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

Taeha Woo1,2¶, Tae-Hui Hong1¶, SangSoo Kim2, Won-Hyong Chung1, Hyo Jin Kang1, Chang-Bae Kim1 and Jungmin Seo1,*

1Korean BioInformation Center, KRB, Daejeon 305-806, Korea, 2Department 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 (Batzar 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.
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 resource for 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.

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 matches the 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
### Information for Human/Chimpanzee

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Human</th>
<th>Chimpanzee</th>
</tr>
</thead>
<tbody>
<tr>
<td>Species</td>
<td>Human</td>
<td>Chimpanzee</td>
</tr>
<tr>
<td>Gene Symbol</td>
<td>ABC2</td>
<td>ABC2</td>
</tr>
<tr>
<td>Gene ID</td>
<td>1244</td>
<td>1244</td>
</tr>
<tr>
<td>mRNA accession</td>
<td>NM_000392</td>
<td>NM_000392</td>
</tr>
<tr>
<td>RNA GI</td>
<td>4557480</td>
<td>-</td>
</tr>
<tr>
<td>Protein GI</td>
<td>4557481</td>
<td>-</td>
</tr>
<tr>
<td>Locus Tag</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Synonyms</td>
<td>ABC30</td>
<td>CMOAT</td>
</tr>
<tr>
<td>Name</td>
<td>ATP-binding cassette, sub-family C (CFTR/MRP), member 2</td>
<td>-</td>
</tr>
</tbody>
</table>

### Pubmed
- 7559771, 8669992, 8797578, 9105779, 9284938, 9425227, 9525973, 9878567, 10361852, 10464142, 10498635, 11004020, 11076395, 11477083, 11677213, 11745434, 11937269, 11952788, 12088294, 12130687, 12222674, 12388192, 12395335, 12576456, 12615054, 12628490, 12702717, 12704183, 12890151, 12842343, 14566249, 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

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Human</th>
<th>Chimpanzee</th>
</tr>
</thead>
<tbody>
<tr>
<td>Species</td>
<td>Human</td>
<td>Chimpanzee</td>
</tr>
<tr>
<td>Chromosome</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Position</td>
<td>101532562 - 101601571</td>
<td>100131505 - 100190709</td>
</tr>
<tr>
<td>Strand</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Length</td>
<td>69010</td>
<td>59205</td>
</tr>
<tr>
<td>Exons</td>
<td>32</td>
<td>32</td>
</tr>
</tbody>
</table>

Fig. 2. (continued)
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


