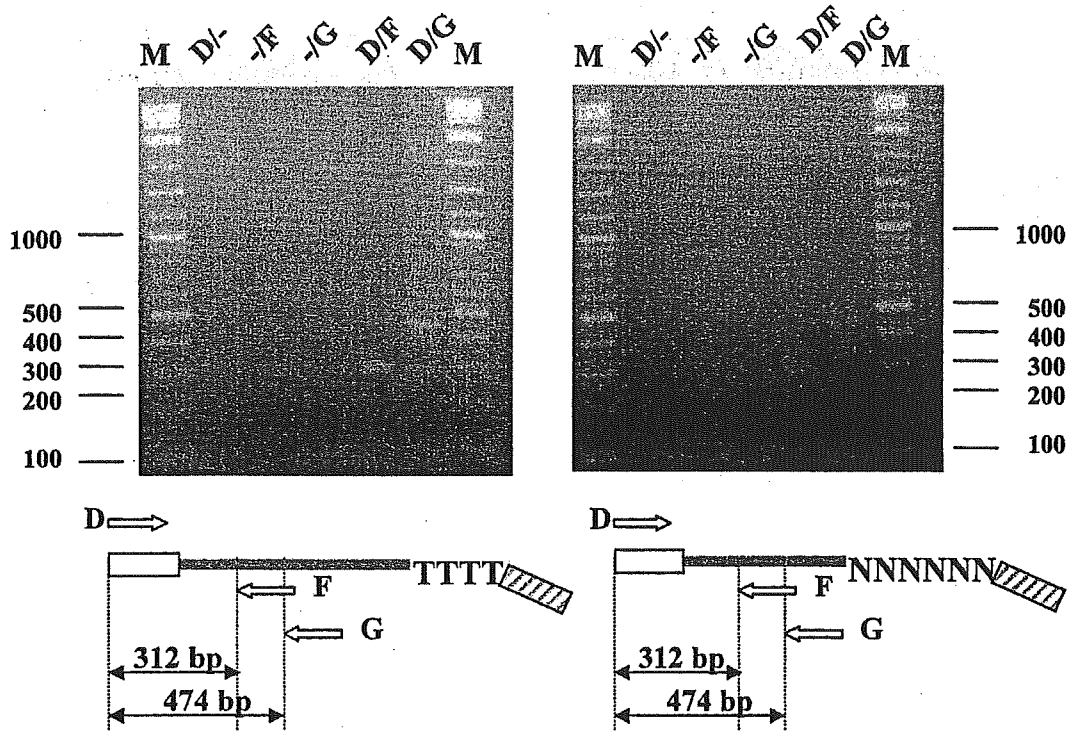

Fig. 2. Confirmation of the first-strand cDNA. The 5′ end of the EF1-α mRNA were amplified using a primer complementary to the cap-replacing oligo (D) and the EF1-α-specific primer (F or G). The first-strand cDNA was synthesized from the dT primer using "oligo-capped" RNA prepared from human intestinal mucosa tissue **(left)** or from the dR primer using "oligo-capped" RNA prepared from HEK293 cells **(right)**. The PCR products of the expected lengths (312 bp for primers B and H and 474 bp for primers B and I) were observed in both cases. The 5′ end of genes of interest could be amplified by a similar method. M: molecular weight marker, 2-Log DNA Ladder (cat. no. 3200S; New England Biolabs).

3.11. *Sfi*I Digestion of the PCR Products

1. Digest the PCR products by combining: 89 μL of the sample, 10 μL of NEBuffer 2,* 1 μL of 100X BSA,* and 2 μL of *Sfi*I.
2. Incubate at 50°C overnight.
3. Extract with phenol:chloroform (1:1) and ethanol precipitate (as described in **Subheading 3.2., steps 3–5**).

3.12. Size Fractionation of the Double-Stranded cDNAs

1. Electrophorese the *Sfi*I-digested PCR products in a 1% (w/v) agarose gel in TAE.
2. Excise the part of the gel containing the DNA fraction longer than 1.5 kb (*see* **Note 13** and **Fig. 3**).
3. Crush and completely dissolve the gel in 800 μL of QG* at RT. Incubation at 50°C and adding isopropanol (as instructed by the supplier) are not necessary.

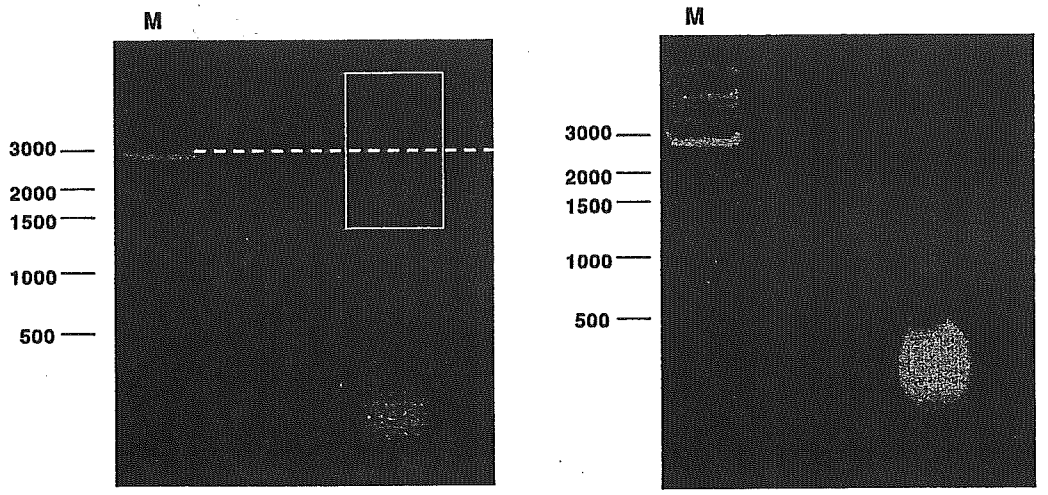

Fig. 3. Image of the smear of the PCR products after gel electrophoresis. The left panel shows the typical smeared profile of the amplified cDNAs. The open box indicates where the gel should be excised. The broken line represent the bottom line of the excision when enrichment of longer cDNAs is attempted (*see* **Note 13**). When the smear is observed only at the lower part of the gel (right), the PCR reaction has failed in amplifying the first strand cDNA. M: Molecular weight marker, 2-Log DNA Ladder (cat. no. 3200S, New England Biolabs).

4. Apply the solution to the column* and centrifuge briefly at RT. Repeat this step until the entire solution has been applied.
5. To wash the column, apply 500 µL of QG* and centrifuge briefly at RT. Discard the flowthrough.
6. To wash the column further more, apply 750 µL of PE* and centrifuge for 1 min at RT. Discard the flowthrough.
7. To elute the DNA, set a fresh tube to collect the eluate and apply 100 µL of dH_2O. Let it stand still for 1 min at RT.
8. Centrifuge for 1 min at RT and collect the elution.
9. Ethanol precipitate (as described in **Subheading 3.2., steps 4** and **5**).
10. Dissolve the sample in 9 µL of dH_2O.

3.13. Preparation of the Cloning Vector (see Note 14)

1. Digest the plasmid vector pME18S-FL3 with *Dra*III by combining 10 µg of pME18S-FL3 in 89 µL of dH_2O, 10 µL of NEBuffer 3,* 1 µL of 100X BSA,* and 2 µL of *Dra*III.
2. Incubate at 37°C overnight.
3. Extract with phenol:chloroform (1:1) and ethanol precipitate (as described in **Subheading 3.2., steps 3–5**).
4. Dissolve the sample in 89 µL of dH_2O and repeat the digestion (**Subheading 3.11., steps 1–3**) two more times, so that the residual uncut vector is completely digested.

5. Electrophorese the *Dra*III-digested vector in a 1% (w/v) agarose gel in TAE. Recover the 3.0-kb vector fragment (as described in **Subheading 3.12.**). Also, keep the 0.4-kb stuffer fragment to use as a mock insert.
6. Before using the prepared vector for the cloning of the PCR-amplified cDNA fragments, estimate the background level of undigested vector. Using the stuffer recovered in **step 5** instead of the cDNAs, follow the procedure described in **Subheading 3.12.** Compare the number of transformed bacterial colonies obtained using the "mock insert plus" and the "mock insert minus" preparations. Until the ratio of the numbers becomes more than 100:1, repeat the digestion and purification.

3.14. Cloning and Transformation of the Library (see Notes 15–18)

1. Using the vector prepared in **Subheading 3.13.**, ligate the PCR-amplified cDNA fragments to the vector by combining: 9 μL of the cDNA fragments, 1 μL of the prepared vector (10 ng/μL), 80 μL of Solution A,* and 10 μL of Solution B* (DNA ligation kit, TaKaRa).
2. Incubate at 16°C for 3 h.
3. Extract with phenol:chloroform (1:1) and ethanol precipitate (as described in **Subheading 3.2., steps 3–5**).
5. Dissolve the sample in 50 μL of dH_2O.
6. Use 1 μL of the library to transform competent *E. coli* cells TOP10 by electroporation according to the standard method.

3.15. Evaluation of the Quality of the Library

3.15.1. Checking the Distribution of the Insert Size

1. Randomly pick up the colonies and culture overnight in 185 μL of LB with ampicilin at 50 ng/mL in a 96-well microtiter plate (MTP) at 37°C.
2. Add 45 μL of 80% (v/v) glycerol and mix well. This solution can be stored at –80°C and used as a glycerol stock.
3. Set up the colony PCR by combining 1 drop of the glycerol stock in 4.93 μL of dH_2O with 1.0 μL of 10X ExTaq Buffer,* 0.9 μL of the four dNTPs at 2.5 m*M* each,* 1.56 μL of each colony PCR primer (sequences H and I), and 0.05 μL of ExTaq DNA polymerase.
4. Thermocycle for 30 cycles at 95°C, 15 s; 55°C, 15 s; 72°C, 4 min.
5. Electrophorese 2 μL of the PCR products in a 1% agarose gel in TAE.
6. Evaluate the length distribution of the inserts (*see* **Fig. 4A**).

3.15.2. Sequencing Analysis and Characterization of the Library

1. Set up the sequencing reaction by combining: 2.0 μL of the PCR products, 3.0 μL of dH_2O, 4.0 μL of BigDye buffer,* and 1.0 μL of the sequencing primer (sequence J or K).
2. Thermocycle for 25 cycles at 96°C, 10 s; 50°C, 5 s; 60°C, 4 min, using a 9700 ABI thermal cycler.

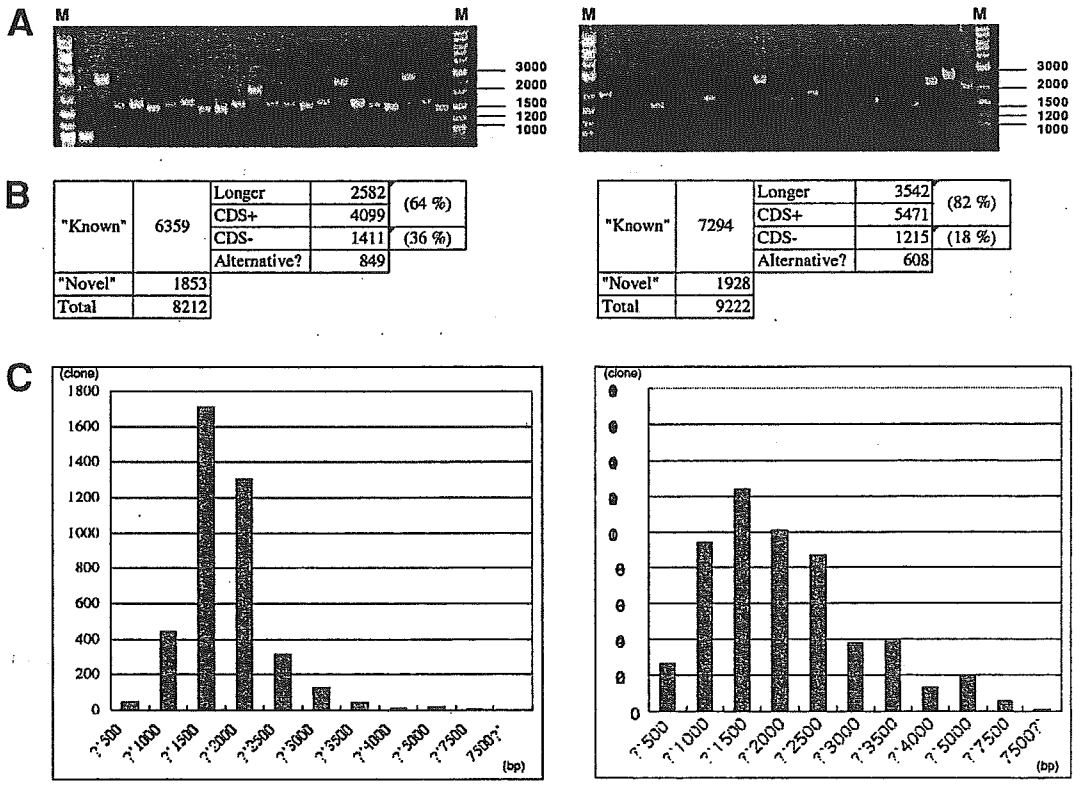

Fig. 4. Characterization and evaluation of a full-length enriched and a 5′-end enriched cDNA library. A full-length enriched and a 5′-end enriched cDNA library from human ileum (right panel) and HEK293 cells (left panel), respectively. (**A**) Results of the colony PCR; (**B**) results of the BLAST search against RefSeq; (**C**) length distribution of the isolated cDNAs of known genes. It is noteworthy that the population of long cDNAs (>5 kb), which is missing from the full-length library, is present in the 5′-end enriched cDNA library.

3. Determine the sequences using a 3700 ABI sequencer.
4. Base-call and trim the vector sequence. Using the processed sequence text data, perform the BLAST search against RefSeq, which is a database of reference human mRNA sequences maintained at NCBI (http://www.ncbi.nlm.nih.gov/LocusLink/refseq.html).
5. Evaluate the population of full-length cDNAs. When the 5′ end of a cDNA covers the translation initiator codon of the known gene according to the RefSeq data, the cDNA is tentatively categorized as "full length." Based on this criterion, the population of the "full-length" cDNAs in a library is usually 50–80% (*see* **Fig. 4B**).

3.15.3. Examples of a Full-Length Enriched and a 5′-End Enriched cDNA Library

A full-length enriched cDNA library and a 5′-end enriched cDNA library were constructed according to the above-described procedure. For the full-

length library, HEK293 cells cultured in twenty 10-cm dishes, each of which contained 1×10^6 cells (2×10^7 cells in total) were used. For the 5′-end library, 1 g of frozen human intestinal mucosa tissue was used. After the first-strand cDNA synthesis, one-fifth of the sample was PCR-amplified and cloned into the vector in both cases. **Figure 4** exemplifies the results of the insert size check by colony PCR and the evaluation of the population of full-length cDNAs by BLAST search. Assessment using the cDNAs matched known genes (**Subheading 3.15., steps 4** and **5**) revealed that the populations of full-length cDNAs constituted 64% and 82% of the full-length and the 5′ end cDNA library, respectively. The distribution of the mRNA size was calculated using the cDNAs of known genes and the reported mRNA lengths. It is noteworthy that mRNAs longer than 5.0 kb is included in the 5′-end library. For further details, *see* **ref. *4***).

3.16. Genomewide Analysis of the "Oligo-Capped" cDNA Libraries

3.16.1. Large-Scale Sequencing of the Library

The full-length cDNA library construction technology enabled us to obtain full-length cDNAs of human genes efficiently. We have constructed more than 150 kinds of full-length enriched and 5′-end enriched cDNA libraries from a wide variety of human tissues and cultured cells. Using these cDNA libraries, we have been performing large-scale one-pass sequencing. So far, we have collected 20,000 different putative full-length cDNAs corresponding to candidate novel genes and more than 20,000 of these cDNAs have been completely sequenced. The analyzed cDNA data and the progress of the project are published at our website (http://cdna.ims.u-tokyo.ac.jp/). The cDNA sequence data have also been deposited in public databases through DNA Data Bank of Japan (DDBJ).

3.16.2. Large-Scale Identification of the Transcriptional Start Sites and Adjacent Promoters

Although motifs important for understanding the transcriptional regulation of human genes are embedded in the promoter, the number of genes whose promoter structure has been determined so far is quite limited. According to the Eukaryotic Promoter Database (http://www.epd.isb-sib.ch/), which accumulates reported promoter sequences, only several hundred human promoters have been characterized (273 human genes in Release 62). In part this may be the result of the fact that the exact mRNA start sites have not been identified for most human genes. The conventional methods for identifying the mRNA start site, such as S1 mapping, primer extension, and 5′-RACE, are technically difficult and often lead to the inaccurate identification of the mRNA start sites.

In order to elucidate the genomewide features of transcriptional regulation, we have initiated a large-scale identification of promoters of human genes. For that purpose, the "oligo-capped" cDNA libraries are good resources, because the 5′ end of a full-length cDNA corresponds to the transcriptional start site (TSS) and, in many cases, the promoter overlaps or is just proximal to the TSS. We computationally aligned the 5′ ends of the full-length cDNAs onto the human draft genomic sequences and obtained the positional information about the TSSs on the genome. Adjacent sequences were retrieved and considered to be the promoters. The latest information on the genomewide identification of TSSs and promoters is available from our website (http://elmo.ims.u-tokyo.ac.jp/dbtss/). For further details, refer to **ref. *7***.

4. Notes

1. Since the "oligo-capping" procedure consists of multi-step enzymatic reactions with long incubation times, the utmost care should be taken for every manipulation to avoid RNase contamination and ensure that all the reagents are kept in an RNase-free condition.
    1. Wear disposable gloves and change them at each step of the procedure.
    2. Use sterile, disposable tubes, tips, and pipets.
    3. Bake glassware in which dH_2O and other reagents will be stored at 150°C for 4 h.
    4. Take special care not to touch inside the cap of the tube whenever it is opened or closed.
    5. Every manipulation should be carried out on ice.
    6. The pH of every reagent should be strictly adjusted.
2. We also avoid using an autoclave because the autoclave is used for sterilizing culture medium in many cases, and airborne particles highly contaminated with RNases could easily be absorbed into the reagents. In addition, RNases cannot be inactivated by prolonged autoclaving. Some procedures employ the DEPC treatment of the plastic and glassware before use. This aims to inactivate the RNase activity by alkylating the histidine residues in the active sites of RNaseA-type enzymes. However, we do not recommend DEPC treatment because it has not been fully assessed how much influence the residual DEPC would have on the enzymatic activities of BAP, TAP, and RNA ligase.
3. The starting RNA material must be of the highest quality obtainable. One of the most popular methods for extracting total RNA is the AGPC method. This is a convenient method, which is applicable for a wide variety of tissues. However, the total RNA isolated with AGPC method contains a lot of fragmented RNA and genomic DNA. RNeasy (Qiagen) contains a column to remove such undesirable fractions. Therefore, we combine these two kits for the isolation of total RNA. If cultured cells are used as an RNA source, the most highly recommended method is the NP-40 method *(8)*. Using this method, only the cytoplasmic RNA can be isolated.

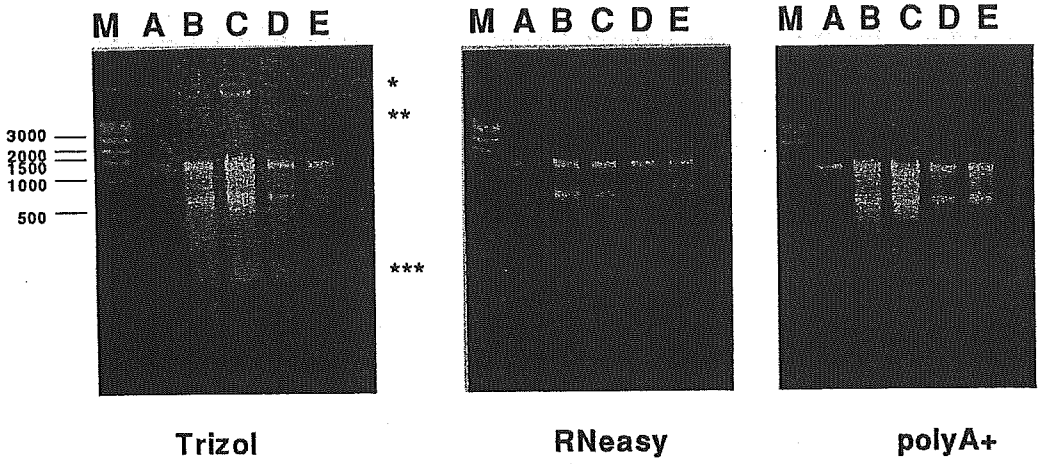

Fig. 5. Assessment of the RNA quality. The panels show gel electrophoretic profiles of the RNAs sampled after Trizol extraction *(left)*, RNeasy purification *(middle)*, and poly(A) selection *(right)*. The RNAs were isolated from human muscle, human ileum, monkey liver, monkey heart, and dog brain (lanes A–E, respectively). In some samples, contaminating genomic DNA (**) is observed. The slow-mobility fractions (*) might indicate that a certain amount of protein is included in the samples. It is noteworthy that these undesirable fractions and fragmented RNA (***) could be removed by RNeasy purification and poly(A) selection. M: Molecular weight marker, 1 kb DNA Ladder (cat. no. N3232S, New England Biolabs).

4. Be fully aware that a lot of genomic DNA is included near the borders of the layers. Residual genomic DNA could interfere with the oligo-capping reaction and/or could be cloned into the vector as an erroneous product. Take approximately two-thirds of the upper layer (or stop taking the solution at a position sufficiently distant from the border) and discard the rest.
5. Before applying the sample to the gel, heat-denature at 65°C for 2 min.
6. For poly(A) selection, many kits, which use latex or magnetic beads for the oligo-(dT) support (e.g., Oligo-Tex [cat. no. W9021B, Nippon-Roche, Tokyo, Japan] and FastTrak [cat. no. 1593-02; Invitrogen]) are commercially available. However, it is difficult to purify large quantities of poly(A)$^+$ RNA in high quality with these kits. We use oligo-(dT) cellulose powder so that we could adjust the bed volume and the washing conditions more flexibly according to the quality and quantity of the sample RNA.
7. In **Fig. 5**, RNA samples from each isolation and purification step are shown. For further details, refer to the figure legend. Now, more precise assessment of RNA quality is possible using Lab-Chip (Agilent 2100 bioanalyzer and RNA 6000 Nano Assay; cat. no. 5065-4476, Agilent Technologies, Waldbronn, Germany), which is a unit for performing the electrophoresis using a microcircuit. For further details, please refer to the instructions for this unit.

8. Drying the pellet is often hazardous to RNA, because RNase may get into the tube. Moreover, sometimes it becomes difficult to redissolve the pellet for the next reaction after the extensive drying.
9. At this step, the intermediate layer should be extremely sticky because of PEG 8000. Do not take this layer. Otherwise, it would be difficult to resuspend the sample in dH_2O for the next reaction.
10. Centrifugation at a higher *g*-force would destroy the column support resin and decrease the yield.
11. In order to avoid the misannealing of the oligo-(dT) primer, we omit the annealing step for the dT primer. However, when the dR primer is used, the incubation at 12°C is indispensable. In both cases, use a long extension time so that the reverse transcription can extend the cDNA maximally.
12. Use ammonium acetate at this step to remove fragmented RNA efficiently. Do not use the ammonium ion for ethanol precipitation until RNA ligation is completed, as ammonium ion interferes with T4 RNA ligase activity.
13. It is possible to excise the part of the gel containing the DNA fraction longer than 3 kb to construct a cDNA library in which a longer population of cDNAs of is enriched.
14. Regarding the vector plasmid pME18S-FL3, cDNA would be inserted downstream of the eukaryotic promoter, SRα, which is derived from the promoter of SV40 large T antigen. Thus, the full-length cDNA could be directly expressed if the vector was to be introduced into mammalian cells.
15. Usually the library size is 10^5–10^6 for 10–20 μg of starting poly(A)$^+$ RNA.
16. Use of PCR is a drawback for this procedure because it sometimes introduces a mutation into a cDNA. We estimate the frequency of such mutations as 1/2,000 bp for substitutions and 1/10,000 bp for insertion/deletion mutations. Also, PCR can cause a bias in the relative abundance of cDNAs because of differences in the efficiency of PCR for different cDNAs. Thus, information about the expression profile of mRNAs in cells may not be maintained in an "oligo-capped" cDNA library.
17. The restriction enzyme *Sfi*I, used for cDNA cloning, could cleave inside a cDNA, resulting in the loss of the cDNA from the library. However, *Sfi*I sites are expected to be rare in cDNAs because their recognition site consists of eight mers 5′GGCCNNNNNGGCC3′.
18. Other methods to construct a full-length enriched cDNA library using a cap-targeted selection step have also been reported by several groups *(9–11)*.

Acknowledgments

The "oligo-capping" method was originally developed in collaboration with K. Maruyama. We thank T. Ota, J. M. Sugano, and T. Isogai for helpful discussions and suggestions, M. Shirota, H. Hata, K. Nakagawa, K. Abe, T. Mizuno, M. Morinaga, M. Ishizawa, and M. Kawamura for their excellent sequencing work, and M. Hida and M. Sasaki for their technical support.

This work was supported by a Grant-in-Aid for Scientific Research on Priority Areas from the Ministry of Education, Science, Sports and Culture of Japan.

References

1. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., et al. (2001) Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921.
2. Venter, J. C., Adams, M. D., Myers, E. M., Li, P. W., Mural, R. J., Sutton, G. G., et al. (2001) The sequence of the human genome. *Science* **291,** 1304–1351.
3. Maruyama, K. and Sugano, S. (1994) Oligo-capping: a simple method to replace the cap structure of eucaryotic mRNAs with oligoribonucleotides. *Gene* **138,** 171–174.
4. Suzuki, Y., Yoshitomo, K., Maruyama, K., Suyama, A., and Sugano, S. (1997) Construction and characterization of a full length-enriched and a 5′-end-enriched cDNA library. *Gene* **200,** 149–156.
5. Suzuki, Y. and Sugano, S. (2001) Construction of full-length-enriched cDNA libraries. The oligo-capping method. *Methods Mol. Biol.* **175,** 143–153.
6. Shinshi, H., Miwa, M., Kato, K., Noguchi, M., Matushima, T., and Sugimura, T. (1976) A novel phosphodiesterase from cultured tobacco cells. *Biochemistry* **15,** 2185–2190.
7. Suzuki, Y., Yamashita, R., Nakai, K., and Sugano, S. (2002) DBTSS: DataBase of human Transcriptional Start Sites and full-length cDNAs. *Nucleic Acids Res.* **30,** 328–331.
8. Sambrook, J., Fritsch, E. F., and Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual*, 2nd ed., Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
9. Carninci, P., Kvam, C., Kitamura, A., Ohsumi, T., Okazaki, Y., Itoh, M., et al. (1996) High-efficiency full-length cDNA cloning by biotinylated CAP trapper. *Genomics* **37,** 327–336.
10. Kato, S., Sekine, S., Oh, S. W., Kim, N. S., Umezawa, Y., Abe, N., et al. (1994) Construction of a human full-length cDNA bank. *Gene* **150,** 243–250.
11. Edery, I., Chu, L. L., Sonenberg, N., and Pelletier, J. (1995) An efficient strategy to isolate full-length cDNAs based on an mRNA cap retention procedure (CAPture). *Mol. Cell. Biol.* **15,** 3363–3371.

Analysis of 5′-End Sequences of Chimpanzee cDNAs

Ryuichi Sakate,[1,5] Naoki Osada,[2] Munetomo Hida,[3] Sumio Sugano,[3] Ikuo Hayasaka,[4] Naoko Shimohira,[1] Shinsuke Yanagi,[1] Yumiko Suto,[1] Katsuyuki Hashimoto,[2] and Momoki Hirai[1]

[1]*Department of Integrated Biosciences, Graduate School of Frontier Sciences, The University of Tokyo, Chiba 277-8562, Japan;* [2]*Division of Genetic Resources, National Institute of Infectious Diseases, Tokyo 162-8640, Japan;* [3]*Department of Genome Structure Analysis, Institute of Medical Science, The University of Tokyo, Tokyo 108-8639, Japan;* [4]*Kumamoto Primate Research Park, Sanwa Kagaku Kenkyusho Co., Ltd., Kumamoto 869-3201, Japan*

We constructed full-length enriched cDNA libraries from chimpanzee brain, skin, and liver tissues by the oligo-capping method to establish a database of sequences of chimpanzee genes. Randomly selected clones from the libraries were subjected to one-pass sequencing from their 5′-ends. As a result, we collected 6813 chimpanzee cDNA sequences longer than 400 bp. Homology search against human mRNA sequences (RefSeq mRNAs) revealed that our collection included sequences of 1652 putative chimpanzee genes. In order to calculate the sequence identity between human and chimpanzee homologs, we constructed 5′-end consensus sequences of 226 chimpanzee genes by aligning at least three sequences for individual genes. Sequence identity was estimated by comparing these consensus sequences and the corresponding sequences of their human homologs. The average sequence identity of the 5′-end cDNAs was 99.30%. Those of the 5′-UTRs and CDSs were 98.79% and 99.42%, respectively. The results confirmed that human and chimpanzee genes are highly conserved at the nucleotide level. As for amino acids, the average sequence identity was 99.44%. The average synonymous (K_S) and nonsynonymous (K_A) divergences were estimated to be 1.33% and 0.28%, respectively.

[Supplemental material is available online at www.genome.org. All of the 1947 sequences used for constructing the consensus sequences of 226 chimpanzee genes have been submitted to DDBJ under accession nos. AU296732–AU298678. Two hundred twenty-six consensus sequences and their detailed annotation descriptions are available at our Web site http://www.prigen.org/.]

Since the draft sequence of the human genome was determined (International Human Genome Sequencing Consortium 2001; Venter et al. 2001), efforts have been under way to construct more comprehensive databases of human genes and their expression patterns. For better understanding of the biological characteristics of humans, a comparative analysis of chimpanzee genes with human genes will yield valuable information. The identity of genomic sequences between humans and chimpanzees was first estimated to be about 98.5% by the DNA–DNA hybridization method (Sibley and Ahlquist 1984, 1987). Recently, a comparative map has been constructed through paired alignment of genome-wide chimpanzee bacterial artificial chromosome end sequences with publicly available human genome sequences (Fujiyama et al. 2002). It revealed that the genomic sequence identity between humans and chimpanzees was 98.77%. Although human and chimpanzee genomes are highly identical at the nucleotide level, morphological traits and cognitive abilities are distinct between these two species. A comparative analysis of mRNA sequences may provide clues to the genetic information that affects the differing phenotypes. In contrast to the great number (about 16,000) of human mRNA sequence entries in public databases such as the RefSeq mRNAs of the National Center for Biotechnology Information (NCBI), only a small number of chimpanzee mRNA sequences and expressed sequence tags (ESTs) have been deposited in public databases. Moreover, the variety of chimpanzee genes in the databases is biased; they contain sequences derived mainly from mtDNAs and genes related to the major histocompatibility complex (MHC), which are known to evolve rapidly and are suitable for the analysis of phylogenetic relationships among closely related species. In this study, we attempted to analyze chimpanzee (*Pan troglodytes verus*) mRNA sequences using a substantial number of 5′-end enriched cDNA clones in order to establish a standard reference between the two species. We constructed cDNA libraries from the brain, skin, and liver tissues of two chimpanzees and sequenced the 5′-ends of 6813 clones. As a result, we were able to annotate the consensus sequences of 226 putative chimpanzee genes by comparing with the sequences of human homologs in public databases.

[5]Corresponding author.
EMAIL sakate@prigen.org; FAX 81-4-7136-3687.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.783103.

RESULTS

We collected 7064 5′-end sequences of chimpanzee cDNAs. These 5′-expressed sequence tags (5′-ESTs) were annotated by the BLAST program (Altschul et al. 1990). Consequently, 163 mitochondrial, 71 repetitive, and 17 vector sequences that were included in our raw sequence data were eliminated. The

13:1022–1026 ©2003 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/03 $5.00; www.genome.org

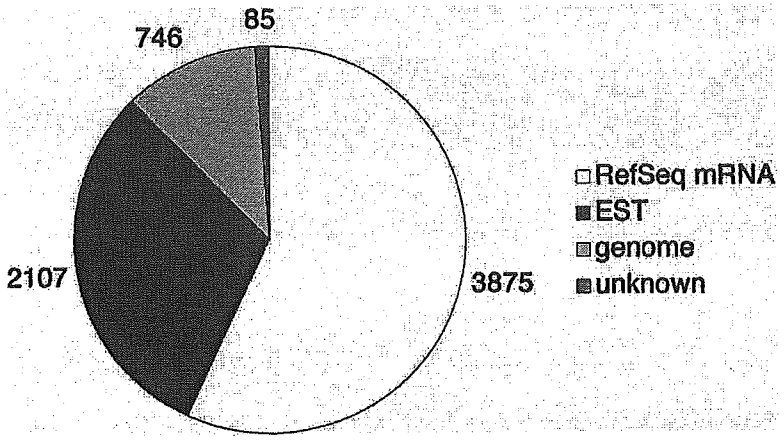

Figure 1 Constitution of 6813 chimpanzee cDNA sequences. Numbers of sequences that matched those of human RefSeq mRNA, EST, and genome sequences are indicated in this figure. "Unknown sequences" denote the sequences that did not match any sequences in public databases.

remaining 6813 ESTs consisted of 3875 sequences that matched human RefSeq mRNAs, 2107 that matched human ESTs, and 746 that matched human genome sequences (Fig. 1). The rest (85 sequences) did not match any sequences in the public databases. Of those that matched human RefSeq mRNAs, 2835 (73.2%) contained the translation start site. These 3875 sequences were clustered into 1652 nonredundant chimpanzee genes. It is worth noting that each of the 1537 sequences (93.0%) were found to correspond to only one human RefSeq mRNA by the BLAST search at a threshold value of $E = 1e^{-120}$, suggesting that these genes are orthologous to the corresponding human genes.

From these 1652 sequences, we constructed a total of 226 consensus sequences of 5′-end cDNAs. As described in the methods, each consensus sequence was established by aligning at least three sequences. The average length of the consensus sequences was 399.9 bp. The molecular functions of the 226 genes were annotated according to the classification system by the Molecular Function of GO categories (http://www.geneontology.org/). The distribution of the functions is shown in the Supplementary Figure 1. These 226 consensus sequences were aligned with 5′-untranslated regions (5′-UTRs) and/or coding sequences (CDSs) of human RefSeq mRNAs. We calculated sequence identities between 226 chimpanzee consensus sequences and the corresponding human RefSeq mRNA sequences (Suppl. Table 1). Among these 226 5′-end cDNA sequences, three consisted solely of 5′-UTR and 29 consisted solely of CDS. One hundred ninety-four sequences contained both 5′-UTRs and CDSs. When we calculated the average sequence identity, we excluded 28 5′-UTR regions and one CDS region of the consensus sequences because they were not sufficiently long enough to calculate reliable values. The distribution of the sequence identities (%) of the 5′-end consensus sequences of cDNAs and their CDS regions are shown separately in Figure 2. Two sequences with an exceptionally low identity were those of the polymorphic MHC-related genes. The remaining genes showed sequence identities greater than 97.0%.

The average sequence identity (%) with standard deviation of the 5′-UTR in 169 genes was 98.79% ± 1.71%. That of the CDS in 222 genes was 99.42% ± 0.62%, and that of the amino acids was 99.44% ± 1.20%. As for all the 5′-end regions of 226 genes, the average sequence identity was 99.30% ± 0.62%. The average K_S and K_A values (%) based on the data of 222 genes were determined to be 1.62% ± 1.75% and 0.26% ± 0.55%, respectively, by the method of Miyata-Yasunaga (1980), and 1.33% ± 1.54% and 0.28% ± 0.59%, respectively, by the method of Li (1993).

When all the bases for 5′-UTRs (197 genes, 15,665 bp) and CDSs (223 genes, 72,708 bp) were combined, their sequence identities were 98.79% and 99.42%, respectively. As for amino acids (24,011 a. a.), the average identity was 99.44%. As for all the 5′-end regions of 226 genes (90,381 bp), the sequence identity was 99.31%. When these combined data were used, the K_S and K_A values were respectively found to be 1.65% and 0.26%, by the method of Miyata and Yasunaga (1980), and 1.35% and 0.27%, by the method of Li (1993). The K_S values obtained by the two methods were different, while the K_A values were the same irrespective of the method used.

DISCUSSION

When analyzing sequence data of cDNAs, limitations in accuracy must be considered. mRNAs are fragile by nature and the final sequencer outputs are apt to include sequence errors generated during the cloning and sequencing processes. Therefore, sequence data derived from only one cDNA clone is not sufficient to obtain reliable information on genes. In this study, we aimed to precisely calculate base-by-base sequence identities. From our collection of 6813 cDNA 5′-end sequences, we were able to construct the consensus sequences of 226 chimpanzee genes based on at least three sequences for each gene. This is probably the first report concerning the comparative sequence analysis between humans and chimpanzees using such a substantial number of genes. In previous studies using 20 GenBank cDNAs (Varki 2000), the sequence identity (%) between human and chimpanzee cDNAs was 99.31% ± 0.38% (mean ± S.D.), and that of amino acid was 99.36% ± 0.66%. These values are not different from those obtained in our study (99.30% ± 0.62% and 99.44% ± 1.20%, respectively). Therefore, the sequence identity in the coding regions between humans and chimpanzees is higher, as expected, than that of the genome sequences (98.77%) reported by Fujiyama et al. (2002). It should be emphasized that we collected the 5′-UTR sequences of mRNAs. Thus far, there is no sufficient information on the 5′-UTR region based on which the substitution rate can be calculated. The average sequence divergence of 5′-UTRs between humans and chimpanzees was found to be 1.21%. This value is the same as those of the genomic sequence differences reported previously (Fujiyama et al. 2002, 1.23%; Chen and Li 2001, 1.22%).

As seen in Figure 2, two genes showed a low sequence identity (96.4%). These were MHC-related genes (PC_061

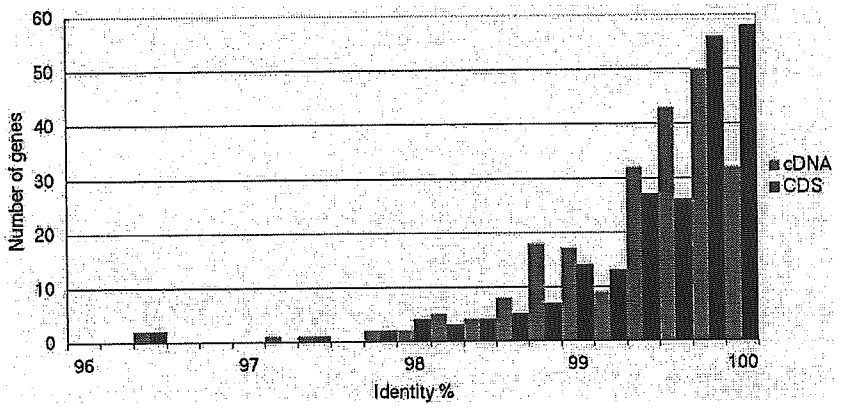

Figure 2 Distribution of sequence identities of 5'-end consensus sequences of chimpanzee cDNAs (226 cDNAs and 222 CDS regions).

[HLA-A homolog] and PC_133 [HLA-B homolog]) as listed in the Supplementary Table 1. The results suggest that these MHC-related genes evolved rapidly. Another MHC-related gene in our collection (PC_134, HLA-E homolog) showed a higher identity (98.2%) with the human homolog than that of the HLA-A and -B genes, and seemed to be relatively conserved. This is consistent with a previous study (Adams and Parham 2001), in which African apes were shown to have orthologs of all human class I MHC-related genes, and HLA-A, B, and C genes were suggested to be highly polymorphic while others (HLA-E, F, and G) conserved.

The synonymous (K_S) and nonsynonymous (K_A) divergences obtained in this study by the method of Miyata and Yasunaga (1980) were 1.62% and 0.26%, respectively. When using the method of Li (1993), the values were 1.33% and 0.28%, respectively. The K_S values calculated by the two methods are different. Li (1993) suggested that the method of Miyata and Yasunaga tends to overestimate the K_S value. Therefore, we considered the values calculated by the method of Li (1993) in this study. In a previous study (Chen et al. 2001), which analyzed 88 GenBank chimpanzee cDNAs, the K_S and K_A values were calculated to be 1.48% and 0.55%, respectively, by the method of Li (1993). These values are larger than our K_S and K_A values obtained by the same method. Distribution of the K_S and K_A values of 88 gene set (Chen et al. 2001) and that of our 226 gene set are shown in the Supplementary Figure 2. In these two gene sets, 15 out of 88 genes (15.9%) and 24 out of 222 genes (10.8%) had a value of $K_A / K_S \geq 1$ (Suppl. Table 1). This implies that the sequence data of 88 GenBank cDNAs include rapidly evolving genes such as immune-related genes and duplicated gene. Our calculation is based on 226 randomly selected genes and may represent a genome-wide average substitution rate of expressed sequences (functional distribution of 226 genes is shown in the Suppl. Fig. 1).

The objective of this study was to precisely calculate sequence diversity between human and chimpanzee homologs. While aligning chimpanzee sequences with those of human RefSeq mRNAs, we noted some problems that need to be clarified. Since studies that address these problems are currently underway, we just briefly report preliminary results.

1. As for the 85 sequences that did not match any human sequences in public databases, they may include those that may match ever-increasing genome and EST sequence databases, those that may be putative chimpanzee-specific transcripts, and those of artifacts. Thus far, we selected ten sequences out of the 85 sequences to confirm the presence of transcripts corresponding to the sequences in human and chimpanzee lymphoblastoid cells by RT-PCR using two sets of primers for each sequence. Primers were designed on the basis of chimpanzee sequences. As a result, three sequences (PorA1155, PstA6283, and PstA7892) were transcribed in both humans and chimpanzees, and one sequence (PccB3689) transcribed only in chimpanzees. The former could be unknown genes and the latter could be a chimpanzee specific transcript, though we did not find any protein domain in the four sequences by using NCBI CD-Search (http://www.ncbi.nlm.nih.gov:80/Structure/cdd/cdd.shtml).
2. We found chimpanzee sequences with ten or more nucleotides at the 5'- or the 3'-end that are completely different from corresponding sequences in human RefSeq. We excluded these inconsistent sequences by computational process and visual inspection because these could affect the results of identity calculation. Thus far, we compared these sequences with the human genome sequences, and found that several sequence inconsistencies could be the product of alternative splicing (Suppl. Table 2). Since a high proportion of human genes have been suggested to undergo alternative splicing (Brett et al. 2002), it is interesting to analyze the possible interspecific alternative splicing affecting gene functions. Recently, Britten has claimed that sequence diversity between human and chimpanzee genomes is as high as 5% when insertions/deletions (indels) are taken into account (Britten 2002). We detected indels in both the 5'-UTR and CDS regions of 226 consensus sequences comparing with human RefSeq mRNAs. In Supplementary Table 3, we listed indels which were confirmed by comparing with corresponding human genome sequences.
3. Recent studies have shown a high frequency of single-nucleotide polymorphisms (SNPs) in the human genome (one SNP per 1.08 kb; The International SNP Map Working Group 2001). Considering threefold to fourfold sequence diversity in chimpanzees compared with that in humans (Kaessmann et al. 1999, 2001), SNPs in chimpanzees are expected to occur frequently. Actually, we found that 16 consensus sequences contained putative SNPs (e.g., PC_121: 210th nucleotide is three As vs. three Gs). Each allele was determined when at least two clones contained the same nucleotide at a polymorphic base position. In total, three SNPs were found in the 5'-UTRs and 14 SNPs in the CDSs (PC_061, HLA-A homolog, containing three SNPs in the CDS region and PC_133, HLA-B homolog, containing 23 SNPs in the CDS region were excluded). The 17 SNPs in the chimpanzee sequences were not found in the human database (dbSNP). Among the 14 SNPs in the CDSs, seven (seven at the third codon) were synonymous and

seven (three at the second codon and four at the first codon) were nonsynonymous (Suppl. Table 4).

For the structural and functional analysis of genes, collection of data on alternative splicing, indels, and SNPs is important. In addition, a comparative analysis of tissue-specific expressions of genes between humans and chimpanzees is expected to shed light on species specificity and evolution of humans. Our collection of full-length cDNA clones and sequence data could be a valuable resource for postgenomic research.

METHODS

Construction of cDNA Libraries and Annotation of cDNA Sequences

Tissue specimens were collected from adult chimpanzees (*Pan troglodytes verus*) kept in the Primate Research Park, Kumamoto, Japan. All the procedures of tissue collection were approved by an institutional board. About 10 g of skin tissues was collected by biopsy from a male chimpanzee. Liver and brain tissues were obtained at autopsy from a female chimpanzee that died of septicemia. The full-length enriched cDNA libraries were constructed by the oligo-capping method (Maruyama and Sugano 1994; Hida et al. 2000). From these libraries, clones were randomly selected and their sequences were determined from the 5′-end by one-pass sequencing using ABI-377 and ABI-3100 sequencers. After eliminating the 5′-end vector sequences and undecided 3′-end sequences using our in-house program, only sequences longer than 400 bp were used for further analysis. Sequence base-calling was performed by the basecaller program attached to the ABI sequencers. The sequence data were subjected to the computer-based homology search against those of vector pME18S-FL3 and those in public databases (human RefSeq mRNAs [including mitochondria] and human ESTs of NCBI [http://www.ncbi.nlm.nih.gov/]; human repetitive sequences of REPBASE [http://www.girinst.org/Repbase_Update.html; Jurka 2000]; and human genome of the University of California Santa Cruz [UCSC, http://genome.ucsc.edu/, April 2001 freeze]) using the BLAST program (Altschul et al. 1990). The threshold values (BLAST expectation values) used to execute the BLAST program for the vector sequences, human RefSeq mRNAs, human mitochondrial sequences, human repetitive sequences, human ESTs, and human genome sequences were $1e,^{-10}$ $1e,^{-120}$ $1e,^{-60}$ $1e,^{-60}$ $1e,^{-60}$ and $1e,^{-30}$ respectively.

Construction of Consensus Sequence and Calculation of Sequence Identity

The chimpanzee sequences that matched the sequences of human RefSeq mRNAs were defined as the sequences of putative chimpanzee genes. Consensus sequences were established from at least three sequences of individual genes. Multiple sequences clustered to each gene were aligned together using the CLUSTAL W program (Thompson et al. 1994) and the output was computationally and visually inspected to remove alignment errors. As a result, we collected 226 consensus sequences from 1947 5′-end cDNA sequences. The average length of the 1947 sequences was 519.6 bp, and the average percentage of undecided bases, denoted as N, was 0.74%. The refined alignment data were processed using our original in-house program that determines individual bases by majority at each nucleotide site. The resulting consensus sequences were aligned with the sequences of the corresponding human RefSeq mRNAs. Then, a base-by-base comparison was conducted to calculate sequence identity of the 5′-end cDNA sequences using our original in-house program. 5′-UTR and CDS included in the 5′-end cDNA sequences were identified and analyzed. 3′-UTRs were omitted from our analysis because of insufficient data set. When the program was executed, gaps and Ns were excluded from the calculation, and alignment mismatches were eliminated by visual inspection. The identity of each amino acid was calculated for sequences spanning from the 5′-end to the 3′-end until an erroneous (not in-frame) gap appeared. The rates of synonymous (K_S) and nonsynonymous (K_A) substitutions were calculated using two previously reported methods (Miyata and Yasunaga 1980; Li 1993).

ACKNOWLEDGMENTS

This study was supported by the health science research grant for the human genome program from the Ministry of Health and Welfare of Japan, and by grant Nos. 13202015 and 13554035 from the Ministry of Education, Culture, Sports, Science, and Technology of Japan. We thank Eiichi Sato and Miho Aruga for their assistance.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Adams, E.J. and Parham, P. 2001. Species-specific evolution of MHC class I genes in the higher primates. *Immunol. Rev.* **20:** 41–64.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.

Brett, D., Pospisil, H., Valcarcel, J., Reich, J., and 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.* **99:** 13633–13635.

Chen, F.-C. and Li, W.-H. 2001. Genomic divergence between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am. J. Hum. Genet.* **68:** 444–456.

Chen, F.-C., Vallender, E.J., Wang, H., Tzeng, C.S., and Li, W.-H. 2001. Genomic divergence between human and chimpanzee estimated from large-scale alignments of genomic sequences. *J. Hered.* **92:** 481–489.

Fujiyama, A., Watanabe, H., Toyoda, A., Taylor, T.D., Itoh, T., Tsai, S.-F., Park, H.-S., Yaspo, M.-L., Lehrach, H., Chen, Z,. et al. 2002. Construction and analysis of a human–chimpanzee comparative clone map. *Science* **295:** 131–134.

Hida, M., Suzuki, Y., Sugano, S., Hashimoto, K., Terao, K., Hayasaka, I., and Hirai, M. 2000. Construction and preliminary characterization of full-length enriched cDNA libraries for nonhuman primates. *Primate Res.* **16:** 95–110.

International Human Genome Sequencing Consortium 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.

The International SNP Map Working Group 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409:** 928–933.

Jurka, J. 2000. Repbase update: A database and an electronic journal of repetitive elements. *Trends Genet.* **16:** 418–420.

Kaessmann, H., Wiebe, V., and Paabo, S. 1999. Extensive nuclear DNA sequence diversity among chimpanzees. *Science* **286:** 1159–1162.

Kaessmann, H., Wiebe, V., Weiss, G., and Paabo, S. 2001. Great ape DNA sequences reveal a reduced diversity and an expansion in humans. *Nat. Genet.* **27:** 155–156.

Li, W.-H. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J. Mol. Evol.* **36:** 96–99.

Maruyama, K. and Sugano, S. 1994. Oligo-capping: A simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides. *Gene* **138:** 171–174.

Miyata, T. and Yasunaga, T. 1980. Molecular evolution of mRNA: A method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. *J. Mol. Evol.* **16:** 23–36.
Sibley, C.G. and Ahlquist, J.E. 1984. The phylogeny of the hominoid primates, as indicated by DNA–DNA hybridization. *J. Mol. Evol.* **20:** 2–15.
Sibley, C.G. and Ahlquist, J.E. 1987. DNA hybridization evidence of hominoid phylogeny: results from an expanded data set. *J. Mol. Evol.* **26:** 99–121.
Thompson, J.D., Higgins, D.G., and 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.
Varki, A. 2000. A chimpanzee genome project is a biomedical imperative. *Genome Res.* **10:** 1065–1070.
Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. 2001. The sequence of the human genome. *Science* **291:** 1304–1351.

WEB SITE REFERENCES

http://www.ncbi.nlm.nih.gov/; NCBI.
http://www.girinst.org/Repbase_Update.html; REPBASE UPDATE.
http://genome.ucsc.edu/; UCSC Genome Bioinformatics.
http://www.geneontology.org/; Gene Ontology (GO) Consortium.
http://www.ncbi.nlm.nih.gov:80/Structure/cdd/cdd.shtml; A conserved Domain Database and Search Service, v1.60.

Received September 10, 2002; accepted in revised form March 5, 2003.

総　説

ヒトとチンパンジー：遺伝子構成の異同を探る

In search of genetic differences between humans and chimpanzees

坂手龍一　  　数藤由美子　  　平井百樹　 

ヒトとチンパンジーの遺伝子を比較解析しその差異を調べることで、ヒトの生物学的特徴を生じさせた進化の道筋の推定や、疾患遺伝子の解析に重要な情報が得られると考えられる。筆者らはチンパンジーの完全長cDNAライブラリーを作製し、精度の高い200種以上の遺伝子配列を得て、ヒトの相同遺伝子配列との比較解析を行った。配列データをさらに充実させ、ヒトとチンパンジーとで異なる遺伝子を発見することを目指している。

キーワード：チンパンジー、完全長cDNA、遺伝子解析

はじめに

ヒトゲノム配列の完全解読は2003年4月に終了した。ポストゲノム研究では、新規遺伝子の探索と並行して、遺伝子の機能および相互作用の解析が急ピッチで進んでいる。ヒト遺伝子の機能解析にあたっては、遺伝子配列の種間比較が有効な方法のひとつと考えられる。しかし、系統的に離れたマウスでは遺伝子配列の対応づけが難しく、比較解析に不適当な場合がある。そこで、類人猿のオーソロガス遺伝子(共通祖先遺伝子から進化した相同な遺伝子)の配列の解析に期待が寄せられている。2001年に、米国NHGRI(National Human Genome Research Institute)は、ヒトゲノム解読後の次の対象となる重点生物種選定に関する指針を白書として公表した。それには、ヒトの疾患遺伝子解析への貢献とヒトの進化や多様性の生物学的理解への貢献が条件として挙げられていた。これを受けてなされた申請のうち、優先順が高い生物種のひとつとして、チンパンジーが選ばれている。チンパンジーについては2001年末までに全ゲノムの1％ほどの配列が解読され、ヒトとの比較の結果、配列の違いは1.23％であることが報告された[1]。2003年7月にはすでにチンパンジー・ゲノム概要は完成したとの報告があり、10月7日にはヒト21番に対応するチンパンジー22番染色体の全塩基配列の解読が終了し、公開された。

通常、塩基配列の比較解析には、置換率が指標として用いられる。しかし、ゲノム中に存在する数塩基から数百塩基にわたる挿入・欠失(いわゆるindel)を考慮すると、両種間には実は4～5％の差があるというのが最近の知見である。このゲノム上の差異は、1億塩基前後の差に相当する。しかし、そのうちどれが機能的に重要で種差に貢献しているかを探索するのは容易ではない。筆者らはここ数年来、ポストゲノム研究として、効率よい遺伝情報集めのための研究、すなわち発現遺伝子配列の解析に取り組んできた。本稿では、

筆者紹介：さかて・りゅういち(SAKATE, Ryuichi)　東京大学大学院新領域創成科学研究科(Dept. of Integr. Biosci., The Univ. of Tokyo)先端生命科学専攻 学術研究支援員　2003年東京大学大学院理学系研究科博士課程修了　博士(理学)　専門：人類進化遺伝学　連絡先：〒277-8562 千葉県柏市柏の葉5-1-5　E-mail　sakate@prigen.org(勤務先)
すうとう・ゆみこ(SUTO, Yumiko)　同上 助手　1995年東京大学大学院理学研究科博士課程修了　博士(理学)　専門：人類細胞遺伝学　連絡先：同上　E-mail　suto@k.u-tokyo.ac.jp(勤務先)
ひらい・ももき(HIRAI, Momoki)　同上 教授　1971年東京大学大学院理学系研究科博士課程中途退学　理学博士　専門：人類細胞遺伝学　連絡先：同上　E-mail　mhirai@k.u-tokyo.ac.jp(勤務先)

チンパンジーの遺伝子構成に関する研究の現状と、筆者らのグループが進めている研究について概説する。

1．ヒトとチンパンジーのDNA配列の比較

チンパンジーとヒトのDNAの種間比較を行った研究としては、1980年代半ばにDNA-DNAハイブリダイゼーションにより、98.5%という高い一致度を持つことが報告されている[2]。1990年代半ばにはミトコンドリアDNA[3]、MHC関連遺伝子や偽遺伝子など進化の早い遺伝子を中心に配列が読まれ、霊長類の遺伝子進化について調べられるようになった。近年の大量のDNA配列の読み取りの効率化により、GenBankなどの公共データベースに登録される配列データは指数関数的に増大した。コンピューターを利用した解析技術の発達により、ヒトやマウスのゲノムが解読され、チンパンジーゲノムプロジェクトについても進行中である。

ヒトとチンパンジーの遺伝的関係は「見方によっては近く、見方によっては遠い」というのが、筆者らの当初からの命題である。巨視的にゲノムを染色体レベルで見ると、確かにあらゆる生物の中で最も核型は近似している。しかし、少なくとも10カ所の再配列部位が分かっており、さらに大規模にゲノム重複や欠失が生じていると考えられる領域が次第に明らかになってきている。そのうちのどれだけが種差に関与し、ヒト特異性に貢献しているかを探ることが重要な課題である。

ゲノム配列の比較が意味を持つには、比較する種のゲノムについても同程度の精度を持つ全配列を得る必要がある。進化の過程で生じた遺伝子の欠失と、遺伝子重複に伴う機能の多様化が種の分化に重要であったと考えられるが、これらを明らかにするには、現在の不完全なチンパンジーのゲノム配列データでは、ヒトの完全な配列に最も一致するように並列させようとしても見失われる情報が多い。例えば、ヒトのいわゆるeuchromatic領域の5%は、90%以上のホモロジーを示す1 kb以上の重複セグメントにより構成されるという。このような機能多様化に関わる領域に関する情報を見失うおそれがある。その意味で、このチンパンジーゲノムプロジェクトは全配列を高精度に読まねばならず、得られる情報が大きいかわりに、かかるエネルギーも大きい。

ヒトとチンパンジーのゲノムの塩基配列のうち、形態学的、生理学的特徴として大きな違いとなって種を分けているのは、各々の種で実際に発現している遺伝子であり、ゲノムの全体の配列の数%にすぎない。約3万2000と推測されるヒト遺伝子のうち、現時点で配列情報としてカタログ化されているのは約2万(NCBI RefSeq)である。遺伝子解析には二つのアプローチがある。第一が、前述のように全ゲノム配列を明らかにし、その中の機能遺伝子を探索して解析する方法である。第二が、実際に機能する遺伝子そのものの解析である。ヒトの多くの遺伝子でスプライシング変異の存在が報告されている[4]こともあり、ゲノム配列だけからエキソン―イントロン構造を予測することは困難である。そこで、転写・発現しているRNAの配列を逆転写によって保存できるcDNAライブラリーを作製し、その配列をシークェンシングによって得ることは、遺伝子配列の蓄積に非常に有効な手法である。

2．cDNAライブラリーの作製とシークェンシング

筆者らのグループでは、厚生労働省研究プロジェクトの一環として、国立感染症研究所遺伝子資源室の橋本雄之室長を主任研究者とする霊長類遺伝子情報の収集と解析を行っている。チンパンジーの遺伝子配列(cDNA配列)については、東京大学医科学研究所・菅野研究室ならびに(株)三和化学研究所熊本霊長類パークと協力し、自然死したチンパンジー(*Pan troglodytes verus*)2個体について、死後ただちに脳や皮膚組織を得て完全長cDNAライブラリーを作製した。これらはオリゴキャッピング法[5]によって作製され、完全長cDNAクローンに富んでいる。このcDNAライブラリーからランダムに単離したクローンについて、5′端からのシークェンシングを行い、チンパンジー遺伝子配列の上流部分について、約1万4000配列を得た。さらに5′端(上流)から3′端(下流)までの全長配列についても、これまでに100以上の配列を得ている(これらの配列のランダム・サンプリングによる品質チェックをphred(http://www.phrap.org/)プログラムを用いて行った結果、QV(Quality Value)値が30以上(P=0.1%)という満足できる値であった)。

チンパンジーについては、ゲノム配列の蓄積に比較して、cDNAやEST(Expressed Sequence Tag)配列は公共データベースにエントリーされはじめたばか

表1　公開されている生物種ごとのEST数
(NCBI dbEST, Sep. 5, 2003)

順位	生物種	登録数
1	ヒト	5,413,050
2	マウス	3,855,716
3	ラット	537,578
…	…	…
132	チンパンジー	4,983
…	…	…
	総数	18,140,083

りである。現在のGenBankのdbESTデータベースでは、チンパンジーは4983配列であり、生物種ごとの登録数は132位にすぎない。1位のヒト541万および2位のマウス386万と比較して、歴然とした違いがある(表1)。現在登録されているチンパンジーの4983配列のうち、1947配列は筆者らのグループが登録したものである(AU296732～AU298678)。

3．配列へのアノテーション情報の付加

シークェンサーから出力された個々の配列がどのような遺伝子なのかを示すアノテーション(生物学的情報)づけには、コンピューターを用いた解析が必須となっており、その方法論の探求がバイオインフォマティクスという新しい学問分野として成立している。筆者らは、配列の相同性検索にはBLAST[6]を用い、アラインメント(複数の似た配列を最も一致するように並列に並べる)にはClustalW[7]を用いた。そして、これらのプログラムの結果をさらに解析するプログラム(Perlスクリプト等)を作製し、独自の解析システムとして配列データの処理にあたった。

シークェンシングによって得られた配列(平均約600塩基)を、ヒト公共データベースのRefSeq mRNA、EST(以上NCBI(生物工学情報センター))、ゲノム配列(UCSC)に対してBLASTで振り分け、ヒト遺伝子の情報をもとにして、遺伝子名、配列中の翻訳領域、染色体上の座位などのアノテーション情報を付加した。その結果、総計1万4004の5′cDNA配列のうち、9131配列(3332種類)がヒト遺伝子(RefSeq mRNA)と高い相同性を持つことが分かった。約3万2000とされるヒトの遺伝子のうち、筆者らの完全長cDNAライブラリーからはチンパンジーの遺伝子として約10%、またヒトRefSeq mRNA約2万

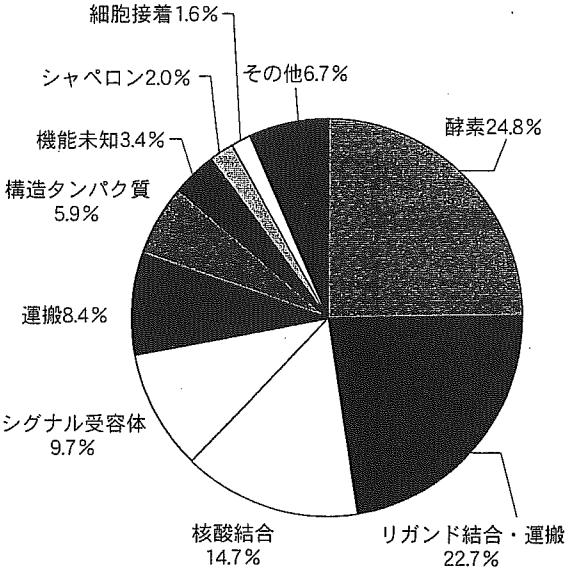

図1　チンパンジーcDNAライブラリーから得られた遺伝子の機能分類(Gene Ontology, Molecular Function)

の約17%にあたる対応配列が得られたことになる。

筆者らのcDNAライブラリー作製で得られた3332遺伝子をGene Ontology(http://www.geneontology.org/)のカテゴリーで機能分類した結果を図1に示す。機能分類としては全体として、酵素、リガンド結合・運搬、核酸結合タンパク質をコードする遺伝子が多く得られた。なお、脳と皮膚由来の遺伝子を分けても、機能の分布にそれほど大きな違いは見られなかった。

4．塩基置換率から遺伝子進化を探る

遺伝子の塩基配列を種間で比較することによって系統関係を求め、分岐年代を推定することができる。これまでにミトコンドリアゲノムの塩基配列を用いて霊長類種間の遺伝距離が推定されている[3]。細胞核にコードされている遺伝子については、現在GenBankに登録されている霊長類の遺伝子塩基配列は少ないが、これらを用いて霊長類(およびマウス)の系統関係を推定した報告があるので、最近の結果を図2に紹介する[8]。図では後述のK_S値によって系統樹を描いている。これらの研究結果から、ヒトとチンパンジーの分岐は約500～600万年前と推定されており、チンパンジーが他の霊長類と比較して最もヒトに近い種であることが確認されている。

筆者らは、チンパンジーcDNAライブラリーの作製とシークェンシングによる配列の蓄積の結果、従来

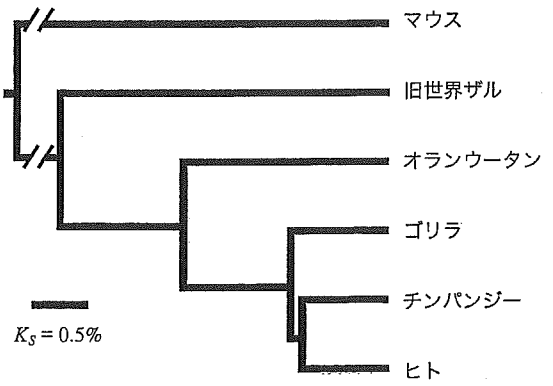

図2　塩基配列の比較によるヒトと他のサルおよびマウスとの系統関係(文献8)より改変)
翻訳領域の同義置換(コードするアミノ酸が変化しない)座位での塩基置換率(K_S)をもとに作図。

よりはるかに多種類の遺伝子について、ヒトとチンパンジーの比較を行うことができた。ヒトとチンパンジーのような近縁種間での一致度の高い塩基配列を比較するには、特に高精度の配列データが要求される。そこで筆者らは、同一遺伝子について3クローン以上得られた226遺伝子について、コンセンサス配列を作製したのちに、チンパンジーのヒトとの塩基置換率を解析した。

その結果、5′端配列全体、5′非翻訳領域(5′UTR)、翻訳領域(CDS)それぞれについて順に、0.69%、1.21%、0.58%の置換率を得た。アミノ酸配列の置換率は0.56%であった。同義置換(コードするアミノ酸が変化しない)座位および非同義置換(アミノ酸の変化をもたらす)座位における塩基置換率(K_SとK_A)は、1.33%および0.28%であった(図3)。これらの結果は、種間の遺伝距離および遺伝子配列の領域ごとの特性を表していると考えられる。特にK_Sは塩基置換率が進化の過程で一定であると考えられるため、種間の分岐年代を推定する指標として用いられることが多い(図2)。また、チンパンジーのK_S値(1.33%)は、ゲノム配列全体の種差1.23%[1]に近い値が得られた。最近の報告[9]では、ヒトとチンパンジー間で5′端の非翻訳領域(5′UTR)の方が、翻訳領域の同義座位(K_S)より保存性が低いとの結果が示されている。しかし筆者らのデータでは、これまでに他生物種においても一般にいわれている結果と同じく、5′UTR(1.21%)は同義座位(1.33%)よりわずかだが保存性が高かった。今後、5′UTR配列の機能解析を含めて検討されていくべき課題であろう。

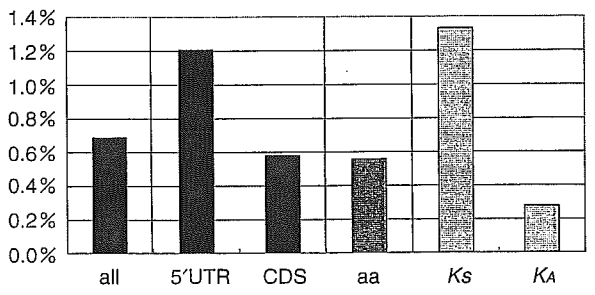

図3　チンパンジーとヒトの遺伝子配列の比較
cDNA配列全体(all)、5′非翻訳領域(5′UTR)、翻訳領域(CDS)における塩基とアミノ酸(aa)の置換率および同義座位と非同義座位での塩基置換率(K_S, K_A)。

比較に用いた226遺伝子のうち、最も塩基置換率が大きかったのはHLA-AとHLA-Bで、ともに翻訳領域で3.6%であった。これに対して、同じHLA関連遺伝子でもHLA-Eでは1.8%という比較的低い値であった。この結果は、類人猿のClass I MHC関連遺伝子のヒトとのオーソログス遺伝子のうち、HLA-A, B, Cは変異が大きく、HLA-E, F, Gは保存性が高いという結果[10]と矛盾しない。

チンパンジーの226遺伝子のコンセンサス配列を作製するもととなった1947クローンの配列については、DDBJ(日本DNAデータバンク)に登録を行った。これによって、公共データベースにおけるチンパンジー遺伝子配列(EST)の登録数が一挙に増加することになった。

5．遺伝子の構造変異

この研究で、ヒトとチンパンジーの遺伝子の塩基配列の置換率が小さいことが確認された。最近の他の研究では、ヒトとチンパンジーは同じ属に含めるべきという考え方[8]もある。他方、両種のゲノムDNA配列について、並列比較した時に見つかった挿入・欠失(indel)を考慮した場合、5%もの違いがあるという報告もなされている[11]。筆者らのcDNAでの解析結果からも挿入・欠失が見つかっており、これらはヒトゲノム(UCSC)との比較で配列の違いが確認されたものである。ほとんどが非翻訳領域における1～2塩基のindelであったが、翻訳領域でのチンパンジーに大きな挿入がある例も見つかった(図4)。なお、これらはすべてアミノ酸の挿入に対応した。筆者らのデータでも、このようなindelを考慮すると、塩基配列の違

(a) NM_000424 KRT5

```
PTR     ..............GGCTCCAAGAGGATATCCATCAGCACTAGTGGTGGCAGCTTCAGGAACCG
HSA     ..............GGCTCCAAGAGGATATCCATCAGCACTAGAGGAGGCAGCTTCAGGAACCG
                      ************************** ** **********************

PTR     GTTTGGTGCTGGTGCTGGTGCTGGAGGCGGCTATGGCTTTGGAGGTGGTGCCGGTAGTGG
HSA     GTTTGGTGCTGGTGCT-------GGAGGCGGCTATGGCTTTGGAGGTGGTGCCGGTAGTGG
        ***************       ***************************************

PTR     ATTTGGTTTCGGCGGTGGAGCTGGTGGTGGCTTTGGGCTCGGTGGCGGAG..........
HSA     ATTTGGTTTCGGCGGTGGAGCTGGTGGTGGCTTTGGGCTCGGTGGCGGAG..........
        ************************************************
```

(b) NM_003754 EIF3S5

```
PTR     ..............GACAAGATGGCCACACCGGCGGTACCAGTAAGTGCTCCTCCGGCCACGCC
HSA     ..............GACAAGATGGCCACACCGGCGGTACCAGTAAGTGCTCCTCCGGCCACGCC
                      ************************************************

PTR     AGCCCCAGTCCCGGCGGCGGCCCCAGTCTCGGCTCCAGCCTCAGTTCCAGCGCCAACGCC
HSA     AACCCCAGTCCCGGCGGCGGCCCCAG----------CCTCAGTTCCAGCGCCAACGCC
        * **********************            ********************

PTR     AGCACCGGGCTGCGGCTCCGGTTCCCGCTGCGGCTCCAGCCTCATCCTCAG.........
HSA     AGCACCGGGCTGCGGCTCCGGTTCCCGCTGCGGCTCCAGCCTCATCCTCAG.........
        ************************************************
```

図4　ヒト(HSA)とチンパンジー(PTR)のcDNA配列の比較で見つかった挿入・欠失(indel)の例

いは大きくなる。特に5′UTRでは、単純に見積もって2.1倍になった(CDSで1.2倍、5′cDNA全体では1.5倍)。

転写される遺伝子の配列において、数十～数百塩基単位でヒトとチンパンジーの間でブロック状に配列が挿入・欠失しているという相違が見られるケースがある。これらについては、ゲノム配列で種間の違いがあるのか、スプライシング変異なのかを解析する必要がある。スプライシング変異の場合は、種間で対応する変異なのかどうかも興味深い。このような大規模な配列の変異が機能に対してどのような影響を与えているのかを探索していくことが今後の課題である。

6．一塩基多型(SNP)

塩基配列の種内変異の代表的な例として、一塩基多型(SNP)が知られている。ヒトでは遺伝子コード領域で約1 kbに一つのSNPがあるとされているが、チンパンジーのDNA配列では種内変異がヒトよりも大きいという報告[12]があることから、SNPについても頻度が高いのではないかと推察される。筆者らは前に述べた226遺伝子の解析の一環として、cDNAにおけるSNP(cSNP)について調べた。ヒトでも多型性の高いHLA-AとHLA-Bホモログではそれぞれ、14と23の1塩基ごとの種内変異が見つかっている。これ以外でも、14遺伝子(vimentin, oncogene family, heat shock proteinなど)から17カ所のSNPと考えられる塩基の種内変異を発見した。17 SNPのうち14は翻訳領域にあり、その半分の7 SNPは非同義座位であった。非同義置換の割合が多いことから、チンパンジーでは遺伝子機能の種内変異が大きいことが示唆される。筆者らのcDNAライブラリーは2個体をもとに作製していることから、同一塩基座位において最大4パターンのバリエーションが存在する可能性がある。これらのチンパンジー遺伝子配列で見つかったSNPについては、ヒトでの相同部位での対応は確認できなかった。したがって、これらのSNPは両種が分岐したのち、種内で独自に広がっていったものがほとんどであろうと推察される。

7．遺伝子機能と発現プロファイル

脳と皮膚組織由来の遺伝子の機能分類では特に大きな分布の差は見られなかったことは前に述べた。一方、個々の遺伝子で見てみると、全体ではtubulin, eukaryotic elongation factorなどが多い。脳由来のクローンからはdyneinやsynaptosomal-associated proteinが、皮膚由来のクローンではkeratin類が多く得られ、組織特異的なスペクトルが見られた。遺伝子の種類数としては、脳組織から2658、皮膚組織から1166が同定されたが、両組織で重複するものは505だけであった。このことから、かなりの割合で遺伝子が組織特異的に発現している可能性が示唆される。ヒトとチンパンジーでの発現パターンが脳において特に異なり、ヒトで昂進しているとの報告もなされている[13]。しかしこれには技術的、あるいは統計処理上の問題が残っているので、今後の重要課題として筆者らもマイクロアレイを用いた解析などに取り組んでいく。

おわりに

本稿で紹介した解析結果は、2003年5月に筆者らが報告した内容が中心となっている[14]。配列データはその後も増加しているので、それらをもとにしてヒト配列との相違が遺伝子の機能にどう関わっていくのかを、より高次のレベルで解明を進める予定である。ヒトの遺伝子機能やその相互作用を含むパスウェイに関してはKEGG(Kyoto Encyclopedia of Genes and Genomes, http://www.genome.ad.jp/kegg/)などの

データベースが知られている。また、配列上の機能モチーフからのタンパク質の立体構造の予測なども、徐々に行えるようになってきている。このような遺伝子ネットワーク中のヒトmRNAに対し、対応するチンパンジーのcDNA全長配列を当てはめていくことによって、それぞれのケースでのシミュレーションが可能になるのではないかと考えられる。例えば、ヒトの疾患原因遺伝子群での対応を見るなど、様々なヒト遺伝子のプロテオーム解析に対して非常に有用であると考えられる。実際例としては、言語機能障害と関連するFOXP2 forkhead trascription factor遺伝子に関するチンパンジー遺伝子の多型解析などから、ヒトを特徴づける遺伝子の進化に関する重要な情報がもたらされている[15]。また、HIV感染による重篤なAIDSの発症や、ある種のマラリアへの感染性、また神経疾患や自己免疫疾患等などのように、ヒトに特徴的な罹患性を示す疾患の遺伝的背景を知ることも重要である。さらに、ヒトを含む霊長類の進化に関わる遺伝子の探求という側面でも重要なデータを与えると考えられる。

ヒトとチンパンジーの間で見られる遺伝子塩基配列の差異、すなわち調節領域の変化、遺伝子欠失、アミノ酸変化、遺伝子重複とそれに引き続く機能の多様化などは、表現型に結びつく候補である。それがどちらの系統で生じたかを知るうえで、質の良いアウトグループの配列データが必要となる。この点、我が国では前述の国立感染症研究所・橋本らによるカニクイザルの成果がある。したがって、ヒトならびにカニクイザルの配列データとの比較解析により、筆者らのチンパンジー遺伝子解析が重要な情報をもたらすと期待している。

本研究で得られた成果については、現在公開中の"PRIGEN-Primate Genes"(http://www.prigen.org/)のコンテンツとして順次追加して公開していく計画である。

○参考文献

1) Fujiyama, A. *et al.*: Construction and Analysis of a Human-Chimpanzee Comparative Clone Map, *Science*, 295, 131～134(2002)
2) Sibley, C. G. and Ahlquist, J. E.: The phylogeny of the hominoid primates, as indicated by DNA-DNA hybridization, *J. Mol. Evol.*, 20, 2～15(1984)
3) Horai, S. *et al.*: Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs, *Proc. Natl. Acad. Sci. USA*, 92, 532～536(1995)
4) Brett, D. *et al.*: Alternative splicing and genome complexity, *Nat. Genet.*, 30, 29～30(2002)
5) Maruyama, K. and Sugano, S.: Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides, *Gene*, 138, 171～174(1994)
6) Altschul, S. F. *et al.*: Basic local alignment search tool, *J. Mol. Biol.*, 215, 403～410(1990)
7) Thompson, J. D. *et al.*: 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(1994)
8) Wildman, D. E. *et al.*: Implications of natural selection in shaping 99.4% nonsynonymous DNA identity between humans and chimpanzees: Enlarging genus *Homo*, *Proc. Natl. Acad. Sci. USA*, 100, 7181～7188(2003)
9) Hellmann, I. *et al.*: Selection on Human Genes as Revealed by Comparisons to Chimpanzee cDNA, *Genome Res.*, 13, 831～837(2003)
10) Adams, E. J. *et al.*: Species-specific evolution of MHC class I genes in the higher primates, *Immunol. Rev.*, 183, 41～64(2001)
11) Britten, R. J.: Divergence between samples of chimpanzee and human DNA sequences is 5%, counting indels, *Proc. Natl. Acad. Sci. USA*, 99, 13633～13635(2002)
12) Kaessmann, H. *et al.*: Great ape DNA sequencesreveal a reduced diversity and an expansion in humans, *Nat. Genet.*, 27, 155～156(2001)
13) Enard, W. *et al.*: Intra- and Interspecific Variation in Primate Gene Expression Patterns, *Science*, 296, 340～343(2002)
14) Sakate, R. *et al.*: Analysis of 5'-End Sequences of Chimpanzee cDNAs, *Genome Res.*, 13, 1022～1026(2003)
15) Enard, W. *et al.*: Molecular evolution of FOXP2, a gene involved in speech and language, *Nature*, 418, 869～872(2002)