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.**
© 2007 Federation of European Biochemical Societies. Published by Elsevier B.V. All rights reserved.

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′-TTTTTTTTTTTTT-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 © 2007 Federation of European Biochemical Societies. Published by Elsevier B.V. All rights reserved.
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 (27 959 entries) were further analyzed. Of these ESTs, 14 727 (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 27 959 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

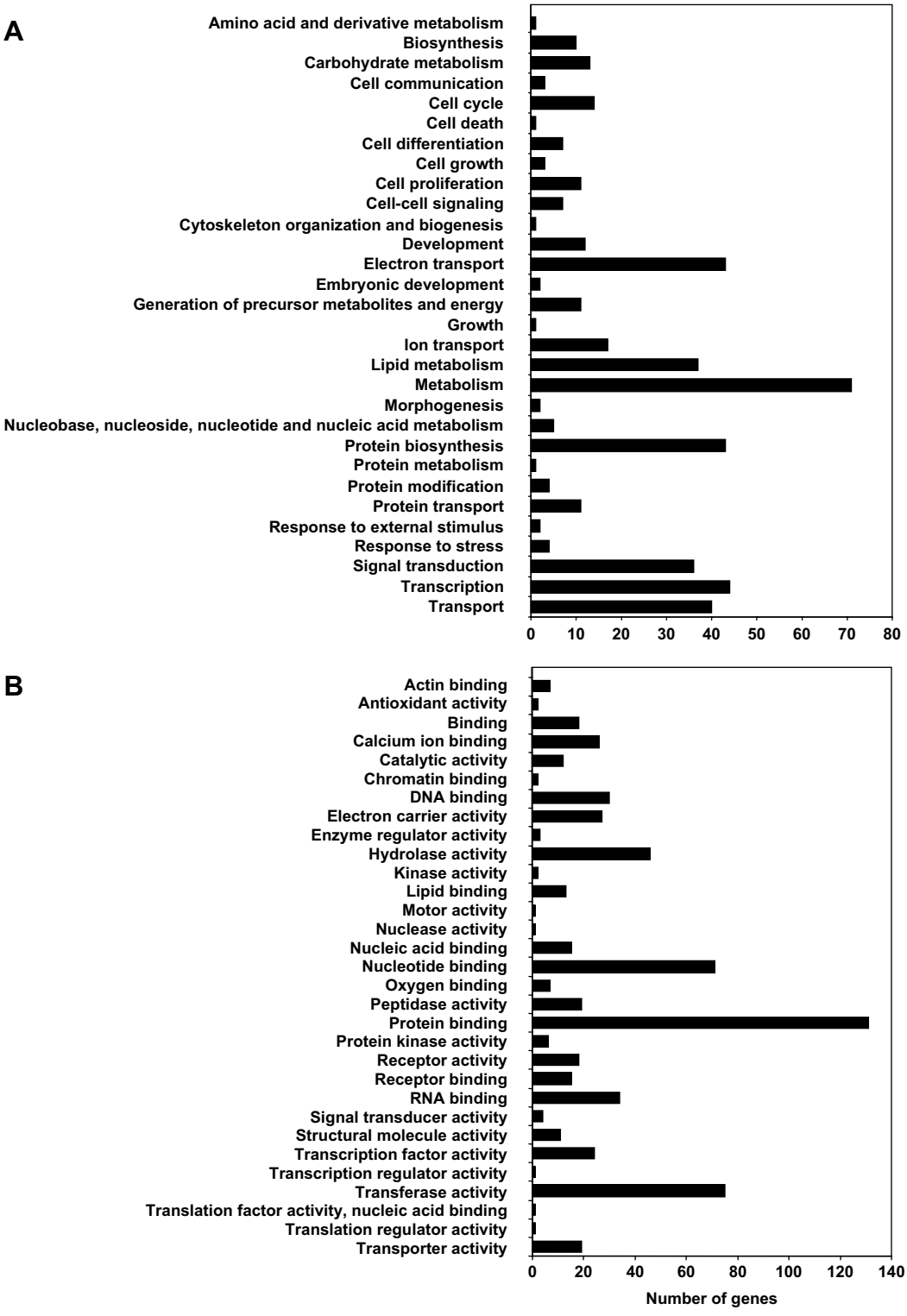

Fig. 1. Functional classification of cynomolgus liver ESTs. All non-redundant ESTs were assigned to each functional category as described in Section 2. Biological process (A) and Molecular function (B) are shown.

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.

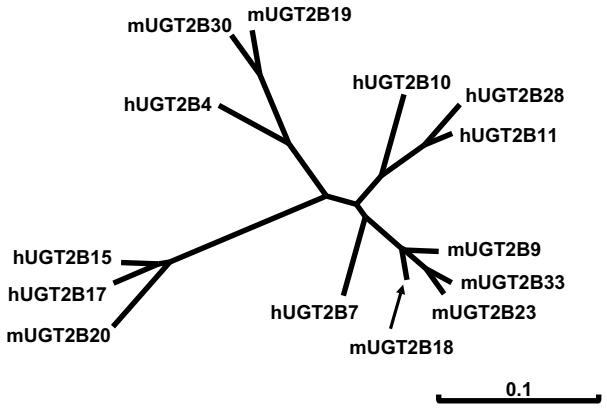

Fig. 2. A phylogenetic comparison of UGTs between macaque and human. The phylogenetic tree was based on amino acid sequence using the Clustal W program. Deduced amino acid sequences were used for cynomolgus monkeys (m) and human (h).

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].

```
mCYP2A23v   1:MLASGLLLVALLACLTVMVLMSVWQQRNSKGKLPPGPTPLPFIGNYLQLNTEQMYNSLMKISERYGPVFTIHLGPRRVVVLCGYDAVKEALVDQAEEFSG 100
mCYP2A23    1:MLASGLLLVALLACLTVMVLMSVWQQRNSKGKLPPGPTPLPFIGNYLQLNTEQMYNSLMKISERYGPVFTIHLGPRRVVVLCGYDAVKEALVDQAEEFSG 100
mCYP2A24    1:MLASGLLLVALLACLTVMVLMSVWQQRNSKGKLPPGPTPLPFIGNYLQLNTEQMCNSIMKISERYGPVFTIHLGPRRVVVLCGYDAVKEALVDQAEEFSG 100
hCYP2A6     1:MLASGMLLVALLVCLTVMVLMSVWQQRKSKGKLPPGPTPLPFIGNYLQLNTEQMYNSLMKISERYGPVFTIHLGPRRVVVLCGHDAVREALVDQAEEFSG 100
hCYP2A7     1:MLASGLLLVALLACLTVMVLMSVWQQRKSRGKLPPGPTPLPFIGNYLQLNTEHICDSIMKFSECYGPVFTIHLGPRRVVVLCGHDAVREALVDQAEEFSG 100
hCYP2A13    1:MLASGLLLVTLLACLTVMVLMSVWRQRKSRGKLPPGPTPLPFIGNYLQLNTEQMYNSLMKISERYGPVFTIHLGPRRVVVLCGHDAVKEALVDQAEEFSG 100
              *****.***.**.***********.** *.**********************....*.**.**.****************** ***.***********

mCYP2A23v 101:RGEQATFDWLFKGYGVVFSNGERAKQLRRFSIATLRDFGVGKRGIEERIQEEAGFLIEALRDTQGANIDPTFFLSRTVSNVISSIVFGDRFDYEDKEFLS 200
mCYP2A23  101:RGEQATFDWLFKGYGVVFSNGERAKQLRRFSIATLRDFGVGKRGIEERIQEEAGFLIEALRDTQGANIDPTFFLSRTVSNVISSIVFGDRFDYEDKEFLS 200
mCYP2A24  101:RGEQATFDWVFKGYGVVFSNGERAKQLRRFSIATLRDFGVGKRGIEERIQEEAGFLIEALRDTHGANIDPTFFLSRTVSNVISSIVFGDRFDYKDKEFLS 200
hCYP2A6   101:RGEQATFDWVFKGYGVVFSNGERAKQLRRFSIATLRDFGVGKRGIEERIQEEAGFLIDALRGTGGANIDPTFFLSRTVSNVISSIVFGDRFDYKDKEFLS 200
hCYP2A7   101:RGEQATFDWVFKGYGVAFSNGERAKQLLRFAIATLRDFGVGKRGIEERIQEESGFLIEAIRSTHGANIDPTFFLSRTVSNVISSIVFGDRFDYEDKEFLS 200
hCYP2A13  101:RGEQATFDWLFKGYGVAFSNGERAKQLRRFSIATLRGFGVGKRGIEERIQEEAGFLIDALRGTHGANIDPTFFLSRTVSNVISSIVFGDRFDYEDKEFLS 200
              ********* ******.**********.**.*****.***************.****.*.* * ****************************.******

mCYP2A23v 201:LLRMMLGSFQFTATSAGQLYEMFSSVMKHLPGPQQQAFKELQGLEDFIAKKVEHNRRTLDPNSPRDFIDSFLIRMQEEEKNPNTEFHLKNLVLTSLNLFF 300
mCYP2A23  201:LLRMMLGSFQFTATSAGQLYEMFSSVMKHLPGPQQQAFKELQGLEDFIAKKVEHNRRTLDPNSPRDFIDSFLIRMQEEEKNPNTEFHLKNLVLTSLNLFF 300
mCYP2A24  201:LLGMMLAIFQFTSTSTGQLYEMFSSVMKHLPGPQQQAFKELQGLEDFIAKKVEHNQRTLDPNSPRDFIDSFLIRMQEEEKNPNTEFYLKNLMMTTLNLFI 300
hCYP2A6   201:LLRMMLGIFQFTSTSTGQLYEMFSSVMKHLPGPQQQAFQLLQGLEDFIAKKVEHNQRTLDPNSPRDFIDSFLIRMQEEEKNPNTEFYLKNLVMTTLNLFI 300
hCYP2A7   201:LLSMMLGIFQFTSTSTGQLYEMFSSVMKHLPGPQQQAFKLLQGLEDFIAKKVEHNQRTLDPNSPQDFIDSFLIHMQEEEKNPNTEFYLKNLMMSTLNLFI 300
hCYP2A13  201:LLRMMLGSFQFTATSTGQLYEMFSSVMKHLPGPQQQAFKELQGLEDFIAKKVEHNQRTLDPNSPRDFIDSFLIRMQEEEKNPNTEFYLKNLVMTTLNLFF 300
              **.***. **** **.**********************..***************.********.********.************.****....****

mCYP2A23v 301:GGTETVSTTLRYGFLLLMKHPEVEAKVHEEIDRVIGKNRQPKFEDWAKMPYTEAVIHEIQRFGDMLPFGVAHRVIKDTKFRDFFLPKGTEVFPMLGSVLK 400
mCYP2A23  301:GGTETVSTTLRYGFLLLMKHPEVEAKVHEEIDRVIGKNRQPKFEDWAKMPYTEAVIHEIQRFGDMLPFGVAHRVIKDTKFRDFFLPKGTEVFPMLGSVLK 400
mCYP2A24  301:AGTETVSTTLRYGFLLLMKYPEVEAKVHEEIDRVIGKNRQPKFEDRVKMPYMEAVIHEIQRFGDVIPMSLARRVNKDTKFRDFFLPKGTEVFPMLGSVLR 400
hCYP2A6   301:GGTETVSTTLRYGFLLLMKHPEVEAKVHEEIDRVIGKNRQPKFEDRAKMPYMEAVIHEIQRFGDVIPMSLARRVKKDTKFRDFFLPKGTEVFPMLGSVLR 400
hCYP2A7   301:AGTETVSTTLRYGFLLLMKHPEVEAKVHEEIDRVIGKNRQPKFEDRTKMPYMEAVIHEIQRFGDVIPMSLARRVKKDTKFRDFFLPKGTEVFPMLGSVLR 400
hCYP2A13  301:AGTETVSTTLRYGFLLLMKHPEVEAKVHEEIDRVIGKNRQPKFEDRAKMPYTEAVIHEIQRFGDMLPMGLAHRVNKDTKFRDFFLPKGTEVFPMLGSVLR 400
               ******************.*************************..**** ************  *. .* ** ************************.

mCYP2A23v 401:DPKFFSNPQDFNPQHFLDEKGQFKKSDAFVPFSIGKRPLFAARPLLTPAGASLTQLPSVLCSLVFLQLGSSC                              472
mCYP2A23  401:DPKFFSNPQDFNPQHFLDEKGQFKKSDAFVPFSIGKRNCFGEGLARMELFLFLTTIMQNFRFKSPQSPKDIDVSPKHMGFATIPPNYTMSFLPR       494
mCYP2A24  401:DPRFFSNPQDFNPQHFLDEKGQFKKSDAFVPFSIGKRNCFGEGLARMELFLFFTTIMQNFRFKSPQSPKDIDVSPKHAGFATIPRNYTMSFLPR       494
hCYP2A6   401:DPSFFSNPQDFNPQHFLNEKGQFKKSDAFVPFSIGKRNCFGEGLARMELFLFFTTVMQNFRLKSSQSPKDIDVSPKHVGFATIPRNYTMSFLPR       494
hCYP2A7   401:DPSFFSNPQDFNPQHFLDDKGQFKKSDAFVPFSIGKRNCFGEGLARMELFLFFTTVMQNFRLKSSQSPKDIDVSPKHVVFATIPRNYTMSFLPR       494
hCYP2A13  401:DPRFFSNPRDFNPQHFLDKKGQFKKSDAFVPFSIGKRYCFGEGLARMELFLFFTTIMQNFRFKSPQSPKDIDVSPKHVGFATIPRNYTMSFLPR       494
              ** *****.********..******************..*..............*. ..... .. ........
```

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.