

# PrimateDB: Development of Primate Genome DB and Web Service

Taeha Woo<sup>1</sup>, Gwangsik Shin<sup>1</sup>, Taewook Kang<sup>1</sup>,  
Byoungchul Kim<sup>1</sup>, Jungmin Seo<sup>1</sup>, Sang Soo Kim<sup>2</sup>  
and Chang-Bae Kim<sup>1\*</sup>

<sup>1</sup>National Genome Information Center, Korea Research Institute of Bioscience and Biotechnology, 52 Eoeun-dong, Yuseong-gu, Daejeon 305-333, Korea, <sup>2</sup>Department of Bioinformatics, Soongsil University, 1-1, Sangdo-dong, Dongjak-Gu, Seoul, 156-743, Korea

## Abstract

The comparative analysis of the human and primate genomes including the chimpanzee can reveal unique types of information impossible to obtain from comparing the human genome with the genomes of other vertebrates. PrimateDB is an open depository server that provides primate genome information for the comparative genome research. The database also provides an easy access to variable information within/between the primate genomes and supports analyzed information, such as annotation and retroelements and phylogeny. The comparative analyses of more primate genomes are also being included as the long-term objective.

**Keywords:** Primate genome, comparative genomics, database, retroelements

**URL:** <http://www.ngic.re.kr:8000/comparative/primatedb/>

## Introduction

After the completion of the human genome project, primate genomes including a common chimpanzee are focused for understanding the origin of humanity and human genetic diseases. These genome data are being produced rapidly and massively. However, information of primate genome are scattered throughout several databases.

There are many families of retroelements in primate genomes. As the most successful short interspersed elements (SINEs) in primate genomes, *Alu* elements have had remarkable proliferation during the primate

radiation and have expanded to more than one million repetitive elements dispersed throughout the human genome (Batzer and Deininger, 2002). Among *Alu* repeat subfamily, *AluY* elements might be contributed the driving force for speciation between human and chimpanzee from the common ancestor (Sakaki *et al.*, 2003; Watanabe *et al.*, 2004)

We constructed a database, PrimateDB that integrates primate genomic data and provides several analyzed data on whole-genome scale by bioinformatics methods. In addition the user can easily search variable information about the genomes of primate including genome sequence, annotation information, retroelement information by using the database.

## Methods

### Primate genome sequence database

The primate genome database was implemented using the Select Query Language (SQL) from the MySQL database server. Whole genome sequence data of primate mainly from common chimpanzee were downloaded from Trace archive of NCBI (<http://www.ncbi.nih.gov/Traces/trace.cgi>) and local alignment with human chromosomes was done by using BLAT (Kent, 2002). The database relies on the build 30 of NCBI human genome and will be updated with each new release. The pre-dataset of each chromosome group was assembled by using ARACHNE (Batzoglou *et al.*, 2002; Jaffe *et al.*, 2003). For analysis of large scale data from the Trace archive an analytical pipeline was designed by Perl language.

The resulting contigs was annotated by using BLAT to UCSC Human reference sequence (hg17). The assembled sequences written by Perl script were added as a customer track to UCSC genome browser and resulting data were stored in the MySQL database server (Fig. 1).

### Phylogenetic information

For the comparison of disease related genes located in human chromosomes with those of primate orthologous chromosomes, complete exon sequences of disease related genes annotated at OMIM (Online Mendelian Inheritance in Man) dataset of NCBI (<http://www.ncbi.nih.gov/entrez/query.fcgi?db=OMIM>) were used for executing

\*Corresponding author: E-mail [changbae@kribb.re.kr](mailto:changbae@kribb.re.kr),  
Tel +82-42-860-4643, Fax +82-42-879-8519  
Accepted 7 May 2005

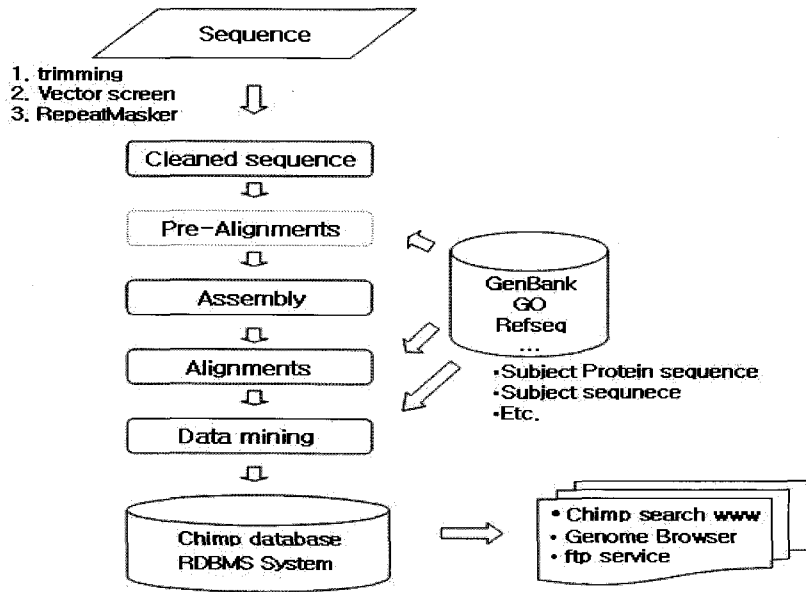


Fig. 1. System architecture of primate genome sequence database.

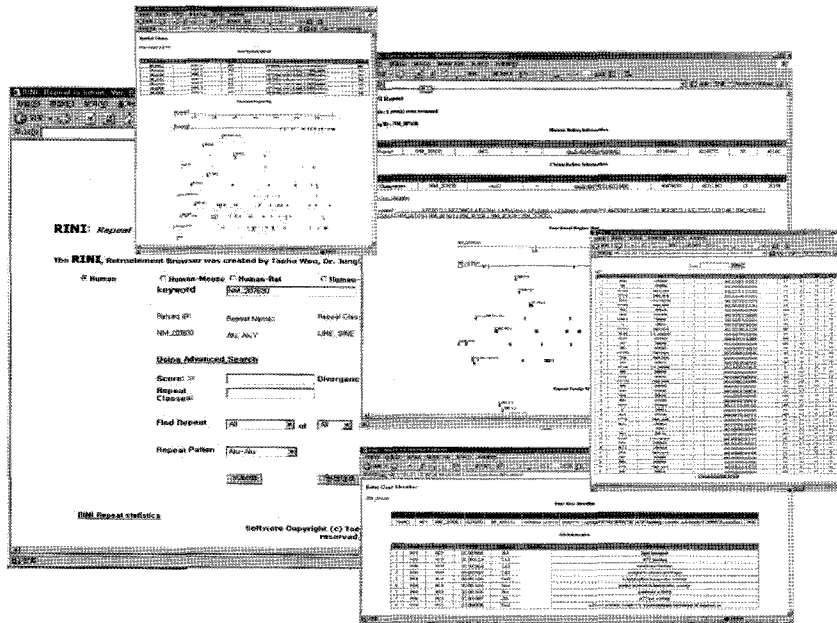


Fig. 2. Primate retroelement database.

phylogenetic analyses by WebPHYLP (Lim and Zhang, 1999) and nucleotide substitution types were examined (Fig. 3).

**Primate retroelement database**

Human, chimpanzee, gorilla, and orangutan genome sequences were retrieved from the UCSC Genome Browser by using human RefSeq. The human and

chimpanzee genome sequences were annotated by using both CENSOR (Version 4.X; <http://www.girinst.org>) and RepeatMasker (Smit and Green, unpublished work; <http://repeatmasker.genome.washington.edu>). The resulting data were stored to MySQL database. Analysis of repeat elements was executed by dividing genome region into UTR, exon and intron, respectively. The outcome of queries can be sequence alignment, genomic locus in

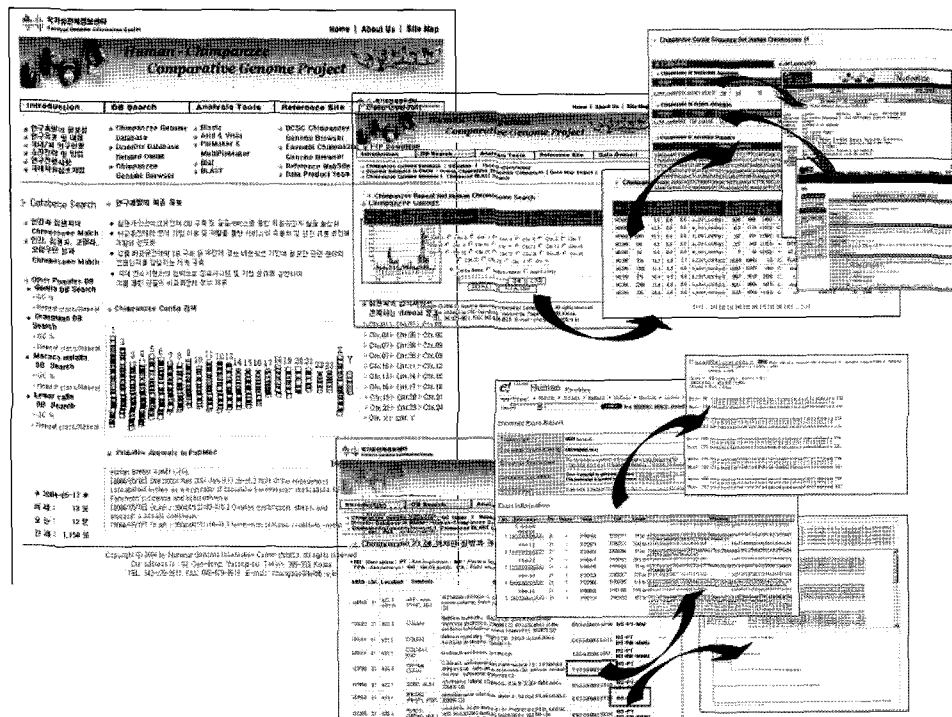


Fig. 3. PrimateDB user interface.

which retroelements are embedded and sequence of families and subfamilies, map image, and gene symbol information (Fig. 2). Image of retroelement was made by using GD library.

### Links to other resources

The database provides links to a variety of related resources on the Internet. Through these links the user can access related human and chimpanzee genome map information available in the UCSC Genome Browser and NCBI and they are used as a reference to the most recent genome map.

### Description of the PrimateDB interface

The HTML interface for PrimateDB (Fig. 3) is made by using PERL and PHP. There are five web pages: The main page, contig search, retroelement search, reference, and analysis tools. The PrimateDB web interface allows access of data through graphical or list browsing, searching by keywords, names or sequences of repeat family and class. The schematic physical map presents the extent and locations of genes and retroelements in the area defined by the user query. Alignment of the nucleotide sequences queried by user can be retrieved.

## Results and Discussion

Analytic pipeline for comparative approach of primate genome information mainly from the NCBI Trace archive of chimpanzee genome was designed and constructed. Through the analysis system other primate genomes were included and analyzed. As human genome as a reference chimpanzee genome was also compared for the comparative analysis of retroelement from which functional repeat can be analyzed. Based upon the present databases evolutionary analysis of genome of primates including retroelements to understand human uniqueness should be stimulated.

### Acknowledgements

This work is supported 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.
- Batzoglou, S., Jaffe, D.B., Stanley, K., Butler, J., Gnerre, S., Mauceli, E., Berger, B., Mesirov, J.P., and Lander, E.S. (2002). ARACHNE: a whole-genome shotgun assembler.

- Genome Res.* 12, 177-189.
- Jaffe, D.B., Butler, J., Gnerre, S., Mauceli, E., Lindblad-Toh, K., Mesirov, J.P., Zody, M.C., and Lander, E.S. (2003). Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res.* 13, 91-96.
- Kent, W. J. (2002). BLAT-the BLAST-like alignment tool. *Genome Res.* 12, 656-664.
- Lim, A. and Zhang, L. (1999). WebPHYLP: a web interface to PHYLIP. *Bioinformatics* 15, 1068-1069.
- Sakaki, Y., Watanabe, H., Taylor, T., Hattori, M., Fujiyama, A., Toyoda, A., Kuroki, Y., Itoh, T., Saitou, N., Oota, S., Kim, C.G., Kitano, T., Lehrach, H., Yaspo, M.L., Sudbrak, R., Kahla, A., Reinhardt, R., Kube, M., Platzer, M., Taenzer, S., Galgoczy, P., Kel, A., Bloecker, H., Scharfe, M., Nordsiek, G., Hellmann, I., Khaitovich, P., Paabo, S., Chen, Z., Wang, S.Y., Ren, S.X., Zhang, X.L., Zheng, H.J., Zhu, G.F., Wang, B.F., Zhao, G.P., Tsai, S.F., Wu, K., Liu, T.T., Hsiao, K.J., Park, H.S., Lee, Y.S., Cheong, J.E., and Choi, S.H. (2003). Chimpanzee Chromosome 22 Sequencing Consortium Human versus chimpanzee chromosome-wide sequence comparison and its evolutionary implication. *Cold Spring Harb Symp Quant Biol.* 68, 455-460.
- Watanabe, H., Fujiyama, A., Hattori, M., Taylor, T.D., Toyoda, A., Kuroki, Y., Noguchi, H., BenKahla, A., Lehrach, H., Sudbrak, R., Kube, M., Taenzer, S., Galgoczy, P., Platzer, M., Scharfe, M., Nordsiek, G., Blocker, H., Hellmann, I., Khaitovich, P., Paabo, S., Reinhardt, R., Zheng, H.J., Zhang, X.L., Zhu, G.F., Wang, B.F., Fu, G., Ren, S.X., Zhao, G.P., Chen, Z., Lee, Y.S., Cheong, J.E., Choi, S.H., Wu, K.M., Liu, T.T., Hsiao, K.J., Tsai, S.F., Kim, C.G., Oota, S., Kitano, T., Kohara, Y., Saitou, N., Park, H.S., Wang, S.Y., Yaspo, M.L., and Sakaki, Y. (2004). DNA sequence and comparative analysis of chimpanzee chromosome 22. *Nature* 429, 382-388.