ORIGINAL ARTICLE

Analysis of full-length mitochondrial DNA D-loop sequences from *Macaca fascicularis* of different geographical origin reveals novel haplotypes

Anjna Badhan[1,*], Christina A. Eichstaedt[2], Neil M. Almond[1], Leslie A. Knapp[2,†] & Nicola J. Rose[1]

1 Division of Virology, National Institute for Biological Standards and Control, Medicines and Healthcare Products Regulatory Authority, South Mimms, Hertfordshire, UK

2 Division of Biological Anthropology, Department of Archaeology and Anthropology, University of Cambridge, Cambridge, UK

* Current address: Public Health England, 61 Colindale Avenue, Colindale, London, NW9 5EQ, UK

† Current address: Department of Anthropology, The University of Utah, Salt Lake City, UT, 84112, USA

Abstract

Background Cynomolgus macaques are indigenous to Asia occupying a range of geographical areas. A non-indigenous population established on Mauritius approximately 500 years ago. Mauritian cynomolgus macaques are recognised as having low genetic diversity compared to Indonesian macaques, from which they originated. As cynomolgus macaques are widely used as a biomedical model, there have been many studies of their genetic relationships. However, population diversity and relationships have only been assessed through analysis of either the hypervariable region I or II separately within the D-loop region of the mitochondrial genome in these macaques.
Methods Using sequencing, we defined haplotypes encompassing the full D-loop sequence for Mauritian and Indonesian cynomolgus macaques.
Results We evaluated the haplotype relationships by constructing a median-joining network based on full-length D-loop sequences, which has not been reported previously.
Conclusion Our data allow a complete D-loop haplotype, including a hereto unreported polymorphic region, to be defined to aid the resolution of populations of cynomolgus macaques and which highlights the value in analysing both D-loop hypervariable regions in concert.

Introduction

Cynomolgus macaques (crab-eating macaques) are indigenous to Asia, inhabiting a wide geographical area, including much of mainland South-East Asia, Malaysia, the Philippines and parts of the Indonesian archipelago. In addition, a non-indigenous population of cynomolgus macaques has expanded following introduction onto the island of Mauritius in the past 400–500 years [36]. Genetic characterisation through mitochondrial DNA (mtDNA) analyses of the Asian and Mauritian macaques has revealed considerable variation between the groups [5, 16, 17, 22, 33, 35]. Most notably, Mauritian-origin cynomolgus macaques (MCM) are less genetically

diverse than macaques from other regions, likely as a result of a small founder population [5, 7, 19, 20, 22, 26, 36, 39]. There have been many speculations regarding the exact origin of Mauritian cynomolgus macaques. The data in support of Sumatra [39] and Java [17] are compelling, but it is possible that Mauritian cynomolgus macaques are of mixed origin [5].

There have been various classifications of cynomolgus macaques based on genetic analysis. Many studies focus on the mitochondrial DNA (mtDNA) control region (D-loop) sequence, particularly the hypervariable regions I and II (HVI, HVII). Haplotype analysis of regions of mtDNA has been used to characterise relationships between cynomolgus macaques. Early studies

used restriction fragment length polymorphism (RFLP) analysis of the D-loop to characterise haplotypes, and less diversity was noted in the Mauritian macaques than Indonesian macaques [22]. Shiina et al. [32] focussed on Indochinese macaques, but were unable to clarify individual subpopulations based on HVII sequences. Blancher et al. [5] defined 70 HVII haplotypes and categorised cynomolgus macaques into two groups termed *insular*, which included Filipino, Mauritian and some Indonesian cynomolgus macaques, and *continental*, which included those that originated from Malaysia, Indochina and Indonesia. This geographical differentiation has been supported through sequence analysis of mtDNA sequences spanning the 3′ end of the 12S ribosomal RNA gene, tRNA-Val, and the 5′ end of the 16S ribosomal gene [39], which included Mauritian and Indonesian cynomolgus macaques (ICM) of Sumatran origin in the insular group; yet Y-chromosomal sequence analysis places Mauritian cynomolgus macaques in the continental group [39]. While studies of chromosomal and mitochondrial DNA yield information on two distinct genetic systems which are not necessarily in conflict, it is clear that additional analysis would add further information to the field of cynomolgus macaque origins.

The extent of mtDNA D-loop haplotype variation is not fully established, and many haplotypes are based on just one hypervariable region. Two groups have reported just two different haplotypes for the HVII region [5, 17], whereas analysis of the more variable HVI region and the cytochrome B gene revealed nine haplotypes [34]. Indonesian cynomolgus macaques have greater diversity than Mauritian cynomolgus macaques. A total of 58 HVI [34] and 35 HVII [5] haplotypes have been reported for Indonesian cynomolgus macaques. To date, analysis of the two hypervariable regions in concert has not been undertaken, and it is unclear how polymorphism in HVI relates to polymorphism in HVII.

Macaques, particularly cynomolgus, rhesus and pigtail macaques, are widely used in infectious disease pathogenesis models [12], vaccine development, including HIV/AIDS, influenza and tuberculosis (reviewed in [2, 6, 13]), and organ transplant research (reviewed in [14]). It is becoming clear that the geographical origin of macaques impacts on their susceptibility to pathogens. Thus, a more complete understanding of macaque genetics will better inform the design of studies and the selection and distribution of individuals used in them.

We predicted that sequence and phylogenetic analyses of the entire mitochondrial D-loop sequence, in addition to considering the hypervariable regions in isolation, would yield a greater depth of information on the genetic diversity of Mauritian- and Indonesian-derived cynomolgus macaques. We therefore sequenced the entire mtDNA D-loop of 30 Mauritian and 15 Indonesian cynomolgus macaques and identified full haplotypes. Five previously unreported haplotypes for Mauritian cynomolgus macaques and nine previously unreported haplotypes for Indonesian cynomolgus macaques were characterised in this way highlighting the value of assessing the full sequence. Furthermore, previously unreported sequences identified for the HVI region in Mauritian cynomolgus macaques, and both HVI and HVII regions in Indonesian cynomolgus macaques add to those previously reported [5, 17, 34].

Materials and methods

Humane care guidelines and origin of biological materials

Archived DNA samples were available from 30 Mauritian-derived cynomolgus macaques and 15 Indonesian-derived cynomolgus macaques which had been housed and maintained in accordance with United Kingdom Home Office guidelines for the care and maintenance of non-human primates. The origin of the macaques has been determined by MHC haplotyping [23, 24]. The precise source of the Indonesian-derived animals that formed the group from which the 15 study animals were derived is as yet unknown; however, mtDNA sequences cluster away from groups which contain samples from animals derived from Mauritius, Java, Sumatra, the Philippines and China (N.J. Rose, unpublished data). Pedigree (family groups) information was available for 22 of these animals: these consisted of four family pedigrees for Mauritian cynomolgus macaques and 19 unrelated macaques, in addition to four family pedigrees for Indonesian-derived cynomolgus macaques and five unrelated macaques.

Amplification and sequencing of mtDNA

DNA was diluted to a concentration of 10 ng/μL in molecular grade water. PCR and sequencing primers were designed by alignment of mtDNA sequences of *Macaca mulatta* (AY612638) and *Macaca sylvanus* (AJ309865). To minimise the chance that amplified sequences were derived from a nuclear DNA pseudogene (numt; reviewed in [4], which are generally less than 1 kb in length (D-loop region is <1.1 kb in length), a 1.231-kb mtDNA fragment encompassing the entire D-loop was amplified from each animal using oligonucleotides M13-LF1-AB-N (5′ GTA AAA CGA CGG CCA GTC YAA CTC YAC CRC CAA CAC 3′) and

M13-AB-C (5′ GCG GAT AAC AAT TTC ACA CAG GAT GYT TGC ATG TGT AAT C 3′). The identity of the amplified sequences was confirmed by alignment against complete mitochondrial genome sequences for *M. fascicularis* (NC_012670) and *M. sylvanus* (AJ309865) as well as against three sequences for *M. mulatta* to represent those derived from Indian- and Chinese-origin macaques (AY612638, 'M.mulatta1'; JN885895, 'M.mulatta2'; JN885890, 'M.mulatta3').

A 10-ng amount of DNA was amplified in a 30 µL volume comprising 1× HF Phusion buffer, 0.6 U Phusion Hot Start High Fidelity DNA Polymerase (both New England Biolabs, Hitchen, Hertfordshire, UK), 200 µM of each dNTP and 0.3 µM each oligonucleotide. Reactions were heated to 95°C for 4 min followed by 35 cycles comprising 95°C for 45 s, 54°C for 40 s and 70°C for 2 min with a final single incubation at 70°C for 8 min.

Sequencing reactions were performed with six overlapping sequencing primer pairs using BigDye v3.1 sequencing chemistry and a 3130xl genetic analyser (Applied Biosystems, Paisley, UK). Contiguous sequences were generated using Ridom TraceEditPro (Ridom GmbH, Münster, Germany) and analysed using MEGA 5.2.2 and pairwise deletion [38].

Phylogenetic analyses

Sequences were aligned with default settings of ClustalW [21] in MEGA 5.2.2 [37]. The best model to approximate sequence evolution was determined with jModelTest 2.0 implemented in Phylemon 2 [27, 31]. According to the Bayesian information criterion (BIC) and the dynamical likelihood ratio test (dLRT), the Hasegawa–Kishino–Yano (HKY) model with gamma distribution and a proportion of invariable sites was selected [15]. The recommendation according the corrected Akaike information criterion (cAIC) was Kimura 81 model (K81) [18], with a proportion of invariable sites and gamma distribution. HKY was selected for the analysis for two reasons. First, K81 has just one additional parameter to HKY and the cAIC scores of both model scores were almost identical (16409.2 and 16409.6), and second, MEGA 5.2.2 model selection also supported HKY.

The Bayesian inference programme MrBayes 3.2.0 (http://mrbayes.sourceforge.net/ index.php) was used to generate the phylogenetic tree using Markov chain Monte Carlo coupling (MCMC; Ronquist, 2012). A total of 500,000 generations were run with a sample frequency of 50 until the convergence diagnostic of the two simultaneous runs approached 0.01. The first 25% of trees and parameters were discarded in the burn-in

phase. The α-parameter of the gamma distribution and the exact proportion of invariable sites were estimated in MrBayes as suggested by the programmers and were set during the run to α = 0.62 and proportion of invariable sites = 0.25, resulting in 1502 trees producing the clade credibility values. The node age of the *M. mulatta* and *M. fascicularis* was calibrated to 1.6 million years before present (Myr BP) [28]; the node of the split of the genera *Chlorocebus* and *Macaca* was estimated between 7.52 and 11.57 Myr BP [30]; *Pan troglodytes* was used as an outgroup, and a traditional mutation rate of 0.01 mutations per million years was used. The tree was visualised with the free available programme, FigTree v1.3.1 [29]. Estimates of genetic diversity and population structure were conducted for the current data set alone and alongside previously reported HVI and HVII region data [5, 34] using ARLEQUIN v.3.1 [10]. Statistical tests were applied to evaluate: nucleotide diversity (π), a measure of genetic differentiation, through calculation of the degree of polymorphisms by the number of nucleotide differences per site between sequences within a group or population [25]; and gene diversity (*h*), through assessment of the probability that two randomly chosen haplotypes are different in the sample [25] and neutrality, and to test whether the frequency distribution of polymorphic sites deviated from the neutral equilibrium expectation, using Tajima's D [37] and Fu's F_s [11] tests. The frequencies of polymorphic sites and of haplotypes within and between populations were also calculated. Using the same sequences included in the phylogenetic tree, relationships between the haplotypes of the 45 macaques in this study were evaluated by constructing a median-joining network based on full-length D-loop sequences (Fig. 2) using NETWORK v. 4.5.1.0 [1].

Results

Haplotype analysis of full mtDNA D-loop sequences

Sequence and haplotype analyses were performed on DNA from 30 Mauritian-derived cynomolgus macaques and 15 Indonesian-derived cynomolgus macaques on the full D-loop region including the HVI and HVII and a previously unreported polymorphic sequence between the hypervariable regions. The chance of amplifying numts was minimised as detailed in the Methods section. Sequencing was bidirectional. Heteroplasmy was not observed in any of the 45 macaques.

The Mauritian cynomolgus macaque D-loops displayed little sequence diversity with just four polymorphic sites identified revealing five haplotypes, of which haplotype MCM 1 was the most frequent (Table 1). While HVI and HVII elements of these haplotypes are

represented in previously reported sequences [5, 34], five previously unreported full-length sequences have been identified. By contrast, the D-loop sequences derived from the 15 Indonesian cynomolgus macaques displayed greater diversity with 175 polymorphic sites, comprising 13 indels and 170 substitutions yielding nine haplotypes, ICM1–ICM9 (Table 2). This group consisted of four pedigrees and five unrelated ICM. Eight of these nine haplotypes are novel. Haplotypes were peculiar to related animals and the same haplotype was not observed in unrelated macaques. There was no overt frequency bias towards any one haplotype. No haplotypes were shared between Mauritian and Indonesian cynomolgus macaques. All nucleotide sequences determined in this study were deposited in DDBJ/EMBL/GenBank with Accession Numbers JN201928–JN201942.

Comparative analysis of mtDNA D-loop hypervariable region sequences

All unique sequences from the HVI sequences identified in the 45 macaques were aligned with HVI sequences previously reported for cynomolgus macaques of different geographical origin, comprising 68 Mauritian and 70 Indonesian cynomolgus macaques as well as macaques of other geographical origins [34], and with two reference sequences for *M. fascicularis* and *M. sylvanus*, and three for *M. mulatta*. In the cohort of 68 Mauritian cynomolgus macaques, eight haplotypes (termed HVI1–HVI8 for this analysis) were defined from 14 polymorphic sites. This diversity is broadly comparable with that of the 30 Mauritian cynomolgus macaques studied here in which three haplotypes were defined from two polymorphic sites including the previously reported HVI1 and HVI8. A third haplotype, termed HVI9, was identified in a single animal and has not been reported previously. Haplotype HVI1 was present at the highest frequency being shared by 93% of the 30 Mauritian cynomolgus macaques analysed and 84% of the cohort of 68 Mauritian cynomolgus macaques [34]. Greater diversity was seen in the Indonesian macaques. A total of 56 haplotypes were defined in the cohort of 70 Indonesian macaques based on 156 polymorphic sites. From analysis of the 15 Indonesian macaques presented here, nine HVI haplotypes were defined from 110 polymorphic sites comprising single nucleotide changes and indels. This group consisted of four pedigrees and five unrelated Indonesian macaques. Eight of these nine

Table 1 Polymorphic mitochondrial DNA sites defining haplotypes in Mauritian cynomolgus macaques for HVI, HVII and the entire D-loop

| | | | | | | | | Polymorphic sites | | | | | | | | | | | | | | | |
| | | D-Loop | | HVI | | HVII | | HVI | | | | | | | | | | | | | | HVII | |
Reference	Sequence	H	F	H	F	H	F	16085	16197	16201	16218	16221	16235	16244	16248	16256	16258	16263	16282	16290	16320	183	314
Smith et al 2007	DQ373530			HVI 1	0.853			T	A	G	T	T	G	T	C	T	G	C	C	T	T		
	DQ373548			HVI 2	0.015			C	.	.	.	.	.	.	.	.	.	.	.	.	.		
	DQ373581			HVI 3	0.029			.	G	.	C	C	A	C	.	C	.	T	T	.	.	N/A	
	DQ373572			HVI 4	0.015			.	.	.	.	.	.	.	T	.	.	.	.	.	.		
	DQ373578			HVI 5	0.015			.	.	.	.	.	.	.	.	.	.	.	.	C	.		
	DQ373582			HVI 6	0.029			.	.	.	.	.	.	.	.	.	.	.	.	.	C		
	DQ373567			HVI 7	0.029			.	.	A	.	.	.	.	.	.	.	.	.	.	.		
	DQ373557			HVI 8	0.015			.	.	.	.	.	.	.	.	.	.	A	.	.	.		
Blancher et al 2008	AB261949					HVII 1	0.987						N/A									T	-
	AB261934					HVII 2	0.010															C	-
Kawamoto et al 2008	AB281366					HVII 1	0.900						N/A									T	-
	AB281367					HVII 3	0.095															T	C
Study data	I315C	MCM 1	0.800	HVI 1	0.933	HVII 1	0.800	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	-
	M895B	MCM 2	0.033	HVI 1	0.933	HVII 2	0.030	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	-
	I007B	MCM 3	0.100	HVI 1	0.933	HVII 3	0.098	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	C
	M066C	MCM 4	0.033	HVI 9	0.033	HVII 1	0.800	.	.	.	.	.	.	.	.	.	.	.	T	.	.	T	-
	I017B	MCM 5	0.033	HVI 8	0.033	HVII 1	0.800	.	.	.	.	.	.	.	.	.	.	A	.	.	.	T	-

Sequences submitted into GenBank from Smith et al. [34] for HVI have been assigned haplotypes (H) HVI 1 to HVI 8. Sequences submitted from Blancher et al. [5] and Kawamoto et al. [17] for HVII have been assigned haplotypes HVII 1 to HVII 3. Frequencies (F) at which these haplotypes occur have been calculated. Haplotypes for the corresponding region from the study MCM have been assigned MCM 1 to MCM 5. Novel sequence HVI 9 has been deposited to the DDBJ/EMBL/GenBank under Accession Number JN201942. Polymorphic sites of HVI and HVII are described.

Table 2 Polymorphic mitochondrial DNA sites that define haplotypes in Mauritian and Indonesian cynomolgus macaques

Polymorphic sites for the entire D-loop of study ICM. Polymorphic site 7 is detected within region nt 1 – 30 which has not been reported previously. Indonesian haplotypes are assigned ICM 1 to 9. Polymorphic sites and indels for MCM are highlighted in grey.

haplotypes have not been previously reported. There were no shared haplotypes between the Indonesian and the Mauritian macaques. The HVI region for Mauritian macaques was highly conserved compared with macaques reported from a range of origins as reported by Smith and McDonough [33].

The HVII region was similarly analysed. The 522-bp HVII region of all 45 macaque sequences was analysed in concert with 47 sequences from Mauritian, Indonesian and Filipino cynomolgus macaques (FCM) [5, 17], and with reference sequences for *M. fascicularis*, *M. sylvanus* and *M. mulatta*. Within the 30 Mauritian macaques studied, only one previously reported polymorphic site, T183C [34], was identified. Three Mauritian macaques had a 313_314insC in a poly C region revealing the presence of the short (7 C) and long (8 C) haplotypes in this cohort [17]. The region from nucleotide 1 to 30 has not been previously reported. Within this region, position 7 was found to be more polymorphic than the remaining sites in this area (Table 2): a G7A transition was identified in four related macaques. The HVII region for Mauritian macaques was noted to be highly conserved in comparison with macaques reported from various origins. As with HVI, greater sequence diversity was seen in the Indonesian macaques. A total of 68 polymorphic sites were identified in the 15 Indonesian macaques presented here, yielding seven haplotypes. When combined with the 87 polymorphic sites from the 137 Indonesian macaques' sequences (yielding 35 haplotypes), a total of 93 polymorphic sites were identified contributing to the definition of 42 haplotypes. The seven haplotypes identified in the cohort of 15 Indonesian macaques have not been previously reported. There was no sharing of HVII haplotypes between the populations analysed here. Across all cynomolgus macaques analysed here, the HVII region is more conserved than the HVI region.

Genetic diversity analysis

To further investigate the sequence diversity of the 45 Indonesian and Mauritian cynomolgus macaques, the genetic diversity indices for the two cohorts were calculated. Nucleotide diversity (π) in the Indonesian group was calculated for HVI (0.072), HVII (0.052) and the entire D-loop (0.059). The nucleotide diversity for the Mauritian macaque cohort was extremely low for both hypervariable regions [HVI (2.5×10^{-4}), HVII (3.5×10^{-4})] and the entire D-loop (1.8×10^{-4}). Genetic diversity (*h*) values for the Indonesian cohort were comparable for the hypervariable regions and the entire D-loop (D-loop 0.924, $P = 0.044$; HVI 0.933, $P = 0.045$; HVII 0.924, $P = 0.044$). Within the Mauritian macaque cohort, the genetic diversity for HVII

(0.333, $P = 0.273$) was higher than that for HVI (0.131). Two tests for neutrality were performed on the Mauritian sequences. For HVI, Fu's F_s test yielded a value of -2.355 ($P = 0.100$) and Tajima's D yielded a value of -1.507 ($P = 0.045$). For HVII, Fu's F_s test yielded a value of -0.574 ($P = 0.207$) and Tajima's D yielded a value of -0.448 ($P = 0.273$). For the D-loop analysis, Fu's F_s test yielded a value of -3.25 ($P = 0.003$) and Tajima's D yielded a value of -1.73 ($P = 0.011$). The full data set is available in Table 3. The average numbers of pairwise differences within and between the Indonesian and Mauritian macaque cohorts were calculated for the individual hypervariable regions and the entire D-loop (Table 4) with the lowest number observed across the full D-loop of the Mauritian macaque cohort (0.39). The individual hypervariable regions also displayed low levels of differences. The highest degree of pairwise differences was obtained from

analysis of full Indonesian macaque D-loop sequences (63.34).

Phylogenetic analysis and population structure

The Bayesian tree for the entire D-loop sequences for Indonesian and Mauritian cynomolgus macaque cohorts is shown in Fig. 1. Sequences for macaques *M. fascicularis*, *M. mulatta*, *M. thibetana* and *M. sylvanus* were included for reference. In line with other reports, five available Cercopithecine mtDNA sequences, in addition to the *Macaca* sequences, were included (*Papio papio*, *P. anubis*, *P. hamadryas*, *Chlorocebus aethiops* and *C. tantalus*), as was the *Pan troglodytes* sequence as an outlier. The Mauritian macaque-derived sequences displayed low genetic diversity with all sequences clustering into one clade (clade credibility value = 1). Haplotypes ICM 1, 2 and 3 are related to

Table 3 Genetic diversity indices for *M. fascicularis* populations

Diversity parameters	Population	mtDNA region		
		D-Loop	HVI	HVII
Nucleotide diversity $\Pi \times 1000$ ($\pm$SE)	MCM study	0.184 (0.261)	0.249 (0.424)	0.345 (0.517)
	ICM study	59.086 (30.303)	72.228 (37.313)	51.810 (26.700)
	Mauritius	n/a	1.313 (1.111)	0.052 (0.190)
	Indonesia	n/a	84.223 (40.943)	37.800 (18.630)
	Vietnam	n/a	42.312 (20.827)	17.100 (8.70)*
	Philipines	n/a	24.392 (12.284)	10.450 (5.820)
	Malaysia	n/a	71.44 (34.557)	n/a
Gene diversity h ($\pm$SE)	MCM study	0.359 (0.108)	0.131 (0.082)	0.333 (0.100)
	ICM study	0.924 (0.044)	0.933 (0.045)	0.924 (0.044)
	Mauritius	n/a	0.273 (0.071)	0.027 (0.026)
	Indonesia	n/a	0.994 (0.003)	0.924 (0.013)
	Vietnam	n/a	0.978 (0.006)	0.932 (0.016)*
	Philipines	n/a	0.822 (0.014)	0.693 (0.082)
	Malaysia	n/a	0.994 (0.002)	n/a
Neutrality tests				
Tajima's D (P-value)	MCM study	**−1.732 (0.011)**	−1.507 (0.045)	−0.448 (0.273)
	ICM study	0.761 (0.851)	0.466 (0.713)	1.173 (0.917)
	Mauritius	n/a	**−2.200 (0.003)**	−1.060 (0.119)
	Indonesia	n/a	1.087 (0.900)	0.630 (0.815)
	Vietnam	n/a	0.351 (0.722)	−0.330 (0.420)*
	Philipines	n/a	−1.462 (0.047)	0.818 (0.811)
	Malaysia	n/a	0.616 (0.815)	n/a
Fu's F_s (P-value)	MCM study	**−3.255 (0.003)**	**−2.355 (0.007)**	−0.574 (0.207)
	ICM study	8.788 (0.997)	4.436 (0.953)	4.430 (0.969)
	Mauritius	n/a	**−4.105 (0.018)**	−1.980 (0.030)
	Indonesia	n/a	−7.444 (0.053)	2.650 (0.782)
	Vietnam	n/a	−6.466 (0.081)	−3.140 (0.180)*
	Philipines	n/a	−4.202 (0.154)	3.660 (0.923)
	Malaysia	n/a	**−23.876 (0.002)**	n/a

Analyses were carried out for study samples and previously reported data [5, 34]. Fu's F_s and Tajima's D should be considered as significant at the 5% level if $P < 0.02$; significant values are highlighted in bold; n/a, data not available. Values incorporated from Blancher et al. [17] are highlighted by an asterisk as Indochinese sequences were not available for analysis.

Table 4 Average number of pairwise differences between and within *M. fascicularis* populations

	HVI	1	2	3	4	5	6	7	8	9	10
Reported data	1. Vietnam (50)	17.33	60.25	52.33	66.22	49.77	66.00	49.20	13.33	66.78	81.33
	2. Philippines (35)		28.83	35.13	35.91	46.40	35.58	50.00	56.75	65.42	83.75
	3. Malaysia (104)			48.00	39.67	46.09	39.17	49.47	48.75	66.08	85.00
	4. Mauritius (9)				3.50	55.81	2.30	65.76	62.56	74.70	86.11
	5. Indonesia (56)					43.11	55.61	41.63	47.36	66.82	80.46
Study data	6. Mauritius (3)						1.33	65.73	62.33	74.78	85.67
	7. Indonesia (9)							36.75	47.33	65.89	77.93
References	8. *M. fascicularis (1)*								0.00	65.67	79.00
	9. *M. mulatta (3)*									47.33	87.33
	10. *M. sylvanus (1)*										0.00

	HVII	1	2	3	4	5	6	7	8	9
Reported data	1. Philippines (6)	7.27	20.00	22.27	30.00	20.17	36.46	37.50	59.22	77.17
	2. Mauritius (2)		0.67	19.22	11.00	0.83	37.50	33.50	60.17	79.50
	3. Indonesia (35)			23.45	29.15	19.40	33.60	34.40	60.43	80.28
	4. Mauritus[1] (2)				1.00	10.83	47.10	42.50	69.83	87.50
Study data	5. Mauritius (3)					1.33	37.67	33.67	60.33	79.67
	6. Indonesia (9)						26.58	35.67	66.49	85.80
References	7. *M. fascicularis (1)*							0.00	55.33	74.00
	8. *M. mulatta (3)*								33.00	85.33
	9. *M. sylvanus (1)*									0.00

	D-Loop	1	2	3	4	5
Study data	1. Mauritian (5)	1.60	101.22	97.00	132.07	166.20
	2. Indonesian (9)		70.64	80.67	131.04	166.89
References	3. *M. fascicularis (1)*			0.00	120.67	156.00
	4. *M. mulatta (3)*				80.33	170.67
	5. *M. sylvanus (1)*					0.00

Average number of pairwise differences between sequences for both hypervariable regions and the entire D-loop. Differences are displayed within (on diagonal) and between (above diagonal) populations of cynomolgus macaques. Previously reported data are included for comparison. HVI table includes sequences submitted by Smith et al. [34] excluding cytochrome B region. HVII table includes sequences reported by Blancher et al. [5], Mauritian[1] group is submitted by Kawamoto et al. [17]. References include single sequences for *M. fascicularis* (NC_012670), *M. mulatta* (AY612638), and *M. sylvanus* (AJ309865).

the Mauritian macaque sequences. ICM1 (clade ICM A), derived from one animal (504F), clusters most closely with the Mauritian sequences, whereas ICM2 and ICM3 form a distinct but related clade (ICM B). Six of the Indonesian haplotypes (ICM4–ICM9) form a further clade (ICM C) which is distinct from the Mauritian cluster. With one exception, related animals displayed identical sequences. The phylogenetic analysis showed the *M. fascicularis* reference sequence to cluster with Indonesian sequences, while the more distantly related *M. mulatta*, *M. thibetana* and *M. sylvanus* sequences were placed separately. Interpopulational divergences were calculated for all unique sequences for 45 macaques. The tree constructed in Fig. 1 was used to add the divergence times assuming a constant rate of mutation. Calibration of the tree was carried out at the split between *M. mulatta* and *M. fascicularis*; the date selected was 1.6 million years before present (Myr BP),

consistent with that used by others [5, 39]. Divergence times for the separation event between the *M. fascicularis* groups were 0.44 Myr BP and for the separation of the Indonesian clades A and B, 0.84 Myr BP. The divergence time between ICM A and the Mauritian animals is calculated at 0.486 Myr BP. Relationships between the haplotypes of the 15 Indonesian and 30 Mauritian animals were evaluated by analysing the entire D-loop in a median-joining network constructed using NETWORK v. 4.5.10 (Fig. 2). The haplotypes clustered into two groups. Group 1 included six ICM haplotypes and the reference sequence for *M. fascicularis*. Group 2 included three Indonesian haplotypes, five Mauritian haplotypes and the reference sequences for *M. mulatta* and *M. sylvanus*. The circle size is scaled to the number of individuals presenting a particular haplotype, and the branches are scaled to the number of polymorphic sites between each haplotype.
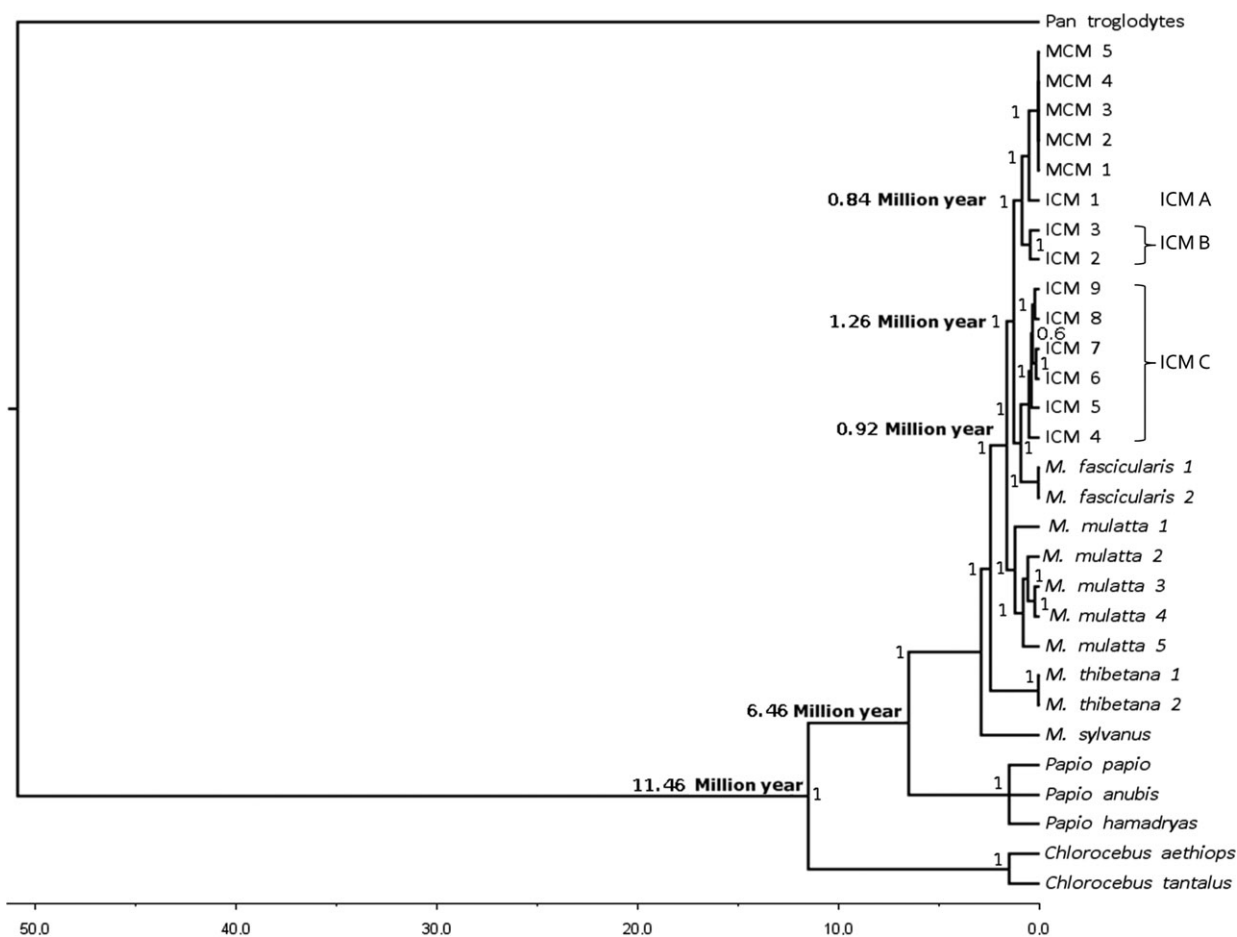
Fig. 1 Phylogenetic tree of the entire D-loop for MCM and ICM. Phylogenetic tree constructed using the Bayesian inference programme MrBayes 3.2.0. Analysis is of the mtDNA D-loop for *M. fascicularis* of Mauritian and Indonesian origin. 50,000 generations were run with a sample frequency of 50 until the convergence diagnostic of the two simultaneous runs approached 0.01. A total of 1502 trees were generated to produce the clade credibility values. The tree is divided into three clades: the Mauritian clade consists of all study Mauritian macaques, ICM A and ICM B comprise animals of insular origin, and ICM C is a cluster of continental animals. The node age of the *M. mulatta* and *M. fascicularis* was calibrated to 1.6 Myr BP and the node age of *Macaca* and *Chlorocebus* to 7.52–11.57 Myr BP; a traditional mutation rate of 0.01 mutations/Myr was implemented. Non-*Macaca* Cercopithecine sequences were included and *Pan troglodytes* sequence was used as an outgroup.

Discussion

We wished to study the genetic diversity within two cohorts of cynomolgus macaques of Mauritian and Indonesian origin. We analysed HVI and HVII sequences for both cohorts in isolation and in relation to data reported by others for cynomolgus macaques of known geographical origin. To increase the power of analysis, mtDNA analysis of the entire D-loop (encompassing the HVI and HVII regions) was also performed; to date, this approach has not been reported. We wished to minimise the chance of amplifying a numt. If this nuclear DNA, which has derived from the mitochondrial genome and become established as a pseudogene, is amplified in preference to the corresponding mito-

chondrial sequence, the data analysis will be confounded. We developed an approach by first assessing the likely length of a numt and then designing a PCR amplification which would yield an amplicon in excess of this size. Analysis of data published by Calabrese et al. [8] on 751 numt sequences identified in the rhesus macaque (*M. mulatta*) genome revealed that the average nuclear sequence spanned 697 bp (median 396 bp) and the average identity to mtDNA sequences was 77% (median 78%). The median identity to mtDNA sequences of the 19.3% of all numts which exceeded 1200 bp in length was 74.62%, thus attaining a percentage identity greater than this for a 1.2-kb fragment minimises the chance of it being a numt sequence. Thus, our approach of amplifying the entire D-loop gave us
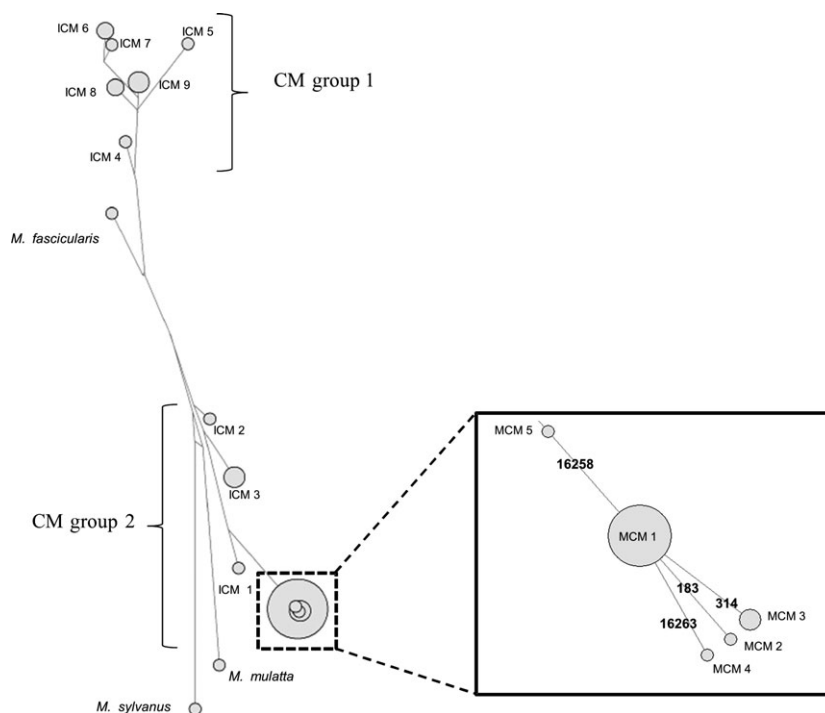
Fig. 2 Haplotype network based on the number of polymorphic sites among the haplotypes of cynomolgus macaques of Mauritian and Indonesian origin. Polymorphic sites for MCM are presented in bold. The circle size is scaled to the number of individuals exhibiting the particular haplotype, and the branches are scaled to the number of polymorphic sites between each haplotype. References are also included to represent outliers of the data set. The two differential groups of cynomolgus macaques, CM Group 1, continental and CM Group 2, insular, are indicated and are comparable to the Fas1 and Fas2 groups reported by Smith et al. [34].

confidence that we had minimised the chance of amplifying a numt.

When we focussed on HVI, we found broadly similar, low diversity in sequences derived from our cohort of 30 Mauritian cynomolgus macaques compared to the 68 analysed by Smith et al. [34]. Further, in addition to identification of two of eight previously characterised haplotypes in our cohort, we identified one previously unreported haplotype. The Indonesian-derived cynomolgus macaques were considerably more polymorphic in this region than the Mauritian macaques. From just 15 animals, we detected nine haplotypes of which eight were previously unreported, adding to the 56 currently in the literature [34]. There were no shared haplotypes between the Indonesian-derived and Mauritian macaques, which is supported by high haplotype diversity in ICM and very low haplotype diversity in the Mauritian animals. The HVI region for Mauritian cynomolgus macaques is highly conserved compared with macaques reported from other origins. Animals of Malaysian origin were found to have the highest number of polymorphic sites and haplotypes. Neutrality tests generated for animals of Malaysian origin for HVI revealed a significantly negative F_s value; however, the *D* value was not statistically significant. There were no shared haplotypes between the Indonesian and the Mauritian macaques; however, one Mauritian haplotype (HVI3) was shared between

animals of Malaysian and Filipino origin [35] suggesting a common ancestry between these populations, as reported previously [5, 7, 9, 17, 19, 22, 34, 40].

When we focussed on HVII, we noted this region for Mauritian macaques to be highly conserved when compared with those from Indonesia. The HVII region is also more conserved than the HVI region across all populations analysed. In addition to the detection of one previously reported polymorphic site in the Mauritian animals with comparably frequency [5], we identified an insertion in a poly C region in three macaques. Thus, we were able to define short (7 C) and long (8 C) haplotypes in these animals as previously reported [17]. This poly C region has been noted as difficult to analyse in human samples [3]. However, confirmation of this specific sequence in three macaques from two lineages was readily performed through sequence analysis of two independent amplicons in both the sense and antisense direction. The region between nucleotides 1 and 30 has not been reported before in this species. This region was identical in all 30 Mauritian macaques, but a G7A transition was found in four related Indonesian-derived macaques. Analysis of the 137 GenBank sequences from ICM revealed 87 polymorphic sites defining 35 haplotypes. By contrast, the 15 ICM displayed a higher diversity with 68 polymorphic sites. However, these defined just seven haplotypes. Thus, the overall

haplotype diversity between the two cohorts is comparable. All seven haplotypes were previously unreported and no haplotypes were shared between the populations analysed.

From just 45 macaques, we obtained previously unreported haplotype data for both HVI and HVII regions; thus, we sought to determine whether sequence and genetic diversity analyses of the full-length D-loop sequence could yield greater information which might contribute to a more detailed genetic characterisation of cynomolgus macaques. Analysis of the entire D-loop of the 30 Mauritian macaques revealed just five haplotypes (MCM1–MCM5) consistent with RFLP-based haplotyping in which two haplotypes were defined in a cohort of 19 Mauritian macaques [22]. Furthermore, through the analysis of the entire D-loop for Mauritian macaques, in combination with previously reported data on the separate HVI and HVII regions, we detected a variable region consisting of 16 polymorphic sites (mitochondrial DNA nucleotides 16085 to 314). The future analysis of this fragment of 800 bp would therefore provide more information than the analysis of the individual hypervariable regions. We included some groups of related animals within our data set which were valuable in verifying the presence of previously unreported alleles. Given the known low diversity in the Mauritian cynomolgus macaques, the potentially adverse effect of including related animals in the diversity analysis was not considered to be significant. This low genetic diversity and low nucleotide diversity values expected of the highly conserved genetics of Mauritian macaques were revealed and mismatch distribution analysis supported the occurrence of a genetic bottleneck. Bonhomme et al. [7] have shown that Mauritian macaques have a lower genetic diversity than other populations, as evidenced by analysis of autosomal and sex-linked loci; however, there is no clear and consistent evidence of a genetic bottleneck and the statistical analyses favour an expansion. However, the authors could not exclude the possibility that various cycles of expansion and contraction may have occurred. Our neutrality test data on HVI sequences suggested either neutral polymorphism or population expansion, and those on HVII sequences highlighted neutral polymorphism while analysis of the entire D-loop sequence supported an expansion in the population and thus may reflect the expansion and contraction scenario proposed. Within the Indonesian-derived macaque cohort, π was higher than for the Mauritian animals, despite the inclusion of pedigree groups. Genetic diversity values for Indonesian-derived macaques were comparable for the hypervariable regions and the entire D-loop. This contrasts with previous reports in which the nucleotide diversity was at least fivefold greater for HVI than for HVII [5], although an effect of including related animals on diversity of this region cannot be excluded. The 15 Indonesian-derived macaques as having a higher degree of differences than observed in sequences reported by others.

Phylogenetic analysis of the full D-loop sequence resulted in the clustering of Mauritian macaque sequences, confirming the limited genetic diversity in this population (Figs 1 and 2). A single Indonesian-derived macaque had a high degree of similarity with Mauritian macaques indicating a haplotype more closely related to Mauritian macaques than the remaining Indonesian-derived macaques which clustered into two related clades, reflecting insular and continental groups. The positioning of Mauritian macaques into the continental group by Y-chromosomal sequence analysis [39] is not supported by our data. The divergence time calculated for the separation of the Mauritian macaque haplotypes away from ICM A was 0.84 Myr BP. This value does not corroborate the suggested date of introduction of macaques to Mauritius but may be influenced by the low number of haplotypes identified, as has been found with calculations derived from other studies [5, 39], and the value may reflect the divergence of the subpopulation of Indonesian macaques from which the Mauritian macaques derived. It is unlikely that further bottlenecks have occurred as a result of captive breeding Mauritian macaques, because we and others have reported no difference between the haplotype diversity at the major histocompatibility complex locus between captive and feral Mauritian macaques [23, 41].

Hitherto, D-loop studies have reported the divergence of macaque populations based on HVI or, to a lesser extent, HVII sequence. Our data broadly agree with the definition of insular and continental groups of *M. fascicularis* using these approaches. Through analysis of entire D-loop sequences, however, we have identified a further polymorphic region and we have defined full haplotypes for this region. This approach has allowed us to illustrate readily the close relationship of ICM clade A with the Mauritian macaques and the divergence of ICM clade B away from these, thereby corroborating data from Smith et al. [34]. As with other studies, ICM clade C (Continental) is distinct from these groups. The use of a full D-loop haplotype, including the hereto unreported polymorphic region, may allow resolution of animal genetics in situations where existing HVI and HVII haplotypes are not linked. Thus, analysis of the entire mtDNA D-loop is recommended to define a haplotype and achieve a comprehensive understanding of macaque diversity against geographical origin.

References

1 Bandelt HJ, Forster P, Rohl A: Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol* 1999; **16**:37–48.

2 Baroncelli S, Negri DR, Michelini Z, Cara A: *Macaca mulatta, fascicularis* and *nemestrina* in AIDS vaccine development. *Expert Rev Vaccines* 2008; **7**:1419–34.

3 Bendall KE, Sykes BC: Length heteroplasmy in the first hypervariable segment of the human mtDNA control region. *Am J Hum Genet* 1995; **57**:248–56.

4 Bensasson D, Zhang D, Hartl DL, Hewitt GM: Mitochondrial pseudogenes: evolution's misplaced witnesses. *Trends Ecol Evol* 2001; **16**:314–21.

5 Blancher A, Bonhomme M, Crouau-Roy B, Terao K, Kitano T, Saitou N: Mitochondrial DNA sequence phylogeny of 4 populations of the widely distributed cynomolgus macaque (*Macaca fascicularis fascicularis*). *J Hered* 2008; **99**:254–64.

6 Bodewes R, Rimmelzwaan GF, Osterhaus AD: Animal models for the preclinical evaluation of candidate influenza vaccines. *Expert Rev Vaccines* 2010; **9**:59–72.

7 Bonhomme M, Blancher A, Cuartero S, Chikhi L, Crouau-Roy B: Origin and number of founders in an introduced insular primate: estimation from nuclear genetic data. *Mol Ecol* 2008; **17**:1009–19.

8 Calabrese FM, Simone D, Attimonelli M: Primates and mouse NumtS in the UCSC Genome Browser. *BMC Bioinformatics* 2012; **13**:S15.

9 Campbell KJ, Detmer AM, Karl JA, Wiseman RW, Blasky AJ, Hughes AL, Bimber BN, O'Connor SL, O'Connor DH: Characterization of 47 MHC class I sequences in Filipino cynomolgus macaques. *Immunogenetics* 2009; **61**:177–87.

10 Excoffier L, Laval G, Schneider S: Arlequin (version 3.0): an integrated software package for population genetics data analysis. *Evol Bioinform Online* 2005; **1**:47–50.

11 Fu YX: Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* 1997; **147**:915–25.

12 Gardner MB, Luciw P: Macaque models of human infectious disease. *ILAR J* 2008; **49**:220–55.

13 Gupta UD, Katoch VM: Animal models of tuberculosis for vaccine development. *Indian J Med Res* 2009; **129**:11–8.

14 Hale DA, Dhanireddy K, Bruno D, Kirk AD: Induction of transplantation tolerance in non-human primate preclinical models. *Philos Trans R Soc Lond B Biol Sci* 2005; **360**:1723–37.

15 Hasegawa M, Kishino H, Yano T: Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol* 1985; **22**:160–74.

16 Kanthaswamy S, Smith DG: Effects of geographic origin on captive *Macaca mulatta* mitochondrial DNA variation. *Comp Med* 2004; **54**:193–201.

17 Kawamoto Y, Kawamoto S, Matsubayashi K, Nozawa K, Watanabe T, Stanley MA, Perwitasari-Farajallah D: Genetic diversity of longtail macaques (*Macaca fascicularis*) on the island of Mauritius: an assessment of nuclear and mitochondrial DNA polymorphisms. *J Med Primatol* 2008; **37**:45–54.

18 Kimura M: Estimation of evolutionary distances between homologous nucleotide sequences. *Proc Natl Acad Sci USA* 1981; **78**:454–8.

19 Kondo M, Kawamoto Y, Nozawa K, Matsubayashi K, Watanabe T, Griffiths O, Stanley M-A: Population genetics of crab-eating macaques (*Macaca fascicularis*) on the island of Mauritius. *Am J Primatol* 1993; **29**:167–82.

20 Krebs KC, Jin Z, Rudersdorf R, Hughes AL, O'Connor DH: Unusually high frequency MHC class I alleles in Mauritian origin cynomolgus macaques. *J Immunol* 2005; **175**:5230–9.

21 Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG: Clustal W and Clustal X version 2.0. *Bioinformatics* 2007; **23**:2947–8.

22 Lawler SH, Sussman RW, Taylor LL: Mitochondrial DNA of the Mauritian macaques (*Macaca fascicularis*): an example of the founder effect. *Am J Phys Anthropol* 1995; **96**:133–41.

23 Mee ET, Badhan A, Karl JA, Wiseman RW, Cutler K, Knapp LA, Almond N, O'Connor DH, Rose NJ: 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* 2009; **38**:1–14.

24 Mitchell JL, Mee ET, Almond NM, Cutler K, Rose NJ: Characterisation of MHC haplotypes in a breeding colony of Indonesian cynomolgus macaques reveals a high level of diversity. *Immunogenetics* 2012; **64**:123–9.

25 Nei M: Molecular Evolutionary Genetics. New York: Columbia University Press, 1987.

26 O'Connor SL, Blasky AJ, Pendley CJ, Becker EA, Wiseman RW, Karl JA, Hughes AL, O'Connor DH: Comprehensive characterization of MHC class II haplotypes in Mauritian cynomolgus macaques. *Immunogenetics* 2007; **59**:449–62.

27 Posada D: jModelTest: phylogenetic model averaging. *Mol Biol Evol* 2008; **25**:1253–6.

28 Purvis A: A composite estimate of primate phylogeny. *Philos Trans R Soc Lond B Biol Sci* 1995; **348**:405–21.

29 Rambaut A: FigTree. Institute of Evolutionary Biology, University of Edinburgh, 2009; http://tree.bio.ed.ac.uk/software/figtree/.

30 Roos C, Zinner D, Kubatko LS, Schwarz C, Yang M, Meyer D, Nash SD, Xing J, Batzer MA, Brameier M, Leendertz FH, Ziegler T, Perwitasari-Farajallah D, Nadler T, Walter L, Osterholz M: Nuclear versus mitochondrial DNA: evidence for hybridization in colobine monkeys. *BMC Evol Biol* 2011; **11**:77.

31 Sanchez R, Serra F, Tarraga J, Medina I, Carbonell J, Pulido L, de Maria A, Capella-Gutierrez S, Huerta-Cepas J, Gabaldon T, Dopazo J, Dopazo H: Phylemon 2.0: a suite of web-tools for molecular evolution, phylogenetics, phylogenomics and hypotheses testing. *Nucl Acids Res* 2011; **39**:W470–4.

32 Shiina T, Tanaka K, Katsuyama Y, Otabe K, Sakamoto K, Kurata M, Nomura M, Yamanaka H, Nakagawa H, Inoko H, Ota M: Mitochondrial DNA diversity among three subpopulations of cynomolgus macaques (*Macaca fascicularis*) originating from the Indochinese region. *Exp Anim* 2010; **59**:567–78.

33 Smith DG, McDonough J: Mitochondrial DNA variation in Chinese and Indian rhesus macaques (*Macaca mulatta*). *Am J Primatol* 2005; **65**:1–25.

34 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**:182–98.

35 Stevison LS, Kohn MH: Determining genetic background in captive stocks of cynomolgus macaques (*Macaca fascicularis*). *J Med Primatol* 2008; **37**:311–7.

36 Sussman R, Tattersall I: Distribution, abundance, and putative ecological strategy of *Macaca fascicularis* on the island of Mauritius, Southwestern Indian Ocean. *Folia Primatol* 1986; **46**:28–43.

37 Tajima F: Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 1989; **123**:585–95.

38 Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S: MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* 2011; **28**:2731–9.

39 Tosi AJ, Coke CS: Comparative phylogenetics offer new insights into the biogeographic history of *Macaca fascicularis* and the origin of the Mauritian macaques. *Mol Phlyogenet Evol* 2007; **42**:498–504.

40 Tosi AJ, Morales JC, Melnick DJ: Y-chromosome and mitochondrial markers in *Macaca fascicularis* indicate introgression with Indochinese M. mulatta, and a biogeographic barrier in the Isthmus of Kra. *Int J Primatol* 2002; **23**:161–78.

41 Wiseman RW, Wojcechowskyj JA, Greene JM, Blasky AJ, Gopon T, Soma T, Friedrich TC, O'Connor SL, O'Connor DH: Simian immunodeficiency virus SIVmac239 infection of major histocompatibility complex-identical cynomolgus macaques from Mauritius. *J Virol* 2007; **81**:349–61.