



Mapping of chimpanzee full-length cDNAs onto the human genome unveils large potential divergence of the transcriptome

Ryuichi Sakate^{a,b,*}, Yumiko Suto^c, Tadashi Imanishi^b, Tetsuya Tanoue^d, Munetomo Hida^e, Ikuo Hayasaka^d, Jun Kusuda^f, Takashi Gojobori^{b,g}, Katsuyuki Hashimoto^{f,h}, Momoki Hirai^e

^a Japan Biological Information Research Center, Japan Biological Informatics Consortium:

AIST Bio-IT Research Bldg. 7F, 2-42 Aomi, Koto-ku, Tokyo, 135-0064, Japan

^b Biological Information Research Center, National Institute of Advanced Industrial Science and Technology:

AIST Bio-IT Research Bldg. 7F, 2-42 Aomi, Koto-ku, Tokyo, 135-0064, Japan

^c Department of Research, Japanese Red Cross Central Blood Institute: 2-1-67 Tatsumi, Koto-ku, Tokyo, 135-8521, Japan

^d Kumamoto Primate Research Park, Sanwa Kagaku Kenkyusho Co., Ltd.: 990 Ohtao, Misumi-cho, Uki-shi, Kumamoto, 869-3201, Japan

^e International Research and Educational Institute for Integrated Medical Sciences, Tokyo Women's Medical University:

8-1 Kawada-cho, Shinjuku-ku, Tokyo, 162-8666, Japan

^f Division of Biomedical Research Resources, National Institute of Biomedical Innovation: 7-6-8 Saito-Asagi, Ibaraki, Osaka, 567-0085, Japan

^g Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics: Mishima, Shizuoka, 411-8540, Japan

^h Division of Genetic Resources, National Institute of Infectious Diseases: 1-23-1 Toyama, Shinjuku-ku, Tokyo, 162-8640, Japan

Received 16 December 2006; received in revised form 8 April 2007; accepted 13 April 2007

Available online 20 April 2007

Abstract

The genetic basis of the phenotypic difference between human and chimpanzee is one of the most actively pursued issues in current genomics. Although the genomic divergence between the two species has been described, the transcriptomic divergence has not been well documented. Thus, we newly sequenced and analyzed chimpanzee full-length cDNAs (FLcDNAs) representing 87 protein-coding genes. The number of nucleotide substitutions and sites of insertions/deletions (indels) was counted as a measure of sequence divergence between the chimpanzee FLcDNAs and the human genome onto which the FLcDNAs were mapped. Difference in transcription start/termination sites (TSSs/TTSSs) and alternative splicing (AS) exons was also counted as a measure of structural divergence between the chimpanzee FLcDNAs and their orthologous human transcripts (NCBI RefSeq). As a result, we found that transposons (*Alu*) and repetitive segments caused large indels, which strikingly increased the average amount of sequence divergence up to more than 2% in the 3'-UTRs. Moreover, 20 out of the 87 transcripts contained more than 10% structural divergence in length. In particular, two-thirds of the structural divergence was found in the 3'-UTRs, and variable transcription start sites were conspicuous in the 5'-UTRs. As both transcriptional and translational efficiency were supposed to be related to 5'- and 3'-UTR sequences, these results lead to the idea that the difference in gene regulation can be a major cause of the difference in phenotype between human and chimpanzee. © 2007 Elsevier B.V. All rights reserved.

Keywords: Human evolution; Sequence analysis; Full-length cDNA; Gene structure; Bioinformatics

1. Introduction

It has been expected that comparative analyses of human and chimpanzee provide us with new knowledge about various biological features unique to human. As chimpanzee is the closest living relative to human in the phylogenetic lineage, it has been considered to be the most suitable species for comparison. Earlier microscopic analyses of chromosomes revealed not only the overall resemblance of karyotypes but also the frequent species-specific rearrangements among anthropoidea

Abbreviations: CDS, coding sequence; FLcDNA, full-length cDNA; Myr, million years; SINE, short interspersed element; TSS, transcription start site; TTS, transcription termination site; UTR, untranslated region.

* Corresponding author. Japan Biological Information Research Center, Japan Biological Informatics Consortium: AIST Bio-IT Research Bldg. 7F, 2-42 Aomi, Koto-ku, Tokyo, 135-0064, Japan. Tel.: +81 3 3599 8800; fax: +81 3 3599 8801.

E-mail address: sakate@prigen.org (R. Sakate).

(Yunis et al., 1980). Currently, the complete sequence of the human genome (International Human Genome Sequencing Consortium, 2004) has given us an opportunity to conduct detailed analyses comparing nucleotide by nucleotide. Also for the chimpanzee genome, the whole genome draft sequences (Olson and Varki, 2003; The Chimpanzee Sequencing and Analysis Consortium, 2005), and more accurate sequences of chromosomes 22 (Watanabe et al., 2004) and Y (Kuroki et al., 2006) are available to date. The previous studies have shown that the amount of sequence divergence between the human and chimpanzee genomes is 1.22–1.23% (Chen et al. 2001; Fujiyama et al., 2002). However, inconsistency between the highly conserved genome sequences and the distinctive differences in phenotype is still an issue. In order to address this issue, we considered that detailed analyses of the transcriptome should be prerequisite.

For examining the transcriptomic differences between human and chimpanzee in detail, essential information is mainly available from full-length cDNAs (FLcDNAs). FLcDNA sequences provide information on where transcription starts and terminates, where exons and introns exist, and where 5'- and 3'-untranslated regions (UTRs) and coding sequences (CDSs) are located on the genome. For comparative analyses, chimpanzee FLcDNAs are still insufficient in the public databases, while large number of sequences of human (Imanishi et al., 2004) and mouse (Okazaki et al., 2002) transcripts are available. In our previous study, we constructed chimpanzee FLcDNA libraries and sequenced the 5'-ends of the chimpanzee cDNAs representing 226 protein-coding genes (Sakate et al., 2003). We found that the average nucleotide difference between human and chimpanzee transcripts was 0.70% and those of 5'-UTRs and CDSs were 1.21% and 0.58%, respectively. Several studies have also reported the comparative analyses of human and chimpanzee transcripts (Chen et al., 2001; Hellmann et al., 2003). However, these previous studies analyzed mostly partial sequences such as expressed sequence tags (ESTs) or predicted transcripts deduced *in silico* from the genome.

In the present study, we newly sequenced chimpanzee FLcDNAs and obtained full-length sequences representing 87 protein-coding genes. Using the full-length chimpanzee sequences containing complete 5'-UTRs, CDSs and 3'-UTRs, we could conduct whole-transcript sequence comparison between human and chimpanzee. We analyzed, as "sequence divergence", substitutions and insertions and deletions (indels). We also analyzed, as "structural divergence", transcription start/termination sites and exon–intron boundaries. In particular, structural divergence analysis cannot be conducted without FLcDNA sequences.

2. Materials and methods

2.1. Sequencing full-length cDNA clones

Full-length enriched cDNA (FLcDNA) libraries were constructed as described previously (Sakate et al., 2003). Tissue specimens were collected from adult chimpanzees (*Pan troglodytes verus*) kept in the Primate Research Park, Sanwa Kagaku Kenkyusho Co., Ltd., Kumamoto, Japan. Brain tissue

samples were from cerebellar cortex, cerebral cortex frontal lobe, and cerebral cortex occipital lobe. They were obtained by autopsy from an 11-year-old female chimpanzee that died of septicemia. Skin epidermal tissue samples were collected by biopsy from a 19-year-old male chimpanzee. An institutional board approved all the tissue collection procedures. The FLcDNA libraries were constructed by the oligo-capping method (Maruyama and Sugano, 1994; Hida et al., 2000).

First, the 5'-end of each cDNA (5'-cDNA) clone was randomly chosen and sequenced by using ABI 377 and ABI 3100 sequencers. More than 10,000 redundant 5'-cDNA sequences were compared with NCBI human RefSeq mRNAs (Release 17, 2006) by using BLASTN 2.2.11 (Altschul et al., 1990) with the "-e 1e-120 -F F" options. 5'-cDNA clones that had the longest 5'-UTR sequence and the same translation start site for each gene were selected. Then, full-length sequencing of the selected clones was conducted by using the primer walking method. After trimming the vector and truncating poly-A sequences, the FLcDNA sequences were subjected to BLASTN search against human RefSeq mRNAs using the "-e 1e-120 -F F" options. A human ortholog candidate for a chimpanzee sequence was determined by selecting the most similar human sequence with the highest score from the BLAST result. For each gene, the chimpanzee FLcDNA and the human RefSeq mRNA were aligned by using ClustalW (Thompson et al., 1994). The 5'- and 3'-untranslated regions (UTRs) and coding sequences (CDSs) of the chimpanzee transcripts were determined by assigning the RefSeq mRNA annotation.

2.2. Genome mapping and analyses of sequence divergence and structural divergence

The chimpanzee FLcDNAs and their corresponding human RefSeq mRNAs were mapped onto the human genome (UCSC hg16). They were subjected to BLAT (Kent, 2002) search with default options to determine candidate genomic loci. Supportively BLASTN was used with the "-F F" option and the *E* value options -e (1e-60–1e-10), if the genomic location of 5'- or 3'-end exons were not found by BLAT. After sequence identity, coverage and the number of exons were examined in this order, a gene locus was determined. The exon–intron structure was determined by using both BLAT and est2genome (Rice et al., 2000) with the default options. The exon–intron boundaries were visually inspected. The orthology of each corresponding pair of human and chimpanzee transcripts was finally determined if they were mapped onto the same locus on the human genome.

The amount of sequence divergence between the chimpanzee transcripts and the human genome was estimated by analyzing substitutions and indels in the sequence alignments (Fig. 1a). First, we analyzed substitutions only. The number of substitutions per site was calculated as "mismatch/(match+mismatch)" in the nucleotide and amino acid sequence alignments (*p*-distance). No correction for multiple substitutions was conducted, because the number of nucleotide differences between human and chimpanzee is small. Next, the amount of sequence divergence was estimated by adding the length of indels (*L*) as "(mismatch+*L*)/

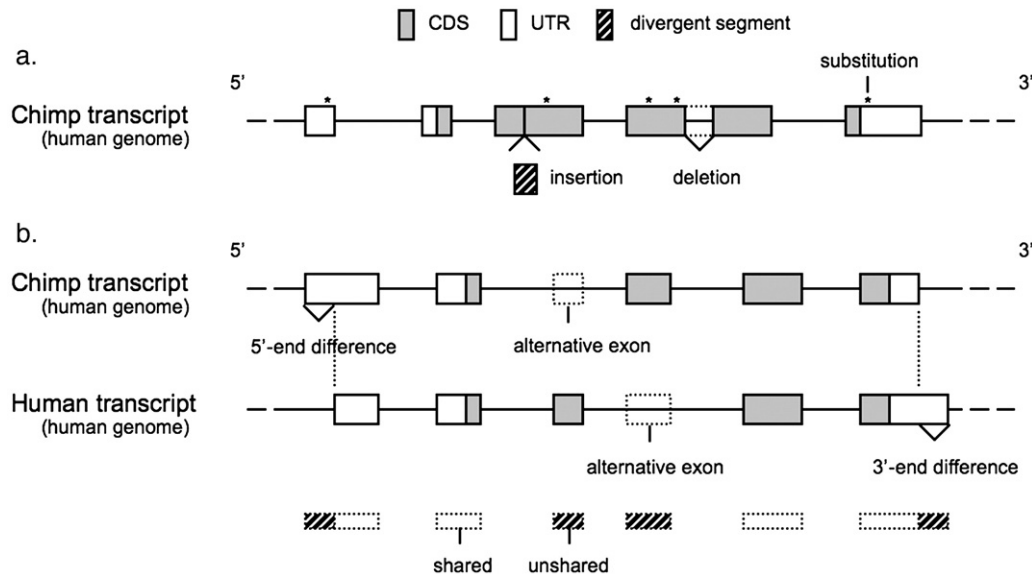


Fig. 1. Two approaches to sequence comparison. a. Comparison between the chimpanzee transcripts and the human genome at the nucleotide level. Substitutions and indels were analyzed to estimate the amount of “sequence divergence”. b. Comparison between the chimpanzee transcripts and the human transcripts at the exon level. Exon–intron boundaries including transcription start/termination sites (TSSs/TTSS) were analyzed. Substitutions and indels within exons were not analyzed. The lengths of shared and unshared regions on the human genome were calculated to estimate the amount of “structural divergence”.

(match + mismatch + L)”. A chimpanzee deletion or human insertion (Pdel/Hins) was not counted as an indel, if it could be a canonical intron (an intron with GT–AG internal boundaries) of the chimpanzee transcript. In addition, the numbers of synonymous (dS) and non-synonymous (dN) substitutions per site were estimated by the Nei–Gojobori method (Nei and Gojobori, 1986) using SNAP (Ota and Nei, 1994). If a large indel was found, the sequence region was analyzed by using the chimpanzee genome sequence (UCSC panTro1).

The amount of structural divergence between the chimpanzee transcripts and the human transcripts was estimated by analyzing corresponding exons on the human genome (Fig. 1b). The exon–intron boundaries on the human genome were compared between the species. The transcription start site (TSS, 5'-end) and the transcription termination site (TTS, 3'-end) on the human genome were also compared between the species. Indels were not counted in the analysis of structural divergence. The lengths of shared and unshared human genomic regions were counted. The amount of structural divergence was calculated as “unshared/(shared + unshared)” (bp).

3. Results

3.1. Chimpanzee FLcDNAs

Chimpanzee transcripts representing 87 protein-coding genes originated from full-length enriched cDNA (FLcDNA) libraries. Of the 87 genes, 49 were from brain tissue samples (clone: PccB, PflB and PorA/B) and 38 were from the skin epidermal tissue samples (PstA) (Supplementary Table 1). The 87 genes were relatively abundant in number and chosen as representative genes expressed in each tissue (keratin genes from skin, etc.). Annotation of the 87 chimpanzee transcripts, assigning gene

names and positions of the coding sequences (CDSs), was performed by comparing them with their orthologous human transcripts (RefSeq). The lengths of the 87 chimpanzee transcripts ranged from 510 bp (PorA0990, NM_000990, *RPL27A*) to 3188 bp (PflB2127, NM_005822, *DSCR1L1*), the average length being 1860 bp. The average lengths of the 5'-UTRs, CDSs and 3'-UTRs were 101 bp, 1167 bp and 592 bp, respectively.

The chimpanzee transcripts were mapped onto the human genome. The 87 genes represented by these transcripts were distributed widely among all the human chromosomes except chromosome Y (Supplementary Table 1). Sequence alignments between the human genome and the chimpanzee transcripts were examined nucleotide by nucleotide. Consequently, 99.4% of the nucleotides of the chimpanzee transcripts without poly-A tails were mapped onto the human genome. The unmapped nucleotides (0.6%) were considered as deletions in human or insertions in chimpanzee.

We used the human genome as a reference genome onto which the chimpanzee transcripts were mapped, because a large part of the transcripts could not be mapped onto the draft chimpanzee genome. In other words, the chimpanzee transcripts provide us with useful information on unsequenced regions of the chimpanzee genome. For example, PorA0463 (NM_001549, *IFIT3*) contained an unsequenced region of the chimpanzee genome. To the contrary, there was also a case where a transcript was located in an unsequenced region of the human genome. A 312 bp CDS segment of PstA5459 (NM_001733, *C1R*) (1096–1407/2285 bp) was located within an unsequenced region of the human genome (chr. 12, 7,059,294–7,132,293). This segment was an exception that was excluded from the analyses.

In the present study, two approaches to sequence comparison were taken. In the analysis of “sequence divergence”, the number of substitutions per site and the length of insertions and

deletions (indels) between the human genome and the chimpanzee transcripts were estimated (Fig. 1a). In the analysis of “structural divergence”, transcription start/termination sites (TSSs/TTSs) and alternative splicing (AS) exons were compared between the chimpanzee transcripts and their orthologous human transcripts, both of which were mapped onto the human genome (Fig. 1b).

3.2. Sequence divergence: substitution and indel

3.2.1. Substitution

In order to survey evolutionary constraint, nucleotide substitution between the human genome and the chimpanzee transcripts was analyzed. The average number of nucleotide substitutions per site for the entire set of transcripts with standard error ($\pm$ SE) was 0.73 ± 0.04 (Fig. 2). Those of the 5'-UTRs, CDSs, 3'-UTRs and amino acid sequences were 0.87 ± 0.15 , 0.59 ± 0.04 , 0.93 ± 0.08 and 0.58 ± 0.07 , respectively. Each of the 87 transcripts contained at least one nucleotide substitution and 65 transcripts (75%) contained at least one amino acid substitution compared to the human genome. The numbers of synonymous (dS) and non-synonymous (dN) substitutions per site were estimated by the Nei–Gojobori method (Nei and Gojobori, 1986). The averages of dS and dN for the entire set of transcripts were 1.77 ± 0.13 and 0.25 ± 0.03 , respectively. As the average numbers of nucleotide substitutions per site of the 5'-UTRs and 3'-UTRs (0.87 and 0.93, respectively) were significantly lower than the average dS (1.77), not only the CDSs but also both the 5'-UTRs and 3'-UTRs were possibly under strong functional constraints, at least in terms of substitutions.

3.2.2. Indel

In addition to substitutions, insertions and deletions (indels) were analyzed. Each gap in a pair-wise sequence alignment between human and chimpanzee was analyzed. A gap in the

human sequence was considered as an insertion in the chimpanzee transcript or a deletion in the human genome (Pins/Hdel), and vice versa for the opposite case (Pdel/Hins). A 5'- or 3'-end unmapped segment of a chimpanzee transcript was defined as an external indel (Pins/Hdel). As a result, out of the 87 alignments, 39 (45%) contained a total of 64 internal indels. There were 17 external indels found in 16 chimpanzee transcripts. When these external indels were included, 51 alignments (59%) contained a total of 81 indels (Supplementary Table 2). No indels causing frame-shifts in the CDS were found.

The difference in distribution of indels among the 5'-UTRs, CDSs and 3'-UTRs was also analyzed (Fig. 3). More than half (52, 64%) of all the 81 indels were found in the 3'-UTRs. For example, PorB0603 (NM_000284, *PDHAI*) and PstA2462 (NM_001424, *EMP2*) contained 4 indels, the largest number of indel occurrences per transcript, in the 3'-UTR (Supplementary Table 2). Among both 6 internal and 11 external indels in the 5'-UTRs, the largest one was a 13 bp external indel (PorB1356, NM_001493, *GDII*). Compared with the small indels in the 5'-UTRs, larger indels were found in the CDSs and 3'-UTRs. Among 12 indels in the CDSs, the largest one was a 126 bp Pins/Hdel-type indel (PstA5646, NM_001614, *ACTG1*) (Fig. 5a). Among both 46 internal and 6 external indels in the 3'-UTRs, the largest one was a 559 bp external indel (PccB1795, NM_000016, *ACADM*) (Fig. 5b). This 559 bp indel was located at the 3'-end and originated from *Alu* insertion (as described in section 3.4).

As the average lengths of the 5'-UTRs, CDSs and 3'-UTRs were different from each other as described in section 3.1 (101 bp, 1,167 bp and 592 bp, respectively), the number of occurrences and the size of indels per 1-kb alignment were estimated for each region. The number of indel occurrences per 1-kb alignment for the entire set of transcripts was 0.40 (0.50, if external indels were added), while those of the 5'-UTRs, CDSs and 3'-UTRs were 0.69 (1.94), 0.12 and 0.90 (1.01),

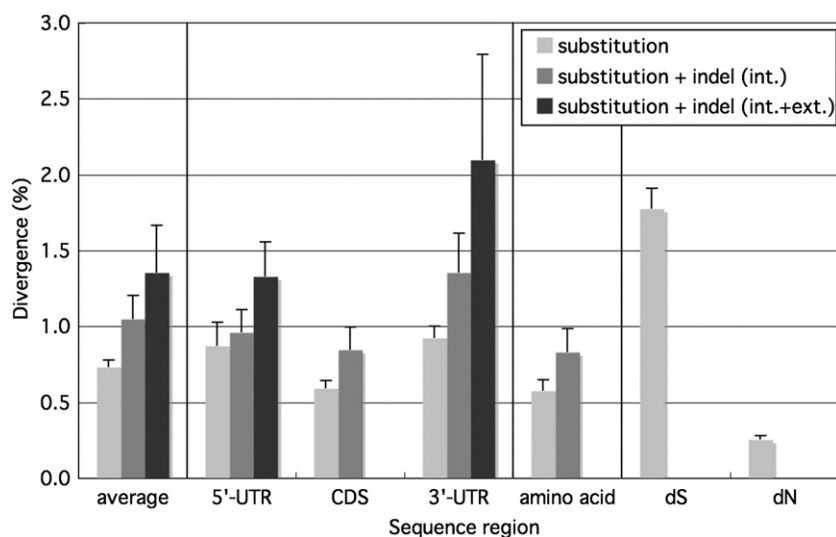


Fig. 2. Sequence divergence between the chimpanzee transcripts and the human genome are shown with standard error (SE). The number of substitutions per site and the length of internal (int.) indels (within alignments), and the length of external (ext.) indels (chimpanzee unmapped segments at 5'- or 3'-end, Pins/Hdel only) were counted in the 5'-UTRs, CDSs, 3'-UTRs and amino acid sequences. Synonymous (dS) and non-synonymous (dN) substitution rates are shown for comparison purposes. Indels influenced the amount of sequence divergence most drastically in the 3'-UTRs.

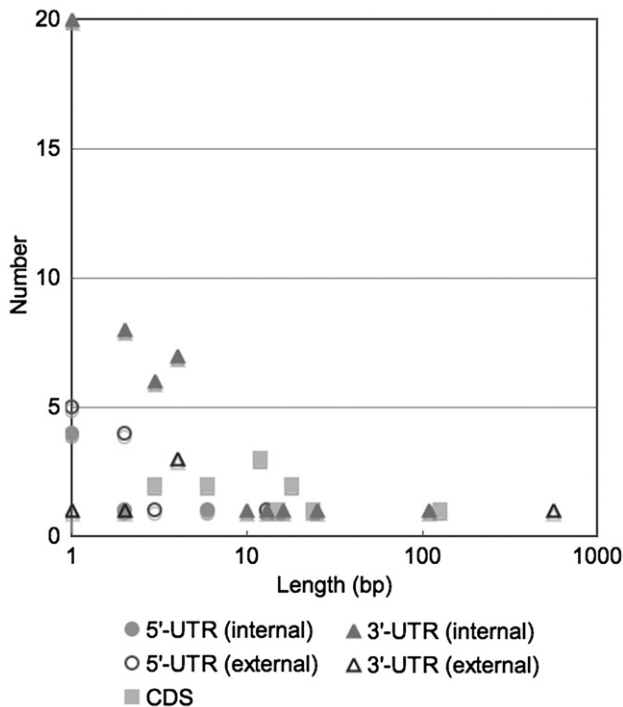


Fig. 3. Distribution of number and length of indels within the 5'-UTRs, CDSs and 3'-UTRs. For the 5'- and 3'-UTRs, both internal indels and external indels were plotted. Out of all the 81 indels, 52 (64%) were found in the 3'-UTRs. Larger indels (>100 bp) were found in the CDSs (126 bp, Fig. 5a) and 3'-UTRs (559 bp, Fig. 5b and 110 bp, Fig. 5c).

respectively. The size of indels per 1-kb alignment for the entire set of transcripts was 3.2 bp (7.0 bp), while those of the 5'-UTRs, CDSs and 3'-UTRs were 1.4 bp (4.7 bp), 2.5 bp and 5.0 bp (16.1 bp), respectively. These results show that indels occurred most frequently in the 3'-UTRs (5'-UTRs, if external indels were added), and that indels in the 3'-UTRs were drastically larger than those in the 5'-UTRs and CDSs.

Indel occurrences in the transcribed sequences (transcripts) and the whole genomic sequence were compared. The average number of indel occurrence (internal indels) for the transcripts was calculated to be 0.40/kb as described above. Since the divergence time between human and chimpanzee has been estimated to be about 5 Myr (million years), the evolutionary rate of indels would be 0.04/kb/Myr. The number of indel occurrences between the two species for the nuclear genome has been reported to be about 0.2/kb/Myr (Saitou, 1991; Saitou and Ueda, 1994). Therefore, the frequency of indel occurrence in transcripts was one-fifth of that in the genomic sequences.

The difference in number and size of indels between human and chimpanzee was also estimated. As external indels (Pins/Hdel-type indels) are located outside of alignments, internal indels were compared. Of all the 64 internal indels, 34 indels were Pins/Hdel-type and 30 indels were Pdel/Hins-type. The total length of the 34 Pins/Hdel-type indels was 374 bp, while the total length of the 30 Pdel/Hins-type indels was 149 bp. The number of Pins/Hdel-type indels was larger than the number of Pdel/Hins-type indels by only 13%. However, the Pins/Hdel-type indels were significantly larger than the Pdel/Hins-type indels (Fisher's exact test, $P < 0.005$).

3.2.3. Sequence divergence

Combining the number of nucleotide substitutions and the length of internal and external indels, the gross amount of sequence divergence was estimated. The lengths of internal indels and external indels were added to the number of nucleotide substitutions as "substitutions+internal indels" and "substitution+internal indels+external indels" (Fig. 2). If internal indels were added, the amount of sequence divergence for the entire set of transcripts became 1.05 ± 0.14 (%). Those of the 5'-UTRs, CDSs, 3'-UTRs and amino acid sequences became 0.96 ± 0.15 , 0.85 ± 0.15 , 1.36 ± 0.25 and 0.83 ± 0.16 , respectively. If both internal and external indels were added, the amount of sequence divergence for the entire set of transcripts became 1.36 ± 0.30 (%). Those of the 5'-UTRs and 3'-UTRs became 1.33 ± 0.22 and 2.10 ± 0.68 , respectively. Compared with the average number of substitutions per site, the addition of both internal and external indels increased the sequence divergence from 0.73% to 1.36% (a factor of 1.9). Indels in the 5'-UTRs, CDSs and 3'-UTRs increased the sequence divergences from 0.87% to 1.33% (a factor of 1.5), from 0.59% to 0.85% (a factor of 1.4) and from 0.93% to 2.10% (a factor of 2.3), respectively. Indels increased the sequence divergence drastically, especially in the 3'-UTRs.

In fact, this large sequence divergence in the 3'-UTRs was produced by a limited number of large indels. The 3'-UTR indels included the largest indel, a 559 bp Pins/Hdel-type external indel. If this exceptionally large indel was excluded from calculation, then the amount of sequence divergence in the 3'-UTRs and the total transcripts turned out to be 1.45% (-0.65%) and 1.08% (-0.28%), respectively. In other words, large-scale sequence divergence at the nucleotide level exists in specific genes. And, as the large driving force for the sequence divergence, indels are particularly prevalent in the 3'-UTRs.

3.3. Structural divergence: TSS/TTS and AS

3.3.1. TSS/TTS and AS

A comparison of the genomic location of exons between the human and chimpanzee transcripts was conducted (Fig. 1b). Orthologous RefSeq mRNA sequences were used as the human transcripts. The chimpanzee unmapped segments (external indels) were not used in this analysis.

First, the chimpanzee and human transcription start sites (TSSs, 5'-end) and termination sites (TTSs, 3'-end) on the human genome were compared (Fig. 4). Among the 87 orthologous transcript pairs, 7 pairs had the same TSS and 32 pairs had the same TTS. The number of pairs that were identical at both TSS and TTS was only 3. These results imply that the transcription process for human–chimpanzee orthologous transcripts tends to start at different sites, and that it tends to terminate at the same site. The distribution pattern was further examined statistically. If both ends of the column (≤ -111 bp and ≥ 111 bp, largely different pairs) in Fig. 4 were excluded, 77 pairs for TSS and 73 pairs for TTS remained. The degrees of kurtosis of the remaining distribution (-110 bp– 110 bp) for TSS and TTS were 5.9 and 10.9, respectively, while the degrees of skewness were 2.2 and 3.3, respectively. The distribution

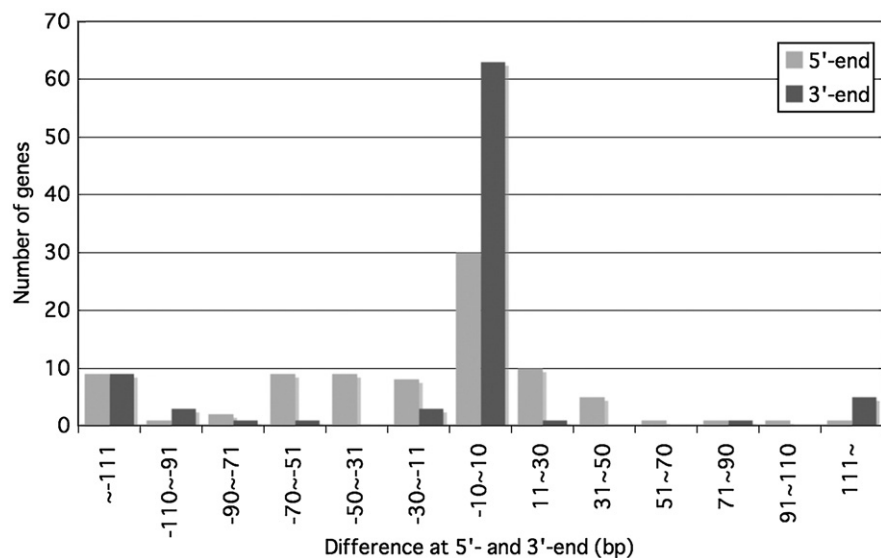


Fig. 4. Comparisons of transcription start/termination sites (TSSs/TTSs) on the human genome between the chimpanzee transcripts and their orthologous human transcripts (NCBI RefSeq). The x-axis corresponds to how much a chimpanzee transcript was “longer (+)” or shorter (–) than its orthologous human transcript at both 5'- and 3'-ends. Within a ±100 bp range, TSSs was found to be more diverged than TTSs (contingency table test, $P<0.001$).

patterns of TSS and TTS were statistically different ($P<0.001$) in contingency table test (Sham and Curtis, 1995).

Second, correspondence of each exon pair was examined between the chimpanzee and human transcripts on the human genome. As a result, for the 87 transcript pairs, the human transcripts had 732 exons in total and the chimpanzee transcripts had 735 exons in total. Of these 732 and 735 exons, 728 belonged to both species. Among the 739 exons that belonged to the human and/or chimpanzee transcripts, 11 exons were unique to one of the species in 7 out of the 87 transcript pairs (Table 1). Of these 11 exons, 6 exons were unique to the chimpanzee transcripts, and 5 exons were unique to the human transcripts. These were considered as candidates for species-specific AS exons. Out of these 11 exons, 6 exons were located in the 5'-UTRs, 2 in the CDSs and 3 in the 3'-UTRs. Considering that the average length of the chimpanzee 5'-UTRs (101 bp) was less than those of the CDSs (1167 bp) and 3'-UTRs (592 bp), AS events occurred most frequently in the 5'-UTRs.

3.3.2. Structural divergence

Combining the difference between the human and chimpanzee transcripts in TSSs/TTSs and AS exons, the amount of structural divergence was estimated. The genomic regions transcribed by both species were regarded as shared regions, while the regions transcribed by only one species were regarded as unshared regions (Fig. 1b). The shared and unshared regions were determined by analyzing the locations of exon–intron boundaries including TSSs/TTSs (5'-/3'-ends). The total lengths of the shared and unshared regions on the human genome were calculated. In this analysis, substitutions and indels within an exon were not included in the calculation. The amount of structural divergence was calculated as “unshared/(shared+unshared)” (bp).

As a result, the amount of structural divergence was estimated to be more than 10% (11.5%) in total between the chimpanzee and human transcripts. This consisted of the differences in TSSs (2.0%), putative species-specific AS exons (1.7%) and TTSs (7.9%). More than two-thirds of the

Table 1
Comparison of AS exons between the chimpanzee transcripts (FLcDNAs) and their orthologous human transcripts (NCBI RefSeq)

cDNA	DDBJ Acc	RefSeq Acc	Gene symbol	AS type	Region	Species	Length
PccB0929	AB222104	NM_001918	DBT	Cassette	CDS	Pins/Hdel	126
PorA0450	AB222122	NM_007177	FAM107A	Cassette	5'-UTR	Pdel/Hins	98
PorA0465	AB222123	NM_012286	MORF4L2	Mutually exclusive	5'-UTR	Pins/Hdel	58
				Mutually exclusive	5'-UTR	Pdel/Hins	38
				Cassette	5'-UTR	Pdel/Hins	153
				Terminal	3'-UTR	Pins/Hdel	97
PorB0603	AB222138	NM_000284	PDHA1	Terminal	3'-UTR	Pins/Hdel	97
PstA1671	AB222146	NM_018479	ECHDC1	Mutually exclusive	5'-UTR	Pdel/Hins	245
				Mutually exclusive	5'-UTR	Pins/Hdel	52
				Cassette	CDS	Pins/Hdel	53 ^a
PstA2462	AB222164	NM_001424	EMP2	Terminal ^b	3'-UTR	Pins/Hdel	1103
PstA2589	AB188282	NM_003016	SFRS2	Retained intron	3'-UTR	Pdel/Hins	995

^a This “not in-frame” exon produces a different stop codon for NM_018479, which is located upstream from where the stop codon of PstA1671 is located.
^b A genomic segment containing *Alu* was inserted into the 3'-UTR of PstA2462 as a putative intron (non-canonical intron).

structural divergence was found at the TSSs. Among the 87 chimpanzee transcripts, 20 were found to contain more than 10% structural divergence in length (Supplementary Table 2). These results show that structural divergence between the human and chimpanzee transcripts could be significantly larger than that expected from genomic sequence comparisons ($\sim 1.2\%$). This implies that large divergence potentially can be produced even from the same genomic sequence by selecting different TSSs, TTSs and exons at the transcriptome level.

3.4. Transposon (*Alu*) and repetitive segment

In the analysis of the 87 chimpanzee transcripts (FLcDNAs), both the sequence divergence and the structural divergence were found to be largely influenced by transposons (*Alu*) and repetitive segments. They were located even in transcribed

genomic regions, especially in the 3'-UTRs. Although they were found only in a few transcripts exemplified below, they could produce large (>100 bp) sequence divergence and possibly functional differences.

Alu is a primate-specific member of the SINE (short interspersed element) family of retroposons (Repbase, Jurka, 2000). By sequence homology search, a 559 bp Pins/Hdel-type external indel in the 3'-UTRs of PccB1795 (NM_000016, *ACADM*) was found to contain one member of this family (Fig. 5b). The transcript segment containing *Alu* was mapped onto the chimpanzee genome but not onto the human genome. This insertion seemed to add at least one more exon of the chimpanzee transcript at the 3'-end on the chimpanzee genome.

Repetitive segments could be the source of indels by changing the number of their copy units between the species. In the present study, two sets of unknown repetitive segments were

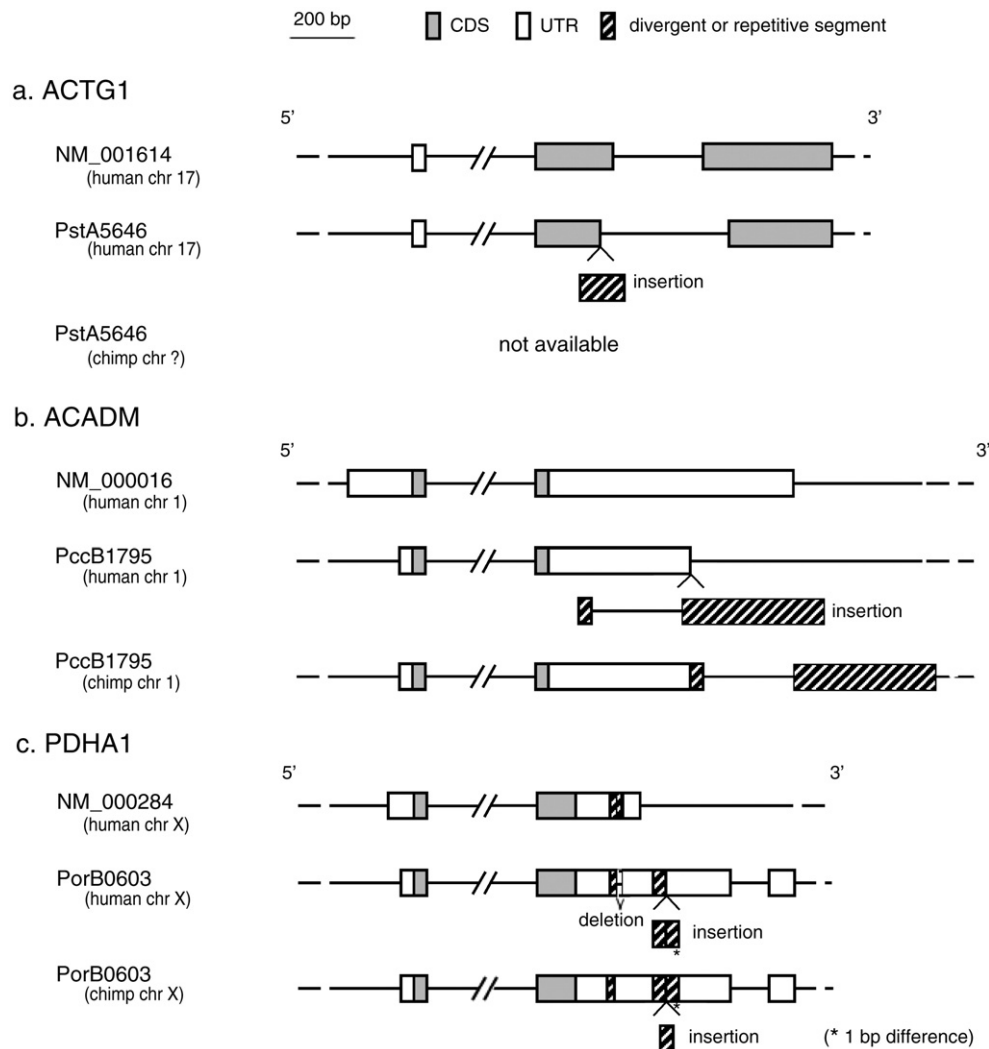


Fig. 5. Three examples of large sequence divergence and structural divergence. a. *ACTG1* (human NM_001614 and chimpanzee PstA5646) had the largest indel in the present study in the CDS. It was a 126-bp Pins/Hdel-type indel. b. *ACADM* (NM_000016 and PccB1795) had the largest indel in the present study in the 3'-UTR (external indel). It was a 559-bp Pins/Hdel-type indel containing *Alu* and it produced an additional exon at 3'-end for the chimpanzee transcript. c. *PDHA1* (NM_000284 and PorB0603) had two sets of indels in the 3'-UTR caused by insertion of repetitive segments. In one case, chimpanzee had one 10-bp copy unit while human had two. In the other case, the chimpanzee transcript had three 55-bp copy units while the human genome had one. One chimpanzee copy unit had 1-bp difference (substitution) compared to other units. There was also a difference in copy number between the chimpanzee transcript (three copy units) and the chimpanzee genome (two copy units). This genomic region seemed to be not transcribed in the human transcript.

found in the 3'-UTR of PorB0603 (NM_000284, *PDHA1*) (Fig. 5c). One consisted of a 10-bp copy unit ("TCAATGAAAT") and the other consisted of a 55-bp unit ("ACTCTAATTATGAAAAGGAAGAACAAATTCCTTG-TATGCCTGTTTCCCCTGCCCC"). In the former case, the number of copy units was different between the human genome (two) and the chimpanzee transcript/genome (one). This caused a 10-bp PdeI/HinfI-type indel. In the latter case, the indel was located in the extended chimpanzee 3'-UTR. The 3'-most copy unit in the chimpanzee transcript was different from the other two copy units. Its 50th nucleotide was "A" (adenine)-type, while the others' nucleotides were "G" (guanine)-type. The number of copy units was different between the human genome (one, "G"-type) and the chimpanzee transcript (three, two "G"-types and one "A"-type in this order). This caused a 110-bp PstI/HinfI-type indel. This result shows that repetitive segments can produce large sequence divergence (>100-bp) especially in the 3'-UTRs. However, polymorphism was also important considering repetitive segments. It was observed in the latter case of the repetitive segments. The numbers of copy units were also different between the chimpanzee transcript (three) and the chimpanzee genome (two, one "G"-type and one "A"-type in this order). Thus, this case involved not only divergence but also polymorphism.

4. Discussion

4.1. Impact of indel in the 3'-UTRs

We found that indels increased the amount of sequence divergence from 0.73% to 1.36% on average, and that they accounted for 30–60% of the sequence divergence among the 5'-UTRs, CDSs and 3'-UTRs (Fig. 2). We found that indels were found to occur differently among the 5'-UTR, CDS and 3'-UTRs, and that indels in the 3'-UTRs were prominent both in number and size (Fig. 3). In the 3'-UTRs, transposable elements and repetitive segments were found to cause large indels (>100 bp). These cases also influenced structural divergence by adding an exon at the TTS (3'-end).

We found a chimpanzee-specific *Alu* insertion (PccB1795, NM_000990, *ACADM*) (Fig. 5b). The frequency of *Alu* insertions into the chimpanzee genome has been reported to be 1.7 times as high as those into the human genome (Hedges et al., 2004). The *ACADM* gene might be an example of such a chimpanzee-specific *Alu* insertion that has occurred after the split of the human-chimpanzee lineages. We also found another example of an *Alu* insertion (PstA2462, NM_001424, *EMP2*). Although *Alu* was inserted into both the human and chimpanzee genomes, only the human transcript was terminated by the *Alu* insertion in the 3'-UTR, while the chimpanzee transcript seemed to have a longer 3'-UTR with an additional exon (non-canonical). These results imply that *Alu* insertions (exonization of *Alu*) had a large influence (>100 bp) on the transcriptome divergence, especially in the 3'-UTRs.

We also found indels caused by insertion of repetitive segments in the 3'-UTR of PorB0603 (NM_000284, *PDHA1*) (Fig. 5c). In this case, the TTSs were different between the two

species. We found that the indels caused by repetitive segments are related to polymorphisms for the *PDHA1* gene. The difference in the copy unit number, copy number variation (CNV), could be a source of large sequence divergence. However, it is difficult to distinguish the difference between inter-species and intra-species types of CNV; further data collection is needed.

As these *Alu* and repetitive segments added exons at the 3'-ends (Fig. 5b and c), nonsense-mediated decay (NMD) (Maquat, 2002) possibly plays a role. NMD recognizes a translation stop codon as a premature stop codon when it is located more than 50–55 bp upstream of the 3'-most exon–exon junction. NMD might occur for these transcripts (PccB1795, PorB0603 and PstA2462), because the relative positions of translation stop sites were moved upstream for these cases. These could be related to the difference in post-transcriptional regulation.

4.2. Divergent TSS and AS in the 5'-UTRs

The number of genes that differed in the TSS was significantly larger than the number of genes that differed in the TTS (Fig. 4). The transcription process could possibly start at different sites in human and chimpanzee, while it tends to stop at the same site in the two species. We also found that AS exons were observed most frequently in the 5'-UTRs. Both the different TSSs and the frequent AS occurrence in the 5'-UTRs might increase the amount of structural divergence in the 5'-UTRs. This resulted in the divergence in transcription initiation mechanism such as different transcription factor binding sites. Thus, they could cause divergence in transcription efficiency and lead to different gene functions between the species.

There have been reports suggesting a large variety of ASs and TSSs among the human transcripts. More than half of human genes undergo alternative splicing (Brett et al., 2002). On average, 2.7 variants per gene locus occur due to alternative splicing in human (Imanishi et al., 2004). Also, more than half of human RefSeq mRNAs have variable transcription start sites, and there are on average 7.7 different splicing types in the first exons in human (Kimura et al., 2006).

Although human RefSeq mRNAs are generally the most authoritative transcript set, there could be the cases where some transcript variants are not found in RefSeq. Even if human RefSeq transcripts exist in a gene locus, other human transcripts of DDBJ/EMBL/GenBank in the same locus could be more properly assigned as the chimpanzee orthologs. Thus, the absence of some transcript variants in RefSeq could result in the overestimation of structural divergence in the present study.

4.3. Different evolutionary constraints on the 5'-UTRs and 3'-UTRs

Using the chimpanzee FLcDNA sequences, we found that the patterns of divergence were different between the 5'-UTRs and 3'-UTRs. In terms of sequence divergence, indels were found most frequently in the 3'-UTRs (Fig. 3), while in terms of structural divergence, AS exons were found most frequently in the 5'-UTRs (Table 1) and TSSs (5'-ends) were found to be more variable than TTSs (3'-ends) between the species (Fig. 4).

In a comparative study between human and cynomolgus monkey (macaque), the number of substitutions per site in the 5'-UTRs was smaller than those in the synonymous sites (Osada et al., 2005). In the present study, we also found that the numbers of substitutions per site in both the 5'- and 3'-UTRs were smaller than those in the synonymous sites, comparing human and chimpanzee (Fig. 2). This shows that evolutionary constraint at the nucleotide level possibly affects both the 5'-UTRs and 3'-UTRs in the primate lineage.

A previous study has reported that a lack of sequence conservation in the transcription regulatory regions (upstream coding sequences and first introns) and an increased rate of gene expression divergence were found between human and chimpanzee (Keightley et al., 2005). Moreover, it was also reported that transcription start sites diverge between mouse transcripts and their orthologous human transcripts (Frith et al., 2006). From these studies, the mechanism of using different transcription start sites to increase diversity was implied in the mammalian lineage. Considering these divergences at the 5'-end, the structural divergence in the 5'-UTRs found in the present study, such as variable TSSs and frequent AS exons, could be a cause of gene expression divergence. This might be because different transcription-factor binding sites, which are located near TSSs, are selected.

The different functions could be derived not only from the structural divergence in the 5'-UTRs but also from the indels in the 3'-UTRs. Transposable elements (*Alu*) and repetitive segments in the 3'-UTRs were found to be the cause of large (>100 bp) sequence divergences. Large indels in the 3'-UTRs might change the translation efficiency of some genes, which is controlled by binding of miRNAs (Lee et al., 1993; He and Hannon, 2004). Different mechanisms of regulation might exist between the 5'-UTRs and 3'-UTRs.

4.4. Large potential divergence of the transcriptome

The amount of sequence divergence between the human and chimpanzee genomes was previously estimated to be as large as 5% when indels are taken into account (Britten, 2002). A chimpanzee whole-genome analysis reported that lineage-specific indels result in about 1.5% difference in the euchromatic sequences of the genomes of the two species (The Chimpanzee Sequencing and Analysis Consortium, 2005). These results show that indels have had a larger impact on the sequence divergence of the genome than substitutions have. Also, for the MHC class I regions between human and chimpanzee, it was reported that sequence divergence increased from 1.4% to 3.3% when indels were taken into account (Anzai et al., 2003).

In the present study, as an initial analysis using a set of chimpanzee FLcDNA sequences, we found that, when indels were taken into account, the sequence divergence between chimpanzee transcripts and the human genome was larger (1.36%) than the number of nucleotide substitutions per site (0.73%). This 1.36% is even larger than the whole-genome sequence divergence calculated by substitutions (1.22–1.23%) (Chen et al., 2001; Fujiyama et al., 2002). This implies that not

only intergenic regions but also genic regions on the genome are highly influenced by indels.

Consequently, the total amount of structural divergence was calculated to be >10% (11.6%). As human RefSeq sequences do not completely cover all human AS variants yet and tend to be longer transcript variants, this divergence could be an overestimate. However, this could still be a large potential divergence in terms of which AS variant was preferably selected in human and chimpanzee, respectively. Also, in what tissue and to what extent a transcript variant is preferably expressed might be important. We could provide the chimpanzee transcript variants expressed in brain or skin tissues. Comparative analysis at the transcript variant level would be essential for functional annotation. The variety of structural divergence could add greater flexibility for the transcriptome, without depending on the genome sequence.

The 87 genes analyzed in the present study were genes expressed in brain or skin tissues (Supplementary Table 1). Phenotypic differences of these tissues are supposed to be large between human and chimpanzee, especially in cognitive or morphological traits. It was reported that genes involved in the development of brain are more divergent between human and macaque than between mouse and rat (Dorus et al., 2004). It was also reported that, although genes expressed in brain were highly conserved between human and chimpanzee, more changes on the human than on the chimpanzee lineage have accumulated (Khaitovich et al., 2005). Phenotypic difference might be derived from slight but critical differences appearing in expression patterns and/or protein structures of specific genes. The specific genes could also be the genes containing large sequence/structural divergences in the 5'-UTRs and 3'-UTRs. Further experimental studies surveying the expressions and functions of the specific genes are expected to verify the effects of the large divergences of the transcriptome.

Acknowledgements

We thank Sumio Sugano and Yutaka Suzuki for the construction of full-length enriched cDNA libraries from chimpanzee tissues. Thanks also go to the members of Japan Biological Information Research Center for their support, discussion and advice. We are also grateful to Craig Gough and Phillip Hilton for critical reading of the manuscript. This work was supported in part by Health and Labor Sciences Research grant for Human Genome Research from Ministry of Health, Labor and Welfare of Japan, and by the Program for Promoting the Establishment of Strategic Research Centers, Special Coordination Funds for Promoting Science and Technology, Ministry of Education, Culture, Sports, Science and Technology of Japan, and by the Ministry of Economy, Trade and Industry of Japan.

Appendix A. Public databases used in the present study

NCBI RefSeq	http://www.ncbi.nih.gov/RefSeq/
UCSC	http://genome.ucsc.edu/
Rebase Update	http://www.girinst.org/Rebase_Update.html
DDJB	http://www.ddbj.nig.ac.jp/Welcome-e.html

Appendix B. Sequence submission

All the 87 chimpanzee FLcDNA sequences were submitted to DDBJ. Their accession numbers are **AB188273–AB188277**, **AB188280–AB188285**, **AB188287**, **AB188288** and **AB222102–AB222175**). The detailed information of the 87 FLcDNA is available at our website “PRIGEN” (<http://www.prigen.org/>).

Appendix C. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.gene.2007.04.013.

References

- Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410.
- Anzai, T., et al., 2003. Comparative sequencing of human and chimpanzee MHC class I regions unveils insertions/deletions as the major path to genomic divergence. *Proc. Natl. Acad. Sci. U. S. A.* 100, 7708–7713.
- Brett, D., Pospisil, H., Valcarcel, J., Reich, J., Bork, P., 2002. Alternative splicing and genome complexity. *Nat. Genet.* 30, 29–30.
- Britten, R.J., 2002. Divergence between samples of chimpanzee and human DNA sequences is 5%, counting indels. *Proc. Natl. Acad. Sci. U. S. A.* 99, 13633–13635.
- Chen, F.C., Vallender, E.J., Wang, H., Tzeng, C.S., Li, W.H., 2001. Genomic divergence between human and chimpanzee estimated from large-scale alignments of genomic sequences. *J. Heredity* 92, 481–489.
- Dorus, S., et al., 2004. Accelerated Evolution of Nervous System Genes in the Origin of *Homo sapiens*. *Cell* 119, 1027–1040.
- Frith, M.C., et al., 2006. Evolutionary turnover of mammalian transcription start sites. *Genome Res.* 16, 713–722.
- Fujiyama, A., et al., 2002. Construction and analysis of a human-chimpanzee comparative clone map. *Science* 295, 131–134.
- He, L., Hannon, G.J., 2004. MicroRNAs: small RNAs with a big role in gene regulation. *Nat. Rev., Genet.* 5, 522–531.
- Hedges, D.J., Callinan, P.A., Cordaux, R., Xing, J., Barnes, E., Batzer, M.A., 2004. Differential alu mobilization and polymorphism among the human and chimpanzee lineages. *Genome Res.* 14, 1068–1075.
- Hellmann, I., Zollner, S., Enard, W., Ebersberger, I., Nickel, B., Paabo, S., 2003. Selection on human genes as revealed by comparisons to chimpanzee cDNA. *Genome Res.* 13, 831–837.
- Hida, M., et al., 2000. Construction and preliminary characterization of full-length enriched cDNA libraries for nonhuman primates. *Primate Res.* 16, 95–110.
- Imanishi, T., et al., 2004. Integrative annotation of 21,037 human genes validated by full-length cDNA clones. *PLoS Biol.* 2, 856–875.
- International Human Genome Sequencing Consortium, 2004. Finishing the euchromatic sequence of the human genome. *Nature* 431, 931–945.
- Jurka, J., 2000. Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet.* 16, 418–420.
- Keightley, P.D., Lercher, M.J., Eyre-Walker, A., 2005. Evidence for widespread degradation of gene control regions in hominid genomes. *PLoS Biol.* 3, e42.
- Kent, W.J., 2002. BLAT — the BLAST-like alignment tool. *Genome Res.* 12, 656–664.
- Khaitovich, P., et al., 2005. Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. *Science* 309, 1850–1854.
- Kimura, K., et al., 2006. Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes. *Genome Res.* 16, 55–65.
- Kuroki, Y., et al., 2006. Comparative analysis of chimpanzee and human Y chromosomes unveils complex evolutionary pathway. *Nat. Genet.* 38, 158–167.
- Lee, R.C., Feinbaum, R.L., Ambros, V., 1993. The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell* 75, 843–854.
- Maquat, L.E., 2002. Nonsense-mediated mRNA decay. *Curr. Biol.* 12, R196–R197.
- Maruyama, K., Sugano, S., 1994. Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides. *Gene* 138, 171–174.
- Nei, M., Gojobori, T., 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* 3, 418–426.
- Okazaki, Y., et al., 2002. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* 420, 563–573.
- Olson, M.V., Varki, A., 2003. Sequencing the chimpanzee genome: insights into human evolution and disease. *Nat. Rev. Genet.* 4, 20–28.
- Osada, N., et al., 2005. Substitution rate and structural divergence of 5'UTR evolution: comparative analysis between human and cynomolgus monkey cDNAs. *Mol. Biol. Evol.* 22, 1976–1982.
- Ota, T., Nei, M., 1994. Variance and covariances of the numbers of synonymous and nonsynonymous substitutions per site. *Mol. Biol. Evol.* 11, 613–619.
- Rice, P., Longden, I., Bleasby, A., 2000. EMBOSS: the European molecular biology open software suite. *Trends Genet.* 16, 276–277.
- Saitou, N., 1991. Reconstruction of molecular phylogeny of extant hominoids from DNA sequence data. *Am. J. Phys. Anthropol.* 84, 75–85.
- Saitou, N., Ueda, S., 1994. Evolutionary rates of insertion and deletion in noncoding nucleotide sequences of primates. *Mol. Biol. Evol.* 11, 504–512.
- Sakate, R., et al., 2003. Analysis of 5'-end sequences of chimpanzee cDNAs. *Genome Res.* 13, 1022–1026.
- Sham, P.C., Curtis, D., 1995. Monte Carlo tests for associations between disease and alleles at highly polymorphic loci. *Ann. Hum. Genet.* 59 (Pt 1), 97–105.
- The Chimpanzee Sequencing and Analysis Consortium, 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437, 69–87.
- Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22, 4673–4680.
- Watanabe, H., et al., 2004. DNA sequence and comparative analysis of chimpanzee chromosome 22. *Nature* 429, 382–388.
- Yunis, J.J., Sawyer, J.R., Dunham, K., 1980. The striking resemblance of high-resolution G-banded chromosomes of man and chimpanzee. *Science* 208, 1145–1148.