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

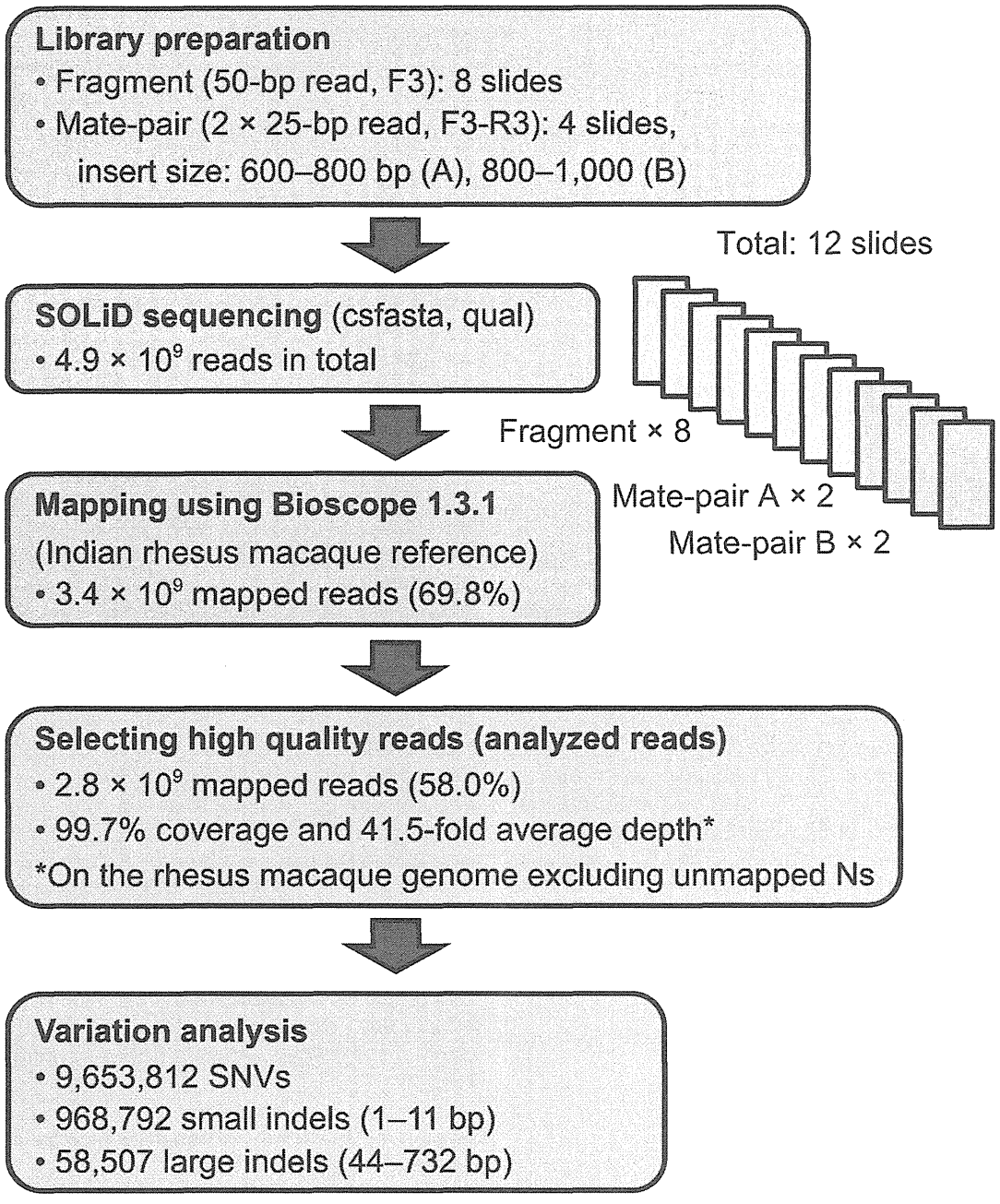


Figure 1 Overview of the Malaysian cynomolgus macaque genome sequencing and analysis. Fragment runs of eight slides and mate-pair runs of four slides (insert size: two slides for 600 to 800 bp and two slides for 800 to 1,000 bp) were performed on the SOLiD 3 Plus System. In total, 4.9 × 10^9 sequence 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.

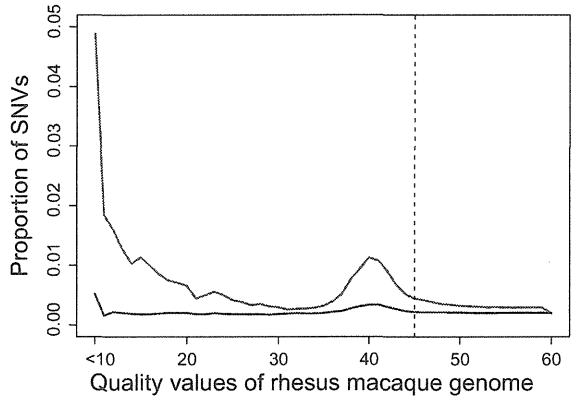


Figure 2 SNV discovery rate and rhesus macaque genome quality. The red and blue lines represent the rates of homozygous and heterozygous SNVs, respectively, with given rhesus macaque genome sequence quality values (QVs). SNVs at sites having QV < 45 (left of the dashed line) were filtered out.

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.

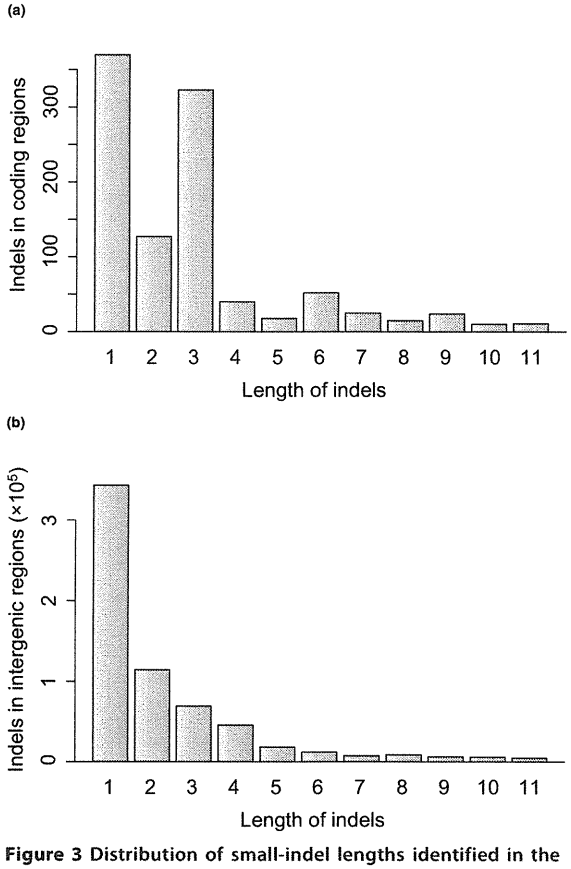


Figure 3 Distribution of small-indel lengths identified in the Malaysian cynomolgus macaque genome. (a) Indels in coding regions. **(b)** Indels in intergenic regions. The indels were identified using the information from high-quality reads.

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

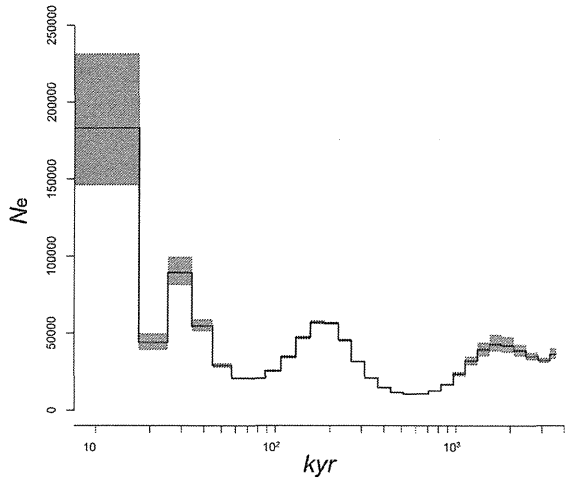


Figure 4 Estimation of ancestral population size of cynomolgus macaques with PSMC software. The *x* and *y* axes show population size and the time (thousand years (kyr)) from present, respectively. The blue rectangles represent 95% confidence interval with bootstrap resampling.

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.

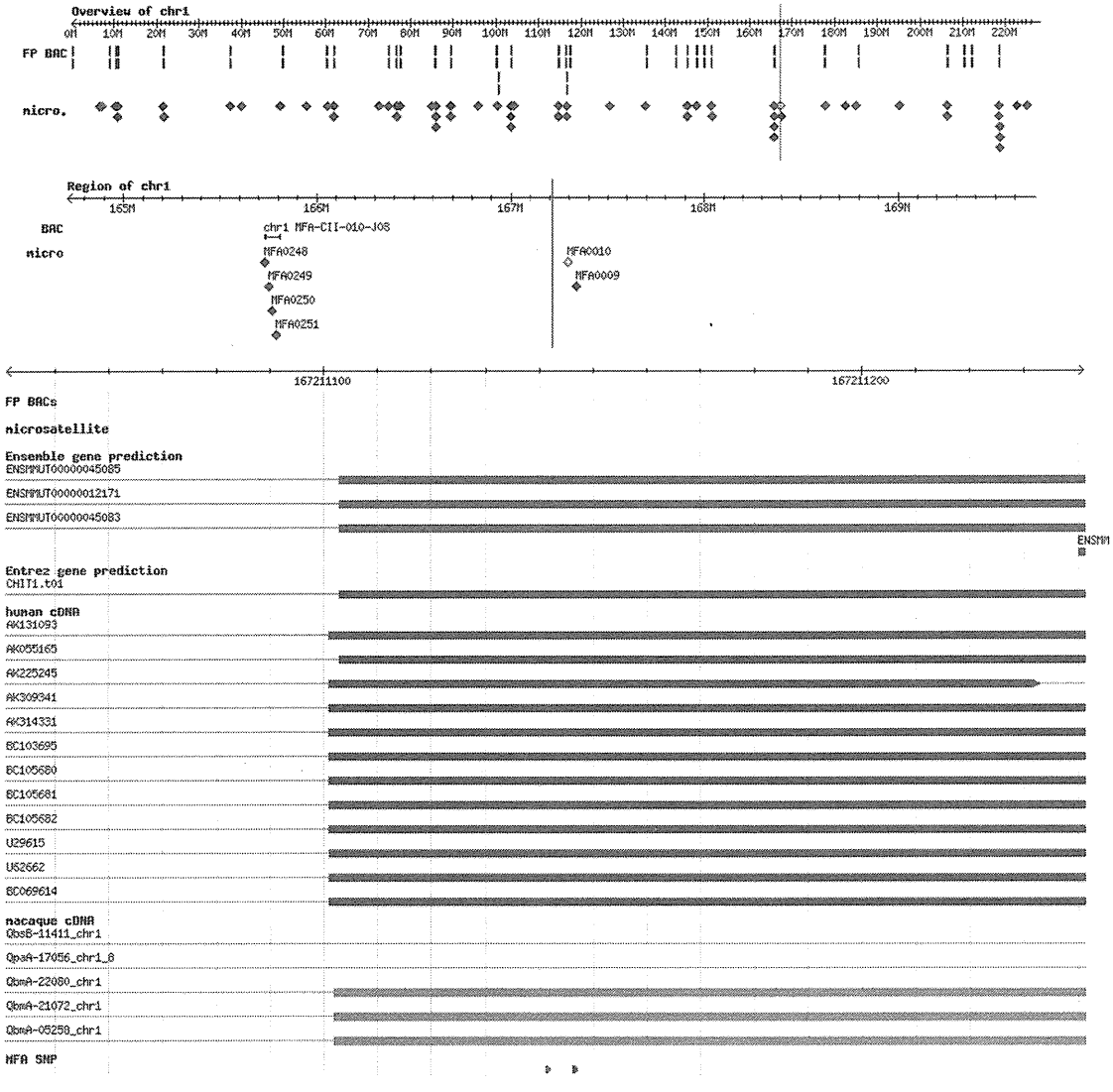
In this study, we have identified about 9.7 million SNVs between Malaysian cynomolgus and Indian rhesus macaques and 8.5 million SNVs between Malaysian and Vietnamese cynomolgus macaques. The total number of SNVs is much higher than that estimated in human genome resequencing studies (approximately 3 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

(a)



(b)

***Macaca fascicularis* MFA SNP**

chr1 167211141 [167211141 .. 167211141]

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.

nature genetics

Plasmodium cynomolgi genome sequences provide insight into *Plasmodium vivax* and the monkey malaria clade

Shin-Ichiro Tachibana[1,13], Steven A Sullivan[2], Satoru Kawai[3], Shota Nakamura[4], Hyunjae R Kim[2], Naohisa Goto[4], Nobuko Arisue[5], Nirianne M Q Palacpac[5], Hajime Honma[1,5], Masanori Yagi[5], Takahiro Tougan[5], Yuko Katakai[6], Osamu Kaneko[7], Toshihiro Mita[8], Kiyoshi Kita[9], Yasuhiro Yasutomi[10], Patrick L Sutton[2], Rimma Shakhbatyan[2], Toshihiro Horii[5], Teruo Yasunaga[4], John W Barnwell[11], Ananias A Escalante[12], Jane M Carlton[2,14] & Kazuyuki Tanabe[1,5,14]

***P. cynomolgi*, a malaria-causing parasite of Asian Old World monkeys, is the sister taxon of *P. vivax*, the most prevalent malaria-causing species in humans outside of Africa. Because *P. cynomolgi* shares many phenotypic, biological and genetic characteristics with *P. vivax*, we generated draft genome sequences for three *P. cynomolgi* strains and performed genomic analysis comparing them with the *P. vivax* genome, as well as with the genome of a third previously sequenced simian parasite, *Plasmodium knowlesi*. Here, we show that genomes of the monkey malaria clade can be characterized by copy-number variants (CNVs) in multigene families involved in evasion of the human immune system and invasion of host erythrocytes. We identify genome-wide SNPs, microsatellites and CNVs in the *P. cynomolgi* genome, providing a map of genetic variation that can be used to map parasite traits and study parasite populations. The sequencing of the *P. cynomolgi* genome is a critical step in developing a model system for *P. vivax* research and in counteracting the neglect of *P. vivax*.**

Human malaria is transmitted by anopheline mosquitoes and is caused by four species in the genus *Plasmodium*. Of these, *P. vivax* is the major malaria agent outside of Africa, annually causing 80–100 million cases[1]. Although *P. vivax* infection is often mistakenly regarded as benign and self-limiting, *P. vivax* treatment and control present challenges distinct from those of the more virulent *Plasmodium falciparum*. Biological traits, including a dormant (hypnozoite) liver stage responsible for recurrent infections (relapses), early infective sexual stages (gametocytes) and transmission from low parasite densities in the blood[2], coupled with emerging antimalarial drug resistance[3], render *P. vivax* resilient to modern control strategies. Recent evidence indicates that *P. falciparum* derives from parasites of great apes in Africa[4], whereas *P. vivax* is more closely related to parasites of Asian Old World monkeys[5–7], although not itself infective of these monkeys.

P. vivax cannot be cultured *in vitro*, and the small New World monkeys capable of hosting it are rare and do not provide an ideal model system. *P. knowlesi*, an Asian Old World monkey parasite recently recognized as a zoonosis for humans[8], has had its genome sequenced[9], but the species is distantly related to *P. vivax* and is phenotypically dissimilar. In contrast, *P. cynomolgi*, a simian parasite that can infect humans experimentally[10], is the closest living relative (a sister taxon) to *P. vivax* and possesses most of the same genetic, phenotypic and biological characteristics—notably, periodic relapses caused by dormant hypnozoites, early infectious gametocyte formation and invasion of Duffy blood group–positive reticulocytes. *P. cynomolgi* thus offers a robust model for *P. vivax* in a readily available laboratory host, the Rhesus monkey, whose genome was recently sequenced[11]. Here, we report draft genome sequences of three *P. cynomolgi* strains and comparative genomic analyses of *P. cynomolgi*, *P. vivax*[12] and *P. knowlesi*[9], three members of the monkey malaria clade.

We sequenced the genome of *P. cynomolgi* strain B, isolated from a monkey in Malaysia and grown in splenectomized monkeys (Online Methods). A combination of Sanger, Roche 454 and Illumina chemistries was employed to generate a high-quality reference assembly at 161-fold coverage, consisting of 14 supercontigs (corresponding to the 14 parasite chromosomes) and ~1,649 unassigned contigs, comprising

[1]Laboratory of Malariology, Research Institute for Microbial Diseases, Osaka University, Suita, Japan. [2]Department of Biology, Center for Genomics and Systems Biology, New York University, New York, New York, USA. [3]Laboratory of Tropical Medicine and Parasitology, Institute of International Education and Research, Dokkyo Medical University, Shimotsuga, Japan. [4]Genome Information Research Center, Research Institute for Microbial Diseases, Osaka University, Suita, Japan. [5]Department of Molecular Protozoology, Research Institute for Microbial Diseases, Osaka University, Suita, Japan. [6]The Corporation for Production and Research of Laboratory Primates, Tsukuba, Japan. [7]Department of Protozoology, Institute of Tropical Medicine (NEKKEN) and Global COE (Centers of Excellence) Program, Nagasaki University, Nagasaki, Japan. [8]Department of Molecular and Cellular Parasitology, Graduate School of Medicine, Juntendo University, Tokyo, Japan. [9]Department of Biomedical Chemistry, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan. [10]Tsukuba Primate Research Center, National Institute of Biomedical Innovation, Tsukuba, Japan. [11]Center for Global Health, Centers for Disease Control and Prevention, Divison of Parasitic Diseases and Malaria, Atlanta, Georgia, USA. [12]Center for Evolutionary Medicine and Informatics, The Biodesign Institute, Arizona State University, Tempe, Arizona, USA. [13]Present address: Career-Path Promotion Unit for Young Life Scientists, Kyoto University, Kyoto, Japan. [14]These authors jointly directed this work. Correspondence should be addressed to K.T. (kztanabe@biken.osaka-u.ac.jp) or J.M.C. (jane.carlton@nyu.edu).

Received 25 January; accepted 9 July; published online 5 August 2012; doi:10.1038/ng.2375

a total length of ~26.2 Mb (**Supplementary Table 1**). Comparing genomic features of *P. cynomolgi*, *P. knowlesi* and *P. vivax* reveals many similarities, including GC content (mean GC content of 40.5%), 14 positionally conserved centromeres and the presence of intrachromosomal telomeric sequences (ITSs; GGGTT(T/C)A), which were discovered in the *P. knowlesi* genome[9] but are absent in *P. vivax* (**Fig. 1**, **Table 1** and **Supplementary Table 2**).

We annotated the *P. cynomolgi* strain B genome using a combination of *ab initio* gene prediction programs trained on high-quality data sets and sequence similarity searches against the annotated *P. vivax* and *P. knowlesi* genomes. Not unexpectedly for species from the same monkey malaria clade, gene synteny along the 14 chromosomes is highly conserved, although numerous microsyntenic breaks are present in regions containing multigene families (**Fig. 2** and **Table 2**). This genome-wide view of synteny in six species of *Plasmodium* also identified two apparent errors in existing public sequence databases: an inversion in chromosome 3 of *P. knowlesi* and an inversion in chromosome 6 of *P. vivax*. The *P. cynomolgi* genome contains 5,722 genes, of which approximately half encode conserved hypothetical proteins of unknown function, as is the case in all the *Plasmodium* genomes sequenced to date. A maximum-likelihood phylogenetic tree constructed using 192 conserved ribosomal and translation- and transcription-related genes (**Supplementary Fig. 1**) confirms the close relationship of *P. cynomolgi* to *P. vivax* compared to five other *Plasmodium* species. Approximately 90% of genes (4,613) have reciprocal best-match orthologs in all three species (**Fig. 3**), enabling refinement of the existing *P. vivax* and *P. knowlesi* annotations (**Supplementary Table 3**). The high degree of gene orthology enabled us to identify specific examples of gene duplication (an important vehicle for genome evolution), including a duplicated homolog of *P. vivax Pvs28*—which encodes a sexual stage surface antigen that is a transmission-blocking vaccine candidate[13]—in *P. cynomolgi* (**Supplementary Table 4**). Genes common only to *P. cynomolgi* and *P. vivax* (n = 214) outnumber those that are restricted to *P. cynomolgi* and *P. knowlesi* (n = 100) or *P. vivax* and *P. knowlesi* (n = 17). Such figures establish the usefulness of *P. cynomolgi* as a model species for studying the more intractable *P. vivax*.

Notably, most of the genes specific to a particular species belong to multigene families (excluding hypothetical genes; **Table 2** and **Supplementary Table 5**). This suggests repeated lineage-specific gene duplication and/or gene deletion in multigene families within the three monkey malaria clade species. Moreover, copy numbers of the genes composing multigene families were generally greater in the *P. cynomolgi*–*P. vivax* lineage than in *P. knowlesi*, suggesting repeated gene duplication in the ancestral lineage of *P. cynomolgi* and *P. vivax* (or repeated gene deletion in the *P. knowlesi* lineage). Thus, the genomes of *P. cynomolgi*, *P. vivax* and *P. knowlesi* can largely be distinguished by variations in the copy number of multigene family members. Examples of such families include those that encode proteins involved in evasion of the human immune system (*vir*, *kir* and *SICAvar*) and invasion of host red blood cells (*dbp* and *rbp*).

In malaria-causing parasites, invasion of host erythrocytes, mediated by specific interactions between parasite ligands and erythrocyte receptors, is a crucial component of the parasite lifecycle. Of great interest are the *ebl* and *rbl* gene families, which encode parasite ligands required for the recognition of host erythrocytes. The *ebl* genes encode erythrocyte binding–like (EBL) ligands such as the Duffy-binding proteins (DBPs) that bind to Duffy antigen receptor for chemokines (DARC) on human and monkey erythrocytes. The *rbl* genes encode the reticulocyte binding–like (RBL) protein family, including reticulocyte-binding proteins (RBPs) in *P. cynomolgi* and *P. vivax*, and normocyte-binding proteins (NBPs) in *P. knowlesi*, which bind to unknown erythrocyte receptors[14]. We confirmed the presence of two *dbp* genes in *P. cynomolgi*[15] (**Supplementary Table 6**), in contrast to the one *dbp* and three *dbp* genes identified in *P. vivax* and *P. knowlesi*, respectively. This raises an intriguing hypothesis that *P. vivax* lost one *dbp* gene, and thus its infectivity of Old World monkey erythrocytes, after divergence from a common *P. vivax*–*P. cynomolgi* ancestor. This hypothesis is also supported by our identification of single-copy *dbp* genes in two other closely related Old World monkey malaria-causing parasites, *Plasmodium fieldi* and *Plasmodium simiovale*, which are incapable of infecting humans[16]. These two Old World monkey species lost one or more *dbp* genes during divergence that confer infectivity to humans, whereas *P. cynomolgi* and *P. knowlesi* retained *dbp* genes that allow invasion of human erythrocytes (**Supplementary Fig. 2**).

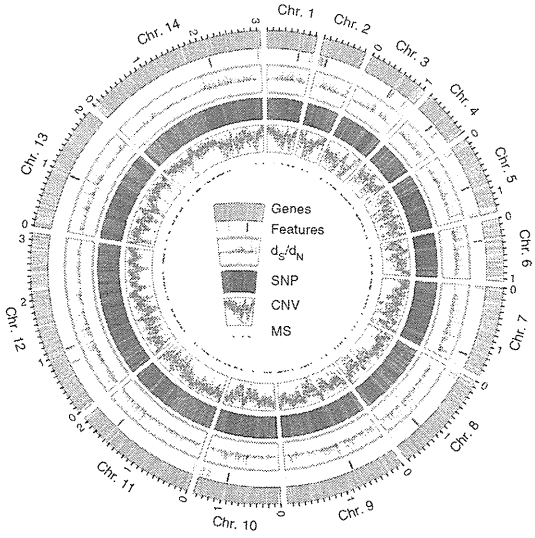


Figure 1 Architecture of the *P. cynomolgi* genome and associated genome-wide variation data. Data are shown for each of the 14 *P. cynomolgi* chromosomes. The six concentric rings, from outermost to innermost, represent (i) the location of 5,049 *P. cynomolgi* genes, excluding those on small contigs (cyan lines); (ii) genome features, including 14 centromeres (thick black lines), 43 telomeric sequence repeats (short red lines), 43 tRNA genes (red lines), 10 rRNAs (dark blue lines) and several gene family members, including 53 *cyir* (dark green lines), 8 *rbp* (brown lines), 13 *sera* (serine-rich antigen; pink lines), 25 *trag* (tryptophan-rich antigen; purple lines), 12 *msp3* (merozoite surface protein 3; light gray lines), 13 *msp7* (merozoite surface protein 7; gray lines), 25 *rad* (silver lines), 8 *etramp* (orange lines), 16 *Pf-fam-b* (light blue lines) and 7 *Pv-fam-d* (light green lines); (iii) plot of d_S/d_N for 4,605 orthologs depicting genome-wide polymorphism within *P. cynomolgi* strains B and Berok (black line) and divergence between *P. cynomolgi* strains B and Berok and *P. vivax* Salvador I (red line); a track above the plot indicates *P. cynomolgi* genes under positive selection (red) and purifying selection (blue), and a track below the plot indicates *P. cynomolgi*–*P. vivax* orthologs under positive selection (red) and purifying selection (blue); (iv) heatmap indicating SNP density of 3 *P. cynomolgi* strains plotted per 10-kb windows: red, 0–83 SNPs per 10 kb (regions of lowest SNP density); blue, 84–166 SNPs per 10 kb; green, 167–250 SNPs per 10 kb; purple, 251–333 SNPs per 10 kb; orange, 334–416 SNPs per10 kb; yellow, 417–500 SNPs per10 kb (regions of highest SNP density); (v) $\log_2$ ratio plot of CNVs identified from a comparison of *P. cynomolgi* strains B and Berok; and (vi) map of 182 polymorphic intergenic microsatellites (MS, black dots). The figure was generated using Circos software (see URLs).

Table 1 Comparison of genome features between *P. cynomolgi*, *P. vivax* and *P. knowlesi*, three species of the monkey malaria clade

Feature	*P. cynomolgi*	*P. vivax*[12]	*P. knowlesi*[9]
Assembly			
Size (Mb)	26.2	26.9	23.7
Number of scaffolds[a]	14 (1,649)	14 (2,547)	14 (67)
Coverage (fold)	161	10	8
GC content (%)	40.4	42.3	38.8
Genes			
Number of genes	5,722	5,432	5,197
Mean gene length (bp)	2,240	2,164	2,180
Gene density (bp per gene)[b]	4,428.2	4,950.5	4,416.1
Percentage coding[b]	51.0	47.1	49.0
Structural RNAs			
Number of tRNA genes	43	44	41
Number of 5S rRNA genes	3	3	0[c]
Number of 5.8S, 18S and 28S rRNA units	7	7	5
Nuclear genome			
Number of chromosomes	14	14	14
Number of centromeres	14	14	14
Isochore structure[d]	+	+	–
Mitochondrial genome			
Size (bp)[e]	5,986 (AB444123)	5,990 (AY598140)	5,958 (AB444108)
GC content (%)	30.3	30.5	30.5
Apicoplast genome			
Size (bp)	29,297[f]	5,064[g]	N/A
GC content (%)	13.0	17.1	N/A

N/A, not available.
[a]Small unassigned contigs indicated in parentheses. [b]Sequence gaps excluded. [c]Not present in *P. knowlesi* assembly version 4.0. [d]Regions of the genome that differ in density and are separable by CsCl centrifugation; isochores correspond to domains differing in GC content. [e]Identified in other studies (GenBank accessions given in parentheses). [f]Partial sequence (~86% complete) generated during this project. [g]Partial sequence of reference genome only published[12]; actual size is ~35 kb.

We found multiple *rbp* genes, some truncated or present as pseudogenes, in the *P. cynomolgi* genome (**Fig. 1** and **Table 2**). Phylogenetic analysis showed that *rbl* genes from *P. cynomolgi*, *P. vivax* and *P. knowlesi* can be classified into three distinct groups, RBP/NBP-1, RBP/NBP-2 and RBP/NBP-3 (**Supplementary Fig. 3**), and suggests that these groups existed before the three species diverged. All three groups of RBP/NBP are represented in *P. cynomolgi*, whereas *P. vivax* and *P. knowlesi* lack functional genes from the RBP/NBP-3 and RBP/NBP-1 groups, respectively. Thus, *rbl* gene family expansion seems to have occurred after speciation, indicating that the three species have multiple species-specific erythrocyte invasion mechanisms. Notably, we found an ortholog of *P. vivax rbp1b* in some strains of *P. cynomolgi* but not in others (**Supplementary Table 6**). To our knowledge, this is the first example of a CNV for a *rbp* gene between strains of a single *Plasmodium* species, highlighting how repeated creation and destruction of *rbl* genes, a signature of adaptive evolution, may have enabled species of the monkey malaria clade to expand or switch between monkey and human hosts.

The largest gene family in *P. cynomolgi*, consisting of 256 *cyir* (cynomolgi-interspersed repeat) genes, is part of the *pir* (plasmodium-interspersed repeat) superfamily that includes *P. vivax vir* genes ($n = 319$) and *P. knowlesi kir* genes ($n = 70$) (**Table 2**). *Pir*-encoded proteins reside on the surface of infected erythrocytes and have an important role in immune evasion[17]. Most *cyir* genes have sequence similarity to *P. vivax vir* genes ($n = 254$; **Supplementary Table 7**) and are found in subtelomeric regions (**Fig. 1**), but, notably, 11 *cyir* genes have sequence similarity to *P. knowlesi kir* genes (**Supplementary Table 7**) and occur more internally in the chromosomes, as do the *kir* genes in *P. knowlesi*. As with 'molecular mimicry' in *P. knowlesi* (mimicry of host sequences by pathogen sequences)[9], one CYIR protein (encoded by PCYB_032250) has a region of 56 amino acids that is highly similar to the extracellular domain of primate CD99 (**Supplementary Fig. 4**), a molecule involved in the regulation of T-cell function. A new finding is that *P. cynomolgi* has two genes whose sequences are similar to *P. knowlesi SICAvar* genes (**Supplementary Table 7**) that are expressed on the surfaces of schizont-infected macaque erythrocytes and are involved in antigenic variation[18].

The ability to form a dormant hypnozoite stage is common to both *P. cynomolgi* and *P. vivax* and was first shown in laboratory infections of monkeys by mosquito-transmitted *P. cynomolgi*[19]. In a search for candidate genes involved in the hypnozoite stage, we identified nine coding for 'dormancy-related' proteins that had the upstream ApiAP2 motifs[20] necessary for stage-specific transcriptional regulation at the sporozoite (pre-hypnozoite) stage (**Supplementary Table 8**). The candidates include kinases that are involved in cell cycle transition; hypnozoite formation may be regulated by phosphorylation of proteins specifically expressed at the pre-hypnozoite stage. Our list of *P. cynomolgi* candidate genes represents an informed starting point for experimental studies of this elusive stage.

We sequenced *P. cynomolgi* strains Berok (from Malaysia) and Cambodian (from Cambodia) to 26× and 17× coverage, respectively, to characterize *P. cynomolgi* genome-wide diversity through analysis of SNPs, CNVs and microsatellites. A comparison of the three *P. cynomolgi* strains identified 178,732 SNPs (**Supplementary Table 9**) at a frequency of 1 SNP per 151 bp, a polymorphism level somewhat

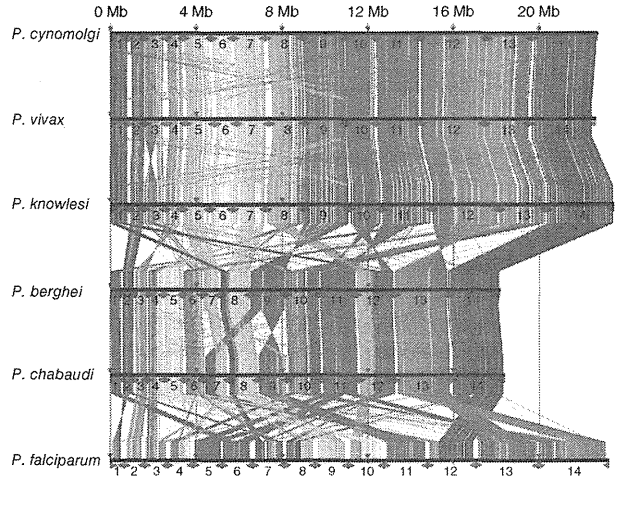


Figure 2 Genome synteny between six species of *Plasmodium* parasite. Protein-coding genes of *P. cynomolgi* are shown aligned with those of five other *Plasmodium* genomes: two species belonging to the monkey malaria clade, *P. vivax* and *P. knowlesi*; two species of rodent malaria, *P. berghei* and *P. chabaudi*; and *P. falciparum*. Highly conserved protein-coding regions between the genomes are colored in order from red (5′ end of chromosome 1) to blue (3′ end of chromosome 14) with respect to genomic position of *P. cynomolgi*.

Table 2 Components of multigene families of *P. cynomolgi*, *P. vivax* and *P. knowlesi* differ in copy number

Family	Multigene family	Localization	Arrangement	*P. cynomolgi*	*P. vivax*	*P. knowlesi*	Putative function and other information
1	*pir* (*vir*-like)	Subtelomeric	Scattered and clustered	254	319[a]	4	Immune evasion
2	*pir* (*kir*-like)	Subtelomeric and central	Scattered and clustered	11	2	66[a]	Immune evasion
3	*SICAvar*	Subtelomeric and central	Scattered and clustered	2	1	242[a]	Antigenic variation, immune evasion
4	*msp3*	Central	Clustered	12	12	3	Merozoite surface protein
5	*msp7*	Central	Clustered	13	13	5	Merozoite surface protein
6	*dbl* (*dbp/ebl*)	Subtelomeric	Scattered	2	1	3	Host cell recognition
7	*rbl* (*rbp/nbp/rh*)	Subtelomeric	Scattered	8[a]	10[a]	3[a]	Host cell recognition
8	*Pv-fam-a* (*trag*)	Subtelomeric	Scattered and clustered	36	36	26[a]	Tryptophan-rich
9	*Pv-fam-b*	Central	Clustered	3	6	1	Unknown
10	*Pv-fam-c*	Subtelomeric	Unknown[b]	1	7	0	Unknown
11	*Pv-fam-d* (*hypb*)	Subtelomeric	Scattered	18	16	2	Unknown
12	*Pv-fam-e* (*rad*)	Subtelomeric	Clustered	27	44	16	Unknown
13	*Pv-fam-g*	Central	Clustered	3	3	3	Unknown
14	*Pv-fam-h* (*hyp16*)	Central	Clustered	6	4	2	Unknown
15	*Pv-fam-i* (*hyp11*)	Subtelomeric	Scattered	6	6	5	Unknown
16	*Pk-fam-a*	Central	Scattered	0	0	12[a]	Unknown
17	*Pk-fam-b*	Subtelomeric	Scattered	0	0	9	Unknown
18	*Pk-fam-c*	Subtelomeric	Scattered	0	0	6[a]	Unknown
19	*Pk-fam-d*	Central	Scattered	0	0	3[a]	Unknown
20	*Pk-fam-e*	Subtelomeric	Scattered	0	0	3[a]	Unknown
21	*PST-A*	Subtelomeric and central	Scattered	9[a]	11[a]	7	αβ hydrolase
22	*ETRAMP*	Subtelomeric	Scattered	9	9	9	Parasitophorous vacuole membrane
23	*CLAG* (*RhopH-1*)	Subtelomeric	Scattered	2	3	2	High-molecular-wieght rhoptry antigen complex
24	*PvSTP1*	Subtelomeric	Unknown[b]	3	10[a]	0	Unknown
25	*PHIST* (Pf-*fam-b*)	Subtelomeric	Scattered and clustered	21	20	15	Unknown
26	*SERA*	Central	Clustered	13[a]	13[a]	8[a]	Cysteine protease

[a]Pseudogenes, truncated genes and gene fragments included. [b]Gene arrangement could not be determined due to localization on unassigned contigs.

similar to that found when *P. falciparum* genomes are compared[21,22]. We calculated the pairwise nucleotide diversity (π) as 5.41×10^{-3} across the genome, which varies little between the chromosomes. We assessed genome-wide CNVs between the *P. cynomolgi* B and Berok strains, using a robust statistical model in the CNV-seq program[23], by which we identified 1,570 CNVs (1 per 17 kb), including 1 containing the *rbp1b* gene on chromosome 7 (**Supplementary Fig. 5**). Finally, mining of the *P. cynomolgi* B and Berok strains identified 182 polymorphic intergenic microsatellites (**Supplementary Table 10**), the first set of genetic markers developed for this species. These provide a toolkit for studies of genetic diversity and population structure of laboratory stocks or natural infections of *P. cynomolgi*, many of which have recently been isolated from screening hundreds of wild monkeys for the zoonosis *P. knowlesi*[24].

We estimated the difference between the number of synonymous changes per synonymous site (d_S) and the number of nonsynonymous changes per nonsynonymous site (d_N) over 4,563 pairs of orthologs within *P. cynomolgi* strains B and Berok and 4,601 pairs of orthologs between these two *P. cynomolgi* strains and *P. vivax* Salvador I, using a simple Nei-Gojobori model[25]. We found 63 genes with $d_N > d_S$ within the two *P. cynomolgi* strains and 3,265 genes with $d_S > d_N$ (**Supplementary Table 11**). Genes with relatively high d_N/d_S ratios include those encoding transmembrane proteins, such as antigens and transporters, among which is a transmission-blocking target antigen, Pcyn230 (encoded by PCYB_042090). Notably, the *P. vivax* ortholog (PVX_003905) does not show evidence for positive selection[26], suggesting species-specific positive selection. We explored the degree to which evolution of orthologs has been constrained between *P. cynomolgi* and *P. vivax* and found 83 genes under possible accelerated evolution but 3,739 genes under possible purifying selection (**Supplementary Table 12**). This conservative estimate indicates that at least 81% of loci have diverged under strong constraint, compared with 1.8% of loci under less constraint or positive selection (**Fig. 1**), indicating that, overall, the genome of *P. cynomolgi* is highly conserved in single-locus genes compared to *P. vivax* and emphasizing the value of *P. cynomolgi* as a biomedical and evolutionary model for studying *P. vivax*.

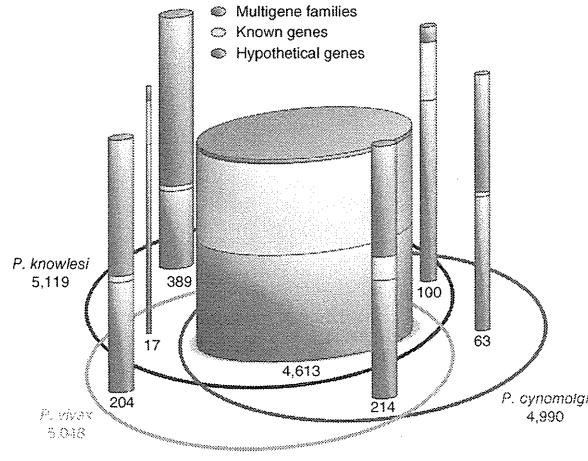


Figure 3 Comparison of the genes of *P. cynomolgi*, *P. vivax* and *P. knowlesi*. The Venn ellipses represent the three genomes, with the total number of genes assigned to the chromosomes indicated under the species name. Cylinders depict orthologous and non-orthologous genes between the three genomes, with the number of genes in each indicated and represented graphically by cylinder relative width. In each cylinder, genes are divided into three categories whose thickness is represented by colored bands proportional to category percentage.

Our generation of the first *P. cynomolgi* genome sequences is a critical step in the development of a robust model system for the intractable and neglected *P. vivax* species[27]. Comparative genome analysis of *P. vivax* and the Old World monkey malaria-causing parasites *P. cynomolgi* and *P. knowlesi* presented here provides the foundation for further insights into traits such as host specificity that will enhance prospects for the eventual elimination of vivax-caused malaria and global malaria eradication.

URLs. PlasmoDB, http://plasmodb.org/; Circos, http://circos.ca/; MIcroSAtelite Identification tool (MISA), http://pgrc.ipk-gatersleben.de/misa/; dbSNP, http://www.ncbi.nlm.nih.gov/projects/SNP/snp_viewBatch.cgi?sbid=1056645.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Sequence data for the *P. cynomolgi* B, Cambodian and Berok strains have been deposited in the DNA Data Bank of Japan (DDBJ), the European Molecular Biology Laboratory (EMBL) and the GenBank databases under the following accessions: B strain sequence reads DRA000196, genome assembly BAEJ01000001–BAEJ01003341 and annotation DF157093–DF158755; Cambodian strain sequence reads DRA000197; and Berok strain sequence reads SRA047950. SNP calls have been submitted to dbSNP (NYU_CGSB_BIO; 1056645) and may also be downloaded from the dbSNP website (see URLs). Sequences of the *dbp* genes from *P. cynomolgi* (Cambodian strain), *P. fieldi* (A.b.i. strain) and *P. simiovale* (AB617788–AB617791) and the *P. cynomolgi* Berok strain (JQ422035–JQ422036) and *rbp* gene sequences from the *P. cynomolgi* Berok and Cambodian strains (JQ422037–JQ422050) have been deposited. A partial apicoplast genome of the *P. cynomolgi* Berok strain has been deposited (JQ522954). The *P. cynomolgi* B reference genome is also available through PlasmoDB (see URLs).

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

We thank H. Sawai for suggestions on genome analysis, D. Fisher for help with genome-wide evolutionary analyses and the NYU Langone Medical Center Genome Technology Core for access to Roche 454 sequencing equipment (funded by grant S10 RR026950 to J.M.C. from the US National Institutes of Health (NIH)). Genome and phylogenetic analyses used the Genome Information Research Center in the Research Institute of Microbial Diseases at Osaka University. This work was supported by grants from the Ministry of Education, Culture, Sports, Science and Technology of Japan (18073013, 18GS03140013, 20390120 and 22406012) to K.T., an NIH grant (R01 GM080586) to A.A.E. and a Burroughs Wellcome Fund grant (1007398) and an NIH International Centers of Excellence for Malaria Research grant (U19 AI089676-01) to J.M.C. The content is soley the responsibility of the authors and does not necessarily represent the official views of the NIH.

AUTHOR CONTRIBUTIONS

K.T., J.M.C., A.A.E. and J.W.B. conceived and conducted the study. S.K., Y.K., Y.Y., S.-I.T. and J.W.B. provided *P. cynomolgi* material. S.N., N.G., T.Y. and H.R.K. constructed a computing system for data processing, and S.-I.T., H.H., P.L.S., S.A.S. and H.R.K. performed scaffolding of contigs and manual annotation of the predicted genes. S.N. performed sequence correction of supercontigs and gene prediction. S.-I.T., S.N., N.G., N.A., M.Y., O.K., K.T., H.R.K., R.S., S.A.S. and J.M.C. analyzed data. S.-I.T., N.M.Q.P., T.T., T.M., K.K., J.M.C., T.H., A.A.E., J.W.B. and K.T. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/doifinder/10.1038/ng.2375.

1. Mendis, K., Sina, B.J., Marchesini, P. & Carter, R. The neglected burden of *Plasmodium vivax* malaria. *Am. J. Trop. Med. Hyg.* **64**, 97–106 (2001).
2. Mueller, I. *et al.* Key gaps in the knowledge of *Plasmodium vivax*, a neglected human malaria parasite. *Lancet Infect. Dis.* **9**, 555–566 (2009).
3. Baird, J.K. Resistance to chloroquine unhinges vivax malaria therapeutics. *Antimicrob. Agents Chemother.* **55**, 1827–1830 (2011).
4. Rayner, J.C., Liu, W., Peeters, M., Sharp, P.M. & Hahn, B.H. A plethora of *Plasmodium* species in wild apes: a source of human infection? *Trends Parasitol.* **27**, 222–229 (2011).
5. Cornejo, O.E. & Escalante, A.A. The origin and age of *Plasmodium vivax*. *Trends Parasitol.* **22**, 558–563 (2006).
6. Escalante, A.A. *et al.* A monkey's tale: the origin of *Plasmodium vivax* as a human malaria parasite. *Proc. Natl. Acad. Sci. USA* **102**, 1980–1985 (2005).
7. Mu, J. *et al.* Host switch leads to emergence of *Plasmodium vivax* malaria in humans. *Mol. Biol. Evol.* **22**, 1686–1693 (2005).
8. Singh, B. *et al.* A large focus of naturally acquired *Plasmodium knowlesi* infections in human beings. *Lancet* **363**, 1017–1024 (2004).
9. Pain, A. *et al.* The genome of the simian and human malaria parasite *Plasmodium knowlesi*. *Nature* **455**, 799–803 (2008).
10. Eyles, D.E., Coatney, G.R. & Getz, M.E. Vivax-type malaria parasite of macaques transmissible to man. *Science* **131**, 1812–1813 (1960).
11. Gibbs, R.A. *et al.* Evolutionary and biomedical insights from the rhesus macaque genome. *Science* **316**, 222–234 (2007).
12. Carlton, J.M. *et al.* Comparative genomics of the neglected human malaria parasite *Plasmodium vivax*. *Nature* **455**, 757–763 (2008).
13. Saxena, A.K., Wu, Y. & Garboczi, D.N. *Plasmodium* p25 and p28 surface proteins: potential transmission-blocking vaccines. *Eukaryot. Cell* **6**, 1260–1265 (2007).
14. Iyer, J., Gruner, A.C., Renia, L., Snounou, G. & Preiser, P.R. Invasion of host cells by malaria parasites: a tale of two protein families. *Mol. Microbiol.* **65**, 231–249 (2007).
15. Okenu, D.M., Malhotra, P., Lalitha, P.V., Chitnis, C.E. & Chauhan, V.S. Cloning and sequence analysis of a gene encoding an erythrocyte binding protein from *Plasmodium cynomolgi*. *Mol. Biochem. Parasitol.* **89**, 301–306 (1997).
16. Coatney, G.R., Collins, W.E., Warren, M. & Contacos, P.G. *The Primate Malarias* (US Department of Health, Education and Welfare, Washington, DC, 1971).
17. Cunningham, D., Lawton, J., Jarra, W., Preiser, P. & Langhorne, J. The *pir* multigene family of *Plasmodium*: antigenic variation and beyond. *Mol. Biochem. Parasitol.* **170**, 65–73 (2010).
18. al-Khedery, B., Barnwell, J.W. & Galinski, M.R. Antigenic variation in malaria: a 3′ genomic alteration associated with the expression of a *P. knowlesi* variant antigen. *Mol. Cell* **3**, 131–141 (1999).
19. Krotoski, W.A. The hypnozoite and malarial relapse. *Prog. Clin. Parasitol.* **1**, 1–19 (1989).
20. Campbell, T.L., De Silva, E.K., Olszewski, K.L., Elemento, O. & Llinas, M. Identification and genome-wide prediction of DNA binding specificities for the ApiAP2 family of regulators from the malaria parasite. *PLoS Pathog.* **6**, e1001165 (2010).
21. Mu, J. *et al.* Genome-wide variation and identification of vaccine targets in the *Plasmodium falciparum* genome. *Nat. Genet.* **39**, 126–130 (2007).
22. Volkman, S.K. *et al.* A genome-wide map of diversity in *Plasmodium falciparum*. *Nat. Genet.* **39**, 113–119 (2007).
23. Xie, C. & Tammi, M.T. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics* **10**, 80 (2009).
24. Lee, K.S. *et al.* *Plasmodium knowlesi*: reservoir hosts and tracking the emergence in humans and macaques. *PLoS Pathog.* **7**, e1002015 (2011).
25. Nei, M. & Gojobori, T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3**, 418–426 (1986).
26. Doi, M. *et al.* Worldwide sequence conservation of transmission-blocking vaccine candidate Pvs230 in *Plasmodium vivax*. *Vaccine* **29**, 4308–4315 (2011).
27. Carlton, J.M., Sina, B.J. & Adams, J.H. Why is *Plasmodium vivax* a neglected tropical disease? *PLoS Negl. Trop. Dis.* **5**, e1160 (2011).

ONLINE METHODS

Parasite material. Details of the origin of the *P. cynomolgi* B, Berok and Cambodian strains, their growth in macaques and isolation of parasite material are given in the **Supplementary Note.**

Genome sequencing and assembly. *P. cynomolgi* B strain was sequenced using the Roche 454 GS FLX (Titanium) and Illumina/Solexa Genome Analyzer IIx platforms to 161× coverage. In addition, 2,784 clones (6.8 Mb) of a ~40-kb insert fosmid library in pCC1FOS (EpiCentre Biotechnologies) was sequenced by the Sanger method. A draft assembly of strain B was constructed using a combination of automated assembly and manual gap closure. We first generated *de novo* contigs by assembling Roche 454 reads using GS *De novo* Assembler version 2.0 with default parameters. Contigs of >500 bp were mapped to the *P. vivax* Salvador I reference assembly[12] (PlasmoDB; see URLs). *P. cynomolgi* contigs were iteratively arrayed through alignment to *P. vivax*–assembled sequences with manual corrections. A total of 1,264 aligned contigs were validated by mapping paired-end reads from fosmid clones using blastn ($e < 1 \times 10^{-15}$; identity > 90%; coverage > 200 bp) implemented in GenomeMatcher software version 1.65 (ref. 28). Additional linkages (699 regions) were made using PCR across the intervening sequence gaps with primers designed from neighboring contigs. The length of sequence gaps was estimated from insert lengths of the fosmid paired-end reads, the size of PCR products and homologous sequences of the *P. vivax* genome. Supercontigs were then manually constructed from the aligned contigs. Eventually, we obtained 14 supercontigs corresponding to the 14 chromosomes of the parasite, with a total length of ~22.73 Mb, encompassing ~80% of the predicted *P. cynomolgi* genome. A total of 1,651 contigs (>1 kb) with a total length of 3.45 Mb was identified as unassigned subtelomeric sequences by searching against the *P. vivax* genome using blastn. Additionally, to improve sequence accuracy, we constructed a mapping assembly of Illumina paired-end reads and the 14 supercontigs and unassigned contigs as reference sequences using CLC Genomics Workbench version 3.0 with default settings (CLC Bio). Comparison of the draft *P. cynomolgi* B sequence with 23 *P. cynomolgi* protein-coding genes (64 kb) obtained by Sanger sequencing showed 99.8% sequence identity (**Supplementary Table 13**). The *P. cynomolgi* Berok and Cambodian strains were sequenced to 26× and 17× coverage, respectively, using the Roche 454 GS FLX platform, with single-end and 3-kb paired-end libraries made for the former and a single-end library only made for the latter. For phylogenetic analyses of specific genes, sequences were independently verified by Sanger sequencing (**Supplementary Table 14** and **Supplementary Note**).

Prediction and annotation of genes. Gene prediction for the 14 supercontigs and 1,651 unassigned contigs was performed using the MAKER genome annotation pipeline[29] with *ab initio* gene prediction programs trained on proteins and ESTs from PlasmoDB Build 7.1. For gene annotation, blastn ($e < 1 \times 10^{-15}$; identity > 70%; coverage > 100 bp) searches of *P. vivax* (PvivaxAnnotated Transcripts_PlasmoDB-7.1.fasta) and *P. knowlesi* (PknowlesiAnnotatedTranscripts_PlasmoDB-7.1.fasta) predicted proteomes were run, and the best hits were identified. All predicted genes were manually inspected at least twice for gene structure and functional annotation, and orthologous relationships between *P. cynomolgi*, *P. vivax* and *P. knowlesi* were determined on synteny. A unique identifier, PCYB_######, was assigned to *P. cynomolgi* genes, where the first two of the six numbers indicate chromosome number. Paralogs of genes that seemed to be specific to either *P. cynomolgi*, *P. vivax* or *P. knowlesi* were searched using blastp with default parameters, using a cutoff *e* value of 1×10^{-16}.

Multiple genome sequence alignment. Predicted proteins of *P. cynomolgi* B strain were concatenated and aligned with those from the 14 chromosomes of 5 other *Plasmodium* genomes: *P. vivax*, *P. knowlesi*, *P. falciparum*, *P. berghei* and *P. chabaudi*, using Murasaki software version 1.68.6 (ref. 30).

Search for sequence showing high similarity to host proteins. Eleven *P. cynomolgi* CYIR proteins (with sequence similarity to *P. knowlesi* KIR) were subjected to blastp search for regions having high similarity to host *Macacca mulatta* CD99 protein, with cutoff *e* value of 1×10^{-12} and compositional adjustment (no adjustment) against the nonredundant protein sequence data set of the *M. mulatta* proteome in NCBI.

Phylogenetic analyses. Genes were aligned using ClustalW version 2.0.10 (ref. 31) with manual corrections, and unambiguously aligned sites were selected for phylogenetic analyses. Maximum-likelihood phylogenetic trees were constructed using PROML programs in PHYLIP version 3.69 (ref. 32) under the Jones-Taylor-Thornton (JTT) amino-acid substitution model. To take the evolutionary rate heterogeneity across sites into consideration, the R (hidden Markov model rates) option was set for discrete γ distribution, with eight categories for approximating the site-rate distribution. CODEML programs in PAML 4.4 (ref. 33) were used for estimating the γ shape parameter, α values. For bootstrap analyses, SEQBOOT and CONSENSE programs in PHYLIP were applied.

Candidate genes for hypnozoite formation. We undertook two approaches. First, genes unique to *P. vivax* and *P. cynomolgi* (hypnozoite-forming parasites) and not found in other non-hypnozoite–forming *Plasmodium* species were identified. We used the 147 unique genes identified in the *P. vivax* genome[12] to search the *P. cynomolgi* B sequence. For the orthologs identified in both species, ~1 kb of sequence 5′ to the coding sequence was searched for four specific ApiAP2 motifs[20]—PF14_0633, GCATGC; PF13_0235_D1, GCCCCG; PFF0670w_D1, TAAGCC; and PFD0985w_D2, TGTTAC—which are involved in sporozoite stage–specific regulation and expression (corresponding to the pre-hypnozoite stage). Second, dormancy-related proteins were retrieved from GenBank and used to search for *P. vivax* homologs. Candidate genes ($n = 128$) and orthologs of *P. cynomolgi* and five other parasite species were searched in the region ~1 kb upstream of the coding sequence for the presence of the four ApiAP2 motifs. Data for *P. vivax*, *P. knowlesi*, *P. falciparum*, *P. berghei*, *Plasmodium chabaudi* and *Plasmodium yoelii* were retrieved from PlasmoDB Build 7.1.

Genome-wide screen for polymorphisms. For SNP identification, alignment of Roche 454 data from strains B, Berok and Cambodian was performed using SSAHA2 (ref. 34), with 0.1 mismatch rate and only unique matches reported. Potential duplicate reads generated during PCR amplification were removed, so that when multiple reads mapped at identical coordinates, only the reads with the highest mapping quality were retained. We used a statistical method[35] implemented in SAMtools version 0.1.18 to call SNPs simultaneously in the case of duplicate runs of the same strain. SNPs with high read depth (>100) were filtered out, as were SNPs in poor alignment regions at the ends of chromosomes (**Supplementary Note**).

Nucleotide diversity (π) was calculated as follows. For each site being compared, we calculated allele frequency by counting the two alleles and measured the proportion of nucleotide differences. Letting π be the genetic distance between allele *i* and allele *j*, then the nucleotide diversity within the population is

$$\pi = \sum_{i,j} P_i P_j \pi_{ij}$$

where P_i and P_j are the overall allele frequencies of *i* and *j*, respectively. Mean π was calculated by averaging over sites, weighting each by $\sum_{i=1}^{n-1} \frac{1}{i}$, where *n* is the number of aligned sites. Average d_N/d_S ratios were estimated using the modified Nei-Gojobori/Jukes-Cantor method in MEGA 4 (ref. 36).

CNV-seq[23] was used to identify potential CNVs in *P. cynomolgi*. Briefly, this method is based on a statistical model that allows confidence assessment of observed copy-number ratios from next-generation sequencing data. Roche 454 sequences from *P. cynomolgi* strain B assembly were used as the reference genome, and the *P. cynomolgi* Berok strain was used as a test genome; the sequence coverage of the Cambodian strain was considered too low for analysis. The test reads were mapped to the reference genome, and CNVs were detected by computing the number of reads for each test strain in a sliding window. The validity of the observed ratios was assessed by the computation of a probability of a random occurrence, given no copy-number variation.

Polymorphic microsatellites (defined as repeat units of 1–6 nucleotides) between *P. cynomolgi* strains B and Berok were identified by aligning contigs