Journal of General Virology (2012), 93, 594–602 DOI 10.1099/vir.0.038075-0

Geographical, genetic and functional diversity of antiretroviral host factor TRIMCyp in cynomolgus macaque (*Macaca fascicularis*)

Akatsuki Saito,[1]† Ken Kono,[2]† Masako Nomaguchi,[3] Yasuhiro Yasutomi,[4] Akio Adachi,[3] Tatsuo Shioda,[2] Hirofumi Akari[1,4] and Emi E. Nakayama[2]

Correspondence
Hirofumi Akari
akari@pri.kyoto-u.ac.jp
Emi E. Nakayama
emien@biken.osaka-u.ac.jp

[1]Primate Research Institute, Kyoto University, Inuyama 484-8506, Japan

[2]Department of Viral Infections, Research Institute for Microbial Diseases, Osaka University, Suita 565-0871, Japan

[3]Department of Microbiology, Institute of Health Biosciences, The University of Tokushima Graduate School, Tokushima 770-8503, Japan

[4]Tsukuba Primate Research Center, National Institute of Biomedical Innovation, Tsukuba 305-0843, Japan

The antiretroviral factor tripartite motif protein 5 (*TRIM5*) gene-derived isoform (TRIMCyp) has been found in at least three species of Old World monkey: rhesus (*Macaca mulatta*), pig-tailed (*Macaca nemestrina*) and cynomolgus (*Macaca fascicularis*) macaques. Although the frequency of TRIMCyp has been well studied in rhesus and pig-tailed macaques, the frequency and prevalence of TRIMCyp in cynomolgus macaques remain to be definitively elucidated. Here, the geographical and genetic diversity of TRIM5α/TRIMCyp in cynomolgus macaques was studied in comparison with their anti-lentiviral activity. It was found that the frequency of TRIMCyp in a population in the Philippines was significantly higher than those in Indonesian and Malaysian populations. Major and minor haplotypes of cynomolgus macaque TRIMCyp with single nucleotide polymorphisms in the cyclophilin A domain were also found. The functional significance of the polymorphism in TRIMCyp was examined, and it was demonstrated that the major haplotype of TRIMCyp suppressed human immunodeficiency virus type 1 (HIV-1) but not HIV-2, whilst the minor haplotype of TRIMCyp suppressed HIV-2 but not HIV-1. The major haplotype of TRIMCyp did not restrict a monkey-tropic HIV-1 clone, NL-DT5R, which contains a capsid with the simian immunodeficiency virus-derived loop between α-helices 4 and 5 and the entire *vif* gene. These results indicate that polymorphisms of TRIMCyp affect its anti-lentiviral activity. Overall, the results of this study will help our understanding of the genetic background of cynomolgus macaque TRIMCyp, as well as the host factors composing species barriers of primate lentiviruses.

Received 2 October 2011
Accepted 22 November 2011

INTRODUCTION

Human immunodeficiency virus type 1 (HIV-1) barely replicates in Old World monkeys such as cynomolgus macaques (CMs; *Macaca fascicularis*) and rhesus macaques (RMs; *Macaca mulatta*). This species barrier has long hampered the use of Old World monkeys for human immunodeficiency virus type 1 (HIV-1) research. Recently, a number of intrinsic anti-HIV-1 cellular factors, including tripartite motif protein 5α (TRIM5α), cyclophilin A (CypA), the apolipoprotein B mRNA-editing enzyme catalytic polypeptide-like 3 (APOBEC3) family and tetherin were identified in Old World monkey cells (Nomaguchi *et al.*, 2008; Sauter *et al.*, 2010). Of these factors, TRIM5α was found to strongly suppress HIV-1 replication, mainly by affecting the virus disassembly step, resulting in a decrease in reverse-transcription products (Nakayama & Shioda, 2010; Stremlau *et al.*, 2004). TRIM5α contains a RING domain, a B-box domain, a coiled-coil domain and a PRYSPRY (B30.2) domain. Importantly, the PRYSPRY domain recognizes the capsid of incoming retroviruses, leading to post-entry restriction of infection. RM and CM TRIM5α restrict HIV-1 but not simian immunodeficiency virus isolated from an infected rhesus macaque (SIVmac) (Nakayama *et al.*, 2005; Stremlau *et al.*, 2004; Yap *et al.*, 2004). In the case of HIV-2

†These authors contributed equally to this work.

The GenBank/EMBL/DDBJ accession numbers for the sequences of CM TRIMCyp-major (DK) and RM TRIMCyp are AB671588 and AB671589, respectively.

Two supplementary tables are available with the online version of this paper.

infection, viruses carrying proline (P) at aa 120 of the capsid protein are sensitive to CM TRIM5α, whereas those with either alanine or glutamine (Q) are resistant (Song *et al.*, 2007). Both CM TRIM5α-sensitive and -resistant HIV-2 strains are restricted by RM TRIM5α, and three amino acid residues – threonine (T), phenylalanine (F) and P at aa 339, 340 and 341, respectively – of RM TRIM5α are important for restricting particular HIV-2 strains, which are still resistant to CM TRIM5α (Kono *et al.*, 2008). It is also known that TRIM5α exhibits a high degree of sequence variation, even within macaque species. In some individual RMs, the TFP residues at aa 339–341 of TRIM5α are replaced with a single Q (Newman *et al.*, 2006) and this TFP→Q polymorphism affects the anti-lentiviral activity of RM TRIM5α (Kirmaier *et al.*, 2010).

Although pig-tailed macaques (PMs; *Macaca nemestrina*) have long been thought to exhibit a higher susceptibility to HIV-1 infection than RMs and CMs (Agy *et al.*, 1992), the underlying mechanism determining this difference remained unclear. Recently, a TRIM5–CypA chimeric protein, referred to as TRIMCyp, was discovered in PMs, and the monkeys exclusively expressed TRIMCyp but not TRIM5α (Brennan *et al.*, 2008; Liao *et al.*, 2007). TRIMCyp is an alternatively spliced isoform of the *TRIM5* gene in which the PRYSPRY domain of TRIM5α is replaced with a retrotransposed *CypA* gene. The retrotransposition of the *CypA* sequence in the 3′ UTR of the *TRIM5* gene correlates with a single nucleotide polymorphism (SNP) at the exon 7 splice-acceptor site, leading to skipping of exons 7 and 8 encoding the PRYSPRY domain and splicing to the *CypA* insertion. Thus, the presence or absence of the *CypA* sequence in the 3′ UTR results in expression of TRIMCyp or TRIM5α, respectively.

In vitro analyses demonstrated that cells expressing PM TRIMCyp restricted HIV-2 but not HIV-1 infection (Brennan *et al.*, 2008; Liao *et al.*, 2007), suggesting that the characteristic isoform of the *TRIM5* gene in PMs may be one of the reasons for their greater susceptibility to HIV-1 infection. Furthermore, TRIMCyp was also identified in some individual RMs and CMs (Brennan *et al.*, 2008; Newman *et al.*, 2008; Wilson *et al.*, 2008a). RM TRIMCyp, as well as that of PMs, is unable to restrict HIV-1 infection (Wilson *et al.*, 2008a). This report also showed that the frequency of TRIMCyp in Indian RMs was approximately 25 %, whilst it was not found in the Chinese RM population, suggesting a geographical deviation in the frequency of TRIMCyp (Wilson *et al.*, 2008a). In the case of CMs, although the existence of TRIMCyp has been reported (Brennan *et al.*, 2008), the allele frequency, geographical distribution and relevance in antiviral activity of TRIMCyp remain to be elucidated. As the *TRIM5* gene-related factors are expected to have an impact on the replication of retroviruses, information about the genetic background of CM TRIMCyp will contribute to our understanding of host factors composing the species barrier. In the present study, we studied the geographical, genetic and functional diversity of CM TRIMCyp originating from South-West Asia (Indonesia, Malaysia and the Philippines). We showed a geographical deviation in the frequency of TRIMCyp. Moreover, we found SNPs in CM TRIMCyp and analysed their impact on the anti-lentiviral functions, including their effect against HIV-1, HIV-2, SIVmac and monkey-tropic HIV-1 (HIV-1mt).

RESULTS

Geographical deviation in the frequency of CM TRIMCyp

Initially, we analysed the frequencies of TRIM5α and TRIMCyp genotypes in 126 CMs originating from three different regions – Indonesia, Malaysia and the Philippines – using a PCR-based assay designed to differentiate between the presence and absence of the *CypA* insertion (Fig. 1a) (Wilson *et al.*, 2008a). Insertion of the *CypA* gene in TRIMCyp resulted in a PCR product larger than the expected size for TRIM5α (Fig. 1b).

As shown in Table 1, 35 of the 46 Philippine individuals were homozygous for TRIMCyp, ten were heterozygous and only one was homozygous for TRIM5α. In contrast, only three of the 33 Indonesian individuals were homozygous for TRIMCyp, 17 were heterozygous and 13 were homozygous for TRIM5α. Interestingly, the Malaysian population was of intermediate proportions: ten TRIMCyp homozygotes, 26 heterozygotes and 11 TRIM5α homozygotes. As shown in Fig. 2, the percentages of individuals having each *TRIM5* genotype indicated that the frequency of TRIMCyp homozygotes in Malaysian CMs was twice that in Indonesian CMs. In contrast, the frequency of TRIM5α homozygotes in Indonesian CMs was twice that in Malaysian CMs. Taken together, the calculated allele frequencies of TRIMCyp in the Philippine, Indonesian and Malaysian CM populations were 87.0, 34.8 and 48.9 %, respectively (Table 1). Statistical analyses using a χ^2 test followed by Bonferroni correction demonstrated that the frequency of TRIMCyp in the Philippine population was significantly higher than that in the Indonesian ($P<0.0001$) and Malaysian ($P<0.0001$) populations. In contrast, there was no significant difference between the Indonesian and Malaysian populations ($P=0.2295$).

It should be noted that our method failed to distinguish homozygotes from hemizygotes, especially when the subjects exhibited no polymorphisms in the *TRIM5* gene. However, hemizygosity for the *TRIM5* gene is highly unlikely for the following reasons: (i) the *TRIM5* gene is on an autosomal chromosome, (ii) there is no precedent of deletion of the *TRIM5* gene in humans or primates, and (iii) all of the three CM populations in Table 1 are in Hardy–Weinberg equilibrium for *TRIM5* genotypes.

Polymorphisms in the *CypA* domain of CM TRIMCyp

Previously, it was reported that aa 357 of CM TRIMCyp, corresponding to aa 54 counting from the methionine of

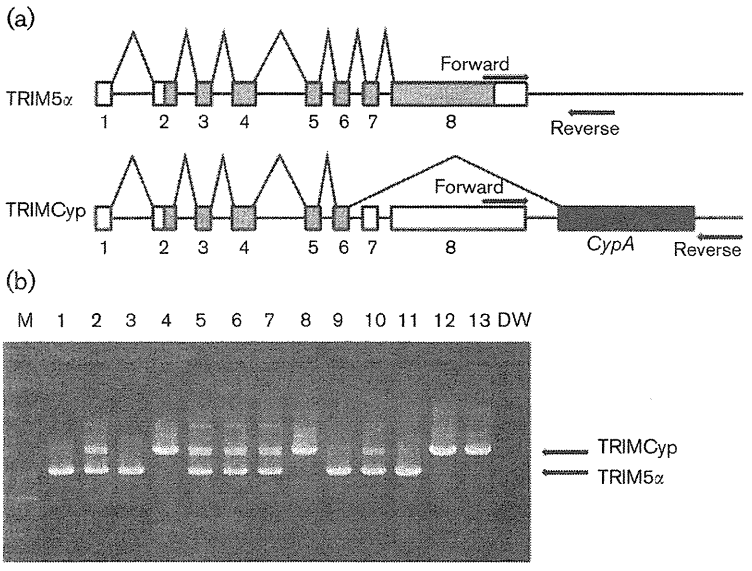

Fig. 1. Determination of *CypA* insertion. (a) Diagram indicating the splicing of TRIM5α or TRIMCyp. Non-coding and coding exons (numbered) and *CypA* sequences are indicated as open, shaded and filled boxes, respectively. The primers used in this study are indicated by arrows. (b) Genomic DNA was extracted from PBMCs. To test for *CypA* insertion, the 3′ region of the *TRIM5* gene was amplified by PCR with primers spanning the 3′ UTR and the putative *CypA* insertion. DW, Distilled water control.

CypA, was arginine (R) (Brennan *et al.*, 2008). Subsequently, Ylinen *et al.* (2010) reported another allele of CM TRIMCyp encoding histidine (H) at this position. To determine the frequency of this R→H polymorphism, we examined 34 TRIM5α/TRIMCyp heterozygotes and 30 TRIMCyp homozygotes for sequence variations in the CypA domain. The results showed that there was no TRIMCyp allele encoding R at position 357 (Cyp 54R). All 94 CM chromosomes carrying the TRIMCyp gene encoded TRIMCyp with H at this position. This result was consistent with the results reported recently by Dietrich *et al.* (2011).

Dietrich *et al.* (2011) also reported CM TRIMCyp polymorphisms at aa 369 and 446, corresponding to aa 66 and 143 in the CypA domain, respectively. Both Brennan *et al.* (2008) and Ylinen *et al.* (2010) reported that aa 369 (Cyp66) and 446 (Cyp143) are aspartic acid (D) and lysine (K), respectively (denoted as the DK haplotype), whilst Dietrich *et al.* (2011) showed the presence of another haplotype encoding asparagine (N) and glutamic acid (E) at positions 369 (Cyp66) and 446 (Cyp143), respectively (denoted as the NE haplotype). Our results showed that 12 CM chromosomes carried TRIMCyp with the NE haplotype, whilst the remaining 82 TRIMCyp were all the DK haplotype (Table 2). Residues 369N (Cyp 66N) and 446E (Cyp 143E) were also found in PM and RM TRIMCyps, and the CypA portion of the NE haplotype of CM TRIMCyp has the same amino acid sequence as RM TRIMCyp (GenBank accession no. EU157763). These results indicate that the previously recognized interspecies variations of the CypA sequence of TRIMCyp were in fact intraspecies variation within CMs. With respect to the geographical distribution of these haplotypes, we found no significant deviation in the frequencies of the haplotypes among the three origins (Table 2).

Polymorphisms in the RING, B-box, coiled-coil, linker and PRYSPRY domains of CM TRIM5α and TRIMCyp

To identify polymorphisms that are in possible linkage disequilibrium with either the DK or NE haplotype in regions other than the CypA domain, we determined nucleotide sequences of TRIM5α and TRIMCyp cDNAs encoding the RING, B-box, coiled-coil and linker domains of six TRIMCyp homozygotes (three homozygotes of the DK haplotype and three heterozygotes for the DK and NE haplotypes) and three TRIM5α homozygotes (see Supplementary Table S1, available in JGV Online). We found polymorphisms at positions 4 [E→glycine (G)] in

Table 1. Frequencies of TRIMCyp alleles in CM populations

Origin	No. of animals	Genotype (no. of animals)			Allele frequency (%)	
		TRIM5α homozygote	TRIM5α/TRIMCyp heterozygote	TRIMCyp homozygote	TRIM5α	TRIMCyp
Philippines	46	1	10	35	13.0	87.0
Malaysia	47	11	26	10	51.1	48.9
Indonesia	33	13	17	3	65.2	34.8

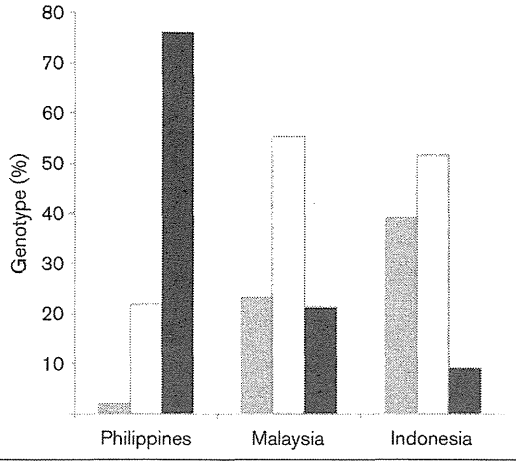

Fig. 2. Frequency of individuals having each *TRIM5* genotype. The percentages of TRIM5α homozygotes and heterozygotes and TRIMCyp homozygotes in each population were calculated. Grey bars, TRIM5α homozygote; white bars, heterozygote; black bars, TRIMCyp homozygote.

the N-terminal region, 44 (K→E) in the RING domain, 178 [H→tyrosine (Y)] and 209 (K→E) in the coiled-coil domain, and 247 (E→D) and 285 (G→R) in the linker domain (Fig. 3a). We found only one chromosome for minor allele 4G, two for 44E, four for 178Y, nine for 209E, five for 247D and four for 285R among 18 chromosomes from the six TRIMCyp homozygotes and three TRIM5α homozygotes. Among the six TRIMCyp homozygotes, we also found three E→Q substitutions at aa 296, which was present in TRIMCyp but absent from TRIM5α. There was no polymorphism that showed strong linkage disequilibrium with either the DK or NE haplotype except for G285R. The NE haplotype tended to link with 285G, although several DK haplotypes also linked with 285G (Supplementary Table S1). The numbers of polymorphic positions were relatively small among CMs compared with RMs (Fig. 3b). It is known that the coiled-coil region of *TRIM5* genes shows a high degree of genetic diversity in RMs (Johnson & Sawyer, 2009). In contrast, the coiled-coil domain of CM TRIM5α and TRIMCyp showed no polymorphism at aa 184, 196, 208, 222, 230 and 236, which were all highly polymorphic in RMs (Newman *et al.*, 2006). These results suggest that the evolutionary pressures targeting the coiled-coil domain of the *TRIM5* gene were weaker in CMs than in RMs.

We also determined the nucleotide sequences of exon 8 encoding the PRYSPRY domain, of 12 TRIM5α homozygotes including the three TRIM5α homozygotes analysed above (see Supplementary Table S2, available in JGV Online). We found polymorphisms at aa 311 [serine (S)→leucine (L)], 327 (P→T), 330 [valine (V)→methionine (M)], 350 [V→isoleucine (I)] and 435 (I→V) in the PRYSPRY domain (Fig. 3a). Among the 12 TRIM5α homozygotes, we did not find a TFP allele at aa 339–341, which is a major determinant for different virus specificity between CM and RM TRIM5αs (Kono *et al.*, 2008) and is also critical for SIV from sooty mangabeys (SIVsm) (Kirmaier *et al.*, 2010) and SIVmac (Lim *et al.*, 2010) restriction by RM TRIM5α. We found only one chromosome for minor allele 311L, one for 327T, one for 350I and four for 435V among 11 TRIM5α homozygotes. We previously cloned CM TRIM5α cDNA from HSC-F cells (GenBank accession no. AB210052) (Nakayama *et al.*, 2005) and found that it contained 330V; however, all of the sequences determined in the present study showed M at this position. In contrast, exon 8 of the TRIMCyp gene of seven TRIMCyp homozygotes (all were heterozygotes for the DK and NE haplotypes), which encoded the PRYSPRY domain but was absent from the mRNA due to splicing, showed a uniform sequence identical to that of the Mamu 7 haplotype of RMs (307P, 313V, 327P, 332R, 333T, 334Q, 339Q, 345I, 383P, 414V, 420S and 488M). We only found one F→L substitution at position 454 among the seven TRIMCyp homozygotes. The Mamu 7 sequence is thus likely to be an ancient prototype sequence of TRIMCyp before the separation of CMs from RMs.

Anti-lentiviral activity of CM TRIMCyps

To elucidate the impact of CM TRIMCyp and its SNPs on the anti-lentiviral activity, we constructed a recombinant

Table 2. Frequencies of DK and NE haplotypes in CM TRIMCyps

Origin	No. of animals	Genotype (no. of chromosomes)				Frequency (%)	
		TRIM5α/TRIMCyp heterozygote*		TRIMCyp homozygote†		DK	NE
		DK	NE	DK	NE		
Philippines	28	6	1	36	6	85.7	14.3
Malaysia	21	14	1	10	2	88.9	11.1
Indonesia	15	12	0	4	2	88.9	11.1

*Haplotypes were determined by direct sequencing of the PCR products.
†Haplotypes were inferred by maximum-likelihood estimation using the results of direct sequencing of the PCR products.

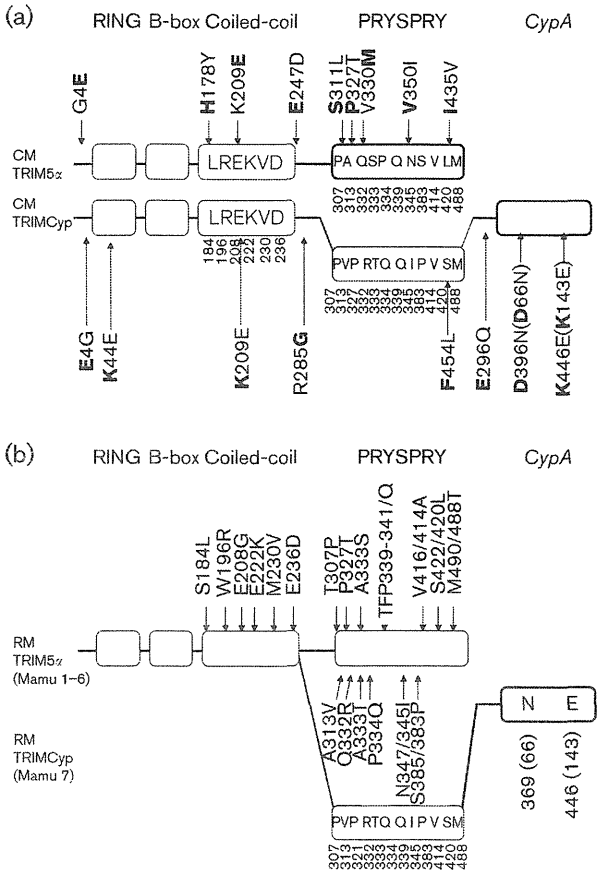

Fig. 3. Sequence variations in TRIM5α and TRIMCyp. (a) Sequence variations in CM TRIM5α and TRIMCyp. The RING, B-box, coiled-coil, PRYSPRY and *CypA* domains of CM TRIM5α and TRIMCyp are indicated by open boxes. The box with thin lines shows exons 7 and 8 of the TRIMCyp gene, which is absent from the mRNA. Polymorphisms are shown outside the boxes, with downward and upward arrows indicating the polymorphisms observed among TRIM5α homozygotes and TRIMCyp homozygotes, respectively. Amino acid residues found in HSC-F cells are shown in front of the amino acid positions, followed by the observed polymorphisms. Major alleles are shown in bold. Numbers in parentheses indicate amino acid positions counting from the initiation methionine codon of the *CypA* ORF. Amino acid residues in the boxes are polymorphic in the RM *TRIM5* gene but lack polymorphism in CM TRIM5α or TRIMCyp. Positions of these amino acid residues are shown below the boxes. (b) Sequence variations of RM TRIM5α (Mamu 1–6) and TRIMCyp (Mamu 7). Downward and upward arrows indicate the polymorphisms observed in TRIM5α and TRIMCyp, respectively. Amino acid residues in boxes indicate those of RM TRIMCyp. Positions of these amino acid residues are shown below the boxes.

Sendai virus (SeV) expressing a series of TRIM5α/TRIMCyp: CM TRIM5α, the DK and NE haplotypes of the CM TRIMCyp [CM TRIMCyp-major (DK) and CM TRIMCyp-minor (NE)], CM SPRY (−) in which the PRYSPRY domain was deleted as a negative control for functional TRIM5α and TRIMCyp, and an RM TRIMCyp. We also constructed a recombinant SeV expressing a CM TRIMCyp-minor (NE) carrying G at position 285 (CM TRIMCyp-minor R285G), as the NE haplotype seemed to be in linkage disequilibrium with G at this position (Supplementary Table S1). As shown in Fig. 4[(a), upper panels], TRIMCyp-major (DK) completely restricted HIV-1 NL4-3, weakly restricted HIV-2 GH123 and failed to restrict SIVmac239. In contrast, TRIMCyp-minor (NE) and TRIMCyp-minor R285G inefficiently restricted HIV-1 NL4-3, barely restricted SIVmac239 and completely restricted HIV-2 GH123. These results indicated that the sequence variations in CM TRIMCyp greatly altered the spectrum of its anti-lentiviral activity. It should be noted that HIV-1 NL4-3 attained slightly higher titres at day 3 in cells expressing TRIMCyp-minor R285G than in those expressing TRIMCyp-minor (NE). The difference was small but reproducible in six independent experiments. This result indicated that aa 285 of TRIMCyp also affected its antiviral activity. In the case of RMs, in which the CypA domain of TRIMCyp has the same amino acid sequence as CM TRIMCyp-minor (NE), RM TRIMCyp showed the same spectrum of anti-lentiviral activity as CM TRIMCyp-minor (NE) (Fig. 4a, lower panels), consistent with previous reports (Price *et al.*, 2009; Wilson *et al.*, 2008a).

Finally, we examined whether HIV-1mt NL-DT5R (Kamada *et al.*, 2006) could evade restriction by CM TRIM5α/TRIMCyp. HIV-1mt possesses core antigen (CA) with the SIVmac239-derived loop between α-helices 4 and 5 (L4/5), which corresponds to a CypA-binding loop of HIV-1, the entire SIVmac239 *vif* gene and two non-synonymous substitutions in the *env* gene (Fig. 4b). As shown in Fig. 4(c), NL-DT5R was restricted by TRIM5α but completely evaded restriction by CM TRIMCyp-major (DK), suggesting that replacement of the CypA-binding loop of HIV-1 CA with the corresponding SIVmac239-derived sequence was sufficient to render HIV-1 resistant to the major haplotype of CM TRIMCyp but not TRIM5α.

DISCUSSION

In the present study, we analysed the geographical, genetic and functional diversity of CM TRIMCyp and found: (i) a clear geographical deviation in the frequency of TRIMCyp, (ii) no typical geographical deviation in the frequency of the DK/NE haplotypes in the CypA domain, and (iii) sequence variations in the CypA domain of CM TRIMCyp, which greatly altered the spectrum of its anti-lentiviral activity.

We first demonstrated that the allele frequency of TRIMCyp in CMs from the Philippines was significantly higher than those in Indonesian and Malaysian CMs. It is possible that some pathogen(s) resistant to the antiviral effect of either TRIM5α or TRIMCyp may contribute to this deviation as a selective pressure. As primate lentiviruses such as HIV-1 and

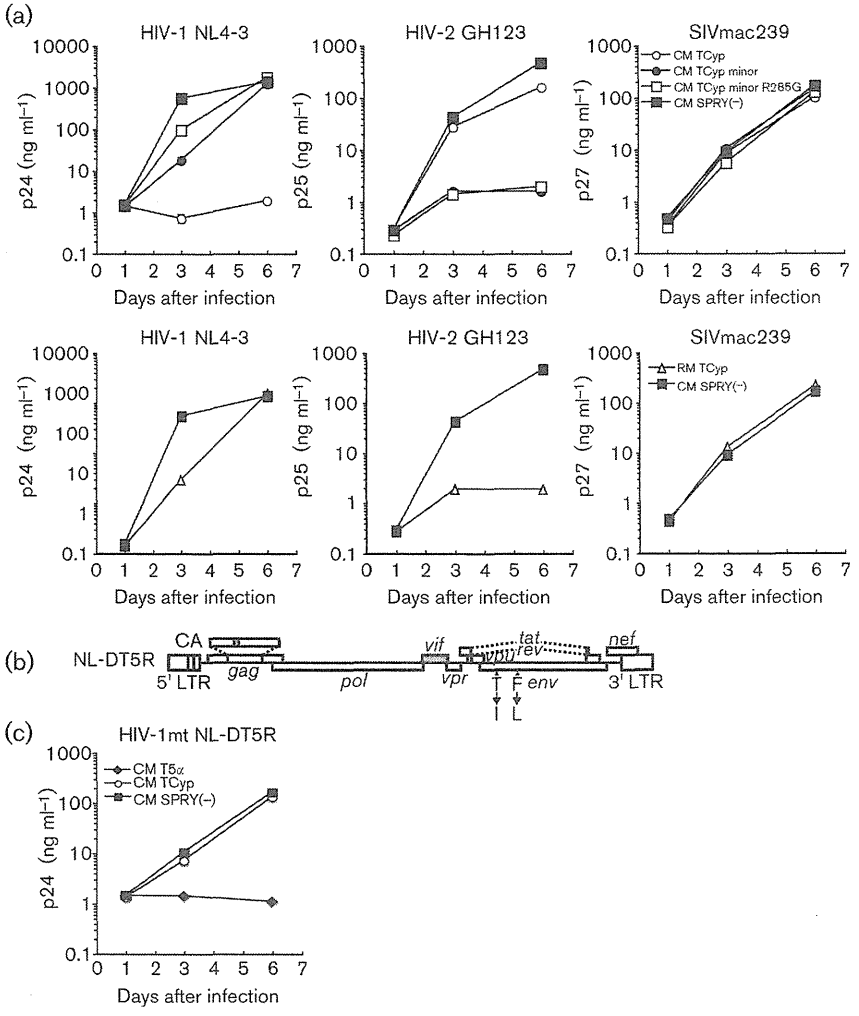

Fig. 4. Anti-lentiviral activity of various CM TRIMCyp. (a) MT4 cells were infected with recombinant SeV expressing CM TRIMCyp-major (DK) (CM TCyp; ○), CM TRIMCyp-minor (NE) (CM TCyp minor; ●), CM TRIMCyp-minor R285G (CM TCyp minor R285G; □), CM SPRY (−) (■) or RM TRIMCyp (RM TCyp; △). Data for CM SPRY (−) in the upper and lower panels were identical. Nine hours after infection, cells were superinfected with HIV-1 NL4-3, HIV-2 GH123 or SIVmac239. Culture supernatants were assayed separately for levels of p24, p25 or p27. (b) Structure of HIV-1mt NL-DT5R used in the experiment shown in Fig. 3(c). Open boxes denote HIV-1 (NL4-3) and shaded boxes denote SIVmac239 sequences. (c) MT4 cells were infected with recombinant SeV expressing CM TRIM5α (CM T5α; ◆), CM TRIMCyp-major (DK) (CM TCyp; ○) or CM SPRY (−) (■). Nine hours after infection, cells were superinfected with HIV-1mt NL-DT5R. Culture supernatants were assayed separately for levels of p24. Error bars show actual fluctuations between duplicate samples. Data from a representative of three (a) or two (c) independent experiments are shown.

SIV originated in African primates, it is unlikely that these viruses could contribute directly to this deviation, and some exogenous and endogenous retroviruses may thus play a critical role in this selection. Alternatively, it is possible that this deviation could come from bottleneck effects. It is estimated that the Philippine CMs were derived from Indonesian CM stocks via sea rafting or terrestrial access through Borneo during periods of low sea level in South-East Asia around 110 000 years ago (Abegg & Thierry, 2002; Blancher *et al.*, 2008; Kita *et al.*, 2009). Furthermore, phylogenetic analyses of mitochondrial DNA sequences of four CM populations distributed in South-East Asia suggested that Philippine CMs were derived from the small founding populations of Indonesian CMs, resulting in low genetic and nucleotide diversities (Blancher *et al.*, 2008). Importantly, however, as the Philippine CMs involved in this study at least originated from Luzon and Mindanao, the results in this study may reflect the frequency of TRIMCyp

in Philippine CMs as a whole, but do not represent local TRIMCyp distribution. In addition, hybridization with RMs may affect the prevalence of TRIMCyp. As Chinese RMs have been reported to have a low frequency of TRIMCyp (Newman *et al.*, 2008; Wilson *et al.*, 2008a), it is possible that interspecies mating with Chinese RMs might result in a lower prevalence of TRIMCyp in the Malaysian and Indonesian populations. In any case, it will be of great interest to determine the allele frequency of TRIMCyp in wild CMs to confirm whether our results reflect the observations in nature.

It is worth noting that the habitat of PMs is close to that of CMs, and in fact both species inhabit Indonesia; however, PMs reportedly express TRIMCyp but not TRIM5α (Brennan *et al.*, 2008; Liao *et al.*, 2007). In contrast, the allele frequency of TRIMCyp in Indonesian CMs was shown to be markedly lower (Table 1). This discrepancy in frequency of TRIMCyp between PMs and CMs suggests that the two species have independently evolved antiretroviral factors to counteract some pathogen(s) existing in their habitats. It is possible that unidentified co-factors that interact with TRIM5α/TRIMCyp may have a role in this discrepancy. Alternatively, the pathogen(s) could develop severe diseases in either monkey species. In the case of RMs, whilst the allele frequency of TRIMCyp was approximately 25 % in the Indian population, TRIMCyp was not detected in the Chinese population (Wilson *et al.*, 2008a). Although the precise reason(s) for these geographical deviations in CMs and RMs is still unknown, it is reasonable to speculate that the possible pathogens, including exogenous and endogenous retroviruses, are/were heterogeneously disseminated, depending on their habitats.

The amino acid sequence of the CypA domain of our CM TRIMCyp-major (DK) is identical to that of Mafa TRIMCyp2 cloned by Ylinen *et al.* (2010); thus, CM TRIMCyp-major (DK) showed almost identical antiviral properties to those of Mafa TRIMCyp2. However, CM TRIMCyp-major (DK) slightly restricted HIV-2 GH123, although Mafa TRIMCyp2 failed to restrict HIV-2 ROD. This discrepancy is possibly due to differences in assays; Ylinen and co-workers performed a single-round infection assay using replication-incompetent virus, whereas we performed a multiple-round replication assay using replication-competent virus and thus our assay could detect weak restriction activities. It is also possible that differences in HIV-2 strains or TRIMCyp amino acid differences outside the CypA domain could affect the result.

In the case of CM TRIMCyp-minor (NE), the amino acid sequence of the CypA domain was identical to that of RM TRIMCyp, and antiviral properties of CM TRIMCyp-minor (NE) were the same as those of RM TRIMCyp. In addition, exon 8 of both TRIMCyp genes showed a uniform sequence, identical to that of the Mamu 7 haplotype of RMs. Exon 8 of TRIMCyp would have been free from selection pressures, as it is absent from the mRNA due to splicing, and the ancestral sequences in exon 8 would have been preserved. Taken together, it is reasonable to speculate that this minor haplotype of CM TRIMCyp was the ancestor when CMs separated from RMs, and the major haplotype of CM TRIMCyp has arisen due to a specific evolutionary pressure on CMs. It should be noted that CM TRIM5α has Q at aa 339, where RM TRIM5α has a Q→TFP polymorphism. This Q→TFP polymorphism in the PRYSPRY domain also altered the spectrum of anti-lentiviral activity of TRIM5α (Kirmaier *et al.*, 2010; Kono *et al.*, 2008; Lim *et al.*, 2010; Wilson *et al.*, 2008b). Therefore, it is tempting to speculate that the selection pressure in CMs drove amplification and diversification in TRIMCyp, whilst that in RMs drove diversification of the PRYSPRY domain of TRIM5α.

In parallel with our study, Dietrich *et al.* (2011) recently reported the prevalence and functional diversity of TRIMCyp in CMs. They analysed populations from Indonesia, Indochina, Mauritius and the Philippines, and found that TRIMCyp was present in populations from Indonesia, Indochina and the Philippines, but not in populations from Mauritius. As they mentioned, the low genetic diversity, probably due to founder effects, may have led to the absence of TRIMCyp in the Mauritian population. In contrast, the small number of animals analysed may have resulted in the absence of TRIM5α in their Philippine population. They also analysed the effects of DK→NE substitution in CM TRIMCyp on antiretroviral activity by mutagenesis techniques. Furthermore, they found a unique individual with the DE haplotype in the CypA domain of TRIMCyp, whilst we did not identify such a haplotype in our study. Their results were essentially in accordance with ours, and we further demonstrated that Philippine CMs possessed TRIM5α as well as TRIMCyp, suggesting that maintenance of both TRIM5α and TRIMCyp in the CM population is beneficial to counteract challenges by retroviruses that are susceptible to TRIM5α and by those susceptible to TRIMCyp. Consistent with this, Reynolds *et al.* (2011) demonstrated that heterozygotes of RMs with TRIM5α and TRIMCyp showed higher resistance to repeated intrarectal challenge of SIVsmE660 compared with homozygotes for TRIM5α or TRIMCyp. Interestingly, this different outcome was not observed in the case of intrarectal challenge with SIVmac239. As RM TRIMCyp restricts SIVsm but not SIVmac (Kirmaier *et al.*, 2010), the combination of TRIM5α and TRIMCyp may function more efficiently as an antiviral factor against SIVsm.

We saw a small difference in anti-HIV-1 activity between CM TRIMCyp-minor (NE) and TRIMCyp-minor R285G. Dietrich *et al.* (2011) suggested that either of two polymorphic amino acid residues, K209E and R285G, might be responsible for attenuated anti-feline immunodeficiency virus activity of a certain haplotype of CM TRIMCyp. Our CM TRIMCyp-minor (NE) had K at aa 209, and an additional R285G mutation slightly attenuated the anti-HIV-1 activity of CM TRIMCyp-minor (NE). Residue 285 is in the linker region between the coiled-coil and CypA domains. The precise mechanism of how aa 285 affects anti-HIV-1 activity is unclear at present, but our result was consistent with those of Dietrich *et al.* (2011) and further revealed the importance of a single amino acid

substitution at aa 285 on the antiviral activity of CM TRIMCyp.

We showed that a prototypic HIV-1mt, named NL-DT5R, encoding L4/5 of SIVmac239 CA instead of that derived from HIV-1, evaded restriction by the major haplotype of CM TRIMCyp. As only HIV-1-derived L4/5 but not the SIVmac-derived L4/5 is expected to bind to CypA (Franke *et al.*, 1994), the substitution of L4/5 results in loss of binding of the capsid from CypA as well as TRIMCyp. Moreover, we recently demonstrated that HIV-1mt has the ability to grow in CMs (Saito *et al.*, 2011). Retrospective analysis of the *TRIM5* genotypes of the infected CMs revealed that they were homozygous for TRIMCyp (data not shown), suggesting that TRIMCyp homozygotes allow the replication of HIV-1mt *in vivo*. These findings will be helpful not only to understand the molecular mechanisms of the species barrier of primates to lentiviruses, but also to emphasize the importance of *TRIM5* genotypes for future studies regarding non-human primate models for HIV-1 infection.

METHODS

Sample collection. Blood samples were obtained from CMs kept in the Tsukuba Primate Research Center (TPRC), National Institute of Biomedical Innovation, Tsukuba, Japan. CMs have been maintained in indoor facilities as closed colony monkeys in TPRC since 1978 (Honjo, 1985). CMs in TPRC were obtained from Indonesia, Malaysia and the Philippines. Although the detailed local information of their origin is unclear, more than 100 animals were introduced to each colony by dividing it several times. Basically, the monkeys have been bred as pure blood of each origin without interbreed crossing. The generation number of animals involved in this study ranged from two to four when we consider the wild-caught founders (introduced monkeys) as zero. These animals were maintained according to the rules of the National Institute of Biomedical Innovation and guidelines for experimental animal welfare. Bleeding was performed under ketamine hydrochloride anaesthesia.

PCR amplification and sequence analysis. Genomic DNA was extracted from peripheral blood mononuclear cells (PBMCs) of 126 CMs using a QIAamp DNA Blood Mini kit (Qiagen). To test for the *CypA* insertion, the 3′ region of the *TRIM5* gene was amplified by PCR using LA *Taq* (TaKaRa) with primers TC forward (5′-TGACTCTGTGCTCACCAAGCTCTTG-3′) and TC reverse (5′-ACCCTACTATGCAATAAAACATTAG-3′), as described by Wilson *et al.* (2008a). The amplified products of *CypA* from 30 TRIMCyp homozygotes and 32 TRIMCyp/TRIM5α heterozygotes were gel-purified and subjected to direct sequencing using the forward and reverse primers.

To determine the sequences of the RING, B-box, coiled-coil and linker domains of TRIM5α and TRIMCyp, which span >15 kb of genomic DNA, we prepared phytohaemagglutinin (PHA)-stimulated PBMCs from six TRIMCyp homozygotes and three TRIM5α homozygotes. Total RNA was extracted from these cells using TRIzol (Invitrogen), and the RNA was reverse-transcribed using SuperScript III reverse transcriptase (Invitrogen) with TC reverse primer for TRIMCyp or TRIM5 reverse primer (5′-GAATT-CTCAAGAGCTTGGTGA-3′) for TRIM5α. The resultant cDNA was then PCR-amplified with LA *Taq* and forward primer TRIM5-235F (5′-GCAGGACCAGTGGAATAGC-3′). The amplified products were purified and subjected to direct sequencing using primers TRIM5-235F, TRIM-N (5′-AGGCAGAAGCAGCAGGAA-3′), TRIM-Nrev (5′-TTCCTGCTGCTTCTGCCT-3′) and TRIM-E (5′-ACCTCCCAG-TAATGTTTC-3′). As the direct sequencing results of exons 5 and 6 of TRIMCyp were ambiguous because of the existence of the other splicing variant containing exons 1–4 combined with *CypA* (Brennan *et al.*, 2008), amplified products were then cloned into the vector pCR-2.1TOPO (Invitrogen) and the nucleotide sequences of numerous independent clones (between three and nine) for each TRIMCyp were determined.

Exon 8 (PRYSPRY domain) was PCR-amplified from 12 TRIM5α homozygotes and seven TRIMCyp homozygotes by using TRIM-genotyping forward (5′-CTTCTGAACAAGTTTCCTCCCAG-3′) and reverse (5′-ATGAGATGCACATGGACAAGAGG-3′) primers. The amplified products were purified and subjected to direct sequencing using the TRIM genotyping forward and reverse primers.

Cloning and expression of TRIMCyp. cDNA of the major haplotype of CM TRIMCyp, CM TRIMCyp-major (DK), was amplified by RT-PCR of mRNA extracted from the TRIM5α/TRIMCyp-heterozygous CM T-cell line HSC-F using Not7TRIM5 (5′-GCGGCCGCAGCTACTATGGCTTCTG-3′) as the forward primer (*Not*I site underlined) and CypA Rev (5′-ACGGCGGTCT-TTTCATTCGAGTTGTCC-3′) as the reverse primer. RM TRIMCyp cDNA was amplified by RT-PCR of mRNA extracted from the TRIMCyp homozygous RM T-cell line HSR5.4 using Not7TRIM5 as the forward primer and CypA Rev as the reverse primer. The amplified products were then cloned into pCR-2.1TOPO and the authenticity of the nucleotide sequence was verified. To generate TRIMCyp cDNAs carrying a haemagglutinin (HA; YPYDVPDYAA) tag at the C terminus, the TRIMCyp cDNA clones were used as templates for PCR amplification with a primer including a *Not*I site and an HA tag.

To generate the minor haplotype, CM TRIMCyp-minor (NE), the C-terminal portion of RM TRIMCyp (*Sal*I–*Not*I) and the N-terminal portion of CM TRIMCyp-major (DK) (*Not*I–*Sal*I) were assembled in the pcDNA3.1 (−) vector (Invitrogen). CM TRIMCyp-minor R285G was generated by site-directed mutagenesis by a PCR-mediated overlap primer-extension method.

The entire coding sequences of these TRIMCyps were then transferred to the *Not*I site of the pSeV18+b (+) vector. Recombinant SeVs carrying various TRIMCyp were recovered according to a previously described method (Nakayama *et al.*, 2005). The viruses were passaged twice in embryonated chicken eggs and used as stocks for all experiments.

Virus propagation. Virus stocks were prepared by transfection of 293T cells with HIV-1 NL4-3, HIV-2 GH123, SIVmac239 and HIV-1mt NL-DT5R (Kamada *et al.*, 2006) using a calcium phosphate co-precipitation method. Virus titres were measured using p24 (for HIV-1 and HIV-1mt) or p27 (for HIV-2 and SIVmac239) RetroTek antigen ELISA kits (ZeptoMetrix).

Virus infection. Aliquots of 2×10^5 MT4 cells were infected with SeV expressing CM TRIM5α or each TRIMCyp at an m.o.i. of 10 and incubated at 37 °C for 9 h. Cells were then superinfected with 20 ng HIV-1 NL4-3 or HIV-1mt DT5R p24, 20 ng HIV-2 GH123 p25 or 20 ng SIVmac239 p27. The culture supernatants were collected periodically, and the levels of p24, p25 and p27 were measured with a RetroTek antigen ELISA kit.

ACKNOWLEDGEMENTS

The authors wish to thank Tomoko Ikoma, Setsuko Bandou and Noriko Teramoto for their helpful assistance. This work was supported by grants from the Ministry of Education, Culture,

Sports, Science, and Technology, the Ministry of Health, Labor, and Welfare in Japan, Global COE Program A06 of Kyoto University and Environment Research and Technology Development Fund (D-1007) of the Ministry of the Environment, Japan.

REFERENCES

Abegg, C. & Thierry, B. (2002). Macaque evolution and dispersal in insular south-east Asia. *Biol J Linn Soc Lond* **75**, 555–576.

Agy, M. B., Frumkin, L. R., Corey, L., Coombs, R. W., Wolinsky, S. M., Koehler, J., Morton, W. R. & Katze, M. G. (1992). Infection of *Macaca nemestrina* by human immunodeficiency virus type-1. *Science* **257**, 103–106.

Blancher, A., Bonhomme, M., Crouau-Roy, B., Terao, K., Kitano, T. & Saitou, N. (2008). Mitochondrial DNA sequence phylogeny of 4 populations of the widely distributed cynomolgus macaque (*Macaca fascicularis fascicularis*). *J Hered* **99**, 254–264.

Brennan, G., Kozyrev, Y. & Hu, S.-L. (2008). TRIMCyp expression in Old World primates *Macaca nemestrina* and *Macaca fascicularis*. *Proc Natl Acad Sci U S A* **105**, 3569–3574.

Dietrich, E. A., Brennan, G., Ferguson, B., Wiseman, R. W., O'Connor, D. & Hu, S.-L. (2011). Variable prevalence and functional diversity of the antiretroviral restriction factor TRIMCyp in *Macaca fascicularis*. *J Virol* **85**, 9956–9963.

Franke, E. K., Yuan, H. E. & Luban, J. (1994). Specific incorporation of cyclophilin A into HIV-1 virions. *Nature* **372**, 359–362.

Honjo, S. (1985). The Japanese Tsukuba Primate Center for Medical Science (TPC): an outline. *J Med Primatol* **14**, 75–89.

Johnson, W. E. & Sawyer, S. L. (2009). Molecular evolution of the antiretroviral *TRIM5* gene. *Immunogenetics* **61**, 163–176.

Kamada, K., Igarashi, T., Martin, M. A., Khamsri, B., Hatcho, K., Yamashita, T., Fujita, M., Uchiyama, T. & Adachi, A. (2006). Generation of HIV-1 derivatives that productively infect macaque monkey lymphoid cells. *Proc Natl Acad Sci U S A* **103**, 16959–16964.

Kirmaier, A., Wu, F., Newman, R. M., Hall, L. R., Morgan, J. S., O'Connor, S., Marx, P. A., Meythaler, M., Goldstein, S. & other authors (2010). *TRIM5* suppresses cross-species transmission of a primate immunodeficiency virus and selects for emergence of resistant variants in the new species. *PLoS Biol* **8**, e1000462.

Kita, Y. F., Hosomichi, K., Kohara, S., Itoh, Y., Ogasawara, K., Tsuchiya, H., Torii, R., Inoko, H., Blancher, A. & other authors (2009). MHC class I A loci polymorphism and diversity in three Southeast Asian populations of cynomolgus macaque. *Immunogenetics* **61**, 635–648.

Kono, K., Song, H., Shingai, Y., Shioda, T. & Nakayama, E. E. (2008). Comparison of anti-viral activity of rhesus monkey and cynomolgus monkey TRIM5αs against human immunodeficiency virus type 2 infection. *Virology* **373**, 447–456.

Liao, C.-H., Kuang, Y.-Q., Liu, H.-L., Zheng, Y.-T. & Su, B. (2007). A novel fusion gene, TRIM5–Cyclophilin *A* in the pig-tailed macaque determines its susceptibility to HIV-1 infection. *AIDS* **21** (Suppl. 8), S19–S26.

Lim, S.-Y., Rogers, T., Chan, T., Whitney, J. B., Kim, J., Sodroski, J. & Letvin, N. L. (2010). TRIM5α modulates immunodeficiency virus control in rhesus monkeys. *PLoS Pathog* **6**, e1000738.

Nakayama, E. E. & Shioda, T. (2010). Anti-retroviral activity of TRIM5α. *Rev Med Virol* **20**, 77–92.

Nakayama, E. E., Miyoshi, H., Nagai, Y. & Shioda, T. (2005). A specific region of 37 amino acid residues in the SPRY (B30.2) domain of African green monkey TRIM5α determines species-specific restriction of simian immunodeficiency virus SIVmac infection. *J Virol* **79**, 8870–8877.

Newman, R. M., Hall, L., Connole, M., Chen, G.-L., Sato, S., Yuste, E., Diehl, W., Hunter, E., Kaur, A. & other authors (2006). Balancing selection and the evolution of functional polymorphism in Old World monkey TRIM5α. *Proc Natl Acad Sci U S A* **103**, 19134–19139.

Newman, R. M., Hall, L., Kirmaier, A., Pozzi, L. A., Pery, E., Farzan, M., O'Neil, S. P. & Johnson, W. (2008). Evolution of a TRIM5-CypA splice isoform in Old World monkeys. *PLoS Pathog* **4**, e1000003.

Nomaguchi, M., Doi, N., Kamada, K. & Adachi, A. (2008). Species barrier of HIV-1 and its jumping by virus engineering. *Rev Med Virol* **18**, 261–275.

Price, A. J., Marzetta, F., Lammers, M., Ylinen, L. M., Schaller, T., Wilson, S. J., Towers, G. J. & James, L. C. (2009). Active site remodeling switches HIV specificity of antiretroviral TRIMCyp. *Nat Struct Mol Biol* **16**, 1036–1042.

Reynolds, M. R., Sacha, J. B., Weiler, A. M., Borchardt, G. J., Glidden, C. E., Sheppard, N. C., Norante, F. A., Castrovinci, P. A., Harris, J. J. & other authors (2011). The TRIM5α genotype of rhesus macaques affects acquisition of simian immunodeficiency virus SIVsmE660 infection after repeated limiting-dose intrarectal challenge. *J Virol* **85**, 9637–9640.

Saito, A., Nomaguchi, M., Iijima, S., Kuroishi, A., Yoshida, T., Lee, Y.-J., Hayakawa, T., Kono, K., Nakayama, E. E. & other authors (2011). Improved capacity of a monkey-tropic HIV-1 derivative to replicate in cynomolgus monkeys with minimal modifications. *Microbes Infect* **13**, 58–64.

Sauter, D., Specht, A. & Kirchhoff, F. (2010). Tetherin: holding on and letting go. *Cell* **141**, 392–398.

Song, H., Nakayama, E. E., Yokoyama, M., Sato, H., Levy, J. A. & Shioda, T. (2007). A single amino acid of the human immunodeficiency virus type 2 capsid affects its replication in the presence of cynomolgus monkey and human TRIM5αs. *J Virol* **81**, 7280–7285.

Stremlau, M., Owens, C. M., Perron, M. J., Kiessling, M., Autissier, P. & Sodroski, J. (2004). The cytoplasmic body component TRIM5α restricts HIV-1 infection in Old World monkeys. *Nature* **427**, 848–853.

Wilson, S. J., Webb, B. L., Ylinen, L. M., Verschoor, E., Heeney, J. L. & Towers, G. J. (2008a). Independent evolution of an antiviral TRIMCyp in rhesus macaques. *Proc Natl Acad Sci U S A* **105**, 3557–3562.

Wilson, S. J., Webb, B. L., Maplanka, C., Newman, R. M., Verschoor, E. J., Heeney, J. L. & Towers, G. J. (2008b). Rhesus macaque TRIM5 alleles have divergent antiretroviral specificities. *J Virol* **82**, 7243–7247.

Yap, M. W., Nisole, S., Lynch, C. & Stoye, J. P. (2004). Trim5α protein restricts both HIV-1 and murine leukemia virus. *Proc Natl Acad Sci U S A* **101**, 10786–10791.

Ylinen, L. M., Price, A. J., Rasaiyaah, J., Hué, S., Rose, N. J., Marzetta, F., James, L. C. & Towers, G. J. (2010). Conformational adaptation of Asian macaque TRIMCyp directs lineage specific antiviral activity. *PLoS Pathog* **6**, e1001062.

Higashino et al. Genome Biology 2012, 13:R58
http://genomebiology.com/2012/13/7/R58

Genome Biology

RESEARCH Open Access

Whole-genome sequencing and analysis of the Malaysian cynomolgus macaque (*Macaca fascicularis*) genome

Atsunori Higashino[1,2], Ryuichi Sakate[1*], Yosuke Kameoka[1], Ichiro Takahashi[1], Makoto Hirata[1], Reiko Tanuma[1], Tohru Masui[1], Yasuhiro Yasutomi[3] and Naoki Osada[4,5*]

Abstract

Background: The genetic background of the cynomolgus macaque (*Macaca fascicularis*) is made complex by the high genetic diversity, population structure, and gene introgression from the closely related rhesus macaque (*Macaca mulatta*). Herein we report the whole-genome sequence of a Malaysian cynomolgus macaque male with more than 40-fold coverage, which was determined using a resequencing method based on the Indian rhesus macaque genome.

Results: We identified approximately 9.7 million single nucleotide variants (SNVs) between the Malaysian cynomolgus and the Indian rhesus macaque genomes. Compared with humans, a smaller nonsynonymous/synonymous SNV ratio in the cynomolgus macaque suggests more effective removal of slightly deleterious mutations. Comparison of two cynomolgus (Malaysian and Vietnamese) and two rhesus (Indian and Chinese) macaque genomes, including previously published macaque genomes, suggests that Indochinese cynomolgus macaques have been more affected by gene introgression from rhesus macaques. We further identified 60 nonsynonymous SNVs that completely differentiated the cynomolgus and rhesus macaque genomes, and that could be important candidate variants for determining species-specific responses to drugs and pathogens. The demographic inference using the genome sequence data revealed that Malaysian cynomolgus macaques have experienced at least three population bottlenecks.

Conclusions: This list of whole-genome SNVs will be useful for many future applications, such as an array-based genotyping system for macaque individuals. High-quality whole-genome sequencing of the cynomolgus macaque genome may aid studies on finding genetic differences that are responsible for phenotypic diversity in macaques and may help control genetic backgrounds among individuals.

Background

Cynomolgus macaque (*Macaca fascicularis*) is one of the most commonly used nonhuman primates in biomedical research worldwide [1]. It is also called the crab-eating or long-tailed macaque and belongs to the *fascicularis* group of the genus *Macaca* [2]. A number of pharmaceutical companies use cynomolgus macaques for drug development and, thus, identifying genetic components that contribute to their drug metabolism is a key issue in biomedical genomic research [3,4].

Rhesus macaque (*Macaca mulatta*), whose draft genome sequence was determined by the Sanger sequencing method with a BAC clone assembly [5], is genetically closely related to the cynomolgus macaque. Whereas rhesus macaques occur from India to southern China and in some neighboring areas, cynomolgus macaques can be found throughout Southeast Asia. Vital hybrids of the two macaques have been observed around northern Thailand, supporting their very close genetic relationship [6]. Previous studies have shown that cynomolgus and rhesus macaques share a considerable number of single

* Correspondence: rsakate@nibio.go.jp; nosada@nig.ac.jp
[1]Laboratory of Rare Disease Biospecimen, Department of Disease Bioresources Research, National Institute of Biomedical Innovation, 7-6-8 Saito-asagi, Ibaraki, Osaka 567-0085, Japan
[4]Division of Evolutionary Genetics, Department of Population Genetics, National Institute of Genetics, 1111 Yata, Mishima, Shizuoka 411-8540, Japan
Full list of author information is available at the end of the article

nucleotide variants (SNVs) [7,8]. Their genetic divergence is estimated to be approximately 0.4% [8,9].

Recently, several genome sequences of macaques have been determined using next-generation sequencing platforms. These include Mauritian and Vietnamese cynomolgus macaques [4,10], two independent Chinese rhesus macaques [10,11] and one Indian rhesus macaque [12]. The two cynomolgus macaque individuals (Mauritian and Vietnamese), however, were derived from two genetically distinct populations that have experienced peculiar demographic histories. Previous studies have suggested that cynomolgus macaques are genetically clustered into Indonesian-Malaysian, Philippine, Indochinese, and Mauritian macaques [8,13]. Mauritian macaques have been known to show extremely low genetic diversity that is associated with their recent colonization [14], whereas Indochinese macaques have experienced a considerable amount of gene flow with rhesus macaques [15,16]. Therefore, the whole-genome sequencing of Indonesian-Malaysian cynomolgus macaques, which show the highest genetic diversity and, according to the fossil evidence, originate from a putative ancestral population [17], would provide significant insight into the genetic differentiation of cynomolgus and rhesus macaques at the species level.

Recent advances in DNA sequencing technologies have enabled rapid and economical determination of whole-genome sequences of organisms. Although *de novo* assemblies of large and complicated genomes, such as mammalian genomes, remain difficult, whole-genome resequencing has become a powerful method for identifying genetic variation within a biological species. Human genome variation is of particular interest for medical and evolutionary studies, and a dozen human genome sequences have thus far been determined using resequencing methods [18-24]. Whole-genome resequencing is not only efficient for identifying variations within a species, but also applicable to closely related species. Because the current methods of mapping short DNA sequence reads have been developed to amend relatively high sequencing errors in massively parallel sequencing, they are also expected to be useful for small sequence divergence. Thus, the strategy of resequencing species that are closely related to model organisms of known genome sequence may be an efficient and important method for detecting genomic diversity.

In this study, we determined and analyzed the Malaysian cynomolgus macaque genome sequence using the massively parallel sequencer SOLiD 3 Plus System (Life Technologies). The sequenced reads were mapped to the Indian rhesus macaque (reference) genome sequence with more than 40-fold coverage. A total of approximately 9.7 million SNVs and 1 million small (< 12 bp) indels and 60,000 large indels (44 to 732 bp) were identified. The identified SNVs were compared with SNVs previously determined for other cynomolgus and rhesus macaque genomes. These SNVs have been deposited in the cynomolgus macaque genome resources database (QFbase [25]). High-quality resequencing of the cynomolgus macaque genome will facilitate further studies directed towards dissecting genetic differences that are responsible for phenotypic divergence among macaque species.

Results

Sequencing and mapping

Blood samples from a 25-year-old male Malaysian cynomolgus macaque were used for genome resequencing. Figure 1 outlines the procedure of the cynomolgus macaque genome resequencing. We performed eight cycles of fragment library sequencing (50 bp) and four cycles of mate-pair library sequencing (2 × 25 bp) using the SOLiD 3 Plus System. The mate-pair libraries of two different insert sizes (600 to 800 bp and 800 to 1,000 bp) were constructed and analyzed. Table 1 summarizes the results of genome sequencing and mapping. A total of 2.6×10^9 reads of fragment sequence and 2.2×10^9 reads of mate-pair sequence data were obtained. The mapping program implemented in BioScope software v1.3.1 (Life Technologies) was used for mapping the reads. A total of 3.4×10^9 reads (69.8%) were successfully mapped on the Golden Path genome assembly, which was derived from an Indian rhesus macaque (mmu_120505). Finally, analyzed reads totaled 1.1×10^{11} bp, and the average coverage depth was 41.5-fold. All chromosomes exceeded 37-fold (Figure S1 in Additional file 1). The analyzed reads covered 99.7% of the reference genome (unmapped Ns were excluded), and 95.8% of the reference genome was covered by at least 10 reads (Figure S2 in Additional file 1). In order to examine whether our mapping statistics depended on the genome assembly, we also mapped our reads to the recently determined Vietnamese cynomolgus macaque genome, which was constructed by *de novo* assembly of short reads [10]. As a result, a similar mapping rate level (67.2%) and genome coverage (42.5-fold) were obtained (Table S1 in Additional file 1). We primarily focus on the results obtained using the Golden Path genome assembly throughout the rest of the paper because the reference genome had more detailed genome annotations, and the results are comparable with those of other studies. Hereafter, we refer to the Golden Path genome assembly as the "reference" genome.

Single nucleotide variant detection

SNVs were called with SAMtools [26] using the mapped reads on the reference genome. SNVs at low (< 5) coverage sites and with low call quality values (QV < 40) were excluded. Because the reference genome sequence has

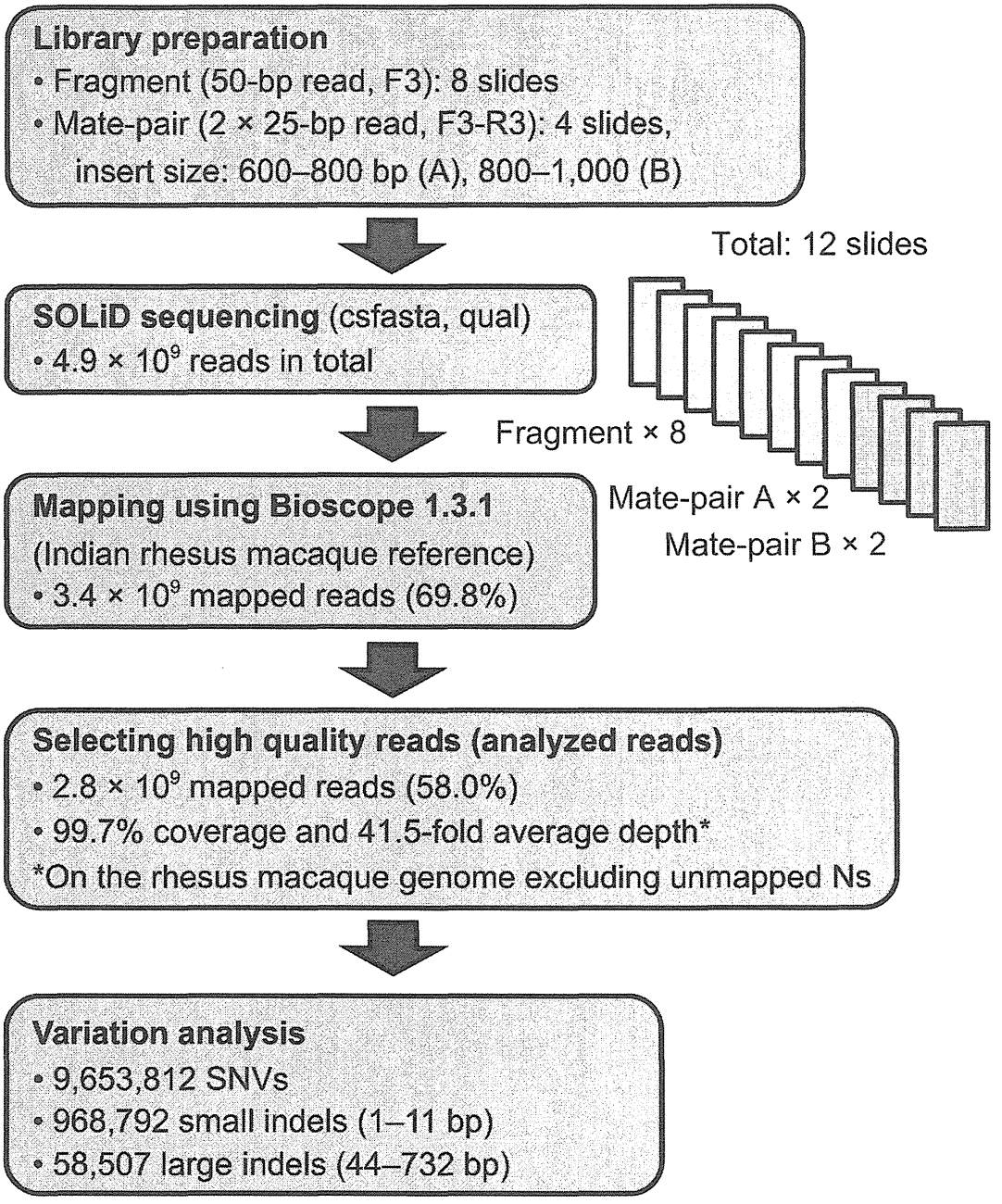

Figure 1 Overview of the Malaysian cynomolgus macaque genome sequencing and analysis. Fragment runs of eight slides and mate-pair runs of four slides (insert size: two slides for 600 to 800 bp and two slides for 800 to 1,000 bp) were performed on the SOLiD 3 Plus System. In total, 4.9×10^9 sequence reads were generated and mapped on the reference genome. After the high-quality reads were selected, single nucleotide variants (SNVs) and indel analyses were conducted.

not yet been finalized, we examined the relationship between the quality of the reference genome assembly and the SNV discovery rate. We expected that homozygous SNVs in low-quality genomic regions were possible errors in the reference genome sequence and that heterozygous SNVs were robust in genome quality. As shown in Figure 2, we plotted the proportions of homozygous and heterozygous SNVs against the reference genome QVs. Although the heterozygous SNV discovery rate was nearly constant across genome quality, homozygous

Table 1 Summary of SOLiD libraries and sequence reads

Library	Read length (bp)	Insert size (bp)	Runs	Reads	Mapped reads	Analyzed reads[a]	Coverage depth of analyzed reads
Fragment	50	-	8	2,648,128,521	1,976,720,560 (74.7%)	1,974,496,337 (74.6%)	33.4
Mate-pair A	25 (×2)	600-800	2	906,783,481	621,175,871 (68.5%)	355,589,008 (39.2%)[b]	3.4
Mate-pair B	25 (×2)	800-1,000	2	1,335,583,547	814,866,634 (61.0%)	508,168,736 (38.0%)[b]	4.8
Total	-	-	12	4,890,495,549	3,412,763,065 (69.8%)	2,838,254,081 (58.0%)	41.5

[a]Reads mapped on chrM and chrUr were removed. [b]'PCR or optical duplicates' (defined by Bioscope; mapped more than 100 loci) were removed, and properly paired reads were selected; each read of a pair was mapped on the same chromosome in a proper direction at a proper distance from each other.

SNV rates in low-quality regions were relatively high, suggesting that those SNVs were probably due to errors in the genome sequence and should be filtered out. In addition, we observed a slight peak in homozygous SNV rates at QV around 40. This pattern was also observed when we removed SNVs within repeat regions (data not shown) and may have been due to unknown problems in the assembly process of the reference genome sequence. Based on this observation, we decided to filter out SNVs at sites having QVs < 45 in the reference genome sequence. This filtering did not significantly sacrifice our SNV detection power, because > 94% of the reference rhesus macaque genome had QV = 60.

Using the above criteria, we identified 4,880,874 heterozygous and 4,527,169 homozygous SNVs on autosomes. The number of estimated SNVs is summarized in Table 2. Note that the numbers in this table are underestimates because SNVs ambiguously assigned as either homozygous or heterozygous were not included (see Materials and methods). In autosomal non-coding regions, 42,930 untranslated exonic (5'/3' UTR), 2,878,903 intronic, and 6,422,898 intergenic SNVs were identified. Among them, 3,707,670 SNVs were mapped to repeat regions. The nucleotide change pattern of the SNVs is shown in Table S2 in Additional file 1. The transition-to-transversion ratio was 2.39, which is close to the estimated value in humans [27]. SNV densities on chromosomes are summarized in Figure S3 in Additional file 1. Using the same SNV-detecting criteria, we identified about 8.5 million SNVs by mapping Malaysian cynomolgus macaque reads on the Vietnamese cynomolgus macaque genome sequences.

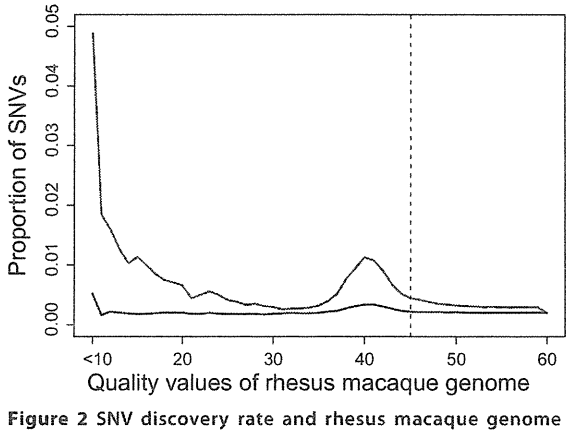

Figure 2 SNV discovery rate and rhesus macaque genome quality. The red and blue lines represent the rates of homozygous and heterozygous SNVs, respectively, with given rhesus macaque genome sequence quality values (QVs). SNVs at sites having QV < 45 (left of the dashed line) were filtered out.

Among 18,912 annotated autosomal protein-coding genes, 14,560 carried at least one coding SNV, consisting of 25,079 nonsynonymous and 38,233 synonymous SNVs. We found that 9,753 autosomal genes contained at least one heterozygous or homozygous amino acid variation in the Malaysian cynomolgus macaque genome, compared with the reference rhesus macaque genome. In addition, 108 and 200 autosomal genes harbored nonsense mutations that were homozygous and heterozygous, respectively. We also estimated the number of SNVs on the X; chromosome. Only homozygous SNVs on the X chromosome were counted. In total, we identified 245,769 SNVs on the X; chromosome, including 1,145 coding (444 nonsynonymous and 701 synonymous SNVs in 662 protein-coding genes), 986 UTR, 50,877 intronic, and 192,761 intergenic homozygous SNVs (Table 2).

Comparisons with previously determined macaque genomes

The newly identified whole-genome SNVs between Malaysian cynomolgus and Indian (reference) rhesus macaques were compared with previously identified SNVs. We downloaded short-read sequences of Vietnamese cynomolgus and Chinese rhesus macaques that had comparable coverage depth to ours (> 40-fold) and mapped on the reference genome [10]. Using the same SNV-detecting pipeline, we identified 13,244,140 and 10,662,418 SNVs in the Vietnamese cynomolgus and Chinese rhesus macaque genomes, respectively. The Malaysian cynomolgus macaque shared 5,181,509 SNVs

Table 2 Number of single nucleotide variants

Chromosome	Heterozygous SNVs	Homozygous SNVs	A[a]	S[b]	UTR[c]	Intronic	Intergenic
Autosomes	4,880,874	4,527,169	25,079	38,233	42,930	2,878,903	6,422,898
X chromosomes	-[d]	245,769	444	701	986	50,877	192,761
Total	4,880,874	4,772,938	25,523	38,934	43,916	2,928,970	6,615,659

[a]Number of nonsynonymous SNVs. [b]Number of synonymous SNVs. [c]Number of SNVs in untranslated regions. [d]Only homozygous SNVs were considered on the X chromosome.

with the Vietnamese cynomolgus macaque, either homozygous or heterozygous, showing that > 50% of our SNVs were shared between the two cynomolgus macaque individuals. Merging the two cynomolgus macaque genomes yielded 17,716,443 SNVs in cynomolgus macaques. Furthermore, we found that 2,519,988 SNVs were restricted to the Malaysian cynomolgus macaque, and 1,368,528 SNVs were completely differentiated between the two cynomolgus and two rhesus macaque genomes. Because sequencing platforms and coverage depth differed among the studies, we could not directly compare the number of inferred SNVs. We therefore compared the fraction of heterozygous SNVs shared between two genomes. About 8% of Malaysian and 11% of Vietnamese heterozygous SNVs were also heterozygous SNVs in the Chinese rhesus macaque, supporting the contention that Indochinese cynomolgus macaques have been more vulnerable to gene introgression from rhesus macaques than Indonesian-Malaysian macaques.

We next searched for immune- and drug-response genes that carried nonsynonymous SNVs in the Malaysian cynomolgus macaque, because these genes are of particular interest in biomedical research. In total, 72 and 42 autosomal genes, of which the human orthologs had been annotated as immune-response (GO: 0006955) and drug-response (GO: 0042493) genes, respectively, had at least one homozygous amino acid change in the Malaysian cynomolgus macaque genome. We further checked whether these homozygous SNVs were likely to be differentiated between the two macaque species. A handful of genes, 29 immune- and 18 drug-response genes, carried completely segregating nonsynonymous SNVs between cynomolgus and rhesus macaques, for a total of 60 nonsynonymous SNVs (Table S3 in Additional file 1).

Population genetic inferences from resequenced data

In contrast to previous resequencing studies, the reference genome and the resequenced genome in this study were from highly differentiated but not completely isolated populations. The average genetic diversity in cynomolgus macaques (nucleotide diversity) corresponded to the fraction of heterozygous SNVs (differences between two sequenced chromosomes) if there was no consanguinity effect, whereas the average genetic divergence between species (Nei's d_{xy}) [28] corresponded to the fraction of homozygous SNVs plus one-half of the heterozygous SNVs.

In order to infer the strength of natural selection within and between macaque species, we estimated the ratio of nonsynonymous to synonymous SNVs. The ratio of nonsynonymous to synonymous heterozygous SNVs within cynomolgus macaques was 0.68. In order to compare the ratios in macaques and humans, a diploid human genome sequence determined by a short-read sequencer with similar read depth (African genome, NA19839) was retrieved from the public database. The human SNVs were determined using the same SNV-detecting pipeline described above. The ratio of nonsynonymous to synonymous heterozygous SNVs in the African human genome was 0.89, significantly higher than the ratio in the macaque ($P < 10^{-15}$, chi-square test). This pattern agrees well with the nearly neutral theory, in which slightly deleterious mutations tend to be segregated within small populations [29], because these macaques have four to five times larger effective population sizes than extant humans. In addition, the ratio within cynomolgus macaques (0.68) was slightly but statistically and significantly higher than that between cynomolgus and rhesus macaques (0.65; P = 0.002, chi-square test). If most of the nonsynonymous SNVs between cynomolgus and rhesus macaques were due to diversifying selection between species, the ratio of nonsynonymous to synonymous SNVs between species should be higher than that within species. This pattern also could be explained by the nearly neutral theory, wherein slightly deleterious mutations are short-lived and cannot contribute to species differentiation.

Small indels detected by sequence mapping

Using the mapping information of sequence reads, we also estimated the number of small indels (< 12 bp) in the Malaysian cynomolgus macaque genome. Interestingly, we observed a slight increase in small indels around QV = 40 of the reference genome sequence (Figure S4 in Additional file 1). We therefore filtered out small indels at sites with QV < 45 in the reference sequence. In total, we identified 365,581 insertions and 587,456 deletions on autosomes and 7,023 insertions and 8,732 deletions on the X chromosome. Only

homozygous indels were counted on the X chromosome. Out of 372,604 small insertions and 596,188 small deletions in total, 154,649 (42%) and 250,398 (42%) were assigned to repeat regions, respectively. Among 1,139 indels within autosomal protein-coding regions, 705 were frameshifting and 434 were non-frameshifting (3x-bp-length) indels. The proportion of 3x-bp-length indels (38%) was significantly higher than the value expected from intergenic indels (14%; $P < 10^{-15}$, chi-square test), suggesting purifying selection on frameshifting indels in coding regions. The distribution of small indel lengths is shown in Figure 3.

Large indels detected by mate-pair distance

An early chromosome-banding study suggested a paracentric chromosomal inversion in the long arm of chromosome 5 between cynomolgus and rhesus macaques [30]. In order to examine the occurrence of inversion at the chromosome-banding level (> 1 Mb), we surveyed mate-pair sequence reads that were not properly aligned on chromosome 5. The number of mate-pair reads showing the signature of inversion was counted within 500-bp-length windows with 250-bp sliding steps. In total, 28 windows that contained ≥ 50 incongruent reads were found. However, all of the windows included alpha satellite repeats and none showed evidence of the large inversion.

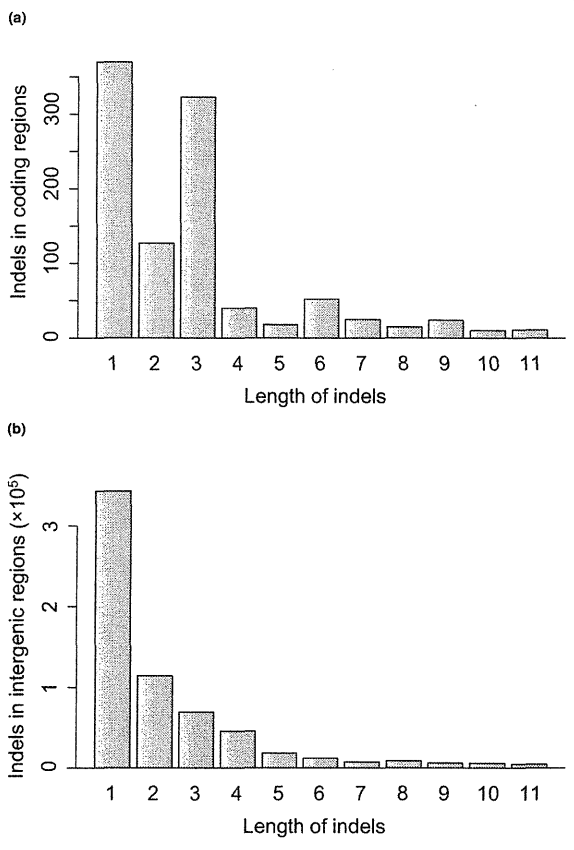

Figure 3 Distribution of small-indel lengths identified in the Malaysian cynomolgus macaque genome. (a) Indels in coding regions. **(b)** Indels in intergenic regions. The indels were identified using the information from high-quality reads.

We further analyzed the pattern of large insertions and deletions using the information from the mate-pair libraries of different insert sizes (mate-pair library A, 600 to 800 bp; library B, 800 to 1,000 bp). A total of 29,009 and 50,945 indels were identified using libraries A and B, respectively. Merging these indels yielded 8,301 insertions and 50,206 deletions; the insertion and deletion size ranges were 77 to 732 bp and 44 to 601 bp, respectively. Although the reference genome assembly has consecutive indices for each chromosome, the assembled genome sequences were built from scaffolds and contigs connected with assembly gaps (stretches of Ns). Among the 50,206 deletions, 45,821 and 22,774 encompassed repeat sequences and ambiguous sequences, respectively. Similarly, among the 8,301 insertions, 7,886 and 1,729 were within repeat sequences and ambiguous sequences, respectively. The distributions of insertion and deletion lengths that were not associated with gaps are shown in Figure S5 in Additional file 1.

Inference of demography

Recently, Li and Durbin [31] developed a novel method for inferring the demography of species from single diploid genome data. The demography is inferred from a distribution of coalescence time between two haploid genomes. We applied this method to our Malaysian cynomolgus genome data, with a generation time of 6 years and a mutation rate per generation of 2.5×10^{-8}. Figure 4 shows the inferred demography of the cynomolgus macaque with bootstrap 95% confidence interval. Although the scaling parameters affect the estimation of time and population size scales, the result showed at least three population bottlenecks in the past. In agreement with the previous estimates, the cynomolgus macaque population size expanded more than several fold during a million-year period [8,10,32].

Database resource

The Malaysian cynomolgus macaque genome sequence reads have been deposited to public databases (DDBJ Sequence Read Archive: DRA000430), and identified SNVs have been registered to the *Macaca fascicularis* genome database (QFbase [25]), which was previously built by our research group. The database was constructed based on the reference genome sequence of the Indian rhesus macaque, and the annotation of cynomolgus macaques was implemented, including cDNA

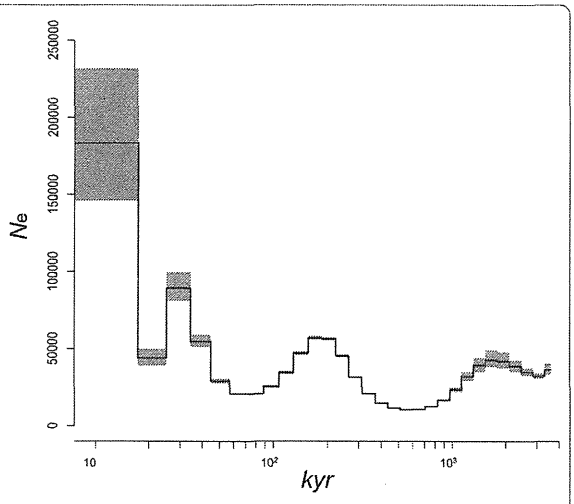

Figure 4 Estimation of ancestral population size of cynomolgus macaques with PSMC software. The *x* and *y* axes show population size and the time (thousand years (kyr)) from present, respectively. The blue rectangles represent 95% confidence interval with bootstrap resampling.

sequences, BAC clones, and microsatellite markers [9,33]. An example of a graphical view of SNVs in the browser is shown in Figure 5. Because cynomolgus macaques are frequently used in animal experiments, these resources will be valuable for researchers who are not familiar with large-scale data manipulation.

Discussion

Controlling the genetic background of experimental animals is a key issue for the efficiency and reliability of preclinical trials in biomedical research. Previous studies have shown that macaques, which are the most popular primates for biomedical research, harbor much higher genetic diversity than humans, even if they are collected from a limited area [8,15,32]. Thus, high-quality whole-genome sequences of cynomolgus macaques are necessary for future biomedical studies in order to control and quantify differences in genetic backgrounds. In addition, many morphological and physiological differences have been reported between the macaque species, including behaviors, tail lengths, body sizes, and susceptibility to pathogens and drugs [34,35]. Determining genetic differences between cynomolgus and rhesus macaques that contribute to phenotypic differences between them is an important subject for both biomedical and evolutionary research.

In this study, we have identified about 9.7 million SNVs between Malaysian cynomolgus and Indian rhesus macaques and 8.5 million SNVs between Malaysian and Vietnamese cynomolgus macaques. The total number of SNVs is much higher than that estimated in human genome resequencing studies (approximately 3 million). Although we cannot directly compare the number of SNVs determined with different platforms and different inference methods, the high level of genetic diversity within macaque species is in agreement with previous multi-locus sequencing studies using the Sanger method [8,32] and with the whole-genome sequencing study using a different platform with a similar level of genome coverage [10]. Despite the high level of genetic diversity within and between macaque species, the number of SNVs potentially responsible for species delimitation may be limited, partly owing to frequent gene flow between Indochinese cynomolgus and Chinese rhesus macaques. Only about 10% of SNVs were completely segregated between the two cynomolgus and two rhesus macaque genomes, which were further narrowed down to 60 nonsynonymous SNVs in drug- and immune-related genes.

The number of nonsynonymous SNVs was also higher in macaques than in humans. Whereas about 10,000 nonsynonymous SNVs were segregated in humans, about 30,000 nonsynonymous SNVs were segregated within and between macaque species. Interestingly, the level of protein diversity relative to background genetic diversity in macaques was significantly smaller than that expected from human data. This difference is probably due to the large effective population size of macaques, which removes slightly deleterious mutations in populations with relatively better efficiency.

Although we found a considerable number of SNVs and indels with high mapping support, we should be careful of some aspects of the quality of the reference genome assembly. In the large indel analysis using the mate-pair libraries, ≥ 90% of large indels included repeat sequences in the genome, indicating that these are potential repeat regions for genome-size change. Unfortunately, because the data we obtained using the SOLiD platform are not suitable for *de novo* assembly of a whole-genome sequence, we cannot conclude whether or not these hotspots are due to artifacts stemming from the reference genome quality. *De novo* assembly of a whole mammalian genome sequence remains costly, but studies using multiple genomes with *de novo* assembly would elucidate the complex pattern of genome-size changes [10].

The demography of the Malaysian cynomolgus macaque reveals the complex history of macaque genomes. As geological and fossil evidence has suggested, ancestors of the cynomolgus macaque lived in Sundaland, which was created by sea-level lowering during the glacial period [17,36]. The most recent population bottleneck around 20,000 years ago may correspond to the last glacial maximum, when average temperatures were 2 to 6°C lower than the present temperatures. The change in population

Position	Reference allele	Alternative allele	Reference count	Alternative count	Zygosity	Transcript	AA change	Rhesus QV
167211141	T	C	0	30	Homozygous	CHIT1.t01:447:+:	I:T;	60

Figure 5 Screenshots of the macaque genome database. (a) cDNA clones, BAC clones, microsatellite markers, gene predictions, and SNVs on chromosome 1 are shown in the genome browser. **(b)** Detailed information on each SNV is linked from the browser.

size is possibly associated with admixture with the rhesus macaque, since their habitats were largely connected by the formation of Sundaland. However, it should be noted that the time estimation largely depends on the generation time parameter of macaques. If we adopt a longer generation time parameter - for example, 10 to 12 years

as the median age of females giving offspring - the most recent bottleneck event would shift earlier, 33,000 to 40,000 years ago.

Conclusions

We identified 9.7 million high-quality SNVs between the Malaysian cynomolgus and the reference (Indian rhesus) macaque genomes. The list of whole-genome SNVs will be useful for many future applications, such as an array-based genotyping system of macaque individuals. In contrast to humans, the genetic variation of experimental animals, especially of monkeys, is largely unexplored. The whole-genome sequence of a Malaysian cynomolgus macaque has unveiled hidden genetic variations among these widely used experimental animals and will benefit future evolutionary and biomedical studies.

Materials and methods

Animal and blood sampling

Whole blood cells for genomic DNA were obtained from a 25-year-old male cynomolgus macaque (Malaysian), housed at the Tsukuba Primate Research Center (TPRC), National Institute of Biomedical Innovation (NIBIO), Tsukuba, Ibaraki, Japan, in accordance with the TPRC guidelines. The sampled macaque was an F1 progeny of unrelated wild individuals captured in the south of Kuala Lumpur. These macaques were cared for and handled according to the guidelines established by the Institutional Animal Care and Use Committee of NIBIO and the standard operating procedures for macaques at the TPRC. Blood collection was conducted at the TPRC in accordance with the guidelines of the Laboratory Biosafety Manual, World Health Organization. Genomic DNA was isolated from 10 ml of peripheral blood with EDTA using a Qiagen Genomic DNA purification kit (Qiagen K. K., Tokyo, Japan). The isolated DNA samples were kept at -80°C until use.

Genome sequencing

Genome sequencing was performed using the SOLiD 3 Plus System (Life Technologies, Gaithersburg, MD, USA). Fragment (50 bp) and mate-pair (25 bp × 2) libraries were generated using the macaque genomic DNA. Mate-pair libraries of 600 to 800 bp and 800 to 1,000 bp insert sizes were prepared, and each library was run in two slides. Library preparations and all SOLiD runs were performed as per the standard manufacturer's protocols.

Mapping sequence data on the Indian rhesus macaque genome

SOLiD sequence data were mapped on the rhesus macaque draft genome sequence (GenBank accession numbers NC_007858 to NC_007878). The assembly QV of the genome was retrieved from the UCSC website [37]. The reads were mapped using the BioScope (Life Technologies) local alignment algorithm with parameters of 25 bp seed length, 2 mismatches in a seed, and mismatch penalty score -2.0 (default threshold). The algorithm finds genomic regions that match to the first 25 bp of each read, allowing at most 2 mismatches, and extending the regions until the score exceeds the threshold. 'PCR and optical duplicates' reads (defined by BioScope; mapped to more than 100 loci, duplicates) and mate-pair reads incongruently mapped on the reference genome (unpaired reads) were filtered out. All mapped sequence reads were deposited to public databases (DNA data bank of Japan (DDBJ) Sequence Read Archive: DRA000430). Chinese rhesus macaque and Vietnamese cynomolgus macaque genome sequences were downloaded from the public database (accession numbers SRA023855 and SRA023856) and aligned to the rhesus macaque genome sequence using the Bowtie 2 program [38] with a local alignment algorithm. A pre-aligned African genome sequence (NA19239) was retrieved from the 1000 Genomes project website [39]. In all resequenced genomes, SNVs were called using SAMtools with a default parameter setting, except for a mismatch tuning parameter (option -C) of 50.

Indel detection

The detection and calling of small and large indels were performed using the software implemented in BioScope software v1.3.1. Briefly, small indels were identified using sequence reads mapped with alignment gaps, and large indels were identified using incongruent distances between mate-pair reads. The small indel-finding algorithm could detect deletions shorter than 12 bp and insertions shorter than 4 bp. In both analyses, a default setting of parameters was applied.

Gene annotation

Entrez Gene annotations in the National Center for Biotechnology Information database were used for classifying SNVs into annotations [40]. Genes assigned to multiple genomic loci were excluded from the analysis. Among 27,424 annotated transcripts in the Indian rhesus macaque genome, 944 showed inconsistencies with the draft genome sequence and were removed from further analyses. When we counted the number of variants at a site with overlapping annotations, we assigned an order of priority as follows: coding exon > non-coding exon > intron > intergenic. For example, when a site was annotated as a coding exon of some transcripts and as an intron of the others, the site was classified as a coding exon. In total, 19,574 protein-coding genes, consisting of 26,480 transcripts, were analyzed. Orthologous genes between human and macaque were determined using the

annotations of the Ensembl database [41]. Only one-to-one orthologs were used for subsequent analyses.

Estimation of demographic parameters

We used PSMC (pairwise sequentially Markovian coalescent) software to infer the demographic history of the Malaysian cynomolgus macaque [31]. Briefly, the program estimates the distribution of coalescent time between two haploid genomes, deduced from the rate of heterozygous SNVs across the genome sequence, with ancestral recombination events inferred by the hidden Markov model. The following parameters were used: time interval = 6 + 29*2, generation time = 6, mutation rate per generation = 2.5×10^{-8}, and the number of iterations = 25. The 95% confidence intervals were estimated using 200 times bootstrap resampling of 5 Mb genome blocks.

Additional material

Additional file 1: Figures S1 to S5 and Tables S1 to S3. Figure S1: chromosomal distribution of fold coverage of quality controlled mapped reads (duplicates and unpaired mate-pair reads were filtered out) on the reference rhesus macaque genome are shown. All chromosomes exceed 37-fold. Figure S2: minimum coverage of quality controlled mapped reads (duplicates and unpaired mate-pair reads were filtered out) on the reference rhesus macaque genome is shown. Genomic regions with at least five-fold coverage were used in the SNV analysis. Figure S3: SNV density along each chromosome. The red and blue lines represent the number of heterozygous and homozygous SNVs in 1 Mb windows, respectively. The step size of window sliding was 100 kb. Figure S4: small indel discovery rate and rhesus macaque genome quality. The red and blue lines represent the rate of small deletions and insertions, respectively, with given rhesus macaque genome sequence quality values (QVs). Small indels at sites having QV < 45 in the rhesus macaque genome sequence were filtered out. Figure S5: distribution of large-indel lengths identified in the cynomolgus macaque genome. Indels were identified using the distance information from the mate-pair libraries. Indel regions containing ambiguous genome sequences were excluded. Table S1: summary of SOLiD libraries and sequence reads (mapped to the Vietnamese cynomolgus macaque genome sequence). Table S2: pattern of nucleotide changes. Table S3: immune- and drug-response genes with completely segregating nonsynonymous SNVs between cynomolgus and rhesus macaques.

Abbreviations

BAC: bacterial artificial chromosome; QV: quality value; SNV: single nucleotide variant; UTR: untranslated region.

Acknowledgements

This study was conducted through the Cooperative Research Program at the Tsukuba Primate Research Center, National Institute of Biomedical Innovation (supported by the Ministry of Health, Labour and Welfare, Japan). This work was partially supported by a Grant-in-Aid for Young Scientists (B) KAKENHI 22700460 and 24700428.

Author details

[1]Laboratory of Rare Disease Biospecimen, Department of Disease Bioresources Research, National Institute of Biomedical Innovation, 7-6-8 Saito-asagi, Ibaraki, Osaka 567-0085, Japan. [2]Center for Human Evolution Modeling Research, Primate Research Institute, Kyoto University, Inuyama, Aichi 484-8506, Japan. [3]Tsukuba Primate Research Center, National Institute of Biomedical Innovation, 1-1 Hachimandai, Tsukuba, Ibaraki 305-0843, Japan. [4]Division of Evolutionary Genetics, Department of Population Genetics, National Institute of Genetics, 1111 Yata, Mishima, Shizuoka 411-8540, Japan. [5]Department of Genetics, The Graduate University for Advanced Studies (SOKENDAI), 1111 Yata, Mishima, Shizuoka 411-8540, Japan.

Authors' contributions

AH, RS, TM, YY and NO contributed to the design of this research. AH, YK, IT, RT and NO performed the experiments. AH, RS, MH and NO contributed to data analysis. AH, RS and NO wrote the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 9 December 2011 Revised: 20 June 2012
Accepted: 2 July 2012 Published: 2 July 2012

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. Fooden J: **Provisional classifications and key to living species of macaques (primates: *Macaca*).** *Folia Primatol (Basel)* 1976, **25**:225-236.
3. Uno Y, Iwasaki K, Yamazaki H, Nelson DR: **Macaque cytochromes P450: nomenclature, transcript, gene, genomic structure, and function.** *Drug Metab Rev* 2011, **43**:346-361.
4. Ebeling M, Kung E, See A, Broger C, Steiner G, Berrera M, Heckel T, Iniguez L, Albert T, Schmucki R, Biller H, Singer T, Certa U: **Genome-based analysis of the nonhuman primate *Macaca fascicularis* as a model for drug safety assessment.** *Genome Res* 2011, **21**:1746-1756.
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. Fooden J: **Rhesus and crab-eating macaques: intergradation in Thailand.** *Science* 1964, **143**:363-364.
7. Street SL, Kyes RC, Grant R, Ferguson B: **Single nucleotide polymorphisms (SNPs) are highly conserved in rhesus (*Macaca mulatta*) and cynomolgus (*Macaca fascicularis*) macaques.** *BMC Genomics* 2007, **8**:480.
8. Osada N, Uno Y, Mineta K, Kameoka Y, Takahashi I, Terao K: **Ancient genome-wide admixture extends beyond the current hybrid zone between *Macaca fascicularis* and *M. mulatta*.** *Mol Ecol* 2010, **19**:2884-2895.
9. Osada N, Hashimoto K, Kameoka Y, Hirata M, Tanuma R, Uno Y, Inoue I, Hida M, Suzuki Y, Sugano S, Terao K, Kusuda J, Takahashi I: **Large-scale analysis of *Macaca fascicularis* transcripts and inference of genetic divergence between *M. fascicularis* and *M. mulatta*.** *BMC Genomics* 2008, **9**:90.
10. Yan G, Zhang G, Fang X, Zhang Y, Li C, Ling F, Cooper DN, Li Q, Li Y, van Gool AJ, Du H, Chen J, Chen R, Zhang P, Huang Z, Thompson JR, Meng Y, Bai Y, Wang J, Zhuo M, Wang T, Huang Y, Wei L, Li J, Wang Z, Hu H, Yang P, Le L, Stenson PD, Li B, *et al*: **Genome sequencing and comparison of two nonhuman primate animal models, the cynomolgus and Chinese rhesus macaques.** *Nat Biotechnol* 2011, **29**:1019-1023.
11. Fang X, Zhang Y, Zhang R, Yang L, Li M, Ye K, Guo X, Wang J, Su B: **Genome sequence and global sequence variation map with 5.5 million SNPs in Chinese rhesus macaque.** *Genome Biol* 2011, **12**:R63.
12. Fawcett GL, Raveendran M, Deiros DR, Chen D, Yu F, Harris RA, Ren Y, Muzny DM, Reid JG, Wheeler DA, Worley KC, Shelton SE, Kalin NH, Milosavljevic A, Gibbs R, Rogers J: **Characterization of single-nucleotide variation in Indian-origin rhesus macaques (*Macaca mulatta*).** *BMC Genomics* 2011, **12**:311.
13. Kanthaswamy S, Satkoski J, George D, Kou A, Erickson BJ, Smith DG: **Interspecies hybridization and the stratification of nuclear genetic variation of rhesus (*Macaca mulatta*) and long-tailed macaques (*Macaca fascicularis*).** *Int J Primatol* 2008, **29**:1295-1311.

14. 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.
15. Stevison LS, Kohn MH: **Determining genetic background in captive stocks of cynomolgus macaques (*Macaca fascicularis*).** *J Med Primatol* 2008, **37**:311-317.
16. Bonhomme M, Cuartero S, Blancher A, Crouau-Roy B: **Assessing natural introgression in 2 biomedical model species, the rhesus macaque (*Macaca mulatta*) and the long-tailed macaque (*Macaca fascicularis*).** *J Hered* 2009, **100**:158-169.
17. Delson E: **Fossil macaques, phyletic relationships and a scenario of deployment.** In *The Macaques: Studies in Ecology, Behavior, and Evolution.* Edited by: Lindburg DG. New York: Van Nostrand Reinhold Co; 1980:10-30.
18. Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G, Lin Y, MacDonald JR, Pang AW, Shago M, Stockwell TB, Tsiamouri A, Bafna V, Bansal V, Kravitz SA, Busam DA, Beeson KY, McIntosh TC, Remington KA, Abril JF, Gill J, Borman J, Rogers YH, Frazier ME, Scherer SW, Strausberg RL, *et al*: **The diploid genome sequence of an individual human.** *PLoS Biol* 2007, **5**:e254.
19. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Keira Cheetham R, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu X, Maisinger KS, Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, *et al*: **Accurate whole human genome sequencing using reversible terminator chemistry.** *Nature* 2008, **456**:53-59.
20. Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, Fan W, Zhang J, Li J, Zhang J, Guo Y, Feng B, Li H, Lu Y, Fang X, Liang H, Du Z, Li D, Zhao Y, Hu Y, Yang Z, Zheng H, Hellmann I, Inouye M, Pool J, Yi X, Zhao J, Duan J, Zhou Y, Qin J, *et al*: **The diploid genome sequence of an Asian individual.** *Nature* 2008, **456**:60-65.
21. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen Y-J, Makhijani V, Roth GT, Gomes X, Tartaro K, Niazi F, Turcotte CL, Irzyk GP, Lupski JR, Chinault C, Song X-z, Liu Y, Yuan Y, Nazareth L, Qin X, Muzny DM, Margulies M, Weinstock GM, Gibbs RA, Rothberg JM: **The complete genome of an individual by massively parallel DNA sequencing.** *Nature* 2008, **452**:872-876.
22. Ahn S-M, Kim T-H, Lee S, Kim D, Ghang H, Kim D-S, Kim B-C, Kim S-Y, Kim W-Y, Kim C, Park D, Lee YS, Kim S, Reja R, Jho S, Kim CG, Cha J-Y, Kim K-H, Lee B, Bhak J, Kim S-J: **The first Korean genome sequence and analysis: Full genome sequencing for a socio-ethnic group.** *Genome Res* 2009, **19**:1622-1629.
23. Kim J-I, Ju YS, Park H, Kim S, Lee S, Yi J-H, Mudge J, Miller NA, Hong D, Bell CJ, Kim H-S, Chung I-S, Lee W-C, Lee J-S, Seo S-H, Yun J-Y, Woo HN, Lee H, Suh D, Lee S, Kim H-J, Yavartanoo M, Kwak M, Zheng Y, Lee MK, Park H, Kim JY, Gokcumen O, Mills RE, Zaranek AW, *et al*: **A highly annotated whole-genome sequence of a Korean individual.** *Nature* 2009, **460**:1011-1015.
24. Fujimoto A, Nakagawa H, Hosono N, Nakano K, Abe T, Boroevich KA, Nagasaki M, Yamaguchi R, Shibuya T, Kubo M, Miyano S, Nakamura Y, Tsunoda T: **Whole-genome sequencing and comprehensive variant analysis of a Japanese individual using massively parallel sequencing.** *Nat Genet* 2010, **42**:931-936.
25. **QFbase.**. [http://genebank.nibio.go.jp/cgi-bin/gbrowse/rheMac2/].
26. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Subgroup GPDP: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25**:2078-2079.
27. McKernan KJ, Peckham HE, Costa GL, McLaughlin SF, Fu Y, Tsung EF, Clouser CR, Duncan C, Ichikawa JK, Lee CC, Zhang Z, Ranade SS, Dimalanta ET, Hyland FC, Sokolsky TD, Zhang L, Sheridan A, Fu H, Hendrickson CL, Li B, Kotler L, Stuart JR, Malek JA, Manning JM, Antipova AA, Perez DS, Moore MP, Hayashibara KC, Lyons MR, Beaudoin RE, *et al*: **Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding.** *Genome Res* 2009, **19**:1527-1541.
28. Nei M: *Molecular Evolutionary Genetics* Columbia University Press; 1987.
29. Ohta T: **The nearly neutral theory of molecular evolution.** *Annu Rev Ecol Systematics* 1992, **23**:263-286.
30. 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.
31. Li H, Durbin R: **Inference of human population history from individual whole-genome sequences.** *Nature* 2011, **475**:493-496.
32. 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.
33. Higashino A, Osada N, Suto Y, Hirata M, Kameoka Y, Takahashi I, Terao K: **Development of an integrative database with 499 novel microsatellite markers for *Macaca fascicularis*.** *BMC Genet* 2009, **10**:24.
34. Matsumoto J, Kawai S, Terao K, Kirinoki M, Yasutomi Y, Aikawa M, Matsuda H: **Malaria infection induces rapid elevation of the soluble Fas ligand level in serum and subsequent T lymphocytopenia: possible factors responsible for the differences in susceptibility of two species of *Macaca* monkeys to *Plasmodium coatneyi* infection.** *Infect Immun* 2000, **68**:1183-1188.
35. Hamada Y, Urasopon N, Hadi I, Malaivijitnond S: **Body size and proportions and pelage color of free-ranging *Macaca mulatta* from a zone of hybridization in Northeastern Thailand.** *Int J Primatol* 2006, **27**:497-513.
36. Heaney LR: **A synopsis of climatic and vegetational change in Southeast-Asia.** *Climatic Change* 1991, **19**:53-61.
37. **UCSC Genome Browser.**. [http://ucsc.genome.edu/].
38. **Bowtie 2.**. [http://bowtie-bio.sourceforge.net/bowtie2/index.shtml].
39. **1000 Genomes.**. [http://www.1000genomes.org/].
40. Maglott D, Ostell J, Pruitt KD, Tatusova T: **Entrez Gene: gene-centered information at NCBI.** *Nucleic Acids Res* 2011, **39**:D52-D57.
41. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Chen Y, Clapham P, Coates G, Fairley S, Fitzgerald S, Gordon L, Hendrix M, Hourlier T, Johnson N, Kähäri A, Keefe D, Keenan S, Kinsella R, Kokocinski F, Kulesha E, Larsson P, Longden I, McLaren W, Overduin B, Pritchard B, Riat HS, Rios D, Ritchie GRS, Ruffier M, Schuster M, *et al*: **Ensembl 2011.** *Nucleic Acids Res* 2011, **39**:D800-D806.

doi:10.1186/gb-2012-13-7-r58
Cite this article as: Higashino *et al.*: Whole-genome sequencing and analysis of the Malaysian cynomolgus macaque (*Macaca fascicularis*) genome. *Genome Biology* 2012 **13**:R58.