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.

PFKL in HSA21q was confirmed to be human-specific and even locus-specific by searching for similar sequences in the nr database. Four expressed sequence tag (EST)/complementary DNA (cDNA) sequences (BQ711940, BQ706616 and AA453553 on the *PFKL* strand and AJ003358 on the complementary strand) are mapped specifically onto this region.

We did not observe significant correlations between the frequencies of base substitution and indels. Two large indel hotspots were found at around 9.5–11.5 Mb and 16.5–17.5 Mb from the centromere, as previously suggested[3,14] (Fig. 1b). In addition, we found large human insertions/chimpanzee deletions in the first introns of the *NCAM2* (~10 kb) and *GRIK1* (~4 kb) genes, which are both related to neural functions.

One of the largest structural changes identified is a 54-kb region located 11.4 Mb from the centromere in HSA21q but that is absent in PTR22q. This region is flanked by HSAT5 satellite repeats and consists of 164 fragments from 64 different long terminal repeat (LTR) elements with inversions of small portions at even intervals, suggesting rearrangement driven by interspersed repetitive elements. The most distal 25,242-bp region of HSA21q ending with a TTAGGG telomeric repeat did not align to the last 80,879 bp of PTR22q, which instead shares high similarity to human chromosomes 2, 9 and 10. This suggests that the subtelomeric region of PTR22q might be larger than that of HSA21q. The pericentromeric regions of these chromosomes have complex structures: those of PTR22q show a high degree of similarity to human acrocentric chromosomes 13, 15, and 18 (submetacentric only in humans but acrocentric in all great apes) as well as HSA21 (data not shown).

Base substitutions

The overall nucleotide substitution level in aligned regions between PTR22q and HSA21q is about 1.44% (excluding indels), which is significantly higher than the calculation based on high-quality BAC end sequence data (1.23%)[3]. Similar trends have also been observed for HSA21q based on smaller comparisons with the chimpanzee genome[18].

The distribution of base substitution rate along the chromosome is shown in Fig. 1a. The frequency of base substitution is distributed around 1.44% 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[12]. Notably, no protein-coding genes have been identified in this highly conserved region. The correlation between the G+C content and the base substitution rate increases along the chromosome, and is especially high in the last 5 Mb of the telomeric region of PTR22q (data not shown). The frequency of base substitution in repetitive sequences also tends to vary, increasing from CR1/long interspersed element (LINE) (divergence, 1.13%; G+C percentage, 39.0%) and L1/LINE repeats (1.38%; 35.6%) to Alu/short interspersed element (SINE) (1.81%; 51.7%), ERVK/LTR (1.88%; 44.6%), CpG islands (2.26%; 65.1%) and simple repeats (4.06%; 44.8%), as discussed previously[19]. The CG dinucleotide frequencies are significantly ($P < 0.01$) different between PTR22q and HSA21q (Table 1).

Repetitive elements

As described above, HSA21q is about 1.2% longer in size than PTR22q. This difference can mostly be explained by the fact that several subfamilies of transposable elements, such as L1Hs (11 versus 2), MER83B (11 versus 0), AluYa5 (23 versus 3) and AluYb8 (37 versus 2), are more common in human than in chimpanzee[20–23]. Five LTR subfamilies (LTR/ERV1) are more abundant in HSA21q. All MER4A1-int and MER83B-int elements are specific to HSA21q and are clustered in limited regions (positions 21,173,756–21,180,021 bp and 40,651,714–40,657,562 bp, respectively) with overlapping tandem repeats.

According to the sequence alignments, single copies of L1Hs and AluYa5 and two copies of AluYb8 are found in both HSA21q and PTR22q, indicating their presence in the genome of the last common ancestor. All of the seven AluYb9 elements found in HSA21q and the one in PTR22q are lineage-specific, suggesting that these elements have been integrated after speciation. Although the AluYa8 subfamily is thought to be a recent derivative of AluYa5 (ref. 24), we found a few AluYa8 units in both species.

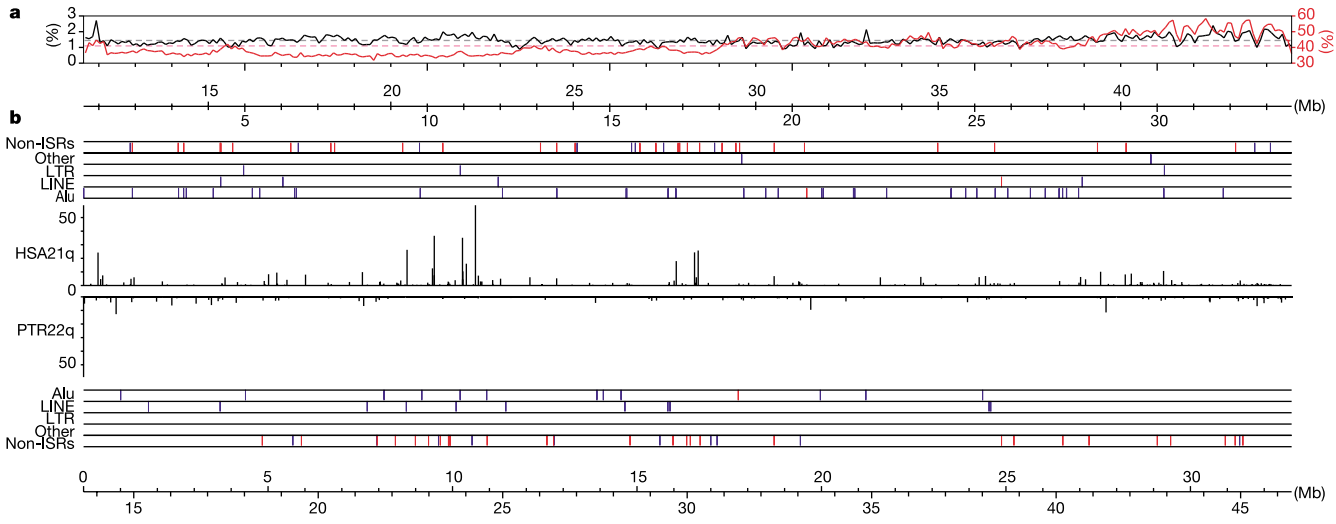

Figure 1 Overview of the differences between HSA21q and PTR22q. **a**, Nucleotide divergence level between HSA21q and PTR22q (bin size = 100 kb) is shown in black and G+C content (%) is in red. Dashed black and red lines indicate average values for divergence level and G+C content, respectively. Abscissa origin is the centromere–euchromatin boundary position in HSA21q (11,000,001 bp) and the scale is in megabases. **b**, *In silico* D-loop (a computer simulation of a hypothetical hybridization experiment using two chromosomal DNAs). The top half represents HSA21q and the bottom half represents PTR22q. Each hypothetical loop-out (regions of DNA that do not hybridize each other) is shown as a vertical line. The *y* axis indicates size of loop-outs (kb). Positions of experimentally verified lineage-specific insertions (blue) and deletions (red) are shown in five lanes along the chromosomes. The lanes, from inside to outside, represent the content of the indels: Alu elements, LINEs, LTRs, other ISRs (Other) and non-ISRs. The scale bar indicates the physical position in megabases relative to the telomere of the short arm (outside track) and to the centromere (inside track).

Lineage-specific insertions and deletions

Through alignment of the high-quality chromosomal sequences of HSA21q and PTR22q we identified about 68,000 indels in total. Greater than 99% of the indels are shorter than 300 bp, but there is a clear abundance of those around 300 bp in size (Fig. 2). These sites are probably produced either through human insertions/chimpanzee deletions or vice versa. Thus the precise identification of these molecular events in the two genomes is essential to understand the processes underlying human and chimpanzee evolution. For this purpose, we tested 567 indels larger than 300 bp using DNA samples from five human, five chimpanzee, one gorilla and two orang-utan individuals by polymerase chain reaction (PCR) amplification using the same primer sets to classify in which lineage these indels arose (see Methods and Supplementary Information). We compared the size of the successfully amplified DNA fragments from 219 indels, of which 193 showed lineage-specific changes in size. Thus, we were able to distinguish insertion from deletion events independently in human and chimpanzee lineages, and to estimate the original state of these regions in the genome of the last common ancestor.

We then classified the indels based on their contents (Fig. 1b). Insertions were mostly produced by the integration of Alu and L1 elements, whereas deletions were not related to particular repetitive structures except in a few cases. We observed different distributions of newly integrated Alu elements between HSA21q and PTR22q: 56% of new Alu elements in HSA21q are inserted in the half of the chromosome with high G+C content, whereas 70% in PTR22q are in the half with low G+C content; new LINEs are more frequent in the half with low G+C content of both chromosomes.

The plots of human insertions and chimpanzee insertions show different multimodal curves (Fig. 3). On the basis of the positions of the insertion sites on HSA21q and PTR22q, we found that most (41 and 13, respectively) of the insertions (300–350 bp in length) were members of the AluY family in both chromosomes. In contrast, only a smaller number of insertions, mostly L1 and LTR elements, were found in the 370–1,000-bp size range. Notably, human and chimpanzee deletion plots form a similar linear line, suggesting a relationship between logarithmic size of deletions and the cumulative frequency in both species (Fig. 3).

We also identified integration of L1PA2 elements after human–chimpanzee speciation, indicating that L1PA2 has been active in both human and chimpanzee lineages, although the activity seems to be lower in the human lineage. Two L1PA2 elements reside in different strands but overlap in a single inserted region in PTR22q, suggesting a single L1PA2 integration event followed by an inversion event within the same region. We also found that some insertions in PTR22q lie within Alu elements (mostly AluSx) on the same strand.

Unlike the insertions, deletions do not correspond exactly to any ISR elements, indicating that deletion events are independent of ISRs. However, one of the deleted regions in HSA21q corresponds perfectly to a single AluY element in PTR22q, and a deleted region in PTR22q corresponds almost perfectly to a single AluYb8 element in HSA21q. In the former case, there are two identical 10-mer segments around the deleted AluY element, and in the latter case, the AluYb8 element is embedded within a single AluSx element at a site in the 14-bp A-rich region in the middle of the Alu element, generating 14–15-bp poly-A/T stretches around AluYb8. Thus, the deletion of these elements may also have been generated by homologous recombination between these relatively short identical or similar flanking segments.

Calculations from the indels in the 300–5,000-bp range indicate that both chromosomes have undergone a net loss in size since speciation despite frequent insertion events: HSA21q has gained 32 kb but lost 39 kb, whereas PTR22q has gained 25 kb and lost 53 kb. This suggests that the ancestral chromosome was larger than both HSA21q and PTR22q, and that PTR22q has suffered more losses than HSA21q since speciation. The large indels (>5 kb) detected in the sequences, which were experimentally confirmed, are found in the pericentromeric, 10 Mb, 17 Mb and 29 Mb regions. HSA21q has more indels greater than 10 kb than PTR22q.

With the knowledge of which Alu family element was inserted after speciation, we carried out an evolutionary analysis of the AluY families that have been inserted into HSA21q and PTR22q. A neighbour-joining analysis revealed that such AluY elements can be largely separated into chimpanzee and human groups and suggesting contribution from a few active elements (Fig. 4). Taken as a whole, these results indicate that the expansion of particular elements was repeated several times during the course of evolution. Humans seem to have experienced such expansions more frequently and more recently than chimpanzees. If we could determine the oldest expansion event through genomic comparison, we might be able to identify whether such an Alu burst was the driving force for speciation between the two species from the common ancestor. Amplification of Alu elements during the evolution of primates and alteration of gene functions through the insertion of repetitive elements has been discussed in many previous studies[25–34]. However, further wide-ranging analyses comparing the chimpanzee and other primate genomes is necessary to clarify these points.

Chimpanzee and human single nucleotide polymorphisms

Single nucleotide polymorphisms (SNPs) provide important clues for detecting ancestral and mutant alleles within the human

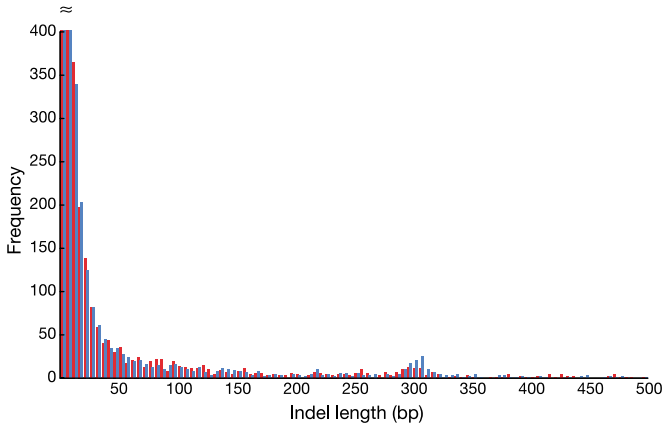

Figure 2 Size distribution of indels. All of the indels are calculated as insertions either in HSA21q (blue) or PTR22q (red). The first two bins are off the scale.

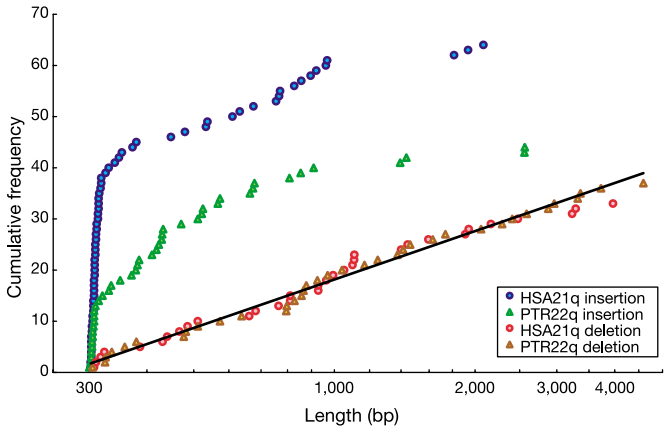

Figure 3 Size-dependency of indel frequency. Cumulative counts of experimentally determined lineage-specific insertions and deletions. The *x* axis represents the log-scaled size of indels (bp), whereas the *y* axis represents the cumulative counts of lineage-specific insertions and deletions.

population. The chimpanzee genomic sequence is the best resource for inferring the ancestral allele of any human SNP locus. We thus unambiguously reconstructed 19,985 ancestral states from 21,435 human SNP sites in HSA21q[7].

Transitional changes are more frequent than transversions, as expected. Among transitions, G→A and C→T changes (19.6% and 20.3%, respectively)[35] are more frequent than A→G and T→C changes (14.0% and 15.1%, respectively). This substitution pattern is compatible with the fact that the G+C content of the human genome is lower than 50%. We conducted the same analysis for 5,781 chimpanzee SNPs obtained from overlapping BAC sequences (Supplementary Table 9A, B), and found that both A→G and T→C transitions (12.6% and 13.2%, respectively) are slightly lower than those for human.

Because SNPs have been created relatively recently during human and chimpanzee evolution (approximately 0.5–1.0 million years ago), nucleotide substitution patterns predicted from SNP data may be slightly different than those based on current G+C content (that is, 41% for both human and chimpanzee). As the estimated equilibrium G+C content for the human genome is 0.422, versus 0.405 for chimpanzee, modern humans seem to have undergone a slight increase in G+C content compared with their ancestors, whereas chimpanzees have undergone no clear change, although the equilibrium values for humans and chimpanzees are not far from the current G+C contents (Supplementary Table 9C, D).

We then conducted an *H* test[36] (Supplementary Table 10), and detected 18 10-kb DNA regions as candidates of positive selection. Three known genes, *KCNE1, DSCR2* and *B3GALT5*, were located in these regions. *KCNE1* and *B3GALT5* were also identified as relatively rapidly evolving genes through analysis of the ratio of the number of amino acid substitutions per site to silent substitutions per site (K_A/K_S), as discussed below (see Supplementary Table 6).

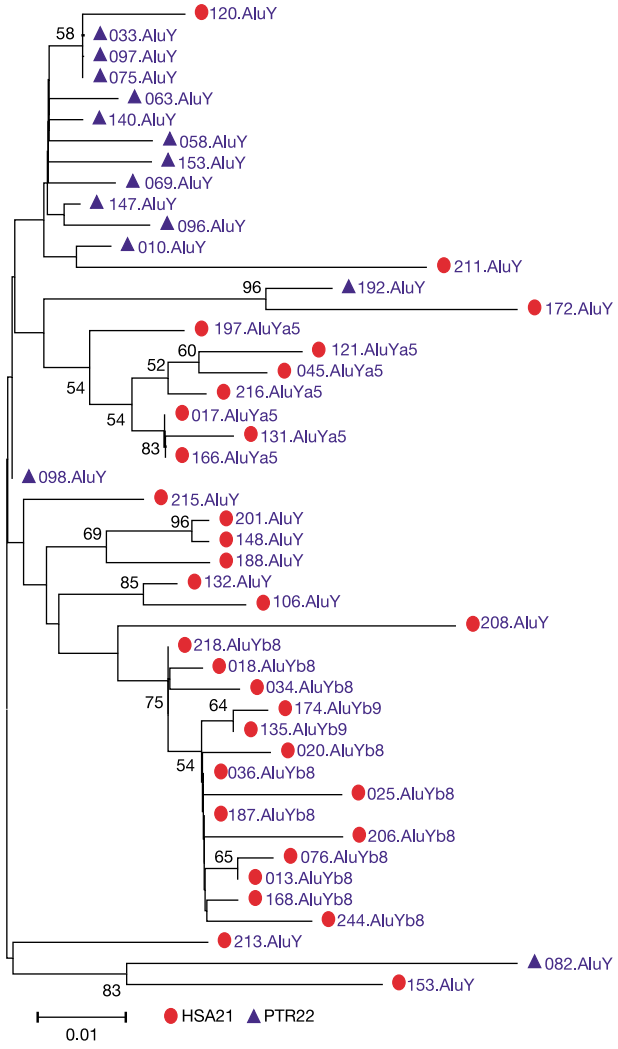

Figure 4 Evolutionary relationships among young AluY elements. Neighbour-joining tree of all AluY families that have been inserted into HSA21q (red circles) or PTR22q (blue triangles) after speciation. Bootstrap proportions greater than 50% are indicated at the corresponding branches (1,000 replicates). The scale indicates the evolutionary distance of 0.01.

Gene catalogue and characterization of coding sequences

We have annotated 284 protein-coding genes and 98 pseudogenes for HSA21q[12] (http://chr21.molgen.mpg.de), and 272 genes and 89 pseudogenes for PTR22q (the comparative gene catalogue of HSA21q and PTR22q, including pseudogenes, is given in Supplementary Table 3). We lacked information for six genes and 11 pseudogenes on HSA21q located partly or completely in sequencing gaps of PTR22q, and one pseudogene, *OR7E92P*, was absent in PTR22q (Supplementary Table 3). All of the conserved pseudogenes were matched in size between human and chimpanzee, except for *KRTAP21P1*, which is non-processed in HSA21q but processed in PTR22q.

Six HSA21q genes displaying hallmarks of retrogenes (see ref. 37 for review) were not found in PTR22q and were probably inserted during human evolution (or less likely, deleted during chimpanzee evolution). These are *H2BFS*, a member of the histone family S, and five members of the keratin-associated protein (KAP) gene cluster in 21q22.1. *H2BFS* was not found in the syntenic region of mouse chromosome 16, suggesting that it was inserted in HSA21q. The KAP gene cluster is too divergent in mouse to establish a conclusive comparison. Three intronless open reading frames (ORFs) have been inserted in chimpanzee PTR22q (or deleted in HSA21q). These are *HNRPA1LK1*, a heterogeneous nuclear ribonucleoprotein, *RPLP1LK1*, a ribosomal protein, and *FAM28ALK1*, a gene of unknown function; all three ORFs are absent in the mouse syntenic region. All but one of the ORFs found in either human insertions/chimpanzee deletions or human deletions/chimpanzee insertions are intronless and probably represent retrogenes.

We found a ribosomal protein pseudogene (*RPL13AP*) on HSA21q with an intact ORF (*RPL13ALK1*) in PTR22q, allowing for possible functionality. Four HSA21q coding sequences (*C21orf81* and three intronless genes: *C21orf115, C21orf104* and *C21orf19*) have interrupted ORFs in PTR22q, precluding their functionality, and these were annotated as pseudogenes in chimpanzee (Supplementary Table 3).

All other HSA21q genes are potentially active in PTR22q, even though several genes show human-specific transcriptional isoforms due to alteration or inactivation of some of the transcript isoforms in chimpanzee (see below). The minimum nucleotide sequence identity is 83% (*KRTAP6-3*). Of the 272 annotated chimpanzee genes, we compared the human and chimpanzee coding sequences in 231 genes for which we could define a non-ambiguous ORF in both species. We omitted for that comparison the 41 entries of the chimpanzee catalogue for which we had no ORF in one of the two species (for example, ambiguous ORFs associated to short ESTs containing mostly untranslated regions (UTRs)) and genes corresponding to pseudogenes in the other species. Among the 231 genes associated to a canonical ORF, 179 show a coding sequence of identical length in human and chimpanzee and exhibit similar intron–exon boundaries. For those 179 genes, 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 sequence between human and chimpanzee, including seven in which the nucleotide sequence of the coding region is also identical (Supplementary Table 3). Examples of biological processes that are perfectly conserved involve transcriptional regulators (*RUNX1*,

PKNOX1, ERG, GABPA, U2AF1), metabolic enzymes (*ATP5O, CBR1, PFKL, CRYZL1, SOD1, ABCG1*), gene products associated with signal transduction generally showing patterned expression during development (*GRIK1, CXADR, DSCR5, PCP4, DSCAM, DYRK1A, S100B, C21orf4*) and gene products involved in protein folding and degradation (*SMT3H1, UBE2G2*)[38]. A total of 140 of these 179 genes show amino acid replacements, but no gross structural changes are expected.

In contrast, 47 PTR22q genes show significant structural changes affecting at least one of their transcript isoforms. Fifteen genes have indels within their coding region yet retain frame consistency in all but one case (*TCP10L*) (Supplementary Table 4). Marked changes are observed in *PCNT2*, a component of the filamentous matrix of the centrosome initiating the nucleation of spindle microtubules, and in *TCP10L*, a t-complex protein. The third exon of *PCNT2* is shorter in chimpanzee owing to a deletion of 195 bp corresponding to five highly similar copies of a 13-mer repeating unit. Human *PCNT2* has seven repeats, the orthologous chimpanzee gene has only two whereas the mouse has none, suggesting that the repeats were inserted during primate evolution. *TCP10L* has a deletion of 17 nucleotides within the fourth exon in chimpanzee. This generates a transcript that uses the last 16 nucleotides of exon 4 and the adjacent 23 nucleotides that are intronic in human, the gene product of which is predicted to have a short frame shift in the middle of the protein (Supplementary Table 4). Thirty-two genes show changes modifying either the first ATG or the stop codon in at least one of their associated transcripts (Supplementary Table 5).

Five chimpanzee genes could not be classified because they displayed structural changes caused by indels (*SH3BGR, SYNJ1, C21orf96* and *TMPRSS3*) or a substitution in the ATG codon (*C21orf18*). These changes correspond to polymorphisms in human and may not be specific to chimpanzee.

Our data suggest that indels within coding regions represent one of the major mechanisms generating protein diversity and shaping higher primate species. We observed an additional level of functional diversity generated through the occurrence of multiple alternative transcripts for a single gene, some of the isoforms being structurally modified or not functional in chimpanzee. We are currently lacking cDNA information to define precisely the structure of these genes but we provide here a framework allowing for experimental verification of the transcript isoforms.

Taken together, gross structural changes affecting gene products are far more common than previously estimated (20.3% of the PTR22 proteins, as listed in Supplementary Tables 4 and 5). In addition, 87 genes in the catalogue show mutations in at least one of the splice sites. However, in all cases the modified sequence still fits the consensus sequence and is therefore expected to be functional. More subtle effects, such as changes in transcript stability or processing, cannot be ruled out.

K_A/K_S analysis

The neutrality of sequence differences found between orthologous pairs of human and chimpanzee genes can be assessed using K_A and K_S values ($K_A/K_S \cong 1$ indicates neutral evolution; $K_A/K_S > 1$ indicates positive selection, and $K_A/K_S < 1$ indicates negative selection; that is, purifying selection). K_A, K_S and other related values were calculated for 231 genes on PTR22q, with 10% of the genes having a K_A/K_S ratio >1, the highest value being 3.37 for the human hair keratin-associated protein 23-1 gene (*KRTAP23-1*; data are summarized in Supplementary Table 6), although these values were not statistically confirmed (Supplementary Fig. 1).

Relatively rapidly evolving genes may be estimated from K_A, $K_A + K_S$, or just nucleotide divergence values. Three KAP genes, *KCNE1* (human cardiac delayed rectifier potassium channel protein), *TCP10L* (t-complex protein 10A-2), *B3GALT5* (UDP-Gal:betaGlcNAc beta 1,3-galactosyltransferase, polypeptide 5), *IGSF5* (immunoglobulin superfamily-5 like) and several other genes are found in this category (Supplementary Table 6). Genes showing statistically significant low K_A/K_S values are thought to be evolving under purifying selection; for example, *C21orf113, PFKL, AIRE, ITGB2, TMPRSS3* and *AGPAT3*. *PFKL* has a K_A/K_S ratio of 0, and is likely to be under strong purifying selection over the entire gene region.

Recently, Clarke *et al.*[39] reported a similar analysis against 7,645 chimpanzee gene sequences including 70 PTR22q gene sequences, in which *KRTAP23-1* was not included. Furthermore, we found discrepancies with a previous analysis reporting K_A/K_S values for some of the HSA21q genes[40]. One explanation is that the previous work used only PCR-amplifiable exons for the analysis.

Comparative gene expression analysis

Using Affymetrix HG U95 arrays, we compared the gene expression profiles of HSA21q genes between humans and chimpanzees in two tissues: 202 genes in brain (cingulate cortex, HG U95Av2, B, C, D and E) and 96 genes in liver (HG U95Av2). We detected 60 genes expressed in brain (this study) and 40 in liver[41] in at least one species. Of these, 9 in the brain and 12 in the liver showed a significant change in expression level between humans and chimpanzees in the range of a 1.5–10-fold difference (Supplementary Table 7). Overall, the proportion of genes showing changes in expression level on HSA21 was not significantly different from the rest of the genome (liver: $\chi^2 = 1.87$, $P = 0.172$; brain: $\chi^2 = 1.12$, $P = 0.290$)[41].

We explored whether the sequence divergence in the different gene regions can predict how a gene is differently expressed between the two species using a Mann–Whitney *U*-test. The data show that there is a trend for regions with a high divergence in the 5′ UTR to differ in messenger RNA levels between humans and chimpanzees ($N = 18, 47$; $P = 0.043$). Notably, genes in which the sequence divergence of an associated CpG islands is high are also more likely to have changed their expression ($N = 13, 39$; $P = 0.046$). However, as CpG islands often stretch into the first exon, the correlations of CpG island and 5′ UTR divergence with expression changes are not independent. In contrast, we found no significant association between the divergence of non-degenerate sites, 3′ UTRs, intergenic and intronic regions and variation in expression levels (Supplementary Table 8). It has been proposed that 5′ UTRs in humans might have been under positive selection[42], possibly due to their involvement in the regulation of gene expression levels.

Some of the 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[43]. For instance, *IFNAR2, IFNGR2, CXADR, ITSN1* and *CRYZL1* are directly or indirectly involved in the immune response against various pathogens. *SH3BGR* is strongly expressed in the developing heart, *C21orf2* is expressed in the peripheral nervous system, *SYNJ1* and *ANKRD3* are signalling molecules acting in early brain development, *MCM3AP* is associated with cell cycle progression, *ETS2* is a transcription factor essential for embryonic development and *COL18A1* is a collagen gene that is mutated in human Knobloch syndrome associated with encephalocele[44].

Promoter analysis

We analysed the upstream region of genes that showed significant expression changes in liver and brain between human and chimpanzee, as well as the upstream region of each corresponding mouse gene. Computational analysis of the transcription-factor-binding sites within the 1-kb upstream region of each gene is summarized in Supplementary Table 11. Transcription-factor-binding sites were compared in the three species, and those specific to either human or chimpanzee were found in most genes. All of the specific transcription-factor-binding sites were caused by base substitutions in either human or chimpanzee, but these may not clearly account for the expression changes observed in this study. To assess precisely the

expression changes, further analysis of promoter regions as well as other factors, such as enhancers and suppressors located outside the promoter region and the effects of 5′ and 3′ UTRs on mRNA stability, are needed.

Conclusion

This study shows a chromosome-wide comparison between human and chimpanzee based on high-quality sequences, and provides the first integrated picture of genetic changes during human evolution. The data presented here suggest that the biological consequences due to the genetic differences are much more complicated than previously speculated. We hope that our work offers a framework for the design of future studies to examine differences between the two species. □

Methods

Mapping, sequencing and data availability

The details concerning mapping and sequencing are summarized in Supplementary Tables 1 and 2 (see also http://chimp22pub.gsc.riken.jp). Briefly, three BAC libraries—PTB1 (ref. 3), RPCI-43 and CHORI-251 (http://bacpac.chori.org), constructed from three male individuals—and a chimpanzee chromosome 22 fosmid library—PTF22 (12-fold coverage)—were used to isolate clones for this analysis. The first set of the seed clones was selected from BAC end data[3] and PCR screening of expanded BAC libraries using high-density human STS primers placed at roughly 20-kb intervals. Only the clones having multiple STS markers and sharing common STSs with neighbouring clones were considered as a member of a clone contig. These clones were then subjected to partial sequencing, from which new chromosome-walking primers were designed for further screening. Clone overlaps were then examined before choosing a set of minimum-tiling-path clones for full-scale high-quality sequencing. These steps were repeated until all gaps were closed or no additional clones could be identified. Some representative clones in contigs were also examined cytogenetically to confirm their localization on the particular chromosome; however, not all such clones were included in the minimum tiling path. Most of the pericentromeric and subtelomeric regions are not included in this study because of the difficulty in identifying clones with good sequence matches. Two PTR22 clones, PTB-242K04 and CH251-010A09, have a terminal region that extends into one of the three clone gaps in HSA21q by 11,546 bp in total. Another clone, PTB-190I13, spans another HSA21q gap with a 9,284-bp region for the gap where the G+C content is high (54.2%,) and a sequence gap of 1,165 bp in length still remains.

All of the clone data have been released from DDBJ/EMBL/GenBank (accession numbers are listed in Supplementary Table 1).

Alignment between HSA21q and PTR22q and contig construction

Each PTR22 clone sequence was aligned to the HSA21q data using NCBI BLAST2. From the BLAST hits, we chose the best match for each site of the corresponding region in HSA21q. Details of the alignment procedure can be seen in Supplementary Information. The clone contig was generated by aligning the overlapping regions of each neighbouring clone. We chose to include the sequence of the clone for the overlap with which the alignment with the counterpart region in HSA21q showed a lower divergence level. On the basis of this clone contig map, the region represented in the contig is extracted from each clone alignment to produce the whole chromosome alignment. To fix any small inconsistencies around the clone boundaries, the final alignment was checked manually. The nucleotide divergence level was calculated from the regions that aligned best with HSA21q.

Detection of lineage-specific insertions and deletions

We selected 567 apparent insertions ≥300 bp from both the human and chimpanzee sequences and designed PCR primers from their flanking sequences. Using genomic DNA samples from five chimpanzees, five humans, one gorilla and two orang-utans as a template, the size of the PCR products was examined. After amplification, reaction products were separated through 1% agarose gel electrophoresis for comparison of the product sizes. We used only the indels that showed a significant size difference between chimpanzee and human, and one of them was of equivalent size to gorilla and orang-utan, for the analysis (*t*-test).

SNP analysis

The HSA21q SNPs[7] in dbSNP (http://www.ncbi.nlm.nih.gov/SNP/) were compared to the corresponding PTR22q sites to infer the ancestral states of the two human alleles. We developed a modified version of the *H* test[36] based on the computer program at http://crimp.lbl.gov/htest.html so as to detect positive selection disregarding singletons (details in Supplementary Information). The pattern of nucleotide substitutions was also estimated using the human SNP data and chimpanzee BAC sequence overlap data generated in the present study. We estimated the substitution pattern[35] and equilibrium frequencies from the nucleotide substitution matrix[45,46].

Genomic annotation

Protein-coding genes in the PTR22q sequence were annotated by extracting the orthologous regions on HSA21q (in the gene catalogue) from the genomic sequence. In addition, PTR22 sequences with no match to the HSA21 gene catalogue were blasted against the complete nr-db, and cDNA matches were retained as potential PTR22-specific genes. Gene orthology with mouse, rat, zebrafish, pufferfish and *Ciona* was calculated from the HSA21 genes, and conceptual translations and multiple alignments were constructed using CLUSTALW. Detailed information is provided in the Supplementary Information.

Comparative gene expression analysis

For all arrays only the oligonucleotides that matched perfectly between the human and chimpanzee sequences were used for analysis. Gene expression levels were compared separately for brain and liver in all nine possible pairwise comparisons between the three individuals of each species.

Received 16 February; accepted 14 April 2004; doi:10.1038/nature02564.

1. King, M. C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees. *Science* **188,** 107–116 (1975).
2. Pennacchio, L. A. & Rubin, E. M. Genomic strategies to identify mammalian regulatory sequences. *Nature Rev. Genet.* **2,** 100–109 (2001).
3. Fujiyama, A. *et al.* Construction and analysis of a human–chimpanzee comparative clone map. *Science* **295,** 131–134 (2002).
4. Olson, M. V. & Varki, A. Sequencing the chimpanzee genome: insights into human evolution and disease. *Nature Rev. Genet.* **4,** 20–28 (2003).
5. Boffelli, D. *et al.* Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* **299,** 1391–1394 (2003).
6. Hacia, J. G. *et al.* Evolutionary sequence comparisons using high-density oligonucleotide arrays. *Nature Genet.* **18,** 155–158 (1998).
7. Patil, N. *et al.* Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294,** 1719–1723 (2001).
8. Thomas, J. W. *et al.* Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* **424,** 788–793 (2003).
9. Britten, R. J. Divergence between samples of chimpanzee and human DNA sequences is 5%, counting indels. *Proc. Natl Acad. Sci. USA* **99,** 13633–13635 (2002).
10. Nickerson, E. & Nelson, D. L. Molecular definition of pericentric inversion breakpoints occurring during the evolution of humans and chimpanzees. *Genomics* **50,** 368–372 (1998).
11. Bailey, J. A. *et al.* Human-specific duplication and mosaic transcripts: the recent paralogous structure of chromosome 22. *Am. J. Hum. Genet.* **70,** 83–100 (2002).
12. Hattori, M. *et al.* The DNA sequence of human chromosome 21. *Nature* **405,** 311–319 (2000).
13. Frazer, K. A. *et al.* Evolutionarily conserved sequences on human chromosome 21. *Genome Res.* **11,** 1651–1659 (2001).
14. Frazer, K. A. *et al.* Genomic DNA insertions and deletions occur frequently between humans and nonhuman primates. *Genome Res.* **13,** 341–346 (2003).
15. McClure, H. M., Belden, K. H., Pieper, W. A. & Jacobson, C. B. Autosomal trisomy in a chimpanzee: resemblance to Down's syndrome. *Science* **165,** 1010–1012 (1969).
16. Ewing, B., Hillier, L., Wendl, M. C. & Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8,** 175–185 (1998).
17. Orti, R. *et al.* Conservation of pericentromeric duplications of a 200-kb part 25 of the human 21q22.1 region in primates. *Cytogenet. Cell Genet.* **83,** 262–265 (1998).
18. Ebersberger, I., Metzler, D., Schwarz, C. & Paabo, S. Genomewide comparison of DNA sequences between humans and chimpanzees. *Am. J. Hum. Genet.* **70,** 1490–1497 (2002).
19. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420,** 520–562 (2002).
20. Sawada, I. *et al.* Evolution of Alu family repeats since the divergence of human and chimpanzee. *J. Mol. Evol.* **22,** 316–322 (1985).
21. Myers, J. S. *et al.* A comprehensive analysis of recently integrated human Ta L1 elements. *Am. J. Hum. Genet.* **71,** 312–326 (2002).
22. Liu, G. *et al.* Analysis of primate genomic variation reveals a repeat-driven expansion of the human genome. *Genome Res.* **13,** 358–368 (2003).
23. Kim, M., Carman, C. V. & Springer, T. A. Bidirectional transmembrane signaling by cytoplasmic domain separation in integrins. *Science* **301,** 1720–1725 (2003).
24. Roy, A. M. *et al.* Potential gene conversion and source genes for recently integrated Alu elements. *Genome Res.* **10,** 1485–1495 (2000).
25. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
26. Britten, R. J., Baron, W. F., Stout, D. B. & Davidson, E. H. Sources and evolution of human Alu repeated sequences. *Proc. Natl Acad. Sci. USA* **85,** 4770–4774 (1988).
27. Batzer, M. A. & Deininger, P. L. A human-specific subfamily of Alu sequences. *Genomics* **9,** 481–487 (1991).
28. Leeflang, E. P., Chesnokov, I. N. & Schmid, C. W. Mobility of short interspersed repeats within the chimpanzee lineage. *J. Mol. Evol.* **37,** 566–572 (1993).
29. Zietkiewicz, E., Richer, C., Makalowski, W., Jurka, J. & Labuda, D. A young Alu subfamily amplified independently in human and African great apes lineages. *Nucleic Acids Res.* **22,** 5608–5612 (1994).
30. Batzer, M. A. *et al.* Dispersion and insertion polymorphism in two small subfamilies of recently amplified human Alu repeats. *J. Mol. Biol.* **247,** 418–427 (1995).
31. Roy, A. M. *et al.* Recently integrated human Alu repeats: finding needles in the haystack. *Genetica* **107,** 149–161 (1999).
32. Chou, H. H. *et al.* Inactivation of CMP-N-acetylneuraminic acid hydroxylase occurred prior to brain expansion during human evolution. *Proc. Natl Acad. Sci. USA* **99,** 11736–11741 (2002).
33. Roy-Engel, A. M. *et al.* Non-traditional Alu evolution and primate genomic diversity. *J. Mol. Biol.* **316,** 1033–1040 (2002).
34. Deininger, P. L. & Batzer, M. A. Mammalian retroelements. *Genome Res.* **12,** 1455–1465 (2002).
35. Gojobori, T., Li, W. H. & Graur, D. Patterns of nucleotide substitution in pseudogenes and functional genes. *J. Mol. Evol.* **18,** 360–369 (1982).
36. Fay, J. C. & Wu, C. I. Hitchhiking under positive Darwinian selection. *Genetics* **155,** 1405–1413 (2000).
37. Brosius, J. RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. *Gene* **238,** 115–134 (1999).

38. The HSA21 expression map initiative. A gene expression map of human chromosome 21 orthologues in the mouse. *Nature* **420,** 586–590 (2002).
39. Clark, A. G. *et al.* Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science* **302,** 1960–1963 (2003).
40. Shi, J. *et al.* Divergence of the genes on human chromosome 21 between human and other hominoids and variation of substitution rates among transcription units. *Proc. Natl Acad. Sci. USA* **100,** 8331–8336 (2003).
41. Enard, W. *et al.* Intra- and interspecific variation in primate gene expression patterns. *Science* **296,** 340–343 (2002).
42. Hellmann, I. *et al.* Selection on human genes as revealed by comparisons to chimpanzee cDNA. *Genome Res.* **13,** 831–837 (2003).
43. Varki, A. A chimpanzee genome project is a biomedical imperative. *Genome Res.* **10,** 1065–1070 (2000).
44. Suzuki, O. T. *et al.* Molecular analysis of collagen XVIII reveals novel mutations, presence of a third isoform, and possible genetic heterogeneity in Knobloch syndrome. *Am. J. Hum. Genet.* **71,** 1320–1329 (2002).
45. Dayhoff, M. O., Schwartz, R. M. & Orcutt, B. C. in *A Model of Evolutionary Change in Proteins* (ed. Dayhoff, M. O.) (Natl Biomed. Res. Found., Silver Springs, Maryland, 1978).
46. Tajima, F. & Nei, M. Biases of the estimates of DNA divergence obtained by the restriction enzyme technique. *J. Mol. Evol.* **18,** 115–120 (1982).

Supplementary Information accompanies the paper on **www.nature.com/nature**.

Acknowledgements We are grateful to T. Ito, C. Kawagoe, T. Kojima, X. Son, A. Beck, K. Borzym, S. Gelling, V. Gimmel, K. Heitmann, A. Kel, S. Klages, N. Lang, I. Mueller, M. Sontag, R. Yildirimman, J. Wickings, C. Baumgart, O. Mueller, T. T. Liao, H. Tsai, Y. Huang, Y. Liu and all the technical staff of the contributing genome centres. This work was supported in part by a Special Fund for RIKEN Genomic Sciences Center and Grant-in-Aid for Scientific Research on Priority Areas 'Genome Science' from the Ministry of Education, Culture, Sports, Science and Technology, Japan; the Ministry of Education and Research, Germany; the Chinese International Science and Technology Cooperation Project, Ministry of Science and Technology, China; the Chinese High-Tech Research and Development Program, Shanghai Commission for Science and Technology; the National Research Program for Genomic Medicine of National Science Council, Taiwan; and the Ministry of Science and Technology, Korea.

Authors' contributions H. Watanabe, A. Fujiyama, M. Hattori, N. Saitou, H.-S. Park, S.-Y. Wang, M.-L. Yaspo and Y. Sakaki contributed to the consortium leadership.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to Y.S. (sakaki@gsc.riken.jp). The accession numbers of the BAC clone sequences have been deposited in DDBJ/EMBL/GenBank and can be found in Supplementary Table 1.

H. Watanabe[1,2]*, A. Fujiyama[1,3], M. Hattori[1,4], T. D. Taylor[1], A. Toyoda[1], Y. Kuroki[1], H. Noguchi[1], A. BenKahla[5], H. Lehrach[5], R. Sudbrak[5], M. Kube[5], S. Taenzer[6], P. Galgoczy[6], M. Platzer[6], M. Scharfe[7], G. Nordsiek[7], H. Blöcker[7], I. Hellmann[8], P. Khaitovich[8], S. Pääbo[8], R. Reinhardt[5], H.-J. Zheng[9], X.-L. Zhang[9], G.-F. Zhu[9], B.-F. Wang[9], G. Fu[9], S.-X. Ren[9], G.-P. Zhao[9], Z. Chen[9,10], Y.-S. Lee[11], J.-E. Cheong[11], S.-H. Choi[11], K.-M. Wu[12], T.-T. Liu[13], K.-J. Hsiao[13,14], S.-F. Tsai[12,13], C.-G. Kim[15], S. OOta[15], T. Kitano[15], Y. Kohara[15], N. Saitou[15], H.-S. Park[11], S.-Y. Wang[9], M.-L. Yaspo[5] & Y. Sakaki[1]

Affiliations for authors: 1, RIKEN, Genomic Sciences Center, Yokohama 230-0045, Japan; 2, Nara Institute of Science and Technology, Ikoma 630-0101, Japan; 3, National Institute of Informatics, Tokyo 101-8430, Japan; 4, Kitasato University, Sagamihara 228-8555, Japan; 5, Max-Planck-Institut für Molekulare Genetik, D-14195 Berlin-Dahlem, Germany; 6, Institut für Molekulare Biotechnologie, Genomanalyse, D-07745 Jena, Germany; 7, German Research Centre for Biotechnology (GBF), D-38124 Braunschweig, Germany; 8, Max-Planck-Institut für Evolutionäre Anthropologie, D-04103 Leipzig, Germany; 9, Chinese National Human Genome Center at Shanghai, Shanghai 201203, China; 10, State Key Laboratory of Medical Genomics, Rui Jin Hospital, Shanghai 200025, China; 11, Genome Research Center, Korea Research Institute of Basic and Biotechnology, Daejeon 305-333, Korea; 12, National Health Research Institutes, Taipei 115, Taiwan; 13, National Yang-Ming University, Taipei 112, Taiwan; 14, Taipei Veterans' General Hospital, Taipei 112, Taiwan; 15, National Institute of Genetics, Mishima 411-8540, Japan

*Present address: Hokkaido University, Sapporo 060-0814, Japan