Supplementary Information

Resources and mapping/sequencing technologies

Construction of the PTB1 library, 10-fold coverage, was described previously (Fujiyama *et al*., *Science* **295**, 131-134, 2002). RPCI-43 and CHORI-251 BAC libraries were obtained from BACPAC Resources (CHORI; http://bacpac.chori.org). In addition, a chimpanzee chromosome 22 fosmid library, PTF22 (coverage=12-fold), was constructed from randomly sheared DNA fragments isolated from flow-sorted chromosomes of the same cell resource as PTB1.

A clone-contig map of chimpanzee chromosome 22 was initially constructed from the BAC-end sequence (BES) map that covers 67% of human chromosome 21 (Fujiyama *et al*., *Science* **295**, 131-134, 2002). STS primers located at roughly 20kb intervals based on the human chromosome 21 sequence, and BES-based primers were used to pick up additional clones. Remaining gaps were filled by a standard chromosome-walking strategy, except for the centromeric and telomeric regions, where the PTF22 library was extensively used as a source of candidate clones that were then tested by both sequencing and FISH analysis. Some gaps were filled by the products of long-range PCR reactions (Expand High Fidelity PCR system, Roche Diagnostics), in which we used purified chimpanzee chromosome 22 DNA as a template to minimize non-specific amplification. Other conditions for the PCR reactions are as recommended by the supplier. FISH analysis was carried out using either human or chimpanzee chromosome spreads prepared by the HANABI system (ADScience, Chiba Japan) and a standard visualization procedure.

We performed paired-end shotgun sequencing for both BAC and fosmid clones. The insert sizes of the shotgun clones were 1 ~ 2-kb or 5 ~ 10-kb. The Phred/Phrap software package (Ewing *et al*., *Genome Res.* **8,** 175-185, 1998) was used for base-calling and assembly. Editing was performed with Consed (Gordon *et al*., *Genome Res.* **8,** 195-202, 1998) or Gap4 (Staden *et al*., *Molecular Biotechnology* **5**, 233-241 1996) to check all low quality bases and the correct assembly. Seven bacterial transposon-derived sequences were identified and manually removed from the original data. Final coverage is estimated to be 98.6% of the

chimpanzee chromosome 22 q-arm euchromatic region (33.3 Mb). We then performed quality assessment as follows. At first, we scrutinized 249 overlapping regions (9,890,299 bases which represents about 15% of the entire non-redundant sequence) between the sequenced chimpanzee clones for errors by looking at the trace data; 165 errors were detected including substitutions and insertions/deletions (INDELs). In addition, several clones were sequenced in duplicate to monitor the quality of the finished data produced from different centers. Some of the differences observed in exons and promoter regions were re-sequenced to confirm the correctness of the sequence. Details are summarized in supplementary Tables 1 and 2.

Alignment between HSA21q and PTR22q

Since we are interested in an in-depth comparison between the genomes of human and chimpanzee, we used a local instead of a global alignment. In some studies, large-scale nucleotide sequence comparisons are carried out with a global alignment method (Liu *et al*., *Genome Res* **13**, 358-368, 2003). However, such comparisons might miss some chromosomal rearrangements in the sequences because of the alignment algorithm, in which every sequence site is aligned to only a single site of the other sequence including gaps in the same order from the 5' to the 3' ends of the sequences compared. Thus, we used a local alignment algorithm to align the PTR22q and the HSA21q sequences in which duplicated regions in a sequence are aligned to the same corresponding region in the other sequence. For the construction of the initial alignments of PTR22q clones to the HSA21q sequences, NCBI BLAST2 (Altschul *et al*. *Nucleic Acids Res* **25**, 3389-3402, 1997) was used. The highest score was chosen for the same query region if there were multiple hits. In some cases, BLAST missed alignments or produced artifactual alignments between evidently orthologous regions between PTR22q and HSA21q, especially between highly repetitive regions. To address this problem, we conducted additional alignment steps for each missed region with MegaBLAST (Zhang *et al*., *J Comput Biol.* **7**, 203-214, 2000), which produces alignments of less precision than BLAST2 but does not have the bugs of BLAST2. The sequencing gaps, represented by N's, in these sequences

were manually aligned to seemingly corresponding regions sand were not treated as species-specific regions in the following analyses.

Gene annotation of the chimpanzee PTR22q sequence

In the first step, we annotated the protein-coding genes of chimpanzee PTR22q by comparative analysis with the updated gene catalog of HSA21q. For this, we have re-annotated the initial HSA21qgene catalog (Hattori *et al*. *Nature* **405**:311-319, 2000). The HSA21qsequence (build July 2002) was cut into overlapping fragments (fragment length=10 kb, overlap=1 kb), and repeats were masked with RepeatMasker (A.F.A. Smit & P. Green, unpublished data); the masked fragments were blasted using blastn against dbEST and nr-db (July 2003). Matches were grouped and localized on the genomic coordinates of the HSA21qsequence. Exon/intron junctions were deduced from the matching cDNAs aligned to HSA21q using the EST2GENOME program (Mott, *Comput. Appl. Biosci*., 1997). We used six different categories for the gene entries in the HSA21qcatalog: known genes (cDNA present in LocusLink at NCBI), except for the proximal KAP gene cluster which was annotated by using the data described (Rogers *et al*. *J. Biol. Chem.* **277**: 48993-49002, 2003), novel genes (sequences showing >96% identity with an annotated ORF, but not present in LocusLink), novel transcripts (spliced ESTs with an ambiguous ORF but with a CDS >50 aa), putative genes (spliced EST without a ORF), pseudogenes (disrupted ORF) and gene predictions exons predicted by GeneScan (Burge and Karlin. *J. Mol. Biol*. **268**:78-94, 1997) overlapping with a cDNA match > 96% identity). We annotated as pseudogenes sequence matches showing an interrupted ORF, either corresponding to a complete human or non-human cDNA or to gene fragments (>70% of the length of the ORF, minimal matching window 31 bp; blast parameter "-w 31"). Later, we added 13

pseudogenes identified by Harrison et al. 2002 (*Genome Res.* **12**:272-280) and obtained using more relaxed parameters (>50% of the length of the ORF). For gene catalogue definitions, we followed the HAWK standard nomenclature (initiated at the Human Annotation Workshops at the Sanger Institute). First methionines were annotated using the NCBI-description of the corresponding mRNAs. During this process, we added 55 genes (known and putative), eliminated 13 orphan gene predictions, fused 5 genes with neighbouring predictions or ESTs changed, two predicted genes to pseudogenes, and changed two pseudogenes to genes. In summary, the updated HSA21qcatalog contains 284 genes and 98 pseudogenes.

For annotating the genes in PTR22q, we blasted all HSA21q transcripts and pseudogenes against the PTR22q genomic sequence. PTR22q genomic segments corresponding to HSA21q genes were extracted and the gene structures were deduced by aligning the human cDNAs with the chimpanzee genomic sequence, using the EST2GENOME program (Mott, *Comput. Appl. Biosci*., 1997). In order to detect potential transcripts specific to PTR22q, the DNA regions of PTR22q that did not match any HSA21q gene were blasted against nr-db at NCBI (all species). We retained as genes only sequence matches corresponding to cDNAs showing > 80% identity with PTR22q; we detected three intronless ORFs inserted in PTR22q and absent from HSA21q (HNRPA1LK1, RPL1LK1, and FAM28ALK1; displayed in Supplementary Table 3) that we annotated as chimpanzee genes.

Multiple alignments of PTR22q genes with orthologs in other species

We calculated strict orthologs of the HSA21q genes with their counterparts (if any) in five species: mouse, rat, zebrafish, pufferfish and ciona. For the mouse, we used the gene orthologs that we had previously identified (Gitton *et al. Nature* **420**: 586-590,

2002). We computed a blast (blastp) of the HSA21q polypeptides against the gene catalogs of rat, zebrafish, pufferfish, and ciona polypeptides. The rat (*Rattus norvegicus*, 22159 genes), zebrafish (*Danio rerio*, 20062 genes), and pufferfish (*Fugu rubripes*, 34615 genes) polypeptides were downloaded from the ENSEMBL ftp server (v.18). The ciona polypeptide catalog (*Ciona intestinalis*, 15852 polypeptides) was downloaded from the JGI ftp server (v.1.0). Strict orthologs to HSA21q genes were identified by the reciprocal best-hit comparison without setting thresholds for either minimal size or percent identity (matches ranged from 28% to 83% for Ciona, for instance). We computed CLUSTALW alignments of the HSA21qand PTR22qpolypeptides and their orthologs in mouse, rat, zebrafish, fugu and ciona. The corresponding multiple alignments are shown in Supplementary table 3.

Comparative gene expression analysis between human and chimpanzee

Gene expression profiles in brain and liver were taken using three adult male humans and three adult male chimpanzees, and Affymetrix complete HG U95 array set and HG U95Av2 array set containing probes to 189 and 99 genes from human chromosome 21, respectively. Out of 18 genes that showed significant changes between humans and chimpanzees (see supplementary Table 4), IFNAR2, TTC3 and ZNF294 showed significant differences in brain and liver. Overall, within brain and liver the proportion of genes which showed different levels of expression is not significantly different on chromosome 21 than in the rest of the genome (liver: χ^2=1.25, p=0.263; brain: χ^2=3.15, p=0.076) (Enard *et al*., *Science* **296**, 340-343, 2002).

Brain samples. Postmortem tissue samples were taken from the frozen brains of three adult male humans (two 45 years old and one 70 years old), three adult male chimpanzees (two 12 years old and one approximately 40 years old), and one adult orangutan female (59 years old). No individuals had history of brain related disease and all suffered sudden deaths without

associated brain damage. Approximately 200 mg of gray matter was dissected from the anterior cingulated cortex of each individual (Brodmann area 24) without thawing the tissues.

Liver samples. The human samples consisted of three liver resections and the chimpanzee and orangutan samples were taken from postmortem liver dissections. The three humans were 27, 29 and 49 years old; the chimpanzees were 12, 34 and 36 years old; and the orangutan was 21 years old. None of the samples showed tissue abnormalities. For each individual, two independent tissue samples, approximately 400 mg each, were dissected and used independently for RNA isolation and further analysis.

Microarray data collection. Total RNA was isolated from all samples using TRIZol® reagent according to the manufacturer instructions and purified with the QIAGEN RNeasy kit following the manufacturer RNA clean-up protocol. All RNAs were of similarly high quality as gauged by the ratio of 28S to 18S ribosomal RNA bands on agarose gels. Also the signal ratio of the probes for the 3' and 5' mRNA ends of GAPDH and β-actin genes, which is the internal mRNA quality control of Affymetrix® microarrays, confirmed the high RNA quality. Labeling of 5 mg of RNA, hybridization, staining, washing and array scanning were carried out following the Affymetrix® protocol.

Expression data were collected using the Affymetrix® HG U95Av2 arrays for the liver samples and the Affymetrix® HG U95Av2, B, C, D, and E array set for the brain samples. The data were analyzed with the Affymetrix Microarray Suite v5.0 using default parameters. For all arrays only the oligonucleotides that matched perfectly between the human and chimpanzee sequences were used for analysis, the rest were masked using a custom designed masking file. The mask was generated using all chimpanzee sequences as available in February 2003 using BLAT for finding the most similar chimpanzee sequence to the probe. After masking, the average number of oligonucleotide probes per transcript mapped to human chromosome 21 decreased on average to 10 from the original 16 present on the array. All arrays were scaled to the same average intensity using all non-masked probes on the array.

Gene expression analysis. Gene expression levels were compared separately for brain and liver in all nine possible pairwise comparisons between three individuals of each species.

From the output-parameters of this software we chose selection criteria for differentially expressed genes to minimize the rate of experimental and biological false positive results. The latter can be caused by the large number of gene expression differences among the individuals of the same species. To assess the number of such false positives among the selected genes and thus to choose the best selection criteria, we looked at all nine different pairings of the human and chimpanzee samples.

A gene was classified as differently expressed between humans and chimpanzees if: *i*. the gene was reliably detected in at least one of the species in all three individuals (detection p-value <=0.05); *ii*. the base-two logarithm of the expression ratio between the species was greater than or equal to 0.2 or less or equal to –0.2 in all nine comparisons. Note that although the minimal signal log ratio difference of 0.2 corresponds only to a 1.15-fold change, the average-fold change was much higher than the minimal threshold since this cutoff was used for all nine comparisons; *iii*. the gene showed a significant and consistent change in expression in all nine pairwise comparisons (change p-value<=0.5 or >=0.5). The change p-value was the only parameter which was varied to optimize the false positive rate of expression differences, 6.4% in brain and 1.0% in liver, found between humans and chimpanzees. Please note that for liver, we calculated the signal log ratios, the detection and the change p-values for two experimental duplicates from independent RNA isolations of the same tissue. The data presented here are based on averages of those values.

Associating Affymetrix probesets with genes: First we used BLAT (Kent WJ, *Genome Res* **12**, 656-664, 2002) to find all Affymetrix target sequences (probesets) mapping to human chromosome 21. Those probesets were matched to human transcripts using BLAST (Altschul *et al*., *J Mol Biol* **215**, 403-410, 1990). Divergence was estimated as the number of basepairs changed over the number of basepairs compared.

Gene-by-gene statistics. For the statistical analysis, we considered only those genes for which at least one of the associated probesets was detected in all three chimpanzees or in all three humans (see i. Gene expression analysis).

Variant genes that may be involved in biomedical processes

Some of the genes that showed differences in structure or expression profile or had a high Ka/Ks ratio may be directly or indirectly involved in biomedical processes which differ between human and chimpanzee. Those include *IFNAR2*, involved in antiviral activities for hepatitis B and C (Han C.S. *et al. Proc Natl Acad Sci USA,* **98**, 6138-6143, 2001), *ETS2,* a broad-range transcription factor which affects the transcription of the *APP* gene associated with the pathogenesis of Alzheimer's disease (Wolvetang E.W. *et al. Biochim Biophys Acta.* **1628**, 105-110, 2003), *CXADR*, a trans-membrane receptor protein for viruses (Peters A.H. *et al. Mol Ther.* **4**, 603-613, 2001), *ITSN1*, a member of a conserved family of proteins involved in clathrin-mediated endocytosis by which internalization of virus particles such as influenza virus into cells is mediated (Guipponi M. *et al. Genomics*, **53**, 369-376, 1998), *DSCR1 (Adapt78)*, an inhibitor of the serine/threonine phosphatase calcineurin which dephosphorylates the tau protein associated with the assembly of paired helical filaments in Alzheimer's disease (Ermak G. *et al. J Biol. Chem.*, **276,** 38787-38794, 2001), *IFNGR2* involved in defense immunity against mycobacterial infections (Dupuis S. *et al. Immunol. Rev*. **178**, 129-137, 2000), *CRYZl1*, a quinone reductase similar to chloroquine-binding proteins identified in malaria-infected erythrocytes (Petri W.A. Jr. *Trends. Pharmacol. Sci*. **24**, 210-212, 2003), *C21orf127,* a member of a DNA (5-cytosine)-methyltransferases (DNMT) family involved in tumor development (Robert M.F. *et al. Nature Genet.* **33**, 61-65, 2003), *TMPRSS2*, a membrane-bound serine protease expressed in a subset of cancers (Lin B. *et al. Cancer Res*. **59**, 4180-4184, 1999), *ABCC13,* a member of the ABC transporter super-family affecting various physiological functions (Yabuuchi H. *et al. Biochem. Biophys. Res. Commun.* **299**, 410-417, 2002), and TTC3, a protein mapped to the *DSCR*, of which the mouse ortholog is expressed predominantly in the central nervous system (Tsukahara F. *et al. J. Biochem. (Tokyo)*, **123,** 1055-63, 1998). In addition to the previous analysis of the expression profiles in mouse (Reymond A. *et al. Nature* **420**, 582-586, 2002; Gitton Y. *et al. Nature* **420**, 586-590, 2002), the present data will be useful for exploring genes that may be correlated with biomedical differences exhibited between human and chimpanzee (Olson M.V. & Varki A. *Nature Rev. Genetics* **4,** 20-28, 2003).

Detection of lineage-specific insertions and deletions

To assign insertion or deletion events on the human and chimpanzee lineages after speciation, we selected all insertions larger than 300bp from both the human and chimpanzee sequences. This size cutoff was chosen only for practical reasons, since the size of the insertions varies widely. In total, 247 sites for human and 181 sites for chimpanzee were selected out of 569 insertion sites (265 for human and 304 for chimpanzee) and PCR primers were designed at both sides of each insertion site for further testing. We could not design proper primers for the remaining 141 sites because of repetitive sequences. Each PCR reaction mixture contained 100ng genomic DNA and a pair of primers (0.3 M each). Other conditions were set according to the recommendation of the supplier (Expand High Fidelity PCR system, Roche Diagnostics). As template, we used five genomic DNA samples each from chimpanzee and human, one from gorilla, and two from orangutan. After PCR amplification, each set of reaction products were separated through 1% agarose gel-electrophoresis for comparison of the product sizes from human and chimpanzee with those of gorilla and orangutan to determine which line contains the ancestral type.

Chimpanzee intra- and inter-individual differences

The chimpanzee BAC sequence overlap differences are a source of SNP (single nucleotide polymorphism) information. We used BAC clone libraries made from three chimpanzees (Gon for PTB1, PTF22, and LA-PCR; Clint for CH251; and Donald for RPCI-43), so intra- and inter-individual differences were obtained:

Clone overlaps	Overlaps with diffs (only for different haplotypes)						All overlaps (incl. same haplotypes)			
	#	Avg length	# of subst.	% subst	# indels	% indels	#	Avg length	% subst	% indels
Clint-Donald	9	83,971	206	0.25	38	0.05	10	85,249	0.24	0.04
Clint-Gon	14	242,199	342	0.14	95	0.04	14	242,199	0.14	0.04
Donald-Donald	22	550,889	1331	0.24	292	0.05	43	1,183,737	0.11	0.02
Donald-Gon	108	2,249,229	3368	0.15	770	0.03	108	2,249,229	0.15	0.03
Gon-Gon	42	1,438,141	932	0.06	273	0.02	95	2,164,036	0.04	0.01
Totals	195	4,564,428	6179	0.14	1468	0.03	270	5,924,449	0.10	0.02

There are no Clint-Clint overlaps. Only one overlap (1278 bp) with no differences was found between two different individuals, all the other (74) overlaps with no differences came from a single individual (either Donald or Gon).

Supplementary Table 1. Center contribution.

Clones Finished per Center					
Center	Minimum tiling path clones	Additional clones	All clones	Total bases	Accession Numbers
CHGC	39	1	40	6,016,694	BS000165-204
KRIBB	21	0	21	2,810,557	BS000205-225
Germany	47	12	59	4,979,927	AL954200-256, 258-259
NIG	5	2	7	1,050,886	BS000158-164
RIKEN	143	14	157	19,540,921	BS000001-157
YMGC	18	2	20	3,412,759	BS000226-245
Total	273	31	304	42,666,429	AL954200-256, 58-59; BS000001-245

Germany: MPI, IMB and GBF

Center-specific sequencing procedures

CHGC

BAC DNA preparation: QIAGEN Large-Construct Kit Shotgun library: sonication and subcloning into pUC18/SmaI Plasmid DNA preparation: Millipore Multiscreen filter plate Sequencing: Big-Dye terminator on ABI 3700 sequencer, DYEnamic ET terminator on Megabase 1000 and 4000 Assembly and finishing: Phrap/Consed

GBF

DNA fragmentation: HydroShear shotgun vector: pTZ 18R prep: TempliPhi chemistry: dideoxy terminator from Amersham machines: MegaBace 1000 or 4000 (semi automated with Watrex plate feeder, "Caddy"), few finishing runs on Licor slab gel assembly and finishing with pregap/Gap4

IMB

Qiagen BAC prep sonication and subcloning into pUC18/SmaI Qiagen template prep ABI BigDye cycle sequencing and separation on ABI3700 Phred/Phrap assembly manual proofreading & editing using Gap4

KRIBB

BAC prep: Ultra-centrifugation by using CsCl2 gradient DNA fragmentation: Hydroshear Shotgun vector: pUC118 Plasmid prep: Millipore 96-well prep kit Chemistry: ET (Amersham), BD-Terminator (ABI) Machine: RISA384 or ABI3730 Data management for shotgun data: Phred/Phrap assembling Data management for finishing data: Sequencher V4.14

MPI

BAC DNA was isolated by alkaline lysis and purified on CsCl by standard procedures. For subcloning DNA was sonicated, fragment ends polished with T4 and Klenow polymerase, size selected for 1.5 and 3.5 kb, ligated in pUC19/SmaI, transformed into E. coli DH10B. Plasmid prep with Millipore kit. Inserts of the libraries were amplified by PCR as templates for sequencing. Sequencing was performed using Big Dye chemistry, M13 primers and ABI 3700 capillary sequencers resulting in more than 10-fold coverage. All raw sequences were processed by PHRED, controlled for vector or E. coli contamination and assembled by Phrap. Analyzed regions were manually edited in GAP4.

NIG

Qiagen BAC prep.,DNA fragmentation by HydroShear, Subcloned by pUC118, PCR by M13F&M13R primers, Sequencing by BigDye terminator and ABI3700, PCR and TA cloning for finishing, phredPhrap assembling, manual editing by consed.

RIKEN

BAC DNA was purified through EtBr/CsCl equilibrium centrifugation, then fragmented by sharing (Hattori *et al*. *Nature* **405**:311-319, 2000).

YMGC

BAC DNA was isolated by alkaline lysis and followed by phenol/Chloroform extraction. The DNA was then sheared by Hydroshear and ends polished with T4 DNA polymerase and Bal 31nuclease. Fragment sizes ranging from 2.5 to 3.5 kb were selected and ligated to pUC18/SmaI followed by transformation to E. coli DH5a. Plasmid DNA was isolated by alkaline lysis and purified by MultiScreen-HV filter (Millpore). Plasmid DNA was sequenced by Big Dye version 1 and analyzed by ABI 3700. For each BAC clone, at least 10-fold coverage was sequenced and the vector and E. coli contamination were below 15%. The sequencing data were processed and assembled by Phred/Pharp/Consed.

Supplementary Table 2. Summary of the chimpanzee22 sequence quality.

(See SupplementaryTable2)

Supplementary Table 3. Comparative gene catalogue.

(See another file)

Supplementary Table 4. Genes with internal amino acid insertions or deletions in chimpanzee.

Human Gene references	Genomic changes in PTR22q	AA indels in chimp	Mouse ortholog
IFNAR1 Interferon-alpha receptor*:* **A26593, A26595, A32389, A32391, BC021825, J03171**	ins (CCT)	p.P148_G149insP	NP_034638 no insertion
IFNAR2 Interferon alpha/beta receptor*:* **L41942, X89772,** X77722, L41944	ins (GAA)	p.L443_E444insE **L41942, X89772**	NP_034639 divergent structure in this region
C21orf2 Nuclear encoded mitochondrial protein, cDNA A2-YF5: **BC031300**, Y11392, U84569, Z93322	ins (GAG)	p.G323_L324insL **BC031300**	BAB23134 divergent structure in this region
MCM3AP MCM3 import factor*:* **AJ010089,** AB005543, BC013285, BC032750	ins (GAT)	p.V1000_S1001insI **AJ010089**	NP_062307 Insertion I
ANKRD3 ankyrin like, dual-specificity Ser/Thr/Tyr kinase domain*:* **AB047783, AJ278016, AK027424**	ins (AGACAC)	p.P315_A316insVS	AAG30871 no insertion
USP16 Ubiquitin processing protease, EC3.1.2.15: **AF113219, AK023247, AF126736, AK025104**	ins (ACTGACTGT)	p.P374_T375insTDC: **AF113219, AK023247** p.P553_T554insTDC: **AF126736, AK025104**	NP_077220 Insertion. EC
ETS2 Erythroblastosis virus oncogene homolog 2*:* **AK096841**, BC017040, J04102	ins (CCCTCGCC CTCG)	S116_P117insPSPS **AK096841**	AAA37581 divergent structure in this region
KRTAP10-10 Hair keratin-associated protein 10.10*:* **AJ566387**	ins (TGCTGCGCC CCCAGC)	P34_A35insSCCAP	no clear ortholog
C21orf45 Unknown function*:* **AF231921, AF387845**	del (CTC)	p.E72del	AAF72945 No deletion
TRPM2 Transient receptor potential-related channel 7, a novel putative Ca2+ channel protein*:* **AB001535**	del (GAG)	p.E15del	NP_036165 divergent structure in this region
SLC19A1 Reduced folate carrier **U19720, AF004354, U15939, U17566**	del (TGG)	p.P234del	AAC53287 No deletion
C21orf22 Unknown function **AY040089**	del (TGCAGC)	p.A44_A45del	No ortholog
COL18A1 Human type XVIII collagen **AF018081, AF018082**	ins (GGCCCCCCC) del (GGCCCCCCA)	p.P1176_S1177ins (GPP) p.G1125_P1127del (GPP)	NP_034059 no insertion del (GPPGPR)
TCP10L * T-complex protein 10A-2 **AF115967, AK058078, BC022024**	del (ACAAAGATCGTCATCTA) corresponds to a duplication of 17 bp in HSA21	p.F156fsS168	No ortholog
PCNT2 *; Pericentrin, kendrin **U52962**	del (195 bp)	p.128Q_E192del (65aa)	AAA17886 Missing exon

Accession numbers in bold refer to the transcript isoforms predicted to be modified in chimpanzee. Coordinates of the genomic changes in the PTR22q sequence, and nucleotide/amino acid sequence alignments are given in supplementary Table 3. Genes with an asterisk are described in the text.

Supplementary Table 5. Genes with altered start or stop codons in chimpanzee.

Human Gene references	Genomic Changes in PTR22q	Modifed START	Modified STOP	Mouse ortholog
C21orf9 Unknown function: **AY077697**	g.29579632G>C	p.M1_I15del		No ortholog
C21orf122 Unknown function: **NM_032653**	g.44721800T>G	p.M1?		No ortholog
C21orf86 Unknown function: **AF426264**	g.44946998T>C	p.M1_V34del		No ortholog
KRTAP23-1 human hair keratin-associated protein 23.1: (KAP ref)	g.30149103C>T g.30149105T>G	p.M1?		No clear ortholog
C21orf97 Unknown function: **AK024977 BC003651**	g.43399866-43399867del (C)	p.M1_D14del		No ortholog
DSCR6 Unknown function: **AB037158 AB037159**	g.36751787C>T		Premature STOP: p.R162X **AB037158** p.R78X **AB037159**	BAB60891 divergent structure in this region
PSMD15 Proteasome 26S subunit gene: **AF050199**	g.36229865C>T		Premature STOP: p.R183X	NP_032977 divergent structure in this region
C21orf118 Unknown function: **AF304442**	g.26044865C>A		Premature STOP: p.S36X	No ortholog
C21orf128 Unknown function: **NM_152507**	g.41815522A>T		Premature STOP: p.Y33X	No ortholog
KRTAP15-1 human hair keratin-associated protein 15.1: (KAP ref)	g.30216795C>T		Premature STOP : p.Q72X	No clear ortholog
IGSF5 putative gene, immunoglobulin superfamily 5 like: **AK092516**	g.39483930C>T		Premature STOP: p.Q160X	NP_082354 The mouse and the human proteins have a stop at the same position but have a different methionine.
ABCC13 putative gene, multidrug resistance associated protein like: **AF418600 AF518320 AY063514 AY063515**	g.14356247ins (G) **AF418600 AY063514 AY063515** g.14385690T>C g.14385691G> **AF518320**		Frameshift + premature STOP: p.R16fsX20 **AF418600, AY063515, AY063514** p.Q180X **AF518320**	AAB80938 divergent structure in this region
C21orf124 Unknown function: **AK056502 BC008008 BC021550**	g.43375146_43375147del (A)		Frameshift + premature STOP: p.S72fsX80	No ortholog
C21orf90 Unknown function: **NM_153204**	g.44162879_44162880del (C)		Frameshift + premature STOP: p.G23fsX26	No ortholog
KRTAP19-4 human hair keratin-associated protein 19.1: (KAP ref)	g.30265157_30265158ins (GA)		Frameshift + premature STOP: p.Q61fsX64	No clear ortholog
ICOSL Unknown function: **AB014553**	g.43879282_43879283del (CGCGGAGACCTCGGGG)		Frameshift + premature STOP: p.P35fsX66	AAF34738 divergent structure in this region

C21orf121 Unknown function: **NM_032653**	g.41735691_41735692del (AGTCCCACCGTCGTC TTCTAGCCCCACCAT CGTCGTCT)		Frameshift + premature STOP: p.I30fsX102	No ortholog
C21orf74 Unknown function: **AY077696**	g.22190917-22190918del (1287 bp)		Frameshift + premature STOP: p.T29fsX49	No ortholog
C21orf114 Unknown function	g.17810821ins (A)		Frameshift + premature STOP: p.I17fsX37	No ortholog
C21orf79 Unknown function	g.26054348ins (T)		Frameshift + premature STOP:p.V29fsX39	No ortholog
C21orf71 Unknown function: **AF086441**	g.25411102T>G		STOP deletion : p.X84CinsA84+1_X+21	No ortholog
C21orf30 Unknown function: **AL117578**	g.44105796T>C g.44105797G>C		STOP deletion: p.X248PinsC248+1_X+20	No ortholog
LIPI putative gene, lipase (EC 3.1.1.3) like: **BC028732**	g.14247293_14247294del (A)		Frameshift and use of another STOP: p.L271fsX294	BB663289 divergent structure in this region
C21orf70 Unknown function: **AF391113** AF391114 **BC009341**	g.44590485_44590486del (corresponding to a duplication of 40nt in the human)		Deletion p.E81_G95del (EAGSSRSVPSIRRG) **AF391113 BC009341**	AAL34506 The duplication is also absent in the mouse
LSS human lanosterol synthase, EC 5.4.99.7.: **AK092334** D63807 S81221 U22526	g.45868808_45868809ins (TA)		Frameshift + premature STOP: p.R20fsX35 **AK092334**	XP_109587 divergent structure in this region
PDE9A CGMP-specific 3', 5'-cyclic phosphodiesterase type 9, EC 3.1.4.17.: AF048837 AF067223 AF067224 AF067225 **AF067226** BC009047	g.42390873C>T		Premature STOP: p.R38X **AF067226**	AAC24344 divergent structure in this region
ABCG1 white protein homolog (ATP-binding cassette transporter 8): **AY048757** X91249	g.41964344_41964345del (A)		Frameshift + premature STOP: p.V216fsX247 **AY048757**	AAB47738 divergent structure in this region
FAM3B Unknown function: AF375989 **AJ409094**	g.41022241_41022242del (TTTGTTGGTT)		Frameshift + premature STOP: p.C29fsX31 **AJ409094**	BAB31283 divergent structure in this region
DSCR5 human Down syndrome critical region protein C: AB035742 AB035743 AB035744 **AB035745** AB037162 **AB037163** AB037164 AF216305 BC011007	g.36807416ins (C)		Frameshift + premature STOP: p.T9fsX20 **AB035745 AB037163**	NP_062416 divergent structure in this region
AIRE autoimmune regulator (APECED protein): AB006682 **AB006683 AB006685** Z97990	g.43928300A>G	p.M1_V25del **AB006683 AB006685**		CAB66141 divergent structure in this region
D21S2056E human NNP-1/Nop52 (NNP-1), novel nuclear protein 1: **AY033999** BC014787 U79775	g.43432783G>A	p.M1_I17del **AY033999**		NP_035055 divergent structure in this region
C21orf66 Unknown function: AF231920 AY033903 AY033905 AY033906 BC030539 HSA279080	g.32483288ins (C) g.32483291ins (A)		STOP deletion: p.X816Lins?	BAB27645 divergent structure

HSA279081 **AF153208 AY033904**			**AY033904** p.X248LinsE248+1_X+2 **AF153208**	in this region

Supplementary Table 6. Genes with extreme Ka/Ks values.

Group	Group character	Minimum Ka/Ks value	Gene name (Hs)	LocusLink ID (Hs)	Description
Ka/Ks>1		3.37	KRTAP23-1	337963	keratin associated protein 23-1
		2.78	C21orf87	257357	chromosome 21 open reading frame 87
		1.98	C21orf81	114035	chromosome 21 open reading frame 81
		1.79	C21orf128	150147	hypothetical protein
		1.76	C21orf119	84996	chromosome 21 open reading frame 119
		1.73	RPS5L	54022	ribosomal protein S5-like
		1.71	PRED62		
		1.67	KRTAP15-1	254950	keratin associated protein 15-1
		1.57	KRTAP21-1	337977	keratin associated protein 21-1
		1.47	ABCC13	150000	ATP-binding cassette, sub-family C (CFTR/MRP), member 13
		1.45	C21orf94	246705	chromosome 21 open reading frame 94
		1.42	C21orf93	246704	chromosome 21 open reading frame 93
		1.40	C21orf86	257103	chromosome 21 open reading frame 86
		1.37	ANKRD21	317754	Expressed in prostate, ovary, testis, and placenta
		1.34	C21orf111	378823	chromosome 21 open reading frame 111
		1.16	C21orf129	150135	hypothetical protein
		1.15	TMPRSS2	7113	transmembrane protease, serine 2
		1.10	C21orf126	84210	hypothetical protein
		1.08	C21orf22	54089	chromosome 21 open reading frame 22

		1.04	C21orf115	378827	chromosome 21 open reading frame 115
		1.01	DSCR6	53820	Down syndrome critical region gene 6
		1.01	PRED61		
(Nd + Sd) >= 10 and (Ka + Ks) / 2 >2% (> 1.44%)	relatively rapidly evolving genes	0.56	KRTAP13-3	337960	keratin associated protein 13-3
		0.56	KRTAP6-3	337968	keratin associated protein 6-3
		0.27	KRTAP19-7	337979	keratin associated protein 19-7
		0.32	KCNE1	3753	potassium voltage-gated channel, Isk-related family, member 1
		0.19	ATP5J	498	ATP synthase, H+ transporting, mitochondrial F1 complex, alpha subunit, isoform 1, cardiac muscle
		0.35	UMODL1	89766	uromodulin-like 1
		0.72	TCP10L	140290	t-complex 10 (mouse)-like
		0.34	B3GALT5	10317	UDP-Gal:betaGlcNAc beta 1,3-galactosyltransferase, polypeptide 5
		0.54	IGSF5	54046	immunoglobulin superfamily, member 5
p-value <0.05, (Ka + Ks) / 2 > 2%	evolving under purifying selection	0.55	KRTAP13-4	284827	keratin associated protein 13-4
		0.00	PFKL	5211	phosphofructokinase, liver
		0.08	C21orf113	378825	chromosome 21 open reading frame 113
		0.27	COL18A1	80781	collagen, type XVIII, alpha 1
		0.19	AIRE	326	autoimmune regulator (autoimmune polyendocrinopathy candidiasis ectodermal dystrophy)
		0.05	ITGB2	3689	integrin, beta 2 (antigen CD18 (p95), lymphocyte function-associated antigen 1; macrophage antigen 1 (mac-1) beta subunit)
		0.06	TMPRSS3	64699	transmembrane protease, serine 3

		0.05	AGPAT3	56894	1-acylglycerol-3-phosphate O-acyltransferase 3

Nd: # of nonsynonymous substitutions; Sd: # of synonymous substitutions

Supplementary Table 7. Genes on human chromosome 21 that show significant expression differences between humans and chimpanzees.

Gene ID	Transcript ID	brain	liver	lineage-specific changes	
				brain	liver
C21orf18	NM-017438			chimp	
C21orf33	Y07572				nonspecific
C21orf5	AJ237839				chimp
C21orf97	BC003651, AK024977			nonspecific	
CRYAA	U05569, U66584				chimp & human
CRYZL1	AK001293			?	
CXADR	Y07593				human
DSCR1	U85267				human
ETS2	J04102			chimp & human	
IFNAR2	X77722			chimp	chimp & human
IFNGR2	BC003624			chimp & human	
ITSN1	U61166				nonspecific
LSS	AK092334				nonspecific
PDXK	BC000123				chimp
PTTG1IP	Z50022				chimp & human
TTC3	D84296			nonspecific	nonspecific
USP16	AK025104			nonspecific	
ZNF294	AB018257			chimp	chimp

	higher in chimp
	higher in human
	not expressed
	no difference, but expressed
	not on chip

The assignment of lineage-specific differences was done using the expression values from one orangutan. If the expression measured for the orangutan differed significantly (t-test p<=0.05) from chimpanzee, but not from human then we assume that the change occurred on the chimpanzee lineage and if the orangutan differed significantly from human but not from chimpanzee then the change occured on the human lineage.

Supplementary Table 8. Ranking of all genes that were expressed in either brain or liver.

	# of genes with expression		mean divergence		
	different	not different	different	not different	p-Value*
non-degenerate sites	18	47	0.0060	0.0043	0.2150
3'UTR	18	47	0.0100	0.0090	0.1240
5'UTR	18	46	0.0115	0.0085	0.023*
CpG-island	12	40	0.0161	0.0135	0.0590
intron	17	46	0.0129	0.0125	0.1290
3' intergenic (10kb)	18	48	0.0129	0.0129	0.9310
5' intergenic (10kb)	18	48	0.0131	0.0127	0.6970
5' intergenic (1kb)	18	48	0.0140	0.0132	0.4080

* Mann-Whitney U -test, two-tailed significance

Calculation is according to the nucleotide divergence at non-degenerate sites, 3'UTR, 5'UTR, CpG-island, introns, 3' and 5' intergenic regions and tested whether the genes with differing mRNA levels between humans and chimpanzees are non-randomly distributed within that list.

Supplementary Table 9. Nucleotide substitution patterns in human and chimpanzee.

(A) Proportion of nucleotide substitution in the human genome based on SNP data

		NEW			
		A	T	C	G
OLD	A	-	2.88 (703)	3.60 (880)	14.00 (3422)
	T	2.80 (681)	-	15.05 (3654)	3.45 (838)
	C	4.43 (747)	20.31 (3424)	-	4.51 (760)
	G	19.56 (3292)	4.53 (762)	4.88 (822)	-

Note. An element at the i-th row and j-th column means nucleotide substitution from nucleotide i to j. Figures designate standardized proportion of substitution in %, while figures in parentheses are numbers of mutations observed out of 19,985 SNP loci.

(B) Proportion of nucleotide substitutions in the chimpanzee genome based on BAC-overlap SNP data

		NEW			
		A	T	C	G
OLD	A	-	3.55 (252)	3.68 (261)	12.58 (893)
	T	2.92 (206)	-	13.18 (930)	4.02 (284)
	C	5.08 (250)	19.22 (946)	-	5.24 (258)
	G	20.83 (1024)	4.37 (215)	5.33 (262)	-

Note. An element at the i-th row and j-th column means nucleotide substitution from nucleotide i to j. Figures designate the standardized proportion of substitution in %, while figures in parentheses are numbers of mutations observed out of 5,781 SNP loci.

(C) Nucleotide transition probability matrices for human and chimpanzee.

	Human				Chimpanzee			
	A	T	C	G	A	T	C	G
A	0.950	0.007	0.008	0.034	0.985	0.002	0.003	0.009
T	0.007	0.948	0.036	0.008	0.002	0.985	0.010	0.003
C	0.012	0.050	0.929	0.011	0.004	0.014	0.978	0.004
G	0.049	0.011	0.011	0.930	0.015	0.004	0.004	0.978

(D) Equilibrium frequencies of four nucleotides based on the nucleotide transition probability matrices for human and chimpanzee shown above. Tajima and Nei (1982)'s method was used. Figures in parentheses are observed values.

	Human	Chimpanzee
A	0.290 (0.297)	0.303 (0.296)
T	0.288 (0.295)	0.293 (0.294)
C	0.214 (0.205)	0.205 (0.205)
G	0.208 (0.204)	0.200 (0.205)
GC content	0.422 (0.409)	0.405 (0.410)

<References sited in this table>

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 Table 10. List of 18 DNA regions that showed statistically significant results based on a Modified H test suggesting positive selection (bin size = 10kb)

Position	*P*(*H*)	*S*	Gene name; description (in any)
16299112	0.0430	2(0)	CHODL; A type I membrane protein with a carbohydrate recognition domain characteristic of C-type lectins in its extracellular portion
17880261	0.0320	2 (0)	Noncoding region
18735154	0.0320	1 (0)	Noncoding region
20045321	0.0270	2 (0)	Noncoding region
20724640	0.0070	8 (0)	Noncoding region
20794640	0.0340	5 (0)	Noncoding region
21949854	0.0450	31 (0)	Noncoding region
25259390	0.0480	7 (0)	Noncoding region
27461495	0.0100	4 (0)	Noncoding region
28230953	0.0450	3 (0)	Noncoding region
30396956	0.0370	4 (0)	C21orf108; Chromosome 21 open reading frame 108
32522462	0.0430	4 (0)	KCNE; Potassium voltage-gated channel, Isk-related family, member 1
37213005	0.0210	3 (0)	DSCR2; Down syndrome critical region gene 2
37694053	0.0220	4 (2)	B3GALT5; UDP-Gal:betaGlcNAc beta 1,3-galactosyltransferase, polypeptide 5; Exon 1: Synonymous difference (ACG/ACA); Exon 3: Difference in 3' UTR
40131716	0.0250	3 (0)	Noncoding region
40243527	0.0340	7 (0)	Noncoding region
41688508	0.0480	4 (1)	KIAA0179; KIAA0179 protein; Exon 7: Synonymous difference (GCG/GCA)
42533280	0.0380	2 (0)	C21orf90, Chromosome 21 open reading frame 90; C21orf29, Chromosome 21 open reading frame 29

Note. Position: the mid-point of a bin (10 kb for the alignment) along HSA21q. *S*: the number of segregating sites. The number of segregating sites in exons are in parentheses, and the SNP information is given when it resides in an exon. *P*(*H*): the probability of observing a more negative *H* value under neutrality (significance level = 5%, one sided). If an SNP locus was detected with both maximum and minimum sample sizes, the smaller *P*(*H*) was shown.

28