BIOINFORMATICS

Vol. 21 Suppl. 1 2005, pages i186–i194
doi:10.1093/bioinformatics/bti1000

Identification of nine human-specific frameshift mutations by comparative analysis of the human and the chimpanzee genome sequences

Yoonsoo Hahn and Byungkook Lee*

Laboratory of Molecular Biology, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, MD 20892, USA

Received on January 15, 2005; accepted on March 27, 2005

ABSTRACT

Motivation: The recent release of the draft sequence of the chimpanzee genome is an invaluable resource for finding genome-wide genetic differences that might explain phenotypic differences between humans and chimpanzees.

Results: In this paper, we describe a simple procedure to identify potential human-specific frameshift mutations that occurred after the divergence of human and chimpanzee. The procedure involves collecting human coding exons bearing insertions or deletions compared with the chimpanzee genome and identification of homologs from other species, in support of the mutations being human-specific. Using this procedure, we identified nine genes, *BASE*, *DNAJB3*, *FLJ33674*, *HEJ1*, *NTSR2*, *RPL13AP*, *SCGB1D4*, *WBSCR27* and *ZCCHC13*, that show human-specific alterations including truncations of the C-terminus. In some cases, the frameshift mutation results in gene inactivation or decay. In other cases, the altered protein seems to be functional. This study demonstrates that even the unfinished chimpanzee genome sequence can be useful in identifying modification of genes that are specific to the human lineage and, therefore, could potentially be relevant to the study of the acquisition of human-specific traits.

Availability:

Contact: bk@nih.gov

1 INTRODUCTION

Humans have many features that make them distinct from the great apes, for example, bipedalism, facilitated encephalization and use of complex language. These must be the result of genetic changes accumulated in the genome during evolution of the great apes and/or after the divergence of human and chimpanzee lineages (Gagneux and Varki, 2001; Varki, 2004). These include changes in the expression level (Khaitovich *et al*., 2004), duplication (Fortna *et al*., 2004) and amino acid substitutions (Enard *et al*., 2002) of existing genes. Lineage-specific traits can also be achieved by gaining new genetic materials through various mechanisms such as segmental duplication (Bailey *et al*., 2002) and retrotransposition (Burki and Kaessmann, 2004). However, the 'less-is-more' hypothesis asserts that loss-of-function mutations are also important for the establishment of a species (Olson, 1999). The most striking example that supports this hypothesis is the inactivation of the myosin heavy chain 16 (*MYH16*) gene in human lineage caused by 2 bp deletion within the coding sequence (Stedman *et al*., 2004). The frameshift mutation resulted in the loss of the protein and a marked reduction of masticatory muscle mass, which may have allowed humans to have bigger brains. Other examples of human-specific gene inactivations are: complete gene loss of the sialic acid binding Ig-like lectin 13 (*SIGLEC13*) gene (Angata *et al*., 2004), Alu repeat-mediated exon deletion of the cytidine monophospho-*N*-acetylneuraminic acid hydroxylase (*CMAH*) gene (Hayakawa *et al*., 2001), a nonsense mutation of the *KRTHAP1* gene in the type I hair keratin gene cluster (Winter *et al*., 2001) and a 1 bp deletion of the EGF-like module-containing mucin-like receptor 4 (*EMR4*) gene (Hamann *et al*., 2003).

The recent release of the chimpanzee (*Pan troglodytes*) genome by the Chimpanzee Genome Sequencing Consortium provides an invaluable resource for the identification of human-specific genetic changes that occurred after human and chimpanzee divergence (Olson and Varki, 2003). A genome-wide comparison should disclose sequence differences between the two genomes. If an insertion or deletion event occurred in one of the two lineages, it will show up as an alignment gap. When such a gap is located within the coding sequence in a gene, it results in the insertion or deletion of one or more amino acids or in a reading frame change. However, the draft-quality of the chimpanzee genome sequence and the lack of supporting chimpanzee mRNA sequences hinder identification of lineage-specific genomic alterations. Since the current chimpanzee genome assembly (NCBI Build 1 Version 1, November 13, 2003 release) is based on 4× sequencing coverage, it is expected to contain sequencing errors including gaps. In contrast, the near-perfect 'finished' human genome

*To whom correspondence should be addressed.

sequence is considered to be highly accurate, exceeding the 99.99% accuracy standard (International Human Genome Sequencing Consortium, 2004; Schmutz *et al*., 2004). Furthermore, the coding sequences can be crosschecked with several mRNA sequences isolated from various independent sources.

Here, we report the development of a simple and rapid procedure for identifying putative human-specific frameshift mutations and the discovery of nine such mutations. The procedure involves the collection of chimpanzee coding exons containing an insertion or deletion event compared with orthologous human mRNAs and comparison of human/chimpanzee protein pairs with corresponding homologs from other species. If a non-human/non-chimpanzee homolog shows substantial identity with the predicted chimpanzee protein in its entire length, the chimpanzee sequence is regarded as reliable and accurate, and, in turn, the presumed human-specific frameshift mutation is considered true. The potential functional consequences of the nine putative human-specific frameshift mutations are discussed. Owing to the limitations of the chimpanzee genome sequence mentioned above, the reciprocal approach was not performed.

2 METHODS

2.1 Data sources and sequence analysis

The human mRNA-to-human genome alignments, the human mRNA-to-chimpanzee genome alignments, the human genome sequence, and the chimpanzee genome sequence were downloaded from the Genome Browser Database (Karolchik *et al*., 2003) at the University of California, Santa Cruz (ftp://hgdownload.cse.ucsc.edu) in August 2004. The human mRNA-to-human genome alignment data were found in the tables ‘refSeqAli’ and ‘all_mrna’ of the database ‘hg17.’ The human mRNA-to-chimpanzee genome alignment data were found in the tables ‘xenoRefSeqAli’ and ‘xenoMrna’ of the database ‘panTro1’. The non-human vertebrate protein database was prepared from the non-redundant protein database ‘nr’ (ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nr.gz) with the assistance of taxonomy information (ftp://ftp.ncbi.nlm.nih.gov/pub/taxonomy/). Domain searches were performed using Pfam database (Bateman *et al*., 2004) at Washington University in St. Louis website (http://pfam.wustl.edu).

2.2 Collection of coding exons containing insertion or deletion

In order to identify human-specific frameshift mutations, we first collected GenBank accession numbers of the human mRNA sequences that were aligned in both the human and the chimpanzee genomes. The sequences without a complete coding region were excluded. If a sequence was aligned in more than one place in a genome, only the best alignment was kept to ensure that a sequence was mapped to a single locus. A total of 75 856 human mRNA sequences met the condition to build the initial dataset.

In the next step, we selected human coding exons bearing nucleotide insertions or deletions when aligned to the current chimpanzee genome sequence. The coding region information and the genome alignment data were used to define the coding exons of each human mRNA sequence. For each human coding exon, a series of human mRNA-to-chimpanzee genome alignment blocks corresponding to the human coding exon was searched. If more than one alignment block was found for a single human coding exon, the corresponding chimpanzee exon was considered to contain insertions or deletions (see the legend to Figure 1 for an example). A total of 6517 non-redundant coding exons were found to have at least one insertion or deletion. The coding exons that contain more than one insertions or deletions were then removed. This condition was applied because multiple insertion or deletion in a single exon could arise from sequencing errors in the chimpanzee genome. It is possible to miss some rapidly decaying genes by this step, but we assume that the two species diverged sufficiently recently (5–7 million years ago) so that the probability of an exon accumulating multiple mutations during this period is small. There were 4628 coding exons with a single insertion or deletion event.

2.3 Collection of human-specific premature termination candidates

In order to investigate the consequence of an insertion or deletion event on a chimpanzee protein, we generated chimpanzee-like mRNA sequences. These are human sequences modified by applying insertions or deletions according to the human mRNA-to-chimpanzee genome alignment data. The coding region of each chimpanzee-like mRNA was then re-assigned using the same start codon as the corresponding human mRNA. Of the 4628 chimpanzee-like mRNAs, 3428 sequences contained complete coding regions from a start codon to a stop codon.

With the aim of finding human-specific frameshift mutations leading to premature termination, we selected cases where a frameshift made the chimpanzee-like protein longer than the human counterpart. The chimpanzee-like mRNAs producing shorter polypeptides were removed. The amino acid insertion cases where the same open reading frame was retained after an insertion event were also discarded. A total of 289 chimpanzee-like mRNAs were identified to encode longer peptides than did the corresponding human mRNAs as a result of the reading frame change.

2.4 Identification of the non-human/non-chimpanzee homologs

In order to identify non-human/non-chimpanzee homologs of each human/chimpanzee protein pair, BLAST searches of locally prepared non-redundant non-human vertebrate protein database were carried out by using human and chimpanzee-like protein sequences as queries. When the BLAST outputs were parsed, non-human/non-chimpanzee homologs were found for 240 of the 289 human/chimpanzee-like protein pairs. The hit list of each human/chimpanzee-like protein pair was sorted by the BLAST score between the chimpanzee-like protein and the non-human/non-chimpanzee homolog. Assuming that the chimpanzee-like proteins were ‘wild types’ (without a frameshift mutation), 38 cases where the chimpanzee-like protein showed a score increase of at least 10 bits, compared with the human protein, within the top five hits were saved for further analysis.

2.5 Collection of the human-specific frameshift mutations

Each of the 38 potential human-specific frameshift mutations was manually inspected. For each case, all human mRNAs and all homologous non-human mRNAs were obtained from BLAST searches at the NCBI website (http://www.ncbi.nlm.nih.gov/BLAST). The human and the chimpanzee genomic fragments that contained a given gene were derived from BLAT searches at the Genome Browser Database (http://genome.ucsc.edu/cgi-bin/hgBlat). The genuine chimpanzee mRNA sequence for each case was predicted from the chimpanzee genome sequence by assembling exons defined by an alignment of the human mRNA and the chimpanzee genomic fragment using the SIM4 program (Florea *et al*., 1998). The chimpanzee protein sequences were deduced by translation of the predicted chimpanzee mRNA sequences. Multiple sequence alignment analyses of human and chimpanzee protein sequences along with respective homologs were performed using the T-COFFEE program (Notredame *et al*., 2000). To be confirmed as a human-specific frameshift mutation: (1) the human genomic sequence of a gene should agree with all of its mRNA sequences and expressed sequence tags currently available and (2) the genuine chimpanzee protein should show significant identities with a non-human/non-chimpanzee homolog in the C-terminal side from the potential human-specific frameshift mutation site. This resulted in nine genes, each with a human-specific frameshift mutation.

3 RESULTS

3.1 The human-specific 1 bp deletion of *BASE*

A simple procedure for comprehensive identification of human-specific frameshift mutations was devised after a thorough examination of the *BASE* (breast cancer and salivary gland expression) gene. *BASE* was discovered in a search for genes expressed in breast cancer (Egland *et al*., 2003). In normal tissue, its expression was almost exclusively detected in the salivary gland. BASE, a 179 amino acid protein, shares sequence similarity with horse Latherin, a 228 amino acid protein. Sequence comparison raised the possibility that BASE is truncated as a result of a 1 bp deletion in exon 6, creating a premature stop codon in exon 6. Insertion of a nucleotide ‘restored’ the reading frame to produce a 229 amino acid protein and led to extended similarity with Latherin. Bingle *et al*. also pointed out the single nucleotide deletion and suggested that *BASE* represents a ‘dying gene’ (Bingle *et al*., 2004).

The chimpanzee draft genome sequence enabled us to identify the chimpanzee *BASE* gene. Analysis of the alignment of the human *BASE* mRNA and the chimpanzee genome obtained from the Genome Browser Database revealed that exon 6 that spans from nt 555 to 635 of the *BASE* mRNA sequence was split into two alignment blocks, one from 555 to 566 and the other from 567 to 635 (Fig. 1A). This interrupted alignment is because of an extra adenine nucleotide at position 33 233 688 of the chimpanzee chromosome 21 (the position is based on the chimpanzee genome November 13, 2003 release), extending the coding region of chimpanzee *BASE*. The single nucleotide deletion between 566 and 567 in the human *BASE* mRNA results in a reading frame shift and truncation of the BASE protein (Fig. 1B). Peptide sequence alignment of human BASE and predicted chimpanzee BASE along with horse Latherin (Fig. 1C) clearly indicates that chimpanzee *BASE* encodes an intact protein and the 1 bp deletion mutation occurred in the human lineage after divergence of human and chimpanzee.

3.2 Design and application of a procedure for identification of the human-specific frameshift mutations

The human-specific single nucleotide deletion of *BASE* provided a clue for designing a procedure for genome-wide identification of human-specific frameshift mutations. It involves collection of chimpanzee coding exons bearing an insertion or a deletion compared with human mRNAs and identification of non-human/non-chimpanzee homologs, which confirms that the mutations are human-specific. A simple procedure was designed to filter the human mRNA-to-chimpanzee genome alignment data downloaded from the Genome Browser Database. Starting with 75 856 alignment datasets, we identified nine highly plausible human-specific frameshift mutations. These are *BASE*, *DNAJB3*, *FLJ33674*, *HEJ1*, *NTSR2*, *RPL13AP*, *SCGB1D4*, *WBSCR27* and *ZCCHC13*. Table 1 shows a summary of the results. Pairwise comparisons of each human and chimpanzee gene pair around the frameshift mutation are shown in Figures 1B, 2A and 3. Multiple sequence alignment analyses of human/chimpanzee protein pairs with their respective homologs are presented in Figures 1C, 2B and 4.

3.3 *DNAJB3*

Human *DNAJB3* (reported as *HCG3*) encodes a DnaJ (Hsp40) homolog, subfamily B, member 3 protein. It has a single guanine nucleotide insertion, inducing an open reading frame shift, when compared with its chimpanzee ortholog (Fig. 2). Human and chimpanzee *DNAJB3* genes are predicted to

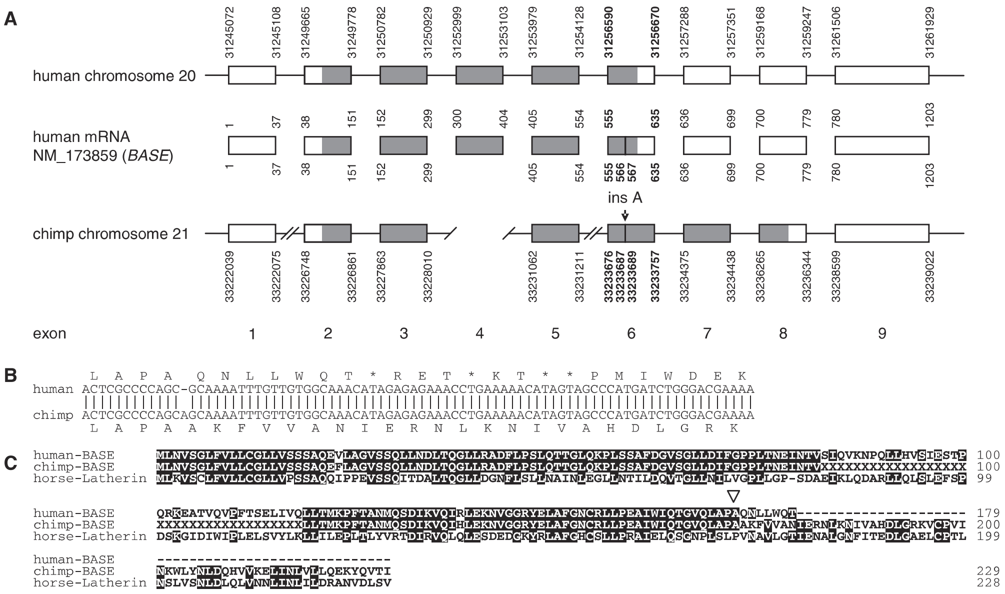
A
31245072
31245108
31249665
31249778
31250782
31250929
31252999
31253103
31253979
31254128
31256590
31256670
31257288
31257351
31259168
31259247
31261506
31261929
human chromosome 20
1
37
38
151
152
299
300
404
405
554
555
635
636
699
700
779
780
1203
human mRNA
NM_173859 (BASE)
1
37
38
151
152
299
405
554
555
566
567
635
636
699
700
779
780
1203
ins A
chimp chromosome 21
33222039
33222075
33226748
33226861
33227863
33228010
33231062
33231211
33233676
33233687
33233689
33233757
33234375
33234438
33236265
33236344
33238599
33239022
exon
1
2
3
4
5
6
7
8
9
B
L A P A Q N L L W Q T * R E T * K T * * P M I W D E K
human ACTCGCCCCAGC-GCAAAATTTGTTGTGGCAAACATAGAGAGAAACCTGAAAAACATAGTAGCCCATGATCTGGGACGAAAA
chimp ACTCGCCCCAGCAGCAAAATTTGTTGTGGCAAACATAGAGAGAAACCTGAAAAACATAGTAGCCCATGATCTGGGACGAAAA
L A P A A K F V V A N I E R N L K N I V A H D L G R K
C
human-BASE MLNVSGLFVLLCGLLVSSSAQEVLAGVSSQLLNDLTQGLLRADFLPSLQTTGLQKPLSSAFDGVSGLLDIFGPPLTNEINTVSIQVKNPQLLHVSIESTP 100
chimp-BASE MLNVSGLFVLLCGLLVSSSAQEFLAGVSSQLLNDLTQGLLRADFLPSLQTTGLQKPLSSAFDGVSGLLDIFGPPLTNEINTVXXXXXXXXXXXXXXXXXX 100
horse-Latherin MLKVSCLFVLLCGLLVPSSAQQIPPEVSSQITDALTQGLLDGNFLSLLNAINLEGLLNTILDQVTGLLNILVGPLLGP-SDAEIKLQDARLLQLSLEFSP 99
▽
human-BASE QRKEATVQVPFTSELIVQLLTMKPFTANMQSDIKVQIRLEKNVGGRYELAFGNCRLLPEAIWIQTGVQLAPAQNLLWQT------------------- 179
chimp-BASE XXXXXXXXXXXXXXXXXXLLTMKPFTANMQSDIKVQIHLEKNVGGRYELAFGNCRLLPEAIWIQTGVQLAPAAKFVVANIERNLKNIVAHDLGRKVCPVI 200
horse-Latherin DSKGIDIWIPLELSVYLKLLILEPLTLYVRTDIRVQLQLESDEDGKYRLAFGHCSLLPRAIELQSGNPLSLPVNAVLGTIENALGNFITEDLGAELCPTL 199
human-BASE -----------------------------
chimp-BASE NKWLYNLDQHVVKELINLVLLQEKYQVTI 229
horse-Latherin NSLVSNLDLQLVNNLINLILDRANVDLSV 228

Fig. 1. A 1 bp deletion mutation in the coding region of human *BASE* gene. (**A**) A schematic representation showing a sequence comparison of human and chimpanzee *BASE* genes. The coordinates of alignment blocks of the human *BASE* gene mRNA in the human genome (top) and in the chimpanzee genome (bottom) are shown. Boxes represent alignment blocks or exons. The coding region is in gray. Human *BASE* mRNA exon 6 is split into two blocks when aligned to the chimpanzee genome owing to an additional adenine nucleotide, extending the coding region of chimpanzee *BASE*. The exon 4 of chimpanzee *BASE* is missing owing to a sequencing gap in the chimpanzee genome sequence. Additional gaps are in introns 1 and 5. (**B**) Nucleotide and deduced amino acid sequence of the exon 6 of human and chimpanzee *BASE* genes. The 1 bp deletion in human *BASE* results in a premature termination (indicated by an asterisk) of BASE protein (**C**) A multiple sequence alignment of human BASE, chimpanzee BASE and horse Latherin protein (GenBank accession number AF491288) sequences. Residues identical in two or more species are highlighted with black background. Chimpanzee BASE shows similarity with horse Latherin at the C-terminal from the human-specific mutation position. This position is marked by an open triangle. X denotes an unidentified amino acid.

Table 1. List of human-specific frameshift insertion or deletion mutations

No.	Gene	Accession	Mutation type[a]	Chromosome[b]	Exon[c]	Description
1	*BASE*	NM_173958	del A	20/21	6/9	Breast cancer and salivary gland expression
2	*DNAJB3*	NM_001001394	ins G	2/13	1/1	DnaJ (Hsp40) homolog, subfamily B, member 3
3	*FLJ33674*	NM_207351	del A	3/2	3/3	Hypothetical protein FLJ33674
4	*HEJ1*	AF396440	ins AA	1/1	2/2	Transcribed pseudogene of *DNAJA1*
5	*NTSR2*	NM_012344	del C	2/12	4/4	Neurotensin receptor 2
6	*RPL13AP*	BC067891	ins A	14/15	1/1	Transcribed pseudogene of *RPL13A*
7	*SCGB1D4*	NM_206998	del T	11/9	2/3	Secretoglobin family 1D member 4
8	*WBSCR27*	NM_152559	ins CTGTGGACCGC	7/6	6/6	Williams Beuren syndrome chromosome region 27
9	*ZCCHC13*	NM_203303	ins C	X/X	1/1	Zinc finger, CCHC domain containing 13

[a] del, deletion; ins, insertion.
[b] Chromosome number (human/chimpanzee).
[c] Exon number harboring the mutation/total number of exons.

encode 145 and 224 amino acid proteins, respectively. Interestingly, a comparison of human and chimpanzee *DNAJB3* with macaque and mouse orthologs revealed that deletion of a thymine nucleotide near the C-terminus is common in human and chimpanzee. Macaque and mouse orthologs encode 242 amino acid proteins, and share many common amino acid residues at the C-terminus. This suggests that the mutation occurred in an ancestral species of both human and chimpanzee after divergence of monkeys and great apes. Macaque *DNAJB3* (reported as *MFSJ1*) and mouse *Dnajb3* (reported as *MSJ-1*) are specifically expressed in testis (Yu and Takenaka, 2003; Berruti and Martegani, 2005). It is unclear whether the chimpanzee *DNAJB3* that lacks 18 C-terminal residues compared with the macaque ortholog produces a functional protein, whereas the human *DNAJB3* is likely to be inactive as almost half of its residues are missing.

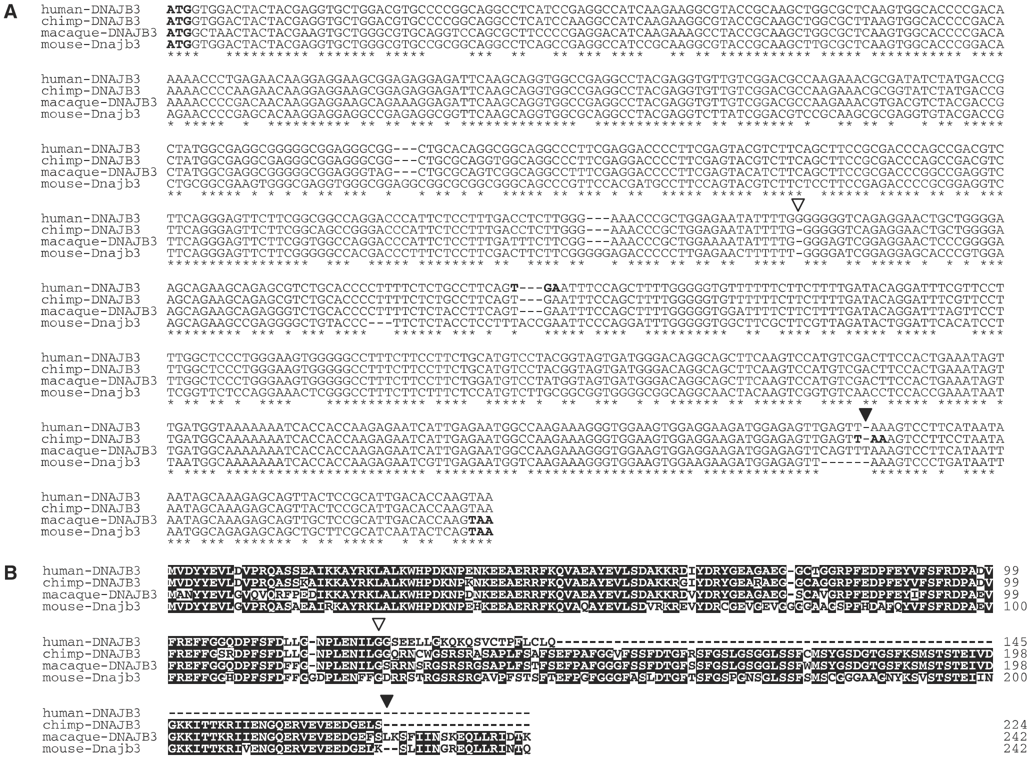

Fig. 2. Multiple sequence alignments of coding sequences (**A**) and deduced amino acid sequences (**B**) of *DNAJB3* genes from human, chimpanzee, macaque and mouse. Open and closed triangles indicate the positions of human-specific and human/chimpanzee common insertion and deletion, respectively. Translation start and stop codons are in bold type. Nucleotides conserved in all organisms are marked by asterisks. Amino acids conserved in two or more organisms are in white on black background. GenBank accession numbers for macaque *DNAJB3* and mouse *Dnajb3* are AB095737 and NM_008299, respectively.

3.4 *FLJ33674*

FLJ33674 is a hypothetical secreted protein (Clark *et al.*, 2003). A deletion of an adenine residue occurs near the C-terminus (Fig. 3A). The deleted adenine residue is embedded in a short poly guanine tract. The C-terminal region is relatively less conserved in the mouse homolog, B230206N24Rik (Fig. 4A). It is possible that the C-terminus is not critical for the function of the protein.

3.5 *NTSR2*

Comparison of human *NTSR2* (neurotensin receptor 2) with orthologs from chimpanzee and mouse revealed a human-specific cytosine (C) deletion in a short poly(C) tract near the C-terminus (Fig. 3B). The deduced C-terminal protein sequence of the chimpanzee NTSR2 shows a high level of homology to mouse Ntsr2 (Fig. 4B), demonstrating that the cytosine deletion is unique to human. *NTSR2* encodes a levocabastine-sensitive G protein-coupled neurotensin receptor 2 (Chalon *et al.*, 1996). It has seven transmembrane domains. The mutation in human NTSR2 occurs within the cytoplasmic tail after the seventh transmembrane domain. The frameshift mutation does not seem to exert an

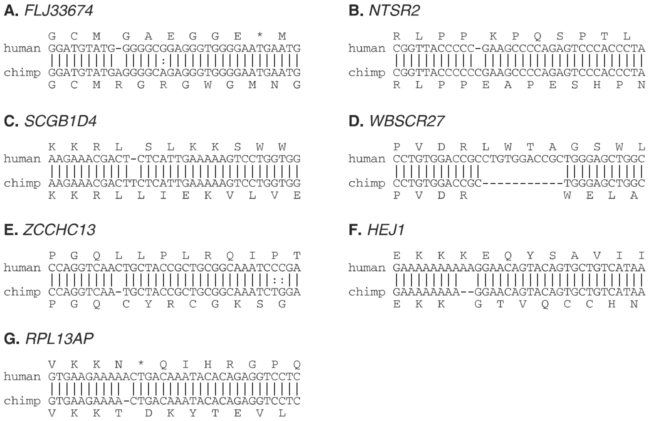
A. FLJ33674
G C M G A E G G E * M
human GGATGTATG-GGGGCGGAGGGTGGGAATGAATG
chimp GGATGTATGAGGGGCAGAGGGTGGGAATGAATG
G C M R G R G W G M N G
B. NTSR2
R L P P K P Q S P T L
human CGGTTACCCCC-GAAGCCCCAGAGTCCCACCCTA
chimp CGGTTACCCCCCGAAGCCCCAGAGTCCCACCCTA
R L P P E A P E S H P N
C. SCGB1D4
K K R L S L K K S W W
human AAGAAACGACT-CTCATTGAAAAAGTCCTGGTGG
chimp AAGAAACGACTTCTCATTGAAAAAGTCCTGGTGG
K K R L L I E K V L V E
D. WBSCR27
P V D R L W T A G S W L
human CCTGTGGACCGCCTGTGGACCGCTGGGAGCTGGC
chimp CCTGTGGACCGC-----------TGGGAGCTGGC
P V D R W E L A
E. ZCCHC13
P G Q L L P L R Q I P T
human CCAGGTCAACTGCTACCGCTGCGGCAAATCCCGA
chimp CCAGGTCAA-TGCTACCGCTGCGGCAAATCTGGA
P G Q C Y R C G K S G
F. HEJ1
E K K K E Q Y S A V I I
human GAAAAAAAAAAGGAACAGTACAGTGCTGTCATAA
chimp GAAAAAAAA--GGAACAGTACAGTGCTGTCATAA
E K K G T V Q C C H N
G. RPL13AP
V K K N * Q I H R G P Q
human GTGAAGAAAAACTGACAAATACACAGAGGTCCTC
chimp GTGAAGAAAA-CTGACAAATACACAGAGGTCCTC
V K K T D K Y T E V L

Fig. 3. Pair wise comparisons of human and chimpanzee mRNA sequences. Nucleotide sequences surrounding the putative mutation positions of seven human genes, (**A**) *FLJ33674*, (**B**) *NTSR2*, (**C**) *SCGB1D4*, (**D**) *WBSCR27*, (**E**) *ZCCHC13*, (**F**) *HEJ1* and (**G**) *RPL13AP* were aligned with respective chimpanzee counterparts. Gaps resulting from insertion or deletion in the human genes are indicated by horizontal lines. Matched and mismatched bases are marked by vertical bars and colons, respectively. Deduced amino acids are shown above (human) and below (chimpanzee) each mRNA sequence.

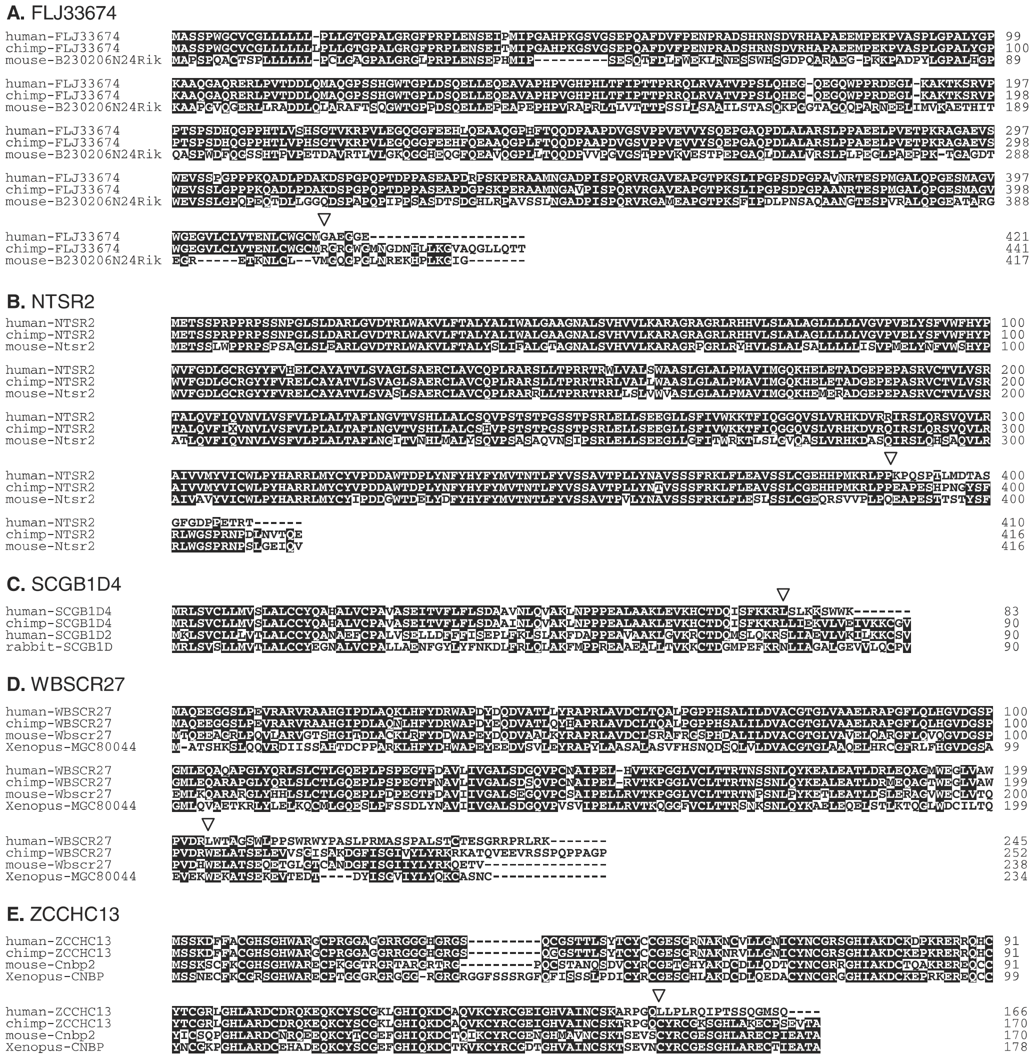

Fig. 4. Multiple sequence alignments of protein sequences showing human-specific frameshift mutations. Six human proteins, (**A**) FLJ33674, (**B**) NTSR2, (**C**) SCGB1D4, (**D**) WBSCR27 and (**E**) ZCCHC13 are aligned with respective homologs. Conserved amino acids in majority are highlighted with black background. A triangle indicates the location of the human-specific frameshift mutation in each case. GenBank accession numbers for the sequences except noted in Table 1 are as follows: mouse B230206N24Rik, NM_172487; mouse Ntsr2, AB041826; human SCGB1D2, NM_006551; rabbit SCGB1D, AY303698; mouse Wbscr27, NM_024479; *Xenopus* MGC80044, BC068654; mouse Cnbp2, AJ421478; *Xenopus* CNBP, Y07751.

effect on the functional domains of the neurotensin receptor 2 (Vita *et al*., 1998; Martin *et al*., 2002).

3.6 *SCGB1D4*

SCGB1D4 (secretoglobin family 1D member 4; also known as IIS for inteferon-γ-inducible SCGB) is a member of secretoglobin superfamily of genes. Its expression is inducible by inteferon-γ, a cytokine that stimulates the immune system (Choi *et al*., 2004). It is expressed in virtually all tissues with the highest level in lymph nodes, tonsil and ovary. A multiple sequence alignment of human SCGB1D4, predicted chimpanzee SCGB1D4, human SCGB1D2 and rabbit SCGB1D shows that the difference in the C-terminal region is unique to human SCGB1D4 (Fig. 4C). The deletion of a thymine residue at the 75th codon, CTT, is responsible for this difference (Fig. 3C). The frameshift mutation abolishes the last cysteine, which is conserved in other closely related secretoglobin proteins. Although the functional consequence of the removal of the last cysteine is yet to be elucidated, the resultant SCGB1D4 has been reported to be still functional in chemotactic migration and in invasion of lymphoblast cells (Choi *et al*., 2004).

3.7 *WBSCR27*

Comparison of human *WBSCR27* coding exons with the chimpanzee genome uncovered an 11 bp insertion in human *WBSCR27* (Fig. 3D). The exact 11 residues are repeated in tandem in exon 6 of human *WBSCR27*. Sequence comparison of human and chimpanzee WBSCR27 protein with mouse Wbscr27 and *Xenopus* hypothetical protein MGC80044 verified that the insertion occurred specifically in human lineage (Fig. 4D). *WBSCR27* is one of the genes assigned to the Williams-Beuren syndrome (WBS) critical region. The WBS is a neurodevelopmental disorder caused by a chromosomal microdeletion at 7q11.23 (Tassabehji, 2003). The biological function or the relationship of *WBSCR27* gene with WBS has not been reported.

3.8 *ZCCHC13*

ZCCHC13 (also known as *CNBP2*) is mapped in the X-inactivation region in mouse and human (Chureau *et al*., 2002). A Pfam search identified 5 CCHC zinc finger domains for human ZCCHC13, whereas it predicted 6 fingers for chimpanzee ZCCHC13. The loss of the 6th finger of the putative human/chimpanzee ancestral *ZCCHC13* is caused by an insertion of cytosine nucleotide between the 150th and the 151st codons in human (Figs 3E and 4E). The mouse Cnbp2 and *Xenopus* CNBP have seven CCHC zinc fingers. The second finger in these proteins is missing in human and chimpanzee ZCCHC13 owing to a point mutation that replaces histidine (H) with arginine (R) within the corresponding region (Fig. 4E). Mouse *Cnbp2* was detected only in adult mouse testis by RT–PCR (Chureau *et al*., 2002), and all human and mouse expressed sequence tags with defined tissue source were isolated from testis (UniGene Clusters Hs.157231 and Mm.159414), implying its involvement in the reproductive process. It is interesting that there is a progressive loss of fingers from 7 to 6 to 5 as one moves up the evolutionary ladder from frog and mouse to chimpanzee and then to human.

3.9 *HEJ1* and *RPL13AP*

Two genes, *HEJ1* and *RPL13AP*, showed two human-specific and one adenine nucleotides insertions, respectively, within the predicted coding regions (Fig. 3F and G). Comparison of HEJ1 and the mouse homolog Dnaja1 [DnaJ (Hsp40) homolog, subfamily A, member 1, GenBank accession number NM_008298] raises the possibility that *HEJ1* is a pseudogene since the predicted protein sequence lacks the J domain, a hallmark of the DnaJ family of proteins. Comparison of *HEJ1* and human *DNAJA1* (GenBank accession number NM_001539) verifies that *HEJ1* is a retrotransposed pseudogene derived from a partially processed *DNAJA1* mRNA. Similarly, comparison of the cDNA clone BC067891, which is named RPL13AP in this study, with mouse Rpl13a (ribosomal protein L13a, GenBank accession number NM_009438) and human RPL13A (GenBank accession number NM_012423), suggests that it is also a retrotransposed pseudogene derived from a fully processed *RPL13A* mRNA. The retrotransposition of these two pseudogenes occurred before divergence of the human and the chimpanzee lineage. It is possible that these two cases represent 'decaying' pseudogenes and do not produce functional proteins.

4 DISCUSSION

Many distinct human traits are presumably the results of many genetic modifications (Gagneux and Varki, 2001). The most direct way to find human-specific genetic alterations would be a comparative analysis of the human and the chimpanzee genomes, which begins to be possible with the release of the chimpanzee genome sequence. We have developed a simple method to collect putative human-specific frameshift mutations that occurred after the divergence of human and chimpanzee. The method involves collection of chimpanzee coding exons showing interrupted alignment with corresponding human coding exons. We focused on human-specific frameshift mutations because the human genomic sequence is nearly complete and many mRNA sequences are available for crosschecking its validity. Chimpanzee genome sequence could contain more errors but the ancestral nature of a particular gene sequence can still be discerned if it is conserved in non-human/non-chimpanzee homologs. By development and application of sequential computational filters to sort out publicly available data including human mRNA-to-chimpanzee genome alignments, we have identified nine human genes each with a human lineage-specific frameshift mutation.

Two of the genes identified in this study, *BASE* and *DNAJB3*, may represent 'dying genes' during human or great ape evolution. The frameshift mutations in these genes result in truncation of a substantial portion of their C-termini, possibly leading to inactivation of the proteins (Figs 1C and 2B). BASE belongs to the PLUNC (secreted proteins, palate, lung and nasal epithelium clones) family of proteins, which seem to mediate host defense functions in the mouth, nose and upper airways (Bingle and Gorr, 2004). More than 10 evolutionarily related PLUNC genes, including *BASE*, are found on the human chromosomal band 20q11.21 and on the orthologous chromosomal region in rodents (Bingle *et al*., 2004). The *DNAJB3* encodes a sperm-specific member of DnaJ protein family for spermiogenesis in macaque and mouse (Yu and Takenaka, 2003; Berruti and Martegani, 2005). This case is intriguing in that, besides the human-specific guanine insertion, there is a human/chimpanzee common thymine deletion near the C-terminus compared with the macaque ortholog (Fig. 2A). The biological outcome of the loss of the short C-terminal tail in chimpanzee or the half of the protein in human, and the exact timing of the thymine deletion in the great ape evolution are yet to be determined.

Two other proteins with altered C-termini, NTSR2 and SCGB1D4, were experimentally proved to be functional in human (Vita *et al*., 1998; Martin *et al*., 2002). The biological functions of the other three proteins with altered C-termini, FLJ33674, WBSCR27 and ZCCHC13, have not been reported. Since the mutations occur near the C-termini, they may exhibit full or at least limited functionality. The remaining two genes, *HEJ1* and *RPL13AP*, are likely to be transcribed pseudogenes produced by retrotransposition. Although they manifested human-specific frameshift mutations in the predicted coding region, they may not encode any functional protein. Retrotransposition and gene decay are common evolutionary processes observed in mammalian genomes (Zhang *et al*., 2002).

The sequences at the insertion or deletion sites suggest the molecular mechanism of the frameshift mutations. Six of nine frameshift mutations reported in this study, 2 in *DNAJB3* and 1 in each of *NTSR2*, *SCGB1D4*, *HEJ1* and *RPL13AP*, occurred within mononucleotide runs. The deleted adenine residue in *FLJ33674* is embedded in a poly-guanine tract. The frameshift mutation in *BASE* involves a deletion of an adenine residue within a doublet of trinucleotide GCAGCA, resulting in a doublet of dinucleotide GCGC. *WBSCR27* has a human-specific duplicated 11mer within the coding sequence. The existence of short monomeric or multimeric nucleotide repeats suggests emergence of frameshift mutations by replication slippage errors (Kunkel and Bebenek, 2000).

Although a handful of human-specific mutations have been reported previously, none of them were found in this study. Examination of database records of those genes revealed the reason they were not detected and innate limitations of the current approach. *MYH16* has a 2 bp deletion in exon 18 (Stedman *et al*., 2004). However, no *MYH16* mRNA sequence has yet been deposited in GenBank. It was identified as a pseudogene in the human genome. *EMR4* has a 1 bp deletion in exon 8 (Hamann *et al*., 2003) but no coding region information has been put in the GenBank entry (AF489700). *KRTHAP1* was identified as a pseudogene and has no mRNA entry in GenBank. Furthermore, it had been inactivated by a nonsense mutation (Winter *et al*., 2001). Inactivation of *CMAH* (GenBank accession number BC022302) occurred by an Alu-mediated exon deletion (Hayakawa *et al*., 2001). The procedure developed in this study requires defined coding sequence information in the database and is designed to find an insertion or a deletion mutation within a coding exon. None of the above cases meets these conditions.

The procedure adopted in this study is designed to find novel frameshift mutations based on predicted coding sequences. A reciprocal study to find chimpanzee-specific frameshift mutations can be made by a slight modification of the reported method. However, the current chimpanzee genome sequence may contain sequence errors that impede proper interpretation of the result. We await the high quality finished chimpanzee genome sequence such as that for the chromosome 22 (Watanabe *et al*., 2004).

Watanabe *et al*. (2004) reported 32 cases where the start ATG or the stop codon is different from their human counterparts. When we reviewed alignment data of 24 cases where the stop codon was changed, we found only 3 cases showing obvious human-specific mutations. These were *C21orf30*, *C21orf 71* and *LIPI* (GenBank accession numbers AL117578, AF086441 and BC028732, respectively). The first two involved nonsense mutations that were not considered in this study. The third case was a frameshift mutation. However, the corresponding GenBank record does not contain coding sequence information. And furthermore, the record had been removed according to submitter's request and it did not produce alignment data in the Genome Browser Database.

In summary, we developed a simple method for a genome-wide detection of human-specific frameshift mutations based on publicly available databases, and identified nine genes that had been specifically modified in the human lineage. The procedure is readily applicable to any species for which a high quality genome sequence is available. It is also possible to collect nonsense mutation-mediated human-specific premature terminations by a minimal modification of the current method. This study demonstrates that even the draft-quality chimpanzee genome sequence delivers useful information for the study of the human evolution.

ACKNOWLEDGEMENT

We thank Ira H. Pastan for reading the manuscript and for making valuable suggestions.

REFERENCES

Angata,T., Margulies,E.H., Green,E.D. and Varki,A. (2004) Large-scale sequencing of the CD33-related Siglec gene cluster in five mammalian species reveals rapid evolution by multiple mechanisms. *Proc. Natl Acad. Sci. USA*, **101**, 13251–13256.

Bailey,J.A., Yavor,A.M., Viggiano,L., Misceo,D., Horvath,J.E., Archidiacono,N., Schwartz,S., Rocchi,M. and Eichler,E.E. (2002) Human-specific duplication and mosaic transcripts: the recent paralogous structure of chromosome 22. *Am. J. Hum. Genet.*, **70**, 83–100.

Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L. *et al*. (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–D141.

Berruti,G. and Martegani,E. (2005) The deubiquitinating enzyme mUBPy interacts with the sperm-specific molecular chaperone MSJ-1: the relation with the proteasome, acrosome, and centrosome in mouse male germ cells. *Biol. Reprod.*, **72**, 14–21.

Bingle,C.D., LeClair,E.E., Havard,S., Bingle,L., Gillingham,P. and Craven,C.J. (2004) Phylogenetic and evolutionary analysis of the PLUNC gene family. *Protein Sci.*, **13**, 422–430.

Bingle,C.D. and Gorr,S.U. (2004) Host defense in oral and airway epithelia: chromosome 20 contributes a new protein family. *Int. J. Biochem. Cell Biol.*, **36**, 2144–2152.

Burki,F. and Kaessmann,H. (2004) Birth and adaptive evolution of a hominoid gene that supports high neurotransmitter flux. *Nat. Genet.*, **36**, 1061–1063.

Chalon,P., Vita,N., Kaghad,M., Guillemot,M., Bonnin,J., Delpech,B., Le Fur,G., Ferrara,P. and Caput,D. (1996) Molecular cloning of a levocabastine-sensitive neurotensin binding site. *FEBS Lett.*, **386**, 91–94.

Choi,M.S., Ray,R., Zhang,Z. and Mukherjee,A.B. (2004) IFN-gamma stimulates the expression of a novel secretoglobin that regulates chemotactic cell migration and invasion. *J. Immunol.*, **172**, 4245–4252.

Chureau,C., Prissette,M., Bourdet,A., Barbe,V., Cattolico,L., Jones,L., Eggen,A., Avner,P. and Duret,L. (2002) Comparative sequence analysis of the X-inactivation center region in mouse, human, and bovine. *Genome Res.*, **12**, 894–908.

Clark,H.F., Gurney,A.L., Abaya,E., Baker,K., Baldwin,D., Brush,J., Chen,J., Chow,B., Chui,C., Crowley,C. *et al*. (2003) The secreted protein discovery initiative (SPDI), a large-scale effort to identify novel human secreted and transmembrane proteins: a bioinformatics assessment. *Genome Res.*, **13**, 2265–2270.

Egland,K.A., Vincent,J.J., Strausberg,R., Lee,B. and Pastan,I. (2003) Discovery of the breast cancer gene BASE using a molecular approach to enrich for genes encoding membrane and secreted proteins. *Proc. Natl Acad. Sci. USA*, **100**, 1099–1104.

Enard,W., Przeworski,M., Fisher,S.E., Lai,C.S., Wiebe,V., Kitano,T., Monaco,A.P. and Pääbo,S. (2002) Molecular evolution of *FOXP2*, a gene involved in speech and language. *Nature*, **418**, 869–872.

Florea,L., Hartzell,G., Zhang,Z., Rubin,G.M. and Miller,W. (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.*, **8**, 967–974.

Fortna,A., Kim,Y., MacLaren,E., Marshall,K., Hahn,G., Meltesen,L., Brenton,M., Hink,R., Burgers,S., Hernandez-Boussard,T. *et al*. (2004) Lineage-specific gene duplication and loss in human and great ape evolution. *PLoS Biol.*, **2**, E207.

Gagneux,P. and Varki,A. (2001) Genetic differences between humans and great apes. *Mol. Phylogenet. Evol.*, **18**, 2–13.

Hamann,J., Kwakkenbos,M.J., de Jong,E.C., Heus,H., Olsen,A.S. and van Lier,R.A. (2003) Inactivation of the EGF-TM7 receptor EMR4 after the Pan-Homo divergence. *Eur. J. Immunol.*, **33**, 1365–1371.

Hayakawa,T., Satta,Y., Gagneux,P., Varki,A. and Takahata,N. (2001) Alu-mediated inactivation of the human CMP-*N*-acetylneuraminic acid hydroxylase gene. *Proc. Natl Acad. Sci. USA*, **98**, 11399–11404.

International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931–945.

Karolchik,D., Baertsch,R., Diekhans,M., Furey,T.S., Hinrichs,A., Lu,Y.T., Roskin,K.M., Schwartz,M., Sugnet,C.W., Thomas,D.J. *et al*. (2003) The UCSC Genome Browser Database. *Nucleic Acids Res.*, **31**, 51–54.

Khaitovich,P., Muetzel,B., She,X., Lachmann,M., Hellmann,I., Dietzsch,J., Steigele,S., Do,H.H., Weiss,G., Enard,W. *et al*. (2004) Regional patterns of gene expression in human and chimpanzee brains. *Genome Res.*, **14**, 1462–1473.

Kunkel,T.A. and Bebenek,K. (2000) DNA replication fidelity. *Annu. Rev. Biochem.*, **69**, 497–529.

Martin,S., Vincent,J.P. and Mazella,J. (2002) Recycling ability of the mouse and the human neurotensin type 2 receptors depends on a single tyrosine residue. *J. Cell. Sci.*, **115**, 165–173.

Notredame,C., Higgins,D.G. and Heringa,J. (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, **302**, 205–217.

Olson,M.V. (1999) When less is more: gene loss as an engine of evolutionary change. *Am. J. Hum. Genet.*, **64**, 18–23.

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

Schmutz,J., Wheeler,J., Grimwood,J., Dickson,M., Yang,J., Caoile,C., Bajorek,E., Black,S., Chan,Y.M., Denys,M. *et al*. (2004) Quality assessment of the human genome sequence. *Nature*, **429**, 365–368.

Stedman,H.H., Kozyak,B.W., Nelson,A., Thesier,D.M., Su,L.T., Low,D.W., Bridges,C.R., Shrager,J.B., Minugh-Purvis,N. and Mitchell,M.A. (2004) Myosin gene mutation correlates with anatomical changes in the human lineage. *Nature*, **428**, 415–418.

Tassabehji,M. (2003) Williams-Beuren syndrome: a challenge for genotype–phenotype correlations. *Hum. Mol. Genet.*, **12**, R229–R237.

Varki,A. (2004) How to make an ape brain. *Nat. Genet.*, **36**, 1034–1036.

Vita,N., Oury-Donat,F., Chalon,P., Guillemot,M., Kaghad,M., Bachy,A., Thurneyssen,O., Garcia,S., Poinot-Chazel,C., Casellas,P. *et al*. (1998) Neurotensin is an antagonist of the human neurotensin NT2 receptor expressed in Chinese hamster ovary cells. *Eur. J. Pharmacol.*, **360**, 265–272.

Watanabe,H., Fujiyama,A., Hattori,M., Taylor,T.D., Toyoda,A., Kuroki,Y., Noguchi,H., BenKahla,A., Lehrach,H., Sudbrak,R. *et al*. (2004) DNA sequence and comparative analysis of chimpanzee chromosome 22. *Nature*, **429**, 382–388.

Winter,H., Langbein,L., Krawczak,M., Cooper,D.N., Jave-Suarez,L.F., Rogers,M.A., Praetzel,S., Heidt,P.J. and Schweizer,J. (2001) Human type I hair keratin pseudogene φhHaA has functional orthologs in the chimpanzee and gorilla: evidence for recent inactivation of the human gene after the Pan-Homo divergence. *Hum. Genet.*, **108**, 37–42.

Yu,S.S. and Takenaka,O. (2003) Molecular cloning, structure, and testis-specific expression of MFSJ1, a member of the DNAJ protein family, in the Japanese monkey (*Macaca fuscata*). *Biochem. Biophys. Res. Commun.*, **301**, 443–449.

Zhang,Z., Harrison,P. and Gerstein,M. (2002) Identification and analysis of over 2000 ribosomal protein pseudogenes in the human genome. *Genome Res.*, **12**, 1466–1482.