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 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 $4N_e u =$

0.00327 ± 0.00025 where t, N_e, 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 $4N_eu = 0.00311 \pm 0.00024$ (Table 4). If we consider $u = 10^{-9}$ (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 ± 3300 with a mutation rate per generation of 1.8×10^{-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 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. Elucidating 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).

Conclusions

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 probe sets. RNA samples from the cerebrum, cerebellum, liver, and testis of a 3-year-old male cynomolgus monkey were extracted using TRIZOL (Invitrogen) and hybridized to the GeneChip with duplication in a single experiment. The *M. fascicularis* GeneChip contains at most 11 perfect-match probes (25-mers complete matches to

the cDNA sequences) and 11 mismatch probes (containing one mismatched oligo) for each probe set, similar to other GeneChip formats. Normalization, signal detection, and signal intensity calculation of the microarrays were performed using Affymetrix MAS5.0 software. Transcripts were considered as expressed when the probe set of both the duplicates agreed for the significant expression ($P \leq 0.05$) [33]. The raw array data were deposited in Gene Expression Omnibus [GEO: GSM201873-201880]. The array design and the sequences of oligonucleotide probes were deposited in the public database [GEO: GPL5396].

RT-PCR

Templates of the human brain RNA were purchased from Clontech. The macaque brain RNA was obtained from a 21-year-old male cynomolgus monkey. One microgram of total mRNA was amplified using the PrimeSTAR® RT-PCR Kit (TakaraBio). The temperature and time schedules were 30 cycles at 94°C for 20 s, 60°C for 30 s, and 72°C for 1 min. All primer sequences are presented in Additional file 6.

Human–cynomolgus cDNA sequence alignment

Human-macaque orthologous gene pairs were assigned by the reciprocal best BLAST hit with an e-60 cut-off value. We aligned only that part of the coding sequences (CDS) that was homologous to a BLAST search, because representative human and macaque cDNAs do not necessarily have the same splicing isoforms. The sequences of human and macaque cDNAs were aligned using CLUSTAL W [51], and an unaligned macaque nucleotide was marked by the letter X in the database. Alignments shorter than 100 bp (≤33 codons) were filtered for further analysis. In the database, the positions including deletion in the human sequence (or insertion in the macaque sequence) were dropped for estimating the substitution rates. The non-synonymous substitution rate per non-synonymous site (K_a) and the synonymous substitution rate per synonymous site (K_s) were estimated using the Li-Pamilo-Bianchi method [52,53]. K_a/K_s ratios were set to 100 in the database when the K_s value was zero.

Human–rhesus–cynomolgus cDNA sequence alignment

The predicted cDNA sequences of rhesus macaques were downloaded from Ensembl (MMUL1.0) and aligned with the human RefSeq sequences. Orthology between the rhesus and cynomolgus genes was confirmed again using the cynomolgus–rhesus reciprocal BLAST hit, and human–rhesus–cynomolgus cDNA alignments were compiled. Alignments

containing any frameshifting indels and those shorter than 100 bp (≤33 codons) were filtered, which resulted in 2655 alignments (dataset I). The rhesus cDNA sequences were then mapped on the draft genome sequence of the rhesus macaque (rheMac2). The rhesus macaque genes showing >80% homology to more than one locus on the rhesus genome were removed from the alignments, which yielded 1499 human–rhesus–cynomolgus cDNA alignments (dataset II). To estimate the divergence among the three species, $K_a(d_n)$ and $K_s(d_s)$ were estimated using the maximum likelihood method implemented in the PAML program package [54]. We estimated the transition/transversion ratios in 4-fold degenerated sites using the concatenated cynomolgus and rhesus alignments in advance, and fixed the value to the observed value. The test of positive selection was conducted using the branch-site test of positive selection described by Zhang et al. [41], applying the critical values of 2.71 and 5.41 at 5% and 1% significance level without a Bonferroni correction, respectively.

Estimation of the divergence time between cynomolgus and rhesus macaques

Maximum likelihood estimation (MLE) of the divergence time and ancestral population size was performed using the method of Takahata and Satta [11,14]. MLE was determined using the Newton-Raphson algorithm with many possible initial values. The

standard error was determined from the numerically evaluated Fisher information matrix. In order to correct a PCR error rate, we estimated the PCR error rate to be 5.40×10^{-4}, which was derived from the difference in the synonymous substitution rates of the cynomolgus and rhesus lineages. We assumed that the generation time and the effect of selection on the synonymous sites of the two macaques were the same, and that the erroneous nucleotide incorporated by PCR did not skew. Therefore, when a synonymous substitution in the cynomolgus lineage was found, it was considered that the substitution is because of the PCR error with a probability of 0.178 ($5.40 \times 10^{-4}/3.04 \times 10^{-3}$). We randomly corrected the number of substitutions in the raw data, generated pseudo data for 1000 times, and estimated the evolutionary parameter for each time.

Abbreviations

CDS, coding sequence; EST, expressed sequence tag; MLE, maximum likelihood estimation; ORF, open reading frame; UTR, untranslated region; MRCA, most recent common ancestor

Authors' contributions

NO contributed to the designing of the research, performed the experiments and data analysis, and wrote the manuscript. KH, KT, and JK designed the research and contributed to

the manuscript. MH performed the computational analysis. YK contributed to the microarray experiments. RT, YU, II, and IT were involved in the cDNA sequencing. YS and SS constructed the oligo-capped cDNA libraries. All authors read and approved the final manuscript.

Acknowledgements

This study was supported by a Health Science Research Grant from the Ministry of Health, Labor and Welfare of Japan and a Grant for Scientific Research from the Ministry of Education, Culture, Sports, Science and Technology, Japan (19770073)..

References

1. Carlsson HE, Schapiro SJ, Farah I, Hau J: Use of primates in research: a global overview. *Am J Primatol* 2004, **63**:225-237.

2. Melnick DJ, Hoelzer GA, Absher R, Ashley MV: mtDNA diversity in rhesus monkeys reveals overestimates of divergence time and paraphyly with neighboring species. *Mol Biol Evol* 1993, **10**:282-295.

3. Hayasaka K, Fujii K, Horai S: Molecular phylogeny of macaques: implications of

nucleotide sequences from an 896-base pair region of mitochondrial DNA. *Mol Biol Evol* 1996, **13**:1044-1053.

4. Magness CL, Fellin PC, Thomas MJ, Korth MJ, Agy MB, Proll SC, Fitzgibbon M, Scherer CA, Miner DG, Katze MG, Iadonato SP: Analysis of the Macaca mulatta transcriptome and the sequence divergence between Macaca and human. *Genome Biol* 2005, **6**:R60.

5. Gibbs RA, Rogers J, Katze MG, Bumgarner R, Weinstock GM, Mardis ER, Remington KA, Strausberg RL, Venter JC, Wilson RK, Batzer MA, Bustamante CD, Eichler EE, Hahn MW, Hardison RC, Makova KD, Miller W, Milosavljevic A, Palermo RE, Siepel A, Sikela JM, Attaway T, Bell S, Bernard KE, Buhay CJ, Chandrabose MN, Dao M, Davis C, Delehaunty KD, Ding Y et al.: Evolutionary and biomedical insights from the rhesus macaque genome. *Science* 2007, **316**:222-234.

6. Dutrillaux B, Biemont MC, Viegas Pequignot E, Laurent C: Comparison of the karyotypes of four Cercopithecoidae: Papio papio, P. anubis, Macaca mulatta, and M. fascicularis. *Cytogenet Cell Genet* 1979, **23**:77-83.

7. Tosi AJ, Morales JC, Melnick DJ: Paternal, maternal, and biparental molecular markers provide unique windows onto the evolutionary history of macaque monkeys. *Evolution Int J Org Evolution* 2003, **57**:1419-1435.

8. Ferguson B, Street SL, Wright H, Pearson C, Jia Y, Thompson SL, Allibone P, Dubay CJ, Spindel E, Norgren RB Jr: Single nucleotide polymorphisms (SNPs) distinguish Indian-origin and Chinese-origin rhesus macaques (*Macaca mulatta*). *BMC Genomics* 2007, **8**:43.

9. Hernandez RD, Hubisz MJ, Wheeler DA, Smith DG, Ferguson B, Rogers J, Nazareth L, Indap A, Bourquin T, McPherson J, Muzny D, Gibbs R, Nielsen R, Bustamante CD: Demographic histories and patterns of linkage disequilibrium in Chinese and Indian rhesus macaques. *Science* 2007, **316**:240-243.

10. Smith DG, McDonough JW, George DA: Mitochondrial DNA variation within and among regional populations of longtail macaques (*Macaca fascicularis*) in relation to other species of the fascicularis group of macaques. *Am J Primatol* 2007, **69**:182-198.

11. Edwards SV, Beerli P: Perspective: gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies. *Evolution Int J Org Evolution* 2000, **54**:1839-1854.

12. Takahata N, Satta Y: Evolution of the primate lineage leading to modern humans: phylogenetic and demographic inferences from DNA sequences. *Proc Natl Acad Sci U S A* 1997, **94**:4811-4815.

13. Tosi AJ, Morales JC, Melnick DJ: Y-Chromosome and Mitochondrial Markers in

Macaca fascicularis Indicate Introgression with Indochinese M. mulatta and a Biogeographic Barrier in the Isthmus of Kra. *Int J Primatol* 2002, **23**:161-178.

14. Osada N, Wu CI: Inferring the mode of speciation from genomic data: a study of the great apes. *Genetics* 2005, **169**:259-264.

15. Ota T, Suzuki Y, Nishikawa T, Otsuki T, Sugiyama T, Irie R, Wakamatsu A, Hayashi K, Sato H, Nagai K, Kimura K, Makita H, Sekine M, Obayashi M, Nishi T, Shibahara T, Tanaka T, Ishii S, Yamamoto J, Saito K, Kawai Y, Isono Y, Nakamura Y, Nagahari K, Murakami K, Yasuda T, Iwayanagi T, Wagatsuma M, Shiratori A, Sudo H et al.: Complete sequencing and characterization of 21,243 full-length human cDNAs. *Nat Genet* 2004, **36**:40-45.

16. Osada N, Hida M, Kusuda J, Tanuma R, Iseki K, Hirata M, Suto Y, Hirai M, Terao K, Suzuki Y et al.: Assignment of 118 novel cDNAs of cynomolgus monkey brain to human chromosomes. *Gene* 2001, **275**:31-37.

17. Osada N, Hida M, Kusuda J, Tanuma R, Hirata M, Hirai M, Terao K, Suzuki Y, Sugano S, Hashimoto K: Prediction of unidentified human genes on the basis of sequence similarity to novel cDNAs from cynomolgus monkey brain. *Genome Biol* 2002, **3**:RESEARCH0006.

18. Osada N, Hida M, Kusuda J, Tanuma R, Hirata M, Suto Y, Hirai M, Terao K, Sugano

S, Hashimoto K: Cynomolgus monkey testicular cDNAs for discovery of novel human genes in the human genome sequence. *BMC Genomics* 2002, **3**:36.

19. Osada N, Kusuda J, Hirata M, Tanuma R, Hida M, Sugano S, Hirai M, Hashimoto K: Search for genes positively selected during primate evolution by 5'-end-sequence screening of cynomolgus monkey cDNAs. *Genomics* 2002, **79**:657-662.

20. Osada N, Hirata M, Tanuma R, Kusuda J, Hida M, Suzuki Y, Sugano S, Gojobori T, Shen CK, Wu CI, Hashimoto K: Substitution rate and structural divergence of 5'-UTR evolution: comparative analysis between human and cynomolgus monkey cDNAs. *Mol Biol Evol* 2005, **22**:1976-1982.

21. Wang HY, Chien HC, Osada N, Hashimoto K, Sugano S, Gojobori T, Chou CK, Tsai SF, Wu CI, Shen CK: Rate of Evolution in Brain-Expressed Genes in Humans and Other Primates. *PLoS Biol* 2007, **5**:e13.

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

23. Maglott D, Ostell J, Pruitt KD, Tatusova T: Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res* 2005, **33**:D54-58.

24. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel Tarver L, Kasarskis A,

Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000, **25**:25-29.

25. URL: QFbase *http://genebank.nibio.go.jp/qfbase/*

26. Benson DA, Karsch Mizrachi I, Lipman DJ, Ostell J, Wheeler DL: GenBank. *Nucleic Acids Res* 2007, **35**:D21-25.

27. Hubbard TJ, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, Down T, Dyer SC, Fitzgerald S, Fernandez Banet J, Graf S, Haider S, Hammond M, Herrero J, Holland R, Howe K, Howe K, Johnson N, Kahari A, Keefe D, Kokocinski F, Kulesha E, Lawson D, Longden I, Melsopp C, Megy K et al.: Ensembl 2007. *Nucleic Acids Res* 2007, *35*:D610-617.

28. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA: Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* 2005, **33**:D514-517.

29. Imanishi T, Itoh T, Suzuki Y, O'Donovan C, Fukuchi S, Koyanagi KO, Barrero RA, Tamura T, Yamaguchi Kabata Y, Tanino M, Yura K, Miyazaki S, Ikeo K, Homma K, Kasprzyk A, Nishikawa T, Hirakawa M, Thierry Mieg J, Thierry Mieg D, Ashurst J, Jia L, Nakao M, Thomas MA, Mulder N, Karavidopoulou Y, Jin L, Kim S, Yasuda