Phylogenomics reveals incomplete lineage sorting and ancient hybrid drove the radiation of macaques

Xinxin Tan[1,2#], Jiwei Qi[1#], Zhijin Liu[3#], Pengfei Fan[4], Gaoming Liu[1], Liye Zhang[5], Ying Shen[1], Jing Li[6], Christian Roos[5,7*], Xuming Zhou[1*], Ming Li[1*]

[1]CAS Key Laboratory of Animal Ecology and Conservation Biology, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China

[2] Geneplus-Beijing Institute, Beijing 102206, China

[3] College of Life Sciences, Capital Normal University, Beijing 100049, China

[4] School of Life Sciences, Sun Yat-Sen University, Guangzhou 510275, China

[5] Primate Genetics Laboratory, German Primate Center, Leibniz Institute for Primate Research, Kellnerweg 4, 37077, Göttingen, Germany

[6]Key Laboratory of Bio-resources and Eco-environment of Ministry of Education, College of Life Sciences, Sichuan University, Chengdu 610064, China

[7]Gene Bank of Primates, German Primate Center, Leibniz Institute for Primate Research, Kellnerweg 4, 37077, Göttingen, Germany

# Contributed equally to this work.

*Correspondence and requests for materials should be addressed to M.L. (lim@ioz.ac.cn), X.Z. (E-mail: zhouxuming@ioz.ac.cn) and C.R. (croos@dpz.eu).

Running heads: ILS and ancient hybrid in *Macaca*

Abstract

The genus *Macaca*, with 23 species assigned into four to seven species groups, exhibits the largest geographic range and represents the most successful adaptive radiation of nonhuman primates. Here, we conducted phylogenomic analyses of 16 newly generated and eight published macaque genomes and found a strong support for the division of this genus into seven species groups. Both

ancient hybrid and incomplete lineage sorting (ILS) have contributed to the radiation and evolution of macaques. Particularly, the contradicting phylogenetic positions among *silenus/nigra*, *fascicularis/mulatta* and *arctoides*/*sinica* lineages were likely resulted from high level of ILS and potential hybridization between the ancestors of the *arctoides*/*sinica* and *silenus/nigra* lineages. Furthermore, an integrated scenario for macaque radiation is reconstructed by the help of the dated phylogenetic tree combined with documented history records. This study provides insights into ancient introgression involved in the radiation of macaques, which may help us to understand the rapid speciation of nonhuman primates.

Keywords: *Macaca*, phylogenomics, incomplete lineage sorting, hybridization.

Introduction

Introgression and hybridization are important driving forces in speciation and evolution, and over the past decade increasing attention has been paid to their role in evolutionary diversification [1, 2, 3, 4, 5, 6, 7]. Hybridization has been recognized as one of the most interesting topics in evolutionary biology, and the genetic introgression, caused by hybridization, is not always maladaptive and could be a potent evolutionary force [3, 8, 9, 10]. With the development of high-throughput sequencing technology and new computational tools, phylogenomics emerged and allows researchers to explicitly account for hybridization when constructing species trees [11, 12]. However, incomplete lineage sorting (ILS), an alternative explanation for contradicting gene trees, needs to be considered as well when reconstructing a species tree [13]. ILS occurs mainly during rapid speciation events, when new lineages descend within a short time period from ancestors, particularly when these have large effective population sizes (*Ne*) [14]. As a result, ILS will cause ancestral polymorphisms to persist even when descendant lineages diverged [15]. Therefore, both ILS and gene flow or hybridization can lead to topological discordance between gene trees and species trees [16, 17]. For example, Edelman et al. [18] used 20 *de novo* genome assemblies to explore the speciation history and gene flow in rapidly radiating *Heliconius* butterflies and distinguish ILS from introgression. They discovered that introgressed loci are underrepresented in low-recombination and gene-rich regions [18]. For primates, information about a genome-wide view on hybridization is still limited but genome sequencing efforts are on the way [19]. For example, Rogers, et al. [20] confirmed reticulation in baboons and documented multiple episodes of admixture and introgression during the radiation of *Papio* using a comparative genomic approach.

The genus *Macaca* (Primates: Cercopithecidae) exhibits the largest geographic range and represents the most successful adaptive radiation of nonhuman primates [21] (**Fig. 1A**). The 23 currently recognized macaque species are subdivided into four to seven species groups according to similarities in morphology, ecology, behavior, distribution and genetics [21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33] (**supplementary Fig. S1**). However, the number and composition of the species groups is disputed. Before the whole-genome age, phylogenetic analyses were mainly based on a few mitochondrial and/or nuclear markers resulting in contradicting phylogenetic

relationships [26, 34, 35, 36], which are partially caused by introgression and hybridization events among various macaque species.

Introgression and hybridization between macaque species are common, for instance, among Sulawesi macaques, between *M. mulatta* and *M. fascicularis*, and between *M. mulatta* and *M. thibetana* [37, 38, 39, 40, 41, 42, 43, 44, 45]. Moreover, for *M. arctoides* a hybrid origin was suggested as different molecular markers place the species into different clades [33, 34, 35, 36, 46]. However, most analyses have focused on gene flow within macaque species groups, or between species of the *fascicularis*/*mulatta* and *arctoides*/*sinica* species groups. Few studies included samples from all species groups when identifying genus-wide hybridization, and few attempted to detect ancient hybridization events in macaques [36, 47]. A recent study reported the history of the genus *Macaca* using 14 individuals from nine species and detected extensive gene flow signals, with the strongest signals between the *fascicularis/mulatta* and *silenus* groups [48]. Introgression signals among island species, such as Sulawesi macaques and other species, were also observed. This put forward a variety of possible reasons for this phenomenon including the genomic similarity of closely related species or ancestral introgression. Therefore, whether hybridization occurred among all macaques, especially among species that are geographically distant still needs to be investigated. Furthermore, it remains unclear how much ancient hybridization or ILS have contributed to the diversity of the macaque genome.

In this study, we aimed to trace the complex, likely reticulated evolutionary history of macaques and newly sequenced whole genomes of 16 individuals representing 13 Asian (*M. mulatta*, *M. cyclopis*, *M. fuscata*, *M. arctoides*, *M. assamensis*, *M. leucogenys*, *M. sinica*, *M. leonina*, *M. silenus*, *M. nemestrina*, *M. maura*, *M. tonkeana*, *M. nigra*) and the only African macaque species (*M. sylvanus*). The sequences were mapped to the latest *Papio anubis* reference genome (Panu_3.0) to obtain single nucleotide variants (SNVs). Combined with eight published macaque genomes, we performed a comprehensive genomics analysis among 24 macaque individuals, representing 16 species and all species groups, aiming to reconstruct the phylogeny of the genus *Macaca*, investigate ILS and hybridization among macaque species, and trace the evolutionary history and rapid radiation of the genus *Macaca*.

Results

Genome Sequencing and Variant Discovery

Whole genome sequencing was performed with the Illumina HiSeq 2500 and 4000 platforms. In total, more than 1,364 Gbp clean data (quality-controlled reads) were obtained and the statistics results of the clean data for each species are shown in **supplementary table S1**. Published genome data of eight macaques (*M. mulatta* (Chinese rhesus macaque, CR), *M. mulatta* (Indian rhesus macaque, IR), *M. fascicularis*, *M. arctoides*, *M. assamensis*, *M. thibetana*, *M. nemestrina* and *M. tonkeana*) were downloaded from NCBI (**supplementary table S1**) and two of them overlapped with [48]. Clean data for each species were mapped to the *Papio anubis* reference genome (Panu_3.0) and yielded genome-wide coverage from 16.85 to 42.34× with an average effective depth at 28.24× (**supplementary table S1**). We picked a more distant reference genome (*Papio anubis*) rather than the more recently evolved macaque (*Macaca mulatta*) to avoid biases that would arise from mapping to an in-group reference genome. Even though this decision likely led to some degree of lost data (in regions that lacked ortholog to *Papio anubis*), however, the mapping coverage (4×) for macaques range from 93.55% to 97.59%, which is very similar to mapping coverage (4×) for *Papio* (98.75%), suggesting a potential bias to be limited. The mapping rate for *M. nemestrina* (2) is lower than 90%, which could be caused by low DNA quality.

The number of high-quality SNVs for each individual ranged from 15,943,244 to 21,778,812, counting both homozygous and heterozygous alternatives (**supplementary table S2**). Genome-wide heterozygosity varied considerably among macaques (**supplementary Fig. S2** and **supplementary table S2**) which may partially reflect different demographic histories. Based on gene annotations from the reference genome, the exonic SNVs in macaques ranged from 225,336 to 257,712 (**supplementary tables S3** and **S4**). The species-specific SNVs for each macaque ranged from 1,279,649 to 8,688,344 (**supplementary table S2**).

Phylogeny and divergence among macaques

In order to reconstruct the phylogeny of macaques, we generated consensus sequences of all genomes for each macaque individual. We split the aligned chromosomes into 6,351 non-overlapping genome fragments (GFs), each 200-kilo basepair (kbp) long, which representing 46%

of the genome sequence. The multispecies coalescent (MSC) species tree based on the 6,351 GFs was supported with posterior probabilities (PPs) of 1.0 for all branches (**Fig. 1B** and **supplementary Fig. S3**). This species tree, consistent with neighbor joining (NJ) tree and maximum-likelihood (ML) tree based on autosomal SNVs (**supplementary Fig. S4** and **supplementary Fig. S5**), supported the segregation of macaques into seven well-supported clades or lineages that correspond to the classification of the genus into seven species groups [21, 29]. Among macaques, the African *sylvanus* lineage diverged first, followed by an initial separation of extant Asian macaques into a clade with the *silenus* and *nigra* lineages and a clade including all other Asian lineages. The latter segregated into a clade including the *arctoides* and *sinica* lineages and another clade with the *fascicularis* and *mulatta* lineages (**Fig. 1B** and **supplementary Figs. S3-S5**). The ML tree based on mitochondrial genomes (mitogenomes), consistent with previous results [32, 33], showed that among Asian macaques, the *silenus/nigra* clade separated first, followed by the *sinica* lineage, while among the remaining lineages, *fascicularis* split first and *mulatta* and *arctoides* afterwards (**supplementary Fig. S6** and **supplementary table S5**).

The quartet score 'q', which is the support value for possible phylogenetic arrangements, identified a conflict in resolving the branch leading to the ancestor of the *fascicularis/mulatta* and *arctoides/sinica* groups (**Fig. 1C-D**, branch no. 2 in **Fig. 1B** and **supplementary Fig. S3**). In addition, the topology cluster found that these two alternative topologies are most common among all 6,351 GF trees and represented by the colors red and blue, respectively in **Fig. 3A-B**. However, the reason for this conflict is unclear.

A consensus network analysis [40] of the GF trees recovered a network with bifurcating branches that emerged from a central cycle in the center of the network that indicates conflicting signals for the position of the *silenus/nigra* clade among Asian macaques (**Fig. 1E** and **supplementary Fig. S7**). At a threshold for conflicting edges of 30%, the position of the *silenus/nigra*, *arctoides/sinica* and *fascicularis/mulatta* clades were unascertainable (**supplementary Fig. S7**). When the threshold for conflicting edges was reduced, the phylogenetic signal became even more complex, indicating additional phylogenetic conflict and extensive gene flow in the data set (**supplementary Fig. S7**). F_{ST} analyses suggested significant differences within (medians for F_{ST} are 0.199) and between (medians for F_{ST} are 0.467) different macaque

groups ($P < 0.0001$) (**supplementary table S6** and **Fig. 1F**). A clear division of macaques into different clusters was also obtained by a principal component analysis (PCA) (**supplementary Fig. S8** and **supplementary table S7**).

The genus *Macaca* was estimated to have originated ca. 7.76 million years ago (Mya) during the Late Miocene using MCMCTREE in PAML v4.9 [49] (**Fig. 2** and **supplementary Fig. S9**). We revealed a divergence time for the split of *M. sylvanus* from the Asian macaques at 5.29 Mya (95% highest posterior density, HPD, 3.65–6.70), close to the Messinian Salinity Crisis [50]. In Asia, divergence of species groups occurred in the Pliocene, starting with the split between the *silenus/nigra* clade and the other four species groups 4.01 Mya (95% HPD 5.59–2.61), followed shortly afterwards by the separation of the ancestors of the *fascicularis/mulatta* and *arctoides/sinica* clades (3.79 Mya, 95% HPD 5.28–2.50). These three clades finally diverged into species groups 3.20 Mya (95% HPD 4.79–2.07; *silenus* and *nigra* groups), 3.01 Mya (95% HPD 4.52–1.84; *sinica* and *arctoides* groups) and 2.94 Mya (95% HPD 4.25–1.52; *mulatta* and *fascicularis* groups). Speciation within species groups took place in the Pleistocene, 2.56–1.39 Mya.

Contributions of ILS to Gene Tree Heterogeneity

Our study was aimed at investigating the extent to which levels of gene tree heterogeneity found in macaques can be attributed to incomplete lineage sorting (ILS), especially for the conflict among the *fascicularis/mulatta*, *arctoides/sinica* and *silenus/nigra* clades showed in **Fig, 1C** and **Fig. 3A-B.** To differentiate between ILS and introgression, which are both potential causes of phylogenetic discordance, we utilized QuIBL, a recently developed tree-based method introduced by Edelman et al. [18]. QuIBL estimates internal branch length distribution in discordant topologies for triplets of species and calculates the likelihood that this distribution is consistent with either ILS and introgression or ILS alone. Since this method is sensitive to recombination [18], we generated 3,175 trees based on small, non-overlapping windows (20 kb) separated by 400-kb steps. Our study revealed that within the *fascicularis/mulatta*, *silenus/nigra*, and *sinica/arctoides* clades, the Bayesian information criterion (BIC) test determined that 60.8% (73 out of 120 triplets) of triplets with internal branches of species exhibited phylogenetic discordances solely due to incomplete lineage sorting (ILS) (ΔBIC>10). Furthermore, only 9.1%

(11 out of 120 triplets) of our tested triplets showed strong evidence of introgression (ΔBIC<-10) (**supplementary table S8**). For instance, using QuIBL on the triplet *M. leonina_M. mulatta_M. thibetana*, we inferred that only 1.33% of loci across the entire genome are introgressed. Moreover, we discovered that a mere 0.29% of genetic loci supported discordant topologies and were introgressed, indicating that interspecific introgression was limited among the species examined (**Fig. 3F, supplementary table S8).** Consequently, QuIBL analysis suggests that incomplete lineage sorting, rather than post-speciation introgression, is the predominant factor underlying the phylogenetic discordance among these species.

Contributions of hybrid to Gene Tree Heterogeneity

Previous studies suggested a vital role of hybridization in shaping the evolutionary history of macaques [35, 36, 37, 38, 41, 42, 43, 44, 45, 46, 51]. In this study, we set out to infer footprints of ancient introgression across macaques. To this end, we first applied the approach based on D-statistics [52] implemented in *Dsuite* [53] to detect introgression and hybridization (**Fig. 3A-B, supplementary table S9)**. This approach estimates D and f4 statistics for all possible combinations of trios in macaques based on two alternative topologies described in **Fig. 3A-B,** respectively, and then performs an f-branch test to assign gene flow to specific internal branches (**Fig. 3C-D)**. The f-branch test based on the species tree suggested an event of introgression between the ancestor of *fascicularis/mulatta* and the ancestor of *silenus/nigra* lineages **(Fig. 3C)**. While the f-branch test based on another topology suggested an event of introgression between the ancestor of *fascicularis/mulatta* and the ancestor of *arctoides/sinica* lineages **(Fig. 3D)**. The above results suggest that the genetic background of *fascicularis/mulatta* lineage is highly heterogeneous and the ancestor of *fascicularis/mulatta* lineage may be a hybrid lineage. In addition, ancient hybrid speciation was also be studied by topology-based ML analysis using PhyloNet [54], which identifies a reticulate node in the scenarios among the ancestor of *fascicularis/mulatta, arctoides/sinica* and *silenus/nigra* lineages (**supplementary Fig. S10**).

Next, we corroborated the evidence for hybridization of the ancestor of *fascicularis/mulatta* in HyDe [55], which can examine genome-scale data for a large number of taxa and detect the population that may have arisen through hybrid speciation, as well as its putative parental populations, by estimates the inheritance parameter (γ value) to quantify the genomic

contributions of the parents to the hybrid. In this study, HyDe was used to verify for significant *p*-values in all combinations of the four taxa, which included one outgroup (*P. anubis*) and a triplet of in-group taxa. However only one of the four significant hybridization events showed a γ value as 0.55 (Z–score =306.84; $P = 0$), which indicated that the *fascicularis/mulatta* lineage is a hybrid speciation and the genomic contributions of its parents (*arctoides/sinica* and *silenus/nigra* lineages) is about 55% vs 45% (**supplementary table S11**). We further verified the hybridization for *fascicularis/mulatta* lineage at the individual level and found that all 5 samples showed a γ value as ranges from 0.534 to 0.571, which further declared a similar, but more nuanced, characterization of high levels of hybridization for *fascicularis/mulatta* lineage (**supplementary table S11)**. Finally, bootstrap resampling of individuals within hybrid populations were performed to obtain a distribution of gamma values to assess heterogeneity in levels of introgression (**Fig. 3F**).

Finally, F_d statistic [56] was initially used to calculate the potential introgression in 10-kb windows firstly and genome regions with top 0.05 high F_d values were selected. Given it is of challenge to accurately identify the introgressed regions in macaque genomes, we applied a two-layer hidden Markov model [57] to infer local ancestry of admixed individuals of the *fascicularis/mulatta* lineage. We selected the top 1% windows with high sum of the proportion of source population (*silenus/nigra* lineage) ancestry for all SNPs in a 10 kb windows with a high F_d values as the final potential introgression regions (**Supplementary tables S12**) and found some functional genes (e.g., *CERS3*, *ACER3*, *GSK3B*) present in these potential introgressed regions (**supplementary Fig. S11** and **supplementary table S13**). We then reconstructed a phylogenetic tree based on autosomal SNVs including only SNVs located in all putatively introgressed regions for *fascicularis/mulatta* lineage form *silenus/nigra* lineage and there is no doubt that the *arctoides/sinica* lineages diverged first among Asian macaques (**supplementary Fig. S12**). In addition, we reconstructed a phylogenetic tree excluding SNVs located in all putatively introgressed regions for *fascicularis/mulatta* lineage form *silenus/nigra* lineage (**supplementary Fig. S12**). The result upheld the topology displayed by the species tree (**Fig. 1B**). The phylogenetic tree based on introgressed and non-introgressed regions form *silenus/nigra* lineage resulted in different phylogenetic positions of the *fascicularis/mulatta* lineage.

Population size history estimate

The history of the effective population size (N_e) for macaque species were modeled from the distribution of heterozygous sites across the genome using a pairwise sequentially Markovian coalescent (PSMC) model [58] (**supplementary Fig. 13A-D** and **supplementary Fig. S14**). With the exception of *M. sylvanus* (**supplementary Fig. 13D**), which showed a truncated plot and most likely due to insufficient sequencing depth, the N_e of remaining macaque species are very similar from 5 Mya until ~ 2.5 Mya, indicating that all Asian macaque species shared the same demographic history during the Pliocene. The Pliocene was a period of global cooling after the warmer Miocene and the cooling and drying of the global environment may have contributed to the enormous spread of savannas during this time [59]. The change in vegetation undoubtedly makes a huge impact on arboreal animals, such as macaques. Subsequently, the ancestor of most macaque species showed an upward trend of N_e after ~2.5 Mya. However, the N_e of three species of the *nigra* group remained stable until 1.5 Mya with a subsequent upward trend at different rates, possibly corresponding to population growth and expansion once ecological conditions allowed. Macaques except of the *nigra* group experienced a population growth until the Pleistocene glaciations and the *Ne* of different macaque species declined during different glaciations indicating that they likely were restricted to smaller refugia during glacial periods. During the last interglacial period, N_e for *M. mulatta* (CR), *M. assamensis*, *M. nemestrina* and *M. tonkeana* increased rapidly but subsequently declined again.

Discussion

Phylogeny and classification of species groups

The present study enables us to verify the validity of species groups within the *Macaca* genus and to reconstruct their evolutionary history, which represents an exciting evolutionary scenery as a primate taxon whose "Out of Africa" spread mirrors that of non-African humans. The species tree deduced by ASTRAL supports the division of macaques into seven well-supported clades or lineages that correspond to the classification of the genus into seven species groups [21, 29] (**Fig. 1B**). *M. sylvanus* diverged first and F_{ST} between *M. sylvanus* and the Asian macaques is higher

than any pairs of Asian macaque groups (**supplementary table S6**). The results support the separation of *sylvanus* from the *silenus* group, which also confirmed the classification of Delson [23] based on morphology and geographic distribution. Asian macaques contained the *silenus* and *nigra* lineages and a clade including all other Asian lineages (**Fig. 1B**). The s*ilenus* and *nigra* lineages were grouped into a single species group (*silenus* group) by Fooden [22] and Delson [23], but recently separated into two distinct species groups: *silenus* and *nigra* or Sulawesi [21, 25, 29, 33]. Based on the species trees, the six investigated species of the *silenus* and *nigra* groups cluster into two clades: *silenus* groups (*M. leonina, M. silenus, M. nemestrina*) and *nigra* groups (*M. maura, M. tonkeana* and *M. nigra*) (**Fig. 1B**).

Except for the *silenus/nigra* lineage, the Asian macaques include two further clades. The first one includes *M. fascicularis, M. fuscata, M. cyclopis* and *M. mulatta* (**Fig. 1B**). Initially, all these four species were grouped into the *fascicularis* group [22], but Groves [25] separated *M. fuscata, M. cyclopis* and *M. mulatta* apart from *fascicularis* group into the *mulatta* group. The F_{ST} values vary widely between *M. fascicularis* vs *M. mulatta* (0.171) and *M. fascicularis* vs *M. fuscata* (0.415). Yan, et al. [41] indicated that about 30% genome of mainland *M. fascicularis* are of Chinese *M. mulatta* origin, which shows that the lower genetic distance between *fascicularis* and *mulatta* groups is reasonable and likely caused by hybridization events. Therefore, they should still be divided into two different groups, particularly as *M. fascicularis* from the Sundaland region might not contain any genomic contribution of *M. mulatta*. The second clade contains *M. arctoides, M. assamensis, M. thibetana, M. leucogenys* and *M. sinica* (**Fig. 1B**). *M. arctoides* was assigned to the *arctoides* group [22], but based on molecular data, later integrated into the *sinica* group [23, 34, 35]. In contrast, Groves [25] recognized *M. arctoides* as member of the *fascicularis* group. Due to all these uncertainties, *M. arctoides* was recently separated into its own *arctoides* group [21, 29, 33].

Contributions of ILS and ancient hybird to Gene Tree Heterogeneity

A recent study has reported gene flow events between the *fascicularis* and *silenus* groups and concluded that the introgression signals between these two groups are mainly due to genome similarity in closely related species [48]. In this study, we utilized QuIBL and discovered that among the 120 triplets examined, 60.8% exhibited phylogenetic discordances exclusively resulting

from ILS. Moreover, evidence of introgression was only observed in 9.1% of the tested triplets (**Supplementary tables S8**), indicating that ILS rather than introgression is the primary factor contributing to the discordance among these species. Our research also suggests that ancient hybrid speciation may have occurred among the ancestral populations of *silenus/nigra* and *arctoides/sinica* clades. However, these analyses do not differentiate between ILS and introgression, both of which can create similar phylogenetic signals [60]. Thus, a plausible explanation is that *fascicularis/mulatta* clade may have originated from an ancient hybrid speciation of the two parental lineages (the ancestral populations *silenus/nigra* and *arctoides/sinica* clades) and is accompanied by ILS. The rationale is, first, according to PSMC analysis, all species in the genus *Macaca* were experiencing a population decline during the time window of the potential hybridization event (3.79 – 4.01 Mya). This decline of the ancestral population most likely resulted in the separation of subpopulations leading to rapid ILS [61]. Secondly, our *Dsuit* results based on two most common topologies (**Fig 3C** and **3D**), phylonet (**Supplementary tables S9-S10**) and HyDe results (**Fig. 3E** and **Supplementary tables S11**) all suggested a potential formation of a hybrid species. However, the authenticity of this event could be disturbed due to two reasons. First, it is difficult to detect the superposition of multiple introgression events throughout macaque history accurately. Additionally, ancient hybridization events may have unique evolutionary histories, such as ongoing gene flow, distinct introgression histories, maintenance of assortative mating, and rate heterogeneity [62]. Secondly, the power of HyDe and D-statistic in detecting hybridization events under high ILS and ancient rapid radiations is limited [63]. Simulation analyses have found that in the case of high level of ILS, the power of HyDe decreases to 0-0.1, and the power for the D-statistic was relatively poor either (0.03–0.7). Moreover, ancient hybridization can reduce the accuracy of the γ estimates [63]. However, though they are not the sole source, both ILS and potential hybrid speciation have contributed to the conflicts of macaques.

D-statistics also reveal asymmetric results for *fascicularis/mulatta* lineages which suggested introgression for *fascicularis/mulatta* lineages. In addition, PhyloNet and HyDe analyses suggested high levels of introgression or hybridization occurred for all macaques from *fascicularis/mulatta* lineage whether its current distribution is isolated or not. Furthermore, the

phylogenetic tree based on SNVs from introgressed and non-introgressed regions, which was identified by F_d statistic and a two-layer hidden Markov model, resulted in different phylogenetic positions of the *fascicularis/mulatta* lineage. And the phylogenetic tree excluding SNVs located in all putatively introgressed regions for the *fascicularis/mulatta* lineage from *silenus/nigra* lineage (**supplementary Fig. S12**) upheld the topology displayed by the species tree (**Fig. 1B**). All these results suggested that potential ancient hybridization occurred in the *fascicularis/mulatta* lineage.

Ancient admixture of ancestral variation is a powerful means for rapid radiations to occur [64]. Hybridization analyses and QuIBL tests support this idea. However, apart from ancient introgression between macaque groups, secondary introgression and recent gene flow can also result in differences in D-statistics. Though HyDe and the D-statistics are powerful for detecting hybridization in most of scenarios [63], to what extent the ILS or potential hybrid have shaped the macaque history may require large-scale genomic data sets to perform optimally under complex evolutionary scenarios in future.

The radiation of macaques

Fossil data suggest that the genus *Macaca* arose about 7 Mya in North Africa during the Late Miocene [23, 59, 60, 61], which is consistent with our MCMCTree result (7.76 Mya, **Fig. 2**). The Messinian Salinity Crisis (5.9-5.3 Mya), in which the level of the Mediterranean Sea fluctuated and occasionally dried out, may have facilitated macaque dispersal from North Africa [51]. A large number of macaque fossils have been reported in Europe near the Mediterranean [33, 51]. Macaque ancestors likely invaded Europe via the arid Mediterranean during Messinian stage [26, 33, 46, 51]. The ancestor of *M. sylvanus* split from the main stem ~5.29 Mya and settled around the Mediterranean (**Fig. 2** and **Fig. 4**). Dental specimens of the oldest known Asian fossil referred to the genus *Macaca* are *M. palaeindicus* from the Late Pliocene Tatrot formation of the Siwalik Hills [23, 59, 62]. Hence, the ancestor of Asian macaques spread eastward step by step likely along the Siwalik Hills to the Hengduan Mountains, which are a potential diversification hotspot for primates [1, 65, 66, 67, 68], and diverged into different lineages. Firstly, the ancestor of *silenus/nigra* split ~4.01 Mya and then hybridized with the ancestor of other Asian macaques and formed a pattern of three clades (**Fig. 4**).

Conclusions

In this study, we conducted a phylogenetic analysis of 24 whole macaque genomes of which 16 are newly generated. We identified ~20 million high-quality SNVs in each macaque individual, which can be invoked as important baseline data in other studies. We support the division of macaques into seven species groups and found that there were visible differences in F_{ST} within and between species groups. Importantly, our results reveal *fascicularis/mulatta* clade may have originated from an ancient hybrid speciation of the two parental lineages (the ancestral populations *silenus/nigra* and *arctoides/sinica* clades) and is accompanied by high levels of ILS. Most prominently, we have used the available fossil data and the literature to reconstruct the radiation scenario of the genus macaques. In summary, this study suggests that high level of ILS and potential hybrid speciation play important roles in the radiation of macaques.

Materials and methods

Sample collection and library preparation

Samples of 16 macaque individuals were obtained from the Institute of Zoology, Chinese Academic of Sciences and the German Primate Center (**supplementary table S1**). In addition, genome data of eight other macaque individuals and one *Papio anubis* individual were obtained from the NCBI Short Read Archive (accession no. show in **supplementary table S1**). The present study was approved by the Animal Ethics Committee of the Institute of Zoology, Chinese Academy of Sciences. The procedure of blood and tissue collection was in strict accordance with the Animal Ethics Procedures and Guidelines of the People's Republic of China. DNA was isolated from blood or tissue samples with the Qiagen Blood and Tissue Kit. Sequencing libraries were prepared based on KAPA library preparation kit with insert sizes of 300-500 bp and sequenced using Illumina Hi-Seq 2500 and 4000 technologies. Quality control was performed using FastQC with default parameters (www.bioinformatics.babraham.ac.uk/projects/fastqc/).

Data Processing

Paired-end reads for all samples were mapped to the *Papio anubis* reference genome (Panu_3.0, GCA_000264685.2) with Burrows-Wheeler Aligner (BWA) mem v0.7.17 [69], except for *M. nemestrina* (2), which was trimmed by cutadapt.2.4 [70] firstly because of the poor sequencing quality due to DNA degradation. The output bam files were sorted using samtools [71] and duplicates were marked with GATK-v4.1.2.0 [72]. Single-nucleotide variants (SNVs) were called using GATK 4.1 [72] following best practice. We obtained the GVCF file for each individual using the “HaplotypeCaller” method in GATK and then all samples were merged based on “CombineGVCFs”. We used the GenotypeGVCFs-based method with the “includeNonVariantSites” flag to get the vcf files. After that, we applied the “SelectVariants” to exclude indels and split the variant and nonvariant sites. All nonvariant sites were filtered in subsequent analyses. We applied the hard filter command “VariantFiltration” to exclude potential false-positive variant calls with the following criteria: “-filterExpression QD< 5.0 || FS > 60.0 || MQ < 40.0 || ReadPosRankSum< -8.0 || MQRankSum < -12.5” and “-genotypeFilterExpression DP < 4.0” [73]. All variant calls that failed the filters were removed. All SNVs were annotated by ANNOVAR v2013–06-21 [74]. (**supplementary tables S3** and **S4**) based on gff3 file from

Ensembl (Papio_anubis.Panu_3.0.104.gff3). Heterozygosity is defined as the rate of heterozygous SNVs in the genome, which were further filter by the following parameters: "--minGQ 20" and "--minDP 8" to ensure data quality based on VCFtools. We calculated the heterozygosity rate in a nonoverlapping window (window = 100 kb) for all 24 macaques. Genome-wide distribution of heterozygosity was shown in a box plot using boxplot in R package (**supplementary table S2** and **supplementary Fig. S2**).

Phylogenetic reconstruction using the window-based strategy

We constructed the consensus sequences for all individual genomes and aligned them per chromosome (sex chromosomes, mitogenomes and scaffolds were excluded) based on VCF, which did not include variants present in the Papio reference but were invariant among macaques. In addition, indels, repetitive regions and low-quality regions were removed from the consensus sequences. Per-chromosome alignments were split into nonoverlapping windows of 200 kbp. The maximum likelihood (ML) trees for each window were performed with RAxML v8 [75] with 100 bootstrap replicates and the GTR + CAT substitution model. From all 200 kbp window trees, we reconstructed a species tree with ASTRAL v4.10.5 [76] under the multispecies coalescent model returning quartet scores and posterior probabilities. The species tree was rooted with *Papio anubis*. Consensus networks of the window trees were generated using SplitsTree5 [40] with different thresholds. Genealogical discordance across the genome of rhesus macaque was assessed by Twisst [77]. All GF trees were clustered based on PhyBin (https://github.com/rrnewton/PhyBin) and the top two topologies were visualized by DensiTree [78].

Phylogenetic analysis of concatenated whole-genome SNV data

SNVs were extracted from the filtered VCF file and converted into individual FASTA sequences using a python script (https://doi.org/10.5281/zenodo.2540861). Merging all FASTA sequences into a single file provided a multiple sequence alignment of all individuals and all concatenated SNVs. Phylogenetic trees were reconstructed with maximum-likelihood (ML) using RAxML v8 [75] with 100 bootstrap replicates and the GTR + CAT substitution model. In addition, a neighbor-joining (NJ) tree was constructed using the program TreeBeST V.1.9.2 (http://treesoft.sourceforge.net/treebest.shtml), which has a build-in algorithm to build the best tree that reconciled with the species tree and rooted with the minimized number of duplications and losses. To further visualize genetic relationships among macaques, we performed principal

components analysis (PCA) on the filtered autosomal SNVs using the Eigensoft package (version 5.0) [79].

Phylogeny of whole mitochondrial genomes

We used NOVOPlasty [80] to *de novo* reconstruct the mitochondrial genomes (mitogenomes) for 18 macaques representing 15 species (**supplementary table S5**) and then annotated the genome sequences using MitoZ [81]. In addition, we downloaded 15 mitogenomes for the genus *Macaca* (**supplementary table S5**) from NCBI based on the filter criteria mentioned in Roos, et al. [33]. The 33 mitogenomes were aligned using MAFFT version 7 [82] and indels and poorly aligned positions were removed with Gblocks version 0.91b [83]. A ML tree was reconstructed using RAxML v 8 [75] with a GTR + CAT substitution model and applying 1000 bootstrap replicates to verify nodal support.

QuIBL

We used QuIBL [18] to assess the likelihood of introgression and ILS for each locus in every species triplet under evaluation. QuIBL provides estimates of the proportion of introgression and the probability of a locus falling into a model with either introgression or ILS alone. To begin, QuIBL estimated the internal branch length distribution for each locus in the triplet of species. Given its sensitivity to recombination [18], we extracted 20-kb windows separated by 400-kb from each sample to minimize the risk of including a recombination breakpoint in the window. We filtered the inferred ML trees based on the number of parsimony-informative sites (≧10), resulting in a total of 3,175 trees that served as QuIBL inputs. We used the species tree topology to assign the outgroup to each triplet and also calculated the percentage of loci that supported discordant topologies and showed significant evidence for introgression.

In order to determine whether ILS (scenario 1) or a combination of ILS and introgression reflected the inner branch lengths, QuIBL evaluated the probability values (Bayesian Information Criterion test, BIC) (scenario 2). The scenario of ILS only with the lower BIC value is preferred when Delta. BIC ($BIC_{scenario2}$-$BIC_{scenario1}$) is larger than 10 and the BIC value is less than 0. The combination of ILS and introgression with the lower BIC value, however, is preferred when Delta.BIC is smaller than -10.

D-statistics

We implemented D-statistics with *Dsuite* [53] across all combinations of the 16 macaques described above. The topology required for *Dsuite* was consistent with two alternative topologies described in **Fig. 3A-B**, respectively. Then, the D-statistics of all possible combinations with *Papio anubis* as the outgroup (Out) and three different macaque species with a phylogenetic topology following (((M1, M2), M3), Out) were calculated with the Dtrios module. In total, 566 groupings were analyzed for each topology. After that, D-statistics were calculated for each branch of the alternative topologies using the f-branch module, and visualized the statistical results using dtools.py script provided in the *Dsuite* software.

Hybridization Detection

To obtain further support for the ancient introgression suggested by the above approach based on the D-statistics, we applied HyDe (Hybridization Detection) [55], which can automate the detection of hybridization across large numbers of species despite ILS and can perform hypothesis tests at the population or individual level by estimating the amount of admixture (γ). The detailed analysis procedure is as follows: We applied the population approach in "run_hyde_mp.py" script to test hybridization events on all triples among macaque lineages in **Fig. 1C** (*fascicularis*/*mulatta* clade, *arctoides/sinica* clade, *silenus/nigra* clade, and *sylvanus* clade) with *Papio anubis* as the outgroup. Results are first filtered based on whether there are significant *P* values (*P* values <0.05) for hybridization. In addition, different γ values represent different events. For example, a 50:50 hybrid is characterized by a γ value of about 0.5 and a very low levels of admixture (e.g., 0.01 means close to parent P1 while 0.99 means close to parent P2) can be indicators of different processes such as ILS and more ancient hybridization. We found that a 50:50 hybrid was significantly concentrated among the *silenus/nigra*, *fascicularis*/*mulatta*, and *arctoides/sinica* clades. We then used the script "run_hyde_mp.py" to test each individual within a putative hybrid population (*fascicularis*/*mulatta* clade) using specified triples and then performed bootstrap resampling (500 replicates) of the individuals within the putative hybrid lineages for each specified triple based on "bootstrap_hyde.py" script.

Maximum likelihood inference for reticulation with PhyloNet

We used the InferNetwork_MPL program in PhyloNet [54], which is based on Maximum Pseudo Likelihood (MPL) algorithm, to analyze a set of every 3th GF ML tree, that is, 2150 trees in a coalescent framework that accounts for incomplete lineage sorting (ILS) while allowing different

numbers of reticulation events. In addition, the individual "2" in all macaque species were pruned from the input gene trees to reduce complexity and computational demand. We set the numbers of reticulation events as 1 and 2 to conduct two independent analyses and each analysis was run with 100 iterations, yielding the five networks with the highest likelihood scores. We visualized the optimal networks in Dendroscope v3 [84] and present it in **supplementary Fig. S10.**

F_d value

To identify putatively introgressed regions in macaques, we screened these segments according to the strategy of Martin, et al. [56] and van der Valk, et al. [6]. The method predicts that introgression between species at a specific genome region should reduce the between-species divergence in this region as compared to the rest of the genome [6, 56]. We computed the F_d value, which is sensitive to the number of informative sites within windows, for each 10-kb window across the whole genome of each macaque.

Introgression analysis with efficient local ancestry inference

The two-layer admixture model implemented in efficient local ancestry inference [57] was used to infer introgressed segments in *fascicularis/mulatta* (the admixture population, -p 1) with the *silenus/nigra* as source population 2 (-p 11) and *arctoides*/*sinica* were used as source population 1 (-p 10). We run 30 iterations with mixture generation values of 2,910,000, 3,200,000 and 3,450,000, which was estimated based on MCMCTree result, and the averaged results over the three independent runs were used for detecting introgression tracts. Sites with a proportion of source population 2 ancestry greater than 1.5 were defined as introgression sites from *silenus/nigra* lineage which may cause position conflict of *fascicularis/mulatta* lineage in species tree. We sum the proportion of source population 2 ancestry in each 10 kb windows, which is same to F_d and the top 1% windows were defined as putatively introgression tracts.

Analysis of demographic history using genome-wide data

We applied the pairwise sequentially Markovian coalescent (PSMC, v0.6.5-r67) analysis [58] to estimate the demographic history of each macaque individual (**supplementary Fig. S2**). We use the Samtools mpileup [71] to generate the input files for PSMC with the parameter "-C" set to be 50 to improve mapping quality. In addition, only autosomal SNVs were reserved to generate the diploid sequence for PSMC. PSMC was run with 25 iterations (-N), a maximum 2N0 coalescent time (-t) of 15, an initial theta/rho ratio (-r) of 5, and the 64 time-intervals were parameterized as

"4 + 25*2 + 4 + 6". PSMC plots were scaled with a mutation rate (μ) of 0.9×10^{-8} and a generation time (g) of 11 years for all macaque species [85]. Bootstrap analyses (100 bootstrap replications) were performed for each species (**supplementary Fig. S14**).

Divergence time calibration

Coding sequences (CDSs) and protein sequences for different representative outgroup species and three macaques (*M. mulatta*, *M. fascicularis*, *M. nemestrina*, *Papio anubis*, *Homo sapiens*, *Pan troglodytes* and *Pongo abelii*) were retrieved from Ensembl (release-98). The single-copy orthologue gene list shared by these species was obtained using OrthoFinder [86]. The macaques were mapped to the Panu_3.0 genome and thus their CDSs have the same genomic coordinates. Therefore, the CDSs of all macaques were extracted after ortholog detection. CDSs of these single-copy genes were aligned by PRANK (v.170427) [87] under codon model and concatenated into a single super-matrix. Fourfold degenerate (4D) sites were extracted from them using MEGA7 [88] and concatenated to calibrated the divergence time of the genus *Macaca* with MCMCTree program in PAML v4.9 (Phylogenetic Analysis of ML, v4.9) package [49]. Four calibration points were applied to constrain the divergence time of the nodes: Catarrhini 20–38 Ma, Hominidae 13–18 Ma, *Homo-Pan* 6–7 Ma and *Papio-Macaca* 5–8 Ma [33, 89]. To check for convergence of the stationary distribution, the analysis was run in duplicate, and the results were compared between runs and similar results for divergence time of macaques were obtained (**supplementary Fig. S9**).

Funding

The work was supported by the National Natural Science Foundation of China (31821001 and 31530068), the Strategic Priority Research Program of Chinese Academy of Sciences (XDB31000000, XDA23080000 and XDA19050000), and Sino-German Mobility Programme (M-0084).

Acknowledgements

The authers thank Prof. Luo Shujin and Dr. Sun Xin for their assistance in DNA isolation, library preparation and quality control for *M. nemestrina (2)*.

Author contributions

ML conceived this study, and X.Z. and C.R. contributed to the idea for this study. X.T., J.Q. and Z.L. performed experiments, analyzed data and wrote the draft manuscript. G.L., L.Z. and Y. S. help to analyze data. P.F. C.R. and J.L. provided samples. M.L., X.M. and C.R. revised the draft manuscript. All authors contributed to the final text.

Data accessibility

Whole genome resequencing data produced in this study were deposited at the NCBI Sequence Read Archive (SRA) under Bioproject PRJNA732172.

Conflicts of interest

The authors declare that there is no conflict of interest.

References

1. Roos C, Zinner D, Kubatko LS, Schwarz C, Yang M, Meyer D, Nash SD, Xing J, Batzer MA, Brameier M, et al. 2011. Nuclear versus mitochondrial DNA: evidence for hybridization in colobine monkeys. *BMC Evol Biol.* 11:77.

2. Zinner D, Arnold ML, Roos C. 2011. The strange blood: natural hybridization in primates. *Evol Anthropol.* 20:96-103.

3. Abbott R, Albach D, Ansell S, Arntzen JW, Baird SJ, Bierne N, Boughman J, Brelsford A, Buerkle CA, Buggs R, et al. 2013. Hybridization and speciation. *J Evol Biol.* 26:229-246.

4. Tung J, Barreiro LB. 2017. The contribution of admixture to primate evolution. *Curr Opin Genet Dev.* 47:61-68.

5. Fontsere C, de Manuel M, Marques-Bonet T, Kuhlwilm M. 2019. Admixture in Mammals and How to Understand Its Functional Implications: On the Abundance of Gene Flow in Mammalian Species, Its Impact on the Genome, and Roads into a Functional Understanding. Bioessays 41:e1900123.

6. van der Valk T, Gonda CM, Silegowa H, Almanza S, Sifuentes-Romero I, Hart TB, Hart JA, Detwiler KM, Guschanski K. 2020. The Genome of the Endangered Dryas Monkey Provides New Insights into the Evolutionary History of the Vervets. *Mol Biol Evol.* 37:183-194.

7. Vanderpool D, Minh BQ, Lanfear R, Hughes D, Murali S, Harris RA, Raveendran M, Muzny DM, Hibbins MS, Williamson RJ, et al. 2020. Primate phylogenomics uncovers multiple rapid radiations and ancient interspecific introgression. *PLoS Biol.* 18:e3000954.

8. Abi-Rached L, Jobin MJ, Kulkarni S, McWhinnie A, Dalva K, Gragert L, Babrzadeh F, Gharizadeh B, Luo M, Plummer FA, et al. 2011. The shaping of modern human immune systems by multiregional admixture with archaic humans. *Science* 334:89-94.

9. Abbott RJ, Barton NH, Good JM. 2016. Genomics of hybridization and its evolutionary consequences. *Mol Ecol.* 25:2325-2332.

10. Zheng Z, Wang X, Li M, Li Y, Yang Z, Wang X, Pan X, Gong M, Zhang Y, Guo Y, et al. 2020. The origin of domestication genes in goats. *Sci Adv.* 6:eaaz5216.

11. Figueiró HV, Li G, Trindade FJ, Assis J, Pais F, Fernandes G, Santos SHD, Hughes GM, Komissarov A, Antunes A, et al. 2017. Genome-wide signatures of complex introgression and

adaptive evolution in the big cats. *Sci Adv.* 3:e1700299.

12. Árnason Ú, Lammers F, Kumar V, Nilsson MA, Janke A. 2018. Whole-genome sequencing of the blue whale and other rorquals finds signatures for introgressive gene flow. *Sci Adv.* 4:eaap9873.

13. Avise JC, Robinson TJ. 2008. Hemiplasy: a new term in the lexicon of phylogenetics. *Syst Biol.* 57:503-507.

14. Maddison WP, Knowles LL. 2006. Inferring phylogeny despite incomplete lineage sorting. *Syst Biol.* 55:21-30.

15. Szöllősi GJ, Tannier E, Daubin V, Boussau B. 2015. The inference of gene trees with species trees. *Syst Biol.* 64:e42-62.

16. Heled J, Drummond AJ. 2010. Bayesian inference of species trees from multilocus data. *Mol Biol Evol.* 27:570-580.

17. Feng S, Bai M, Rivas-González I, Li C, Liu S, Tong Y, Yang H, Chen G, Xie D, Sears KE, et al. 2022. Incomplete lineage sorting and phenotypic evolution in marsupials. *Cell* 185:1646-1660.e1618.

18. Edelman NB, Frandsen PB, Miyagi M, Clavijo B, Davey J, Dikow RB, García-Accinelli G, Van Belleghem SM, Patterson N, Neafsey DE, et al. 2019. Genomic architecture and introgression shape a butterfly radiation. *Science* 366:594-599.

19. Wu DD, Qi XG, Yu L, Li M, Liu ZJ, Yoder AD, Roos C, Hayakawa T, Rogers J, Marques-Bonet T, et al. 2022: Initiation of the Primate Genome Project. *Zool Res.* 43: 147-149.

20. Rogers J, Raveendran M, Harris RA, Mailund T, Leppälä K, Athanasiadis G, Schierup MH, Cheng J, Munch K, Walker JA, et al. 2019. The comparative genomics and complex population history of *Papio* baboons. *Sci Adv.* 5:eaau6947.

21. Zinner D, Fickenscher GH, Roos C. 2013. Family Cercopithecidae (Old World Monkeys). In: Mittermeier RA, Rylands AB, Wilson DE, editors. Handbook of the Mammals of the World. Volume 3, Primates. Barcelona, Lynx Edicions, p.550-627.

22. Fooden J. 1976. Provisional classifications and key to living species of macaques (Primates: *Macaca*). *Folia Primatol.* 25:225-236.

23. Delson E. 1980. Fossil macaques, phyletic relationships and a scenario of deployment. In: D. Lindburg D, editor. The Macaques. Studies in Ecology, Behavior, and Evolution. New York: van Nostrand Rheinhold, p.10-30.

24. Morales JC, Melnick DJ. 1998. Phylogenetic relationships of the macaques (Cercopithecidae: *Macaca*), as revealed by high resolution restriction site mapping of mitochondrial ribosomal genes. *J Hum Evol.* 34:1-23.

25. Groves CP. 2001. Primate Taxonomy. Washington DC: Smithsonian Institution Press.

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

27. Sinha A, Datta A, Madhusudan MD, Mishra C. 2005. *Macaca munzala*: A new species from western Arunachal Pradesh, northeastern India. *Int J Primatol.* 26:977-989.

28. Chakraborty D, Ramakrishnan U, Panor J, Mishra C, Sinha A. 2007. Phylogenetic relationships and morphometric affinities of the Arunachal macaque *Macaca munzala*, a newly described primate from Arunachal Pradesh, northeastern India. *Mol Phylogenet Evol.* 44:838-849.

29. Roos C, Boonratana R, Supriatna J, Fellowes JR, Groves CP, Nash SD, Rylands AB, Mittermeier RA. 2014. An updated taxonomy and conservation status review of Asian primates. *Asian Primates J.* 4(2):1-38.

30. Li C, Zhao C, Fan PF. 2015. White-cheeked macaque (*Macaca leucogenys*): A new macaque species from Medog, southeastern Tibet. *Am J Primatol.* 77:753-766.

31. Hou W, Liu S, Jiang J, Fan Z, Fan P, Li J. 2016. The complete mitochondrial genome of White-cheeked macaque (*Macaca leucogenys*). *Mitochondrial DNA B Resour.* 1:374-375.

32. Fan P, Liu Y, Zhang Z, Zhao C, Li C, Liu W, Liu Z, Li M. 2017. Phylogenetic position of the white-cheeked macaque (*Macaca leucogenys*), a newly described primate from southeastern Tibet. *Mol Phylogenet Evol.* 107:80-89.

33. Roos C, Kothe M, Alba DM, Delson E, Zinner D. 2019. The radiation of macaques out of Africa: Evidence from mitogenome divergence times and the fossil record. *J Hum Evol.* 133:114-132.

34. Tosi AJ, Morales JC, Melnick DJ. 2000. Comparison of Y chromosome and mtDNA phylogenies leads to unique inferences of macaque evolutionary history. *Mol Phylogenet Evol.* 17:133-144.

35. Li J, Han K, Xing J, Kim HS, Rogers J, Ryder OA, Disotell T, Yue B, Batzer MA. 2009. Phylogeny of the macaques (Cercopithecidae: *Macaca*) based on *Alu* elements. *Gene* 448:242-249.

36. Fan Z, Zhou A, Osada N, Yu J, Jiang J, Li P, Du L, Niu L, Deng J, Xu H, et al. 2018. Ancient hybridization and admixture in macaques (genus *Macaca*) inferred from whole genome sequences. *Mol Phylogenet Evol.* 127:376-386.

37. Ciani AC, Stanyon R, Scheffrahn W, Sampurno B. 1989. Evidence of gene flow between Sulawesi macaques. *Am J Primatol*. 17:257-270.

38. Evans BJ, Supriatna J, Melnick DJ. 2001. Hybridization and population genetics of two macaque species in Sulawesi, Indonesia. *Evolution* 55:1686-1702.

39. Evans BJ, Supriatna J, Andayani N, Setiadi MI, Cannatella DC, Melnick DJ. 2003. Monkeys and toads define areas of endemism on Sulawesi. *Evolution* 57:1436-1443.

40. Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol.* 23:254-267.

41. Yan G, Zhang G, Fang X, Zhang Y, Li C, Ling F, Cooper DN, Li Q, Li Y, van Gool AJ, et al. 2011. Genome sequencing and comparison of two nonhuman primate animal models, the cynomolgus and Chinese rhesus macaques. *Nat Biotechnol.* 29:1019-1023.

42. Hamada Y, Yamamoto A, Kunimatsu Y, Tojima S, Mouri T, Kawamoto Y. 2012. Variability of tail length in hybrids of the Japanese macaque (*Macaca fuscata*) and the Taiwanese macaque (*Macaca cyclopis*). *Primates* 53:397-411.

43. Fan Z, Zhao G, Li P, Osada N, Xing J, Yi Y, Du L, Silva P, Wang H, Sakate R, 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.

44. Evans BJ, Tosi AJ, Zeng K, Dushoff J, Corvelo A, Melnick DJ. 2017. Speciation over the edge: gene flow among non-human primate species across a formidable biogeographic barrier. *R Soc Open Sci.* 4:170351.

45. Ito T, Kanthaswamy S, Bunlungsup S, Oldt RF, Houghton P, Hamada Y, Malaivijitnond S. 2020. Secondary contact and genomic admixture between rhesus and long-tailed macaques in the Indochina Peninsula. *J Evol Biol.* 33:1164-1179.

46. Tosi AJ, Disotell TR, Morales JC, Melnick DJ. 2003. Cercopithecine Y-chromosome data provide a test of competing morphological evolutionary hypotheses. *Mol Phylogenet Evol.* 27:510-521.

47. Matsudaira K, Hamada Y, Bunlungsup S, Ishida T, San AM, Malaivijitnond S. 2018. Whole

Mitochondrial Genomic and Y-Chromosomal Phylogenies of Burmese Long-Tailed Macaque (*Macaca fascicularis aurea*) Suggest Ancient Hybridization between *fascicularis* and *sinica* Species Groups. *J Hered.* 109:360-371.

48. Song Y, Jiang C, Li KH, Li J, Qiu H, Price M, Fan ZX, Li J. 2021. Genome-wide analysis reveals signatures of complex introgressive gene flow in macaques (genus Macaca). *Zool Res.* 42:433-449.

49. Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 24:1586-1591.

50. Elton S, O'Regan HJ. 2014. Macaques at the margins: the biogeography and extinction of *Macaca sylvanus* in Europe. *Quat Sci Rev.* 96:117-130.

51. Hamada Y, Urasopon N, Hadi I, Malaivijitnond S. 2006. Body size and proportions and pelage color of free-ranging *Macaca mulatta* from a zone of hybridization in northeastern Thailand. *Int J Primatol.* 27:497-513.

52. Durand EY, Patterson N, Reich D, Slatkin M. 2011. Testing for ancient admixture between closely related populations. *Mol Biol Evol.* 28:2239-2252.

53. Malinsky M, Matschiner M, Svardal H. 2021. Dsuite fast D statistics and related admixture evidence from VCF files. *Mol Ecol Resour.* 21:584–595.

54. Yu Y, Dong J, Liu KJ, Nakhleh L. 2014. Maximum likelihood inference of reticulate evolutionary histories. Proceedings of the National Academy of Sciences of the United States of America 111:16448-16453.

55. Blischak PD, Chifman J, Wolfe AD, Kubatko LS. 2018. HyDe: A Python Package for Genome-Scale Hybridization Detection. *Syst Biol.* 67(5):821-829.

56. Martin SH, Davey JW, Jiggins CD. 2015. Evaluating the use of ABBA-BABA statistics to locate introgressed loci. *Mol Biol Evol.* 32:244-257.

57. Guan Y. 2014. Detecting structure of haplotypes and local ancestry. *Genetics* 196:625-642.

58. Li H, Durbin R. 2011. Inference of human population history from individual whole-genome sequences. *Nature* 475:493-U484.

59. Salzmanna U, Williams M, Haywoodd AM, Johnsone A, Kenderc S, Zalasiewicz J. 2011. Climate and environment of a Pliocene warm world. *Palaeogeogr Palaeocl.* 309:1-8

60. Fontaine MC, Pease JB, Steele A, Waterhouse RM, Neafsey DE, Sharakhov IV, Jiang X, Hall

AB, Catteruccia F, Kakani E, et al. 2015. Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science* 347(6217):1258524.

61, Whitfield JB, Lockhart PJ. 2007. Deciphering ancient rapid radiations. *Trends Ecol Evol*. 22(5):258–265.

62. Sankararaman S, Mallick S, Dannemann M, Prüfer K, Kelso J, Pääbo S, Patterson N, Reich D. 2014. The genomic landscape of Neanderthal ancestry in present-day humans. *Nature* 507(7492):354-357.

63. Kong S, Kubatko LS. 2021 Comparative Performance of Popular Methods for Hybrid Detection using Genomic Data. *Syst Biol.* 70(5):891-907.

64. Marques DA, Meier JI, Seehausen O. 2019. A combinatorial view on speciation and adaptive radiation. *Trends Ecol Evol*. 34(6): 531–544

65. Peng YZ, Pan RL, Jablonski NG. 1993. Classification and evolution of Asian colobines. Folia primatologica; *Int J Primatol*. 60:106-117.

66. Jablonski NG. 1998. The Natural History of the Doucs and Snub-Nosed Monkeys. Singapore: World Scientific Press.

67. Meijaard E, Groves CP. 2006. The geography of mammals and rivers in mainland Southeast Asia. In: Primate Biogeography: Progress and Prospects. Boston: Springer, p.305-329.

68. Thinh VN, Mootnick AR, Geissmann T, Li M, Ziegler T, Agil M, Moisson P, Nadler T, Walter L, Roos C. 2010. Mitochondrial evidence for multiple radiations in the evolutionary history of small apes. *BMC Evol Biol.* 10:74.

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

70. Kechin A, Boyarskikh U, Kel A, Filipenko M. 2017. cutPrimers: A New Tool for Accurate Cutting of Primers from Reads of Targeted Next Generation Sequencing. *J Comput Biol.* 24:1138-1143.

71. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078-2079.

72. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20:1297-1303.

73. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, et al. 2013. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* 43:11.10.11-11.10.33.

74. Wang K, Li M, Hakonarson H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 38:e164-e164.

75. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30:1312-1313.

76. Sayyari E, Mirarab S. 2016. Fast Coalescent-Based Computation of Local Branch Support from Quartet Frequencies. *Mol Biol Evol.* 33:1654-1668.

77. Martin SH, Van Belleghem SM. 2017. Exploring Evolutionary Relationships Across the Genome Using Topology Weighting. *Genetics* 206:429-438.

78. Bouckaert RR. 2010. DensiTree: making sense of sets of phylogenetic trees. *Bioinformatics* 26:1372-1373.

79. Patterson N, Price AL, Reich D. 2006. Population structure and Eigenanalysis. *PLoS Genet.* 2:e190.

80. Dierckxsens N, Mardulyn P, Smits G. 2017. NOVOPlasty: de novo assembly of organelle genomes from whole genome data. *Nucleic Acids Res.* 45:e18.

81. Meng G, Li Y, Yang C, Liu S. 2019. MitoZ: a toolkit for animal mitochondrial genome assembly, annotation and visualization. *Nucleic Acids Res.* 47:e63.

82. Yamada KD, Tomii K, Katoh K. 2016. Application of the MAFFT sequence alignment program to large data-reexamination of the usefulness of chained guide trees. *Bioinformatics* 32:3246-325

83. Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol.* 17:540-552.

84. Huson DH, Scornavacca C. 2012. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. *Syst Biol.* 61:1061-1067.

85. Xue C, Raveendran M, Harris RA, Fawcett GL, Liu X, White S, Dandouli M, Deiros DR, Below JE, Salerno W, et al. 2016. The population genomics of rhesus macaques (*Macaca mulatta*) based on whole-genome sequences. *Genome res.* 26:1651-1662.

86. Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* 20:238.

87. Löytynoja A. 2014. Phylogeny-aware alignment with PRANK. *Methods Mol Biol.* 1079:155-170.

88. Kumar S, Stecher G, Tamura K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol Biol Evol.* 33:1870-1874.

89. Perelman P, Johnson WE, Roos C, Seuánez HN, Horvath JE, Moreira MAM, Kessing B, Pontius J, Roelke M, Rumpler Y, et al. 2011. A Molecular Phylogeny of Living Primates. *PLoS Genet.* 7:e1001342.

Figures and Legends

Figure 1. Macaque phylogenetics and differentiation.

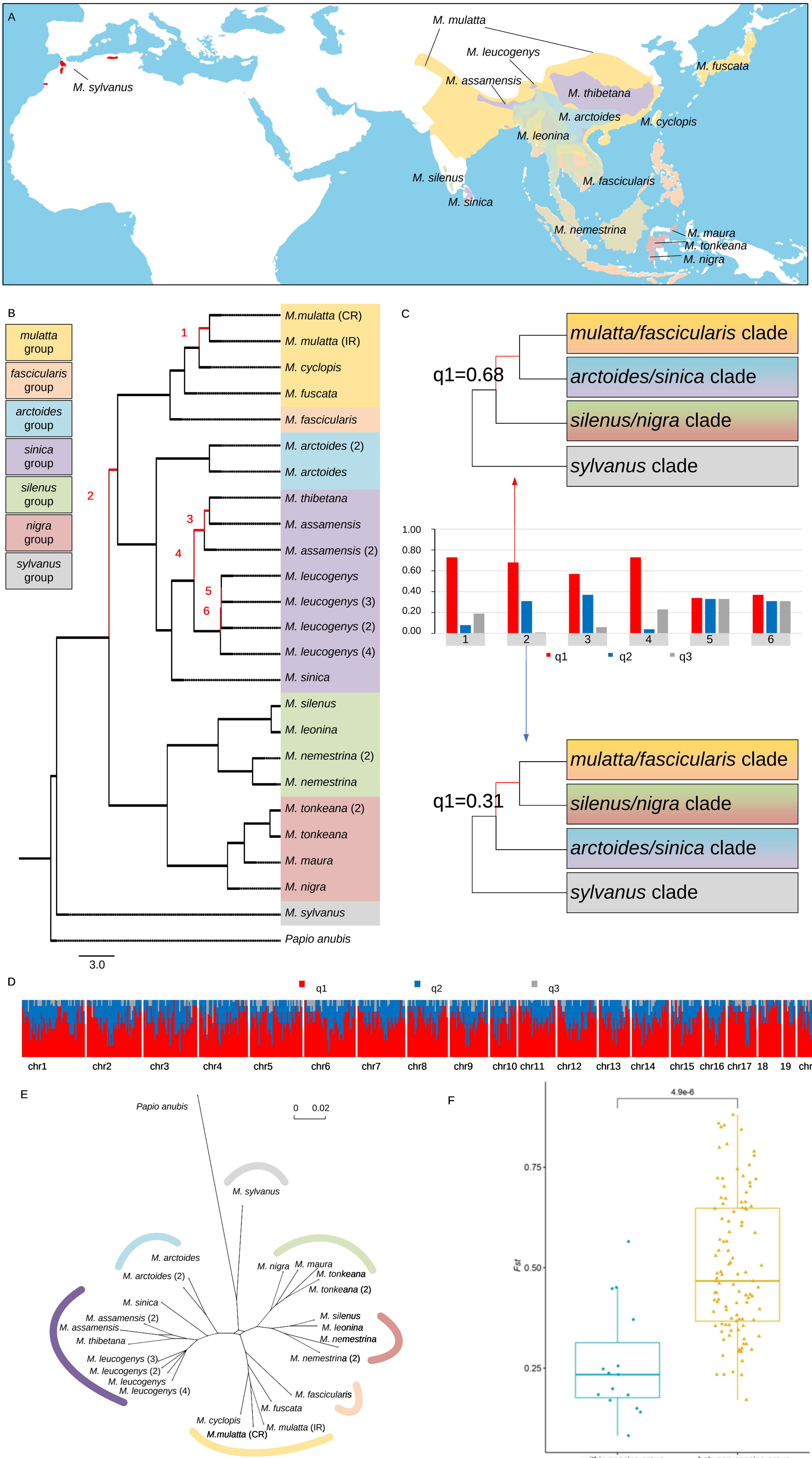
(A) Distribution ranges of the 16 macaque species investigated in this study. The range overlaps are shown in blended colors (orange, purple, green, blue and yellow). (B) A MSC species tree was constructed from 6,351 individual genome fragments. Branches with low q1 (<0.8) are numbered 1 to 6. All branches receive maximal support (P = 1.0, ASTRAL analysis). Branch lengths were calculated from a ML analysis. Different species groups are represented by different background colors which are show to the left. (C) ASTRAL quartet-score analyses for branches 1 to 6 in Fig. 1B. Quartet scores were calculated for the three possible arrangements (q1 to q3) for the respective branch. The principal quartet trees are depicted, with q1 representing the species tree. Two alternative topologies for branch no. 2 as shown in Fig. 1C. (D) Genealogical discordance across the genome of rhesus macaque, demonstrated by a window analysis (200-kb window size) of a full-genome alignment mapped to Panu_3.0 (chromosomes showed at the bottom). The y axis indicates the percentage of windows within a given interval (every 5 windows next to each other) that conform to (red) or reject (blue and grey) the species tree. (E) Consensus networks for macaques from 6,351 genome fragments with a minimum threshold of 30% to form an edge. (F) Distributions of F_{ST} within (blue) and between (yellow) different macaque species groups. The exact *P* value of above were computed based on Wilcoxon test.

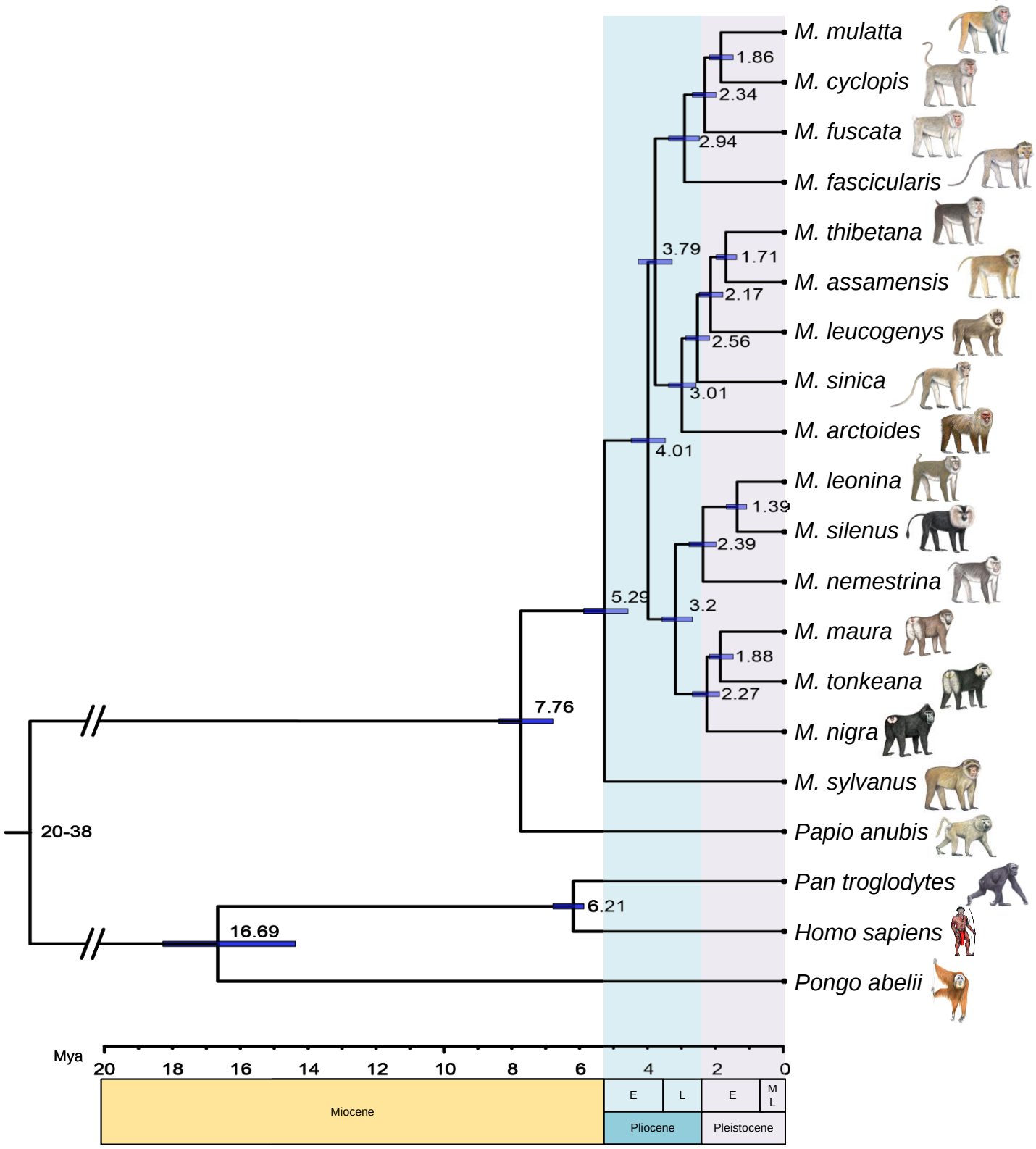
Figure 2. Divergence time tree of genus *Macaca*, estimated from 3,319,411 4d sites from 10,641 single-copy genes. Blue bars indicate 95% HPDs of divergence times and the time scale below shows million years ago (Mya). Abbreviations: L = late, M = middle, E = early. Four fossil-based calibration points are as follows: Catarrhini 20–38 Mya, Hominidae 13–18 Mya, *Homo-Pan* 6–7 Mya and *Papio-Macaca* 5–8 Mya.

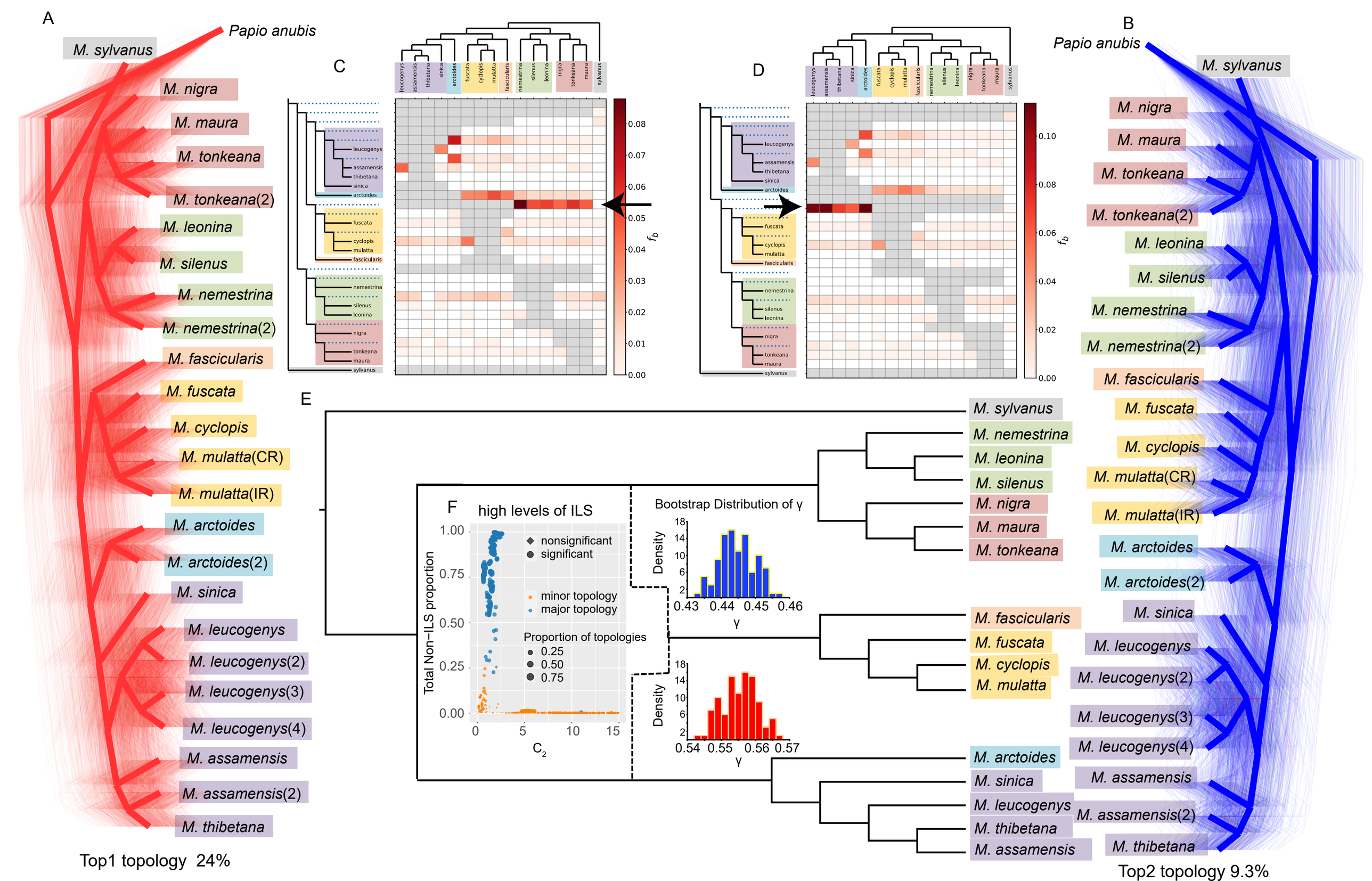
Figure 3 High levels of ILS and footprints of introgression contribute to gene tree heterogeneity. (A), (B) All samples' DensiTree plots were created using 200 bp window trees spanning the whole genome. The top two most common topologies are represented by the colors red (A) and blue (B), respectively. (C), (D) The top two most popular topologies are taken as the

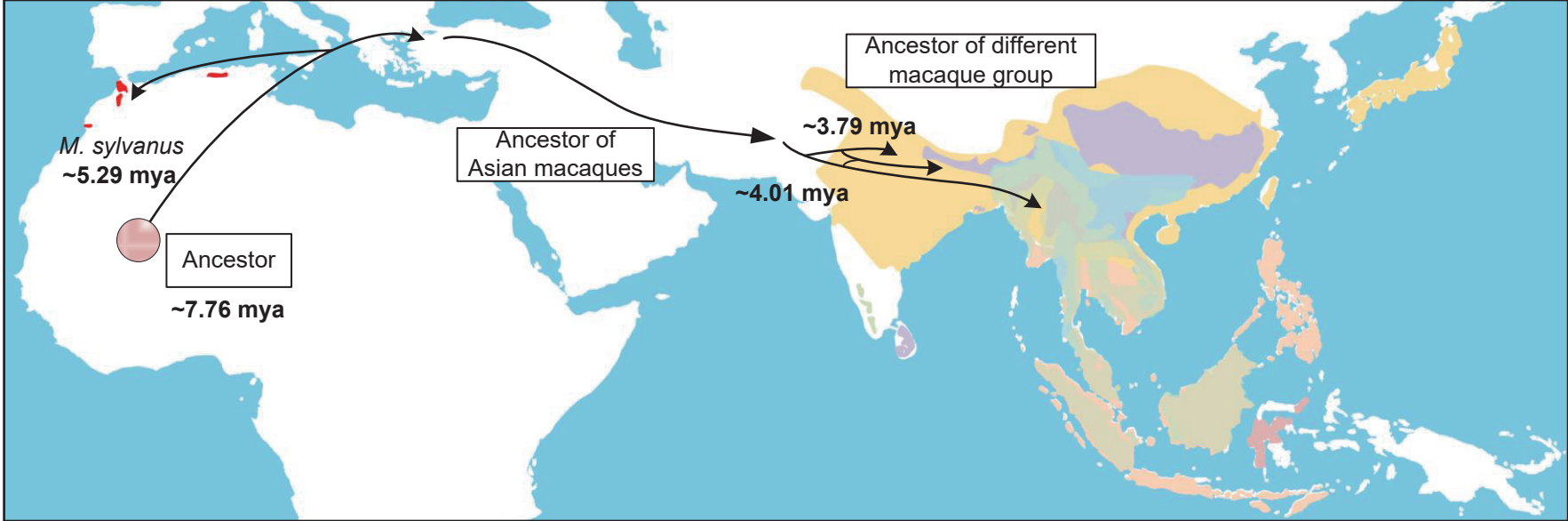
reference trees, and a heatmap is shown to illustrate the statistical support for introgression between species pairs inferred from the Dsuit software. The Fbranch function was used to further process the Dtrios results for every branch on the tree, The black arrows refer to the infiltration of the *fascicularis/mulatta* clade with the *silenus/nigra* (C) and *sinica/arctoides* (D) clades, respectively. (E) The diagram illustrates the complex evolutionary relationships between macaque species, highlighting instances of hybridization and incomplete lineage sorting (ILS). The *fascicularis/mulatta* clade have a hybrid origin, while the reticulated patterns on the blue (*silenus/nigra*) and red (*sinica/arctoides*) γ correspond to the estimated probabilities of inheritance from the respective ancestral clades. (F) QuIBL results for the triplets. Relationships between internal branch length in coalescent units (C_2) and total proportion of introgressed loci, blue symbols represent triplets with true topology, orange symbols represent triplets with discordant topology. Total proportion of introgressed loci was obtained by multiplying the probability that each topology corresponds to introgression by its genomic frequency.

Figure 4 Scenario for the radiation of macaques. Macaques originated in Africa and after the split of *M. sylvanus* the ancestor of Asian macaques invaded Asia and initially split into three lineages, *silenus/nigra*, *arctoides/sinica* and *fascicularis/mulatta*. *fascicularis/mulatta* clade may have originated from an ancient hybrid speciation of the two parental lineages (the ancestral populations *silenus/nigra* and *arctoides/sinica* clades) at about 3.79 – 4.01 Mya ago.


A
M. mulatta
M. leucogenys
M. fuscata
M. sylvanus
M. assamensis
M. thibetana
M. arctoides
M. cyclopis
M. leonina
M. silenus
M. fascicularis
M. sinica
M. nemestrina
M. maura
M. tonkeana
M. nigra
B
mulatta group
fascicularis group
arctoides group
sinica group
silenus group
nigra group
sylvanus group
M.mulatta (CR)
M. mulatta (IR)
M. cyclopis
M. fuscata
M. fascicularis
M. arctoides (2)
M. arctoides
M. thibetana
M. assamensis
M. assamensis (2)
M. leucogenys
M. leucogenys (3)
M. leucogenys (2)
M. leucogenys (4)
M. sinica
M. silenus
M. leonina
M. nemestrina (2)
M. nemestrina
M. tonkeana (2)
M. tonkeana
M. maura
M. nigra
M. sylvanus
Papio anubis
3.0
C
q1=0.68
mulatta/fascicularis clade
arctoides/sinica clade
silenus/nigra clade
sylvanus clade
1.00
0.80
0.60
0.40
0.20
0.00
1
2
3
4
5
6
q1
q2
q3
q1=0.31
mulatta/fascicularis clade
silenus/nigra clade
arctoides/sinica clade
sylvanus clade
D
q1
q2
q3
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
18
19
chr20
E
Papio anubis
0
0.02
M. sylvanus
M. arctoides
M. arctoides (2)
M. sinica
M. assamensis (2)
M. assamensis
M. thibetana
M. leucogenys (3)
M. leucogenys (2)
M. leucogenys
M. leucogenys (4)
M. nigra
M. maura
M. tonkeana
M. tonkeana (2)
M. silenus
M. leonina
M. nemestrina
M. nemestrina (2)
M. fascicularis
M. fuscata
M. cyclopis
M. mulatta (IR)
M.mulatta (CR)
F
4.9e-6
0.75
0.50
0.25
Fst
within-species-group
between-species-group


M. mulatta
M. cyclopis
M. fuscata
M. fascicularis
M. thibetana
M. assamensis
M. leucogenys
M. sinica
M. arctoides
M. leonina
M. silenus
M. nemestrina
M. maura
M. tonkeana
M. nigra
M. sylvanus
Papio anubis
Pan troglodytes
Homo sapiens
Pongo abelii
1.86
2.34
2.94
3.79
1.71
2.17
2.56
3.01
4.01
1.39
2.39
5.29
3.2
1.88
2.27
7.76
20-38
6.21
16.69
Mya
20
18
16
14
12
10
8
6
4
2
0
Miocene
E
L
E
M
L
Pliocene
Pleistocene


A
Papio anubis
M. sylvanus
M. nigra
M. maura
M. tonkeana
M. tonkeana(2)
M. leonina
M. silenus
M. nemestrina
M. nemestrina(2)
M. fascicularis
M. fuscata
M. cyclopis
M. mulatta(CR)
M. mulatta(IR)
M. arctoides
M. arctoides(2)
M. sinica
M. leucogenys
M. leucogenys(2)
M. leucogenys(3)
M. leucogenys(4)
M. assamensis
M. assamensis(2)
M. thibetana
Top1 topology 24%
C
D
leucogenys
assamensis
thibetana
sinica
arctoides
fuscata
cyclopis
mulatta
fascicularis
nemestrina
silenus
leonina
nigra
tonkeana
maura
sylvanus
f_b
B
Papio anubis
M. sylvanus
M. nigra
M. maura
M. tonkeana
M. tonkeana(2)
M. leonina
M. silenus
M. nemestrina
M. nemestrina(2)
M. fascicularis
M. fuscata
M. cyclopis
M. mulatta(CR)
M. mulatta(IR)
M. arctoides
M. arctoides(2)
M. sinica
M. leucogenys
M. leucogenys(2)
M. leucogenys(3)
M. leucogenys(4)
M. assamensis
M. assamensis(2)
M. thibetana
Top2 topology 9.3%
E
M. sylvanus
M. nemestrina
M. leonina
M. silenus
M. nigra
M. maura
M. tonkeana
M. fascicularis
M. fuscata
M. cyclopis
M. mulatta
M. arctoides
M. sinica
M. leucogenys
M. thibetana
M. assamensis
F
high levels of ILS
nonsignificant
significant
minor topology
major topology
Proportion of topologies
0.25
0.50
0.75
Total Non−ILS proportion
C_2
Bootstrap Distribution of γ
Density
γ


Ancestor of different macaque group
~3.79 mya
Ancestor of Asian macaques
~4.01 mya
M. sylvanus
~5.29 mya
Ancestor
~7.76 mya