Transposable elements are found in a large number of human protein-coding genes

Anton Nekrutenko and Wen-Hsiung Li

To study the genome-wide impact of transposable elements (TEs) on the evolution of protein-coding regions, we examined 13 799 human genes and found 533 (~4%) cases of TEs within protein-coding regions. The majority of these TEs (~89.5%) reside within 'introns' and were recruited into coding regions as novel exons. We found that TE integration often has an effect on gene function. In particular, there were two mouse genes whose coding regions consist largely of TEs, suggesting that TE insertion might create new genes. Thus, there is increasing evidence for an important role of TEs in gene evolution. Because many TEs are taxon-specific, their integration into coding regions could accelerate species divergence.

Transposable elements (TEs) have been thought to have a negligible role in the evolution of protein-coding sequences because a TE insertion into the protein-coding region of a gene is most likely to be deleterious and eliminated from the population. However, as there are more than 4 million TEs in the human genome[1,2], most of which are retrotransposable elements, TEs might have had a strong impact on the evolution of transcribed protein-coding regions. Previous studies[3,4] listed many instances of TEs in mammalian coding regions, but it is not clear what the frequency of TEs in human coding sequences is, especially on a genome-wide scale. Moreover, the mechanisms of TE integration into protein-coding sequences and the effect of integration on gene function have not been well explored. How frequently are TEs found inside protein-coding regions? How are TEs introduced into protein-coding regions? What are the effects of taxon-specific TEs on the evolution of orthologous genes?

To address the above questions we used the UniGene database at the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/UniGene) because the building procedure of this database is less likely to exclude TE-containing sequences than the RefSeq database. Only sequences representing protein-coding regions of genes were selected for the analysis; all expressed sequence tag (EST)-derived entries were excluded. The resultant dataset contained 13 799 human sequences. To find TEs, we compared the sequences against a collection of TE consensus sequences using the RepeatMasker program (http://genome.washington.edu). We found 533 (~4% out of 13 799) human protein-coding regions containing TEs or TE fragments (Table 1; the entire dataset is available at http://nekrut.uchicago.edu/research.html). Extrapolating this result, if there are 30 000 genes in the human genome, then 1200 genes (4%) contain transposable elements in their protein-coding regions!

How do TEs integrate into coding regions?
Next, we considered how TEs are integrated into coding regions. There are two possibilities: a TE can be inserted directly into a protein-coding exon, or it can be inserted into a noncoding region (e.g. an intron) and is subsequently recruited as a new exon (Fig. 1). To determine which of the two routes is more common, we compiled a dataset of 6351 human genes (71 964 protein-coding exons) from GenBank. We found 189 genes (~3%) that contained at least one exon homologous to a TE (306 exons in total). Among these 306 exons, we found only 32 (~10.5%) that actually contained the TE. In the other cases (89.5%), the TE was recruited as a novel exon. This high rate of exon recruitment from TEs is possible because many TEs carry potential splice sites; for example, the consensus sequence of Alu elements (short interspersed elements [SINES] carrying the recognition site for the restriction enzyme *Alu* I)contains eight putative donor sites and three acceptor sites[5]. Thus, TE insertion might be one important cause for the high frequency of alternative splicing in human protein-coding genes[6]. 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.

How important is TE integration for functional divergence between orthologous genes?
To address this question, we extracted 2109 groups of putative orthologous genes from human, mouse, and rat from the HomoloGene database at NCBI (http://www.ncbi.nlm.nih.gov/HomoloGene). We found 123 (6%) cases where one of the sequences in a group contained a TE-insertion within the coding region. Unfortunately, in the vast majority of cases there were only sequence data but no functional studies. Table 2 lists the cases for which we were able to use published data to detect structural or expressional differences between species. For example, insertion of an Alu at the C-terminus of human hematopoietic progenitor kinase is fixed (i.e. the Alu is present in all transcripts). Activity assays demonstrated that the C-terminus is important for the proper functioning of this enzyme, enhancing its ability to activate the transcription factor AP1 (Ref. 7). As another example, insertion of an LTR-containing element at the 5′ end of the human 8-oxo-7,8-dihydrodeoxyguanosine triphosphatase gene generated three additional in-frame start codons that are used with almost equal frequency[8]. This enzyme is important in repairing DNA damage caused by oxygen radicals and is thought to be most pertinent to mutagenesis and carcinogenesis. The

Table 1. Transposable elements (TEs) in coding regions of 13 799 human genes exctracted from the UniGene database[a]

TE class	No. found
SINEs (*Alu* only)[b]	213 (177)
LINEs (*L1* only)[c]	146 (73)
LTRs	130
DNA transposons	48
Others	21
Total	558
Total no. of affected genes[d]	533 (~4%)

[a]TEs were indentified using the RepeatMasker program using a Smith–Waterman cutoff score of 225.
[b]SINE, short interspersed element
[c]LINE, long interspersed element
[d]This number is smaller than the total number of TEs found because some genes contain more than one TE.

 0168-9525/01/$ – see front matter PII: S0168-9525(01)02445-3

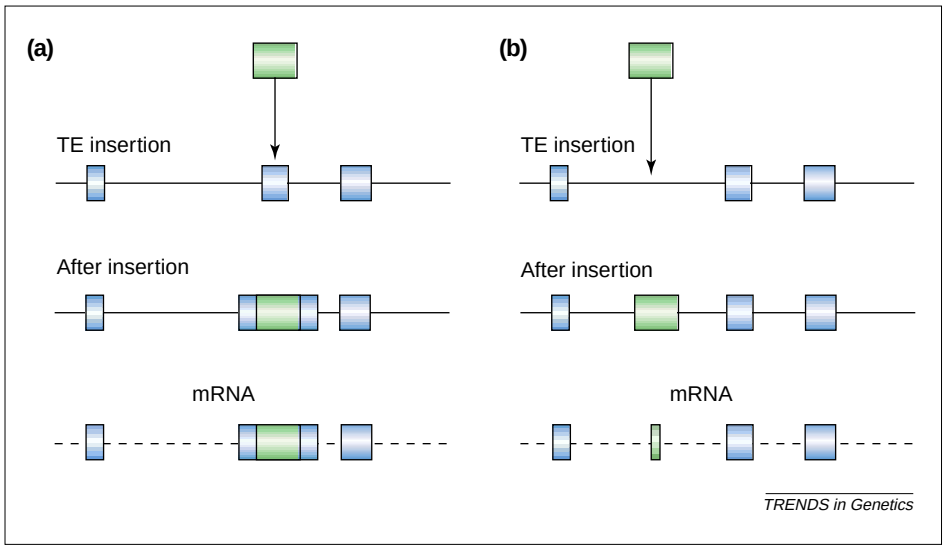

Fig. 1. Two possible mechanisms of transposable element (TE) insertion into protein-coding region. (a) TE is inserted directly into a protein-coding exon. The effects of the direct insertion are likely to be deleterious because TEs often contain multiple stop codons and would destroy the target exon. (b) TE is inserted into an intron. Later a portion of the TE is recruited as a novel exon. This scenario is preferred for two reasons: first, in many instances the novel exon is alternatively spliced and may not destroy the function of the gene; and second, typically only a fragment of the TE insert is recruited, which is less likely to contain stop codons.

Table 2. Coding regions modified by transposable element (TE) insertion in human, mouse or rat

Species[a]	UniGene ID[b]	Gene	TE type	TE family	Effect on coding region[c]
Hs	Hs.86575	Human hematopoietic progenitor kinase (HPK1)	*Alu*S	SINE	Ext
Hs	Hs.25051	Plakophilin 2a and b	*Alu*S	SINE	Ext
Hs	Hs.85302	Adenosine deaminase	*Alu*J	SINE	Ext
Hs	Hs.5648	Proteasome subunit p27	L1MB3	LINE	Stop
Hs	Hs.239	Hepatocyte nuclear factor-3/fork head protein	L1PB1	LINE	Ext
Hs	Hs.25674	Methyl-CpG binding protein	L2	LINE	Stop
Hs	Hs.66493	Down syndrome critical region gene 5	L2	LINE	Start
Hs	Hs.23205	Drosophila tumor suppressor homolog (*DLG2*)	LIMC4	LINE	Ext
Hs	Hs.388	8-oxo-dGTPase	LTR1	LTR/Ret	Start
Hs	Hs.169309	Myelin-associated oligodendrocytic basic protein	LTR50	LTR	Stop
Mm	Mm.3785	Protein phosphatase 2A regulatory subunit	IAP	LTR/Ret	Start
Mm	Mm.3257	Growth arrest specific protein 7	L2	LINE	Start
Mm	Mm.1202	Myeloblastosis viral oncogene	MULV	LTR/Ret	Start
Mm	Mm.4788	Retinoic acid receptor-alpha	RLTRETN	LTR/Ret	Stop
Mm	Mm.42233	Serotonin *N*-acetyltransferase	URR1B	DNA	Stop
Mm, *Rn*	Mm.20935 Rn.24511	Interleukin enhancer binding factor 3	MER44D	DNA	Start
Rn	Rn.10765	Lung-derived *c-ros*-1 proto-oncogene	B4A	SINE	Ext
Rn	Rn.9963	Tyrosine protein kinase (trkC)	L1	LINE	Start
Rn	Rn.6032	Voltage-gated sodium channel	L1	LINE	Start

[a]TE insertion is restricted to this species only: *Hs*, *Homo sapiens*; *Mm*, *Mus musculus*; *Rn*, *Rattus norvegicus*.
[b]As the HomoloGene database does not provide unique accession numbers, we use UniGene IDs for reference.
[c]Abbreviations: ext, extension; start, introduction of alternative start; stop, introduction of alternative stop.

presence of additional initiation sites increases the cellular production of this important protein.

In addition, we found two mouse mRNAs that contain one or many TEs (Fig. 2) and have no orthologs in human or rat, suggesting that TE insertion might create new genes. The first gene, lungerkine (Accession number: AF082859), is a novel CXC chemokine (a basic heparin-binding protein), whose expression is restricted to the lung[9]. Chemokines are key in the process of lymphocyte recruitment to specific sites of inflammation or injury. Mouse lungerkine exhibits only marginal sequence similarity at the N-terminus to known CXC chemokines and contains a long C-terminus that is derived from a long terminal repeat of retroviral-like element (Fig. 2a) and is important for the lungerkine stability. The second gene, mNSC1 (Accession number: D50656), belongs to a group of Ca^{2+}-activated nonselective cation channels found in neurons, macrophages, insulin secreting, and ganglion cells[10]. The mNSC1 mRNA contains nine different SINE elements in both orientations, five of which are interspersed within the coding region (Fig. 2b). Remarkably, 263 (62%) of the 424 amino acid residues in this protein are encoded by fragments of B1, B3 and B4 elements, which are rodent SINEs.

Conclusions

Our results have important evolutionary implications, namely that TE insertion can accelerate the evolution of genes and provide a means for rapid species divergence. For example, the existence of ~1.4 million Alu elements interspersed throughout our genome[1,2], with each Alu carrying several potential splicing sites[5], provides numerous possibilities for the formation of alternative transcripts. A certain, albeit small, fraction of the alternative transcripts might be advantageous; for instance, by increasing the functional versatility of the gene. Compared with the slow process of gene evolution by nucleotide substitution, a TE insertion can suddenly expand or truncate the coding region. So, orthologous genes in two different species can encode functionally different proteins or can differ in terms of expression because of TE insertion in one lineage but not in the other.

Let us consider Alu insertion again. It has been estimated that Alu insertion occurs once in every 200 human births[11].

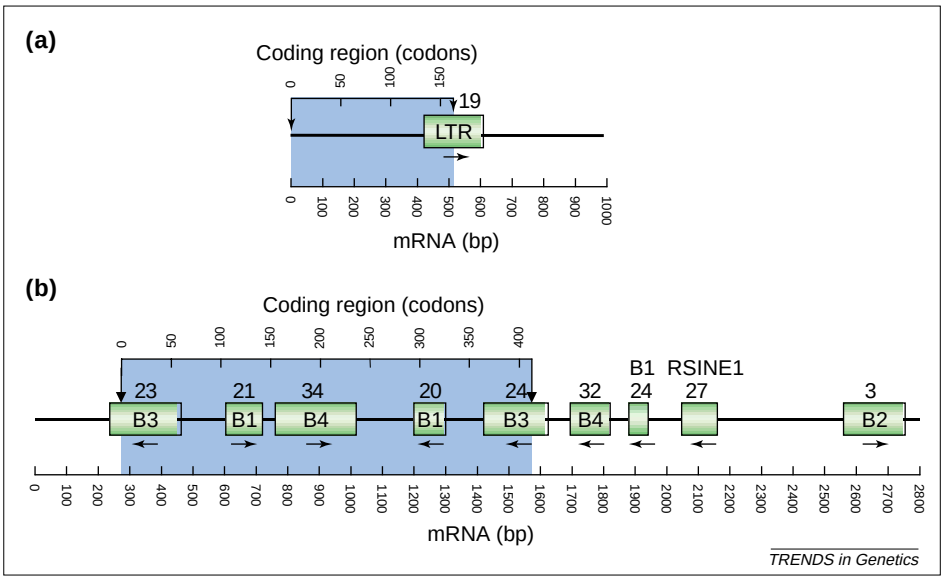

Fig. 2. Transposable elements (TEs) in lungerkine mRNA (a) and mNSC1 mRNA (b). The direction of transcription is from left to right. Boxes indicate TEs and arrows indicate TE orientation (relative to the corresponding consensus sequences). Shaded area represents the coding region. Numbers above boxes show % divergence from the TE consensus sequence. B1, B2, B3, B4 and RSINE1 are families of rodent short interspersed elements and were likely introduced by independent insertion events. LTR is a long terminal repeat of a rodent retrovirus-like element. In each case RepeatMasker hits were examined manually to exclude the possibility of misprediction.

Thus, many of the Alu elements in the human genome would not be found in the chimpanzee genome (and vice versa) and might be partly responsible for the conspicuous morphological differences between human and chimpanzee. Over a longer evolutionary period, Alu elements might have contributed significantly to the divergence between primates and other mammals because Alu elements are not found in non-primates. In our dataset (Table 1), Alu elements are found in ~1.3% of human coding regions, so if the human genome contains 30 000 genes, then ~400 human genes contain Alu inserts in their coding regions that are not found outside the primate lineage. By generalization, a large number of TEs might not be junk, but could have had a significant role in gene evolution[12] or species divergence.

Acknowledgements

We thank Richard Blocker for his help with the UNIX system and Arian Smit for providing the RepeatMasker program. We are also grateful to Zhenglong Gu for help in preparing the manuscript. This work was supported by NIH grants GH30998, GM55759 and HD38287.

References

1 Smit, A.F.A. (1999) Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr. Opin. Genet. Dev.* 9, 657–663
2 Li, W-H. *et al.* (2001) Evolutionary analysis of the human genome. *Nature* 409, 847–849
3 Brosius, J. (1999) Genomes were forged by massive bombardments with retroelements and retrosequences. *Genetica* 107, 209–238
4 Makalowski, W. (2000) Genomic scrap yard: how genomes utilize all that junk. *Gene* 259, 61–67
5 Makalowski, W. *et al.* (1994) Alu sequences in the coding regions of mRNAs: a source of protein variability. *Trends Genet.* 10, 188–193
6 Brett, D. *et al.* (2000) EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. *FEBS Lett.* 474, 83–86
7 Hu, M.C. *et al.* (1996) Human HPK1, a novel human hematopoietic progenitor kinase that activates the JNK/SAPK kinase cascade. *Genes Dev.* 18, 5514–5524
8 Oda, H. *et al.* (1997) Regulation of expression of the human MTH1 gene encoding 8-Oxo-dGTPase. *J. Biol. Chem.* 272, 17843–17850
9 Rossi, D.L. *et al.* (1999) Lungerkine, a novel CXC chemokine, specifically expressed by lung bronchoepithelial cells. *J. Immunol.* 162, 5490–5497
10 Suzuki, M. *et al.* (1998) Primary structure and functional expression of a novel non-selective cation channel. *Biochem. Biophys. Res. Commun.* 242, 191–196
11 Deininger, P.L. and Batzer, M.A. (1999) Alu repeats and human disease. *Mol. Genet. Metab.* 67, 183–193
12 Brosius, J. and Gould, S.J. (1992) On 'genomenclature': a comprehensive (and respectful) taxonomy for pseudogenes and other 'junk DNA'. *Proc. Natl. Acad. Sci. U. S. A.* 89, 10706–10710

Anton Nekrutenko
Wen-Hsiung Li*
1101 East 57th Street, Dept. of Ecology and Evolution, University of Chicago, Chicago, IL 60637, USA.
*e-mail: whli@uchicago.edu