7	1.27	45 (14.5%)	715
	LTR/ERV1	29	5.25		
	LTR/MaLR	13	2.36		
	LTR/ERVL	31	5.62		
DNA	Arthur1	1	0.18	18 (3.43%)	19
	Charlie8	4	0.72		
	MER	12	2.17		
	Tigger	2	0.36		

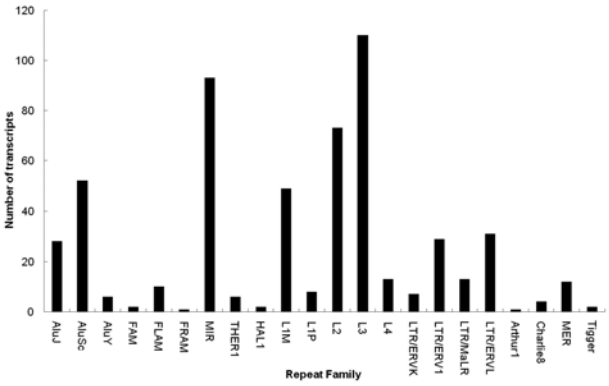

Fig. 1. Distribution of transposable elements detected in human dual coding regions

vation that TE exonizations in human protein coding regions might be a key reason behind the high frequency of alternative splicing in human protein coding genes.

A total of 309 of 486 TE fusion transcripts were predicted to be TE-derived dual coding genes (Table 1). TE-derived dual coding regions were analyzed to investigate the classification of TEs fusion events in the coding regions (CDS) of genes. Many kinds of TEs were identified in the CDS region of functional genes. There were 226 (42.6%) SINE-fusion transcripts, 189 (36.87%) LINE-fusion transcripts, 71 (14.5%) LTR-fusion transcripts, and 19 (3.43%) DNA-fusion transcripts identified in the protein coding regions (Fig. 1). TE insertions in human protein coding regions are possible because many TEs carry potential splice sites. For example, the consensus sequence of Alu elements contains eight putative donor sites and three acceptor sites (Makalowski et al., 1994). In a previous study, Alu-containing internal exons appeared frequently within the CDS of human transcripts (Nekrutenko et al., 2001; Sorek et al., 2002).

It is commonly recognized that TEs have negative effects by causing insertion mutations in the protein coding region. Therefore, in most cases, the host genome must control the expression of TEs to protect cellular components from accumulating mutations in the TE coding regions. However, some TE families have been inserted and recruited into protein coding regions as novel exons during human evolution (Britten, 2006). Also in a previous study, an older subfamily of TEs showed a tendency for a higher retention ratio than a younger subfamily of TEs (Sorek et al., 2002). Generally, most TEs in the human genome are defective by the accumulation of mutations. The bias toward old subfamilies in the set of TEs fusion transcripts may reflect the necessity of accumulating many substitutions to create a functional splice site within the TEs sequence to allow for its exonization. Meaning, the original characters of TE needed to be changed from TE to non-TE so it may be recognized by the human genome as a cellular component for gene-like evolution. The mechanism of gene-like evolution as derived by the old transposable element could be a main source of human genome diversification using limited resources. TEs could have a strong impact on the evolution of the human genome by providing unlimited resources for human genetic diversity. Our results show that a small fraction of relatively young exonized TEs have the potential to contribute to protein functionality.

Functional analysis of TE fusion genes in the dual coding region

To investigate the functional and structural impact of TEs in dual coded proteins, InterPro was applied (Mulder et al., 2002; Quevillon et al., 2005). First, we searched protein domains in the TE-derived dual coding regions. In approximately half of these, the alterations could potentially affect the number of conserved domains. Functional domains of 121 genes were changed by the TEs insertion events in the protein coding regions. As shown in Table 2, TE fusion transcripts influenced a wide variety of protein motifs. In total, among 309 genes with TE cassette-exons into protein coding regions, alterations of the domains due to TEs were observed in 121 genes (39%; Table 2). Currently, only few functional analyses of TE-derived proteins have been conducted, and no comparative functional analysis of TE-derived dual coding genes has been conducted. However, our results showed that more than 100 dual coding genes were affected in protein motif regions by TEs (Fig. 2). Only 4.3 % of the exonized TE elements potentially contributed a new function to the proteome, consistent with previous reports (Gotea et al., 2006). A well known example is the new Alu exon of human gene RNA specific adenosine deaminase 2. This exon was derived from a primate specific Alu transposable element. It inserts a 40 amino acid peptide segment into the catalytic domain of ADAR2, altering the enzymatic activity of the protein product (Gerber et al., 1997). Our findings also suggest that the potential contribution of exons originated from TE elements to proteome complexity is low. Therefore, TEs in protein motif regions could modify the original gene functions by the production of dual coded proteins and functional analysis of TE-derived dual coded genes should be investigated.

Comparative analysis of TE-derived dual coding genes in human, chimpanzee, and mouse

We performed a comparative analysis of orthologous se-

Table 2. Functional analysis of coding regions modified by the insertion of transposable elements in the human genome

Gene	Repeat family	Database		InterPro ID	Discription
		Pfam	SMART		
ADAM15	MIR_Mars	PF08516	SM00608	IPR006586	ADAM, cysteine-rich
ADAM28	L1M4	PF01421		IPR001590	Peptidase M12B, ADAM/reprolysin
ADAR	L4	PF00035	SM00552	IPR001159	Double-stranded RNA binding
ADARB1	AluJb	PF02137	SM00552	IPR002466	Adenosine deaminase/editase
ADRA1A	AluSc	PF00001		IPR000276	Rhodopsin-like GPCR superfamily
ADSSL1	TAR1	PF00709		IPR001114	Adenylosuccinate synthetase
AES	MIRb	PF03920		IPR005617	Groucho/TLE, N-terminal Q-rich region
ALDH5A1	L1PB1	PF00171		IPR002086	Aldehyde dehydrogenase
APBB1	LTR50	PF00640	SM00462	IPR006020	Phosphotyrosine interaction region
APOBEC3F	MER3	PF08210		IPR013158	APOBEC-like, N-terminal
ARHGAP8	AluJo	SM00516		IPR001251	Cellular retinaldehyde-binding/triple function, C-terminal
ATP1A1	L3	PF00122		IPR008250	E1-E2 ATPase-associated region
ATRX	L2	PF001761	SM00490	IPR000330	SNF2-related
BAX	AluSx	PF00452	SM00337	IPR000712	Apoptosis regulator Bcl-2, BH
BIRC5	AluY	PF00653	SM00238	IPR001370	Proteinase inhibitor I32, inhibitor of apoptosis
C15orf21	LTR17	PF01101	SM00527	IPR000079	High mobility group protein HMG14 and HMG17
C20orf30	AluSg	PF05915		IPR008590	Protein of unknown function DUF872, eukaryotic
CACNA1G	MIRb	PF00520		IPR005821	Ion transport
CACNA1H	L3	PF00520		IPR005821	Ion transport
CADPS2	L3_Mars	PF00169	SM00233	IPR001849	Pleckstrin-like
CASP6	MIRm	PF00656	SM00115	IPR011600	Peptidase C14, caspase catalytic
CASP7	AluSx	PF00656		IPR011600	Peptidase C14, caspase catalytic
CD34	L3	PF06365		IPR008083	CD34 antigen
CD8B1	L4	PF07686	SM00409	IPR013106	Immunoglobulin V-set
CHEK2	FLAM_C	PF00069	SM00220	IPR000719	Protein kinase
CLEC12A	MER4E1	PF00059	SM00034	IPR001304	C-type lectin
CNN2	MIR	PF00307	SM00033	IPR001715	Calponin-like actin-binding
COMMD6	L1MA9	PF07258		IPR009886	HCaRG
DKFZP586H2123	MLT1A0	PF00084	SM00032	IPR000436	Sushi/SCR/CCP
DPM2	MIR	PF07297		IPR009914	Dolichol phosphate-mannose biosynthesis regulatory
DPM3	L1M1	PF08285		IPR013174	Dolichol-phosphate mannosyltransferase subunit 3
DYX1C1	LTR7	PF04969		IPR007052	CS
ETFB	MIR,L3	PF01012		IPR000049	Electron transfer flavoprotein beta-subunit
EVI2A	MER5A1	PF05399		IPR008608	Ectropic viral integration site 2A
EYA2	L2	PF00702		IPR005834	Haloacid dehalogenase-like hydrolase
FBLIM1	AluY	PF00412	SM00132	IPR001781	LIM, zinc-binding
FBXL5	L3	SM00367		IPR006553	Leucine-rich repeat, cysteine-containing subtype
FTSJ2	AluSq	PF01728		IPR002877	Ribosomal RNA methyltransferase RrmJ/FtsJ
FUSIP1	L3_Mars	PF00076	SM00360	IPR000504	RNA-binding region RNP-1 (RNA recognition motif)
GABRB2	L3	PF02931		IPR006202	Neurotransmitter-gated ion-channel ligand-binding
GCK	L2	PF03727		IPR001312	Hexokinase
GFRA1	L3_Mars	PF02351		IPR003438	Glial cell line-derived neurotrophic factor receptor
GPR110	L2	PF01825	SM00303	IPR000203	GPS
GRIK2	L2	PF00060	SM00079	IPR001320	Ionotropic glutamate receptor
HAP1	MIRb	PF04849		IPR006933	HAP1, N-terminal
HNF4A	MIR	SM00430		IPR000536	Nuclear hormone receptor, ligand-binding
IFT122	FLAM_C	SM00320		IPR001680	WD-40 repeat
IGSF3	L3	PF07686	SM00409	IPR013106	Immunoglobulin V-set
INADL	MSTD	PF00595	SM00228	IPR001478	PDZ/DHR/GLGF
ITGB1	AluSx	PF00362	SM00187	IPR002369	Integrin, beta chain N-terminal
KCNK2	L3	PF07885		IPR013099	Ion transport 2, bacterial

(continued)

Gene	Repeat family	Database		InterPro ID	Discription
		Pfam	SMART		
LOC285989	L1M5	SM00349		IPR001909	KRAB box
LRP8	MIRm	PF00057	SM00192	IPR002172	Low density lipoprotein-receptor, class A
MAD2L1BP	AluSp	PF06581		IPR009511	MAD2L1 binding protein
MAGI3	AluSc	PF00397		IPR001202	WW/Rsp5/WWP
MBD1	MLT2F	PF01429	SM00391	IPR001739	Methyl-CpG binding
MGC3207	L2	PF01008		IPR000649	Initiation factor 2B related
MKNK1	AluJo	PF00069	SM00220	IPR000719	Protein kinase
MMP16	LTR40a	PF01471		IPR002477	Peptidoglycan-binding domain 1
MTO1	AluSp	PF01134		IPR002218	Glucose-inhibited division protein A
NALP1	AluJb	PF05729		IPR007111	NACHT nucleoside triphosphatase
NARF	AluSx	PF02906		IPR004108	Hydrogenase large subunit, C-terminal
NMNAT2	L3	PF01467		IPR004820	Cytidylyltransferase
NRAP	L4	PF00880	SM00227	IPR000900	Nebulin
NTRK3	MIR3.L1MEe	PF01462	SM00013	IPR000372	Leucine-rich repeat, cysteine-rich flanking region, N-terminal
OGG1	MLT1K	PF00730	SM00478	IPR003265	HhH-GPD
OPN4	MIRb	PF00001		IPR000276	Rhodopsin-like GPCR superfamily
OSBPL8	L3_Mars	PF01237		IPR000648	Oxysterol-binding protein
OSCAR	MSR1	SM00409		IPR003599	Immunoglobulin subtype
PCBD1	MLT2B4	PF01329		IPR001533	Transcriptional coactivator/pterin dehydratase
PCDH11X	MIR_Mars	PF08374	SM00112	IPR013585	Protocadherin
PCDHGA4	L1M4	PF00028	SM00112	IPR002126	Cadherin
PCDHGA9	AluSx	PF00028	SM00112	IPR002126	Cadherin
PDE1A	L3	PF00233	SM00471	IPR002073	3′5′-cyclic nucleotide phosphodiesterase
PHF19	MIRb	SM00333		IPR002999	Tudor
PIGF	MIRb	PF06699		IPR009580	Phospho-ethanolamine N-methyltransferase
PKP2	AluSg	PF00514	SM00185	IPR000225	Armadillo
PLAUR	MLT1G	PF00021	SM00134	IPR001526	CD59 antigen
POLR1C	AluSc	SM00662		IPR011263	RNA polymerase, RpoA/D/Rpb3-type
PPEF1	L3	SM00054		IPR002048	Calcium-binding EF-hand
PPIL3	AluSx	PF00160		IPR002130	Peptidyl-prolyl cis-trans isomerase, cyclophilin type
PPP2R4	L1ME4a	PF03095		IPR004327	Phosphotyrosyl phosphatase activator, PTPA
PRDX2	THER1_MD	PF00578		IPR000866	Alkyl hydroperoxide reductase/ Thiol specific antioxidant/ Mal allergen
PRDX5	MIR	PF08534		IPR013740	Redoxin
PSEN1	MER52A	PF01080	SM00730	IPR001108	Peptidase A22A, presenilin
PSG4	MER65-int	PF07686		IPR013106	Immunoglobulin V-set
PSG6	AluJb	PF07686		IPR013106	Immunoglobulin V-set
PTGES	AluJb	PF01124		IPR001129	Membrane-associated proteins in eicosanoid and glutathione metabolism (MAPEG)
PTGES2	L2	PF00043		IPR004046	Glutathione S-transferase, C-terminal
PTK2	AluSx	PF03623		IPR005189	Focal adhesion targeting region
PTPRD	L3	PF00102	SM00194	IPR000242	Tyrosine specific protein phosphatase
RALGPS2	L3_Mars	SM00147		IPR001895	Guanine-nucleotide dissociation stimulator CDC25
RBM10	L3	PF00641	SM00547	IPR001876	Zinc finger, RanBP2-type
RPE	AluJ/FRAM	PF00834	PF00834	IPR000056	Ribulose-phosphate 3-epimerase
RYK	L2	SM00220		IPR002290	Serine/threonine protein kinase
SCMH1	MIR_Mars	PF02820	SM00561	IPR004092	Mbt repeat
SCN1B	MIRm	PF00047		IPR013151	Immunoglobulin
SIGLEC12	L1MD2	SM00409		IPR003599	Immunoglobulin subtype
SLC11A1	AluSx	PF01566		IPR001046	Natural resistance-associated macrophage protein
SMARCA1	L1ME4a	PF00271	SM00490	IPR001650	Helicase, C-terminal
SMN1	L1MC5	PF06003		IPR010304	Survival motor neuron
SMN2	L1MC5	PF06003		IPR010304	Survival motor neuron

(continued)

Gene	Repeat family	Database		InterPro ID	Discription
		Pfam	SMART		
SMOX	L2	PF01593		IPR002937	Amine oxidase
ST7	L2	PF04184		IPR007311	ST7
SULT1C1	AluJb	PF00685		IPR000863	Sulfotransferase
SUV420H1	MIR_Mars	PF00856	SM00317	IPR001214	Nuclear protein SET
SYNE1	L3_Mars	PF00435	SM00150	IPR002017	Spectrin repeat
TBXA2R	AluSg/x	PF00001		IPR000276	Rhodopsin-like GPCR superfamily
TIPRL	AluY	PF04176		IPR007303	TIP41-like protein
TMPRSS3	L1MDb	SM00192		IPR002172	Low density lipoprotein-receptor, class A
TRIM35	MIR3	PF00097	SM00184	IPR001841	Zinc finger, RING-type
TRIM5	AluSq	PF00097	SM00184	IPR001841	Zinc finger, RING-type
UBE2L3	L2	PF00179	SM00212	IPR000608	Ubiquitin-conjugating enzyme, E2
UBE2V1	MIRb	PF00179	SM00212	IPR000608	Ubiquitin-conjugating enzyme, E2
USP4	MIR_Mars	PF00443		IPR001394	Peptidase C19, ubiquitin carboxyl-terminal hydrolase 2
WNK3	L1MC	PF00069	SM00220	IPR000719	Protein kinase
YME1L1	LTR41	PF01434		IPR000642	Peptidase M41
ZNF197	L1MC5	SM00349		IPR001909	KRAB box
ZNF21	MSTB	PF00096	SM00355	IPR007087	Zinc finger, C2H2-type
ZNF265	L3	PF00641	SM00547	IPR001876	Zinc finger, RanBP2-type
ZNF398	L3_Mars	PF00096	SM00355	IPR007087	Zinc finger, C2H2-type

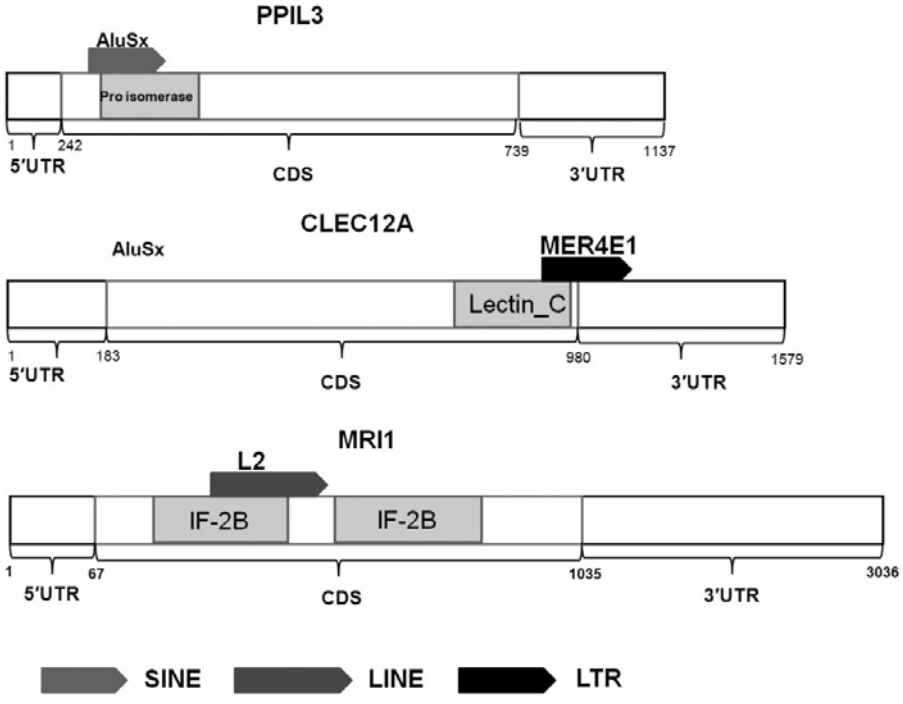

Fig. 2. Schematic representation of transposable elements into protein domain. Arrowed boxes indicate the protein domain.

quences of the human dual coding regions in the chimpanzee and mouse genomes. In previous studies, the splicing signals of TE fusion exons tend to be conserved over a long period of evolutionary time, suggesting stable exonization and potential acquisition of functional properties by TE fusion exon (Krull et al., 2007). Therefore, we inferred the presence of dual coding regions by using orthologous genomic sequences. Data for the murine orthologs were extracted using the HomoloGene (http://www.ncbi.nlm.nih.gov/HomoloGene/) database. Based on the HomoloGene catalog of gene pairs orthologous between human and mouse, we matched the exons of each human gene by sequence homology to exons of the orthologous mouse gene. The mouse ortholog gene structures were identified using the best hit method: the mouse mRNA was aligned to the mouse genome and we extracted the position information of the exon and genome sequences to be matched. Next, we used RepeatMasker (http://repeatmasker.org) to search for TE sequences in the genome. Based on this information, the loca-

Table 3. Comparative analysis of the insertion of transposable elements within human and mouse dual coding regions

Gene	Human				Mouse			
	RefSeq ID	Direct	Repeat family	Family	RefSeq ID	Direct	Repeat family	Family
ABL2	NM_007314	+	L2	LINE/L2	NM_009595	C	L2	LINE/L2
APBB1	NM_001164	+	LTR50	LTR/ERVL	NM_009685	+	LTR50	LTR/ERVL
ATP1A1	NM_000701	+	L3	LINE/CR1	NM_144900	C	L3	LINE/CR1
BTBD11	NM_001017523	C	MIR	SINE/MIR	NM_001017525	C	THER1_MD	SINE/MIR
BTBD7	NM_018167	C	HAL1	LINE/L1	XM_976481	C C +	B1_Mur2 B1_Mus1 Lx9	SINE/Alu LINE/L1
CASP6	NM_032992	+	MIRm	SINE/MIR	NM_009811	C	MIRm	SINE/MIR
CNOT1	NM_206999	C	L2	LINE/L2	XM_901788	+	B1_Mur1	SINE/Alu
CREM	NM_182719	C	MIR	SINE/MIR	NM_013498	+	MIRb	SINE/MIR
CSF2RA	XM_942501	C	LTR5B MIRb	LTR/ERVK SINE/MIR	XM_977426	C	RLTR11A	LTR/ERVK
FUSIP1	NM_006625	C	L3_Mars	LINE/CR1	NM_010178	+	L3_Mars	LINE/CR1
HIF3A	NM_152796	C	AluJb	SINE/Alu	XM_989804	C	MIR B2_Mm1a B1_Mm	SINE/MIR SINE/B2 SINE/Alu
IFT81	NM_014055 NM_031473	+ C	HAL1 AluJo	LINE/L1 SINE/Alu	NM_009879	C	HAL1	LINE/L1
IQWD1	NM_001017977	C	L3	LINE/CR1	XM_917541 XM_991018 NM_001025358	+ + C	L3 RLTR17-int MIRb	LINE/CR1 LTR/ERVK SINE/MIR
KNS2	NM_005552	C	MIR_Mars	SINE/MIR	XM_992143	+	B1_Mur4 MT2B	SINE/Alu LTR/ERVL
					XM_992253	+	B3	SINE/B2
L3MBTL3	NM_032438	+	L2	LINE/L2	NM_172787	C	L2	LINE/L2
LRP8	NM_004631	+	MIRm	SINE/MIR	NM_053073	C	L2	LINE/L2
MACF1	NM_012090	C	L2	LINE/L2	XM_991844	C	RLTR25A B1_Mus1 MT2B	LTR/ERVK SINE/Alu LTR/ERVL
MLL3	NM_021230	C	THER1_MD	SINE/MIR	XM_987930 XM_982326	+ C	Lx3B RSINE1	LINE/L1 SINE/B4
MOBP	NM_006501	+	LTR33A	LTR/ERVL	NM_001039365	+	LTR50	LTR/ERVL
NALP1	NM_014922	+	AluJb	SINE/Alu	NM_001004142	+	L1MA9	LINE/L1
NIN	NM_182946	+	MER101-int	LTR/ERV1	NM_008697	+	MER66-int	LTR/ERV1
NRAP	NM_006175	+	L4	LINE/RTE	NM_198059	+	L4	LINE/RTE
NTRK3	NM_001012338	+	L1MEe	LINE/L1	NM_182809	+	MIR3	SINE/MIR
	NM_001007156	+	MIR3	SINE/MIR	XM_981510	+	Lx5 L1M5	LINE/L1
OSBPL8	NM_020841	+	L3_Mars	LINE/CR1	NM_175489	C	L3_Mars	LINE/CR1
RBM10	NM_005676	C	L3	LINE/CR1	NM_145627	C	L3	LINE/CR1
RFX4	NM_032491	+	L3	LINE/CR1	NM_027689	+	L3	LINE/CR1
RNF7	NM_183063	C	MIR	SINE/MIR	XM_905787	+	Lx8	LINE/L1
ST7	NM_021908	C +	L3 L2	LINE/CR1 LINE/L2	NM_022332	C	L3	LINE/CR1
TEX11	NM_001003811	+	AluSg	SINE/Alu	NM_031384	C	B3	SINE/B2
USH2A	NM_206933	C	MIR3	SINE/MIR	NM_021408	+	L2	LINE/L2

Table 4. Comparative analysis of the insertion of transposable elements within human and chimpanzee coding regions

Gene	Human				Chimpanzee			
	RefSeq ID	Direct	Repeat family	Family	RefSeq ID	Direct	Repeat family	Family
ABCA6	NM_080284	C	L3_Mars	LINE/CR1	XM_001146278	C	L3_Mars	LINE/CR1
ABL2	NM_007314	+	L2	LINE/L2	XM_001156169	+	L2	LINE/L2
ACOT11	NM_015547	+	MER9	LTR/ERVK	XM_001152717	+	HERVK9	LTR/ERVK
		+	HERVK9					
AES	NM_198969	+	MIR	SINE/MIR	XM_001153657	+	MIRb	SINE/MIR
AKAP9	NM_147166	C	L1MA4	LINE/L1	XM_527814	+	L1M4c	LINE/L1
ALG3	NM_001006940	+	L2	LINE/L2	XM_001148398	+	L2	LINE/L2
APBB1	NM_145689	+	LTR50	LTR/ERVL	XM_521814	+	LTR50	LTR/ERVL
ATP1A1	NM_001001586	+	L3	LINE/CR1	XM_513679	C	L3	LINE/CR1
ATRX	NM_138270	+	L2	LINE/L2	NM_001009018	+	L2	LINE/L2
BCL2L11	NM_138625	C	AluSx	SINE/Alu	XM_001142733	C	L1MB2,L1MD2	LINE/L1
	NM_207002	C	L1MB2	LINE/L1				
	NM_207002	C	L1MD2	LINE/L1				
CASC5	NM_144508	+	L2	LINE/L2	XM_510312	+	L2	LINE/L2
	NM_170589	C	AluSq	SINE/Alu				
CCNT2	NM_058241	C	L3	LINE/CR1	XM_001152297	C	L3	LINE/CR1
	NM_001241	+	L4	LINE/RTE		+	L4	LINE/RTE
CD163	NM_203416	+	L3	LINE/CR1	XM_508986	+	L3	LINE/CR1
CDK5RAP2	NM_001011649	+	L2	LINE/L2	NM_001042436	+	L2	LINE/L2
					NM_001042436	C	AluJb	SINE/Alu
CREM	NM_183013	C	MIR	SINE/MIR	XM_001147643	C	MIR	SINE/MIR
CUGBP2	NM_006561	+	L4	LINE/RTE	XM_507653	+	L4	LINE/RTE
ETFB	NM_001985	+	L3	LINE/CR1	XM_001164201	C	MIR	SINE/MIR
	NM_001014763	C	MIR	SINE/MIR				
FBLIM1	NM_001024215	C	MER21B	LTR/ERV1	XM_001150414	C	MER21A	LTR/ERV1
		+	AluY	SINE/Alu		+	AluY	SINE/Alu
FBXO18	NM_032807	+	MIR3	SINE/MIR	XM_507640	+	L4	LINE/RTE
FOXM1	NM_202002	+	L1MC	LINE/L1	XM_508929	+	L1PB1	LINE/L1
GABRB2	NM_021911	+	L3	LINE/CR1	XM_001143894	+	L3	LINE/CR1
GFRA1	NM_145793	C	L3_Mars	LINE/CR1	XM_001150865	C	L3_Mars	LINE/CR1
GPR110	NM_153840	+	L2	LINE/L2	XM_518521	+	L2	LINE/L2
	NM_025048	+	AluSq	SINE/Alu				
GRIK2	NM_021956	C	L2	LINE/L2	XM_518653	C	L2	LINE/L2
HAP1	NM_003949	C	MIRb	SINE/MIR	XM_001167791	+	MIRb	SINE/MIR
HISPPD2A	NM_014659	+	L3	LINE/CR1	XM_510352	C	L4	LINE/RTE
						+	L3	LINE/CR1
IFT122	NM_052985	+	FLAM_C	SINE/Alu	XM_001142824	C	MIRb	SINE/MIR
IFT81	NM_014055	+	HAL1	LINE/L1	XM_001141206	+	HAL1	LINE/L1
	NM_031473	C	AluJo	SINE/Alu				
IL22RA2	NM_052962	C	MSTB2	LTR/MaLR	XM_001171428	C	MSTB2	LTR/MaLR
IL5RA	NM_175725	C	THER1_MD	SINE/MIR	XM_001138951	+	L1M4	LINE/L1
ING4	NM_016162	+	L3_Mars	LINE/CR1	XM_001169091	+	L3_Mars	LINE/CR1
IQWD1	NM_001017977	C	L3	LINE/CR1	XM_001174797	C	L3	LINE/CR1
KLRC3	NM_002261	+	AluJb	SINE/Alu	NM_001009017	+	AluJb	SINE/Alu
KNS2	NM_005552	C	MIR_Mars	SINE/MIR	XM_001140159	C	MIR_Mars	SINE/MIR

(continued)

Gene	Human				Chimpanzee			
	RefSeq ID	Direct	Repeat family	Family	RefSeq ID	Direct	Repeat family	Family
L3MBTL3	NM_032438	+	L2	LINE/L2	XM_518737	+	L2	LINE/L2
LRP8	NM_004631	+	MIRm	SINE/MIR	XM_513416	+	MIRm	SINE/MIR
MAD2L1BP	NM_001003690	C	AluSp	SINE/Alu	XM_001139524	C	AluSp	SINE/Alu
MAGI3	NM_020965	C	AluSc	SINE/Alu	XM_001161602	C	L2	LINE/L2
MTO1	NM_133645	C	AluSp	SINE/Alu	XM_527435	C	AluSp	SINE/Alu
MYO18A	NM_078471	+	L3_Mars	LINE/CR1	XM_511371	C	L3_Mars	LINE/CR1
NALP1	NM_033007	+	AluJb	SINE/Alu	XM_001167078	+	AluJb	SINE/Alu
						C	L3_Mars	LINE/CR1
					XM_001167200	+	L2	LINE/L2
NEK10	NM_001031741	C	MSTA	LTR/MaLR	XM_001165354	C	MSTA	LTR/MaLR
NFAT5	NM_006599	C	L3_Mars	LINE/CR1	XM_001168930	C	L3_Mars	LINE/CR1
NMNAT2	NM_170706	C	L3	LINE/CR1	XM_001162779	C	L3	LINE/CR1
NRAP	NM_198060	+	L4	LINE/RTE	XM_521607	+	L4	LINE/RTE
OPN4	NM_001030015	+	MIRb	SINE/MIR	XM_001135445	+	MIRb	SINE/MIR
OSBPL9	NM_148909	C	L3_Mars	LINE/CR1	XM_001143018	C	L3_Mars	LINE/CR1
PAX7	NM_013945	C	MIR	SINE/MIR	XM_513156	C	MIR	SINE/MIR
PIGF	NM_173074	+	MIRb	SINE/MIR	XM_001147744	+	MIRb	SINE/MIR
PIP5K2B	NM_138687	+	AluSc	SINE/Alu	XM_001172797	C	AluSc	SINE/Alu
PSEN1	NM_007319	C	MER52A	LTR/ERV1	XM_001149419	C	MER52A	LTR/ERV1
RAD51L1	NM_133509	C	AluJo	SINE/Alu	XM_001139007	C	AluJo	SINE/Alu
RFX4	NM_032491	+	L3	LINE/CR1	XM_001161711	+	AluJo	SINE/Alu
					XM_001161866	+	L3	LINE/CR1
RPE	NM_006916	+	L4	LINE/RTE	XM_001145990	C	AluSx	SINE/Alu
		C	AluJ/FRAM	SINE/Alu	XM_001145546	C		
SCMH1	NM_001031694	C	MIR_Mars	SINE/MIR	XM_001172755	C	MIR_Mars	SINE/MIR
SLC3A2	NM_001012662	C	AluSq	SINE/Alu	XM_001159413	C	FLAM_C	SINE/Alu
	NM_001012661		FLAM_C		XM_001159363		AluSq	
SMN1	XM_936952	+	L1MC5	LINE/L1	XM_001156435	+	L1MC5	LINE/L1
SPAM1	NM_003117	C	MLT1B	LTR/MaLR	XM_527873	C	MER41B	LTR/ERV1
ST7	NM_021908	C	L3	LINE/CR1	XM_001157113	+	AluY	SINE/Alu
		+	L2	LINE/L2		+	MIRb	SINE/MIR
	NM_138728	C	MIRb	SINE/MIR	XM_001139362	C	L3	LINE/CR1
					XM_001157439	C	AluSq	SINE/Alu
SULT1C1	NM_176825	C	AluJb	SINE/Alu	XM_001135224	C	AluJb	SINE/Alu
SUV420H1	NM_017635	+	MIR_Mars	SINE/MIR	XM_001173457	+	MIR_Mars	SINE/MIR
SYNE1	NM_182961	C	L3_Mars	LINE/CR1	XM_001138443	C	L3_Mars	LINE/CR1
TMPRSS3	NM_032405	+	L1MDb	LINE/L1	XM_531474	+	L1MDb	LINE/L1
USH2A	NM_206933	C	MIR3	SINE/MIR	XM_001171992	+	L1P3	LINE/L1
					XM_001171804	+	LTR9	LTR/ERV1
					XM_001171818	+	Tigger1	DNA/MER2_type
USP4	NM_003363	C	MIR_Mars	SINE/MIR	XM_001162976	C	MIR_Mars	SINE/MIR
UTY	NM_182660	+	AluSx	SINE/Alu	NM_001009002	+	AluSx	SINE/Alu
VPS13D	NM_015378	+	L2	LINE/L2	XM_514406	+	L2	LINE/L2
YIF1B	NM_001031731	C	AluY	SINE/Alu	XM_001166472	C	AluY	SINE/Alu
YME1L1	NM_139312	+	LTR41	LTR/ERVL	XM_507710	+	LTR41	LTR/ERVL

tion of the TEs and exons were calculated from their position on the genome. Finally, 30 of the mouse ortholog genes contained TEs in the coding region corresponding to the human coding region (Table 3), but 219 of the mouse orthologs contained no TE sequences. Previous studies show that approximately 7% of alternatively spliced genes contain dual (multiple) coding regions in known human gene (Liang et al., 2005). Also, Sela et al*.,* did not focused their analysis in the dual coding gene, but described TE exonizations in protein coding regions (dual coding gene). In humans, 0.12% of the TEs exonized within protein coding genes and 0.18% of the TEs exonized within non-protein-coding genes. In contrast, in mouse, 0.06% rate of exonization within protein coding genes and 0.08% in non-protein-coding genes (Sela et al., 2007). And also, about 94% of TE exonization in human and 88% of the TE exonizations in mouse generated an internal cassette exon (Sela et al., 2007). Moreover, the effect of TEs on the human transcriptome was several times greater than the effect on the mouse transcriptome, mostly because of the contribution of the primate-specific *Alu* elements. Therefore, we could assume that TE-derived dual coding events are mainly identified in human lineages.

To analyze the lineage specific TE-derived dual coding events, chimpanzee orthologs of all 151 genes were added to our analysis. Intriguingly, 71 of the orthologous chimp genes contained TEs in the protein coding regions corresponding to the human protein coding regions (Table 4) and 80 of the chimp orthologs contained no TEs sequences. Although, the available gene information of the chimpanzee is limited, we could conclude that the human specific TE-derived dual coding events. A recent survey of all vertebrate protein coding sequences showed that TEs from all categories contribute to protein variability (Britten, 2006; Kreahling et al., 2004; Li et al., 2001). For example, the existence of 1.4 million Alu elements interspersed throughout our genome, each with Alu carrying several potential splicing sites, provides numerous possibilities for alternative transcripts (Kreahling et al., 2004).

Therefore, lineage specific TEs, specifically primate Alu elements, could contribute to protein diversity by providing different protein coding regions through dual coding events. Our results can also could help us to understand the evolutionary phenomenon of lineage specific new exons and functional diversity in the limited gene sources.

ACKNOWLEDGEMENT

This study was supported by a research grant from the Ministry of Education, Science and Technology (KBM4311022, KGC 5401011), Republic of Korea.

REFERENCES

Ahn, K., and Kim, H.S. (2009). Structural and quantitative expression analyses of HERV gene family in human tissues. Mol. Cells *28*, 99-103.

Andersen, J.N., Jansen, P.G., Echwald, S.M., Mortensen, O.H., Fukada, T., Del Vecchio, R., Tonks, N.K., and Moller, N.P. (2004). A genomic perspective on protein tyrosine phosphatases: gene structure, pseudogenes, and genetic disease linkage. FASEB J. *18*, 8-30.

Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L., et al. (2004). The Pfam protein families database. Nucleic Acids Res. *32*, D138-D141.

Britten, R. (2006). Transposable elements have contributed to thousands of human proteins. Proc. Natl. Acad. Sci. USA *103*, 1798-1803.

Christensen, T. (2005). Association of human endogenous retroviruses with multiple sclerosis and possible interaction with herpes viruses. Rev. Med. Virol. *15*, 179-211.

Florea, L., Hartzell, G., Zhang, Z., Rubin, G.M., and Miller, W. (1998). A computer program for aligning a cDNA sequence with a genomic DNA sequence. Genome Res. *8*, 967-974.

Gerber, A., O'Connell, M.A., and 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., and Makalowski, W. (2006). Do transposable elements really contribute to proteomes? Trends Genet. *22*, 260-267.

Halling, K.C., Lazzaro, C.R., Honchel, R., Bufill, J.A., Powell. S.M., Arndt, C.A.S., and Lindor, N.M. (1999). Hereditary desmoid disease in a family with a germline Alu I repeat mutation of the APC gene. Hum. Hered *49*, 97-102.

Jurka, J. (2000). Repbase update: a database and an electronic journal of repetitiveelements. Trends Genet. *16*, 418-420.

Kreahling, J., and Graveley, B.R. (2004). The origins and implications of Aluternative splicing. Trends Genet. *20*, 1-4.

Krull, M., Petrusma, M., Makalowski, W., Brosius, J., and Schmitz, J. (2007). Functional persistence of exonized mammalian-wide interspersed repeat elements (MIRs). Genome Res. *17*, 1139-1145.

Li, W.H., Gu, Z., Wang, H., and Nekrutenko, A. (2001). Evolutionary analyses of the human genome. Nature *15*, 847-849.

Liang, H., and Landweber, L.F. (2005). A genome-wide study of dual coding regions in human alternatively spliced genes. Genome Res. *16*, 190-196.

Licatalosi, D.D., and Darnell, R.B. (2010). RNA processing and its regulation: global insights into biological networks. Nat. Rev. Genet. *11*, 75-87.

Lopez, A.J. (1998). Alternative splicing of pre-mRNA: developmental consequences and mechanisms of regulation. Annu. Rev. Genet. *32*, 279-305.

Makalowski, W. (2000). Genomic scrap yard: how genomes utilize all that junk. Gene *23*, 61-67.

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

Modrek, B., and Lee, C. (2002). A genomic view of alternative splicing. Nat. Genet. *30*, 13-19.

Mulder, N.J., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Biswas, M., Bradley, P., Bork, P., Bucher, P., et al. (2002). InterPro Consortium: InterPro: an integrated documentation resource for protein families, domains and functional sites. Brief Bioinform. *3*, 225-235.

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

Pruitt, K.D., Tatusova, T., and Maglott, D.R. (2005). NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes transcripts and proteins. Nucleic Acids Res. *33*, D501-504.

Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., and Lopez, R. (2005). InterProScan: protein domains identifier. Nucleic Acids Res. *33*, W116-W120.

Schultz, J., Milpetz, F., Bork, P., and Ponting, C.P. (1998). SMART, a simple modular architecture research tool: identification of signaling domains. Proc. Natl. Acad. Sci. USA *95*, 5857-5864.

Scorilas, A., Kyriakopoulou, L., Yousef, G.M., Ashworth, L.K., Kwamie, A., and Diamandis, E.P. (2001). Molecular cloning, physical mapping,and expression analysis of a novel gene, BCL2L12, encoding a proline-rich protein with a highly conserved BH2 domain of the Bcl-2 family. Genomics *72*, 217-221.

Sela, N., Mersch, B., Gal-Mark, N., Lev-Maor, G., Hotz-Wagenblatt, A., and Ast, G. (2007). Comparative analysis of transposed element insertion within human and mouse genomes reveals Alu's unique role in shaping the human transcriptome. Genome Biol. *8*, R127.

Sin, H.S., Huh, J.W., Kim, D.S., Kang, D.W., Min, D.S., Kim, T.H., Ha, H.S., Kim, H.H., Lee, S.Y., and Kim, H.S. (2006). Transcriptional control of HERV-H LTR element of GSDML gene in human tissues and cancer cells. Arch. Virol.*151*, 1985-1994.

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

Szathmáry, E., Jordán, F., and Pál, C. (2001). Molecular biology and evolution. can genes explain biological complexity? Science *292*, 1315-1316.

Xu, H., Wang, P., Fu, Y., Zheng, Y., Tang, Q., Si, L., You, J., Zhang, Z., Zhu, Y., Zhou, L., et al. (2010). Length of the ORF, position of

the first AUG and the Kozak motif are important factors in potential dual-coding transcripts. Cell Res. *20*, 445-457.

Yi, X., and Lee, C. (2006). Alternative splicing and RNA selection pressure-evolutionary consequences for eukaryotic genomes. Nat. Rev. Genet. *7*, 499-509.

Zhao, F.Q., Zheng, Y., Dong, B., and Oka, T. (2004). Cloning, genomic organization, expression, and effect on β-casein promoter activity of a novel isoform of the mouse Oct-1 transcription factor. Gene *326*, 175-187.