FEBS Letters 582 (2008) 351–358

Expressed sequence tags from cynomolgus monkey (*Macaca fascicularis*) liver: A systematic identification of drug-metabolizing enzymes

Yasuhiro Uno[a,*], Yutaka Suzuki[b,*], Hiroyuki Wakaguri[b], Yoshiko Sakamoto[a], Hitomi Sano[a], Naoki Osada[c], Katsuyuki Hashimoto[c], Sumio Sugano[b], Ituro Inoue[a,d]

[a] *Division of Genetic Diagnosis, Institute of Medical Science, The University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan*
[b] *Department of Medical Genome Sciences, Graduate School of Frontier Sciences, University of Tokyo, 4-6-1 Shirokanedai, Minatoku, Tokyo 108-8639, Japan*
[c] *Department of Biomedical Resources, National Institute of Biomedical Innovation, Ibaraki, Osaka, Japan*
[d] *Division of Molecular Life Science, School of Medicine, Tokai University, Shimokasuya 134, Isehara, Kanagawa 259-1193, Japan*

Received 1 October 2007; revised 14 December 2007; accepted 18 December 2007

Available online 31 December 2007

Edited by Takashi Gojobori

Abstract **The liver, a major organ for drug metabolism, is physiologically similar between monkeys and humans. However, the paucity of identified genes has hampered a deep understanding of drug metabolism in monkeys. To provide such a genetic resource, 28655 expressed sequence tags (ESTs) were generated from a cynomolgus monkey liver full-length enriched cDNA library, which contained 23 unique ESTs homologous to human drug-metabolizing enzymes. Our comparative genomics approach identified nine lineage-specific candidate ESTs, including three drug-metabolizing enzymes, which could be important for understanding the physiological differences between monkeys and humans.**

Keywords: Cynomolgus monkey; Drug metabolism; Drug-metabolizing enzyme; Expressed sequence tags; Lineage-specific gene; Liver

1. Introduction

Cynomolgus monkeys have been used as an animal model for the investigation of human physiology and disease because of their close genetic and physiological similarities to humans. Application of this animal model includes predicting metabolic fate of newly developed drugs due to pharmacokinetics similar to humans. However, we now know that differences in metabolic properties are occasionally seen for some drugs between monkeys and humans [1–7] possibly due to differences in genetic components essential for drug metabolism between the two lineages, such as lineage-specific genes and alternatively spliced transcripts. However, limited numbers of lineage-specific genes identified in monkeys have hampered complete knowledge of lineage differences in drug metabolism.

An expressed sequence tag (EST)-sequencing approach has been a rapid and efficient way to identify novel cDNAs that provide a basis to investigate genetic components essential to various physiological functions. In non-human primates, efforts have been made for the comprehensive identification of ESTs in chimpanzees [8], rhesus monkeys [9,10], and cynomolgus monkeys [11–13]. However, liver tissue has not been extensively sequenced for ESTs, thus only limited genetic information is available on liver physiological function such as drug metabolism. With the completion of a draft of the rhesus monkey genome sequence [14], EST analysis of macaques should be more feasible and accurate.

To provide a monkey genetic resource, 28655 ESTs from cynomolgus monkey liver were generated. These macaque ESTs analyzed against the rhesus genome identified 1064 unique ESTs, most of which (77.0%) matched the human RefSeq database. cDNAs highly homologous to human drug-metabolizing genes were identified, including those of cytochrome P450 (CYP), UDP-glucuronosyltransferase (UGT), glutathione *S*-transferase (GST), sulfotransferase (SULT), and flavin-containing monooxygenase (FMO). Moreover, our method to select lineage-specific ESTs successfully identified novel transcripts related to drug metabolism. This genetic information should help in discerning various physiological characteristics, including drug metabolism in monkeys.

2. Materials and methods

2.1. cDNA library construction and EST sequencing

Liver samples were collected from three adult cynomolgus monkeys (two males and one female) and used to generate a full-length enriched cDNA library using the pME18S-FL3 vector by the oligo-capping method as previously described [15]. Purified DNA was sequenced using the ABI PRISM® BigDye™ Terminator Cycle Sequencing Ready Reaction Kit, Version 2.0 (Applied Biosystems, Foster City, CA), followed by electrophoresis with ABI-3700 DNA Analyzer (Applied Biosystems) according to the manufacturer's instructions. Primers (5′-GGATGTTGCCTTTACTTCTA-3′ and 5′-TTTTTTTTTTTT-TTTTTV-3′) were used for single-pass sequencing of 5′ and 3′-ends for each cDNA, respectively.

*Corresponding authors.
E-mail addresses: unoxx001@pharm.hokudai.ac.jp (Y. Uno), ysuzuki@k.u-tokyo.ac.jp (Y. Suzuki).

Abbreviations: CYP, cytochrome P450; EST, expressed sequence tag; FMO, flavin-containing monooxygenase; GST, glutathione *S*-transferase; ORF, open reading frame; SULT, sulfotransferase; UGT, UDP-glucuronosyltransferase

0014-5793/$32.00
doi:10.1016/j.febslet.2007.12.031

2.2. Sequence data analysis

Vector sequence was trimmed and sequence quality was inspected using Phred (University of Washington). Only EST sequences longer than 200 bases were used. Generated EST sequences were first computationally mapped to the *Macaca mulatta* genomic sequence (rheMac2, UCSC Genome Browser). Computational mapping was carried out as previously described by consequential use of sequence alignment programs, BLAT and SIM4 [16]. Only ESTs over the entire sequence length that mapped perfectly at unique positions on the macaque genome were regarded as "mapped". Further information for each cDNA is presented in our database, DBTSS (http://dbtss.hgc.jp), and a user manual has been published [16].

The macaque genomic sequences to which our cynomolgus ESTs mapped were examined for any corresponding human genomic and RefSeq sequence. If any, the corresponding macaque EST was correlated with the human RefSeq gene. Based on information from the correlated human RefSeq gene, GO (Gene Ontology) classification was carried out for macaque ESTs using GO slim (http://www.geneontology.org/) for "Biological Process", Molecular Function", and "Cellular Component".

2.3. Identification of putative macaque-specific transcripts

To identify macaque ESTs that do not match to human genes, the ESTs were analyzed by either a genome- or cDNA-based approach. In the genome-based approach, we selected the EST sequence located outside human-macaque alignable regions according to the genome–genome alignment in the UCSC Genome Browser. In the cDNA-based approach, ESTs were first searched with the human RefSeq database using BLASTN (cut-off = 1.0e−100). Those ESTs with no hits were clustered with each other (cut-off = 0.0; >98% identity) and clusters containing more than 10 ESTs were selected. Those clustered cDNAs were searched against the human RefSeq database again (1.0e−50), and the generated sequence alignments were further manually inspected. For the macaque-specific transcript candidates, complete sequences were determined by primer-walking.

3. Results and discussion

3.1. Sequencing and clustering of macaque liver ESTs

A full-length cDNA library was constructed from cynomolgus monkey (*Macaca fascicularis*) livers using the oligo-capping method [15]. One-pass sequencing at 5′ and 3′-ends of the liver cDNA clones and sequence processing generated a total of 28 655 high quality ESTs (deposited in GenBank under Accession Nos. BB873801–BB902455). Only 3′ ESTs (27959 entries) were further analyzed. Of these ESTs, 14727 (53%) were successfully mapped to 1064 different regions in the rhesus macaque genome. Of the 1064 regions, 819 (77%) reside in genomic regions highly homologous to human RefSeq genes as revealed by a genome–genome comparison, and were annotated with human RefSeq genes (Table 1). Clustering of 27959 ESTs was carried out by calculating the number of ESTs that mapped to the same region, which should represent the cluster size for the corresponding gene. This analysis for the 1064 mapped regions in the genome indicated that these 1064 unique ESTs consisted of 525 contigs (49.3%) and 539 singletons (50.7%). The number of members in each cluster ranged up to 4354, with a 26.9 average. The gene expression profile based on our EST data reflected liver functional characteristics because the most abundantly expressed genes were hepatocyte-specific markers, such as albumin, fibrinogen gamma and beta polypeptides, haptoglobin, and alcohol dehydrogenase, all of which comprised more than half of the identified ESTs (Table 1). Such high redundancy of hepatic ESTs from the non-normalized cDNA libraries has been also seen for human libraries [17–19].

3.2. Functional classification of ESTs

Provisional functional classification was carried out using GO slim terms based on the human RefSeq genes that correlated with our macaque ESTs (Fig. 1). Out of 819 unique ESTs that matched a human RefSeq entry, 786 (96.0%) were assigned to at least one main category; Biological Process, Molecular Function, and Cellular Component, to which 520, 458, and 373 sequences (48.9%, 43.0%, and 35.1%) were classified, respectively. Sequences from 133 ESTs (16.2%) were annotated into all three categories. The largest EST groups include metabolism, transcription, protein biosynthesis, electron transport, transport, signal transduction, and lipid metabolism (Fig. 1A) for Biological Process, and protein binding, transferase activity, and nucleotide binding for Molecular Function (Fig. 1B).

3.3. ESTs relevant to drug metabolism

Our major objective was identification of cDNAs important for drug metabolism, namely those encoding drug-metabolizing enzymes, which belong to CYP, UGT, GST, SULT, and FMO families. The 446 ESTs in 23 clusters were highly homologous to genes for such drug-metabolizing enzymes in humans (Table 2). For the CYP family, only ESTs for the CYP1 to CYP4 subfamilies are indicated in the list because of their importance in drug metabolism. The CYP family contained 231 entries (51.8%), the largest group among the ESTs for drug-metabolizing enzymes. CYP, a phase I drug-metabolizing enzyme, is involved in hydroxylation of a large number of

Table 1
Genes abundantly expressed in liver (>100 reads)

Contig number	ESTs	Human RefSeq ID	Annotation
15043	4354	NM_000477	Albumin
15223	3763	NM_000509	Fibrinogen gamma chain
15577	1097	NM_005141	Fibrinogen beta chain
10429	277	NM_005143	Haptoglobin
19271	253	NM_000668	Alcohol dehydrogenase IB (class I), beta polypeptide
17733	245	NM_001085	Serpin peptidase inhibitor, clade A, member 3
5070	227	NM_000035	Aldolase B, fructose-bisphosphate
6405	158	NM_016413	Carboxypeptidase B2 (plasma, carboxypeptidase U)
5438	154	NM_000638	Vitronectin
11141	130	NM_000354	Serpin peptidase inhibitor, clade A, member 7
12017	125	NM_001622	Alpha-2-HS-glycoprotein
15222	109	NM_000508	Fibrinogen alpha chain
17083	106	NM_001756	Serpin peptidase inhibitor, clade A, member 6

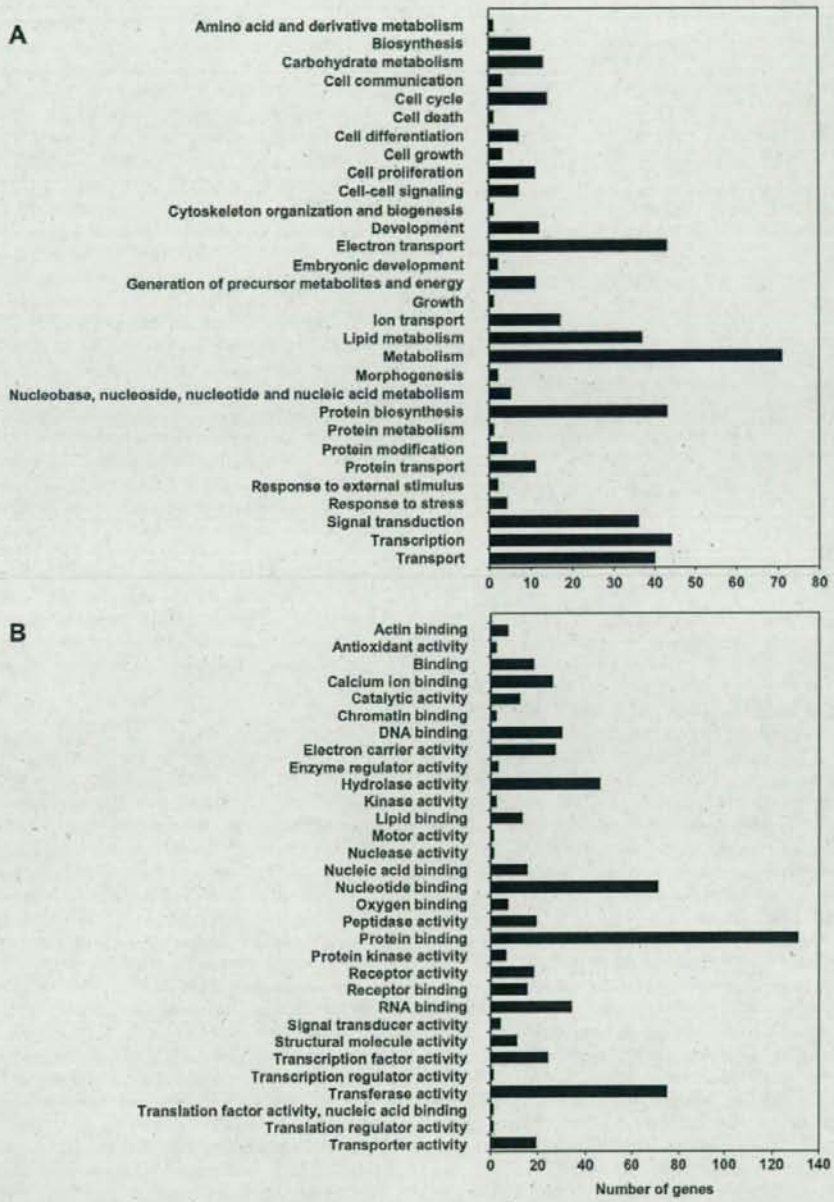

Fig. 1. Functional classification of cynomolgus liver ESTs. All non-redundant ESTs were assigned to each functional category as described in Section 2. Biological process (A) and Molecular function (B) are shown.

drugs [20]. Among the CYP ESTs identified, 124 (53.7%) belonged to the CYP2C subfamily that is important for metabolism of ~20% of all prescribed drugs such as tolbutamide, phenytoin, and warfarin [21]. Fifty-two ESTs belonged to the CYP3A subfamily. In humans, genes in the CYP3A subfamily (especially *CYP3A4*) are essential for drug metabolism, and are involved in the metabolism of more than half the currently prescribed drugs. Moreover, human CYP3A4 and CYP3A5 occupy more than half of the total CYP protein content in liver [20], contributing substantially overall drug

Table 2
Cynomolgus ESTs highly homologous to human drug-metabolizing enzyme families, CYP, UGT, GST, SULT, and FMO

Family	Contig number	Number of ESTs	Accession number	Matched human cDNA
CYP	18729	89	NM_0007700	CYP2C8
	12912	38	NM_017460	CYP3A4
	463	30	NM_000106	CYP2D6
	19109	21	NM_000769	CYP2C19
	19111	14	NM_000771	CYP2C9
	12910	14	NM_000777	CYP3A5
	19262	13	NM_000773	CYP2E1
	7671	8	NM_023944	CYP4F12
	8451	3	NM_000775	CYP2J2
	8410	1	NM_000778	CYP4A11
UGT	15423	75	NM_001074	UGT2B7
	2531	41	NM_019093	UGT1A3
	15424	30	NM_050394	UGT2B28
GST	19270	9	NM_145740	GSTA1
	14165	8	NM_000846	GSTA2
	9619	7	NM_000851	GSTM5
	1616	7	NM_145792	MGST1
	19167	3	NM_004832	GSTO1
	17697	2	NM_145870	GSTZ1
SULT	10631	10	NM_001055	SULT1A1
	19266	2	NM_001054	SULT1A2
	3203	1	NM_006588	SULT1C2
FMO	10051	20	NM_001002294	FMO3

metabolism in humans. Thirty ESTs were highly similar to human CYP2D6. In the human genome, three *CYP2D* genes are present including one functional *CYP2D* gene (*CYP2D6*) and two pseudogenes (*CYP2D7* and *CYP2D8*). CYP2D6 accounts for 5% of the total hepatic CYP content and is responsible for the metabolism of 25% of all drugs oxidized by CYPs [20]. In cynomolgus monkeys, CYP2D17, which is highly homologous to human CYP2D6, has been isolated [22]. Meanwhile, marmoset is known to have two functional CYP2Ds with different metabolic properties, CYP2D19 and CYP2D30 [23]. Further in-depth analysis of our EST clones could reveal whether *CYP2D17* is the only *CYP2D* gene expressed in cynomolgus monkey liver. Characterization of these CYP EST clones is currently in progress, such as full-length sequencing, tissue expression patterns, and metabolic assays, the outcome of which has been partly published [24,25].

Clusters for other drug-metabolizing enzymes of UGT, GST, SULT, and FMO families contained 146, 36, 13, and 20 ESTs, respectively (Table 2). UGT, a phase II drug-metabolizing enzyme, catalyzes the conjugation of various drugs to assist drug excretion and is composed of UGT1A, UGT2A, and UGT2B subfamilies in humans. The 146 ESTs for the UGT family were grouped into three clusters. Forty-one EST sequences were highly homologous to human UGT1A3. The *UGT1A* gene locus contains 13 distinct first exons (promoters) followed by exons 2–5 that are shared among all 13 transcripts, giving rise to nine different proteins (four pseudogenes) in humans [26]. Considering that these ESTs were 3′ cDNAs, the 41 EST clones possibly encode multiple *UGT1A* genes. Because only four *UGT1A* genes have been identified for macaques, further sequence analysis of these UGT1A ESTs could lead to the identification of novel *UGT1A* genes in this lineage. Seventy-five and 30 ESTs matched human UGT2B7 and UGT2B28, respectively. Initial full-length sequencing of the UGT2B EST clones revealed that the clones contained the UGT2B33 cDNA (GenBank Accession No. AB371703) newly identified in cynomolgus monkeys as well as the previously identified cDNAs for cynomolgus UGT2B9, UGT2B18, UGT2B19, UGT2B20, UGT2B23, and UGT2B30 (GenBank Accession Nos. U91582, AF016310, AF112112, AF072223, AF112113, and AF401657, respectively). In contrast to *UGT1A*, *UGT2B* genes have been frequently duplicated in many mammalian lineages [26]; therefore, some of these *UGT2Bs* are possibly lineage-specific genes as discussed below.

Other EST sequences were highly homologous to six human genes, two genes, and one gene in the GST, SULT, and FMO families, respectively (Table 2). GST is another phase II enzyme, catalyzing the conjugation of electrophilic substrates to glutathione and is composed of at least 16 genes for cytosolic, mitochondrial, and microsomal GSTs in humans [27]. SULT is also a gene family comprising at least 10 human genes, catalyzing sulfate conjugation of a wide variety of drugs [28]. FMO is a family of flavoproteins, catalyzing oxygenation of various drugs containing sulfur, nucleophilic nitrogen, and phosphorus heteroatoms [29]. Full-length sequencing and functional characterization of these EST clones are currently under investigation, by which novel genes could be identified because limited numbers of genes have been identified for these enzymes in monkeys. These results suggest that our EST-sequencing approach successfully identified a number of cDNA clones for various drug-metabolizing enzymes in monkeys.

3.4. Identification of lineage-specific genes

In order to better utilize monkeys as an animal model, it is essential to understand similarities or differences in genes expressed between monkeys and humans. The EST data should provide essential information on lineage-specific genes and transcripts. To identify macaque-specific transcripts, 27959 3′ ESTs were analyzed by either a genome- or cDNA-based approach. In the genome-based approach, we found 77 EST clusters, for which at least a part of the sequences were located outside human-macaque alignable regions. In the cDNA-based approach, we identified 12 clusters containing >10 ESTs that were unmatched to any human RefSeq genes according to BLASTN (cut-off = 1.0e−100). Clones available for the 10 remaining candidate clusters after subsequent manual inspection, along with clones for the 29 clusters randomly selected from 77 candidates in the genome-based approach, were subjected to full-length sequencing (excluding 1 overlapping clone). Sequence analysis of these 38 clones confirmed that nine clones contained lineage-specific candidate genes. Of these, six clones matched to human RefSeq sequences (Table 3): two clones lack a portion of human genome sequence and the other four matched to more than one member of a gene family. Thus, these four clones were potentially lineage-specific genes and were further characterized as described below.

One candidate clone (Qlv-U097A-G10) encoded CYP2C76 with a relatively low homology (~80%) to members of the human CYP2C subfamily, CYP2C8, CYP2C9, CYP2C18, and CYP2C19 (Table 3). The extent of homology was much lower than those for other ESTs (~95%). Our characterization

Table 3
Potential lineage-specific ESTs in cynomolgus monkey

Clone ID	GenBank Accession number	Nucleotide (bp)	ORF[a] (Number of amino acids)	Cynomolgus sequence	The most highly homologous human RefSeq cDNAs	Genome- or cDNA-based approach	Aligned location
Novel member of gene family							
Qlv-U042A-F11	AB362497	1637	454	None	CFH, CFHR3/4	Genome/cDNA	Intergenic
Qlv-U097A-G10	AB362507	1986	489	CYP2C76	CYP2C8/9/18/19	cDNA	Intergenic
Qlv-U346A-B11[b]	AB371605	1758	472	CYP2A23	CYP2A6/7/13	cDNA	Intergenic
Qlv-U405A-G11	AB362508	2225	528	UGT2B19	UGT2B4	cDNA	Intergenic
Partially unmatched to human genome							
Qlv-U244A-C6[b]	AB362499	1612	305	None	TSPAN12	Genome	Intergenic
Qlv-U258A-D7[b]	AB362500	1984	89	None	SS18L1	Genome	Intergenic
Unmatched to human genome							
Qlv-U050A-D10	AB362503	1700	34	None	None	Genome	Intron
Qlv-U295A-A3	AB362504	2278	118	None	None	Genome	Intergenic
Qlv-U389A-C1	AB362506	2043	90	None	None	Genome	Intergenic

[a]The longest ORF was selected.
[b]Transcript variants with different exon-intron structure from human homologs.

of CYP2C76 (GenBank Accession No. DQ074807) at the RNA, protein, and genomic level revealed that this CYP2C did not have any human ortholog because the corresponding genes were not found in the human genome [24]. Moreover, this CYP2C76 was at least partly responsible for lineage differences in drug metabolism [30]. These results confirmed that our comparative genomic approach succeeded in identifying macaque-specific transcripts that are absent in humans.

One clone (Qlv-U405A-G11) identified as a lineage-specific candidate contained the cDNA for UGT2B19 previously reported [31]. Cynomolgus UGT2B19 as well as UGT2B30 cDNAs were both highly homologous (92%) to human UGT2B4 cDNA [32]. A phylogenetic comparison (Fig. 2) indicated that the 1-to-1 orthologous relationship to the human UGT2Bs could not be determined for these cynomolgus UGT2Bs, raising the possibility that *UGT2B19* might be a lineage-specific gene. *UGT2B19* is expressed in cynomolgus monkey liver and prostate and has enzymatic activity to xenobiotics (1-naphthol) and steroids (testosterone) [31]. The UGT2B subfamily consisted of a number of member genes including a lineage-specific candidate [26], suggesting that UGT2B19 and other functional UGT2B enzymes in cynomolgus monkeys contribute not only to overall drug metabolism in monkey liver but also possibly to differences in drug metabolism.

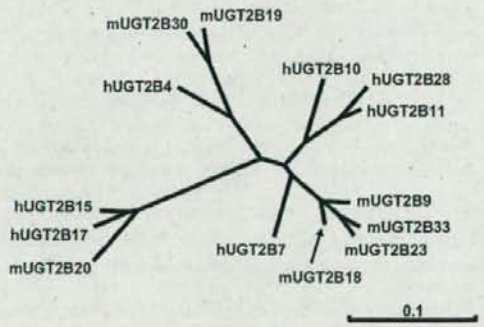

Fig. 2. A phylogenetic comparison of UGTs between macaque and human. The phylogenetic tree was based on amino acid sequence using the Clustal W program. Deduced amino acid sequences were used for cynomolgus monkeys (m) and human (h).

Another lineage-specific candidate clone (Qlv-U346A-B11) was cynomolgus CYP2A23 variant (tentatively named CYP2A23v), containing exons 1–8 with a partial intron 8 sequence and thus, lacking the entire exon 9 as compared to a complete CYP2A23 transcript. CYP2A23 and another cynomolgus CYP2A, CYP2A24, were both highly homologous (~95%) to the three human CYP2As, specifically CYP2A6, CYP2A7, and CYP2A13, indicating the difficulty in determining the orthologous relationship of CYP2A23 and CYP2A24 to human CYP2As [25]. This novel CYP2A23 transcript variant encodes a protein of 472 amino acids and lacks a part of a heme-binding region essential for CYP proteins (Fig. 3). The protein generated from this transcript, therefore, might not function as a drug-metabolizing enzyme. A similar transcript variant was also identified for CYP2C76 and UGT2B19 (data not shown). It remains to be determined whether the presence of these transcripts lacking a functional domain is limited to the animals that provided liver samples for the cDNA library construction and what roles these transcript variants play in drug metabolism.

Other than those for drug metabolism, one lineage-specific candidate (Qlv-U042A-F11) had high sequence homology to complement factor H (CFH) family genes in humans (Table 3). CFH (also called Factor H), an important complement regulator, forms a gene family along with CFH-related proteins (CFHL1-5) in humans [33]. This macaque transcript contained an open reading frame (ORF) of 454 amino acids. CFH and other genes important for immune response and T cell-mediated immunity such as *immunoglobulin-like* genes and MHC-related genes have been identified in macaques as the genes that went under positive selection [13,14,34], and thus, our finding of lineage-specific CFH-like sequence in macaques is not surprising. Further analysis of this CFH-like sequence indicated that the first 19 amino acids and the remaining amino acids were highly similar to CFH-related proteins (CFHR3 and CFHR4) and CFH in humans, respectively (data not shown), raising the possibility that this novel transcript might be a hybrid of CFH and CFH-related genes. In humans, a hybrid transcript of CFH and CFHR1 has been identified and implicated in atypical haemolytic uraemic syndrome [35].

Fig. 3. Alignment of the amino acid sequences deduced from cynomolgus monkey (m) and human (h) CYP2A cDNAs. The sequences were aligned using the Clustal W program. Asterisks and dots under the sequences indicate identical amino acids and conservatively unchanged amino acids, respectively. A black line under the amino acid sequences indicates the putative heme-binding region. The CYP2A23 variant (mCYP2A23v) newly identified lacks half of the putative heme-binding region.

These results suggest that our approach of lineage-specific gene identification successfully identified potential lineage-specific genes or transcripts, possibly relevant to the immune system. Further investigation of other ESTs should help make better use of the macaque for immunological studies.

Three candidate genes were unmatched to any human RefSeq sequence, and thus could be apparent lineage-specific genes (Table 3). The two candidate genes (Qlv-U295A-A3 and Qlv-U389A-C1) reside in intergenic regions of the macaque genome, which might be novel genes in monkeys. This was supported by the fact that these two sequences did not match any human ESTs by BLAST (data not shown). The two transcripts contained relatively small ORFs (<100 amino acids). Transcripts with small ORFs have been identified in mice and humans, some of which could be actually translated in vitro [36,37]. Alternatively, these mRNAs might be functioning as non-coding RNAs. A large proportion of transcripts are non-coding RNAs, including those having essential functions in transcriptional and translational control [38,39].

4. Conclusion

The data presented here provide an overview of genes expressed in cynomolgus liver to investigate liver physiology for macaques. ESTs for genes encoding a variety of drug-metabolizing enzymes hold great promise in deepening our understanding of drug metabolism in monkeys, which in turn helps to elucidate lineage differences between monkeys and humans. Indeed, our characterization of CYP2C ESTs has identified lineage-specific CYP2C76, which is responsible for lineage differences in drug metabolism [24,30]. Furthermore, the ESTs generated in this study can be a resource for the production of microarrays. Given that our cDNA library was generated with RNAs from only three animals, the EST sequencing using the library originated from the RNA samples of more animals would be useful for the identification of the allelic variants expressed in liver.

Many drug-metabolizing enzyme genes are confined to gene families, many of which have been subjected to gene duplication or gene loss during evolution, resulting in family size differences [40]. This indicates that lineage-specific genes could be identified for gene families even between evolutionarily close lineages such as monkeys and humans. Moreover, physiological differences should partly result from differences at the transcriptional level, for example, by alternative splicing and non-coding RNAs [41]. Further investigation of our EST data will lead to the identification of lineage-specific transcripts generated by alternative splicing and lineage-specific gene gain/loss, as the efforts for identifying such transcripts have succeeded partly in macaques [9,13]. The identified lineage-specific transcripts and genes will help lead to a better understanding

of the physiological differences between monkeys and humans, leading to more efficient utilization of monkeys as an animal model.

References

[1] Stevens, J.C., Shipley, L.A., Cashman, J.R., Vandenbranden, M. and Wrighton, S.A. (1993) Comparison of human and rhesus monkey *in vitro* phase I and phase II hepatic drug metabolism activities. Drug Metab. Dispos. 21, 753–760.

[2] Sharer, J.E., Shipley, L.A., Vandenbranden, M.R., Binkley, S.N. and Wrighton, S.A. (1995) Comparisons of phase I and phase II *in vitro* hepatic enzyme activities of human, dog, rhesus monkey, and cynomolgus monkey. Drug Metab. Dispos. 23, 1231–1241.

[3] Guengerich, F.P. (1997) Comparisons of catalytic selectivity of cytochrome *P*450 subfamily enzymes from different species. Chem. Biol. Interact. 106, 161–182.

[4] Shimada, T., Mimura, M., Inoue, K., Nakamura, S., Oda, H., Ohmori, S. and Yamazaki, H. (1997) Cytochrome P450-dependent drug oxidation activities in liver microsomes of various animal species including rats, guinea pigs, dogs, monkeys, and humans. Arch. Toxicol. 71, 401–408.

[5] Weaver, R.J., Dickins, M. and Burke, M.D. (1999) A comparison of basal and induced hepatic microsomal cytochrome P450 monooxygenase activities in the cynomolgus monkey (*Macaca fascicularis*) and man. Xenobiotica 29, 467–482.

[6] Bogaards, J.J., Bertrand, M., Jackson, P., Oudshoorn, M.J., Weaver, R.J., van Bladeren, P.J. and Walther, B. (2000) Determining the best animal model for human cytochrome P450 activities: a comparison of mouse, rat, rabbit, dog, micropig, monkey and man. Xenobiotica 30, 1131–1152.

[7] Narimatsu, S., Kobayashi, N., Masubuchi, Y., Horie, T., Kakegawa, T., Kobayashi, H., Hardwick, J.P., Gonzalez, F.J., Shimada, N., Ohmori, S., Kitada, M., Asaoka, K., Kataoka, H., Yamamoto, S. and Satoh, T. (2000) Species difference in enantioselectivity for the oxidation of propranolol by cytochrome P450 2D enzymes. Chem. Biol. Interact. 127, 73–90.

[8] Sakate, R., Osada, N., Hida, M., Sugano, S., Hayasaka, I., Shimohira, N., Yanagi, S., Suto, Y., Hashimoto, K. and Hirai, M. (2003) Analysis of 5′-end sequences of chimpanzee cDNAs. Genome Res. 13, 1022–1026.

[9] Magness, C.L., Fellin, P.C., Thomas, M.J., Korth, M.J., Agy, M.B., Proll, S.C., Fitzgibbon, M., Scherer, C.A., Miner, D.G., Katze, M.G. and Iadonato, S.P. (2005) Analysis of the *Macaca mulatta* transcriptome and the sequence divergence between *Macaca* and human. Genome Biol. 6, R60.

[10] Li, Y. and Su, B. (2006) No accelerated evolution of 3′UTR region in human for brain-expressed genes. Gene 383C, 38–42.

[11] Osada, N., Hida, M., Kusuda, J., Tanuma, R., Iseki, K., Hirata, M., Suto, Y., Hirai, M., Terao, K., Suzuki, Y., Sugano, S. and Hashimoto, K. (2001) Assignment of 118 novel cDNAs of cynomolgus monkey brain to human chromosomes. Gene 275, 31–37.

[12] Osada, N., Hida, M., Kusuda, J., Tanuma, R., Hirata, M., Suto, Y., Hirai, M., Terao, K., Sugano, S. and Hashimoto, K. (2002) Cynomolgus monkey testicular cDNAs for discovery of novel human genes in the human genome sequence. BMC Genomics 3, 36.

[13] Chen, W.H., Wang, X.X., Lin, W., He, X.W., Wu, Z.Q., Lin, Y., Hu, S.N. and Wang, X.N. (2006) Analysis of 10,000 ESTs from lymphocytes of the cynomolgus monkey to improve our understanding of its immune system. BMC Genomics 7, 82.

[14] The Rhesus Macaque Genome Sequencing and Analysis Consortium (2007) Evolutionary and biomedical insights from the rhesus macaque genome. Science 316, 222–234.

[15] Suzuki, Y. and Sugano, S. (2003) Construction of a full-length enriched and a 5′-end enriched cDNA library using the oligo-capping method. Method Mol. Biol. 221, 73–91.

[16] Yamashita, R., Suzuki, Y., Wakaguri, H., Tsuritani, K., Nakai, K. and Sugano, S. (2006) DBTSS: database of human transcription start sites, progress report 2006. Nucl. Acid Res. 34, D86–89.

[17] Xu, X.R., Huang, J., Xu, Z.G., Qian, B.Z., Zhu, Z.D., Yan, Q., Cai, T., Zhang, X., Xiao, H.S., Qu, J., Liu, F., Huang, Q.H., Cheng, M., Li, N.G., Du, J.J., Hu, W., Shen, K.T., Lu, G., Fu, G., Zhong, M., Xu, S.H., Gu, W.Y., Huang, W., Zhao, X.T., Hu, G.X., Gu, J.R., Chen, Z. and Han, Z.G. (2001) Insight into hepatocellular carcinogenesis at transcriptome level by comparing gene expression profiles of hepatocellular carcinoma with those of corresponding noncancerous liver. Proc. Natl. Acad. Sci. USA 98, 15089–15094.

[18] Yu, Y., Zhang, C., Zhou, G., Wu, S., Qu, X., Wei, H., Xing, G., Dong, C., Zhai, Y., Wan, J., Ouyang, S., Li, L., Zhang, S., Zhou, K., Zhang, Y., Wu, C. and He, F. (2001) Gene expression profiling in human fetal liver and identification of tissue- and developmental-stage-specific genes through compiled expression profiles and efficient cloning of full-length cDNAs. Genome Res. 11, 1392–1403.

[19] Otsuka, M., Arai, M., Mori, M., Kato, M., Kato, N., Yokosuka, O., Ochiai, T., Takiguchi, M., Omata, M. and Seki, N. (2003) Comparing gene expression profiles in human liver, gastric, and pancreatic tissues using full-length-enriched cDNA libraries. Hepatol. Res. 27, 76–82.

[20] Guengerich, F.P. (2005) Human cytochrome P450 enzymes in: (Ortiz de Montellano, P., Ed.), third ed, Cytochrome P450: Structure, Mechanism, and Biochemistry, pp. 377–530, Kluwer Academic/Plenum Publishers, New York.

[21] Goldstein, J.A. (2001) Clinical relevance of genetic polymorphisms in the human CYP2C subfamily. Brit. J. Clin. Pharmacol. 52, 349–355.

[22] Mankowski, D.C., Laddison, K.J., Christopherson, P.A., Ekins, S., Tweedie, D.J. and Lawton, M.P. (1999) Molecular cloning, expression, and characterization of CYP2D17 from cynomolgus monkey liver. Arch. Biochem. Biophys. 372, 189–196.

[23] Hichiya, H., Kuramoto, S., Yamamoto, S., Shinoda, S., Hanioka, N., Narimatsu, S., Asaoka, K., Miyata, A., Iwata, S., Nomoto, M., Satoh, T., Kiryu, K., Ueda, N., Naito, S., Tucker, G.T. and Ellis, S.W. (2004) Cloning and functional expression of a novel marmoset cytochrome P450 2D enzyme, CYP2D30: comparison with the known marmoset CYP2D19. Biochem. Pharmacol. 68, 165–175.

[24] Uno, Y., Fujino, H., Kito, G., Kamataki, T. and Nagata, R. (2006) CYP2C76, a novel CYP in cynomolgus monkey, is a major CYP2C in liver, metabolizing tolbutamide and testosterone. Mol. Pharmacol. 70, 477–486.

[25] Uno, Y., Hosaka, S., Matsuno, K., Nakamura, C., Kito, G., Kamataki, T. and Nagata, R. (2007) Characterization of cynomolgus monkey cytochrome P450 (CYP) cDNAs: Is *CYP2C76* the only monkey-specific *CYP* gene responsible for species differences in drug metabolism? Arch. Biochem. Biophys. 466, 98–105.

[26] Mackenzie, P.I., Bock, K.W., Burchell, B., Guillemette, C., Ikushiro, S., Iyanagi, T., Miners, J.O., Owens, I.S. and Nebert, D.W. (2005) Nomenclature update for the mammalian UDP glycosyltransferase (*UGT*) gene superfamily. Pharmacogenet. Genomics 15, 677–685.

[27] Hayes, J.D., Flanagan, J.U. and Jowsey, I.R. (2005) Glutathione transferases. Annu. Rev. Pharmacol. Toxicol. 45, 51–88.

[28] Blanchard, R.L., Freimuth, R.R., Buck, J., Weinshilboum, R.M. and Coughtrie, M.W. (2004) A proposed nomenclature system for the cytosolic sulfotransferase (SULT) superfamily. Pharmacogenetics 14, 199–211.

[29] Cashman, J.R. and Zhang, J. (2006) Human flavin-containing monooxygenases. Annu. Rev. Pharmacol. Toxicol. 46, 65–100.

[30] Uno, Y., Kumano, T., Kito, G., Nagata, R., Kamataki, T. and Fujino, H. (2007) CYP2C76-mediated species difference in drug metabolism: A comparison of pitavastatin metabolism between monkeys and humans. Xenobiotica 37, 30–43.

[31] Belanger, G., Barbier, O., Hum, D.W. and Belanger, A. (1999) Molecular cloning, expression and characterization of a monkey steroid UDP-glucuronosyltransferase, UGT2B19, that conjugates testosterone. Eur. J. Biochem. 260, 701–708.

[32] Girard, C., Barbier, O., Turgeon, D. and Belanger, A. (2002) Isolation and characterization of the monkey *UGT2B30* gene that encodes a uridine diphosphate-glucuronosyltransferase enzyme active on mineralocorticoid, glucocorticoid, androgen and oestrogen hormones. Biochem. J. 365, 213–222.

[33] Male, D.A., Ormsby, R.J., Ranganathan, S., Giannakis, E. and Gordon, D.L. (2000) Complement factor H: sequence analysis of

221 kb of human genomic DNA containing the entire *fH*, *fHR-1* and *fHR-3* genes. Mol. Immunol. 37, 41–52.
[34] Geraghty, D.E., Daza, R., Williams, L.M., Vu, Q. and Ishitani, A. (2002) Genetics of the immune response: identifying immune variation within the MHC and throughout the genome. Immunol. Rev. 190, 69–85.
[35] Venables, J.P., Strain, L., Routledge, D., Bourn, D., Powell, H.M., Warwicker, P., Diaz-Torres, M.L., Sampson, A., Mead, P., Webb, M., Pirson, Y., Jackson, M.S., Hughes, A., Wood, K.M., Goodship, J.A. and Goodship, T.H. (2006) Atypical haemolytic uraemic syndrome associated with a hybrid complement gene. PLoS Med. 3, e431.
[36] Oyama, M., Itagaki, C., Hata, H., Suzuki, Y., Izumi, T., Natsume, T., Isobe, T. and Sugano, S. (2004) Analysis of small human proteins reveals the translation of upstream open reading frames of mRNAs. Genome Res. 14, 2048–2052.
[37] Frith, M.C., Forrest, A.R., Nourbakhsh, E., Pang, K.C., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y., Bailey, T.L. and Grimmond, S.M. (2006) The abundance of short proteins in the mammalian proteome. PLoS Genet. 2, e52.
[38] Gustincich, S., Sandelin, A., Plessy, C., Katayama, S., Simone, R., Lazarevic, D., Hayashizaki, Y. and Carninci, P. (2006) The complexity of the mammalian transcriptome. J. Physiol. 575, 321–332.
[39] Prasanth, K.V. and Spector, D.L. (2007) Eukaryotic regulatory RNAs: an answer to the 'genome complexity' conundrum. Genes Dev. 21, 11–42.
[40] Demuth, J.P., Bie, T.D., Stajich, J.E., Cristianini, N. and Hahn, M.W. (2006) The evolution of mammalian gene families. PLoS ONE 1, e85.
[41] Mattick, J.S. and Makunin, I.V. (2006) Non-coding RNA. Hum. Mol. Genet. 15, R17–R29.

BMC Genomics

BioMed Central

Research article

Open Access

Large-scale analysis of *Macaca fascicularis* transcripts and inference of genetic divergence between *M. fascicularis* and *M. mulatta*

Naoki Osada*1, Katsuyuki Hashimoto1, Yosuke Kameoka1, Makoto Hirata1, Reiko Tanuma1, Yasuhiro Uno2, Itsuro Inoue3, Munetomo Hida4, Yutaka Suzuki5, Sumio Sugano5, Keiji Terao6, Jun Kusuda1 and Ichiro Takahashi1

Address: 1Department of Biomedical Resources, National Institute of Biomedical Innovation, Ibaraki, Japan, 2Pharmacokinetics and Bioanalysis Center, Shin Nippon Biomedical Laboratories, Ltd., Kainain, Japan, 3Division of Genetic Diagnosis, Institute of Medical Science, University of Tokyo, Tokyo, Japan, 4International Research and Educational Institute for Integrated Medical Sciences, Tokyo Women's Medical University, Tokyo, Japan, 5Department of Medical Genome Sciences, Graduate School of Frontier Sciences, University of Tokyo, Tokyo, Japan and 6Tsukuba Primate Center for Medical Science, National Institute of Biomedical Innovation, Tsukuba, Japan

Email: Naoki Osada* - nosada@nibio.go.jp; Katsuyuki Hashimoto - khashi@nih.go.jp; Yosuke Kameoka - ykameoka@nibio.go.jp; Makoto Hirata - mhirata@nibio.go.jp; Reiko Tanuma - tanumark@nibio.go.jp; Yasuhiro Uno - unoxx001@pharm.hokudai.ac.jp; Itsuro Inoue - ituro@ims.u-tokyo.ac.jp; Munetomo Hida - hida@imcir.twmu.ac.jp; Yutaka Suzuki - ysuzuki@hgc.jp; Sumio Sugano - ssugano@ims.u-tokyo.ac.jp; Keiji Terao - terao@nibio.go.jp; Jun Kusuda - jkusuda@nibio.go.jp; Ichiro Takahashi - ichiro-t@nibio.go.jp

* Corresponding author

Published: 24 February 2008

BMC Genomics 2008, 9:90 doi:10.1186/1471-2164-9-90

Received: 27 September 2007

Accepted: 24 February 2008

This article is available from: http://www.biomedcentral.com/1471-2164/9/90

Abstract

Background: Cynomolgus macaques (*Macaca fascicularis*) are widely used as experimental animals in biomedical research and are closely related to other laboratory macaques, such as rhesus macaques (*M. mulatta*). We isolated 85,721 clones and determined 9407 full-insert sequences from cynomolgus monkey brain, testis, and liver. These sequences were annotated based on homology to human genes and stored in a database, QFbase http://genebank.nibio.go.jp/qfbase/.

Results: We found that 1024 transcripts did not represent any public human cDNA sequence and examined their expression using *M. fascicularis* oligonucleotide microarrays. Significant expression was detected for 544 (51%) of the unidentified transcripts. Moreover, we identified 226 genes containing exon alterations in the untranslated regions of the macaque transcripts, despite the highly conserved structure of the coding regions. Considering the polymorphism in the common ancestor of cynomolgus and rhesus macaques and the rate of PCR errors, the divergence time between the two species was estimated to be around 0.9 million years ago.

Conclusion: Transcript data from Old World monkeys provide a means not only to determine the evolutionary difference between human and non-human primates but also to unveil hidden transcripts in the human genome. Increasing the genomic resources and information of macaque monkeys will greatly contribute to the development of evolutionary biology and biomedical sciences.

Background

Genomic resources and information about primates are valuable for evolutionary and biomedical studies to determine how and why phenotypes specific to humans, as well as human diseases, have been formed. Moreover, they are important for extrapolating the results of laboratory experiments to medical research because the physiology of primates is more similar to that of humans as compared with other common experimental animals such as rodents. The cynomolgus macaque (*Macaca fascicularis*), also known as the long-tailed or crab-eating macaque, is an Old World monkey living in Southeast Asia. It is bred in laboratories worldwide and is one of the most popular primates used for laboratory animal studies, such as those on infectious diseases, immunology, pharmacology, tissue engineering, gene therapy, senescence, and learning [1]. Cynomolgus macaques, rhesus macaques (*M. mulatta*), and Japanese macaques (*M. fuscata*) are widely used for experimental studies and are closely related to each other [2-4]. The US government funded genome sequencing of the rhesus macaque because it is the most common laboratory animal bred in the US, and in 2007, the draft sequence of the rhesus macaque was published [5].

Since cynomolgus and rhesus monkeys are very closely related at the genetic level, we aim to determine the extent to which the rhesus macaque genome sequence can be used as a reference for biomedical studies involving cynomolgus macaques. At the chromosomal level, a previous study suggested that a pericentric chromosome inversion occurred in the cynomolgus lineage after splitting from rhesus macaques [6]. At the nucleotide sequence level, the genetic divergence between cynomolgus and rhesus monkeys has been measured using mitochondrial DNA sequences [2,3] or a limited number of loci on the chromosomes [4,7]. Thus, the divergence of a sufficient number of loci between cynomolgus and rhesus macaques would assist in determining the degree of genetic divergence between them. In addition, recent studies have shown that there is a considerable amount of genetic diversity within the species themselves [5-10], which also hampers the measurement of the genetic divergence. Because the divergence between the two macaques is very recent (much later than the divergence between humans and chimpanzees), we must consider the segregation of polymorphisms in the common ancestral population to estimate the correct species divergence time [11,12]. By analyzing the number of loci in the two species, we can determine the history of divergence between them, including the ancestral population size, divergence time between species, and possible gene flow [13,14].

We have constructed full-length-enriched cDNA libraries from cynomolgus monkey brain, testis, and liver using the oligo-capping method. Many comparative genomics projects have focused on sequencing of the genome or expressed sequenced tags (ESTs), and full-length cDNA sequences are uniquely informative resources for accurately predicting the full structure of transcripts in the genome [15]. Furthermore, because cynomolgus and rhesus macaques are very closely related, transcriptome data from cynomolgus macaques is useful for annotating the genome sequence of other macaques whose transcriptome data is less than 1% of that from humans and whose full-length cDNA data is scarce.

Along with the cynomolgus macaque cDNA sequencing project, we have published a part of our results, such as novel gene findings [16-19], search for fast-evolving genes [19], molecular evolution of 5'-untranslated regions (UTRs) [20], and evolution of brain-expressed genes [21]. In this study, we summarize the final sequencing project and present novel findings with an expanded dataset. In total, 85,721 ESTs and 9407 full-length sequences were determined, annotated, and stored in an in-house database and the public databases (DDBJ/EMBL/GenBank). Our study focused on the divergence between the cynomolgus and rhesus macaque genes. We did not intensively analyze the divergence between humans and cynomolgus monkeys, because a study on rhesus genome has investigated this thoroughly [5]; it also identified and discussed positively selected genes or extensively duplicated genomic regions during the evolution of *Catarrhine* primates.

Results

Summary of cDNA sequences

We constructed several oligo-capped cDNA libraries from cynomolgus monkey testis, liver, and seven anatomical parts of the brain (cerebellar cortex, parietal lobe, occipital lobe, frontal lobe, temporal lobe, medulla oblongata, and brain stem). The oligo-capping method selectively amplifies full-length cDNAs with a cap structure and poly(A) tail [22]. We sequenced the 5'- or 3'-end of 85,721 clones, yielding 63,395 and 22,326 sequences of 5'- and 3'-ESTs, respectively, after filtering the vector and low quality sequences. These EST sequences grouped into 16,466 clusters with 11,016 singletons (BLAST e-value: 1e-30). We classified them based on homology to the 26,575 non-redundant human RefSeq sequences (see methods). Of the 85,721 EST sequences, 68,257 (80%) were homologous to the human RefSeq gene set and were clustered into 9065 types of genes, indicating that our EST sequences would cover about 34% of the known human transcripts (Table 1). In particular, when we limited the human reference genes to the validated protein-coding genes (*i.e.*, RefSeq accession beginning with NM), 47% of the human reference genes were represented in the macaque cDNAs.

Table 1: Summary of cDNA clones

Library	# of isolated clones	# of full-sequenced clones
Brain: Parietal Lobe (QnpA)	8063 (5890)	649 (336)
Brain: Frontal Lobe (QflA)	13,215 (9286)	2493 (1768)
Brain: Temporal Lobe (QtrA)	6797 (6039)	1078 (862)
Brain: Occipital Lobe (QorA)	5458 (4518)	634 (606)
Brain Stem (QbsA, B)	2776 (1993)	359 (301)
Brain: Medulla Oblongata (QmoA)	4485 (3645)	1146 (912)
Brain: Cerebellar Cortex (QccE)	11,734 (9028)	731 (563)
Testis (QtsA)	10,867 (8510)	2316 (2175)
Liver (Qlv)	22,326 (20,833)	0 (0)
Total	85,721 (69,742)	9407 (7523)
Averaged Length		1882 bp

*Numbers of genes with the RefSeq homologs are shown in parentheses.

In parallel to EST sequencing, we determined about 9500 full-insert sequences of the cDNA clones. About 2500 clones whose 5'-EST sequences were not homologous to the public cDNA sequences and 7000 clones whose 5'-EST sequences were homologous to the human RefSeq sequences were chosen [16-21]. Out of the 9407 full-insert sequences, 7407 sequences were homologous to 5384 types of human genes (Table 1). The averaged length of the full-insert sequences was 1864 bp, excluding the length of the poly(A) tail. The macaque sequences were annotated for gene function and homologous locus in the human genome using information from the Entrez Gene [23] and Gene Ontology (GO) databases [24].

Database construction

All cDNA sequences and annotations were deposited in the public databases and stored in a simple in-house database, QFbase [25]. On the QFbase website, users can search the macaque clones by keywords and BLAST searches. For each human gene, the distribution of the macaque homologs is represented graphically and users can easily retrieve information of the objective macaque cDNA clones. The entries are further linked to the gene annotation in the outside databases, GenBank [26], Ensembl [27], OMIM [28], and H-InvDB [29]. The cDNA sequences were mapped on the human and rhesus genome sequences using the UCSC genome browser [30]. Moreover, 4665 human-macaque orthologous alignments are provided in the QFbase. For each alignment, the non-synonymous substitution rate (*Ka*) and the synonymous substitution rate (*Ks*) between the human and macaque cDNA sequences were estimated. Non-synonymous substitutions are nucleotide changes that replace amino acids between species whereas synonymous substitutions cause no amino acid changes. The relative pace of protein evolution was thus determined using *Ka*/*Ks*, assuming that the *Ks* value reflects the neutral mutation rate [31]. Using the database, users can sort the alignments according to the *Ka* and *Ks* values. For example, users can determine the *Ka* and *Ks* values of a particular gene or view the list of the 100 most rapidly evolved genes between humans and cynomolgus monkeys. The cDNA clones are distributed through the Human Science Research Resource Bank in Japan (Tokyo, Japan). Further information is available at the QFbase website.

Analysis of unidentified transcripts

Of the 9407 full-sequenced cDNAs, about 2000 were not homologous to the human reference gene sequences (RefSeq, built on Sep 14, 2006; BLAST: E = 1e-60). These Non-RefSeq transcripts clustered into 1245 non-redundant transcripts, which were further classified as shown in Figure 1. The list of the Non-RefSeq transcripts is provided in Additional file 1. We filtered 11 junk sequences and 210 known transcripts. The 210 transcripts matched with the unannotated human cDNA sequences in the database and were called as orphan transcripts. These may help in further annotation of the human genome.

After removing the junk sequences and orphan transcripts, the remaining 1024 transcripts were referred as the unidentified transcripts although 40% (406/1024) of the transcripts showed homology to human ESTs (BLAST: E = 1e-60), because no full cDNA sequence of humans has been registered in the public databases. One of the advantages of full-length cDNAs is that we can determine the splicing pattern and reading direction of the transcripts in the genome. We categorized the unidentified transcripts as anti-transcript, intronic spliced transcript, intronic single-exon transcript, intergenic spliced transcript, or intergenic single-exon transcript. Among the intergenic transcripts, 82 were located within 5 kb of the genic regions with the same direction as the genes. Of these, 6 were mapped on the upstream regions and 76 were mapped on the downstream regions of the known genes. The result showed they may be hidden extensions of the

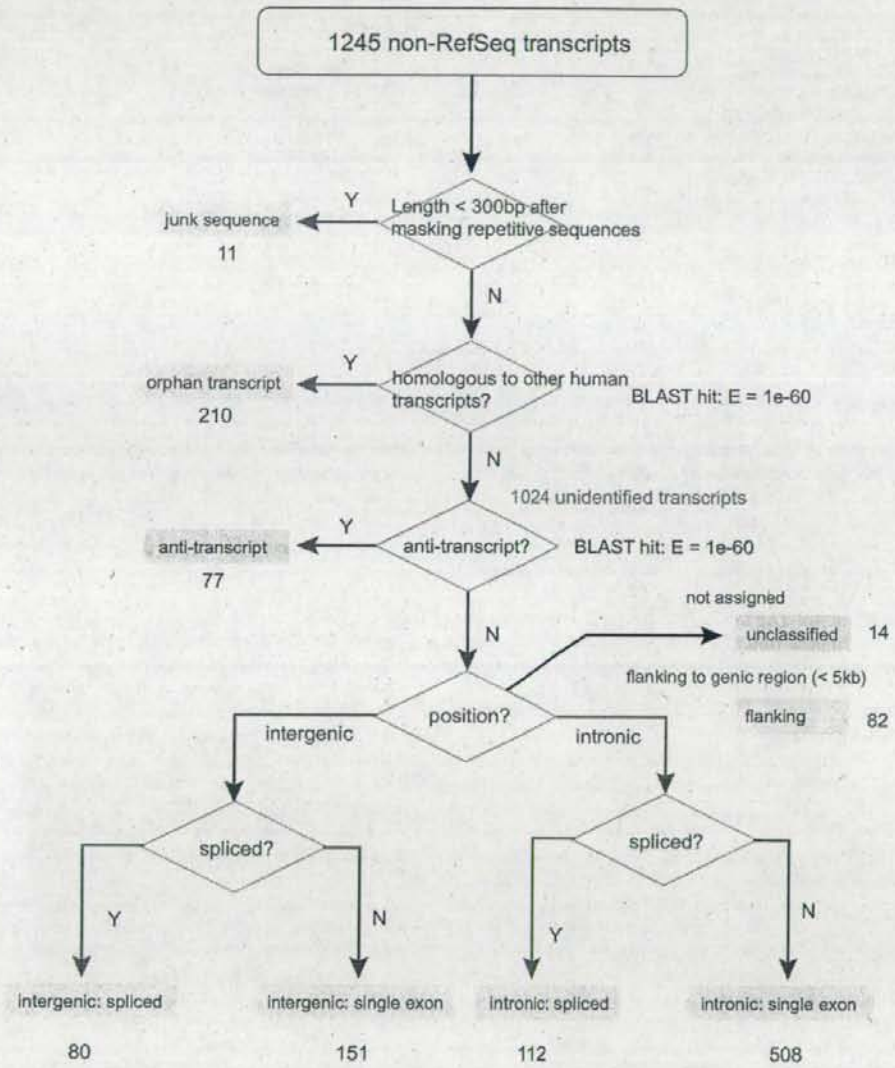

Figure 1
Classification of the 1245 Non-RefSeq transcripts. Transcripts shorter than 300 bp after masking the repetitive sequences were categorized as junk sequences. The remaining sequences were BLAST-searched against all public human cDNA sequences for the forward strand. Homologous sequences to the unannotated human cDNAs were classified as orphan transcripts for the forward strand and anti-transcript for the reverse strand. The remaining 947 clones were mapped on the human genome sequence and arranged according to the annotation from the UCSC genome browser (hg18). The transcripts that overlapped with the genic regions including UTR were classified as intronic transcripts, and the transcripts that were mapped more than 5 kb away from the genic region were classified as intergenic transcripts.

known transcripts, using alternative promoters and/or poly(A) signals in the human genome. These sequences were filtered from the intergenic transcripts and classified as 'flanking' to genic regions. The largest group was the intronic single-exon transcripts. Although they might be acquired from premature mRNA molecules in the cell nucleus, recent studies have revealed the potential abundance of short intronic transcripts in the human genome [32]. Among these classes, anti-transcripts and intergenic spliced transcripts are the most biologically relevant classes, which are unlikely to be derived from contamination by premature mRNAs.

We designed oligonucleotide microarrays (Affymetrix GeneChip) containing probes complementary to the known genes and unidentified transcripts. Hybridizations were performed using the RNA sampled from a 3-year-old macaque cerebrum, cerebellum, liver, and testis with duplications. The significance of expression was determined using Affymetrix MAS5.0 software [33] (see methods). The proportion of the expressed transcripts is presented in Figure 2. In the unidentified transcripts, 544 transcripts were expressed in at least one of the four tissues ($P < 0.05$; Table 2). Because all the unidentified transcripts were isolated from the macaque brain or testis, fewer transcripts were expressed in the liver (14%) than in the cerebrum (31%), cerebellum (41%), and testis (24%). The expressed proportion of the unidentified transcripts was significantly smaller than that of 8428 RefSeq homologs (51% and 81%, respectively; $P < 10\text{-}15$; Fisher's exact test). The orphan transcripts were expressed in an intermediate proportion (72%). The percentages of the expressed unidentified transcripts ranged from 33% to 57% (Fig. 1). A large difference was observed between the intergenic and intronic transcripts; more intronic transcripts displayed significant expression on the microarrays than intergenic transcripts ($P = 0.0005$; Fisher's exact test).

Previous studies have shown that many unannotated transcripts are not conserved at a DNA sequence level in many organisms [34]. In practice, sequence conservation is determined by investigating whether the region is alignable. Here, we directly measure the difference in the DNA sequences between humans and macaques. For protein-coding genes, previous studies have shown large disparities in sequence divergence between brain- and testis-expressed genes, both in the CDS and UTR, owing to the stronger functional constraint on the brain-expressed genes [20,21]. We further inquired whether the trend was observed in the unidentified transcripts. We classified the transcripts into brain-expressed transcripts (expressed in the cerebrum and not in the testis) and testis-expressed transcripts (expressed in the testis and not in the cerebrum). As shown in Figure 3, while the non-synonymous substitution rates of the RefSeq homologs were higher in the testis than in the brain, the DNA sequence divergence

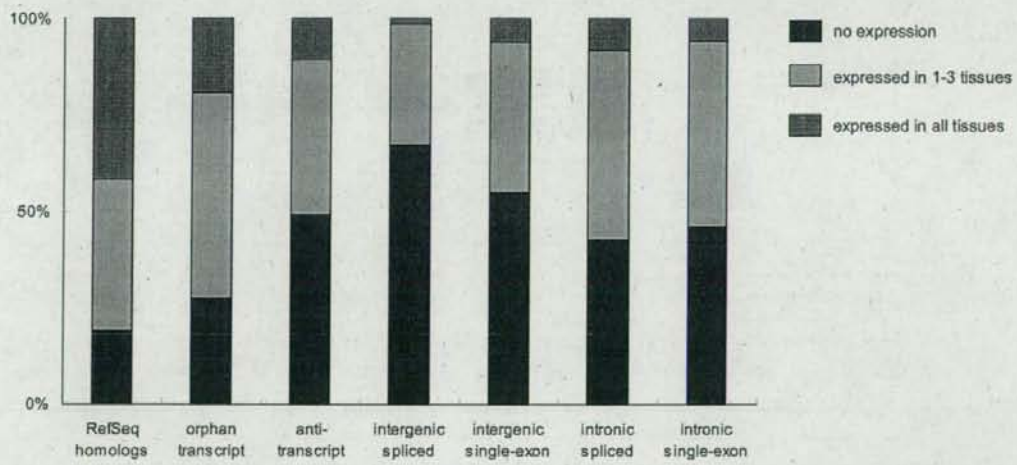

Figure 2
The proportion of the expressed transcripts in the RefSeq homologs (control) and unidentified transcripts. Cerebrum, cerebellum, liver, and testis of a male macaque were used for the microarray experiments with duplicated hybridizations. The transcripts were classified into no expression (blue), expressed in 1–3 tissues (grey), or expressed in all tissues (red).

Table 2: Number of expressed transcripts in the unknown macaque transcripts

	Unidentified transcripts[a]	Intergenic transcripts[b]
Cerebrum	321	54
Cerebellum	417	58
Liver	139	13
Testis	241	52
All tissues	74	10
Any tissue	544	137
Total	1024	231

[a]Transcripts that have no homology to the public human cDNA sequences.
[b]Transcripts that were mapped more than 5 kb away from the annotated genic regions on the human genome (see Fig. 1).

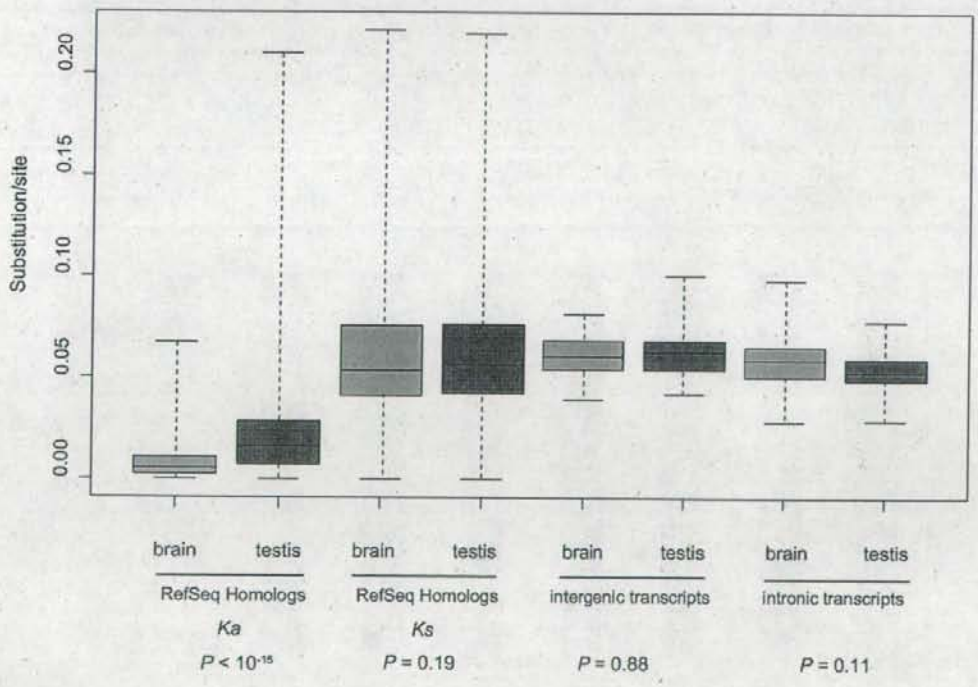

Figure 3
Sequence conservation of the brain-expressed and testis-expressed transcripts between humans and macaques. For the RefSeq homologs (control), the non-synonymous (K_a) and synonymous (K_s) substitution rates were estimated using the Li-Pamilo-Bianchi method [48]. The substitution rates in the intergenic and intronic transcripts were estimated using Kimura's two parameter methods [55]. The heights of the boxes represent the lower and upper quartile points, and the whiskers show the minimum and maximum points.

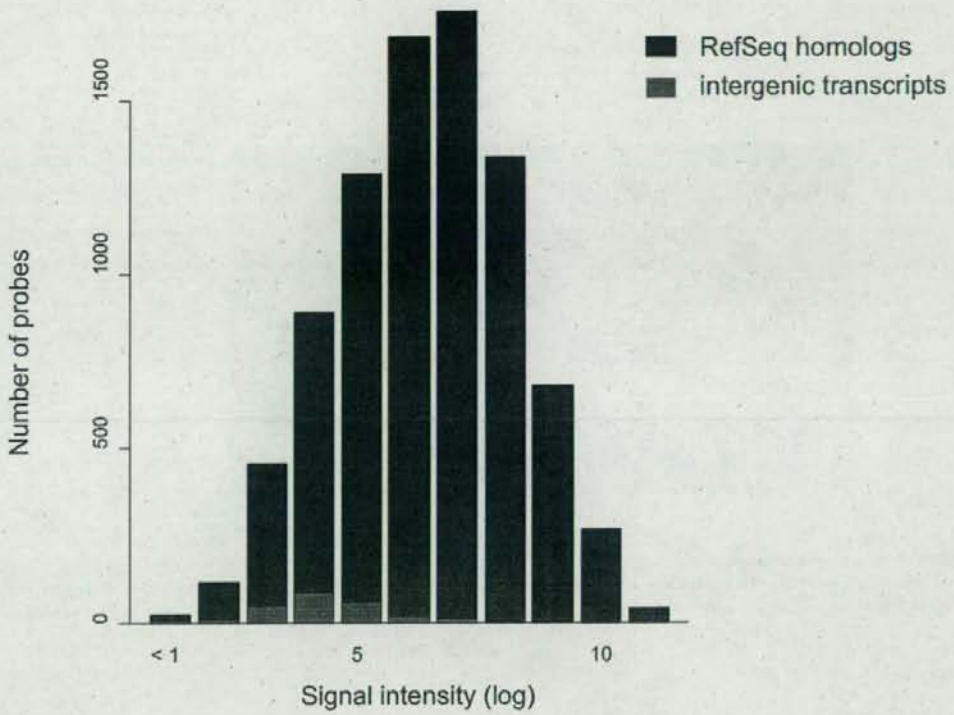

Figure 4
Distribution of transcript expression levels of the RefSeq homologs (blue) and the intergenic transcripts (red). Only the transcripts that were determined as significantly expressed on the microarray are presented in the figure. Log-transformed signal intensity in the tissue with the highest expression was shown. The intergenic transcripts showed significantly lower expression levels than the RefSeq homologs.

of the unidentified transcripts was not associated with the expression pattern. Furthermore, there was no evidence that the unidentified transcripts were more conserved than the synonymous sites of the RefSeq homologs.

We further evaluated the expression level of the 231 intergenic transcripts. We collected the strongest signal intensity of the significantly expressed intergenic transcripts. As shown in Figure 4, even if they were significantly expressed, signal intensities of the intergenic transcripts were significantly weaker than those of the control genes ($P < 10\text{-}13$; Wilcoxon test). Weak expression levels of intergenic sequences have been previously reported [35,36] and these may cause weak detection levels of the intergenic transcripts. To test the reproducibility of the microarray experiments using another method, we selected eight intergenic spliced transcripts and tried to amplify human and macaque transcripts using RT-PCR. We designed the PCR primers that would match both human and macaque sequences and would amplify introns of genomic sequences when the genomic DNA is contaminated. A gel picture of the RT-PCR products is shown in Figure 5. Two transcripts showed positive results, while six showed negative results on the microarray. We confirmed the expression of the two transcripts in the macaque brain using both the microarray and RT-PCR. Furthermore, even though we failed to detect the expression of the six transcripts on the arrays, we recov-

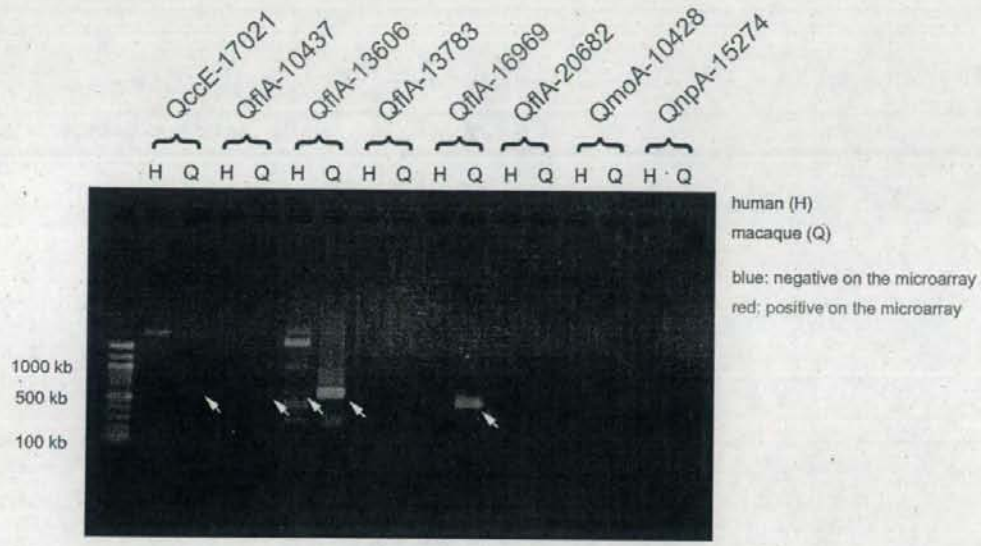

Figure 5
RT-PCR gel images for the expression of the intergenic transcripts in the human (H) and the macaque (Q) brain. Transcript names indicate whether the expression was detected by the microarray experiments (red) or not (blue). Expected PCR products are marked by the white arrows.

ered the expression of the two transcripts by RT-PCR. In these two transcripts, the expression levels detected by RT-PCR resulted in considerably weaker bands on the gel (Fig. 5), indicating that the microarray failed to capture their expression at a very low level. In total, we detected the expression of four transcripts in the macaque brain. Of these four transcripts, two were not detected and one was transcribed in an unspliced form in humans. The other showed multiple extra bands in both humans and macaques. Overall, the expression of the macaque intergenic spliced transcripts was not well conserved between the human and the macaque brain.

Hidden transcript structures in the human genome

Of the 9407 macaque cDNA sequences, 2261 covered the entire CDS of the human RefSeq genes in a single BLAST hit chain. In the 2261 cDNAs, we sought a stretch of UTR sequences (> 50 bp) that did not match any homologous human cDNA sequence. Simple genomic insertion or deletion in the genome was not counted. After filtering the ambiguous entries, in the UTR of macaque cDNAs, we found 262 exons that were not found in the human cDNA data. Out of the 262 unidentified exons, 85 (32%) did not match any human EST sequence. We classified the unidentified UTRs as follows: (A) extended exons and (B) novel exons (Fig. 6). Those unidentified exons were further classified into internal and external exons (Fig. 6). As shown in Figure 1, the distribution of the different types of unidentified exons was not uniform; most of them were external exons.

Because the human transcriptome data is more complex than previously thought, as revealed by genome tiling DNA microarrays [34-36], these unrepresented exons may be expressed at a very low level in human tissues. Moreover, these exons have not been found in the conventional cDNA exploration methods. However, previous studies have suggested a frequent evolutionary turnover of exon sequences [37]. The evolutionary alteration of external exons in the 5'-UTR may be caused by the alternative usage of promoter sequences [38]. The evolutionarily altered exons in the 3'-UTR may be caused by the alternative usage of poly(A) adenylation signals [39]. All the unidentified exons are provided in Additional file 2.

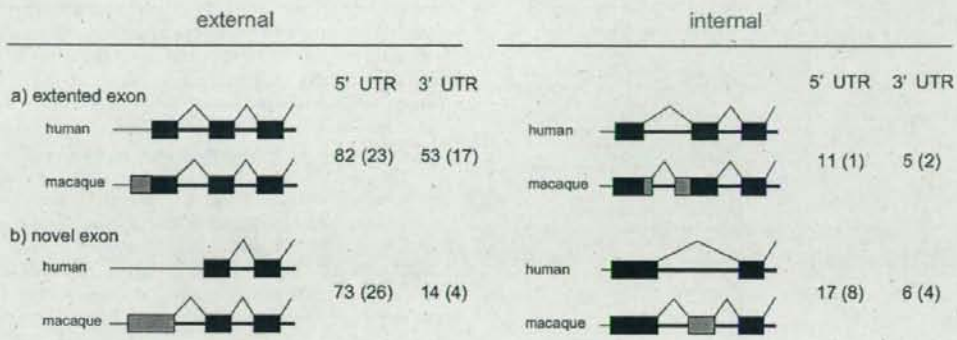

Figure 6
Pattern of the unidentified exons. The closed boxes represent exons in the genomes. Unidentified exons in macaques are presented as blue boxes. Intergenic regions and introns are depicted by thick and thin horizontal lines, respectively. (A) extended exons. (B) novel exons. These exons were further classified into internal (right panel) and external (left panel) exons. The number of genes in each category is shown on the left of each schema. The number of unidentified exons that have not been found even in the EST sequences is shown in parentheses.

Comparison of the human, cynomolgus, and rhesus genes
We compiled 2655 human-rhesus-cynomolgus cDNA alignments (dataset I) using the rhesus macaque genome and the predicted transcript sequences. The phylogenetic relationship among the three species is shown in Figure 7. Because the rhesus and cynomolgus genomes are very similar, we wanted to minimize non-orthologous alignments, which inflate the average and variance of the nucleotide divergence between them. Therefore, the macaque genes showing > 80% homology to more than one locus in the rhesus genome were filtered (dataset II). Although the number of genes analyzed was reduced to 1499 in the second dataset, the subset of the genes would be useful in estimating the divergence among the three species. The results were obtained using dataset II in the following manner. The results using the unfiltered dataset (dataset I), which resulted in the inflation of variance, are provided in Additional file 3. Genes that have evolved under positive selection were searched with the model-based likelihood ratio test [40]. In total, 39, 15, and 22 genes showed evidence of positive selection in the human, cynomolgus, and rhesus lineages, respectively ($P < 0.05$). Thirty-eight genes also showed a positive selection signature between the two macaques and 74 were detected in all the three lineages (Table 3). Note that, in Figure 7, the phylogenetic tree is unrooted. The list of positively selected genes is provided in Additional file 4. Excluding the overlapped genes, we identified 101 out of 1499 genes (6.7%) that underwent positive selection in any lineage at 5% significance level. The number of positively selected genes in each of the two macaque lineages was comparable to that estimated in the human-chimpanzee lineages using the same method [41]. Although these candidates of positively selected genes contain many biologically interesting functions, such as transcriptional regulation (*RELA*, *ZNF263*, and *L3MBTL4*), visual perception (*RGS9*, *GPRC5B*, and *RPGRIP1*), and mitochondrial localization (*PET112L*,*VARS*,*ACAA2*,*YARS2*, *FOXRED1*, and *COQ9*) [19], none of the GO categories were statistically overrepresented probably because of the small sample size.

Genetic divergence between cynomolgus and rhesus macaques
Using the above dataset, we estimated the nucleotide substitution rates of each lineage from the common ancestor of cynomolgus and rhesus macaques (presented in Table 4). Numbers and rates of the non-synonymous and synonymous substitutions for each lineage were estimated using the maximum likelihood method. We assumed that synonymous substitutions are nearly neutral and used

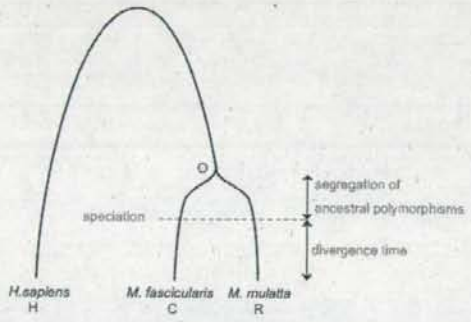

Figure 7
Genealogical relationship (phylogeny of genes) among the humans (H), cynomolgus macaque (C), and rhesus macaque (R). The common ancestor of the two macaques is indicated by the letter O. The time of speciation between the two macaques is shown by the dashed line. Note that the tree is unrooted.

them to estimate the divergence time. If we set the divergence of humans and Old World monkeys to 25–35 million years ago (Mya) [42], genes of the two macaques would be considered to diverge 1.9–2.6 Mya on average. However, since the two species diverged very recently, we have to consider the ancestral polymorphisms segregating at the time of speciation [11,12]. Figure 7 illustrates the impact of the ancestral polymorphisms on the estimation of the divergence time. Suppose that the common ancestor of two macaques had the same population size as the population size of extant chimpanzees, *i.e.*, 1–2 Mya to the most recent common ancestor [43]. In this situation, only one Mya divergence is assigned to the actual divergence time of the two species. Therefore, without considering ancestral polymorphisms, we tend to overestimate the true species divergence. We applied the maximum likelihood method to estimate the divergence time and ancestral population size of the rhesus and cynomolgus monkeys. As a result, we obtained $2tu = 0.00213 \pm 0.00022$ and $4Neu = 0.00327 \pm 0.00025$ where t, Ne, and u represent the divergence time after speciation, ancestral population size at speciation, and mutation rate with standard errors, respectively. We also noticed that a non-negligible number of nucleotide substitutions were erroneously assigned to the cynomolgus macaque lineage owing to PCR errors in the cDNA libraries. Therefore, the actual substitution rate was estimated by correcting the PCR error using a computational method (see methods). After the correction, we obtained $2tu = 0.00181 \pm 0.00021$ and $4Neu = 0.00311 \pm 0.00024$ (Table 4). If we consider $u = 10\text{-}9$ (nucleotide substitution rate per year of Old World monkeys [44]), the divergence time between the two macaques would be estimated as 0.91 ± 0.11 Mya with a standard error. We also estimated the ancestral population size to be $43{,}200 \pm 3300$ with a mutation rate per generation of $1.8 \times 10\text{-}8$ in humans [45]. The result suggests that more than a half of the genetic divergence between the two macaques is derived from the ancestral polymorphisms.

Discussion

In summary, the sequencing project of cynomolgus monkey cDNAs yielded 85,721 ESTs and 9407 full-length sequences. Since our project mainly studied the brain and testis, the dataset is deficient in other tissue-specific genes, e.g., the genes related to the immune system that many medical researchers would want to explore [46]. The construction of cDNA libraries from other tissues and EST sequencing is still ongoing to complement the transcriptome of cynomolgus monkeys. The latest sequencing status can be confirmed from the website. Because of the close relationship between the cynomolgus and rhesus macaques, cDNA resources of cynomolgus macaques not only are useful for research using cynomolgus macaques but also complement the relative paucity of the transcriptome data from rhesus macaques. Using macaque tissues to scan the primate transcriptome is advantageous because RNA molecules are unstable and are instantly degraded in the tissues during sampling. This causes serious problems for RNA sampling from human tissues, especially in the brain, where fresh samples are rarely obtainable. Therefore, we hope to uncover rare transcripts

Table 3: Number of genes under positive selection out of 1499 non-duplicated genes determined using the branch-site test of positive selection

Lineage[a]	$P \leq 0.05$	$P \leq 0.01$
H-O	39	14
C-O	15	10
R-O	22	21
Between the macaques (C-O + R-O)	37	32
All lineages (H-O + C-O + R-O)	74	33

[a]H: human; C: cynomolgus macaque; R: rhesus macaque; O: cynomolgus-rhesus ancestor (see Fig. 7).

Table 4: Divergence among the human, cynomolgus, and rhesus genes

Model without ancestral polymorphisms (Raw data)		
Lineage[a]	K_a (± S.E.)	K_s (± S.E.)
H-O	1.06×10^{-2} (3.20×10^{-4})	6.82×10^{-2} (1.07×10^{-3})
C-O	1.02×10^{-3} (4.79×10^{-5})	3.04×10^{-3} (1.20×10^{-4})
R-O	4.98×10^{-4} (3.36×10^{-5})	2.50×10^{-3} (1.15×10^{-4})
Model with ancestral polymorphisms		
	$2tu$[b] (± S.E.)	$4N_e u$[b] (± S.E.)
Raw data	2.13×10^{-3} (2.24×10^{-4})	3.27×10^{-3} (2.52×10^{-4})
PCR error corrected	1.81×10^{-3} (2.12×10^{-4})	3.11×10^{-3} (2.40×10^{-4})

[a]H: human, C: cynomolgus macaque, R: rhesus macaque, O: cynomolgus-rhesus ancestor (see Fig. 7).
[b]t: time after speciation; u: mutation rate; N_e: effective population size of the cynomolgus-rhesus ancestor.

that would be hidden in the human transcriptome data. In this study, we identified 1024 macaque cDNAs that were not represented in the public human cDNA sequences. Although 51% of the cDNA did not show a positive signal on the microarrays, the following RT-PCR experiments recovered the expression in half (3/6) of the transcripts. The results indicate that these unidentified transcripts were expressed at a low level in the tissues even though the microarray could not detect the expression.

The *M. fascicularis* oligonucleotide microarrays contain probes that matched 8316 known genes and 1024 unidentified transcripts. We determined the number of probe sets for the known genes that overlapped among the commercially available microarray (Affymetrix GeneChip) and previously published microarray of rhesus macaque by Wallace et al., which contains the largest number of probe sets among the published microarrays [47]. Of our 8316 probes for the known genes, 1728 (21%) were not represented in the commercial microarray and 1091 (13%) were not found in the published microarray. Combining the three microarrays, 417 probes for the known genes were represented only in the *M. fascicularis* microarrays [see Additional file 5]. In our preliminary study of the polymorphisms within cynomolgus macaques, we found that the level of polymorphisms in cynomolgus macaques was greater than that in rhesus macaques and slightly smaller than the level of divergence between rhesus and cynomolgus macaques (Osada et al., unpublished data). Therefore, even if we should be careful about sequence mismatches within and between species, the information from both macaque transcripts and the rhesus genome can be combined to build more versatile and comprehensive DNA microarrays that can be used for biomedical surveys using laboratory macaques.

Suppose that we identify positively selected genes in the human lineage after the spilt from chimpanzees. Such genes are useful for understanding the human-specific physiology only when those genes have not been under positive selection in other primate lineages. We identified 37 genes under positive selection between the two macaques at 5% significance level. None of these genes were shared with 387 genes under positive selection in the human or chimpanzee lineages previously determined from the whole genome scan [41], providing support that the method has correctly identified positively selected genes in the specific lineages.

For estimating of the divergence time between cynomolgus and rhesus macaques, we assumed that there is no gene flow between the ancestral species throughout their speciation and divergence time (*i.e.*, allopatric model). However, considering the ancestral polymorphisms and the PCR error rate, we estimated the divergence time to be around 0.9 Mya, which is less than the estimation of the age of MRCA of rhesus macaques [10]. Indeed, more than half of the genetic divergence between the two macaques was derived from ancestral polymorphisms. If continuous gene flow is present during speciation, the variance component would be inflated and we would tend to overestimate the amount of ancestral polymorphisms [14].

In this analysis, we used the rhesus macaque genome sequence to represent rhesus macaques. We should note that the rhesus macaque used for genome sequencing was an Indian rhesus macaque; these macaques have genetically differentiated from Chinese rhesus macaques [9]. In addition, our samples of cynomolgus macaques were obtained from different geographic subpopulations. Previous studies using mitochondrial DNA sequences [10] and our preliminary analysis using nuclear DNA sequences (Osada et al., unpublished data) showed that there is a substantial genetic divergence between cynomolgus monkeys of Sundaland (Indonesia and Indochina) and Philippine populations. Therefore, our phylogenetic inference using two sampled sequences has a technical limitation and may be accurate only if there are no complex population structures among the ancestral cynomolgus and rhesus macaque populations. Elucidat-

ing the polymorphisms and divergence among macaque species would provide further insight into the evolutionary history of macaques and benefit biomedical research using macaque monkeys.

In Table 4, without correcting the PCR error rate, both the non-synonymous and synonymous divergences are greater in the cynomolgus lineage. This may be due to shorter generation time and smaller population size of cynomolgus monkeys. However, a more reasonable explanation is that the cDNA sequences of cynomolgus monkeys might incorporate the errors resulting from PCR amplification during the construction of the oligo-capped cDNA libraries. The synonymous substitution rate in the cynomolgus lineage is about 0.0005 points higher than that in the rhesus lineage, and the non-synonymous substitution rate differs in about 0.0004 points. Assuming that the selective constraint and generation time of the two macaque lineages are the same, excess divergence of 0.04%-0.05% in the cynomolgus lineage may be an artifact introduced by PCR amplification, which is fairly close to the estimation from the experiment by Suzuki and Sugano [48]. If we reflect the substitution rate in the rhesus lineage to that in the cynomolgus lineage for correcting the errors, the total divergence of the two macaques will be reduced to about 90% (Table 4).

Conclusion

Transcript data from Old World monkeys provide us with means to determine not only the evolutionary difference between human and non-human primates but also the hidden transcripts in the human genome. Actual cDNA clones of macaques are also indispensable resources for genetic engineering studies. It is considered that the species divergence between rhesus and cynomolgus macaques would be much later than the previous estimates, and the speciation process between them might have been complex. To use laboratory macaques more efficiently, we need to be more aware of the genetic difference within and among macaque monkeys. Increasing the genomic resources and information of macaque monkeys will greatly contribute to the development of evolutionary biology and biomedical sciences.

Methods

Cynomolgus monkey samples

Samples from two cynomolgus monkeys, a 16-year-old female (Philippine origin) and a 15-year-old male (Cambodian-Thai hybrid), were used for the cDNA libraries, except for the liver cDNA library (Qlv). The liver samples were collected from three adult cynomolgus monkeys of unknown origin. The monkeys were cared for and handled according to the guidelines established by the Institutional Animal Care and Use Committee of the National Institute of Infectious Diseases (NIID) of Japan and the standard operating procedures for monkeys at the Tsukuba Primate Center, NIID (present National Institute of Biomedical Innovation), Tsukuba, Ibaraki, Japan. Tissues were excised in accordance with all the guidelines in the Laboratory Biosafety Manual, World Health Organization, at the P3 facility for monkeys of the Tsukuba Primate Center. Immediately after collection, the tissues were frozen in liquid nitrogen and used for RNA extraction. Oligo-capped cDNA libraries were constructed according to the method described previously [48]. The prefix in each clone name represents the location of the source of the tissue: Qnp (brain, parietal lobe), Qfl (brain, frontal lobe), Qtr (brain, temporal lobe), Qor (brain, occipital lobe), Qbs (brain stem), Qmo (medulla oblongata), Qcc (cerebellar cortex), Qlv (liver), and Qts (testis).

Sequencing of cDNA clones

The cDNA clones were sequenced with ABI 3700 and 3730 automated sequencers. The EST sequences were trimmed to avoid the vector sequence of pME18-FL3 [DDBJ/EMBL/GenBank: AB009864]. Entire sequences of the clones were determined by the primer walking method. The repeat sequences at the 5'- and 3'-ends were masked using the Repbase Update database [49] before BLAST search. The BLAST search was performed with an e-60 cut-off value against non-redundant human RefSeq data. The non-redundant data was based on the annotation in the Ensembl Gene database. The longest transcript in the locus was selected as the representative cDNA [24,50]. The macaque cDNA sequences were deposited in the public DNA databases [DDBJ/EMBL/GenBank: CJ430287–CJ493524; BB873801–BB894695; AB303966–AB303967].

Classification of unidentified transcripts

Classification of the Non-RefSeq transcripts was performed as shown in Figure 1. Transcripts shorter than 300 bp after masking the repetitive sequences were categorized as junk sequences. The remaining sequences were BLAST-searched against all public human cDNA sequences (downloaded on Aug 3, 2007) for the forward strand. Homologous sequences to the human cDNAs were classified as orphan transcripts for the forward strand and anti-transcript for the reverse strand. The remaining 947 clones were mapped on the human genome sequence (build 36.1) by BLAST algorithm and arranged according to the annotation from the UCSC genome browser (hg18). The transcripts that overlapped with the genic regions including UTR were classified as intronic transcripts, and the transcripts that were mapped more than 5 kb away from the genic region were classified as intergenic transcripts.

Expression assays

Affymetrix GeneChip was designed using the available cDNA sequences of *M. fascicularis*. The chip loads 10,307