DNA sequence and comparative analysis of chimpanzee chromosome 22

The International Chimpanzee Chromosome 22 Consortium*

**A list of authors and their affiliations appears at the end of the paper*

Human–chimpanzee comparative genome research is essential for narrowing down genetic changes involved in the acquisition of unique human features, such as highly developed cognitive functions, bipedalism or the use of complex language. Here, we report the high-quality DNA sequence of 33.3 megabases of chimpanzee chromosome 22. By comparing the whole sequence with the human counterpart, chromosome 21, we found that 1.44% of the chromosome consists of single-base substitutions in addition to nearly 68,000 insertions or deletions. These differences are sufficient to generate changes in most of the proteins. Indeed, 83% of the 231 coding sequences, including functionally important genes, show differences at the amino acid sequence level. Furthermore, we demonstrate different expansion of particular subfamilies of retrotransposons between the lineages, suggesting different impacts of retrotranspositions on human and chimpanzee evolution. The genomic changes after speciation and their biological consequences seem more complex than originally hypothesized.

To understand the genetic basis of the unique features of humans, a number of pilot studies comparing the human and chimpanzee genomes have been conducted[1–5]. Estimates of nucleotide substitution rates of aligned sequences range from 1.23% by bacterial artificial chromosome (BAC) end sequencing[3] to about 2% by molecular analysis[1,6–8], whereas the overall sequence difference was estimated to be approximately 5% by taking regions of insertions or deletions (indels) into account[9]. Chromosomal rearrangements including duplications, translocations and transpositions have also been identified[10,11]. However, owing to technological limitations there is not an integrated picture of the dynamic changes of the genome, thus a gold standard is required to evaluate the overall consequence of these genetic changes on human evolution.

To address these issues and to be able to detect molecular blueprints that have shaped the two genomes, we have conducted a human–chimpanzee whole-chromosome comparison at the nucleotide sequence level on human chromosome 21 (HSA21) and its orthologue chimpanzee chromosome 22 (PTR22). HSA21 is one of the most well characterized human chromosomes[7,12–14] and serves as a representative of the human genome by having characteristic features such as uneven distribution of G+C content with a high correlation to gene density, and repetitive/duplicated structures, allowing for detailed long-range comparative studies with PTR22. Moreover, molecular analysis of HSA21 and its genes is of central medical interest because of trisomy 21, the most common genetic cause of mental retardation in the human population. One case of trisomy 22 in chimpanzee has been reported, with phenotypic features similar to human Down's syndrome[15]. Therefore, our analysis of these chromosomes should reveal dynamic changes that may reflect general evolutionary events occurring throughout the human genome.

Mapping, sequencing and overview of PTR22

We used three different BAC libraries prepared from genomic DNA originating from three male chimpanzees (*Pan troglodytes*). Sequence coverage of the euchromatic portion of the long arm of chromosome 22 (PTR22q) is estimated to be 98.6% (33.3 megabases (Mb)). Accuracy was calculated as 99.9983% from the overlapping clone sequences and ≥99.9981% on the basis of Phrap scores[16]. Altogether, these efforts enabled us to produce a sequence with the highest possible accuracy to be used for reliable comparative analysis (see Supplementary Information for details).

The overall structural features of PTR22q are almost the same as those of HSA21q. The G+C content of these chromosomes is around 41% (Table 1). The corresponding regions between HSA21q and PTR22q, where the extra regions (see Methods) are excluded, show a roughly 400-kilobase (kb) or 1.2% difference in size, with HSA21q being larger than PTR22q. The difference is mainly due to interspersed repeats (ISRs) and simple repeats, representing 63.2% (53.7% and 9.5%, respectively) of the regions corresponding to the gaps in PTR22q. The pericentromeric copy of a 200-kb region found duplicated in HSA21q is missing in PTR22q, as reported previously[17]. We also detected human-specific sequences that are neither repetitive nor low complexity and are unique in the nr data set of NCBI (http://www.ncbi.nih.gov/). For example, a 1,245-base-pair (bp) insertion found in the first intron of

Table 1 **Statistics on HSA21q and PTR22q**

Genome characteristic	HSA21q		PTR22q	
Size (bp)*	33,127,944		32,799,845	
Unaligned sites†	25,242		101,709	
Sequencing gaps	14		22	
Clone gaps‡	3		2	
Estimated total clone gap size	73,108		74,311	
G+C content (%)	40.94		41.01	
CG dinucleotides	361,259		358,450	
CpG islands	950		885	
Nucleotide diversity (%)	0.072		0.14	
	HSA21q		PTR22q	
Repeats	Bp	Number	Bp	Number
SINEs	3,649,153	15,137	3,614,825	15,048
Young Alu elements§	21,557	75	2,606	10
LINEs	5,853,821	8,737	5,736,911	8,673
Young L1 elements‖	82,493	48	78,657	55
LTRs	3,621,501	7,282	3,550,807	7,180
Transposons	949,215	3,363	945,129	3,350
RNAs¶	8,830	100	8,722	99
Satellites	19,327	21	14,773	18
Others	30,452	38	34,776	43
Total	14,132,299		13,905,943	
	42.7%		42.4%	

*Size of the contig data after the site where the first base of the PTR22q contig is aligned.
†Regions extended into HSA21q clone gaps and subtelomeric unmatched regions.
‡Excluding pericentromeric and subtelomeric gaps.
§AluYa5, AluYa8, AluYb8 and AluYb9.
‖L1Hs and L1PA2.
¶Small nuclear RNA, small cytoplasmic RNA, 5S ribosomal RNA, transfer RNA, 7SL RNA and other small RNA genes.