Minireview Genomics & Informatics Vol. 2(4) 163-166, December 2004

Comparative Study of Human and Chimpanzee Genome

Sang-Hang Choi, Dae-Soo Kim, Dae-Won Kim, Yong-Seok Lee and Hong-Seog Park*

Genome Structure Research Laboratory, Korea Research Institute of Bioscience and Biotechnology, 52 Oun-dong, Yusong-gu, Daejeon 305-333, Korea

Keywords: human and chimpanzee genome, comparative genomics, evolution, physical map, sequencing

Introduction

Chimpanzee is an essential organism in the earth to understand the human unique features, such as highly developed cognitive functions, bipedalism and the use of complex language. Therefore, a number of pilot studies comparing the human and chimpanzee genome have been conducted to understand the genetic basis of the unique features of humans, although all of their studies are by using partial genome information (Park *et al.*, 2000; King *et al.*, 1975; Fujiyama *et al.*, 2002; Olson *et al.*, 2003; Boffelli *et al.*, 2003). The recently released human genome sequences provide us with reference data to conduct comparative genomic research between human and chimpanzee. One of pioneer group in this field is "The International Chimpanzee Genome Sequencing Consortium." The consortium was organized by 8 centers from 5 countries at 13 March 2001 (Table 1). This is a review article of comparative genome analysis between human and chimpanzee which have been conducted by the consortium.

Construction and Analysis of a Human-Chimpanzee Comparative Clone Map

Our consortium already represented the construction and analysis of a first-generation human-chimpanzee comparative genomic map, which was an initial step for understanding the chimpanzee genome structure (Fujiyama *et al.*, 2002).

The map was constructed through paired alignment of 77,461 chimpanzee bacterial artificial chromosome end sequences (BES) information with publicly available human genome sequences. The 77,461 BESs numbers are to have an alignment longer than 50 base pairs (bp) with ≥90% identity. Out of this number, 49,160 BESs from 24,580 clones formed paired ends where each pair was derived from the same clones. Only one end could be successfully aligned from the remaining 28,301 clones. The remaining 36,960 BESs that were not aligned with human genome were categorized into three different classes, repetitive sequences, matched to several species, and not sequenced human region or lineage specific sequences.

In this effort, 48.6% of the whole human genome was covered by the chimpanzee BACs, the reasons for this apparently low coverage is that we used rather stringent conditions for the calculation, and probably the human reference data was draft stage at that time. This probability can be proved by the fact that the coverage score was substantially higher for the chromosome 14, 20, 21, and 22 was substantially higher, which have very high quality sequences. Based on several tests to estimate the theoretical coverage with the human genome, we concluded that the actual coverage is approximately 70% of the reference genome. Relatively lower coverage of chromosome X is because of the haploid status of the chromosome in the chimpanzee BAC libraries which was constructed from the male chimpanzee). In contrast, the Y chromosome coverage is so much lower (4.8%) as compared with the other chromosomes. One of the possible explanation is that larger genome structure changes have occurred in the genome structure of Y chromosome during evolutional process.

We detected the boundaries of possible genomic rearrangements by searching for candidate clones containing chromosomal breakpoints using the mapping results. One of samples, PTB-053J22, contained the breakpoints corresponding to the human (or vice versa in

Table 1. An international team for chimpanzee genome sequencing

THE CHIMP PLAYERS	
Country	Laboratory
Japan	RIKEN's Genomic Sciences Center, Yokohama
Germany	The Max Planck Institute for Molecular Genetics, Berlin
Korea	Koera Research Institute of Bioscience and Biotechnology, Daejeion
China	National Human Genome Center, Shanghai
Taiwan	National Yang-Ming University, Taipei

*Corresponding author: E-mail hspark@kribb.re.kr, Tel +82-42-879-8132, Fax +82-42-879-8139
Accepted 15 November 2004

chimpanzee) chromosomal inversion between 12p12 and 12q15.

In addition, this region in human chromosome 12 or in chimpanzee chromosome 10 is known to be inverted in gorilla and chimpanzee as opposed to human and orangutan. The effect of genes orders by the inversion should be the target of future studies.

Comparative analysis of human chromosome 21 and chimpanzee chromosome 22

Comparative study between human-chimpanzee genome is an essential research for narrowing down the genetic changes which affected their unique features from a common ancestor in evolution process. Cytological evidence such as duplication, translocations, and transpositions has been reported, but owing to technological limitation there is not an integrated picture of the dynamic changes of the genome (Thomas *et al.*, 2003). Therefore, a gold standard is required to evaluate the overall consequence of these genetic changes on human evolution. To address these issues we have conducted a human-chimpanzee whole-chromosome comparison at the nucleotide sequence level.

Targeting genome was human chromosome 21 (HSA21) which is the orthologue of chimpanzee chromosome 22 (PTR22). HSA21 is one of the most well characterized genome and also the smallest in size among human chromosomes (Hattori *et al.*, 2000). Moreover, in medical implication of HSA21, its genes involved in the chromosome are well known to be related with Down's syndrome (trisomy 21). Interestingly, one case of trisomy 22 in chimpanzee has been reported, which showed very similar phenotypic feature to the Down's syndrome of human (McClure *et al.*, 1969). Therefore, our analysis of these chromosomes should reveal dynamic changes that may reflect general evolutionary events occurring throughout the human genome.

Our consortium used three different BAC libraries from genomic DNA originated from three male chimpanzees (*Pan troglodytes*). After constructing high resolution comparative BAC clone map, we divided sequencing region responsible to each sequencing centers. And we established common web site for organizing all the sequencing data and communication among consortium members. Our basic policy for sequencing data management was not to release all the data before publishing the paper.

Sequencing coverage of the euchromatic portion of the long arm of PTR 22 is estimated to be 98.6% (PTR22/HSA21=32,799,845 bp/ 33,127,944), and sequencing accuracy was calculated as 99.9983% and ≥99.9981% from overlapping clone sequences and Phrap scores, each other.

The overall structural features of PTR22 are almost the same as those of HSA 21q. However three main different features could be found. One is that HSA21q was larger a roughly 400-kilobase (kb) or 1.2% difference in size than PTR 22, the difference of which is mainly due to interspersed repeats (ISRs) and simple repeats. This difference can mostly be explained by the fact that several subfamilies of transposable elements, such as L1, MER83R, AluYa5 and AluYb8 are more common in human than in chimpanzee, and also five LTR subfamilies are more abundant in HSA21q.

The second is the pericentromeric copy of a 200-kb region found duplicated in HSA21 is missing in PTR22q. Finally, one of the largest structural changes is a 54-kb region located 11.4 Mb from the centromere in HSA21q but that is absent in PTR22. We also detected some interesting features that human-specific sequences are neither repetitive nor low complexity and are unique in the nr data set of NCBI, and that two large indel hotspots were found at around 9.5Mb-11.5Mb and 16.5Mb-17.5Mb from the centromere.

The overall nucleotide substitution level in aligned regions between PTR22q and HSA21q is about 1.44% (excluding indel). The base substitution frequency showed almost same distribution along the chromosome, except for elevated regions in the pericentromeric and subtelomeric regions. The most conserved region was at about the 12.5-Mb region (0.87% over 100 kb), corresponding to the distal boundary region of the gene desert. The correlation between the G+C content and the base substitution rate increases along the chromosome, especially high in the last 5Mb of the telomeric region of PTR22, and also base substitution in repetitive sequences also tend to have various frequencies. We found out

Table 2. Summary of annotated genes from chimpanzee chromosome 22

Gene catalogue and characterization of coding sequences			
Annotated Gene(272)	Single ORF or EST (41)		
	Canonical ORF(231)	Nucleotide sequence	
		Identical length(179)	Amino acid sequence
			100% match (39)
			Replacement (140)
		Significant Change(47)	
		Unclassified(5)	
Pseudogene (89)			

#HSA21q protein coding genes (284), pseudognes (98) (http://chr21.molgne.mpg.de)

Table 3. Positional translocated genes probably by retrogenes insertion during human evolution

Chimpanzee			Human			Function
Gene Name	Location	Size(bp)	Gene Name	Location	Size(bp)	
RPLP1LK1	26390907-26391416	510	RPLP1	15q22	512bp	Ribosomal protein, large, P1
HNRPA1LK1	31680423-31681781	1359	HNRPA1	12q13.1	1769bp	Heterogeneous nuclear ribonucleo protein A1
FAM28ALK1	32934790-32937332	2543	MGC10198	4q35.1	2579bp	unknown
H2Bf	5p21	430	H2BFS	21q21.3	126aa	H2B histone family S member
KAP(5)	?		KAP(5)	21q21	Variant	Human hair-keratin associated

Table 4. Probable genes to be changed in their function during human and chimpanzee evolution

Name(H)	Name(C)	Fun.(H/C)	Function
RPL13AP7	RPL13ALK1	-/+	Ribosomal protein RPL13A pseudogene
C21orf81	C21orf81	+/-	Chromosome 21 open reading frame 81(C21orf81), mRNA
C21orf115	C21orf115	+/-	Homo sapiens PP1416 mRNA, Complete cds
C21orf104	C21orf104	+/-	Hypothetical protein
C21orf19	C21orf19	+/-	Homo sapiens C21orf19-like protein mRNA, complete cds

repetitive sequences indicating evolutional processing between human and chimpanzee that the last common ancestors are L1Hs, AluYa5 and AluYb8, and two elements to be integrated after speciation were AluYb9 and L1PA2.

The precise identification of insertion and deletion (indel) in the two genomes is essential to understand the processes underlying human and chimpanzee evolution (Frazer *et al.*, 2003). We identified about 68,000 indels in total, and tested 567 indels larger than 300bp using DNA samples from human, chimpanzee, gorilla and orangutan by PCR amplification. Then we were able to distinguish 193 fragments which showed lineage-specific changes in size, and estimate the original state of these regions in the genome of the last common ancestor. Insertions were mostly produced by the integration of Alu and L1 elements, whereas deletions were not related to particular repetitive structures. An interesting difference between the human and chimpanzee genome evolution was newly integrated Alu element, which is inserted in the high G+C region in human whereas in the low G+C in chimpanzee. Unlike the insertion, deletions do not correspond exactly to any ISR elements, indicating that deletion events are independent of ISRs. Calculation result of 300-5000 bp range indel suggested that the ancestral chromosome was larger than both HSA21q and PTR22q. In conclusion, the expansion of particular elements was repeated several times during the course of evolution, and also AluY elements burst might be contributed the driving force for speciation between the human and the chimpanzee from the common ancestor.

One of interesting question in this research should be how many different genes exist and which kind of the genes are different between human and chimpanzee (Varki, 2000). We have annotated 272 protein-coding genes and 89 pseudogenes in PTR22 by comparing with HSA21 (Table 2). Among the 231 genes associated to a canonical ORF, 179 genes show a coding sequence of identical length in human and chimpanzee, the average nucleotide and amino acid identity in the coding region is 99.29% and 99.18%, respectively. Of these, 39 genes show an identical amino acid sequences, other genes were different in more than one amino acid. Especially, 47 genes were significantly changed, and 5 chimpanzee genes were unclassified because they displayed structural changes caused by indels (*SH3BGR, SYNJ1, C21orf96 and TMPRSS3*) or substitution in the ATG codon (*C21orf18*). Six HSA21q genes displaying hallmark of retro genes were not found in PTR22q and were probably inserted during human evolution or deleted during chimpanzee evolution (Table 3). Moreover, we found out 5 probable genes to be changed in their function, one ribosomal protein pseudogene, and four coding sequences. Our data suggest that indels within coding regions represent one of the major mechanisms generating protein diversity and shaping higher primate species.

The other interesting point is to know the gene expression pattern between human and chimpanzee, and then we compared them in two tissues of brain (202 genes) and liver (96 genes). Of these, 9 in the brain and 12 in the liver showed a significant change in expression level between human and chimpanzees in the range of a 1.5-10-fold difference. Some of genes displaying significant changes in protein sequence or differences in expression between human and chimpanzee might be correlated with physiological or disease susceptibility differences exhibited between the two species, such as immune response (*IFNAR2, IFNGR2, CXADR, ITSN1 and CRYZL1*), developing heart (*SH3BGR*), peripheral nervous system

(*C21orf2*), early brain development (*SYNJ1 and ANKRD3*), cell cycle progression (*MCM3AP*), and Knobloch syndrome (*COL18A1*). Beside these results, we analyzed SNP for interfering the ancestral allele of any human SNP locus, Ka/Ks test, and promoter

Conclusions

Our data revealed a complex set of genetic differences between human and chimpanzee. Whole genome comparative and further experimental studies are required before inferring the impact of these genetic changes for the biological consequences of the organism. Nonetheless, data presented here suggest that the biological consequences derived from the genetic differences may be more complicated than our previous speculation and offer a framework allowing the design of experimental strategies for testing novel hypothesis. In present, whole genome sequencing project of chimpanzee is progressing by USA sequencing groups by shot-gun sequencing, and also our consortium group will finish chimpanzee Y chromosome sequencing with very high quality in a few time. And then I expect to understand the question of what makes us human is not so far.

Acknowledgement

We are grateful to all the technical staff for contributing the genome sequencing. This work was supported by the Ministry of Science and Technology, and the Korea Research Institute of Bioscience and Biotechnology, Korea.

References

Boffelli, D., McAuliffe, J., Ovcharenko, D., Lewis, K.D., Ovcharenko, I., Pachter, L., and Rubin, E.M. (2003). Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Scinece* 299, 1391-1394.

Frazer, K.A., Chen, X., Hinds, D.A., Pant, P.V., Patil, N., and Cox, D.R. (2003). Genomic DNA insertions and deletions occur frequently between humans and nonhuman primates. *Genome Res.* 13, 341-346.

Fujiyama, A., Watanabe, H., Toyoda, A., Taylor, T.D., Itoh, T., Tsai, S.F., Park, H.S., Yaspo, M.L., *et al.* (2002). Construction and analysis of a human-chimpanzee comparative clone map. *Science* 296, 131-134.

Hattori, M.A., Fujiyama, T. D., Taylor, H.. Watanabe, T.. Yada, H.-S., Park, A., *et al.* (2000). The DNA sequence of human chromosome 21. *Nature* 405, 368-372.

International Human Genome Sequencing Consortium (2001). Initial sequencing analysis of the human genome. *Nature* 409, 860-921.

King, M.C. and Wison, A.C. (1975). Evolution at two levels in humans and chimpanzee. *Science* 188, 107-116.

McClure, H.M., Belden, K.H., Pieper, W.A., and Jacobson, C.B. (1969). Autosomal trisomy in a chimpanzee: resemblance to Down's syndrome. *Scinece* 165, 1010-1012.

Olson, M.V. and Varki, A. (2003). Sequencing the chimpanzee genome: insights into human evolution and disease. *Nature Rev. Genet.* 4, 20-28.

Park, H.-S., Nogami, M., Okumura, K., Hattori, M., Sakaki, Y., and Fujiyama A. (2000). Newly identified repeat sequences, derived from human chromosome 21qter, are also localized in the subtelomeric region of particular chromosomes and are conserved in the chimpanzee genome. *FEBS letter* 475, 167-169.

Thomas, J.W., Touchman, J.W., Blakesley, R.W., Bouffard, G.G., Beckstrom-Sternberg, S.M., Margulies, E.H., *et al.* (2003). Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* 424, 788-793.

Varki, A. (2000). A chimpanzee genome project is a biomedical imperative. *Genome Res.* 10, 1065-1070.