Genomics & Informatics Vol. 5(4) 179-187, December 2007

REPEATOME: A Database for Repeat Element Comparative Analysis in Human and Chimpanzee

Taeha Woo[1,2,¶], Tae-Hui Hong[1,¶], SangSoo Kim[2], Won-Hyong Chung[1], Hyo Jin Kang[1], Chang-Bae Kim[1] and Jungmin Seo[1,*]

[1]Korean BioInformation Center, KRIBB, Daejeon 305-806, Korea, [2]Department of Bioinformatics, Soongsil University, 1-1 Sangdo-dong, Dongjak-Gu, Seoul 156-743, Korea

Abstract

An increasing number of primate genomes are being sequenced. A direct comparison of repeat elements in human genes and their corresponding chimpanzee orthologs will not only give information on their evolution, but also shed light on the major evolutionary events that shaped our species. We have developed REPEATOME to enable visualization and subsequent comparisons of human and chimpanzee repeat elements.

REPEATOME (http://www.repeatome.org/) provides easy access to a complete repeat element map of the human genome, as well as repeat element-associated information. It provides a convenient and effective way to access the repeat elements within or spanning the functional regions in human and chimpanzee genome sequences. REPEATOME includes information to compare repeat elements and gene structures of human genes and their counterparts in chimpanzee. This database can be accessed using comparative search options such as intersection, union, and difference to find lineage-specific or common repeat elements.

REPEATOME allows researchers to perform visualization and comparative analysis of repeat elements in human and chimpanzee.

Availability: REPEATOME is freely available at http://www.repeatome.org. The web interface of REPEATOME is supported with JAVA and java scripts that enable formulation of queries against the database. Results are displayed either in tabulated or graphical formats.

Keywords: Repeat element; repeat map; comparative genomics; human repeat; chimpanzee repeat; mobile element

Background

The human genome has acquired a variety of repeat elements through successive retrotranspositions over the past 60 million years of evolution (Batzer and Deininger, 2002). Among these repeat elements, Alu, L1, and SVA are known to be the most prevalent transposon families that have retroposed throughout primate evolution (Smit *et al*., 1995, The Chimpanzee Sequencing and Analysis Consortium, 2005). Together, these mobile elements account for over 30% of the DNA in the human genome (Lander *et al*., 2001). Alu retrotransposition depends on a reverse transcriptase encoded by active L1 retroelements. These insertions may have contributed to the differential evolution of humans and chimpanzees (The Chimpanzee Sequencing and Analysis Consortium, 2005, Mills *et al*., 2006).

Repeat elements such as Alu have previously been considered as junk DNA with no function (Ohno, 1991). However, recent studies suggested that repeats residing within coding regions, promoter regions, introns, untranslated regions (UTRs), and CpG islands are all involved in various functions (Dagan *et al*., 2004, Le *et al*., 2003, Grover *et al*., 2005, Amaud *et al*., 2000, Thornburg *et al*., 2006).

Several existing databases support the identification of repeat sequences. For example, Repbase (Jurka, 2005) is a well-known nucleotide-sequence-repeat database that uses the RepeatMasker program (http://www.repeatmasker.org). A set of repeats can also be retrieved from the public University of California, Santa Cruz (UCSC) Genome Browser (http://genome.cse.ucsc.edu) (Hinrichs *et al*., 2006), but it does not provide extensive repeat information for comparative genome analysis. Other species-specific databases include FREP (Nagashima *et al*., 2004), which stores the functional repeats in mouse cDNAs, and TRbase, which stores tandem repeats related to disease genes in the human genome (Boby *et al*., 2005). However, these databases have not provided in-depth comparative data on repeat elements in the genomes of multiple species.

In this report, we describe a new database of human and chimpanzee repeat elements, called REPEATOME.

[¶]These authors contributed equally to this work.

*Corresponding author: E-mail jmseo@kribb.re.kr
Tel +82-42-879-8531, Fax +82-42-879-8519

Accepted 2 September 2007

This database aims to build a comprehensive repeat element database to provide easy access to the repeat elements-as well as an additional 5kb of flanking DNA of the entire human and chimpanzee genomes. REPEATOME contains all human RefSeq genes (Pruitt *et al*., 2005) from the UCSC Genome Browser that contain repeat sequences, and provides putative repeat elements for both sets of genomes. It should therefore be a valuable resourcefor understanding primate evolution.

Construction & Contents

Database construction

REPEATOME is a relational database built using MySQL (http://www.mysql.com/). The overview of data sources and flows for REPEATOME is described in Fig. 1. The construction steps are composed of 1) data extraction, 2) alignment of human and chimpanzee sequences, 3) gene annotation, 4) categorization of functional regions, and 5) screening of repeat elements.

Data extraction

The location of each gene and its surrounding sequences were determined using the UCSC Genome Browser May 2004 (hg17) and March 2006 (hg18) assemblies for human and the November 2003 (panTro1) and March 2006 (panTro2) assemblies for chimpanzee. We downloaded the sequence and annotation data for hg17, hg18, panTro1, and panTro2 from the UCSC Genome Browser FTP server. The extracted genome sequences spanned all RefSeq genes from the start to the end of the annotation, plus an additional 5 kb upstream and downstream of the start and end, respectively. We used refGene tracks from the UCSC database for gene information.

Alignment of human and chimpanzee sequences

Human and chimpanzee DNA sequences including 5 kb both upstream and downstream were obtained from the UCSC Genome Browser database (http://genome.ucsc.edu), and chimpanzee genome sequences were aligned against human genome sequences using BLAT (Blast-like Alignment Tool) (Kent, 2002), setting a minimum threshold of 98% sequence identity. We selected the comparison with the highest alignment score, which gave us a final working set of 19,577 (hg17/panTro1) and 23,588 (hg18/panTro2) non-redundant alignments for human and chimpanzee, respectively. The alignment was then parsed with a Perl script and stored in a MySQL database for further analysis.

Gene annotation

The mRNA records in refGene were linked to other genetic databases through various accession keys provided by NCBI RefSeq, GI number, and gene name. In addition, each mRNA record was identified by its genome location, repeat name, repeat family, and repeat class. Intron, exon, and promoter records were linked to their corresponding mRNAs through the RefSeq ID. We also mapped mRNA records to the GO, OMIM, and EntrezGene (Maglott *et al*., 2005) databases.

Categorization of potential functional regions in genome sequences

Potential functional elements in human genome sequences were categorized into functional regions: promoter regions, CpG islands, 5' UTRs, translation start sites, splice sites, coding exons, introns, translation stop sites, polyadenylation signal sites, 3' UTRs, and 5-kb downstream regions. In this study, we used RefSeq gene (refGene) tracks from the UCSC database to obtain gene boundary information.

i. Promoter regions

We used the MATCH program in TRANSFAC Professional 8.4 (Matys *et al*., 2006), which is a weight-matrix-based tool for searching putative transcription factor binding sites in DNA sequences, to predict transcription factor binding sites in 5-kb upstream sequences (promoter regions). We used the “minimizing false-positive errors” option for cut-offs and the “high-quality vertebrate” option for the grouping of matrices.

ii. CpG islands

CpG islands are associated with specific genes and are common near transcriptional start sites in vertebrates. CpG islands in the UCSC database are predicted by searching for a sequence and scoring each dinucleotide. We downloaded the complete tracks and filtered them out only if they were located within the 5-kb upstream sequence.

iii. Splice sites

We considered the first and last 2 base pairs (bp) in introns as splice sites (Zhang, 1998). The positions of introns were obtained from refGene tracks in the UCSC database.

iv. 5' UTRs, 3'UTRs, coding exons, introns, start codon, and stop codon

Position information of exons, intron boundaries, and coding sequence (CDS) boundaries is present in refGene tracks in the UCSC database (Pruitt *et al*., 2005). The 5' UTRs are the portion of sequences that are upstream of the CDS start sites, whereas the 3'UTRs are downstream of the CDS stop sites. In this study, coding exons were considered as CDSs within the transcripts, meaning that 5' UTRs are not coding exons. On the contrary, introns are non-transcribed sequences from DNA sequences, and hence can exist between UTRs and coding exons. The translational start and stop sequences are the first and last 3 bp of CDSs, respectively.

v. Polyadenylation signal sites

In this study, we used ERPIN (Lambert *et al*., 2004) to identify polyadenylation signal sites in genes. ERPIN was downloaded from the internet and installed on a local PC. The training set file was also downloaded from the website, and the query database that contained only 3' UTRs was extracted from refGene tracks in the UCSC database (Pruitt *et al*., 2005).

Screening repeat elements

There are two kinds of repeat data in REPEATOME: (i) those obtained from repeat tracks from the UCSC database and (ii) those scanned on a local server using the RepeatMasker program. First, the RepeatMasker annotations of the human (hg17, May 2004 freeze; hg18, March 2006) and chimpanzee (panTro1, November 2003 freeze; panTro2, March 2006) genomes were obtained from the UCSC Genome Bioinformatics Site. Second, repeat element data were created using the RepeatMasker (version open-3.1.5) program installed on a local PC, which screens DNA sequences for interspersed repeats and low-complexity DNA sequences. Sequences from Repbase Update (version 11.01, released on March 2006; http://www.girinst.org/Repbase_Update.html) were used to screen and annotate repetitive elements in human and chimpanzee mRNA sequences, including those 5 kb upstream and 5 kb downstream using the RepeatMasker program. All repeats detected by the program were collected in a MySQL database.

The distribution of repeat elements, over the functional regions of human and chimpanzee genomes, is shown in Table 1. The analysis of the human repeat data showed a general increase in the frequency of repeat elements in the functional regions compared with those of chimpanzee. However, the repeat elements in chimpanzee were likely to possess more splice sites than human elements.

Table 1. Summary statistics of REPEATOME

Data Source	hg17/ panTro1	hg18/ panTro2
RefSeq*	23,821/20,827	25,066/24,966
Functional Regions		
Upstream 5 kb region**	215,325/184,333	215,948/201,937
5' UTRs**	1,919/1,245	1,934/1,490
3' UTRs**	12,547/9,552	12,552/11,221
CpG island**	2,961/1,437	2,934/1,939
Coding exon**	2,701/1,889	2,714/2,184
Intron**	2,195,406/1,933,787	2,201,797/2,208,353
Downstream 5 kb region**	184,373/161,707	184,805/175,133
PolyA Signal***	1,851/1,747	1,840/1,794
Splice sites***	4,319/4,699	4,282/6,088
Start***	435/286	389/365
Stop***	509/366	502/499

* : Numbers of RefSeq genes

** : Numbers of repeat elements within each functional regions

*** : Numbers of functional regions in repeat elements

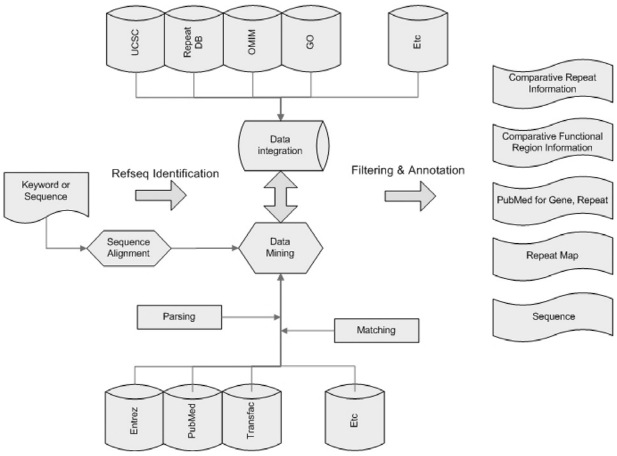


Fig. 1. REPEATOME data sources and construction flows.

Results & Discussion

Data retrieval

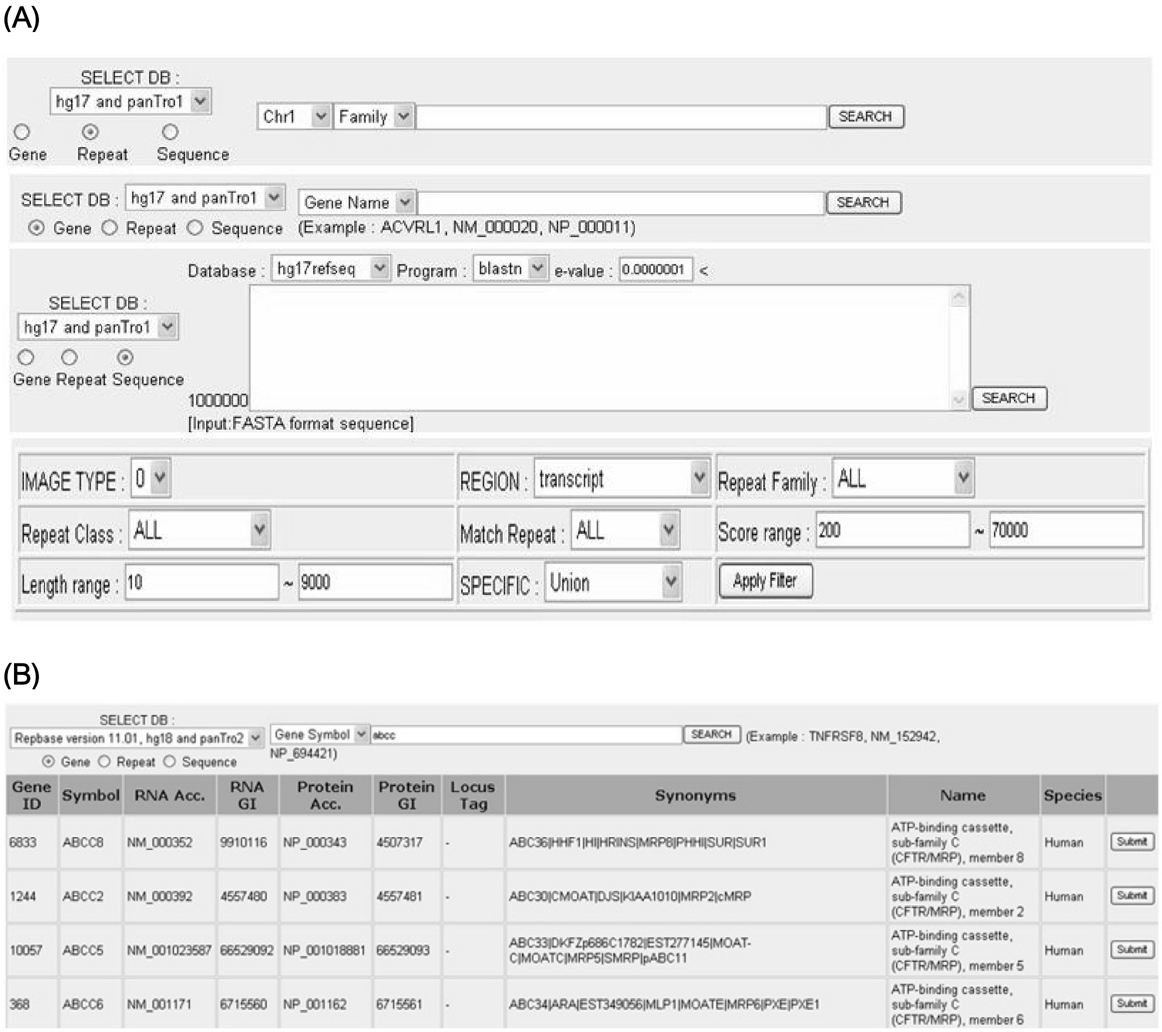
The REPEATOME database provides a set of forms through which researchers can easily access, select, and compare repeats of interest. The REPEATOME database can be accessed at http://www.repeatome.org. It supports various types of queries, such as gene name, mRNA accession number, protein accession number, repeat family name, repeat subfamily name, and repeat element name (Fig. 2A). In addition, a BLAST search (Altschul *et al*., 1990) with a user-defined query sequence can be performed against a certain genome version (hg17, hg18, panTro1, and panTro2). The search-results page shows the gene list that matchesthe query (Fig. 2B).

The results page shows a gene annotation containing the repeat elements, and includes extensive links to relevant resources such as the EntrezGene (Maglott *et al*., 2005), OMIM, GO, and PubMed (McEntyre *et al*., 2001) databases (Fig. 2C). Most repeat databases, except for Repbase, do not provide journal publications, whereas each of these repeat annotations in REPEATOME is accompanied by its PubMed reference link. Repeat sequences are provided in FASTA format (Fig. 2D). Images are provided on the website by the Repeat Viewer program, which shows the alignment of human and chimpanzee sequences within a gene and in the 5-kb upstream and 5-kb downstream regions. In addition, it also shows the name and position of repeat elements in functional regions, and reports the predicted binding sites of transcription factors in their promoter regions (Fig. 2E).

Various filtering options

As one of the important functions of REPEATOME, various filtering methods can be applied to the REPEATOME database to select the functional region, repeat family, repeat subfamily, repeat name, repeat score, and repeat length. In addition, the database uses set operations such as intersection, union, and difference to find lineage-specific or common repeats (Fig. 2D). Using filter options, users can search human-specific or chimpanzee- specific repeat elements.

Fig. 3 shows snapshots from Repeat Viewer that illustrates these examples. The human RefSeq entry for NM_005441 is aligned with chimpanzee orthologs. Both the human and chimpanzee sequences contain the SVA repeat in the intron (Fig. 3A). In the case of the human RefSeq entry for NM_004000 aligned with the chimpanzee orthologs, the SVA repeats appear only in the human intronic regions (Fig. 3B). For RefSeq entry NM_206889 aligned with its chimpanzee orthologs, only chimpanzee intronic sequences contain the SVA repeat (Fig. 3C). It is known that a large fraction of new transposon insertions in humans and chimpanzees are targeted preferentially to specific genes. Species-specific genetic variation may



(A)

SELECT DB : hg17 and panTro1
○ Gene ● Repeat ○ Sequence
Chr1 Family

SELECT DB : hg17 and panTro1
● Gene ○ Repeat ○ Sequence
Gene Name
(Example : ACVRL1, NM_000020, NP_000011)

SELECT DB : hg17 and panTro1
○ Gene ○ Repeat ● Sequence
Database : hg17refseq Program : blastn e-value : 0.0000001 <
1000000
[Input:FASTA format sequence]

IMAGE TYPE : 0
REGION : transcript
Repeat Family : ALL
Repeat Class : ALL
Match Repeat : ALL
Score range : 200 ~ 70000
Length range : 10 ~ 9000
SPECIFIC : Union

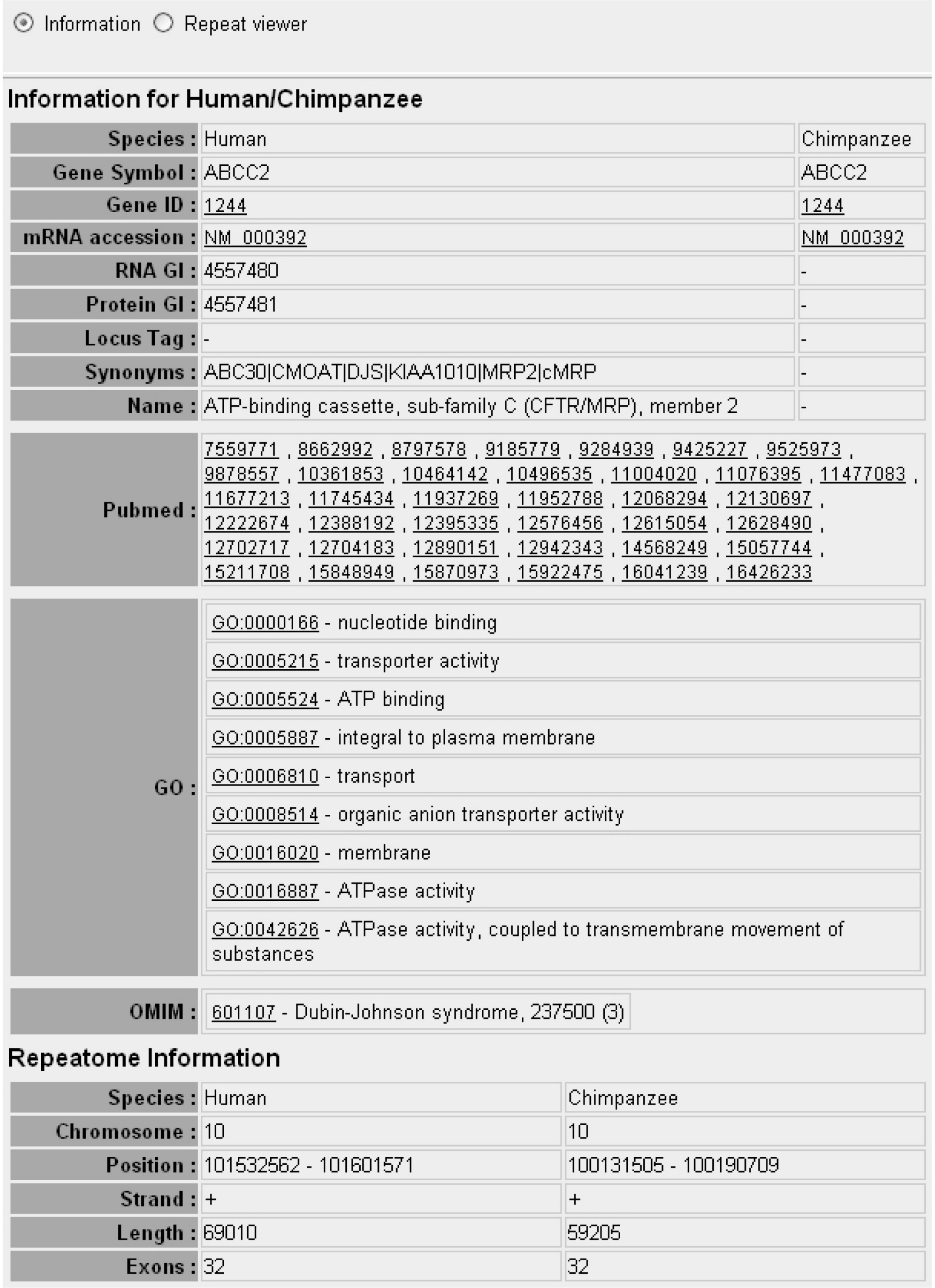
(B)

SELECT DB : Repbase version 11.01, hg18 and panTro2
● Gene ○ Repeat ○ Sequence
Gene Symbol abcc
(Example : TNFRSF8, NM_152942, NP_694421)

Gene ID	Symbol	RNA Acc.	RNA GI	Protein Acc.	Protein GI	Locus Tag	Synonyms	Name	Species	
6833	ABCC8	NM_000352	9910116	NP_000343	4507317	-	ABC36|HHF1|HI|HRINS|MRP8|PHHI|SUR|SUR1	ATP-binding cassette, sub-family C (CFTR/MRP), member 8	Human	Submit
1244	ABCC2	NM_000392	4557480	NP_000383	4557481	-	ABC30|CMOAT|DJS|KIAA1010|MRP2|cMRP	ATP-binding cassette, sub-family C (CFTR/MRP), member 2	Human	Submit
10057	ABCC5	NM_001023587	66529092	NP_001018881	66529093	-	ABC33|DKFZp686C1782|EST277145|MOAT-C|MOATC|MRP5|SMRP|pABC11	ATP-binding cassette, sub-family C (CFTR/MRP), member 5	Human	Submit
368	ABCC6	NM_001171	6715560	NP_001162	6715561	-	ABC34|ARA|EST349056|MLP1|MOATE|MRP6|PXE|PXE1	ATP-binding cassette, sub-family C (CFTR/MRP), member 6	Human	Submit

Fig. 2. (continued)

(C)



● Information ○ Repeat viewer

Information for Human/Chimpanzee

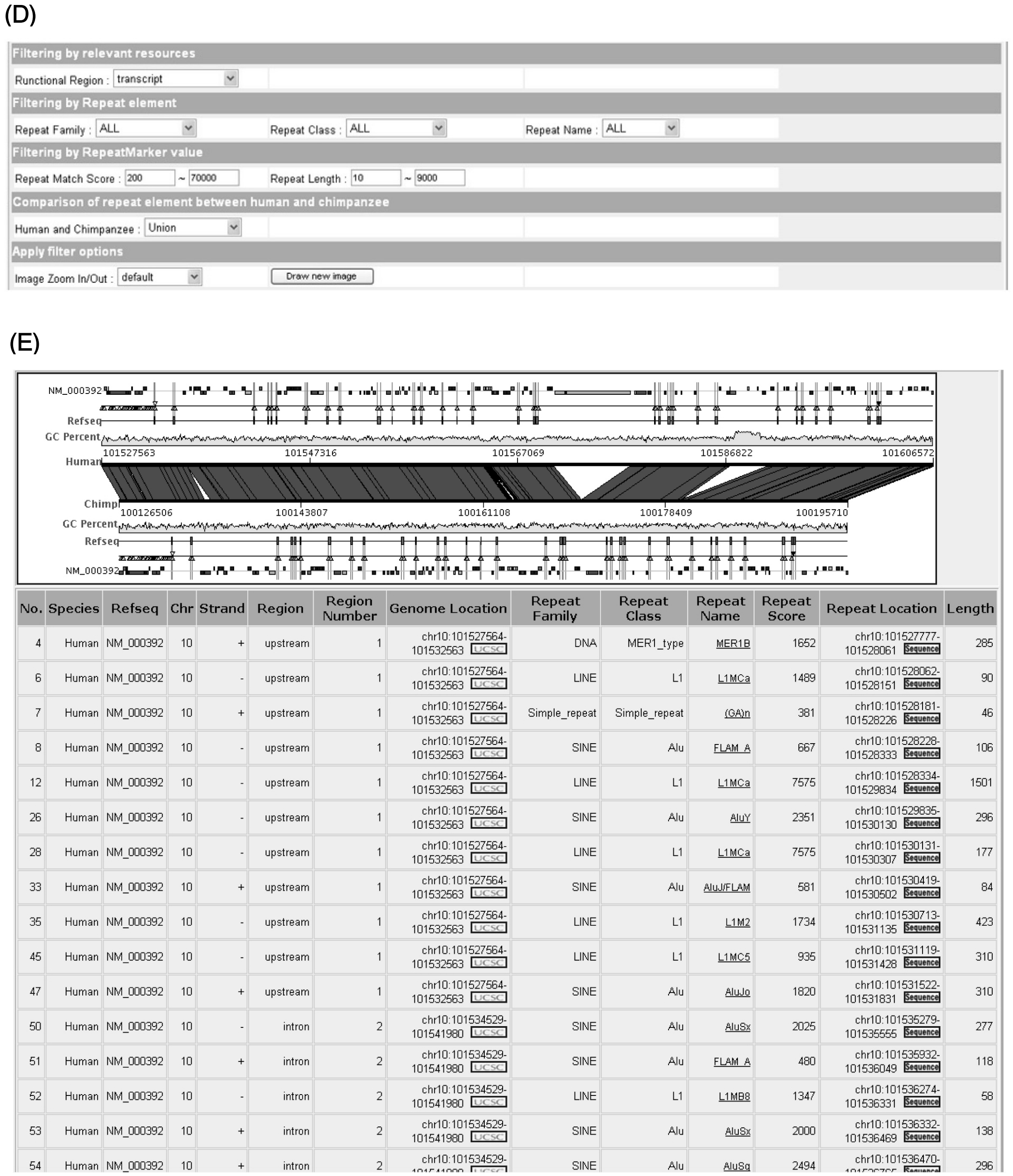
Species :	Human	Chimpanzee
Gene Symbol :	ABCC2	ABCC2
Gene ID :	1244	1244
mRNA accession :	NM_000392	NM_000392
RNA GI :	4557480	-
Protein GI :	4557481	-
Locus Tag :	-	-
Synonyms :	ABC30|CMOAT|DJS|KIAA1010|MRP2|cMRP	-
Name :	ATP-binding cassette, sub-family C (CFTR/MRP), member 2	-

Pubmed :	7559771 , 8662992 , 8797578 , 9185779 , 9284939 , 9425227 , 9525973 , 9878557 , 10361853 , 10464142 , 10496535 , 11004020 , 11076395 , 11477083 , 11677213 , 11745434 , 11937269 , 11952788 , 12068294 , 12130697 , 12222674 , 12388192 , 12395335 , 12576456 , 12615054 , 12628490 , 12702717 , 12704183 , 12890151 , 12942343 , 14568249 , 15057744 , 15211708 , 15848949 , 15870973 , 15922475 , 16041239 , 16426233
GO :	GO:0000166 - nucleotide binding GO:0005215 - transporter activity GO:0005524 - ATP binding GO:0005887 - integral to plasma membrane GO:0006810 - transport GO:0008514 - organic anion transporter activity GO:0016020 - membrane GO:0016887 - ATPase activity GO:0042626 - ATPase activity, coupled to transmembrane movement of substances
OMIM :	601107 - Dubin-Johnson syndrome, 237500 (3)

Repeatome Information

Species :	Human	Chimpanzee
Chromosome :	10	10
Position :	101532562 - 101601571	100131505 - 100190709
Strand :	+	+
Length :	69010	59205
Exons :	32	32

Fig. 2. (continued)



(D)

Filtering by relevant resources

Runctional Region : transcript

Filtering by Repeat element

Repeat Family : ALL   Repeat Class : ALL   Repeat Name : ALL

Filtering by RepeatMarker value

Repeat Match Score : 200 ~ 70000   Repeat Length : 10 ~ 9000

Comparison of repeat element between human and chimpanzee

Human and Chimpanzee : Union

Apply filter options

Image Zoom In/Out : default   Draw new image

(E)

No.	Species	Refseq	Chr	Strand	Region	Region Number	Genome Location	Repeat Family	Repeat Class	Repeat Name	Repeat Score	Repeat Location	Length
4	Human	NM_000392	10	+	upstream	1	chr10:101527564-101532563 UCSC	DNA	MER1_type	MER1B	1652	chr10:101527777-101528061 Sequence	285
6	Human	NM_000392	10	-	upstream	1	chr10:101527564-101532563 UCSC	LINE	L1	L1MCa	1489	chr10:101528062-101528151 Sequence	90
7	Human	NM_000392	10	+	upstream	1	chr10:101527564-101532563 UCSC	Simple_repeat	Simple_repeat	(GA)n	381	chr10:101528181-101528226 Sequence	46
8	Human	NM_000392	10	-	upstream	1	chr10:101527564-101532563 UCSC	SINE	Alu	FLAM_A	667	chr10:101528228-101528333 Sequence	106
12	Human	NM_000392	10	-	upstream	1	chr10:101527564-101532563 UCSC	LINE	L1	L1MCa	7575	chr10:101528334-101529834 Sequence	1501
26	Human	NM_000392	10	-	upstream	1	chr10:101527564-101532563 UCSC	SINE	Alu	AluY	2351	chr10:101529835-101530130 Sequence	296
28	Human	NM_000392	10	-	upstream	1	chr10:101527564-101532563 UCSC	LINE	L1	L1MCa	7575	chr10:101530131-101530307 Sequence	177
33	Human	NM_000392	10	+	upstream	1	chr10:101527564-101532563 UCSC	SINE	Alu	AluJ/FLAM	581	chr10:101530419-101530502 Sequence	84
35	Human	NM_000392	10	-	upstream	1	chr10:101527564-101532563 UCSC	LINE	L1	L1M2	1734	chr10:101530713-101531135 Sequence	423
45	Human	NM_000392	10	-	upstream	1	chr10:101527564-101532563 UCSC	LINE	L1	L1MC5	935	chr10:101531119-101531428 Sequence	310
47	Human	NM_000392	10	+	upstream	1	chr10:101527564-101532563 UCSC	SINE	Alu	AluJo	1820	chr10:101531522-101531831 Sequence	310
50	Human	NM_000392	10	-	intron	2	chr10:101534529-101541980 UCSC	SINE	Alu	AluSx	2025	chr10:101535279-101535555 Sequence	277
51	Human	NM_000392	10	+	intron	2	chr10:101534529-101541980 UCSC	SINE	Alu	FLAM_A	480	chr10:101535932-101536049 Sequence	118
52	Human	NM_000392	10	-	intron	2	chr10:101534529-101541980 UCSC	LINE	L1	L1MB8	1347	chr10:101536274-101536331 Sequence	58
53	Human	NM_000392	10	+	intron	2	chr10:101534529-101541980 UCSC	SINE	Alu	AluSx	2000	chr10:101536332-101536469 Sequence	138
54	Human	NM_000392	10	+	intron	2	chr10:101534529-	SINE	Alu	AluSg	2494	chr10:101536470-	296

Fig. 2. Web interface of REPEATOME. (A) Search interface. Users can search REPEATOME using the gene name, mRNA accession number, protein accession number, repeat family name, repeat class name, repeat name, and specific sequences. (B) The search-results page shows the gene list that matches the query. (C) Detailed information about human and chimpanzee RefSeq. (D) Filter options. This page provides the functionality for the filtering and comparison of repeat elements based on several options. An interesting option in REPEATOME is the "SPECIFIC" option for example, human-specific repeat elements can be selected using "human-specific" options. (E) The Repeat Viewer output. When a gene is selected, the numbers of repeat elements in each region of the human and chimpanzee genomes are displayed in the viewer.

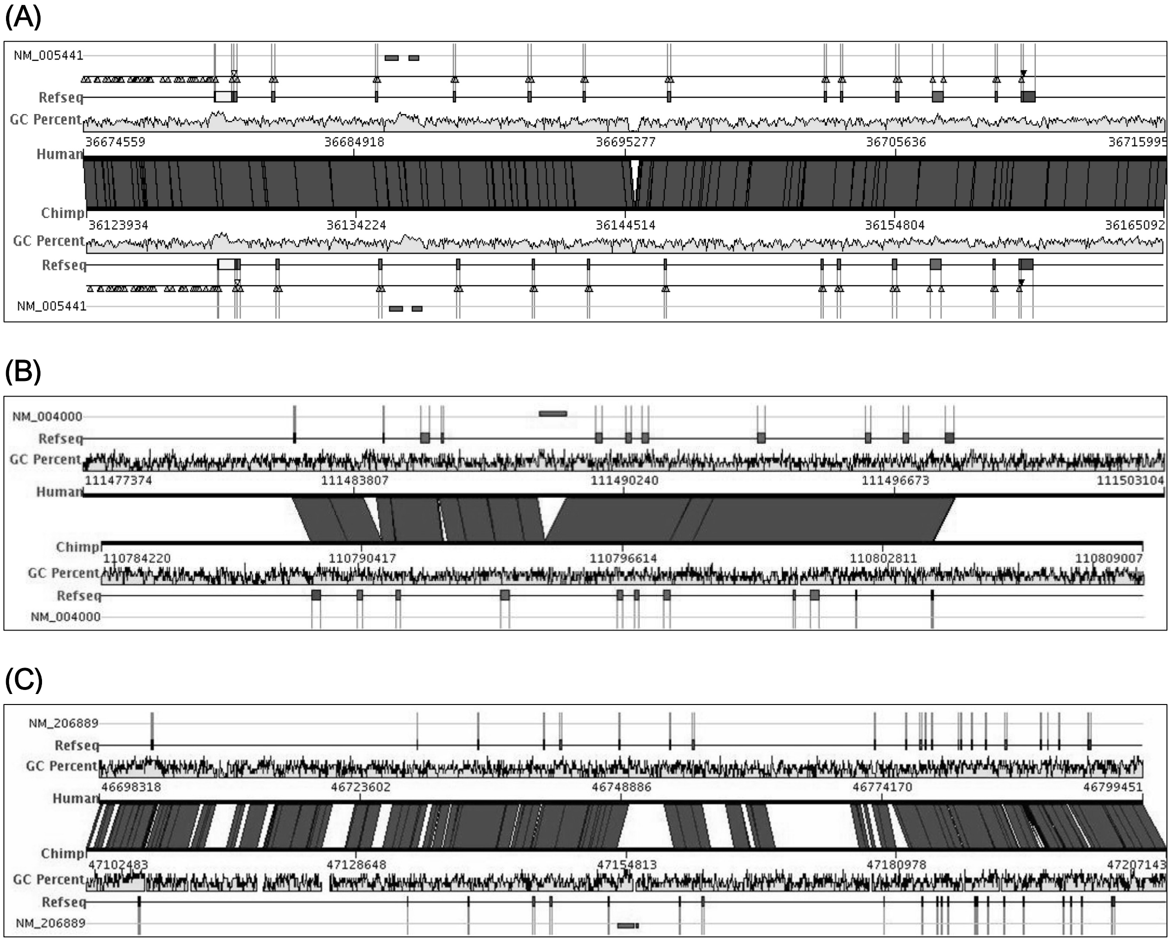


Fig. 3. Output images produced by REPEATOME, which show genes containing SVA repeats in human and chimpanzee orthologs. (A) Both the human RefSeq entry for NM_005441 and its orthologous chimpanzee sequences contain the SVA repeat in their introns. (B) Only the human RefSeq entry for NM_004000 contains the SVA repeat. (C) Only the chimpanzee ortholog (human NM_206889) contains the SVA repeat.

have contributed to the differential evolution of humans and chimpanzees (Mills *et al*., 2006). Functional studies will be necessary to further elucidate the roles of the transposon insertions.

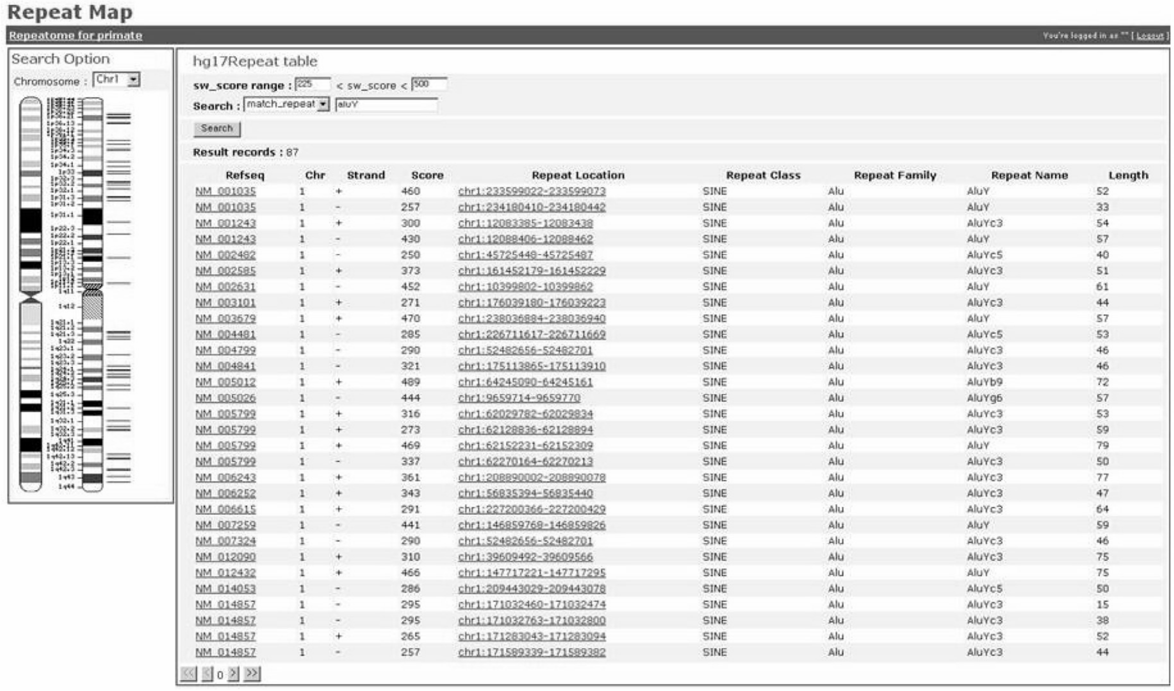
Repeat Map

Repeat Map allows users to search the database using the repeat name, with Repeat Map Viewer presenting a graphical view of the search results. The search-results page shows the location of repeats with blue lines denoting cytogenetic locations, followed by repeat-elements tables linked to the RefSeq ID (Fig. 4).

Most databases containing repeat data do not yet contain detailed comparative information. In contrast, REPEATOME provides such information as the conserved region, functional region, and GC ratio, as well as the functionality of the Repeat Map Viewer and a number of filtering options.

Current Status and Future Directions

Currently, REPEATOME is limited to RefSeq-identified repeats in the human genome, and we are incorporating all primate repeat elements as they are identified. The January 2006 assemblies of the chimpanzee genome (panTro2) have recently been added to the database. Other possible additions include single nucleotide polymorphisms within repeat elements. We will add all published data on repeat elements to the database as they are identified (Woo *et al*., 2007). In addition, we will consider adding orthologous sequences from animals such as macaca and mouse as outgroups to human-chimpanzee



hg17Repeat table

sw_score range : 225 < sw_score < 500

Search : match_repeat aluY

Search

Result records : 87

Refseq	Chr	Strand	Score	Repeat Location	Repeat Class	Repeat Family	Repeat Name	Length
NM_001035	1	+	460	chr1:233599022-233599073	SINE	Alu	AluY	52
NM_001035	1	-	257	chr1:234180410-234180442	SINE	Alu	AluY	33
NM_001243	1	+	300	chr1:12083385-12083438	SINE	Alu	AluYc3	54
NM_001243	1	-	430	chr1:12088406-12088462	SINE	Alu	AluY	57
NM_002482	1	-	250	chr1:45725448-45725487	SINE	Alu	AluYc5	40
NM_002585	1	+	373	chr1:161452179-161452229	SINE	Alu	AluYc3	51
NM_002631	1	-	452	chr1:10399802-10399862	SINE	Alu	AluY	61
NM_003101	1	+	271	chr1:176039180-176039223	SINE	Alu	AluYc3	44
NM_003679	1	+	470	chr1:238036884-238036940	SINE	Alu	AluY	57
NM_004481	1	-	285	chr1:226711617-226711669	SINE	Alu	AluYc5	53
NM_004799	1	-	290	chr1:52482656-52482701	SINE	Alu	AluYc3	46
NM_004841	1	-	321	chr1:175113865-175113910	SINE	Alu	AluYc3	46
NM_005012	1	+	489	chr1:64245090-64245161	SINE	Alu	AluYb9	72
NM_005026	1	-	444	chr1:9659714-9659770	SINE	Alu	AluYg6	57
NM_005799	1	+	316	chr1:62029782-62029834	SINE	Alu	AluYc3	53
NM_005799	1	+	273	chr1:62128836-62128894	SINE	Alu	AluYc3	59
NM_005799	1	+	469	chr1:62152231-62152309	SINE	Alu	AluY	79
NM_005799	1	-	337	chr1:62270164-62270213	SINE	Alu	AluYc3	50
NM_006243	1	+	361	chr1:208890002-208890078	SINE	Alu	AluYc3	77
NM_006252	1	+	343	chr1:56835394-56835440	SINE	Alu	AluYc3	47
NM_006615	1	+	291	chr1:227200366-227200429	SINE	Alu	AluYc3	64
NM_007259	1	-	441	chr1:146859768-146859826	SINE	Alu	AluY	59
NM_007324	1	-	290	chr1:52482656-52482701	SINE	Alu	AluYc3	46
NM_012090	1	+	310	chr1:39609492-39609566	SINE	Alu	AluYc3	75
NM_012432	1	+	466	chr1:147717221-147717295	SINE	Alu	AluY	75
NM_014053	1	-	286	chr1:209443029-209443078	SINE	Alu	AluYc5	50
NM_014857	1	-	295	chr1:171032460-171032474	SINE	Alu	AluYc3	15
NM_014857	1	-	295	chr1:171032763-171032800	SINE	Alu	AluYc3	38
NM_014857	1	+	265	chr1:171283043-171283094	SINE	Alu	AluYc3	52
NM_014857	1	-	257	chr1:171589339-171589382	SINE	Alu	AluYc3	44

Fig. 4. Repeat MapViewer. The search-results page shows the location of repeats with blue lines at cytogenetic locations, followed by repeat-elements tables linked to the RefSeq ID.

comparisons.

REPEATOME would be a useful resource for biologists who are interested in primate evolution, by providing a variety of repeat elements in a comparative perspective that are also associated with functional regions.

List of abbreviations

GO: Gene Ontology
OMIM: Online Mendelian in Man
MySql: My Structured Query Language
BLAST: Basic Local Alignment Search Tool

Acknowledgments

This project was supported by the Korean Ministry of Science and Technology (MOST) under grant number M10407010001-04N0701-00110, and by the KRIBB Research Initiative Program.

References

Batzer, M. A., and Deininger, P. L. (2002). Alu repeats and human genomic diversity. *Nat. Rev. Genet*. 3, 370-379.

Smit, A. F., Toth, G., Riggs, A. D., and Jurka, J. (1995). Ancestral mammalian-wide subfamilies of LINE 1 repetitive sequences. *J. Mol. Biol*. 246, 401-417.

The Chimpanzee Sequencing and Analysis Consortium (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*. 437, 69-87.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., and FitzHugh, W. *et al*. (2001). Initial sequencing and analysis of the human genome. *Nature*. 409, 860-921.

Mills, R. E, Bennett, E. A., Iskow, R. C., Luttig, C. T., Tsui, C., Pittard, W. S., and Devine, S. E. (2006). Recently mobilized transposons in the human and chimpanzee genomes. *Am. J. Hum. Genet*. 78, 671-679.

Ohno, S., and Yomo, T. (1991). The grammatical rule for all DNA: junk and coding sequences. *Electrophoresis*. 12, 103-108.

Dagan, T., Sorek, R., Sharon, E., Ast, G., and Graur, D. (2004). AluGene: a database of Alu elements incorporated within protein-coding genes. *Nucleic Acids Res*. 32, D489-D492.

Le Goff, W., Guerin, M., Chapman, M. J., and Thillet, J. (2003). A CYP7A promoter binding factor site and Alu repeat in the distal promoter region are implicated in regulation of human CETP gene expression. *J. Lipid Res*. 44, 902-910.

Grover, D., Kannan, K., Brahmachari, S. K., and Mukerji, M. (2005). Alu-ring elements in the primate genomes. *Genetica.* 124, 273-289.

Arnaud, P., Goubely, C., Pelissier, T., and Deragon, J. M. (2000). SINE retroposons can be used in vivo as nucleation centers for de novo methylation. *Mol. Cell. Biol*. 20, 3434-3441.

Thornburg, B. G., Gotea, V., and Makalowski, W. (2006). Transposable elements as a significant source of transcription regulating signals. *Gene*. 365, 104-110.

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res*. 110, 462-467.

Hinrichs, A. S., Karolchik, D., Baertsch, R., Barber, G. P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T. S., Harte, R. A., Hsu, F., Hillman-Jackson, J., Kuhn, R. M., Pedersen, J. S., Pohl, A., Raney, B. J., Rosenbloom, K. R., Siepel, A., Smith, K. E., Sugnet, C. W., Sultan-Qurraie, A., Thomas, D. J., Trumbower, H., Weber, R. J., Weirauch, M., Zweig, A. S., Haussler, D., and Kent, W. J. (2006). The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res*. 34, D590-D598.

Nagashima, T., Matsuda, H., Silva, D. G., Petrovsky, N., Konagaya, A., Schönbach, C., Kasukawa, T., Arakawa, T., Carninci, P., Kawai, J., and Hayashizaki, Y. (2004). FREP: a database of functional repeats in mouse cDNAs. *Nucleic Acids Res*. 32, D471-D475.

Boby, T., Patch, A. M., and Aves, S. J. (2005). TRbase: a database relating tandem repeats to disease genes for the human genome. *Bioinformatics*. 21, 811-816.

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.

Kent, W. J. (2002). BLAT-the BLAST-like alignment tool. *Genome Res*. 12, 656-664.

Maglott, D., Ostell, J., Pruitt, K. D., and Tatusova, T. (2005). Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res*. 33, D54-D58.

Matys, V., Kel-Margoulis, O. V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., Voss, N., Stegmaier, P., Lewicki-Potapov, B., Saxel, H., Kel, A. E., and Wingender, E. (2006). TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res*. 34, D108-110.

Zhang, M. Q. (1998). Statistical features of human exons and their flanking regions. *Hum. Mol. Genet*. 7, 919-932.

Lambert, A., Fontaine, J. F., Legendre, M., Leclerc, F., Permal, E., Major, F., Putzer, H., Delfour, O., Michot, B., and Gautheret, D. (2004). The ERPIN server: an interface to profile-based RNA motif identification. *Nucleic Acids Res*. 32, W160-W165.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403-410.

McEntyre, J., and Lipman, D. (2001). PubMed: bridging the information gap. *CMAJ*. 164, 1317-1319.

Woo, T., Kim, Y., Kwon, J., and Seo, J. (2007). RepWeb: A Web-Based Search Tool for Repeat-Related Literatures. *Genomics & Informatics*. 5, 89-91.