DNA Research **20**, 255–262, (2013)
Advance Access publication on 29 March 2013

doi:10.1093/dnares/dst007

Development and Characterization of cDNA Resources for the Common Marmoset: One of the Experimental Primate Models

Shoji Tatsumoto[1,†], Naoki Adati[1,‡], Yasushi Tohtoki[1,¶], Yoshiyuki Sakaki[1,§], Thorsten Boroviak[2], Sonoko Habu[3], Hideyuki Okano[4], Hiroshi Suemizu[5], Erika Sasaki[6], and Masanobu Satake[7,*]

RIKEN Genomic Sciences Center, Suehiro-cho 1-7-22, Tsurumi-ku, Yokohama 230-0045, Japan[1]; Centre for Stem Cell Research, Wellcome Trust and Medical Research Council Stem Cell Institute, University of Cambridge, Tennis Court Road, Cambridge CB2 1QR, UK[2]; Department of Immunology, Juntendo University School of Medicine, Hongo 2-1-1, Bunkyo-ku, Tokyo 113-8421, Japan[3]; Department of Physiology, School of Medicine, Keio University, Shinano-machi 35, Shinjyuku-ku, Tokyo 160-8582, Japan[4]; Biomedical Research Department, Central Institute for Experimental Animals, Tonomachi 3-25-12, Kawasaki-ku, Kawasaki 210-0821, Japan[5]; Department of Applied Developmental Biology, Central Institute for Experimental Animals, Tonomachi 3-25-12, Kawasaki-ku, Kawasaki 210-0821, Japan[6] and Institute of Development, Aging and Cancer, Tohoku University, Seiryo-machi 4-1, Aoba-ku, Sendai 980-8575, Japan[7]

*To whom correspondence should be addressed. Tel. +81-22-717-8477. Fax. +81-22-717-8482. Email: satake@idac.tohoku.ac.jp

Edited by Dr Toshihiko Shiroishi
(Received 9 May 2012; accepted 1 March 2013)

Abstract

The common marmoset is a new world monkey, which has become a valuable experimental animal for biomedical research. This study developed cDNA libraries for the common marmoset from five different tissues. A total of 290 426 high-quality EST sequences were obtained, where 251 587 sequences (86.5%) had homology (1E^{-100}) with the Refseqs of six different primate species, including human and marmoset. In parallel, 270 673 sequences (93.2%) were aligned to the human genome. When 247 090 sequences were assembled into 17 232 contigs, most of the sequences (218 857 or 15 089 contigs) were located in exonic regions, indicating that these genes are expressed in human and marmoset. The other 5578 sequences (or 808 contigs) mapping to the human genome were not located in exonic regions, suggesting that they are not expressed in human. Furthermore, a different set of 118 potential coding sequences were not similar to any Refseqs in any species, and, thus, may represent unknown genes. The cDNA libraries developed in this study are available through RIKEN Bio Resource Center. A Web server for the marmoset cDNAs is available at http://marmoset.nig.ac.jp/index.html, where each marmoset EST sequence has been annotated by reference to the human genome. These new libraries will be a useful genetic resource to facilitate research in the common marmoset.

Key words: common marmoset; cDNA; gene resource

1. Introduction

The mouse is a widely used and well-studied model animal for biomedical research. Many techniques for sophisticated genetic manipulations to model diseases have been established and researchers hope that the results obtained can be extrapolated to humans. Although this assumption is true in some cases, there are several areas of biomedical research

[†] Present address: National Institute of Genetics, Yata 1111, Mishima 411-8540, Japan.

[‡] Present address: Research Equipment Center, Hamamatsu University School of Medicine, Handayama 1-20-1, Higashi-ku, Hamamatsu 431-3192, Japan.

[¶] Present address: Division of Cancer Genomics, National Cancer Center Research Institute, Tsukiji 5-1-1, Chuo-ku, Tokyo 104-0045, Japan.

[§] Present address: Toyohashi University of Technology, Hibarigaoka 1-1, Tenpaku-cho, Toyohashi 441-8580, Japan.

where this proves to be more difficult. These areas include neuroscience, behavioural research, toxicology, drug development, and infectious diseases.[1] To overcome these limitations, efforts have been made to carry out biomedical studies in non-human primate model organisms, as the latter are more closely related to humans.

One of these established non-human primate model organisms is *Callithrix jacchus* (common marmoset). The marmoset is a small new world monkey and offers many advantages as an experimental animal over other non-human primates. It is small in size, which makes it comparatively easy to handle. Furthermore, it has been bred in captivity and its progeny have been maintained for >30 years in laboratory environments. Also, it does not harbour or transmit hazardous infectious agents. Therefore, the common marmoset is increasingly used in biomedical research worldwide. For example, models of autoimmune diseases involving the central nervous system have been developed in the common marmoset and it has been used extensively as a primate model.[2–4] More recently, genetically modified common marmosets have been produced successfully and their transgenes have been transmitted through the germ line.[5] In the future, it would be very useful to develop transgenic marmosets as models of human diseases.

Intensive efforts have been made to develop research tools for using the common marmoset as an experimental animal. For instance, several lines of monoclonal antibodies have been prepared, which are directed against immunity-related antigens of the marmoset.[6–8] Many, but not all anti-human antigen antibodies cross-reacted with the corresponding marmoset antigens, so it was necessary to establish marmoset-specific antibodies.[9] A pilot gene analysis study reported cDNA sequencing of immunity-related genes.[10] Based on genome-wide analyses, a draft sequence of the common marmoset, known as caljac3, was produced and made available to the public via the genome browser of the University of California Santa Cruz (http://genome.ucsc.edu/).

The current study describes the preparation of cDNA libraries for the common marmoset using five different cell types/tissues, which resulted in the identification of 290 426 high-quality EST sequences. These sequences were characterized by comparison with the sequences of six primate species, including humans. Overall, the EST sequences transcribed in the marmoset shared many common features with those from humans, whereas a small fraction was found to be unique to the marmoset.

2. Materials and methods

2.1. RNA extraction and library construction

Cytoplasmic RNA was extracted from the liver (MLI), brain and spinal cord (MSC), spleen (MSP), testis (MTE), and embryonic stem (ES) cells (MES) of the common marmoset using Trizol reagent. Marmoset ES cells were cultured as described previously.[11] Full-length cDNA libraries were constructed from the total RNAs of the aforementioned tissues/cells using a vector-capping method.[12] cDNAs generated from MLI, MSC, MSP, MTE, and MES were ligated into pGCAP1, pGCAPzf3, pGCAPzf3, pGCAP10, and pGCAP10 vectors, respectively. Colonies of *Escherichia coli* transformants were picked randomly, inoculated into 384-well plates using a Flexys colony picker (Genomic Solutions Ltd., Cambridgeshire, UK), and stored at −80°C.

2.2. EST sequencing

Colonies were picked from 99 plates for MSC; 200 plates each for MES, MSP, and MTE; and 201 plates for MLI (Supplementary Table S1). Sequencing templates were prepared using a TempliPhi DNA Amplification Kit (GE Healthcare UK Ltd., Buckinghamshire, UK). The sequencing reactions for the 5′-end directional ESTs were conducted using a BigDye Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems, Inc., CA, USA). The sequence primers used for pGCAP1, pGCAP10, and pGCAPzf3 were 5′-AGGCCTGTACGGAAGTGT-3′, 5′-AGGCCTGTACGGAAGTGT-3′, and 5′-CAAGGCGATTAAGTTGGGT-3′, respectively. The sequencing reaction products were purified by ethanol precipitation and loaded onto 3730 DNA Analyzers (Applied Biosystems Inc.).

2.3. Selection of high-quality EST data

The raw sequence data were basecalled using the KB basecaller program, which identified 345 600 sequences. A cross-match program was applied to the raw data to remove sequences derived from vectors and those added as caps during plasmid construction (−minimach 10, −minscore 20). Low-quality sequences [quality value (QV) = one for each nucleotide ± 3 neighbouring nucleotides measuring <105 in length] were masked by *N* (average QV > 15). If *N* was longer than 10 nucleotides, sequences located 3′ of these *N* were removed. Finally, only high-quality sequences longer than 100 nucleotides were selected and used in further analyses (see the high quality column, Supplementary Table S1). There were 290 426 high-quality ESTs, which represented 84% of the raw sequence data. High-quality EST sequences were obtained frequently and were relatively abundant (93%) in MES and

MTE, whereas their frequencies were comparatively low (72–80%) in MLI, MSC, and MSP.

The sequence length was higher in MES and MTE (702–706 nucleotides), than in MLI, MSC, and MSP (514–551 nucleotides; see the length column in Supplementary Table S1). The differences in the average length of readable sequences and the differences in the frequency of high-quality ESTs probably reflected the quality of each cDNA library. The cDNA libraries were also evaluated for transcription initiation sites by searching for the presence or absence of cap structure-derived guanine nucleotides at the extreme 5′ end of each EST (see the G-cap column, Supplementary Table S1). The frequency of G-cap-positive sequences was >80% in all five cDNA libraries, indicating that the synthesis of most cDNAs was initiated at the 5′ end.

2.4. *Registration of the sequences*

All of the EST sequences were deposited in the DNA Databank of Japan. Their accession numbers are as follows; HX373156 to HX444163 for MES cDNAs, HX444164 to HX500395 for MLI cDNAs, HX500396 to HX529651 for MSC cDNAs, HX529652 to HX591448 for MSP cDNAs, and HX591449 to HX663542 for MTE cDNAs.

2.5. *Tools used for sequence analyses*

The following tools were used for sequence analysis: BLAST (The Basic Local Alignment Search Tool) finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches),[13] BLAT (The BLAST-Like Alignment Tool uses the index to find regions in the genome likely to be homologous to the query sequence),[14] CD-HIT (a widely used program for clustering and comparing protein or nucleotide sequences. CD-HIT helps to significantly reduce the computational and manual efforts in many sequence analysis tasks and aids in understanding the data structure and correct the bias within a dataset),[15] CAP3 (a DNA sequence assembly program),[16] cross_match (this tool uses cross_match to mask vector/adapter sequences and optimally trim vector sequence and/or polyA/T trail. It takes a set of sequences to be masked, and a set of vectors/adapters used to perform masking),[17] EMBOSS (getorf finds and extracts open reading frames),[18] and InterProScan (a protein domains identifier).[19]

3. Results and discussion

3.1. *EST clustering and assembly*

The 290 426 sequences obtained by normalization of ESTs as in Section 2.3 were clustered and/or assembled to estimate how many genes/transcriptional units were read as ESTs (Table 1). Overlapping sequences were clustered within the longest sequence using the CD-HIT program, whereas overlapping sequences were assembled and extended into a contig using the CAP3 program. According to the CD-HIT program, the number of clusters varied from 10 010 in MLI to 29 028 in MTE. The summed number of contigs and singletons ranged from 8831 in MLI to 25 909 in MTE. According to both programs, MTE had the highest number of clusters/assemblies, suggesting that the testis had the greatest number of expressed transcriptional units. Notably, there were some more clusters than assemblies in each cDNA library. In fact, the total and non-redundant numbers in the five different libraries were 62 210 clusters and 60 568 assemblies (Table 1). This small difference (62 210 versus 60 568) suggests that most of the EST sequences obtained correspond to the mRNA 5′ end (as for another indication of similar performance of CD-HIT and CAP3, see Supplementary Fig. S1).

3.2. *Assignment of marmoset ESTs to the Refseq mRNAs of primates*

All 290 426 sequences were examined by comparing them with the known sequences registered at the NCBI as Refseq mRNAs (http://ncbi.nlm.nih.gov/RefSeq/). Refseqs from six different primates were used in the reference dataset, including *Homo sapiens*, *Pan troglodytes*, *Pongo abelii*, *Macaca mulatta*, *Nomascus leucogenys*, and *C. jacchus*. It should be noted that the human Refseqs were based on cDNA

Table 1. Clustering of ESTs by CD-HIT and assembly of ESTs by CAP3

Libraries	Number of ESTs	Number of clusters by CD-HIT	Number of contigs and singlets assembled by CAP3
MES	71 009	17 467	15 837 = 5519 (contig) + 10 318 (singlet)
MLI	56 232	10 010	8831 = 3319 (contig) + 5512 (singlet)
MSC	29 258	12 309	10 617 = 3764 (contig) + 6853 (singlet)
MSP	61 831	16 600	14 268 = 5086 (contig) + 9182 (singlet)
MTE	72 096	29 028	25 909 = 8044 (contig) + 17 865 (singlet)
All	290 426	62 210	60 568 = 17 232 (contig) + 43 336 (singlet)

Parameters used in CD-HIT and CAP3 programs were default.

sequences, whereas the Refseqs from other primate species were based mainly on the predictions of genomic sequences. The search program used was BLASTn and the threshold of significant homology was set to $1E^{-100}$, which was a very strict criterion.

A total of 239 920 and 231 084 sequences shared homology with human and marmoset Refseqs, respectively (Table 2). Homology with the combined Refseqs from the six primates was found for 251 587 sequences (85.6% of 290 426). Therefore, these 251 587 sequences were designated as primate homologues. The average length of the homologous sequences was 528 nucleotides. Of these 251 587 sequences, 4974 sequences were identical to their corresponding Refseq sequences, whereas 94 102 sequences shared 100% nucleotide sequence identity only with the aligned homologous regions. Additionally, 931 sequences shared homology with the Refseqs of non-primates. Most of these sequences were homologous to mouse Refseqs, suggesting that they were probably derived from mouse cells that were used as a feeder layer to culture ES cells.

Out of the above described 251 587 primate homologues, only sequences that shared homology with the Refseqs of all six primates were extracted, and that yielded 199 511 sequences. Then, for each of 199 511 sequences, the alignments between the EST and Refseq with the highest score among the six primates were selected, and 199 511 sequences were grouped into six for each primate species. Finally, the average identities and coverage were calculated using the alignments for each primate species (Table 3). As expected, the highest sequence identity (99%) was between the marmoset EST and marmoset Refseq. The average sequence identities between the marmoset EST and the Refseqs of the other five primate species were in a similar range (94–95%). Thus, the sequences differed by ~5% between the marmoset and the other five primates. The coverage was highest between the marmoset EST and human Refseq (91%), whereas it was lowest between the marmoset EST and marmoset Refseq (84%). This difference of 6% corresponded to 25 nucleotides, so it is likely that the alignment of the EST and marmoset Refseqs started 25 nucleotides downstream from the 5′ end of the marmoset Refseq. This might suggest that the 5′UTR of the marmoset Refseq was not predicted precisely from its genomic sequence.

Table 2. Assignment of common marmoset ESTs to primates' Refseq

Species derivation of Refseq	Number of ESTs homologous to Refseq (number of homologous genes)
Homo sapiens	239 920 (13 825)
Pan troglodytes	233 913 (14 372)
Callithrix jacchus	231 084 (13 499)
Pongo abelii	231 354 (13 898)
Macaca mulatta	229 151 (13 677)
Nomascus leucogenys	228 749 (13 296)
Six primates	251 587
Non-primates	931

EST sequences of common marmoset (total 290 426) were referred to primates' Refseq mRNA that are registered at NCBI. Homology was searched using BLASTn and judged significant at $<1E^{-100}$.

In the above assignment of marmoset ESTs to primates' Refseqs in Table 2, the Refseqs used as references can be re-classified into non-overlapping, distinct gene entities. Homologous sequences of 239 920 and 231 084 corresponded to 13 825 human and 13 499 marmoset genes, respectively (see the numbers in parenthesis in Table 2). In addition, we searched for HomoloGenes (http://www.ncbi.nlm.nih.gov/homologene) that are found in common among *Homo sapiens*, *Pan troglodytes*, and *Macaca mulatta*, and detected 9879 Homologenes. When identities and coverage between marmoset ESTs and primates' Refseq were recalculated for these 9879 Homologenes (see the numbers in parenthesis in Table 3), coverage increased by 2–3%, whereas identity between marmoset and human remained the same (94–95%). Thus, it appears again plausible that the sequences diverge by 5% between marmoset and other primates including human.

3.3. Mapping marmoset ESTs to the human genome

Each marmoset EST sequence (total 290 426 reads) was mapped to the human genome (hg19)

Table 3. Identity and coverage between homologous marmoset ESTs and primates' Refseq

Species derivation of Refseq	Identity for Refseq (for 9879 HomoloGenes)	Coverage for Refseq (for 9879 HomoloGenes)
Homo sapiens	94.88% (94.54%)	91.14% (93.91%)
Pan troglodytes	94.86% (94.54%)	88.84% (92.11%)
Callithrix jacchus	99.55%	84.70%
Pongo abelii	94.77%	87.36%
Macaca mulatta	94.73% (94.44%)	87.20% (89.66%)
Nomascus leucogenys	94.72%	87.80%

Identity (%) represents a degree of identity between the aligned two sequences of high-scoring segment pairs, whereas coverage (%) represents a ratio of aligned sequence length over an entire length of EST. See the text as for the details how identity and coverage were calculated.

using the BLAT search program. This showed that 99.1% of the ESTs (287 849 reads) mapped to the human genome (Table 4). This mapping result was filtered further using the UCSC Genome Browser and a pslCDnaFilter. Finally, 93.2% (270 673 reads) of the marmoset ESTs were assigned specifically and exactly to the human genome.

The nucleotide sequence identity between aligned sequences was calculated after each EST had been mapped to the genome (Supplementary Fig. S2). The majority of ESTs shared 90–96% identity with the human genome. The relatively low identity between the ESTs and human sequences was analogous to the identity between the EST and Refseq shown in Table 3. It should be noted that the degree of mapping to the human genome did not differ significantly among the five cDNA libraries (data not shown).

3.3.1. ESTs that were mapped to exonic regions of the human genome The CAP3 method identified 17 232 contigs (Table 1). The contig sequences represented those that were read multiple times in the overlapping regions and they were considered more reliable than single-read singletons. These 17 232 contigs were mapped onto the human genome according to the filtering method shown in Table 4. Furthermore, information from human Refseqs was used to determine whether the mapped contigs corresponded to exonic sequences of the human gene. This approach identified 15 089 contigs (88% of 17 232) that mapped onto the human genome and that corresponded to human Refseqs. These 15 089 contigs, which consisted of 218 857 ESTs, were considered to be representative of the genes that are commonly transcribed in marmosets and humans.

Of these 15 089 contigs, ESTs that were present in only one type of cDNA library were identified. Of these tissue-specific contigs, those with the highest numbers of constitutive ESTs are shown in Table 5. The contigs with >1000 ESTs were *ALB*, *HPR*, *ORM2*, and *ORM1* from MLI, and *PRM1* from MTE. Most of the genes listed in Table 5 are known to be characteristic of each specific cell type/tissue.

3.3.2. ES cell-specific transcripts Supplementary Table S2 shows previously reported cDNA/EST studies of primates other than the marmoset.[20–24] These species include *Pan troglodytes*, *Macaca fascicularis*, and *Chlorocebus sabaeus*, while the tissues included the brain, skin, liver, B lymphocytes, bone marrow, pancreas, spleen, thymus, and peripheral blood mononuclear cells. As the current study is the first example of the use of monkey-derived ES cells in EST studies, the ES-specific transcripts are mentioned briefly (see Supplementary Table S3 where the contigs containing >5 ESTs are listed).

The most notable were *LIN28A*, *NANOG*, and *SOX2* because, together with *OCT4*, they are known to reprogram human somatic cells to induced pluripotent stem cells.[25] *LIN28A* and *NANOG* contribute to the maintenance of pluripotency in stem cells.[26–28]

Table 4. Mapping of ESTs on the human genome

Libraries	Number of marmoset ESTs	Number of ESTs mapped on the human genome (raw data)	Number of ESTs mapped on the human genome (filtered)
MES	71 009	70 375	66 894 (94.2%)
MLI	56 232	55 602	52 405 (93.2%)
MSC	29 258	28 931	27 253 (93.1%)
MSP	61 831	61 300	58 170 (94.1%)
MTE	72 096	71 641	65 951 (91.5%)
All	290 426	287 849 (99.1%)	270 673 (93.2%)

Marmoset ESTs (290 426) were mapped on the human genome (hg19) by using a BLAT search program (−stepSize = 5, −minScore = 50, −minIdentity = 80, −repMatch = 2253). This initial mapping gave a number of 287 849 (99.1%) ESTs. These ESTs were then filtered, following a UCSC Genome Browser and using a pslCDnaFilter (−minId = 0.85, −minCover = 0.75, −globalNearBest = 0.0025, −minQSize = 20, −minNonRepSize = 16, −ignoreNs, −bestOveralp). Basis for adopting this filtering condition was as follows. A use of the condition such as (−minId = 0.95, −minCover = 0.25) selected only 159 309 ESTs (54.9%), indicating −minId = 0.95 to be extremely strict in the exactness. Therefore, we lowered the −minId to 0.85 (and −minCover = 0.25) and found the reasonably selected numbers of exact ESTs. Thus, under this −minId of 0.85, we then tried to improve the specificity by increasing −minCover to 0.75 (since 0.90 appeared too strict, 0.90 was not used). This −minCover number of 0.75 is roughly equal to the expected coverage of coding sequence [the average length of EST was 619 nt, and the average length of 5′UTR of human transcripts is 170 nt, therefore, an expected coverage between coding sequences and ESTs would be (619−170)/619 = 0.73]. Eventually, a condition of −minId = 0.85, −minCover = 0.75 filtered 270 673 ESTs (93.2%) as exact and specific.

Table 5. Top five genes expressed abundantly in each cDNA library

Libraries	Gene symbols	Descriptions	Number of ESTs
MES	*PYY*	Peptide YY	123
MES	*LIN28A*	Lin-28 homologue A	50
MES	*C6orf221*	Chromosome 6 open reading frame 221	46
MES	*NANOG*	Nanog homeobox	45
MES	*ERVMER34-1*	Endogenous retrovirus group MER34, member 1	28
MLI	*ALB*	Albumin	4492
MLI	*HPR*	Haptoglobin-related protein	1537
MLI	*ORM2*	Orosomucoid 2	1227
MLI	*ORM1*	Orosomucoid 1	1221
MLI	*APOA2*	Apolipoprotein A-II	673
MSC	*SNAP25*	Synaptosomal-associated protein, 25 kDa	109
MSC	*PLP1*	Proteolipid protein 1	58
MSC	*CALCA*	Calcitonin-related polypeptide alpha	51
MSC	*STMN2*	Stathmin-like 2	30
MSC	*THY1*	Thy-1 cell surface antigen	28
MSP	*MS4A1*	Membrane-spanning 4-domains, subfamily A, member 1	62
MSP	*ITGB2*	Integrin, beta 2	29
MSP	*HLA-DPA1*	Major histocompatibility complex, class II, DP alpha 1	26
MSP	*CLEC4F*	C-type lectin domain family 4, member F	20
MSP	*CD53*	CD53 molecule	18
MTE	*PRM1*	Protamine 1	1,198
MTE	*TNP1*	Transition protein 1	161
MTE	*HMGB4*	High mobility group box 4	139
MTE	*PHF7*	PHD finger protein 7	138
MTE	*DKKL1*	Dickkopf-like 1	107

Shown are the contigs that were detected only in one out of five cDNA libraries and possessed larger numbers of constituting ESTs.

C6orf221 (also known as *ECAT1*, ES cell-associated transcript 1) and *DPPA5* belong to the same gene family, and they are expressed specifically in human ES cells.[29,30] The frequent appearance of the claudin family (see *CLDN6*, *CLDN9*, *CLDN7*, and *CLDN4*) was also noted, although its biological significance is not known.

3.3.3. ESTs that were mapped onto the human genome outside exonic regions We noted that there was another category of contigs that was mapped onto the human genome, but not located in exonic regions (i.e. contigs that shared no homology with human Refseqs). These comprised 808 contigs (4.7% of 17 232) with 5578 ESTs. These 808 contigs represented transcribed genes in the marmoset. Although they were conserved in the human genome, they did not appear to be transcribed as genes; hence, they were not characterized as human genes.

Next, the number of marmoset cDNA libraries, in which each contig was expressed, was determined. The number of contigs with ESTs detected in libraries 1 (i.e. only MES), 2 (i.e. MES and MTE), 3, 4, and 5 (all of MES, MLI, MSC, MSP, and MTE) were 562, 195, 28, 7, and 6, respectively (808 in total). Of the 562 contigs, 29 were found in MES, 31 in MLI, 28 in MSC, 87 in MSP, and 387 in MTE (562 in total). It appeared that the genes expressed in the marmoset testes (387/562 = 69%) had an increased likelihood of not being expressed in humans compared with the genes expressed in other tissues (31%/4 = 7.8%).

Supplementary Table S4 shows the 20 contigs whose constituting ESTs' numbers were the largest among the 808 contigs. Interestingly, the five contigs that had the highest numbers of ESTs (ID: 198, 15591, 1258, 1567, and 3721) also had ESTs in all five libraries. In the previous paragraph, it was noted that six contigs were expressed in all five libraries. It was found that five out of these six contigs were widely expressed at a high level in the marmoset. This suggests that actively expressed genes in the marmoset may even be non-transcribed as genes in humans. Many of the remaining genes in Supplementary Table S4 were expressed in a combination of tissues e.g. MLI and MSP, or MLI and MSC, or only in MTE.

3.3.4. Characterization of 808 contigs in non-coding RNA or silent regions of the human genome The 808 contigs described above were mapped onto the human genome, but did not have any corresponding human Refseqs. The human genome browser at UCSC also supports mapping of RNA-seq and large intergenic non-coding RNAs (linc RNA).[31,32] The locations of these RNA-seq/linc RNAs were checked in relation to the marmoset ESTs mapped onto the human genome. The tissues that are used in common in RNA-seq/linc RNA studies[31,32] and the current study are the liver, brain, and testis; hence, 138 contigs expressed in MLI, 125 in MSC, and 516 in MTE (total 779) were used for the analyses (Supplementary Table S5). This showed that 500 contigs (64% of 779) matched RNA-seqs, while 98 contigs had corresponding sequences in RNA-seq and linc RNA. However, 259 contigs (33% of 779) did not have any counterparts in RNA-seq or linc RNA, suggesting that they were not expressed as RNA.

In summary, of the 808 contigs that did not have a corresponding human Refseq, two-thirds were probably expressed as non-coding RNA, whereas the other third appeared to be silent. This potentially suggests that these 808 may have lost their characteristics as genes during human evolution.

3.3.5. Identification of so far unknown genes based on probable full-length cDNA sequences The degree of conservation and/or divergence in marmoset and human genes was analysed using the following approach as well. Supplementary Fig. S3A shows a flow chart of the method used for selecting 'unknown genes.' First, getorf in EMBOSS was applied to extract 60 568 unigenes (see Table 1 where 60 568 represents the sum of contigs and singlets). Thus, sequences were selected that contain an open-reading frame spanning an initiating methionine through to a stop codon, with a poly(A) tail at the 3′ end. The G-cap was preserved in most ESTs, so the selected sequences probably represented 'full-length' cDNAs. This approach selected 3151 unigenes.

Each unigene was mapped onto the human genome. In total, 2595 sequences were mapped onto the human genome. Of the 556 marmoset sequences that could not be mapped onto the human genome, 127 were annotated and identified using the Refseq information of all living species, while 311 (306 + 5) shared homology with sequences registered in EST/nr sequence databases. The remaining 118 unigenes could not be annotated, suggesting that they represent potentially unknown genes.

Notably, only 1 sequence was derived from MLI and MSC, 3 from MSP, 14 from MES, while 104 from MTE. This might indicate comparatively abundant expression of unknown genes in the marmoset testis. The length of the polypeptides encoded by the 'unknown genes' ranged from 11 to 131 amino acids, with an average of 47 (Supplementary Fig. S4). It should be noted that no known domains/motifs were detected in these 'unknown' amino acid sequences using the INTERPRO program.

3.3.6. Further characterization of the 118 'unknown genes' It is possible that the 118 genes with ORFs might not encode polypeptides but may represent non-coding RNAs. In fact, five shared homology (1e-5) with non-coding RNAs in a comprehensive database (http://www.ncrna.org/frnadb/) and a marmoset-derived non-coding RNA database (ftp://ftp.ensembl.org/pub/release 70/fasta/callithrix_jacchus/ncrna/). Furthermore, 66 out of 118 genes could be aligned with the human genome if a 75% sequence similarity was employed, and each of these alignments covered more than half of each sequence. Therefore, a significant portion of the 118 genes had features of both polypeptide-coding genes and non-coding RNAs. Supplementary Fig. S3B and its legend provide further details on the unbiased approach used to identify unknown genes.

3.4. *Availability of the resources*

A Web server of marmoset cDNAs has been constructed (http://marmoset.nig.ac.jp/index.html), in which each marmoset EST sequence is mapped onto the human genome. Information accessible via the human genome browser can be obtained on this server, including the Ensembl gene annotation, Refseqs, and RNA-seq/linc RNA. This Web server can be used as a search engine, so the marmoset EST sequences can be BLASTed and the results displayed. The cDNAs libraries and/or the clones are available upon request from the DNA Bank, RIKEN BioResource Center (RDB no. 6388–6392). These deposited resources are expected to be valuable for future studies that use the common marmoset as an experimental animal model.

Acknowledgements: S.T. and M.S. are very grateful to Atsushi Toyoda (National Institute of Genetics, Mishima, Japan) for his continuous encouragement, which helped us to accomplish this work. We thank members of Sequence Technology Team, RIKEN Genomic Sciences Center for the generation of marmoset ESTs.

Supplementary data: Supplementary Data are available at www.dnaresearch.oxfordjournals.org.

Funding

Sequencing work was supported by a Special Fund for RIKEN Genomic Sciences Center.

References

1. Mansfield, K. 2003, Marmoset models commonly used in biomedical research, *Comp. Med.*, **53**, 383–92.
2. Massacesi, L., Genain, C., Lee-Parritz, D., et al. 1995, Active and passively induced experimental autoimmune encephalomyelitis in common marmosets: a new model for multiple sclerosis, *Ann. Neurol.*, **37**, 519–30.
3. Genain, C. and Hauser, S. 2001, Experimental allergic encephalomyelitis in the new world monkey Callithrix jacchus, *Immunol. Rev.*, **183**, 159–92.
4. 't Hart, B., Laman, J., Bauer, J., et al. 2004, Modeling of multiple sclerosis: lessons learned in a non-human primate, *Lancet Neurol.*, **3**, 588–97.

5. Sasaki, E., Suemizu, H., Shimada, A., et al. 2009, Generation of transgenic nonhuman primates with germline transmission, *Nature*, **459**, 515–6.
6. Izawa, K., Tani, K., Nakazaki, Y., et al. 2004, Hematopoietic activity of common marmoset CD34 cells isolated by a novel monoclonal antibody MA24, *Exp. Hematol.*, **32**, 843–51.
7. Ito, R., Maekawa, S., Kawai, K., et al. 2008, Novel monoclonal antibodies recognizing different subsets of lymphocytes from the common marmoset (Callithrix jacchus), *Immunol. Lett.*, **121**, 116–22.
8. Kametani, Y., Suzuki, D., Kohu, K., et al. 2009, Development of monoclonal antibodies for analyzing immune and hematopoietic systems of common marmoset, *Exp. Hematol.*, **37**, 1318–29.
9. Brok, H., Hornby, R., Griffiths, G., et al. 2001, An extensive monoclonal antibody panel for the phenotyping of leukocyte subsets in the common marmoset and the cotton-top tamarin, *Cytometry*, **45**, 294–303.
10. Kohu, K., Yamabe, E., Matsuzawa, A., et al. 2008, Comparison of 30 immunity-related genes from the common marmoset with orthologues from human and mouse, *Tohoku J. Exp. Med.*, **215**, 167–80.
11. Sasaki, E., Hanazawa, K., Kurita, R., et al. 2005, Establishment of novel embryonic stem cell lines derived from the common marmoset (*Callithrix jacchus*), *Stem Cells*, **23**, 1304–13.
12. Ohtake, H., Ohtoko, K., Ishimaru, Y., and Kato, S. 2004, Determination of the capped site sequence of mRNA based on the detection of cap-dependent nucleotide addition using an anchor ligation method, *DNA Res.*, **11**, 305–9.
13. Altschul, S.F., Gish, W., Miller, W., et al. 1990, Basic local alignment search tool, *J. Mol. Biol.*, **215**, 403–10.
14. Kent, W.J. 2002, BLAT—the BLAST-like alignment tool, *Genome Res.*, **12**, 656–64.
15. Li, W. and Godzik, A. 2006, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, *Bioinformatics*, **22**, 1658–9.
16. Huang, X. and Madan, A. 1999, CAP3: a DNA sequence assembly program, *Genome. Res.*, **9**, 868–77.
17. Ewing, B. and Green, P. 1998, Base-calling of automated sequencer traces using phred. II. Error probabilities, *Genome. Res.*, **8**, 186–94.
18. Rice, P., Longden, I. and Bleasby, A. 2000, EMBOSS: the European Molecular Biology Open Software Suite, *Trends in Genet.*, **16**, 276–7.
19. Quevillon, E., Silventoinen, V., Pillai, S., et al. 2005, InterProScan: protein domains identifier, *Nucleic Acids Res.*, **33**, 116–20.
20. Sakate, R., Osada, N., Hida, M., et al. 2003, Analysis of 5′-end sequences of chimpanzee cDNAs, *Genome. Res.*, **13**, 1022–6.
21. Chen, W.H., Wang, X.X., Lin, W., et al. 2006, Analysis of 10,000 ESTs from lymphocytes of the cynomolgus monkey to improve our understanding of its immune system, *BMC Genomics*, **7**, 82.
22. Uno, Y., Suzuki, Y., Wakaguri, H., et al. 2008, Expressed sequence tags from cynomolgus monkey (*Macaca fascicularis*) liver: a systematic identification of drug-metabolizing enzymes, *FEBS Lett.*, **582**, 351–8.
23. Osada, N., Hirata, M., Tanuma, R., et al. 2009, Collection of *Macaca fascicularis* cDNAs derived from bone marrow, kidney, liver, pancreas, spleen, and thymus. *BMC Res. Notes*, **2**, 199.
24. Tchitchek, N., Jacquelin, B., Winker, P., et al. 2012, Expression sequence tag library derived from peripheral blood mononuclear cells of the chlorocebus sabaeus, *BMC Genomics*, **13**, 279.
25. Yu, J., Vodyanik, M.A., Smuga-Otto, K., et al. 2007, Induced pluripotent stem cell lines derived from human somatic cells, *Science*, **318**, 1917–20.
26. Mitsui, K., Tokuzawa, Y., Itoh, H., et al. 2003, The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells, *Cell*, **113**, 631–42.
27. Darr, H. and Benvenisty, N. 2009, Genetic analysis of the role of the reprogramming gene *LIN-28* in human embryonic stem cells, *Stem Cells*, **27**, 352–62.
28. Xu, B., Zhang, K. and Huang, Y. 2009, Lin28 modulates cell growth and associates with a subset of cell cycle regulator mRNAs in mouse embryonic stem cells, *RNA*, **15**, 357–61.
29. Kim, S.K., Suh, M.R., Yoon, H.S., et al. 2005, Identification of developmental pluripotency associated 5 expression in human pluripotent stem cells, *Stem Cells*, **23**, 458–62.
30. Pierre, A., Gautier, M., Callebaut, I., et al. 2007, Atypical structure and phylogenomic evolution of the new eutherian oocyte- and embryo-expressed *KHDC1/DPPA5/ECAT1/OOEP* gene family, *Genomics*, **90**, 583–94.
31. Wang, E.T., Sandberg, R., Luo, S., et al. 2008, Alternative isoform regulation in human tissue transcriptomes, *Nature*, **456**, 470–6.
32. Cabili, M.N., Trapnell, C., Goff, L., et al. 2011, Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses, *Genes Dev.*, **25**, 1915–27.