Whole-Genome Sequencing of Six Mauritian Cynomolgus Macaques (*Macaca fascicularis*) Reveals a Genome-Wide Pattern of Polymorphisms under Extreme Population Bottleneck

Naoki Osada[1,2], Nilmini Hettiarachchi[2,3], Isaac Adeyemi Babarinde[2,3], Naruya Saitou[2,3], and Antoine Blancher[4,*]

[1]Division of Evolutionary Genetics, National Institute of Genetics, Mishima, Japan

[2]Department of Genetics, SOKENDAI (The Graduate University for Advanced Studies), Mishima, Japan

[3]Division of Population Genetics, National Institute of Genetics, Mishima, Japan

[4]Laboratoire d'Immunogénétique Moléculaire (LIMT, EA3034), Faculté de Médecine Purpan, Université Paul Sabatier, Toulouse III, France

*Corresponding author: E-mail: blancher.a@chu-toulouse.fr.

Accepted: February 16, 2015

Data deposition: All raw read sequences and initial set of variants have been deposited at the public database under the EMBL-EBI accession PRJEB7871.

Abstract

Cynomolgus macaques (*Macaca fascicularis*) were introduced to the island of Mauritius by humans around the 16th century. The unique demographic history of the Mauritian cynomolgus macaques provides the opportunity to not only examine the genetic background of well-established nonhuman primates for biomedical research but also understand the effect of an extreme population bottleneck on the pattern of polymorphisms in genomes. We sequenced the whole genomes of six Mauritian cynomolgus macaques and obtained an average of 20-fold coverage of the genome sequences for each individual. The overall level of nucleotide diversity was 23% smaller than that of the Malaysian cynomolgus macaques, and a reduction of low-frequency polymorphisms was observed. In addition, we also confirmed that the Mauritian cynomolgus macaques were genetically closer to a representative of the Malaysian population than to a representative of the Indochinese population. Excess of nonsynonymous polymorphisms in low frequency, which has been observed in many other species, was not very strong in the Mauritian samples, and the proportion of heterozygous nonsynonymous polymorphisms relative to synonymous polymorphisms is higher within individuals in Mauritian than Malaysian cynomolgus macaques. Those patterns indicate that the extreme population bottleneck made purifying selection overwhelmed by the power of genetic drift in the population. Finally, we estimated the number of founding individuals by using the genome-wide site frequency spectrum of the six samples. Assuming a simple demographic scenario with a single bottleneck followed by exponential growth, the estimated number of founders (~20 individuals) is largely consistent with previous estimates.

Key words: Mauritian cynomolgus macaque, genome sequence, population bottleneck.

Introduction

Nonhuman primates, in particular macaque monkeys, are important biological resources because of their genetic similarity with humans (approximately 94% in nucleotide sequence identity), which is much higher than that of the nonprimate mammal animal models (Gibbs et al. 2007; Shively and Clarkson 2009). However, the difficulty of obtaining genetically homogenous individuals in primates hampers their use in several fields of experimental medicine. Therefore, it is important to elucidate the genetic background of macaques to be able to use these animals for future biomedical research.

The cynomolgus macaque (*Macaca fascicularis*) is one of the most widely used experimental animals in biomedical research, and has been used to study the effect of various medications as well as vaccines against infectious diseases. This species lives in widely distributed range in Southeast

Asia, including areas of Indochina, Malaysia, Indonesia, the Philippines, and also the island of Mauritius, where the animals were only recently introduced by humans (Fooden 1976). Cynomolgus macaques are evolutionarily closely related to rhesus macaques (*Macaca mulatta*), another species which has been extensively studied. Polymorphisms shared between the cynomolgus and rhesus macaques suggest historical gene introgression between the two species, particularly in populations living near the boundary between their geographical distribution areas in the North of the Indochinese peninsula (Bonhomme et al. 2009; Stevison and Kohn 2009; Higashino et al. 2012). The average genetic divergence between the cynomolgus and rhesus macaques is 0.4–0.5% per site in the nuclear genome (Osada et al. 2010), which is considerably close to the average genetic diversity within each species. After the Indian government banned the export of the rhesus macaque to foreign countries in 1978, the importance of the cynomolgus macaques as an alternative resources for biomedical research has been increasingly appreciated (Wade 1978; Pavlin et al. 2009).

Previous studies have shown that cynomolgus macaques are genetically highly heterogeneous (Osada et al. 2010; Yan et al. 2011), and that this genetic heterogeneity could contribute to varied responses to drugs and pathogens (Menninger et al. 2002; Drevon-Gaillot et al. 2006) and influence various biological parameters (Aarnink, Garchon, et al. 2011; Aarnink et al. 2013). Studies using mitochondrial and nuclear genome data have revealed that cynomolgus macaque populations are divided into four major genetic groups (Smith et al. 2007; Blancher et al. 2008; Osada et al. 2010): The Indonesian-Malaysian, Indochinese, Philippine, and Mauritian populations. These four populations show different levels of genetic diversity and have different demographic histories. The Indonesian-Malaysian population is thought to be the ancestral population of cynomolgus macaques, and show the highest level of nucleotide diversity (π), estimated to be $3.0 - 3.2 \times 10^{-3}$ per site (Osada et al. 2010; Higashino et al. 2012; Fan et al. 2014), which is approximately three times higher than that in the entire human population (Prado-Martinez et al. 2013). The macaque population in the Philippines shows slightly reduced genetic diversity, probably because of a recent population size contraction (Osada et al. 2013). Phylogenetic trees of the mitochondrial DNA suggest that the Philippine population was derived from the Indonesian-Malaysian population (Smith et al. 2007; Tosi and Coke 2007; Blancher et al. 2008; Kanthaswamy et al. 2008; Stevison and Kohn 2008). The Indochinese cynomolgus population is thought to have experienced a nonnegligible amount of gene introgression from the rhesus macaques (Kanthaswamy et al. 2008; Bonhomme et al. 2009; Stevison and Kohn 2009), although the historical effect of this interspecies gene flow has not been solely restricted to the Indochinese population (Osada et al. 2010).

Among the four major population groups, the Mauritian population has a particularly interesting demographic history; a small number of individuals were brought to the Mauritian island in the 16th century, where they settled to give rise to a quickly expanding population (Sussman and Tattersall 1986). Consistent with this historical record, the Mauritian population is characterized by a limited number of major histocompatibility complex (MHC) alleles (Leuchte et al. 2004; Krebs et al. 2005; Aarnink, Apoil, et al. 2011; Blancher et al. 2012), a small number of mitochondrial haplotypes (Smith et al. 2007; Tosi and Coke 2007; Blancher et al. 2008), and small numbers of microsatellite alleles at various loci (Bonhomme et al. 2008; Kawamoto et al. 2008). Because of their large population on the island (they are an invasive species in the Mauritian island) and the relatively simple configuration of their MHC alleles, Mauritian cynomolgus macaques have been used in several biomedical studies and their genome was sequenced as the first cynomolgus macaque genome (Ebeling et al. 2011). Although Mauritian cynomolgus macaques are thought to have a highly homogenous genetic background, recent studies using single nucleotide variant (SNV) markers unexpectedly identified genetic structures, indicating that there may be two or three subpopulations within the Mauritian cynomolgus macaques (Ogawa and Vallender 2014). However, these genetic structures do not correspond to their geographic distribution (Satkoski Trask et al. 2013). Through the studies of their MHC locus, it has been demonstrated that their repertoire of MHC haplotypes has been reduced by the founder effect; however, the impact of this population bottleneck on the other nuclear genes has not been well studied at the SNV level.

The extreme population bottleneck in Mauritian cynomolgus macaques also provides evolutionary insight into how, in such circumstances, deleterious mutations can accumulate in genomes. Theoretical studies predict that the reduction of effective population size reduces the efficacy of natural selection and could result in the segregation and fixation of slightly deleterious mutations (Ohta 1973). Extreme population bottleneck reduces genetic diversity; hence, may cause a decline of the average fitness of a population. Mauritian cynomolgus macaques have well thrived in the island and rapidly expanded their population size (Sussman and Tattersall 1986). The well-documented demographic history of Mauritian cynomolgus macaque may provide a good opportunity for investigating the effect of extreme population bottleneck on the genome-wide pattern of polymorphisms.

To date, whole-genome sequences of one Malaysian (Higashino et al. 2012), one Vietnamese (Yan et al. 2011), and one Mauritian (Ebeling et al. 2011) cynomolgus macaque have been analyzed. More recently, large-scale genome sequencing of Mauritian cynomolgus macaques was performed to find genetic causes of viral susceptibility (Ericsen et al. 2014). However, analyzing randomly sampled individuals to infer the past demography of the Mauritian cynomolgus

macaques has not yet been conducted. Clarifying the genetic background of Mauritian cynomolgus macaques is of great importance for both biomedical and evolutionary research. Here, we report the whole-genome sequences of six Mauritian cynomolgus macaques with an approximately 20-fold coverage of the genome.

Materials and Methods

DNA Sequencing

We extracted DNA from blood samples of wild-caught male Mauritian cynomolgus macaques, initially used for studying the sex-matched response to SIV infection (Aarnink, Dereuddre-Bosquet, et al. 2011). At sampling, there was no evidence that they were closely related with each other. Genome sequencing libraries of approximately 450 bp length were constructed for each of the six macaques. Paired-end sequences of 100 bp were determined using HiSeq2000 (Illumina Inc, San Diego, CA). The library construction, sequencing, and initial quality check were performed at Beijing Genomics Institute (Shenzhen, China).

SNV Calling

Reads were mapped on the draft genome of the rhesus macaque (rheMac2), the draft Y chromosome sequence (Hughes et al. 2012), and the mitochondrial genome (DDBJ/GenBank/EMBL accession number: AY612638) using the BWA aln/sampe algorithm with default parameter settings, except for quality trimming score of −15 (Li and Durbin 2009). Among the samples, the average mapping rate of reads was 93.3%. Potential polymerase chain reaction duplicates were marked using Picard software (http://picard.sourceforge.net, last accessed March 5, 2015). SNVs were jointly called on all six samples using the Best Practice pipeline of the Genome Analysis Toolkit software package (Mckenna et al. 2010), which includes base quality score recalibration, insertion/deletion (indel) realignment, SNV calling, and variant quality score recalibration (Van Der Auwera et al. 2002; Depristo et al. 2011). After calling the initial set of variants, further application of the following hard filters was employed: FS > 60.0, HaplotypeScore > 13.0, MQ < 40.0, MQRankSum < −12.5, QD < 2.0, ReadPosRankSum < −8.0. SNVs on fragmented scaffolds (chrUr) were not included in analysis. Heterozygosity within individuals and nucleotide diversity (π) were estimated using only high coverage sites (≥10-folds). All raw read sequences and initial sets of variants are deposited into the public database (EMBL-EBI accession number: PRJEB7871).

Principal Component Analysis and Population Tree

Principal component analysis (PCA) was conducted using the smartpca program in the EIGENSOFT software package (Patterson et al. 2006). Extraction, filtering, and processing of data were performed using custom-made perl scripts. A population tree was constructed using three additional genome sequences of macaques (Yan et al. 2011; Higashino et al. 2012). PHYLIP software (Felsenstein 1989) was used to generate the distance matrix for the macaque individuals by Nei's genetic distance (Nei 1972). For the tree construction, the allele frequency data of SNV sites (coverage ≥ 10) found in all nine individuals were used. The phylogenetic tree was constructed using the neighbor-joining method (Saitou and Nei 1987) implemented in MEGA6 (Tamura et al. 2013).

Pairwise Sequentially Markovian Coalescent Method

The analysis was performed using Pairwise Sequentially Markovian Coalescent (PSMC) software (Li and Durbin 2011). Consensus genome sequences for PSMC input were constructed using samtools and vcf2fq utility (Li et al. 2009). The time interval parameter of 4+25*2+4+6 and the number of iterations of 25 were used for the parameters of PSMC.

Prediction of Disease Causing Mutations

To infer the disease causality of nonsynonymous mutations in the Mauritian cynomolgus macaques, we identified respective nonsynonymous mutations in human orthologs and predicted the functional effect using PolyPhen-2 (Adzhubei et al. 2010). Gene annotation of the rhesus macaque followed the annotation in a previous study (Higashino et al. 2012). The human–macaque ortholog information was retrieved from the Ensembl database (Flicek et al. 2014). Human and macaque protein sequences were aligned using ClustalW (Thompson et al. 1994) and only the sites that have the same amino acid residues between human and macaque reference proteins in the alignment were analyzed through the PolyPhen-2 website (http://genetics.bwh.harvard.edu/pph2/, last accessed March 5, 2015). From this analysis, we had a final functional prediction of 7,976 nonsynonymous mutations.

Site Frequency Spectrum

Folded site frequency spectrum (fSFS) of *i*th occurrence was defined as $C'_i = C_i + C_{n-i}, i : i < n/2$ and $C'_i = C_i, \mathrm{i} : i = n/2$, where C_i represents the number of variants observed for i chromosomes and n is the number of sampled chromosomes. The number of sampled chromosomes in our study is 12 (diploid chromosomes of six individuals). To correct the excess of SNVs that are heterozygous in all individuals, most of which are thought to be due to genotyping error (see also Results and Discussion section), we applied a simple correction method assuming the Hardy–Weinberg equilibrium. We assumed that the allele frequency of all SNV sites that showed heterozygosity in all six individuals was 0.5, for which the highest proportion (1/2) of heterozygotes is expected. It should be noted that this assumption is conservative. We

denoted the observed number of SNV sites that showed heterozygosity in all six individuals by C_{6_H6}, with a corresponding expected probability of 0.5^6. The observed number of SNV sites that have a frequency of 0.5 and are not heterozygous in all samples is C_{6_nH6}. If $\hat{C}_{6\_nH6}$ is the true number of mutations that are heterozygous in all individuals, the following relationship should hold: $(C_{6\_nH6} + \hat{C}_{6\_H6}) \times 0.5^6 = \hat{C}_{6\_H6}$. The number of $\hat{C}_{6\_H6}$ was estimated by solving this equation.

Estimating the Number of Founders

The level of the past population bottleneck was inferred by fitting expected fSFS to the observed fSFS using the analytical formula obtained by Marth et al. (2004). We considered a single bottleneck event, followed by exponential growth, which has two population genetic parameters to be estimated: The time of the bottleneck (T_b) and the size of the bottleneck (N_b). The ancestral population size (N_a) and the current population size (N_0) were fixed for each estimation. Because the model is scalable to any population size, we estimated fSFS for when N_a is 100,000, and scaled the parameters after fitting. The deviance of expected to observed fSFS was evaluated using χ^2 statistics for very small intervals for each T_b and N_b. In addition, we confirmed that the analytical formula and coalescent simulations gave highly consistent expected fSFS using our population growth model (data not shown).

Results and Discussion

Identification of SNVs and Estimation of Nucleotide Diversity

We obtained 100-bp-length Illumina paired-end sequences from six unrelated Mauritian cynomolgus macaques and mapped the reads to the reference rhesus macaque genome (see Materials and Methods section). We did not map the reads to the reference genome of the Mauritian cynomolgus macaque (Ebeling et al. 2011), because the rhesus macaque reference has better gene annotation and previous studies have shown that rhesus macaque genomes are sufficiently close to cynomolgus macaque genomes for read mapping by typical short-read mappers (Yan et al. 2011; Higashino et al. 2012). The average coverage is approximately 20-fold for each individual (table 1). In total, we identified approximately 21.8 million SNVs and 1.9 million indels against the reference genome among the six macaques on the autosomes. Because all samples are males, sex chromosomes and mitochondrial genome are all haploid genomes in our samples. Therefore, we mainly focused on the pattern of SNVs on the autosomes in this study. Summary of SNVs identified on the sex chromosomes is shown in supplementary table S1, Supplementary Material online. Each individual has an average of 5.8 million heterozygous and 8.3 million homozygous SNVs, and these numbers are highly consistent among individuals. Here, homozygous variants are defined against the reference genome sequence of rhesus macaque. Estimation of genetic diversity within the Mauritian cynomolgus macaque population was 2.28×10^{-3} for nucleotide diversity (π). We retrieved the previously published Malaysian cynomolgus macaque genomes and estimated heterozygosity using the same criteria for SNV identification ($\pi = 2.96 \times 10^{-3}$). The heterozygosity of the Mauritian cynomolgus macaques was 23% smaller than the Malaysian cynomolgus macaque, which is thought to have very high genetic diversity.

Genetic Relationship between and within Populations

In addition to the Malaysian cynomolgus macaque genome, we retrieved the two more previously published macaque genomes (Vietnamese cynomolgus macaque and Chinese rhesus macaque). Genetic relationships among the six Mauritian cynomolgus macaque individuals were examined using PCA plot (fig. 1). We confirmed that no individuals were closely overlapped in the plot. A plot including all nine macaque genomes is presented in the supplementary figure S1, Supplementary Material online. We further examined whether the Mauritian cynomolgus macaques are genetically closer to the Malaysian or to the Indochina cynomolgus macaques. Figure 2 shows the phylogenetic relationship of the four macaque populations. Consistent with the results from mitochondrial data (Smith et al. 2007), the Mauritian cynomolgus macaques are genetically closer to the Malaysian cynomolgus macaques. Because genome sequences of the Indonesian populations have not been analyzed, we were not able to determine the detailed origin of the Mauritian cynomolgus macaques.

In addition, the past demography was estimated using the PSMC method (Li and Durbin 2011). The inferred demographic histories are shown in figure 3. The six Mauritian cynomolgus macaques showed a very similar trend of past demography. This result indicates that they are not derived from genetically distinct origins, which agrees with that of the mitochondrial data (Smith et al. 2007). However, we should note that PSMC would not work for the Mauritian cynomolgus macaques to properly scale time and population size because this analysis has a limitation in inferring very recent population size changes. The actual bottleneck in the Mauritian cynomolgus macaques was very recent to be inferred by PSMC. If the genome-wide heterozygosity were dramatically changed by very recent demographic events, scaling parameters would fail. In this study, our purpose of using PSMC was to check whether the population size trajectories overlap with each other, and not for estimating demographic parameters themselves; therefore, in figure 3, we only showed parameter values scaled by $N_0 = 10{,}000$, which was arbitrarily determined. The confidence intervals of population size estimation are shown in the supplementary figure S2, Supplementary Material online. In order to infer the recent

Table 1
Summary of Variant Calling in Six Mauritian Cynomolgus Macaques

Sample ID	Average Sequencing Depth[a]	Total SNV	Heterozygous SNV	Homozygous SNV	Heterozygosity
Tlse-8102 (MCM1)	20.80	14,048,997	5,676,969	8,372,028	0.00225
Tlse-8141 (MCM2)	20.42	14,045,116	5,814,396	8,230,720	0.00231
Tlse-8249 (MCM3)	19.14	14,175,393	5,947,573	8,227,820	0.00236
Tlse-9204 (MCM4)	20.33	14,116,067	5,837,878	8,278,189	0.00232
Tsle-9413 (MCM5)	19.61	14,150,408	5,883,805	8,266,603	0.00234
Tlse-9859 (MCM6)	20.42	14,080,208	5,753,557	8,326,651	0.00229
Malaysian cynomolgus macaque	26.1	12,758,246	7,177,728	5,580,518	0.00296

[a]Reads mapped on autosomes.

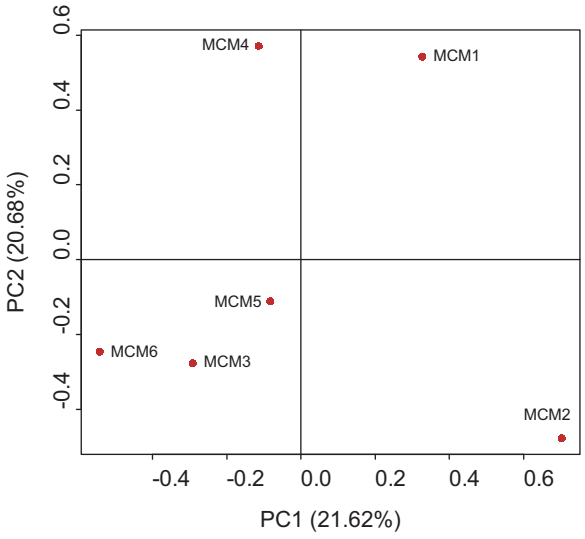

Fig. 1.—PCA plot of the six Mauritian cynomolgus macaque individuals. The individual ID is given beside each data point. The *x*- and *y*-axes represent the first and second principal components, respectively.

demography we applied the method using information on polymorphism frequency, which is described in the later section.

Site Frequency Spectrum of SNVs

To elucidate a more detailed pattern of the polymorphisms in the Mauritian cynomolgus macaques, we calculated the SFS of mutations in the six samples. Because we cannot assume that the reference rhesus macaque genome has ancestral states, the spectrum is folded (see Materials and Methods section). Before evolutionary inference, we carefully examined the potential genotyping errors that could affect the pattern of SFS. We found an excess of SNVs for which all six macaques were heterozygous (H6 sites); this fraction of H6 sites was enriched with nonsynonymous mutations. Fixation of segmental duplication with subsequent mutations may cause false identification of heterozygosity at such sites; alternatively, these sites would be observed when one of the duplicated loci is not present in the reference genome sequence. If either of these cases were the cause of the H6 sites, we would expect these sites to have higher genome sequencing coverage. To evaluate this, we compared the occurrence of H6 sites with the average coverage depth among the six genomes at those sites. Repeat regions of the genome were excluded from this analysis to avoid the complex effect of repetitive sequences. The results showed that the coverage distribution for H6 sites was skewed toward higher coverage, and the coverage distribution for nonsynonymous and synonymous sites among the H6 sites was more strongly biased toward higher coverage (fig. 4). Although we cannot identify the reason for these systematic biases, this pattern of miscalling should be carefully interpreted in future whole-genome sequencing studies. To remove the potential genotyping errors for the analysis of fSFS data, we corrected the miscalling of H6 sites by assuming the Hardy–Weinberg equilibrium (see Materials and Methods).

In figure 5*A*, fSFS for nonsynonymous, synonymous, and noncoding sites is shown. Notably, nonsynonymous and synonymous sites are defined among the six Mauritian cynomolgus macaque alleles. Compared with a neutral expectation with a constant population size, Mauritian cynomolgus macaques harbor significantly fewer low-frequency polymorphisms, particularly singletons ($P < 10^{-15}$; χ^2 test). A reduction of population size is expected to affect low-

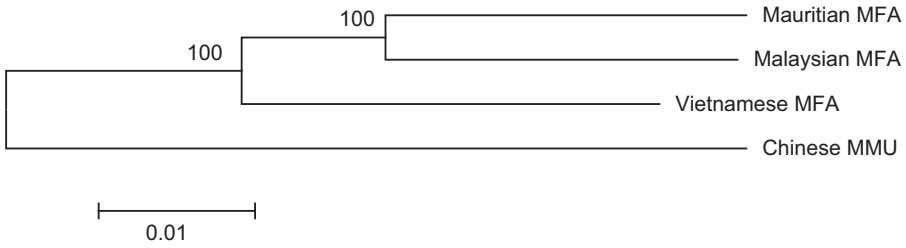

Fig. 2.—Phylogenetic relationships of the four populations. MFA and MMU designate *M. fascicularis* and *M. mulatta*, respectively. The branch length represents Nei's genetic distance. Bootstrap confidence values (percentile) are shown upon the branches.

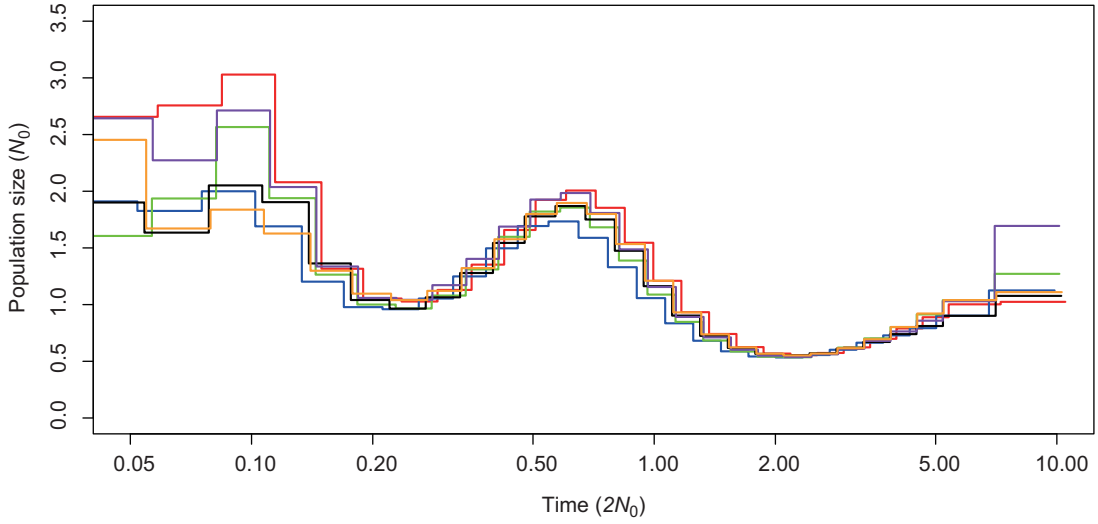

Fig. 3.—The change in population size inferred by PSMC. Six individuals, MCM1–MCM6, are labeled by red, blue, green, black, purple, and orange lines, respectively. The time from the present and effective population size are shown in the *x*- and *y*-axes, respectively. Note that the time and the size were arbitrarily scaled by baseline effective population size (N_0) equal to 10,000 (see Results and Discussion).

frequency polymorphisms more than common-frequency polymorphisms (Luikart et al. 1998). Because π is more sensitive to the difference in common-frequency polymorphisms, we expected that π would not be greatly affected by the very recent population bottleneck (Tajima 1989).

Interestingly, the patterns of fSFS for noncoding, synonymous, and nonsynonymous mutations are not strongly different within the Mauritian cynomolgus macaque population. In particular, most of the large-scale population genetic studies in humans (e.g., Fujimoto et al. 2010) have found an excess of low frequency nonsynonymous mutations; however, this was not observed in the Mauritian macaque, although the difference between synonymous and nonsynonymous mutations was statistically significant ($P < 10^{-15}$; χ^2 test). In addition, nonsynonymous and synonymous mutations showed similar level of singletons ($P = 0.67$; χ^2 test). Because recent population bottleneck mostly affects the pattern of rare polymorphisms, mutations segregating within the macaque population at low frequencies were rapidly lost during the bottleneck period, and the time elapsed since the bottleneck has been short to allow for the appearance of new mutations.

We examined the phenotypic effect of mutations using predictions of disease causality in human genes. Predictions of the functional effect on nonsynonymous SNVs were performed using PolyPhen-2 (Adzhubei et al. 2010), which predicts the potential impact of an amino acid substitution based on protein structure and evolutionary conservation. We excluded all H6 mutations from the disease-causing mutation analyses because most of them are likely genotyping errors. The fraction of potentially damaging mutations for each fSFS category is shown in figure 5*B*. The excess of damaging mutations in the singleton class was not statistically significant ($P = 0.17$; χ^2 test), which is considerably different from the pattern in humans (Andrés et al. 2009).

On an average, Mauritian cynomolgus macaques have 10,565 nonsynonymous and 13,533 synonymous heterozygous SNVs. The ratio of nonsynonymous to synonymous polymorphisms was 0.78, which is significantly higher than the

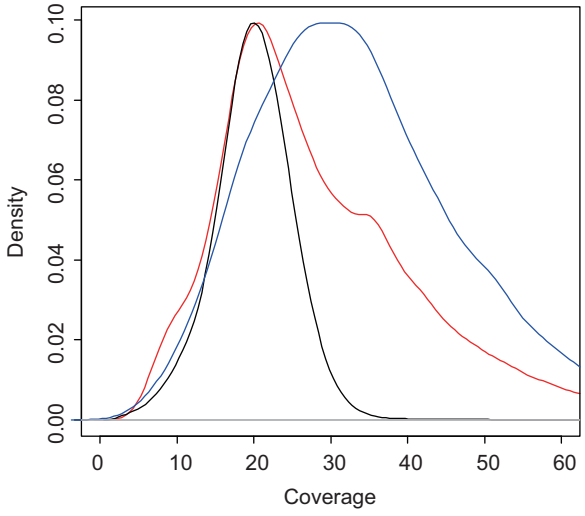

Fig. 4.—Genome sequencing coverage of SNV sites. The height of the lines shows the density estimation of read coverage. The black, red, and blue lines represent the estimated density for all SNV sites, noncoding H6 sites, and coding H6 sites, respectively. H6 sites are the sites where all six samples are heterozygous.

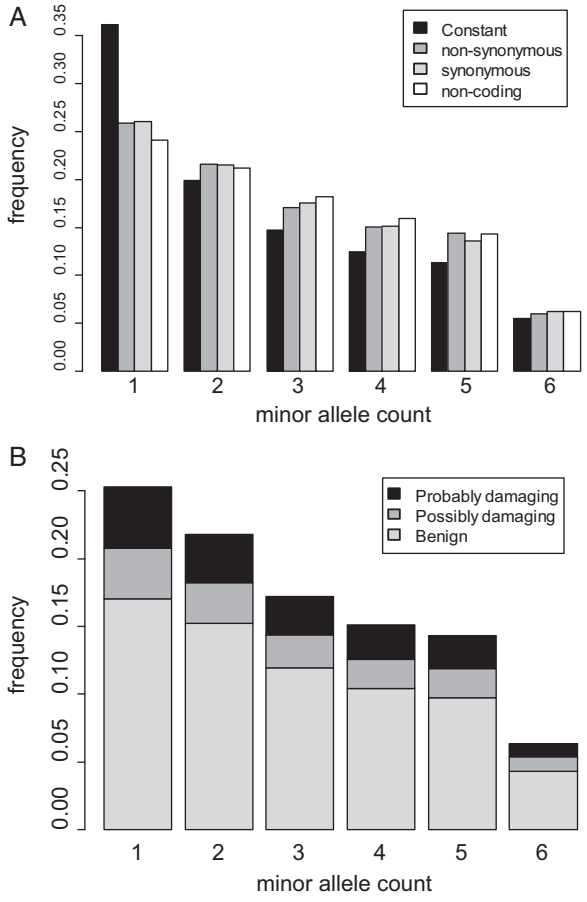

Fig. 5.—fSFS for nonsynonymous (dark gray), synonymous (light gray), and noncoding (white) sites among the six Mauritian cynomolgus macaque individuals. The expected frequency with constant population size is shown by the black bar (*A*). Prediction of disease-causing mutations was performed using PolyPhen-2. Probably damaging, possibly damaging, and benign nonsynonymous mutations are shown in the black, dark gray, and light gray bars, respectively (*B*).

ratio observed in the Malaysian cynomolgus macaque individual (0.68; Higashino et al. 2012; $P < 10^{-15}$; χ^2 test). The higher ratio in the Mauritian cynomolgus macaques indicates that more deleterious mutations are segregating with greater frequency in the population. In addition, we found that 12,467 nonsynonymous and 17,749 synonymous changes are fixed among the Mauritian cynomolgus macaque samples compared with the reference rhesus macaque genome.

Considering that the proportion of heterozygosity at nonsynonymous SNVs/synonymous SNVs is higher in Mauritian individuals than in the Malaysian individual, we concluded that the pattern of polymorphisms in the Mauritian cynomolgus macaques has been predominantly shaped by a strong genetic drift and has overwhelmed by the power of purifying selection during the population bottleneck.

However, the data also showed that, at the same time, low-frequency nonsynonymous polymorphisms have been effectively removed from the population by genetic drift. Therefore, there have been both gain and loss of deleterious mutations in the population. The observation is consistent with recent theoretical and experimental studies in humans, which found that the genetic load of the population is not strongly affected by recent demographic changes (Lohmueller 2014; Simons et al. 2014; Do et al. 2015).

Estimation of Demography

To investigate whether the observed fSFS agrees with the extreme population bottleneck from the known historical record, we estimated the level of population bottleneck by fitting expected fSFS to observed fSFS. To this end, we assume a simple demographic scenario, where a small number of individuals were introduced to the island from the ancestral population with a constant population size, followed by a quickly increased population size with exponential growth (fig. 6). Four parameters: Ancestral effective population size (N_a), current effective population size (N_0), the effective population size at bottleneck (N_b), and timing of bottleneck (T_b), are involved in this model. Our approach estimated N_b for a given N_a, N_0, and T_b by fitting observed fSFS to expected fSFS with grid sampling of N_b and T_b (Marth et al. 2004). Because the number of analyzed sites is so large (approximately 16 million), the confidence intervals of the estimate became small. Therefore, although we need be careful in the interpreting these results, here we present only the

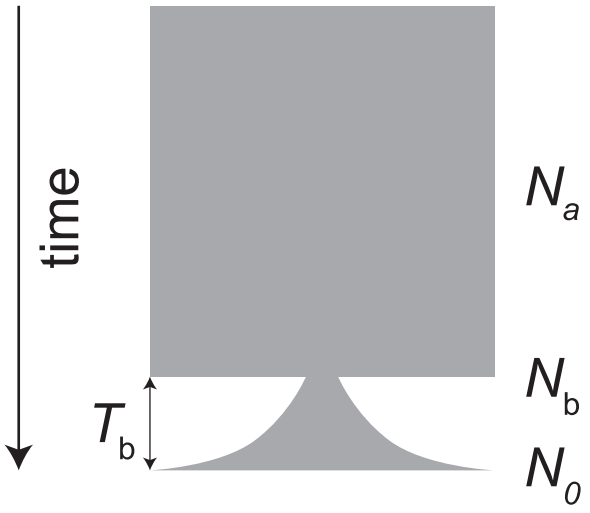

Fig. 6.—Proposed demographic model for estimating the number of founding individuals. The width of the shaded area represents effective population sizes. We assumed that the bottleneck occurred at time T_b and the number of N_b individuals were randomly selected at the bottleneck from an ancestral population with constant population size of N_a. After the bottleneck the size of population recovered to N_0 with exponential growth.

maximum-likelihood point estimates of N_b for a given N_a, N_0, and T_b.

According to the historical record, introduction of macaques to the Mauritian islands was around 400–500 years ago. The estimation of a generation time for macaques is uncertain, ranging from 5 to 12 years (Gage 1998). Any bias in generation time estimation would affect the accuracy of the estimation of N_b. Therefore, we used the two long-term demographic studies of Japanese macaques (*Macaca fuscata*) to calculate the average time of reproduction of females, which yielded a generation time of 9.6–11.4 years (Koyama et al. 1992; Fujimoto et al. 2010). In this study, we applied the estimation of 10-year generation time, which means the population bottleneck occurred 40–50 generations ago. In the following estimation, we assumed $T_b = 40$. Assuming $N_a = 30{,}000$ and $N_0 = 30{,}000$, the estimated number of individuals during the bottleneck is 16. This estimated number of founders is robust against the assumption of N_a and N_0. For example, $N_b = 15$ assuming $N_a = 50{,}000$ and $N_0 = 25{,}000$, $N_b = 14$. In table 2, the estimated numbers of founders with different values of N_a and N_0 are shown. The range of the estimated number of founders does not contradict the previous microsatellite (Bonhomme et al. 2008) and mitochondrial data (Smith et al. 2007).

In addition to the exponential growth model, we examined logistic growth models. These models showed a better fit than the previous study using microsatellite data (Bonhomme et al. 2008). In general, the logistic growth models yielded a smaller number of founding individuals than the exponential growth model. This is because the logistic model has more rapid growth in the early phase, which makes the bottleneck less effective. Using similar parameter settings as the study of Bonhomme et al. (2008), a generation time of 5 years and a growth rate of 0.3, we obtained a slightly smaller number of founding individuals (2–8 founders) than the estimated number by Bonhomme et al. (12 founders). However, in this study, we did not thoroughly apply different growth models because our data have a limited sample size and may not have enough power to infer a very recent demographic history.

Table 2
Estimated Number of Founders with Different N_a and N_0

	$N_a = 50{,}000$	$N_a = 30{,}000$
$N_0 = N_a$	15	16
$N_0 = N_a/2$	17	18
$N_0 = N_a/5$	20	21

The estimated number of founders assuming exponential growth may be overestimated because reports of large-scale haplotyping in the MHC region have identified only seven founding MHC haplotypes in the Mauritian cynomolgus macaque population (Wiseman et al. 2007; Mee et al. 2009; Budde et al. 2010; Blancher et al. 2012; Aarnink et al. 2014) and eight haplotypes in the killer cell immunoglobulin-like receptor (KIR) region (Bimber et al. 2008). In this scenario, the lower limit of the founding individuals is 4, which is closer to our estimated number assuming logistic growth. However, the probability of allele loss is highly dependent on the initial pattern of population growth; this is difficult to accurately estimate, and the process could be highly stochastic with a small number of founders.

Natural selection could have preserved the number of alleles at the MHC locus; in particular, a recent study found that there is MHC class I semi-incompatibility between mother and offspring in cynomolgus macaques; thus, natural selection would act against the loss of MHC alleles in this population (Aarnink et al. 2014). It is of interest to further investigate the effect of natural selection on the genetic diversity at the MHC locus in future studies using whole-genome sequences of more individuals from the populations from which the Mauritian macaques originated.

Conclusions

In this article, we report the genome sequences of six Mauritian cynomolgus macaques. The pattern of polymorphisms in these animals shows a reduced level of genetic diversity, particularly in low-frequency polymorphisms. This pattern agrees well with the historical record of an extreme population bottleneck during the founding of this population. The low efficacy of purifying selection on their genomes may provide the further insight into the specific phenotypic

characteristics in Mauritian cynomolgus macaques. The smaller genetic diversity in this population is of great importance for better reproducibility of drug testing and viral infection experiments. In addition, the whole-genome sequences of the Mauritian cynomolgus macaques provide further insights into the genetic basis of variation among macaques for drug and viral response in future biomedical research.

Supplementary Material

Supplementary table S1 and figures S1 and S2 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

This work was supported by the funds from University Toulouse III (EA3034, Université Paul Sabatier) and French Ministry of Research and by the Grant-in-Aid for Scientific Research (A), Grant number 26251040, to N.S. and N.O. They thank Dr Tomas Marques-Bonet for the assistance of data analysis and the two anonymous reviewers for helpful comments on the manuscript.

Literature Cited

Aarnink A, Apoil PA, Takahashi I, Osada N, Blancher A. 2011. Characterization of MHC class I transcripts of a Malaysian cynomolgus macaque by high-throughput pyrosequencing and EST libraries. Immunogenetics 63:703–713.

Aarnink A, Dereuddre-Bosquet N, et al. 2011. Influence of the MHC genotype on the progression of experimental SIV infection in the Mauritian cynomolgus macaque. Immunogenetics 63:267–274.

Aarnink A, et al. 2013. Comparative analysis in cynomolgus macaque identifies a novel human MHC locus controlling platelet blood counts independently of BAK1. J Thromb Haemost. 11:384–386.

Aarnink A, et al. 2014. Deleterious impact of feto-maternal MHC compatibility on the success of pregnancy in a macaque model. Immunogenetics 66:105–113.

Aarnink A, Garchon HJ, et al. 2011. Impact of MHC class II polymorphism on blood counts of CD4+ T lymphocytes in macaque. Immunogenetics 63:95–102.

Adzhubei IA, et al. 2010. A method and server for predicting damaging missense mutations. Nat Methods. 7:248–249.

Andrés AM, et al. 2009. Targets of balancing selection in the human genome. Mol Biol Evol. 26:2755–2764.

Bimber BN, Moreland AJ, Wiseman RW, Hughes AL, O'Connor DH. 2008. Complete characterization of killer Ig-like receptor (KIR) haplotypes in Mauritian cynomolgus macaques: novel insights into nonhuman primate KIR gene content and organization. J Immunol. 181:6301–6308.

Blancher A, Aarnink A, Savy N, Takahata N. 2012. Use of cumulative Poisson probability distribution as an estimator of the recombination rate in an expanding population: example of the *Macaca fascicularis* major histocompatibility complex. G3 2:123–130.

Blancher A, et al. 2008. Mitochondrial DNA sequence phylogeny of 4 populations of the widely distributed cynomolgus macaque (*Macaca fascicularis fascicularis*). J Hered. 99:254–264.

Bonhomme M, Blancher A, Cuartero S, Chikhi L, Crouau-Roy B. 2008. Origin and number of founders in an introduced insular primate: estimation from nuclear genetic data. Mol Ecol. 17:1009–1019.

Bonhomme M, Cuartero S, Blancher A, Crouau-Roy B. 2009. Assessing natural introgression in 2 biomedical model species, the rhesus macaque (*Macaca mulatta*) and the long-tailed macaque (*Macaca fascicularis*). J Hered. 100:158–169.

Budde M, et al. 2010. Characterization of Mauritian cynomolgus macaque major histocompatibility complex class I haplotypes by high-resolution pyrosequencing. Immunogenetics 62:773–780.

Depristo MA, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 43: 491–498.

Do R, et al. 2015. No evidence that selection has been less effective at removing deleterious mutations in Europeans than in Africans. Nat Genet. 47:126–131.

Drevon-Gaillot E, Perron-Lepage M-F, Clément C, Burnett R. 2006. A review of background findings in cynomolgus monkeys (*Macaca fascicularis*) from three different geographical origins. Exp Toxicol Pathol. 58:77–88.

Ebeling M, et al. 2011. Genome-based analysis of the nonhuman primate *Macaca fascicularis* as a model for drug safety assessment. Genome Res. 21:1746–1756.

Ericsen A, et al. 2014. Whole genome sequencing of SIV-infected macaques identifies candidate loci that may contribute to host control of virus replication. Genome Biol. 15:478.

Fan Z, et al. 2014. Whole-genome sequencing of Tibetan macaque (*Macaca thibetana*) provides new insight into the macaque evolutionary history. Mol Biol Evol. 31:1475–1489.

Felsenstein J. 1989. Phylip (version 3.2): phylogeny inference package. Cladistics 5:164–166.

Flicek P, et al. 2014. Ensembl 2014. Nucleic Acids Res. 42:D749–D755.

Fooden J. 1976. Provisional classifications and key to living species of macaques (primates: *Macaca*). Folia Primatol (Basel). 25:225–236.

Fujimoto A, et al. 2010. Whole-genome sequencing and comprehensive variant analysis of a Japanese individual using massively parallel sequencing. Nat Genet. 42:931–936.

Gage TB. 1998. The comparative demography of primates: with some comments on the evolution of life histories. Annu Rev Anthropol. 27:197–221.

Gibbs RA, et al. 2007. Evolutionary and biomedical insights from the rhesus macaque genome. Science 316:222–234.

Higashino A, et al. 2012. Whole-genome sequencing and analysis of the Malaysian cynomolgus macaque (*Macaca fascicularis*) genome. Genome Biol. 13:R58.

Hughes JF, et al. 2012. Strict evolutionary conservation followed rapid gene loss on human and rhesus Y chromosomes. Nature 483: 82–86.

Kanthaswamy S, et al. 2008. Hybridization and stratification of nuclear genetic variation in *Macaca mulatta* and *M. fascicularis*. Int J Primatol. 29:1295–1311.

Kawamoto Y, et al. 2008. Genetic diversity of longtail macaques (*Macaca fascicularis*) on the island of Mauritius: an assessment of nuclear and mitochondrial DNA polymorphisms. J Med Primatol. 37: 45–54.

Koyama N, Takahata Y, Huffman M, Norikoshi K, Suzuki H. 1992. Reproductive parameters of female Japanese macaques: thirty years data from the Arashiyama troops, Japan. Primates 33:33–47.

Krebs KC, Jin Z, Rudersdorf R, Hughes AL, O'Connor DH. 2005. Unusually high frequency MHC Class I alleles in Mauritian origin cynomolgus macaques. J Immunol. 175:5230–5239.

Leuchte N, et al. 2004. MhcDRB-sequences from cynomolgus macaques (*Macaca fascicularis*) of different origin. Tissue Antigens 63:529–537.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25:1754–1760.

Li H, Durbin R. 2011. Inference of human population history from individual whole-genome sequences. Nature 475:493–496.

Li H, et al. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079.
Lohmueller KE. 2014. The impact of population demography and selection on the genetic architecture of complex traits. PLoS Genet. 10: e1004379.
Luikart G, Allendorf F, Cornuet J-M, Sherwin W. 1998. Distortion of allele frequency distributions provides a test for recent population bottlenecks. J Hered. 89:238–247.
Marth GT, Czabarka E, Murvai J, Sherry ST. 2004. The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. Genetics 166:351–372.
Mckenna A, et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20:1297–1303.
Mee ET, et al. 2009. MHC haplotype frequencies in a UK breeding colony of Mauritian cynomolgus macaques mirror those found in a distinct population from the same geographic origin. J Med Primatol. 38:1–14.
Menninger K, et al. 2002. The origin of cynomolgus monkey affects the outcome of kidney allografts under neoral immunosuppression. Transplant Proc. 34:2887–2888.
Nei M. 1972. Genetic distance between populations. Am Nat. 106: 283–292.
Ogawa L, Vallender E. 2014. Genetic substructure in cynomolgus macaques (*Macaca fascicularis*) on the island of Mauritius. BMC Genomics 15:748.
Ohta T. 1973. Slightly deleterious mutant substitutions in evolution. Nature 246:96–98.
Osada N, et al. 2010. Ancient genome-wide admixture extends beyond the current hybrid zone between *Macaca fascicularis* and *M. mulatta*. Mol Ecol. 19:2884–2895.
Osada N, et al. 2013. Finding the factors of reduced genetic diversity on X chromosomes of *Macaca fascicularis*: male-driven evolution, demography, and natural selection. Genetics 195:1027–1035.
Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. PLoS Genet. 2:e190.
Pavlin BI, Schloegel LM, Daszak P. 2009. Risk of importing zoonotic diseases through wildlife trade, United States. Emerg Infect Dis. 15: 1721–1726.
Prado-Martinez J, et al. 2013. Great ape genetic diversity and population history. Nature 499:471–475.
Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 4:406–425.
Satkoski Trask J, George D, Houghton P, Kanthaswamy S, Smith DG. 2013. Population and landscape genetics of an introduced species (*M. fascicularis)* on the island of Mauritius. PLoS One 8:e53001.
Shively CA, Clarkson TB. 2009. The unique value of primate models in translational research. Am J Primatol. 71:715–721.
Simons YB, Turchin MC, Pritchard JK, Sella G. 2014. The deleterious mutation load is insensitive to recent population history. Nat Genet. 46:220–224.
Smith DG, Mcdonough JW, George DA. 2007. 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. 69:182–198.
Stevison LS, Kohn MH. 2008. Determining genetic background in captive stocks of cynomolgus macaques (*Macaca fascicularis*). J Med Primatol. 37:311–317.
Stevison LS, Kohn MH. 2009. Divergence population genetic analysis of hybridization between rhesus and cynomolgus macaques. Mol Ecol. 18:2457–2475.
Sussman RW, Tattersall I. 1986. Distribution, abundance, and putative ecological strategy of *Macaca fascicularis* on the Island of Mauritius, Southwestern Indian Ocean. Folia Primatol (Basel). 46:28–43.
Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595.
Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. 2013. MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. Mol Biol Evol. 30:2725–2729.
Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673–4680.
Tosi AJ, Coke CS. 2007. Comparative phylogenetics offer new insights into the biogeographic history of *Macaca fascicularis* and the origin of the Mauritian macaques. Mol Phylogenet Evol. 42: 498–504.
Van Der Auwera GA, et al. 2002. From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. Current Protocols in Bioinformatics. p. 11.10.1-11.10.33. Hoboken (NJ): John Wiley & Sons, Inc.
Wade N. 1978. India bans monkey export: U.S. may have breached accord. Science 199:280–281.
Wiseman RW, et al. 2007. Simian immunodeficiency virus SIVmac239 infection of major histocompatibility complex-identical cynomolgus macaques from Mauritius. J Virol. 81:349–361.
Yan G, et al. 2011. Genome sequencing and comparison of two nonhuman primate animal models, the cynomolgus and Chinese rhesus macaques. Nat Biotechnol. 29:1019–1023.

Associate editor: Yoshihito Niimura