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

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

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

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)

ISSN 1471-2164

Article type Research article

Submission date 27 September 2007

Acceptance date 24 February 2008

Publication date 24 February 2008

Article URL http://www.biomedcentral.com/1471-2164/9/90

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

Naoki Osada[1], Katsuyuki Hashimoto[1], Yosuke Kameoka[1], Makoto Hirata[1], Reiko Tanuma[1], Yasuhiro Uno[2], Itsuro Inoue[3], Munetomo Hida[4], Yutaka Suzuki[5], Sumio Sugano[5], Keiji Terao[6], Jun Kusuda[1], Ichiro Takahashi[1]

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

6. Tsukuba Primate Center for Medical Science, National Institute of Biomedical Innovation, Tsukuba, Japan.

Corresponding author

Naoki Osada

E-MAIL: nosada@nibio.go.jp; TEL & FAX: +81 (72) 641-9016.

Email address

NO: nosada@nibio.go.jp

KH: khashi@nih.go.jp

YK: ykameoka@nibio.go.jp

MH: mhirata@nibio.go.jp

RT: tanumark@nibio.go.jp

YU: unoxx001@pharm.hokudai.ac.jp

II: ituro@ims.u-tokyo.ac.jp

MH: hida@imcir.twmu.ac.jp

YS: ysuzuki@hgc.jp

SS: ssugano@ims.u-tokyo.ac.jp

JK: jkusuda@nibio.go.jp

IT: ichiro-t@nibio.go.jp

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.

Conclusions

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 22,395 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.

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 (K_a) and the synonymous substitution rate (K_s) 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 K_a/K_s, assuming that the K_s value reflects the neutral mutation rate [31]. Using the database, users can sort the alignments according to the K_a and K_s values. For example, users can determine the K_a and K_s 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 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; $P < 10^{-15}$; 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 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 ($P < 10^{-13}$; 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 recovered 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 (>50bp) 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.

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.