Genes Genet. Syst. (2007) **82**, p. 145–156

Transposable elements in human cancers by genome-wide EST alignment

Dae-Soo Kim[1], Jae-Won Huh[2] and Heui-Soo Kim[1,2]*

[1]*PBBRC, Interdisciplinary Research Program of Bioinformatics, Pusan National University, Busan 609-735, Republic of Korea*
[2]*Division of Biological Sciences, College of Natural Sciences, Pusan National University, Busan 609-735, Republic of Korea*

(Received 24 November 2006, accepted 23 January 2007)

Transposable elements may affect coding sequences, splicing patterns, and transcriptional regulation of human genes. Particles of the transposable elements have been detected in several tissues and tumors. Here, we report genome-wide analysis of gene expression regulated by transposable elements in human cancers. We adopted an analysis pipeline for screening methods to detect cancer-specific expression from expressed human sequences. We developed a database (TECESdb) for understanding the mechanism of cancer development in relation to transposable elements. A total of 999 genes fused with transposable elements were found to be cancer-related in our analysis of the EST database. According to GO (Gene Ontology) analysis, the majority of the 999 cancer-specific genes have functional association with gene receptor, DNA binding, and kinase activity. Our data could contribute greatly to our understanding of human cancers in relation to transposable elements.

Key words: Transposable elements, Cancer, Fusion gene, Bioinformatics, EST

INTRODUCTION

The human genome is estimated to be composed of 45% transposable elements (International Human Genome Sequencing Consortium 2001). The transposable elements are categorized as four types; long interspersed nuclear elements (LINEs) or non-long terminal repeat retrotransposons, short interspersed nuclear elements (SINEs), LTR retroposons, and DNA transposons (Smit, 1999; Kazazian, 2004). Defects in their structures mainly include multiple stop codon, deletion, and insertion mutations. Nevertheless, some transposable elements have been reported to have a capacity for affecting adjacent genes by altering transcriptional regulation (Nigumann et al., 2002; Sin et al., 2006). Moreover, these insertion mutations could cause genetic diseases and contribute to protein variability or versatility in the human genome (Makalowski et al., 1999; Halling et al., 1999; Christensen, 2005). Most transposable elements are transcriptionally silent in normal human tissues. However, some transposable elements have been found to be expressed specifically in placenta tissues and cancer cell lines (Lower et al., 1996; Mi et al., 2000; Yi and Kim, 2004). Several fragments of transposable elements have also appeared in open-reading frames of functional human genes (Yulug et al., 1995; Makalowski et al., 1999; Nekrutenko and Li, 2001; Huh et al., 2006).

The L1 5'UTR element is known to have an antisense promoter sequence, and that chimeric transcript is initiated from an antisense promoter sequence in the 5'UTR of a full-length LINE-1 element in primary esophageal adenocarcinoma (Lin et al., 2006). Desmoid disease, a retrotransposition event of the *Alu* I element caused the truncation of a protein sequence in the middle region of the APC gene (Medstrand et al., 2001). Another important point is that transposable elements could act as portable carriers of regulator elements, such as promoter or enhancers, around functional genes. L1 antisense promoter-driven transcription has been detected in human tumor cells and normal cells, and HERV LTR elements have been shown to have bidirectional promoter activity (Medstrand et al., 2001; Nigumann et al., 2002; Dunn et al., 2003; Sin et al., 2006). Several genes showed tumor-specific alternative splicing by integration of TEs (Okumura et al., 2005). In addition, genome-wide expressed sequence tag (EST) alignment indicated that alternatively spliced mRNA variants were related to various human cancers (Hui et al., 2004). Those transposable elements have been assumed to be noninfectious replication-defective retroviral fossils passed on during primate evolution. However, recent studies have revealed that at

Edited by Norihiro Okada
* Corresponding author. E-mail: khs307@pusan.ac.kr

least some members of the transposable element families are transcriptionally active and may be capable of causing human cancers through several mechanisms. The association of transposable elements with breast cancer (Wang-Johanning et al., 2003b), melanoma (Buscher et al., 2005), seminoma (Rakoff-Nahoum et al., 2006), germ cell tumors (Galli et al., 2005), leukemia(Depil et al., 2002), ovarian carcinomas (Menendez et al., 2004), and prostate adenocarcinoma (Wang-Johanning et al., 2003a) has been suggested. Therefore, transposable elements may play a biological role in cancer development or organismal complexity by introducing transcriptional diversity (Landry et al., 2003; Landry et al., 2003). The transposable elements in human cancers have been implicated in relation to immune disturbance, recombination excision, altering of gene structure, and abnormal expression. Here, we investigated the relationships between cancers and transposable elements. Moreover, we developed a database for understanding the mechanism of cancer development and analyzed gene expression in relation to transposable elements using human EST sequences.

MATERIALS AND METHODS

Data source The 23,188 mRNA sequences of human gene were downloaded from NCBI database Build 35.1. The useful EST information for tissues and pathology types was obtained from the eVOC ontology, a set of controlled vocabularies for unifying gene expression data (Kelso et al., 2003). Mobile elements in the human genome sequences were identified by RepeatMasker (http://repeatmasker.genome.washington.edu), and transposable element consensus sequences were identified by Repbase Update (Jurka, 2000). The Entrez gene database was used to identify the genomic location of references in RefSeq (Pruitt et al., 2005). RefSeq mRNA was obtained from the NCBI GenBank database. If RefSeq mRNA sequences overlapped, only the longest was considered.

EST classification Tissue source information for human EST libraries was completely examined to produce a consistent cancer and normal/unknown classification. The EST sequences were derived from NCBI's dbEST database that contains 8209 cDNA libraries (Boguski et al., 1993). The useful EST information for tissues and pathology types was obtained from the eVOC ontology, a set of controlled vocabularies for unifying gene expression data (Kelso et al., 2003). If pathology information was unclear, they were included in normal/unknown ESTs. Nevertheless, some of ESTs were excluded by reason of unclear assignment in cancer or normal.

Identification of transposable elements in human cancers by EST sequence analysis We used the publicly available human genome resources of mRNA and dbEST (database for expressed sequence tag) sequences from the INSDC databases (http://insdc.org). First of all, 23,188 mRNA sequences of human gene were downloaded from NCBI database Build 35.1 and aligned with the genomic DNA sequences (Build 35.1) using SIM4 program (Florea et al., 1998). Only alignments having > = 97% sequence identity were used in further stages. As a result, we extracted position information of exon and genome sequences to be matched. Based on this information, we extracted the contig sequences which have an additional 5 kb sequences from 5' UTR and 3' UTR end of genes, respectively. All the sequences were stored as mapping data for each gene. Alignments and EST clustering were produced by SIM4 and MegaBLAST program (Florea et al., 1998, Zhang et al., 2000) using mRNA, EST databases and the human genome assembly of NCBI Build 35.1. We set the criterion of EST sequences for appropriate data set as at least two different overlapping with genomic sequences and exonic region of functional gene. In order to investigate the fusion of transposable elements with functional genes in human cancers, the human expressed sequence tag (EST) and RefSeq mRNAs were screened by RepeatMasker using consensus sequences from Repbase Update (Jurka, 2000).

Screening for specific expressed transposable elements in human cancer Data validation about cancer specificity of fusion gene with transposable elements was performed using normal EST and cancer derived EST dataset. First of all, we divided all EST sequences deposited in the dbEST into three categories of normal EST, cancer EST, and unknown EST from eVOC ontology information (Kelso et al., 2003). Normal EST and unknown EST sequences from eVOC ontology information were used to screen for potential cancer associated transcripts and transposable elements fusion transcripts, which exclusively expressed in cancer tissues and cancer cell lines. We also visualized and validated all aspects of the genomic mapping of our dataset, gene structure, and splices sites by examining the entire feature in the genomic and EST sequences alignment.

RESULTS AND DISCUSSION

Large-scale analysis of fusion gene expression with transposable elements in various cancers using human EST sequences We analyzed fusion gene expression with transposable elements in human cancer tissues using human EST sequences. We also developed a screening procedure to identify putative cancer-specific fusion transcripts with transposable elements as shown in Fig. 1. In order to investigate the pathological influence of such fusion transcripts with transposable elements in human cancer ESTs, SIM4 and MegaBLAST

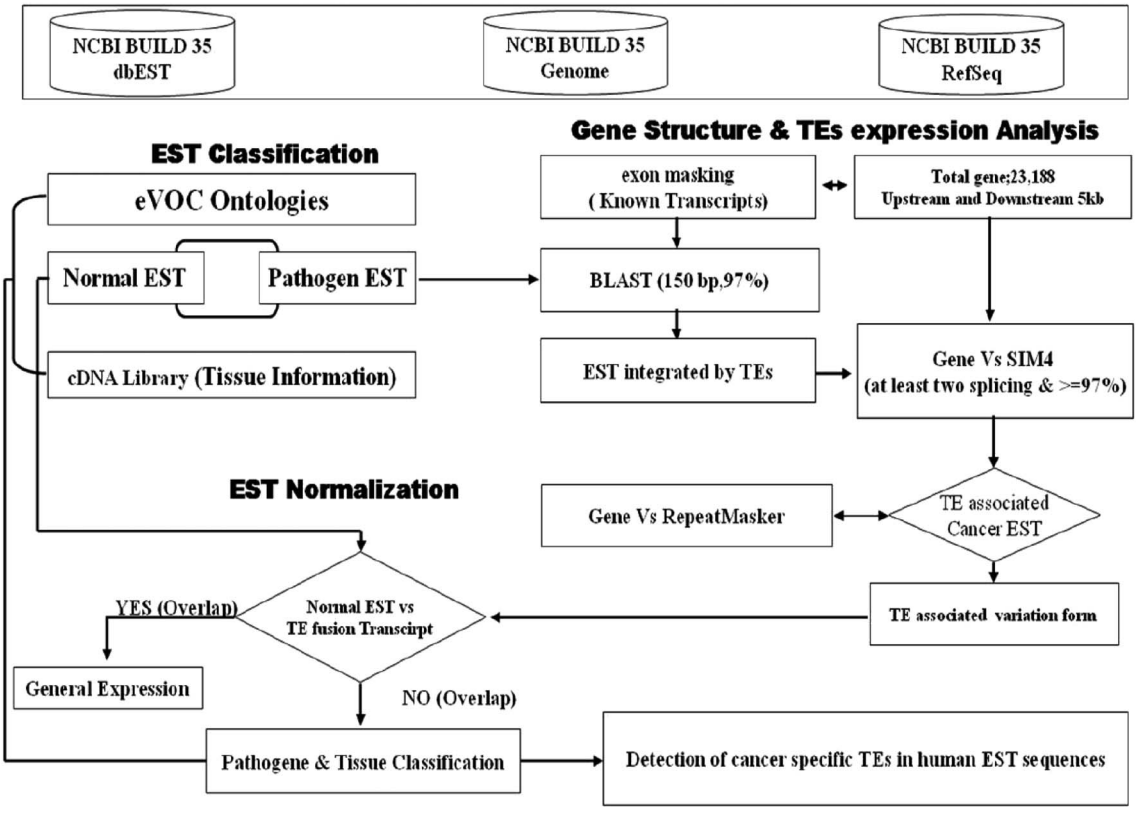

Fig. 1. A flow diagram showing the overall procedure for searching for the transposable elements-specific expression in human cancer ESTs (expressed sequence tags).

programs were used (Florea et al., 1998; Zhang et al., 2000) with the RefSeq mRNA dataset, and EST databases and the human genome were assembled using NCBI Build 35.1. We set the criterion of EST sequences for an appropriate dataset as at least two different overlapping regions with genomic sequences and exonic regions of a functional gene. More specifically, the process included mapping of EST consensus sequences to unique genomic locations and validation of intronic splice site sequences, and this process was conducted very carefully. The EST alignments showing > = 97% sequence identity were used for our analysis, and transposable elements fusion transcripts were identified by RepeatMasker (http://repeatmasker.genome.washington.edu) with transposable elements consensus sequences from Repbase Update (Jurka, 2000). We searched each fusion transcript with transposable elements using alignment information with transposable elements and human genes. As a result, our analysis identified 2798 transposable fusion genes among the 28171 genes; the fusion transcripts were created by integration of transposable elements that act as functional gene transcripts in human cancer tissues. In our previous study, the HERVH-*env* gene was expressed in various human cancer cells (Yi et al., 2006), and *in silico* analysis showed that transposable elements could affect protein sequences, splicing patterns, and expression of genes in human normal tissues or cancer tissues (Kim et al., 2005; Kim et al., 2006). Moreover, transposable elements were detected in several tumor cell lines including teratocarcinoma, bladder carcinomas, testicular tumors, and lung tumors (Wilkinson et al., 1990; Hirose et al., 1993). Most of the characterized transposable elements are defective, and include deletions and stop codons in internal regions. Occasionally, insertional mutations of transposable elements cause a genetic disease in human genes (Wilkinson et al., 1990; Lower et al., 1993; Kazazian, 1998; Armbruester et al., 2002) and also contribute to protein variability or versatility (Venables et al., 1995; Halling et al., 1999; Kjellman et al., 1999; Depil et al., 2002). These transposable elements are worth investigating for potential pathogenic effects related to various human cancers.

Validation of cancer-specific transposable element fusion transcripts Our analysis strategy was designed for searching transposable element exonizations that were not expressed in any normal tissues, but were expressed in cancer libraries in the form of fusion transcripts. Fig. 1 shows a schematic alignment of EST consensus sequences with the genomic sequences, and also shows transposable element fusion transcripts that were classified according to their cancer and normal EST

alignment results. We investigated the number of transposable element fusion transcripts showing cancer-specific expression patterns in 2798 genes using normal and unknown ESTs. In this result, we found that about 1799 transposable element fusion genes are detected inclusively in normal EST (expressed sequences tag), and 999 transposable element fusion genes were only detected in cancer tissues derived from ESTs. The 1799 transposable elements fusion genes might be important cause for the high frequency of alternative splicing in human genes. This result supports the observation that expression of transposable element fusion transcripts might be caused by the transposable elements (TEs) during both carcinogenesis and in normal tissues. A total of 999 genes of the 1329 transposable element fusion transcripts were predicted as cancer-specific alternative splicing forms created by transposable elements. The distribution of cancer-specific EST counts was represented by a percentage of transposable element fusion genes (Fig. 2). Among these genes, approximately 79.8% were represented by the EST singletons. When bioinformatic analysis was conducted for the investigation of alternative splicing, singleton EST sequences were considered to be insufficient data (Brett et al., 2000) because of the sequencing errors or genomic DNA contamination present in the EST database. However, singleton EST sequences could pro-

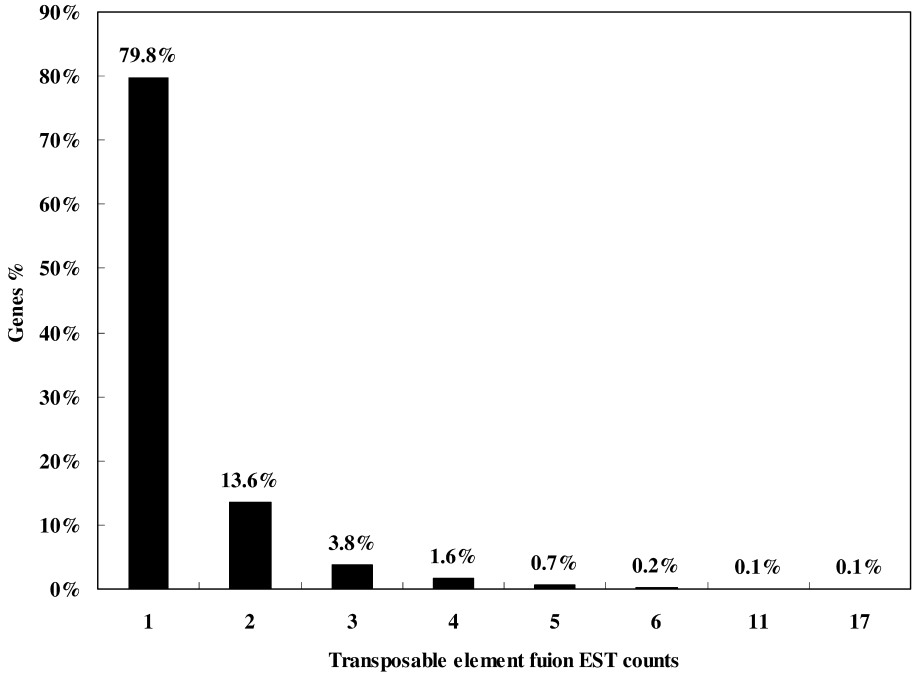

Fig. 2. Distribution of EST counts plotted to the percentage of transposable element fusion transcript variants.

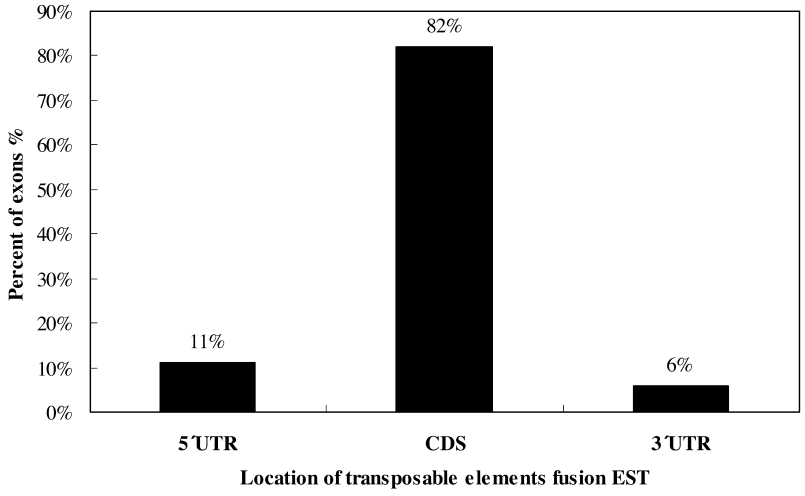

Fig. 3. Location of TEs fusion ESTs expression within the human genes.

vide enough evidence for transcript variants (Hui et al., 2004). On the basis of this report, the singleton EST was included in our analysis of cancer-specific transposable element fusion transcripts and database construction.

Most alternative splicing occurs within protein-coding regions (Pritsker et al., 2005). It has been suggested that alternative splicing is a general mechanism to increase the diversity of protein products or cause genetic disease result in splicing defects. To investigate the distribution of transposable element fusion events in the coding (CDS) and noncoding regions (UTR) of genes, the coding region was defined as an exon region and was subsequently extracted. In the case of transcript variants showing different coding regions, only the longest transcripts were selected for the dataset. Needless to say, in terms of the longest transcripts, multiple alignment anal-

Table 1. Distribution of exonization of transposable elements

Transposable elements fusion region within genes	Transposable elements				
	SINE Family	LINE Family	LTR Family	DNA Family	Others
CDS	619	280	85	76	1
5' UTR	76	30	33	5	0
3' UTR	44	20	14	5	0

*CDS: Coding sequence
*UTR: Untranslated Region
*SINE: Short Interspersed Nuclear Element
*LINE: Long Interspersed Nuclear Element
*LTR: Long Terminal Repeat

Table 2. Distribution of transposable elements detected in exon genes

Family	Subfamily	Transposable elements fusion in gene region		
		5UTR	CDS	3UTR
SINE Family	Alu	0	20	0
	AluJ	20	131	12
	AluS	13	190	15
	MIR	33	198	7
	FAM	0	2	0
	FRAM	0	16	2
	FLAM	7	25	3
LINE Family	HAL	0	11	0
	L1HS	0	1	0
	L1P	1	12	5
	L1M	6	125	6
	L2	22	111	7
	L3	1	20	2
LTR Family	MaLR	16	40	6
	ERV1	13	23	3
	ERVL	4	16	5
	ERVK	0	6	0
DNA Family	Charlie	0	9	0
	HSMAR2	0	2	0
	Kanga1	0	0	1
	MARNA	0	3	0
	MER	5	50	3
	Tigger	0	11	1
	Zaphod2	0	1	0
Others	Charlie	0	1	0

Table 3. EST based expression profiles of transposable elements expressed in human cancer

EST library	Transposable elements fusion transcript NO	Tissue	Total
Cervical carcinoma	7	Uterus	82
Cervix	1		
Endometrial	6		
Endometrium adenocarcinoma	7		
Follicular lymphoma	1		
Juvenile granulosa tumor	2		
Moderately-differentiated endometrial	5		
Ovarian cancer	4		
Ovarian tumor	27		
Poorly-differentiated endometrial	1		
Small cell carcinoma	10		
Uterine	2		
Uterus tumor	8		
Carcinoid	47	Lung	116
Chondrosarcoma Lung Metastasis cell	8		
Large cell carcinoma	17		
Lung carcinoma	2		
Lung Focal Fibrosis	9		
Lung tumor	9		
Primary Lung Cystic Fibrosis Epithelial	14		
Squamous cell carcinoma	10		
Adenocarcinoma	48	Breast	110
Breast	21		
Ductal carcinoma	2		
Mammary adenocarcinoma	11		
Mammary gland tumor	12		
Moderately-differentiated adenocarcinoma	5		
Papillary serous carcinoma	2		
Papillary serous ovarian metastasis	1		
Serous papillary carcinoma	3		
Transitional cell papilloma	5		
Acute myelogenous leukemia	3	Blood	46
B CELLS (RAMOS CELL LINE)	2		
B-cell chronic lymphotic leukemia	20		
Burkitt lymphoma	2		
Leukopheresis	6		
Lymphoma	12		
T CELLS (JURKAT CELL LINE)	1		
Colon	6	Colon	69
Colon tumor	37		
Colon tumor RER+	19		
Colon_est	2		
Colon_ins	5		
Hepatocellular carcinoma	20	Liver	56
Liver	36		
Carcinoma in situ from retromolar trigone	1	Head neck	64
Denis_drash	2		
Head neck	58		
Mucoepidermoid carcinoma	3		
Embryonal carcinoma	12	Embryo	15
Germ cell tumor	3		
Pooled	44	Others	72

Table 3. Continued

EST library	Transposable elements fusion transcript NO	Tissue	Total
Bone marrow	7	Bone	59
Ewing's sarcoma	1		
Marrow	27		
Metastatic Chondrosarcoma	3		
Myeloma	6		
Osteoarthritic Cartilage	9		
Osteosarcoma	6		
Anaplastic oligodendroglioma	25	Brain	70
Astrocytoma	1		
Astrocytoma grade IV	6		
Brain glioblastoma	3		
Glioblastoma	32		
Medulloblastoma	2		
Schizophrenic brain S-11 frontal lobe	1		
Epidermoid	5	Skin	91
Epidermoid carcinoma	23		
Malignant melanoma	1		
Melanoma	2		
Melanotic melanoma	59		
Skin tumor	1		
Chondrosarcoma	92	Cartilage	107
Chondrosarcoma Grade II	8		
Enchondroma	7		
Adrenal adenoma	1	Kidney	42
Adrenal cortex carcinoma	4		
Adrenal gland	1		
Adrenal tumor	1		
Insulinoma	24		
Kidney tumor	8		
Renal cell adenocarcinoma	3		
Alveolar rhabdomyosarcoma	1	Muscle	52
Fibrosarcoma	5		
Fibrotheoma	2		
Leiomyosarcoma	21		
Multiple sclerosis lesions	9		
Rhabdomyosarcoma	14		
Nervous tumor	12	Nervous	52
Neuroblastoma	39		
Schwannoma tumor	1		
Amelanotic melanoma	4	Viscera (Stomach,Heart, Pancreas) Eye	130
Ascites	56		
Duodenal adenocarcinoma	12		
Heart	2		
Pancreas	2		
Stomach	54		
Cornea	23	Eye	41
Retinoblastoma	18		
Bladder tumor	5	Genitalia	18
Prostate tumor	7		
Teratocarcinoma	6		
Parathyroid tumor	12	Thyroid	15
Thyroid	3		

ysis of different transcript variants was conducted. Many kinds of transposable elements are expressed within the CDS region of human transcripts in the form of a fusion transcript with a cellular functional gene (Fig. 3). Among them, 82% of fusion transcript were located within the protein coding region, and 11% and 6% were located within the 5' UTR and 3' UTR end, respectively. In a previous study, internal exons containing *Alu* appeared frequently within the CDS region of human transcripts (Sorek et al., 2002). Additionally, we analyzed the detailed distribution of different transposable elements (SINE, LINE, HERV, and DNA elements) in coding and non-coding regions. As shown in Table 1, the SINE family shows a unique pattern of integration into the coding regions (CDS) and non-coding regions of genes. The SINE family shows higher levels of enrichment than the other transposable elements. Of these, 619 (48.0%) SINE-fusion transcripts, 280 (21.7%) LINE-fusion transcripts, 85 (6.6%) LTR-fusion transcripts, 76 (5.9%) DAN-fusion transcripts were located within the protein coding region. Subfamilies of the transposable elements were also checked for the detection of specific fusion tendency (Table 2). Old subfamilies of transposable elements showed a higher retention ratio tendency

Table 4. Distribution of transposable elements within cancer specific expression transcripts

Transposable elements		Occurrences	Percent (%)
Family	Subfamily		
SINE	Alu	20	1.44
	AluJ	171	12.35
	AluS	244	17.62
	MIR	250	18.05
	FAM	2	0.14
	FRAM	18	1.30
	FLAM	37	2.67
LINE	HAL	13	0.94
	L1HS	1	0.07
	L1P	18	1.30
	L1M	153	11.05
	L2	151	10.90
	L3	25	1.81
LTR	MaLR	67	4.84
	ERV1	40	2.89
	ERVL	27	1.95
	ERVK	6	0.43
DNA	Charlie	9	0.65
	HSMAR2	2	0.14
	Kanga1	1	0.07
	MARNA	3	0.22
	MER	61	4.40
	Tigger	14	1.01
	Zaphod2	1	0.07
Others	Charlie	1	0.07

Table 5. Potential splice site utilized by transposable elements fusion exons

Type of potential splicing site	Transposable elements			
	SINE Family	LINE Family	LTR Family	DNA Family
Accept&Donor	83	68	50	12
Accept Site	271	110	33	28
Donor Site	216	80	43	18

than young subfamilies of transposable elements. In general, most transposable elements present in the human genome are defective due to the accumulation of mutations. The bias toward old subfamilies in the set of transposable element fusion transcripts may reflect that the accumulation of many substitutions was necessary to create a functional splice site within the transposable elements sequence to allow for its exonization. In other words, original characters of transposable element could be changed from transposable elements to non-transposable elements recognized by the human genome as a cellular component for gene-like evolution (Sorek et al., 2002, Huh et al., 2006). The mechanism of gene-like evolution derived from an old transposable element could be one of the main sources of human genome diversification through the use of limited resources. Transposable elements could have a strong impact on the evolution of the human genome by providing unlimited resources for human gene diversity.

Transposable elements integrated into human expressed sequences We investigated how transposable elements are integrated into human expressed sequences during tumorigenesis. From our analysis of 7199 transposable element fusion transcripts, we found that 1329 transposable element fusion transcripts were detected exclusively in human cancer tissues (Table 3). The distribution of transposable elements was found to be strongly correlated with time since evolutionary divergence (Table 4). As shown in Table 4, the ratio of transposable element subfamily expression is higher in old subfamilies of transposable elements than in young subfamilies. As shown in Table 4, among various hybrid transcripts between transposable elements and cancer-associated ESTs, we were able to identify 1329 transposable elements in 22 cancer tissues, indicating that cancer-specific transcripts variants could be derived from transposable elements. We also confirmed that canonical splicing sites were offered by transposable elements (Table 5). The splice sites utilized by the transposable element fusion transcripts were counted, and our counts indicated that transposable elements are capable of driving gene expression during tumorigenesis.

GO analysis of transposable element fusion genes in ESTs originated from human cancers To test whether there is selective association of human cancers with transposable elements, we initially classified the genes of 999 transposable element fusion genes into functional GO (Gene Ontology) categories. The GO describes gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner. We then calculated the density in each functional category. Analysis of transposable element fusion genes between the functional categories revealed that the majority of the genes have functions that are associated with cancer (receptor, DNA binding, kinase activity) (Fig. 4), suggesting that their transposable elements may play roles in tumorigenesis.

Database construction of transposable element-specific expression in human cancers (TECESdb) Our database was created as a biological database using the MySQL database management system, and the data

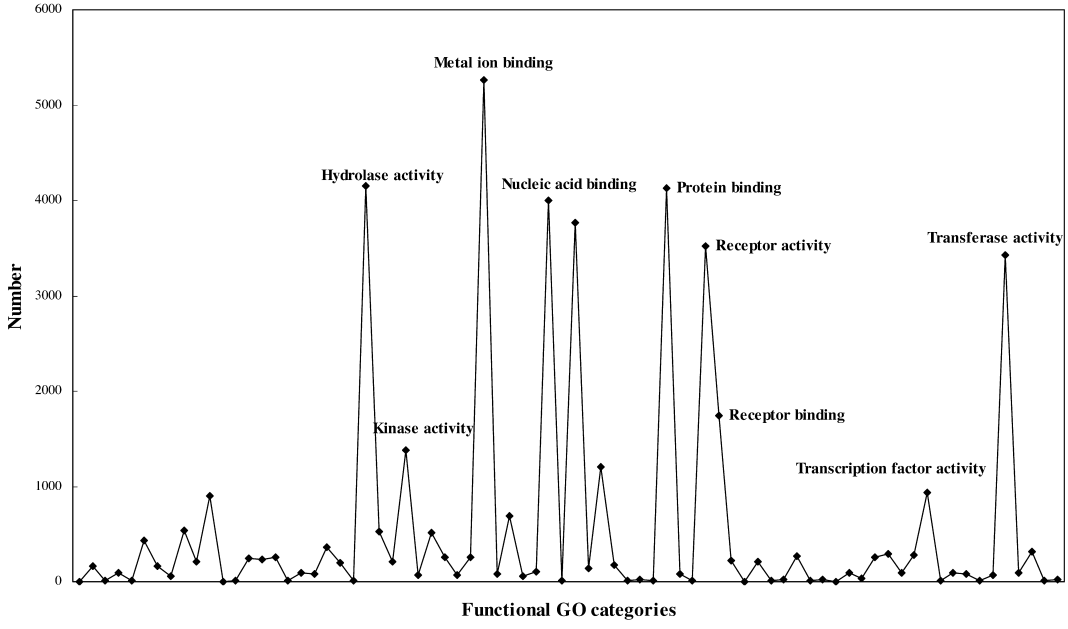

Fig. 4. Classification of the transposable element fusion genes into GO functional categories.

used was obtained from a primary. Users can efficiently retrieve information concerning transposable elements expression within various cancer tissues. The TECESdb can be accessed through a CGI-Python base web interface. The TECESdb web interface provides access to the database contents in four ways through the search database option (Fig. 5A). First, users can search the genes of interest using the accession numbers of the NCBI data bank, or can search using the HUGO symbol name provided on the view page. EST sequences can even be accessed from the NCBI data bank for further studies. Second, users can search for TE expression in a cancer-associated, expressed sequence by clicking on a gene listed on the view page according to chromosome number. Third, it is possible for users to view the results of this search by clicking on pathology information. Fourth, the database provides the type information in which TE expression is classified to four types (SINE class, LINE class, LTR class, DNA class). The type information may make it easier for users to speculate regarding the effects of TE expression within interesting genes.

The result pages are listed in a tabular format that provides evidence and information regarding TE expression within the genes. The results are presented in two different sections, one on the detailed information about TE expression and the other on the graphic viewer. The graphic viewer shows expression of TE-fused transcripts in a human gene, and is represented by the exon-intron splicing structure of mRNAs/ESTs. Moreover, this viewer provides a highlighted viewer indicating the TE fusion region in a purple color within the mRNA structure. Users can also visualize the normalized EST information around the TE fusion regions by using the normalized EST image viewer in the result page (Fig.

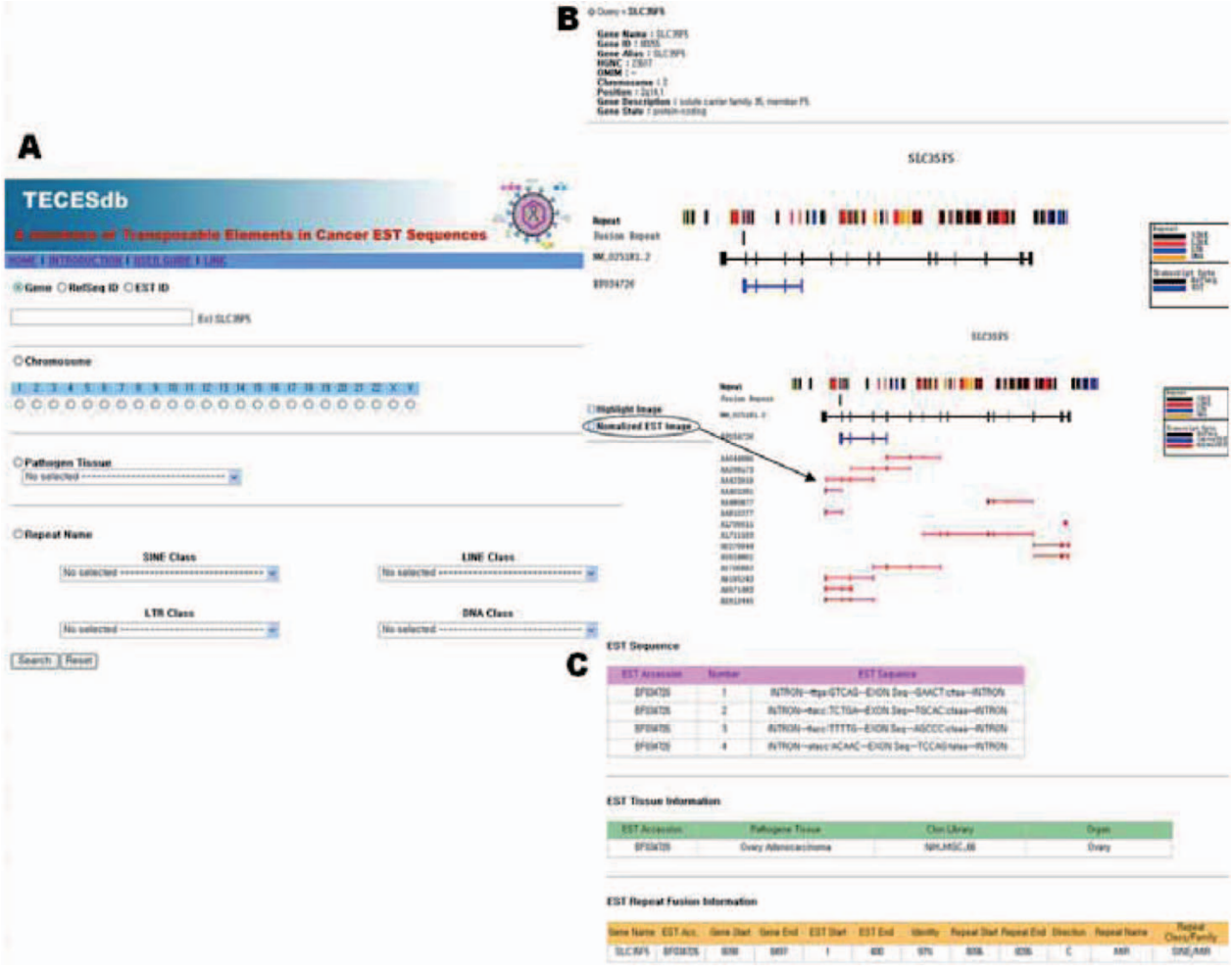

Fig. 5. A snapshot of TECESdb interface showing the advanced search page and the results of search query. (A) Search options for TECESdb. Users can query of a gene by gene name or GenBank ID or UniGene ID or Gene ID. Advanced options enables user search for TEs fusion gene by chromosome and by TEs class. (B) Showing SLC35F5 as an example. This viewer provides graphic view that represent about the TEs fusion region, transcripts orientation, gene structure. Users also can see the normalized EST information in the TEs fusion regions by using the normalized EST image viewer in the result page. (C) The result page includes not only TEs fusion information, but also tissue, pathology, clone library, organ information of the target gene.

5B). In addition, the result page includes not only TE fusion information, but also shows tissue, pathology, clone library, and organ information of the target gene (Fig. 5C). Available at http://www.primate.or.kr/TECESdb.

CONCLUSION

Most transposable elements have been inserted into new genomic locations, and have been truncated and rearranged as inactive copies of the active progenitor elements. Occasionally, these insertional mutations cause a genetic disease, and also contribute to protein variability or versatility. The purpose of TECESdb is to list all of the possible transposable element fusion genes that can be predicted and annotated by current technology. The system could identify transposable element-related exonizations that are not expressed in any normal tissues, but are expressed in only cancer libraries in the form of fusion transcripts. We found that 999 genes were detected in the mRNA sequences with transposable elements. These transposable elements are worth investigating as potential pathogenic elements for various human cancers. The database is being constantly updated and supplemented with new human gene databases from the available sources. Through the continuous update, we will do profile the expression patterns of transposable elements in various cancers and the systematic relationships between transposable elements and neighboring functional genes. We believe that our work will help us to gain insight into implication of transposable element fusion genes in human evolution and diseases.

This study was supported by a grant from the National R&D Program for Cancer Control, Ministry of Health and Welfare, Republic of Korea (0620150-1).

REFERENCES

Armbruester, V., Sauter, M., Krautkraemer, E., Meese, E., Kleiman, A., Best, B., Roemer, K., and Mueller-Lantzsch, N. (2002) A novel gene from the human endogenous retrovirus K expressed in transformed cells. Clin. Cancer Res. **8**, 1800–1807.

Boguski, M. S., Lowe, T. M., and Tolstoshev, C. M. (1993) dbEST–database for "expressed sequence tags". Nat. Gene. **4**, 332–333.

Brett, D., Hanke, J., Lehmann, G., Haase, S., Delbruck, S., Krueger, S., Reich, J., and Bork, P. (2000) EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. FEBS Lett. **474**, 83–86.

Buscher, K., Trefzer, U., Hofmann, M., Sterry, W., Kurth, R., and Denner, J. (2005) Expression of human endogenous retrovirus K in melanomas and melanoma cell lines. Cancer Res. **65**, 4172–4180.

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

Depil, S., Roche, C., Dussart, P., and Prin, L. (2002) Expression of a human endogenous retrovirus, HERV-K, in the blood cells of leukemia patients. Leukemia **16**, 254–259.

Dunn, C. A., Medstrand, P., and Mager, D. L. (2003) An endogenous retroviral long terminal repeat is the dominant promoter for human beta 1,3-galactosyltransferase 5 in the colon. Proc. Natl. Acad. Sci. USA. **100**, 12841–12846.

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.

Galli, U.M., Sauter, M., Lecher, B., Maurer, S., Herbst, H., Roemer, K., and Mueller-Lantzsch, N. (2005) Human endogenous retrovirus rec interferes with germ cell development in mice and may cause carcinoma in situ, the predecessor lesion of germ cell tumors. Oncogene **24**, 3223–3228.

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.

Huh J.W., Kim, T. H., Yi, J. M., Park, E. S., Kim, W. Y., Sin, H. S., Kim, D. S., Min, D. S., Kim, S. S., Kim, C. B., Hyun, B. H., Kang, S. K., Jung, J.S., Lee, W.H., Takenaka, O., and Kim, H.S. (2006) Molecular evolution of the periphilin gene in relation to human endogenous retrovirus M element. J. Mol. Evol. **62**, 730–737.

Hui, L., Zhang, X., Wu, X., Lin, Z., Wang, Q., Li, Y., and Hu, G. (2004) Identification of alternatively spliced mRNA variants related to cancers by genome-wide ESTs alignment. Oncogene **17**, 3013–3023.

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

Kazazian, Jr. H. H. (1998) Mobile elements and disease. Curr. Opin. Genet. Dev. **8**, 343–350.

Kazazian, H. H Jr. (2004) Mobile elements: drivers of genome evolution. Science **303**, 1626–1632.

Kelso, J., Visagie, J., Theiler, G., Christoffels, A., Bardien, S., Smedley, D., Otgaar, D., Greyling, G., Jongeneel, C. V., McCarthy, M. I., Hide, T., and Hide, W. (2003) eVOC: a controlled vocabulary for unifying gene expression data. Genome Res. **13**, 1222–1230.

Kim, D. S., Kim, T. H., Huh, J. W., Kim, I. C., Kim, S. W., Park, H.S., and Kim, H.S. (2006) LINE FUSION GENES: a database of LINE expression in human genes. BMC Genomics. **7**, 139.

Kim, T. H., Jeon, Y.J., Kim, W. Y., and Kim, H. S. (2005) HESAS: HERVs Expression and Structure Analysis System. Bioinformatics **21**, 1699–700.

Hirose, Y. M., Takamatsu, M., and Harada, F. (1993) Presence of env genes in members of the RTVL-H family of human endogenous retrovirus-like elements. Virology **192**, 52–61.

Kjellman, C., Sjogren, H. O., Salford, L. G., and Widegren, B. (1999) HERV-F (XA34) is a full-length human endogenous retrovirus expressed in placental and fetal tissues. Gene **239**, 99–107.

Landry, J. R., Mager, D. L., and Wilhelm, B. T. (2003) Complex controls: the role of alternative promoters in mammalian genomes. Trends Genet. **19**, 640–648.

Landry, J. R., and Mager, D. L. (2003) Functional analysis of the endogenous retroviral promoter of the human endothelin B receptor gene. J. Virol. **13**, 7459–7466.

Lin, L., Zhuwen, W., Michael, S. P., Herman, V. D., Dafydd, G. T., Thomas, J. G., Andrew, C. C., Mark, B. O., Stephen, B. G., John, V. M., Thomas, D. G., Giordano, T. J., Chang, A. C., Orringer, M. B., Gruber, S. B., Moran, J. V., Glover, T. W., and Beer, D. G. (2006) Multiple forms of genetic instability within a 2-Mb chromosomal segment of 3q26.3–q27

are associated with development of esophageal adenocarcinoma. Genes Chrom. Cancer. **45**, 319–331.

Lower, R., Lower, J., and Kurth, R. (1996) The viruses in all of us: characteristics and biological significance of human endogenous retrovirus sequences. Proc. Natl. Acad. Sci. USA. **93**, 5177–5184.

Lower, R., Lower, J., Tondera-Koch, C., and Kurth, R. (1993) A general method for the identification of transcribed retrovirus sequences (R-U5 PCR) reveals the expression of the human endogenous retrovirus loci HERV-H and HERV-K in teratocarcinoma cells. Virology **192**, 501–511.

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.

Medstrand, P., Landry, J. R., and 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.

Menendez, L., Benigno, B. B., McDonald, J. F. (2004) L1 and HERV-W retrotransposons are hypomethylated in human ovarian carcinomas. Mol. Cancer. **3**, 12–12.

Mi, S., Lee, X., Li, X., Veldman, G. M., Finnerty, H., Racie, L., LaVallie, E., Tang, X. Y., Edouard, P., Howes, S., Keith, J. C., and McCoy, M. (2000) Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. Nature **403**, 785–789.

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

Nigumann, P., Redik, K., Matlik, K., and Speek, M. (2002) Many human genes are transcribed from the antisense promoter of L1 retrotransposon. Genomics **79**, 628–634.

Okumura, M., Kondo, S., Ogata, M., Kanemoto, S., Murakami, T., Yanagida, K., Saito A., and Imaizumi K. (2005) Candidates for tumor-specific alternative splicing. Biochem. Biophys. Res. Commun. **334**, 23–29.

Pritsker, M., Doniger, T. T., Kramer, L. C., Westcot, S. E., and Lemischka, I. R. (2005) Diversification of stem cell molecular repertoire by alternative splicing. Proc. Natl. Acad. Sci. USA. **102**, 14290–14295.

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.

Rakoff-Nahoum, S., Kuebler, P. J., Heymann, J. J., Sheehy, M. E., Ortiz, G. M., Ogg, G. S., Barbour, J. D., Lenz, J., Steinfeld, A. D., and Nixon, D. F. (2006) Detection of T lymphocytes specific for human endogenous retrovirus K (HERV-K) in patients with seminoma. AIDS Res. Hum. Retroviruses. **22**, 52–56.

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.

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

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

Venables, P. J., Brookes, S. M., Griffiths, D., Weiss, R. A., and Boyd, M. T. (1995) Abundance of an endogenous retroviral envelope protein in placental trophoblasts suggests a biological function. Virology **211**, 589–592.

Wang-Johanning, F., Frost, A. R., Jian, B., Epp, L., Lu, D. W., and Johanning, G. L. (2003b) Quantitation of HERV-K env gene expression and splicing in human breast cancer. Oncogene **22**, 1528–1535.

Wang-Johanning, F., Frost, A. R., Jian, B., Azerou, R., Lu, D. W., Chen, D. T., and Johanning, G. L. (2003a) Detecting the expression of human endogenous retrovirus E envelope transcripts in human prostate adenocarcinoma. Cancer **98**, 187–197.

Wilkinson, D. A., Freeman, J. D., Goodchild, N. L., Kelleher, C. A., and Mager, D. L. (1990) Autonomous expression of RTVL-H endogenous retroviruslike elements in human cells. J. Virol. **64**, 2157–2167.

Yi, J. M., and Kim, H. S. (2004) Expression analysis of endogenous retroviral elements belonging to the HERV-F family from human tissues and cancer cells. Cancer Lett. **211**, 89–96.

Yi, J. M., Kim, H. M., and Kim, H. S. (2006) Human endogenous retrovirus HERV-H family in human tissues and cancer cells: expression, identification, and phylogeny. Cancer Lett. **231**, 228–239.

Yulug, I. G., Yulug, A., and Fisher, E. M. (1995) The frequency and position of Alu repeats in cDNAs, as determined by database searching. Genomics **27**, 544–548.

Zhang, Z., Schwartz, S., Wagner, L., and Miller, W. (2000) A greedy algorithm for aligning DNA sequences. J. Comput. Biol. **7**, 203–214.