The bear bones: mosaic ancestry of *Macaca arctoides* explains divergent baculum morphology

Laurie S. Stevison[1,*], Zachary A. Szpiech[1], Nick P. Bailey[1], Taylor E. Novak[1], Don J. Melnick[2,†], Ben J. Evans[3], Jeffrey D. Wall[4]

[1]Department of Biological Sciences, Auburn University, Auburn, AL 36849, USA
[2]Department of Ecology, Evolution, and Environmental Biology, Columbia University, 10th floor Schermerhorn Extension, 119th Street and Amsterdam Avenue, New York,NY10027, USA
[3]Biology Department, Life Sciences Building Room 328, McMaster University, 1280 Main Street West, Hamilton, ON, Canada L8S4K1
[4]Institute for Human Genetics, University of California, San Francisco, California 94143

*Corresponding Author
†Deceased April 18, 2019

Abstract

Genital divergence contributes to reproductive barriers between species. Emergence of a novel accessory structure, the baculum, has independently evolved and been lost throughout mammalian evolution, purportedly driven by sexual selection. In primates, the longest recorded baculum belongs to *Macaca arctoides*, the bear macaque. This species has been proposed to be of homoploid hybrid origin via ancient hybridization between representatives from the fascicularis and sinica species groups. To investigate the evolutionary origins of the bear macaque and its unique morphology, we used whole genome sequences to quantify gene flow and phylogenetic relationships in 10 individuals from 5 species, including the bear macaque (n=3), and two species each from the sinica (n=3) and fascicularis (n=4) species groups. The results of these analyses were concordant, and identified 608 genes in the bear macaque that supported both clustering between *M. arctoides* and the sinica group (topo2) and had shared derived alleles between species from the two groups. Similarly, 361 genes supported both clustering between *M. arctoides* and the fascicularis group (topo3) and had shared derived alleles between both groups. Further, sliding window analysis of phylogenetic relationships revealed 53% of the genomic regions supported placement of *M. arctoides* in the sinica species group (topo2), 16% supported placement in the

fascicularis species group (topo3), and 11% supported *M. arctoides* in a grouping distinct from the sinica and fascicularis groups (topo1). Genomic regions with topo1 were intersected with previously identified QTL for mouse baculum morphology, and 47 genes were found, including five of sixteen major candidate loci that govern mouse baculum variation (KIF14, KIAA0586, RHOJ, TGM2, and DACT1). Although baculum morphology in the bear macaque is diverged from its parent taxa, it most closely resembles that of the fascicularis group. Outliers of shared ancestry from the fascicularis species group located within these same QTL regions overlap with the gene BMP4, which is an important component of the hedgehog signaling pathway that controls gonadogenesis. Two additional outlier genes (one shared with each species group) outside of the baculum QTL are known to interact with BMP4, suggesting this pathway may be involved in baculum morphology in primates. These results highlight how the mosaic ancestry of the bear macaque could explain its unique baculum evolution and collectively contribute to reproductive isolation.

Introductory Paragraph

In mammals, the baculum has extreme morphological variability, a dynamic evolutionary history characterized by repeated gain and loss, and is often used in species identification. The bear macaque has divergent genital morphology, including the longest baculum among all primates, and is proposed to have evolved via ancient hybrid speciation. Here, population genetic and phylogenomic approaches were used to examine how ancient hybridization in the bear macaque may have shaped this important component of genital morphology. Results demonstrate extensive mosaicism across the genome, which is consistent with ancient genetic contributions from both putative parental taxa. Genetic regions associated with baculum morphology also had mosaic ancestry for several genes, including KIF14 and KIAA0586, major candidate genes for baculum morphology in mice, and BMP4, a developmental gene involved in gonadogenesis. These results

have important implications for how hybridization may have shaped the evolution of reproductive isolation in this unusual species with complex speciation.

Introduction

For species with internal fertilization, male genital morphology has complex evolution and diverges rapidly[1–3]. The baculum, or penis bone, is a novel male genital structure in placental mammals that evolved ~145 million years ago[4]. Since that time, it has been the subject of recurrent adaptation having been gained or lost multiple times in various mammals, including recent loss in *Homo sapiens*[5]. Despite its remarkable interspecific variation, the penis bone generally does not vary extensively within species[6], and may be the only physical trait that differs consistently between closely related species, making it a key component in species identification[5]. However, the genetics of morphological variation in this novel trait are not well known. In 2016, a morphometric analysis using micro-CT scans of bacula in mice identified three major quantitative trait loci (QTLs), two that explained 50.6% of the variation in baculum size, and one that explained 23.4% of the variation in baculum shape[7]. These regions were further narrowed by differential gene expression in early development of male mice and potential involvement in bone or genital morphogenesis, revealing 16 major candidate loci (see Table S3 in ref[7]). Sequence divergence to rats reduced this list even further to four major candidate genes, all of which have orthologs in primates (PTPRC, KIAA0586, ASPM, and KIF14)[7].

Among primates, the bear macaque, *Macaca arctoides* has the longest baculum (4-6 cm[8]), even when corrected for body weight[8,9]. This species is perhaps the most unusual among macaques, distinguished not only by its extreme baculum length, but also by divergent genital morphology in both males and females (Figure S1), a unique single-mount ejaculatory mating behavior, whitish pelage in newborns, and a bald forehead and cheeks[8]. One possible explanation for the origin of the bear macaque baculum is that natural selection favored the evolution of prezygotic barriers to

76 hybridization via a mechanical "lock and key" mechanism[10]. Interestingly, however, this dramatic
77 derived genital morphology nonetheless may have been influenced by, or evolved despite,
78 extensive genetic exchange with other macaque species. Early molecular work revealed
79 phylogenetic incongruence between mitochondrial and Y-chromosomal genealogies for *M.*
80 *arctoides*[10,11]. Mitochondrial markers place *M. arctoides* as sister to the fascicularis species group,
81 whereas autosomal markers place it as sister to the sinica species group[12–15]. However, both
82 morphological and Y-chromosomal trees support placement in the sinica group[11,15,16] (see Table S1
83 in ref[14] for review). This incongruence was inferred to be consistent with ancient hybridization
84 between the ancestor of the fascicularis species group and a sinica species group member (an
85 ancestor of either *M. assamensis* or *M. thibetana*)[10]. This hypothesized hybrid origin is diagrammed
86 in Figure 1B.

87 In contrast to allopolyploid speciation, homoploid hybrid speciation (HHS) conserves
88 chromosome number and is expected to have similar occurrence frequencies in plants and
89 animals[17,18]. Though expected to be a rare speciation mechanism relative to "regular" bifurcating
90 speciation, diverse taxa such as birds (*Passer italiae*[19]), butterflies (*Heliconius heurippa*[20]) and yeast
91 (*Saccharomyces paradoxus*[21]) are thought to have evolved via HHS. Recent mitochondrial
92 phylogenetics have shown that *M. arctoides* is more closely related to *M. mulatta* as compared to *M.*
93 *fascicularis*[14,22], indicating introgression may have occurred after the split of these two species,
94 which is estimated to be 1.68 mya[14,15,23] (Figure 1B). This suggests an alternative evolutionary
95 scenario to the previously proposed model for the origin of *M. arctoides* by hybrid speciation; this
96 alternative is also presented in Figure 1B.

97 Recent analysis of macaque genomes support introgression between *M. arctoides* and
98 members of the fascicularis and sinica species groups[14], both of whom it overlaps with
99 geographically (Figure 1C). The baculum in *M. arctoides* is more than double the average length of
100 sinica species group taxa and nearly four times as long as fascicularis species group taxa[9,24]. These

101 sister species groups are estimated to have split ~3 million years ago, with hybridization to
102 produce *M. arctoides* purportedly occurring ~1.8 million years ago[14], during which glacial periods
103 during the Pleistocene were associated with reduced expanse of Southeast Asian forests into
104 multiple refugia. Ecological isolation is a common feature to other scenarios of HHS[25], and may have
105 contributed to the extensive interspecific hybridization that gave rise to the unique macaque
106 species, *M. arctoides*. Y-chromosomal data suggests that interbreeding between ancestors of the
107 fascicularis and sinica groups was driven by sinica males. Available evidence suggests that this
108 interbreeding subsided ~1.5 mya[10], followed by little interbreeding until the present[8].
109 Here, we investigated the role hybridization played in shaping the unusual genital
110 morphology of *M. arctoides*. Because a major criteria for HHS[26,27] is that reproductive isolation (RI)
111 be a direct byproduct of hybridization (but see[28]), we focused especially on genetic regions that are
112 associated with morphological variation and candidate genes in mice[5]. Phylogenomic and
113 population genetic approaches were used on several extant macaque species (*M. arctoides;* the
114 sinica group species *M. thibetana* and *M. assamensis*; and the fascicularis group species *M.*
115 *fascicularis* and *M. mulatta*). Also, outlier regions were examined within the bear macaque genome
116 with alternate alleles from the two parental genomes and regions where the bear macaque differs
117 from both parents. Despite its divergent morphology[8,9,29], the baculum of *M. arctoides* is more
118 similar to fascicularis species[8], making these outliers of particular interest.

119 **Results**

120 Genomic analysis of five macaque species revealed a mosaic of evolutionary ancestry with
121 respect to both putative parental species; this generated novel allelic combinations at multiple loci
122 throughout the genome, including several genes associated with baculum morphology. These
123 analyses used newly sequenced whole genome samples from *M. arctoides* (20X coverage) and *M.*
124 *assamensis* (13X), and 8 publicly available genomes ranging in coverage from 4-49X (median 32X;

125 Table 1). While not directly investigated here, these data will be useful for addressing questions
126 related to the putative hybrid origin of this species in combination with targeted simulations (see
127 Discussion).

128 Of the heterozygous sites, the majority are unique to each species group (Figure 1D). *M.*
129 *arctoides* shares the most variants with the sinica species group, *M. assamensis* and *M. thibetana*,
130 consistent with taxonomic placement in this species group[30]. However, the two-way intersection
131 between *M. arctoides* and *M. thibetana* had the fewest number of shared heterozygous sites, which
132 is likely due to the low overall heterozygosity in the *M. thibetana* sample (Figure 1D; Table 1). *M.*
133 *assamensis* had the largest number of heterozygous sites (~4.7 million; Table 1; Figure 1D). This
134 high variability is consistent with the demographic results from smc++, which indicated population
135 size expansion around 10,000 years ago in *M. assamensis* but not *M. thibetana* and *M. arctoides*
136 (Figure 1E). Consistent with this, PSMC analysis in a previous study showed *M. assamensis* has a
137 higher overall effective population size beginning ~250 kya compared to *M. arctoides* and other
138 sinica species (see Figure 3B in[14]).

139 ***Extensive Mosaicism of the M. arctoides Genome***

140 The four-taxon test[31–33] results supported that the *M. arctoides* genome has a mosaic of
141 shared ancestry with both the sinica and fascicularis species groups (Figure 2), which is consistent
142 with extensive introgression from both species groups. Though it is worth noting that windows
143 with shared ancestry could also be due to incomplete lineage sorting between these taxa[23]. The
144 genome-wide average f_{dM}[34,35] value was -0.113, ranging from -0.69 (supporting shared ancestry
145 with the sinica group) and 0.55 (supporting shared ancestry with the fascicularis group) in 50kb
146 sliding windows. This sliding window analysis was repeated with several window sizes (Figure S2),
147 and all had a similar genome-wide mean f_{dM}. In a previous study, estimates of *D* were similar to the
148 f_{dM} estimates here (see Table S8 in[14]).

149 Comparison of f_{dM} in windows on the autosomes and on the X-chromosome revealed the X-
150 chromosome to have significantly lower mean f_{dM} estimate (-0.15), with a statistical difference
151 between the two distributions (Figure S3). There was a non-significant negative relationship
152 between chromosome length and f_{dM}, but the X-chromosome has an intermediate length (149 Mb)
153 compared to the autosomes (mean=134 Mb). Thus, the lower f_{dM}estimate on the X is not likely
154 attributed to the length of the chromosome.

155 Phylogenetic relationships among the species groups were evaluated via the software
156 Twisst[36] (topology weighting by iterative sampling of subtrees). This method quantifies
157 phylogenetic relationships across the genome and returns weights in sliding windows (Figure 1F).
158 The topology weights were used to define a "majority topology" as having two-thirds or more of the
159 sum of weights in any region. Windows that did not meet this criterion were re-classified as
160 "unresolved". This analysis showed the proportion of the genome that supported topo2, which
161 groups *M. arctoides* with the sinica group, as 52.64% (95% bootstrap CI [52.02%, 53.28%]).
162 However, a significant portion, 15.70% (95% bootstrap CI [15.40%, 16.01%]), of the genome
163 supported topo3, which groups *M. arctoides* with the fascicularis species group. Interestingly,
164 11.07% (95% bootstrap CI [10.84%, 11.31%]) of the genome supported *M. arctoides* clustering in a
165 group by itself (topo1). The remaining 20.59% (95% bootstrap CI [20.18%, 21.01%]) of the genome
166 was categorized as unresolved. Still, the 15.70% of the genome that supports topo3 is a rather large
167 fraction of the genome considering this is an ancient hybridization event. Similar to our four-taxon
168 results, this distribution of support for each of the three possible topologies is different on the X
169 chromosome, with an excess of topo2 on the X (70.89%, 95% bootstrap CI [67.49%, 74.18%]). One
170 possible explanation for this difference relates to incomplete lineage sorting which is expected to be
171 lower on the X-chromosome as compared to the autosomes, owing to the lower effective population
172 size of the X.

173 In addition to examining broader patterns of shared ancestry in this species, outliers of

174 shared ancestry with each species group and the *M. arctoides* genome were specifically scrutinized
175 (see below). This set of regions were intersected with the Ensembl gene predictions downloaded
176 from University of California Santa Cruz (UCSC) Genome Browser, resulting in 371 genes with the
177 highest ancestry from the fascicularis group, and 608 genes with the highest ancestry from the
178 sinica group. These results were highly concordant with the majority topology for each of the two
179 possible groupings between *M. arctoides* and the parental taxa. The intersection resulted in no
180 reduction of the 608 sinica-shared genes, whereas the fascicularis-shared genes were reduced to
181 363 from 371. This set of 971 genes were uploaded to the Panther Database. The enrichment gene
182 terms for GO Biological Processes revealed three GO terms enriched from fascicularis species group
183 and 24 from sinica species group (Table S5). A surprising result was that there were three
184 neuronal-related gene categories for ancestry shared between sinica and *M. arctoides* (see below
185 and Discussion).

186 ***Intersection of Regions Associated with Baculum Morphology with Topo1 Regions***

187 Next, the phylogenomic results were intersected with the mouse baculum QTL to investigate
188 the genetic ancestry of the divergent baculum in *M. arctoides*. First, the 11.07% of topo1 regions that
189 support placement of *M. arctoides* into its own lineage that does not share recent ancestry with either
190 the fascicularis or sinica groups were investigated further. To accomplish this, the stringency in
191 defining the majority topology was increased to require 100% of the weights into topo1, which
192 reduced this to 4.5% of the genome. The regions from the Twisst analysis were extracted and
193 intersected with the known Ensembl gene annotations, resulting in 4175 genes. A GO Analysis of
194 these genes revealed enrichment of gene categories involved in developmental processes, immune
195 system processes, and metabolic processes (Figure 4). There were also gene categories involved in
196 reproduction, reproductive processes, behavior, and pigmentation, which have particular
197 importance in *M. arctoides* given its phenotypic changes in these traits.

198 This gene list was intersected with homologous chromosome regions in rheMac8 to the
199 three major mouse baculum QTL, leaving 47 genes. Five of these genes (TGM2, KIF14, RHOJ, DACT1,
200 and KIAA0586) were listed in Table S3 of ref[7] as the 16 major candidate loci that overlap QTL
201 regions for baculum morphology in mice. Two of the five genes, KIF14 and KIAA0586, also had a
202 high sequence divergence between mouse and rat, thus being considered as two of four major
203 candidate loci in explaining the genetic basis of mouse baculum size variation[7]. Similarly,
204 divergence between *M. arctoides* and the sinica and fascicularis species group members was also
205 high for KIF14 (Sin-Arc F_{ST} = 0.45; Fas-Arc F_{ST} = 0.61) and KIAA0586 (Sin-Arc F_{ST} = 0.38; Fas-Arc F_{ST}
206 = 0.52) compared to the genome-wide averages (Sin-Arc F_{ST} = 0.33; Fas-Arc F_{ST} = 0.43). Ten of the
207 47 genes also had missense mutations between the mouse strains used for the QTL mapping (see
208 Table S2 in ref[7]). Two of the 10 genes with missense mutations also had high F_{ST}: DDX59 (Sin-Arc
209 F_{ST} = 0.40; Fas-Arc F_{ST} = 0.52) and SOGA1 (Sin-Arc F_{ST} = 0.38; Fas-Arc F_{ST} = 0.51).
210 For the two top candidate genes, KIF14 and KIAA0586, *M. arctoides* differed from all other
211 lineages at 15 and 13 nucleotide positions, respectively. *M. arctoides* had two fixed and six
212 heterozygous non-synonymous mutations in the protein-coding region of KIAA0586 and one fixed
213 and one heterozygous non-synonymous mutation in KIF14. Neither of the fixed differences were
214 located in any known functional domains. One of the other five genes (DACT1) had a premature
215 stop codon fixed in *M. arctoides*; however, the impact to overall function is unclear (but see
216 Discussion).
217 For fixed non-coding variants and variants immediately upstream or downstream of each
218 gene, 100bp regions surrounding variants for each allele were input into the online tool TomTom[38]
219 to identify potential binding sites for transcription factors (Table S8). One of these SNPs suggested
220 that CTCF binds to the *M. arctoides* allele of KIF14, but not any of the others (Figure 3A). The
221 ENCODE Transcription Factor Target dataset as accessed through Harmonizone[39] suggests that
222 CTCF is a known transcription factor for KIF14.

223 ***Gene Flow in Regions Associated with Baculum Morphology***

224 To understand how the mosaic ancestry of the *M. arctoides* genome contributed to its
225 extreme baculum morphology and subsequent RI, the outliers of shared ancestry were also
226 intersected with regions in rheMac8 homologous to the mice baculum QTLs. ANOVA indicated the
227 distribution of f_{dM} scores in these regions were significantly different from the rest of the genome.
228 Surprisingly, these regions were more negative, indicating more shared ancestry with sinica
229 member species than fascicularis group species (Figure S4), which is the opposite of the
230 expectation based on morphology.

231 Because the baculum of *M. arctoides* is more similar to members of the fascicularis species
232 group than the sinica group, the fascicularis-shared introgression outlier regions (shown in blue in
233 Figure S4) were further examined. Of 363 genes identified genome-wide, three genes were located
234 within one of the three QTL regions on chromosome 7 (Figure S5; Supplementary Methods and
235 Results). One of these genes, *Bone Morphogenetic Protein 4* (BMP4), is a key gene in the hedgehog
236 signaling pathway[40], and has a role in gonadal development[41,42]. As expected for a major
237 developmental gene, no coding differences were observed. However, there were two variants in the
238 3' UTR of BMP4, one of which matched topo3, where *M. arctoides* and fascicularis species had a
239 shared allele (Figure 3C; Figure S6A). Using TomTom as described above, differential binding was
240 predicted for variants at each SNP which identified three genes known to interact with BMP4: Six1,
241 Tead1, and Tead3. Six1 deficiencies lead to reduced BMP4 expression[43] and affect spatial
242 expression patterns[44]. Similarly, Tead proteins are a family of transcription factors that mediate
243 transcription and cell migration of BMP4 via the TAZ component of the Hippo pathway[45]. Based on
244 the alleles present, the lack of Six1 binding and the added binding of Tead1 and Tead3 would
245 predict an increase in expression of BMP4 in *M. arctoides* and fascicularis group species as
246 compared to sinica group species.

247 ***Characterization of Outlier Genes Explain Other Phenotypes Unique to M. arctoides***

248 As mentioned earlier, *M. arctoides* has other traits (e.g. ontogenetic coloration, prolonged
249 intromission with a postejaculatory "pair-sit", etc[8].) that make it unusual among macaques. Of the
250 971 genes identified as shared ancestry outliers, a stringent set of 146 genes, 51 fascicularis and 95
251 sinica origin genes were identified (see Supplementary Methods and Results) and manually
252 curated. The seven manually curated categories had the following number of genes: 11 bone, 12
253 reproductive, 7 sensory, 8 immune and 6 skin/fur. These highly represented groups had: 55 cellular
254 functions and 55 neural functions (Table S6). Some genes were in more than one category. The two
255 highly represented groups corresponded well with the prevalence of neuronal-related gene
256 categories from the GO analysis for sinica outliers. One possibility for this neural enrichment is that
257 interactions between the fascicularis-shared mtDNA and the mostly sinica-shared nucleus may
258 have led to mitonuclear incompatibilities between these species (see Discussion).

259 A moderate number of bone, reproductive, and skin/fur genes is unsurprising based on the
260 unique biology of *M. arctoides*. Notably, two known interactors with BMP4 (DYNC2LI1 and
261 CTDSPL2) were included in this stringent gene list. DYNC2LI1 exhibited shared ancestry between
262 *M. arctoides* and fascicularis (f_{dM} = 0.79) whereas CTDSPL2 exhibited shared ancestry between *M.*
263 *arctoides* and sinica (f_{dM} = -0.82). In humans, DYNC2LI1 mutations are associated with lethal
264 skeletal diseases and have been found to impair the Hedgehog pathway[46]. CTDSPL2 phosphorylates
265 various Smad proteins, which are transcriptional regulators directly involved with the BMP
266 pathway. CTDSPL2 knock-down experiments in mice lead to increased expression of BMP targeted
267 genes[47].

268 TomTom was used to examine putative binding motifs associated with the differing alleles
269 at SNPs within CTDSPL2 and DYNC2LI1. Within CTDSPL2, *M. arctoides*/sinica and the
270 fascicularis/outgroup had 16 fixed differences between them and 5 heterozygous sites. Of these,
271 only one transcription factor seemed to be directly related to regulation of CTDSPL2. Specifically,

272 overexpression of XBP1 in mouse cells up-regulates CTDSPL24[48]. TomTom results suggested that
273 XBP1 binds to the shared *M. arctoides*/sinica allele of CTDSPL2 but not the shared
274 fascicularis/outgroup allele (Figure 3B). The DYNC2LI1 gene had 12 fixed differences between *M.*
275 *arctoides*/fascicularis and the sinica/outgroup, one resulting in a non-synonymous mutation (c.
276 422G>A; p. Arg>Lys141). TomTom results suggest that this SNP changes the binding affinity of
277 RBFOX2 protein, which regulates splicing events[49]. TomTom predicted binding of RBFOX2 to the
278 allele shared by *M. arctoides* and fascicularis species (Figure 3C). See Table S8 for detailed
279 characterization of TomTom results for each SNP.

280 **Discussion**

281 ***A patchwork of ancestral origins for the baculum in the bear macaque***

282 The gene analysis of alleles unique to the bear macaque lineage revealed 47 genes that
283 overlapped three QTL regions associated with baculum morphology in mice. Two of these genes are
284 major candidate genes for baculum size, KIF14 and KIAA0586[7], the latter of which has several
285 nonsynonymous mutations. Three other genes were on a longer list of 16 candidate genes, one of
286 which (DACT1) had a nonsense mutation c. 1317G>A; p. Trp439* fixed in *M. arctoides*. DACT1 is
287 part of the Wnt signaling pathway, interacts with *dishevelled* (DVL), and is associated with Townes-
288 Brocks syndrome 2 (TBS2; MIM# 617466). Interestingly, a heterozygous nonsense mutation c.
289 1256G>A; p. Trp419* in DACT1 in humans has been shown to have a variety of phenotypes
290 including genitourinary malformations. These malformations include hypospadias[50], a birth defect
291 where the urethral opening is on the underside of the penis, which is also characteristic of *M.*
292 *arctoides* male genital morphology[8].

293 While there is a mosaic of ancestries across the genome of *M. arctoides*, there is more
294 support for sinica group ancestry, consistent with previous results[10,14,15]. Interestingly, baculum
295 morphology, which is divergent in both size and shape in *M. arctoides*[9,29], does not support

296 clustering of the bear macaque with the sinica species group, and instead morphologically
297 resembles members of the fascicularis species group[8]. Consistent with these morphological
298 similarities, one of 363 fascicularis-sourced introgression outliers overlaps with a QTL associated
299 with variation in baculum size in mice. This gene, BMP4, is a major regulator of bone formation and
300 gonad development[40,42]. Two SNPs in the 3' UTR have alternative binding motif alleles that suggest
301 regulatory changes in this gene that would result in upregulation of BMP4 in *M. arctoides* as
302 compared to sinica species (Figure 6C). Still, functional validation is needed to ascertain whether
303 this genomic variation is coupled with differences in expression intensity, location, or timing during
304 development.

305 Additionally, among the 47 genes overlapping the mouse QTLs from topo1, one gene with
306 notably high F_{ST} was the gene *BMP/Retinoic Acid Inducible Neural Specific 3* (BRINP3; Sin-Arc F_{ST} =
307 0.48, Fas-Arc F_{ST} = 0.60). There were two other genes in the outlier analysis that are known to
308 interact with BMP4 and bone development, pointing to possible rewiring of this important
309 developmental pathway in the *M. arctoides* genome. Although this pathway was not identified in the
310 mouse baculum morphology study, it was mentioned as a likely candidate worthy of investigation.
311 It is also possible that this pathway is involved in baculum morphology in primates, but not in mice.
312 Future work should focus on these genes to understand the evolution of baculum morphology in *M.*
313 *arctoides*.

314 ***Complex speciation of the bear macaque***

315 Macaques are a diverse primate genus with several examples of purported complex
316 speciation[10,11,50]. Although mtDNA suggests that introgression with the fascicularis group should be
317 driven by *M. mulatta*[14,22], when considering these two species separately, the autosomal estimates
318 of f_{dM} are largely similar (Table S3). One possible explanation for this is that the sample used for *M.*
319 *fascicularis* is Vietnamese in origin and therefore has extensive gene flow from *M. mulatta*[51]. Though

320 it is worth noting that there are 1.6 times more shared heterozygous sites between *M. arctoides* and
321 *M. mulatta* than between *M. arctoides* and *M. fascicularis* (Figure 1D). We also found 11.07% of
322 genomic regions support a clustering of *M. arctoides* separate from other macaque species.
323 Although several taxonomic surveys have suggested the placement of *M. arctoides* into its own
324 species group[30,52,53], recent genome surveys suggest that sequencing of additional sinica species
325 members will be useful in making this determination[54,55].

326 While genomic mosaicism is a key feature of hybrid species, it is also a ubiquitous feature of
327 admixture more generally. For example, in swordtail fish, there is genomic mosaicism that resembles
328 hybrid speciation, although it is not consistent with other information from this species[26,27,56].
329 Similarly, up to 15% of the genome supports clustering of *M. arctoides* with fascicularis species. This
330 finding, though compelling, is also not uncommon in other taxa that do not have hybrid speciation.
331 For example, in human-chimp-gorilla comparisons, ~30% of the human genome does not support
332 the species tree[57]. In principle, an even contribution from both parents to a hybrid genome is possible,
333 but this is often not the case in HHS (e.g., Heliconius[20]). Here, we observed a significant deviation
334 from a 50-50 split of ancestry from the two groups, which could be due to a combination of proposed
335 recent admixture with sinica group taxa[10] and differences in ancestral populations sizes of the
336 putative parental taxa.

337 Here, we also found higher shared ancestry between *M. arctoides* and sinica species group
338 taxa on the X-chromosome as compared to the autosomes, which is consistent with either (1)
339 higher introgression from the sinica group on this chromosome compared to the autosomes, (2)
340 lower effective population size of the X-chromosome compared to the autosomes and/or (3)
341 increased selection against introgression on the X chromosome (which is qualitatively consistent
342 with Haldane's rule). The latter possibility is interesting in light of a recently proposed
343 mitochondrial capture scenario via reduced fitness of male hybrids[14]. Mitochondrial capture could
344 either have followed *M. arctoides* split from sinica, or it could have been simultaneous with an HHS

345 origin (Figure 1B). While there are recent statistics that explicitly test for HHS[58], these have a
346 restrictive set of assumptions that may not apply to this system (e.g., subsequent gene flow with
347 parental taxa). Therefore, a more explicit set of simulations of various speciation scenarios would
348 need to be conducted to determine the timing of ancient introgression from the two parental
349 genomes.

350 ***Unanticipated enrichment of introgression outliers with neurological function***

351 GO terms in the introgression outliers from sinica and *M. arctoides* origin analysis revealed a
352 large number of genes involved in neurological function that were noticeably absent in the
353 fascicularis outliers. A possible explanation is that the mismatch in sinica-sourced mitonuclear genes
354 could lead to mitochondrial dysfunction, which could contribute to neural malfunction and disease.
355 For example, disease-causing mutations in mitochondrial aminoacyl tRNA-synthetases in humans
356 predominantly result in neurological disorders[59]. This is believed to be due to the generally high-
357 energy demand of primate brains and may necessitate coevolution of mitonuclear and mitochondrial
358 genes[60,61]. Moreover, at least two of the sinica-sourced neurological genes characterized here, GFM2
359 and SOD2, are known to be involved in mitochondrial disorders exhibiting neurological effects[62,63].
360 Additionally, hybridization scenarios exhibiting mismatch between mitochondrial and mitonuclear
361 genes have been discovered in a variety of systems[64–66]. A comprehensive analysis of the potential for
362 mitonuclear incompatibilities in the evolution of *M. arctoides* could also shed light on the evolution
363 of the bear macaque.

364 ***Future Directions***

365 This study has focused on the baculum based on its extreme length in *M. arctoides*, which is
366 proposed to play a role in reproductive isolation among species However, we recognize that the
367 baculum is certainly not the only barrier to reproduction between *M. arctoides* and its parental taxa.
368 We anticipate there are other regions of the genome that have contributed to speciation[67]. For

369 example, there have been compensatory changes in female reproductive morphology[29] that are
370 equally important in the mechanical barrier between these taxa (Figure S1). Additionally, changes
371 in their mating behavior also contribute to differences between *M. arctoides* and its parent taxa[9].
372 Therefore, as the genetic basis of such traits become known, they can be further investigated.

373 In addition to this study, there has been extensive interest in the evolution of this species.
374 However, due to the limited number of publicly available whole genome samples (now three with the
375 present study), the conservation status of this species, and the ethical and logistical barriers to
376 experimental research with primates, it is challenging to make functional insights into its evolution.
377 Additional WGS data from more individuals of this species would allow for a better characterization
378 of selection/adaptation within this lineage. Additionally, more WGS samples would be useful in
379 examining the distribution of introgression haplotype lengths to pinpoint the timing of events in this
380 system that would better clarify when hybridization from each parent took place.

381 **Methods**

382 ***Samples and Genome Sequencing***

383 One female each from *M. arctoides* (Malaya) and *M. assamensis* (A20) were sequenced.
384 These two samples were multiplexed and sequenced on one lane of sequencing for initial quality
385 assessment. They were then subsequently run on one additional lane of sequencing each for a final
386 1.5 lane per sample. Sequencing was done at UCSF Medical Center Sequencing Facility on a Illumina
387 HiSeq 2000 machine. Each run was assigned as different read groups to be treated as independent
388 in the variant calling workflow. The raw read data have been deposited in NCBI sequence read
389 archive and are available via the accession PRJNA622565.

390 ***Publicly available sequences***

391 Raw fastq files were downloaded from eight public genome samples for this project (Table

392 1). Additional sample details can be found in the corresponding publications. These were each
393 downloaded from NCBI SRA using sratoolkit version 2.8.1 with options to split data into read pairs
394 and in the original format[68]. Because *M. thibetana* was sequenced on an older platform, we used
395 seqtk[69] to convert from Q64 to Q33. Additionally, SRA files from multiple lanes as indicated in the
396 fastq header were split for independent alignment to the reference genome facilitating the
397 definition of separate read groups to be treated as independent in the variant calling workflow.

398 ***Alignment, genome analysis, and variant calling***

399 Raw fastq files from all samples were aligned to the reference genome rheMac8 (NCBI
400 Accession: PRJNA214746). The masked reference genome was downloaded from UCSC. GATK best
401 practices were followed to obtain high quality variant sites (see Supplementary Methods). Briefly,
402 reads were aligned to the reference genome, duplicates were marked, indels were realigned, base
403 quality scores were recalibrated, variants were called, and variant quality scores were recalibrated.
404 Samples from each species were processed independently and merged into a final callset. The
405 baboon reference genome was added to the variant files via two-way genome alignment and
406 custom scripts. These final filtered files for each chromosome with the baboon information are
407 freely available and were used for all subsequent analyses.

408 ***Four-taxon test for introgression***

409 To analyze patterns of introgression in sliding windows, a modified four-taxon test was
410 used (Figure 2). Rather than separately analyzing gene flow between *M. arctoides* and each parental
411 species group, as has been done previously[14], we relied on the modified statistic f_{dM}, which is better
412 at comparing introgression from both groups[35]. Specifically, hybridization between *M. arctoides*
413 (P3) was tested with members of both the sinica species group (P1) and the fascicularis species
414 group (P2), using Baboon as an outgroup. Calculations were done using scripts available on
415 github[33,34]. To select the appropriate window size for the analysis, a series of different sized

416 windowed analyses were conducted. f_{dM} was computed for overlapping windows of size 5kb, 25kb,
417 50kb, 100kb, 500kb and 1000kb, with a step size of 20% of each window size. The minimum
418 number of sites per window size was based on the distribution of sites available in the callset at
419 each bin size. Specifically, a minimum number of sites of 10, 50, 100, 200, 1000, and 2000,
420 respectively, per window size were used. Because it is recommended to have at least 100 sites per
421 window[34], 50kb sliding windows were selected and are shown in Figure 2 as a heatmap across the
422 genome. Results from the additional bin sizes are shown in Figure S2. This analysis was repeated
423 with individual taxa from each species group (Table S3).

424 ***Topology weighting***

425 To analyze the weights of different phylogenetic topologies across the genome, the software
426 Twisst was used to compare the weights of different species topologies throughout the genome[36,70].
427 First, data used for the introgression analysis were pruned to make sure that each group had at
428 least one member represented in the genotype data. Next, neighbor joining trees were constructed
429 in sliding windows of 50 SNPs each across the genome using Phyml from scripts available on
430 github[34]. This generated newick formatted tree files that were used as input for Twisst.
431 The topology weighted outputs were further explored by collapsing adjacent intervals with
432 the same majority topology supported. Since there were three possible topologies, the majority
433 topology was defined as any one topology that had at least two-thirds of the total weight values.
434 Any intervals without a majority topology were labeled as "unresolved". If adjacent intervals
435 supported the same majority topology, they were collapsed and split into the three major
436 topologies. These were intersected with gene regions from the Ensembl annotations for rheMac8
437 downloaded from UCSC.
438 Bootstrap confidence intervals for topo1, topo2, topo3, and unresolved proportions were
439 computed by resampling 2,792 1Mbp intervals along the genome (149 when considering only
440 chromosome X) for each of 10,000 replicates and taking the 2.5% and 97.5% quantiles.

441 ***Demography Analysis using smc++***

442 The method smc++[71] was used, which uses whole genome sequencing data across multiple
443 samples and does not require phased data. All sequences were masked at the start of a
444 chromosome up to the first genotype call, at the end of a chromosome from the last genotype call,
445 and all rheMac8 assembly gaps. Then, the estimate module was used to estimate population size
446 history with an assumed mutation rate of 2.5e-8[14] and automatic selection of inference timepoints.
447 When generating population size trajectory plots, a generation time of 6 years[14] was used.

448 ***Functional Analysis of Targeted and Outlier Regions***

449 A major criteria of homoploid hybrid speciation is evidence that RI evolved as a by-product
450 of hybridization[26]. To investigate this directly, regions associated with morphological variation in
451 the baculum identified in mice[7] were targeted. The UCSC tool liftover[72] was used to convert
452 coordinates of the three QTL from the mouse study from mm10 to rheMac8.

453 To further identify outlier regions, loci with the highest shared ancestry from each parent
454 taxa were extracted from the sliding window results (see Supplementary Methods). These regions
455 were further intersected with gene lists from the Twisst analysis. Specifically, sinica outliers of
456 shared ancestry were intersected with regions where the majority topologies supported *M.*
457 *arctoides* clustering with sinica species. Similarly, fascicularis outliers of shared ancestry were
458 intersected with regions where the majority topologies supported *M. arctoides* clustering with
459 fascicularis species. These intersected gene lists (Table S4) were used to conduct a GO Analysis
460 using Panther[73] (Table S5).

461 **Acknowledgements**

462 This work was supported by research start-up funds from the Department of Biological
463 Sciences at Auburn University (LSS). Sequencing costs were supported in part by NIH grant R01

464 GM115433 (to JDW). This work was made possible in part by a grant of high performance
465 computing resources and technical support from the Alabama Supercomputer Authority. BJE was
466 supported by a grant from the Natural Science and Engineering Research Council of Canada
467 (RGPIN-2017-05770). We thank members of the Stevison lab for helpful discussions regarding
468 the analysis and results, with particular thanks to Stephen Sefick. We thank Simon Martin both for
469 having a public github repository of high quality, but also for responsiveness regarding inquiries in
470 use of his code. We thank several researchers in attendance at the 2016 Speciation GRC, Evolution
471 2017, and PEQG 2018 for helpful discussions regarding this work. We thank Wendy Hood, Ken
472 Halanych, and Geoff Hill for feedback on early drafts of the manuscript.

473 **Author Contributions**

474 LSS, JDW, DJM, and BJE conceived of the project. DJM supplied the samples and JDW paid for
475 sequencing. LSS executed methods/analyses with substantial input from JDW and BJE, and wrote
476 the draft manuscript. BJE and DJM significantly aided the interpretation of results. ZAS conducted
477 VQSR and SMC++ analyses. NPB and TEN conducted manual curation of gene outliers and TomTom
478 on baculum alleles. All authors were involved in editing and revision of the final draft.

479 **References**

480 1. Langerhans, R. B., Anderson, C. M. & Heinen-Kay, J. L. Causes and Consequences of Genital
481 Evolution. *Integr. Comp. Biol.* **56**, 741–751 (2016).
482 2. Klaczko, J., Ingram, T. & Losos, J. Genitals evolve faster than other traits in Anolis lizards. *J. Zool.*
483 **295**, 44–48 (2015).
484 3. Eberhard, W. G. Evaluating Models of Sexual Selection: Genitalia as a Test Case. *Am. Nat.* **142**,
485 564–571 (1993).
486 4. Brindle, M. & Opie, C. Postcopulatory sexual selection influences baculum evolution in primates
487 and carnivores. *Proc. R. Soc. B Biol. Sci.* **283**, 20161736 (2016).

488 5. Carosi, M. & Scalici, M. Baculum (Os Penis). in *The International Encyclopedia of Primatology* 1–5
489 (American Cancer Society, 2017). doi:10.1002/9781119179313.wbprim0214.
490 6. Ramm, S. A., Khoo, L. & Stockley, P. Sexual selection and the rodent baculum: an intraspecific
491 study in the house mouse (Mus musculus domesticus). *Genetica* **138**, 129 (2009).
492 7. Schultz, N. G. *et al.* The Genetic Basis of Baculum Size and Shape Variation in Mice. *G3*
493 *GenesGenomesGenetics* **6**, 1141–1151 (2016).
494 8. Fooden, J. The bear macaque,Macaca arctoides: a systematic review. *J. Hum. Evol.* **19**, 607–686
495 (1990).
496 9. Dixson, A. F. Observations on the evolution of the genitalia and copulatory behaviour in male
497 primates. *J. Zool.* **213**, 423–443 (1987).
498 10. Tosi, A. J., Morales, J. C. & Melnick, D. J. Paternal, maternal, and biparental molecular markers
499 provide unique windows onto the evolutionary history of macaque monkeys. *Evol. Int J Org*
500 *Evol.* **57**, 1419–35 (2003).
501 11. Tosi, A. J., Morales, J. C. & Melnick, D. J. Comparison of Y Chromosome and mtDNA Phylogenies
502 Leads to Unique Inferences of Macaque Evolutionary History. *Mol. Phylogenet. Evol.* **17**, 133–
503 144 (2000).
504 12. Li, J. *et al.* Comparative Genome-Wide Survey of Single Nucleotide Variation Uncovers the
505 Genetic Diversity and Potential Biomedical Applications among Six Macaca Species. *Int. J. Mol.*
506 *Sci.* **19**, (2018).
507 13. Li, J. *et al.* Phylogeny of the macaques (Cercopithecidae: Macaca) based on Alu elements. *Gene*
508 **448**, 242–249 (2009).
509 14. Fan, Z. *et al.* Ancient hybridization and admixture in macaques (genus Macaca) inferred from
510 whole genome sequences. *Mol. Phylogenet. Evol.* **127**, 376–386 (2018).
511 15. Jiang, J. *et al.* Mitochondrial Genome and Nuclear Markers Provide New Insight into the
512 Evolutionary History of Macaques. *PLoS ONE* **11**, (2016).

513 16. Fan, P. *et al.* Phylogenetic position of the white-cheeked macaque (Macaca leucogenys), a newly
514 described primate from southeastern Tibet. *Mol. Phylogenet. Evol.* **107**, 80–89 (2017).
515 17. Mallet, J. Hybrid speciation. *Nature* **446**, 279–283 (2007).
516 18. Mavárez, J. & Linares, M. Homoploid hybrid speciation in animals. *Mol. Ecol.* **17**, 4181–4185
517 (2008).
518 19. Elgvin, T. O. *et al.* The genomic mosaicism of hybrid speciation. *Sci. Adv.* **3**, e1602996 (2017).
519 20. Jiggins, C. D., Salazar, C., Linares, M. & Mavarez, J. Hybrid trait speciation and Heliconius
520 butterflies. *Philos. Trans. R. Soc. B Biol. Sci.* **363**, 3047–3054 (2008).
521 21. Leducq, J.-B. *et al.* Speciation driven by hybridization and chromosomal plasticity in a wild
522 yeast. *Nat. Microbiol.* **1**, 15003 (2016).
523 22. Roos, C., Kothe, M., Alba, D. M., Delson, E. & Zinner, D. The radiation of macaques out of Africa:
524 Evidence from mitogenome divergence times and the fossil record. *J. Hum. Evol.* **133**, 114–132
525 (2019).
526 23. Stevison, L. S. & Kohn, M. H. Divergence population genetic analysis of hybridization between
527 rhesus and cynomolgus macaques. *Mol. Ecol.* **18**, 2457–2475 (2009).
528 24. Fooden, J. Male external genitalia and systematic relationships of the Japanese macaque
529 (Macaca fuscata Blyth, 1875). *Primates* **12**, 305–311 (1971).
530 25. Gompert, Z., Fordyce, J. A., Forister, M. L., Shapiro, A. M. & Nice, C. C. Homoploid Hybrid
531 Speciation in an Extreme Habitat. *Science* **314**, 1923–1925 (2006).
532 26. Schumer, M., Rosenthal, G. G. & Andolfatto, P. How Common Is Homoploid Hybrid Speciation?
533 *Evolution* **68**, 1553–1560 (2014).
534 27. Schumer, M., Rosenthal, G. G. & Andolfatto, P. What do we mean when we talk about hybrid
535 speciation? *Heredity* **120**, 379 (2018).
536 28. Nieto Feliner, G. *et al.* Is homoploid hybrid speciation that rare? An empiricist's view. *Heredity*
537 **118**, 513–516 (2017).

538 29. Fooden, J. Complementary Specialization of Male and Female Reproductive Structures in the
539 Bear Macaque, Macaca arctoides. *Nature* **214**, 939–941 (1967).
540 30. Fooden, J. Classification and Distribution of Living Macaques (Macaca Lacepede, 1799). in *The*
541 *Macaques: Studies in Ecology, Behavior, and Evolution* (ed. Lindburg, D. G.) 1–9 (Van Nostrand
542 Reinhold Company, 1980).
543 31. Kulathinal, R. J., Stevison, L. S. & Noor, M. A. F. The genomics of speciation in Drosophila:
544 diversity, divergence, and introgression estimated using low-coverage genome sequencing.
545 *PLoS Genet.* **5**, e1000550 (2009).
546 32. Green, R. E. *et al.* A draft sequence of the Neandertal genome. *Science* **328**, 710–722 (2010).
547 33. Martin, S. H., Davey, J. W. & Jiggins, C. D. Evaluating the Use of ABBA–BABA Statistics to Locate
548 Introgressed Loci. *Mol. Biol. Evol.* **32**, 244–257 (2015).
549 34. Martin, S. *General tools for genomic analyses. Contribute to simonhmartin/genomics_general*
550 *development by creating an account on GitHub*. (2019).
551 35. Malinsky, M. *et al.* Genomic islands of speciation separate cichlid ecomorphs in an East African
552 crater lake. *Science* **350**, 1493–1498 (2015).
553 36. Martin, S. H. & Belleghem, S. M. V. Exploring Evolutionary Relationships Across the Genome
554 Using Topology Weighting. *Genetics* **206**, 429–438 (2017).
555 37. Alby, C. *et al.* Mutations in KIAA0586 Cause Lethal Ciliopathies Ranging from a Hydrolethalus
556 Phenotype to Short-Rib Polydactyly Syndrome. *Am. J. Hum. Genet.* **97**, 311–318 (2015).
557 38. Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L. & Noble, W. S. Quantifying similarity between
558 motifs. *Genome Biol.* **8**, R24 (2007).
559 39. Rouillard, A. D. *et al.* The harmonizome: a collection of processed datasets gathered to serve and
560 mine knowledge about genes and proteins. *Database J. Biol. Databases Curation* **2016**, (2016).
561 40. Lochab, A. K. & Extavour, C. G. Bone Morphogenetic Protein (BMP) signaling in animal
562 reproductive system development and function. *Dev. Biol.* **427**, 258–269 (2017).

563 41. Yoshino, T. & Saito, D. Epithelial-to-mesenchymal transition–based morphogenesis of dorsal
564 mesentery and gonad. *Semin. Cell Dev. Biol.* **92**, 105–112 (2019).
565 42. Yoshino, T., Murai, H. & Saito, D. Hedgehog–BMP signalling establishes dorsoventral patterning
566 in lateral plate mesoderm to trigger gonadogenesis in chicken embryos. *Nat. Commun.* **7**, 1–11
567 (2016).
568 43. Chang, W. *et al.* Bmp4 Is Essential for the Formation of the Vestibular Apparatus that Detects
569 Angular Head Movements. *PLoS Genet.* **4**, (2008).
570 44. Nie, X., Xu, J., El-Hashash, A. & Xu, P.-X. Six1 regulates Grem1 expression in the metanephric
571 mesenchyme to initiate branching morphogenesis. *Dev. Biol.* **352**, 141–151 (2011).
572 45. Lai, D. & Yang, X. BMP4 is a novel transcriptional target and mediator of mammary cell
573 migration downstream of the Hippo pathway component TAZ. *Cell. Signal.* **25**, 1720–1728
574 (2013).
575 46. Taylor, S. P. *et al.* Mutations in DYNC2LI1 disrupt cilia function and cause short rib polydactyly
576 syndrome. *Nat. Commun.* **6**, 7092 (2015).
577 47. Zhao, Y. *et al.* C-terminal Domain (CTD) Small Phosphatase-like 2 Modulates the Canonical Bone
578 Morphogenetic Protein (BMP) Signaling and Mesenchymal Differentiation via Smad
579 Dephosphorylation. *J. Biol. Chem.* **289**, 26441–26450 (2014).
580 48. Kakiuchi, C., Ishiwata, M., Hayashi, A. & Kato, T. XBP1 induces WFS1 through an endoplasmic
581 reticulum stress response element-like motif in SH-SY5Y cells. *J. Neurochem.* **97**, 545–555
582 (2006).
583 49. Kuroyanagi, H. Fox-1 family of RNA-binding proteins. *Cell. Mol. Life Sci.* **66**, 3895 (2009).
584 50. Zinner, D., Arnold, M. L. & Roos, C. The strange blood: Natural hybridization in primates. *Evol.*
585 *Anthropol. Issues News Rev.* **20**, 96–103 (2011).
586 51. Yan, G. *et al.* Genome sequencing and comparison of two nonhuman primate animal models, the
587 cynomolgus and Chinese rhesus macaques. *Nat. Biotechnol.* **29**, 1019–1023 (2011).

588 52. Roos, C. *et al.* An Updated taxonomy and Conservation Status Review of Asian Primates. *Asian*
589 *Primates J. Vol. 4 1 2-38* (2014).
590 53. Melnick1, D. J. & Kidd, K. K. Genetic and evolutionary relationships among Asian Macaques. *Int.*
591 *J. Primatol.* **6**, 123–160 (1985).
592 54. Roos, C. Complete mitochondrial genome of a Toque Macaque (Macaca sinica). *Mitochondrial*
593 *DNA Part B* **3**, 182–183 (2018).
594 55. Exome sequences of toque macaques (Macaca sinica) of Sri Lanka reveal many amino acid
595 changes | bioRxiv. https://www.biorxiv.org/content/10.1101/740761v1.
596 56. Schumer, M., Cui, R., Powell, D. L., Rosenthal, G. G. & Andolfatto, P. Ancient hybridization and
597 genomic stabilization in a swordtail fish. *Mol. Ecol.* **25**, 2661–2679 (2016).
598 57. Scally, A. *et al.* Insights into hominid evolution from the gorilla genome sequence. *Nature* **483**,
599 169–175 (2012).
600 58. Hibbins, M. S. & Hahn, M. W. The Timing and Direction of Introgression Under the Multispecies
601 Network Coalescent. *Genetics* **211**, 1059–1073 (2019).
602 59. Sissler, M., González-Serrano, L. E. & Westhof, E. Recent Advances in Mitochondrial Aminoacyl-
603 tRNA Synthetases and Disease. *Trends Mol. Med.* **23**, 693–708 (2017).
604 60. Osada, N. & Akashi, H. Mitochondrial-nuclear interactions and accelerated compensatory
605 evolution: evidence from the primate cytochrome C oxidase complex. *Mol. Biol. Evol.* **29**, 337–
606 346 (2012).
607 61. Hill, G. E. *et al.* Assessing the fitness consequences of mitonuclear interactions in natural
608 populations. *Biol. Rev.* **94**, 1089–1104 (2019).
609 62. Glasgow, R. I. C. *et al.* Novel GFM2 variants associated with early-onset neurological
610 presentations of mitochondrial disease and impaired expression of OXPHOS subunits.
611 *Neurogenetics* **18**, 227–235 (2017).
612 63. Liang, L.-P. *et al.* Mitochondrial oxidative stress and epilepsy in SOD2 deficient mice:

613 attenuation by a lipophilic metalloporphyrin. *Neurobiol. Dis.* **45**, 1068–1076 (2012).
614 64. Trier, C. N., Hermansen, J. S., Sætre, G.-P. & Bailey, R. I. Evidence for Mito-Nuclear and Sex-
615 Linked Reproductive Barriers between the Hybrid Italian Sparrow and Its Parent Species. *PLOS*
616 *Genet.* **10**, e1004075 (2014).
617 65. Meiklejohn, C. D. *et al.* An Incompatibility between a Mitochondrial tRNA and Its Nuclear-
618 Encoded tRNA Synthetase Compromises Development and Fitness in Drosophila. *PLOS Genet.* **9**,
619 e1003238 (2013).
620 66. Ellison, C. K. & Burton, R. S. Cytonuclear conflict in interpopulation hybrids: the role of RNA
621 polymerase in mtDNA transcription and replication. *J. Evol. Biol.* **23**, 528–538 (2010).
622 67. Jiggins, C. D. & Martin, S. H. Glittering gold and the quest for Isla de Muerta. *J. Evol. Biol.* **30**,
623 1509–1511 (2017).
624 68. *ncbi/sra-tools*. (NCBI - National Center for Biotechnology Information/NLM/NIH).
625 69. Li, H. *Toolkit for processing sequences in FASTA/Q formats: lh3/seqtk*. (2019).
626 70. Martin, S. H., Davey, J. W., Salazar, C. & Jiggins, C. D. Recombination rate variation shapes
627 barriers to introgression across butterfly genomes. *PLOS Biol.* **17**, e2006288 (2019).
628 71. Terhorst, J., Kamm, J. A. & Song, Y. S. Robust and scalable inference of population history from
629 hundreds of unphased whole-genomes. *Nat. Genet.* **49**, 303–309 (2017).
630 72. Hinrichs, A. S. *et al.* The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res* **34**,
631 D590-8 (2006).
632 73. Mi, H. *et al.* Protocol Update for large-scale genome and gene function analysis with the
633 PANTHER classification system (v.14.0). *Nat. Protoc.* **14**, (2019).
634 74. Waal, F. B. M. de. Peace Lessons from an Unlikely Source. *PLOS Biol.* **2**, e101 (2004).
635 75. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–8 (2011).
636 76. Conway, J. & Gehlenborg, N. *UpSetR: A More Scalable Alternative to Venn and Euler Diagrams for*
637 *Visualizing Intersecting Sets*. (2019).

Figures

Figure 1. Summary of focal species, *Macaca arctoides*, its evolutionary relationships, and the data used. (a) *M. arctoides* has the largest baculum among primates and is a putative homoploid hybrid species. Image "Stumptail Monkeys" from[74] by Frans de Waal licensed under CC. (b) Model based on literature[10,14,15] depicts possible hybrid speciation between sinica and fascicularis species groups to form *M. arctoides*, with split time uncertainty relative to split of fascicularis group species, *M. mulatta* and *M. fascicularis*[22]. Approximate split times from ref[15]. This study included WGS from *M. arctoides* and representative members of each species group. Baboon was used as an outgroup species. (c) Geographic range of *M. arctoides* and other focal species in this study. *M. arctoides* has present day geographic overlap with members of both species groups (redrawn from[30]). Background image "Topographic map of East Asia" by Ksiom is licensed under CC BY-SA 3.0. (d) Intersection of heterozygous sites between the five species. The output of vcf-compare from vcftools[75] was input into the UpSetR package[76] to plot the intersection of sites. Values in the left plot match count of heterozygous sites per sample in Table 1. (e) Inferred recent effective population size trajectories for each species using smc++. Years before present based on a generation time of 6 years per generation. arc - *M. arctoides*, fas - *M. fascicularis*, mul - *M. mulatta*, thi - *M. thibetana*, ass - *M. assamensis*. (f) Topology weight distribution of three rooted topologies of species relationships in sliding windows along the genome. bab - *P. anubis*.

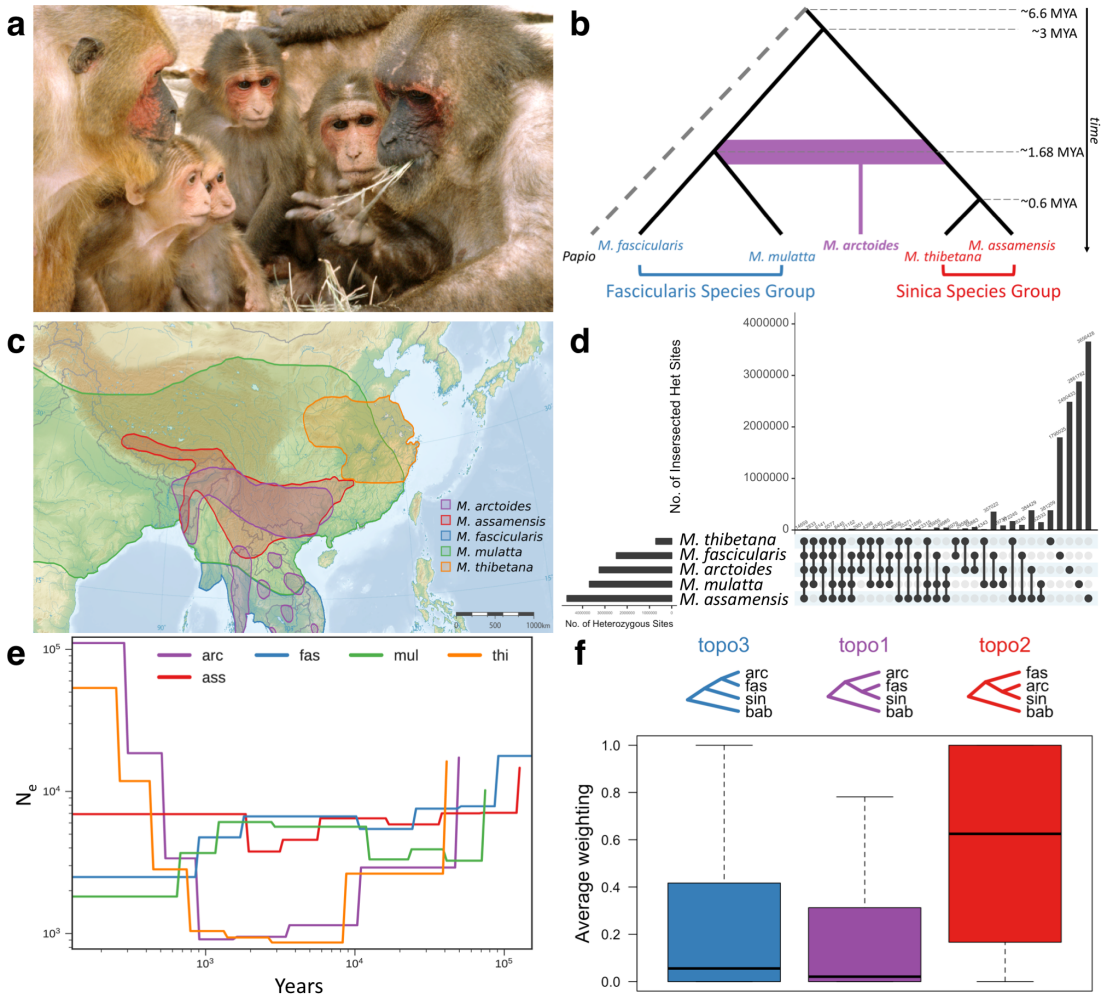
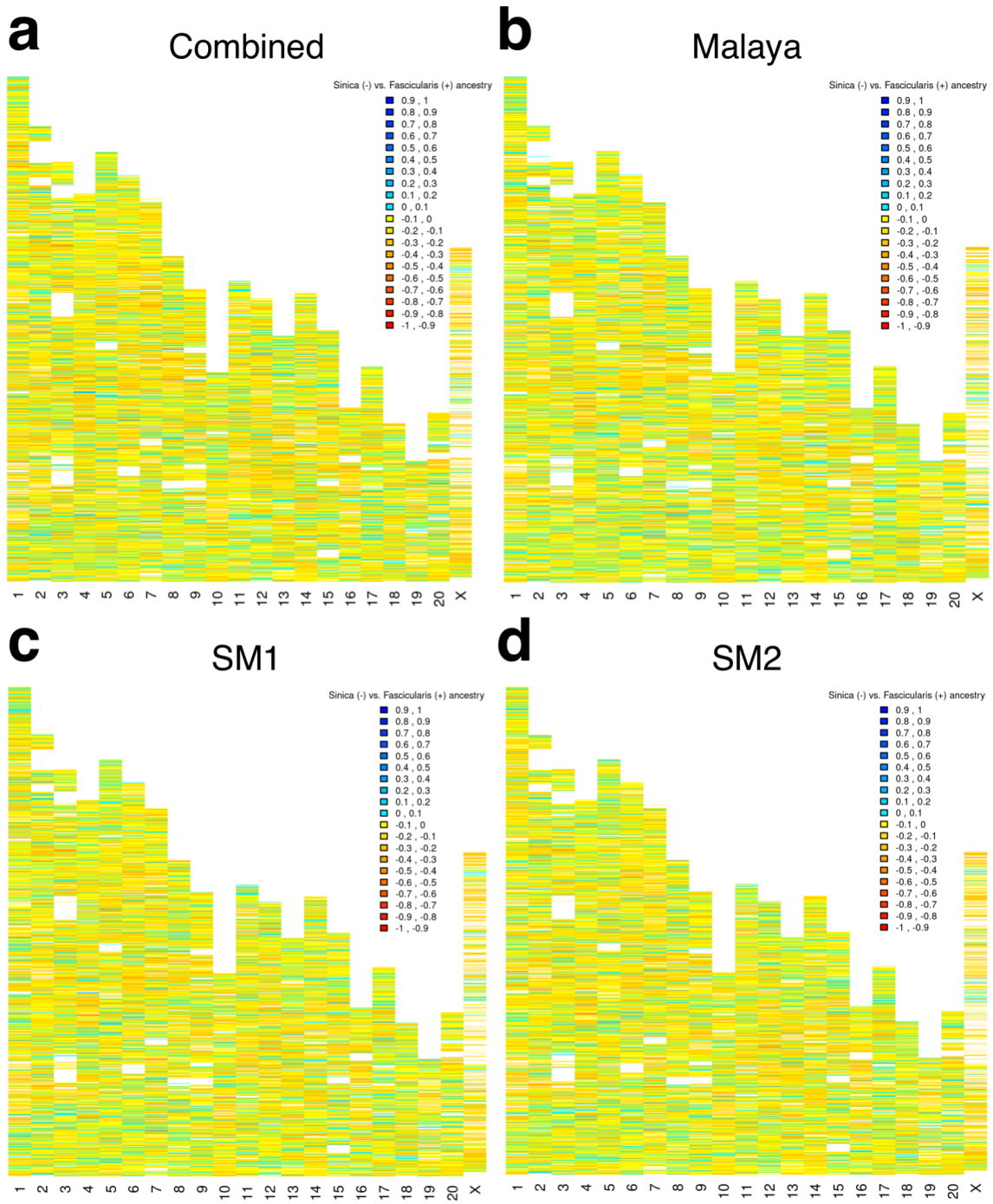
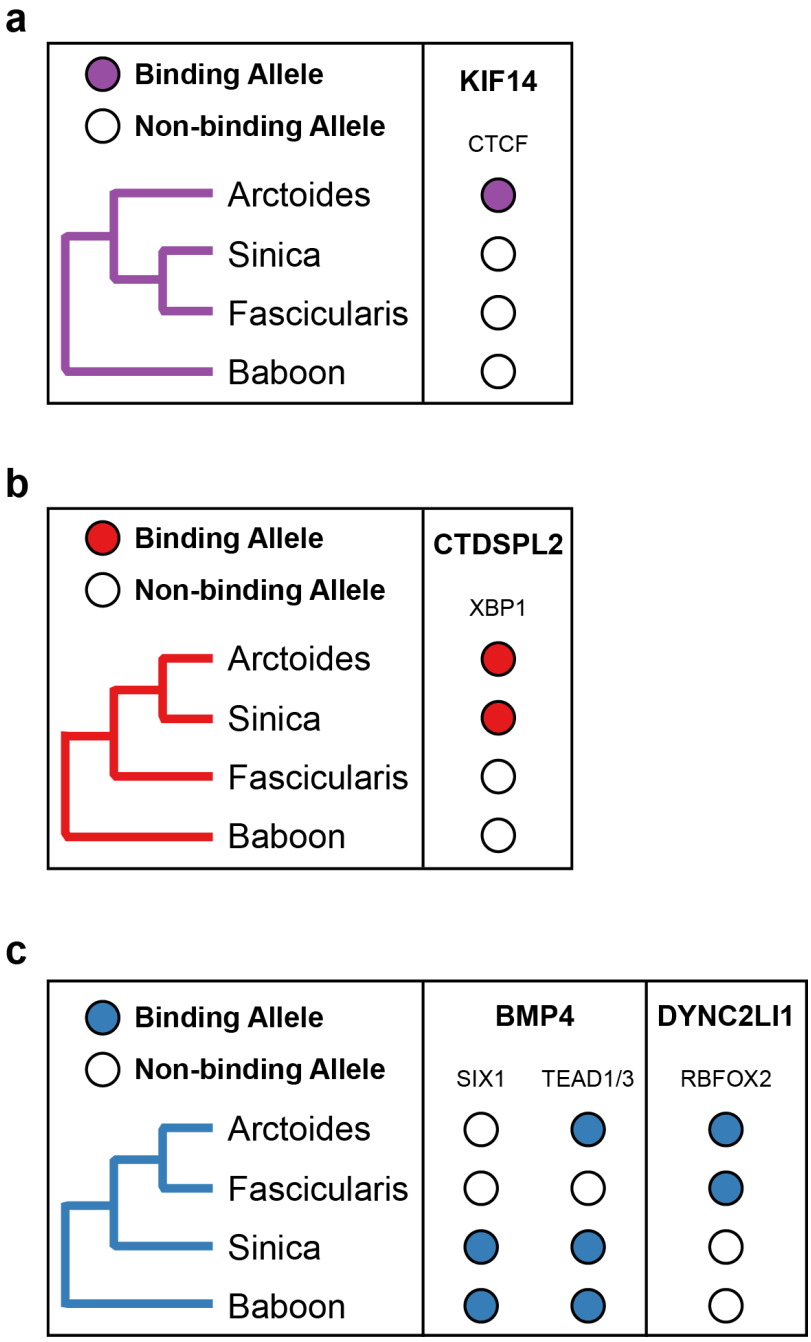


Figure 2. Introgression analysis. Results from analysis of introgression (f_{dM}) in 50kb sliding windows (a) using all three *M. arctoides* samples, (b) only the Malaya sample, (c) only the SM1 sample, and (d) only the SM2 sample. Regions where *M. arctoides* has most recent ancestry with the sinica group are displayed as negative values and regions where *M. arctoides* shares most recent common ancestry with the fascicularis group are displayed as positive values. Other window size results are in Figure S2.

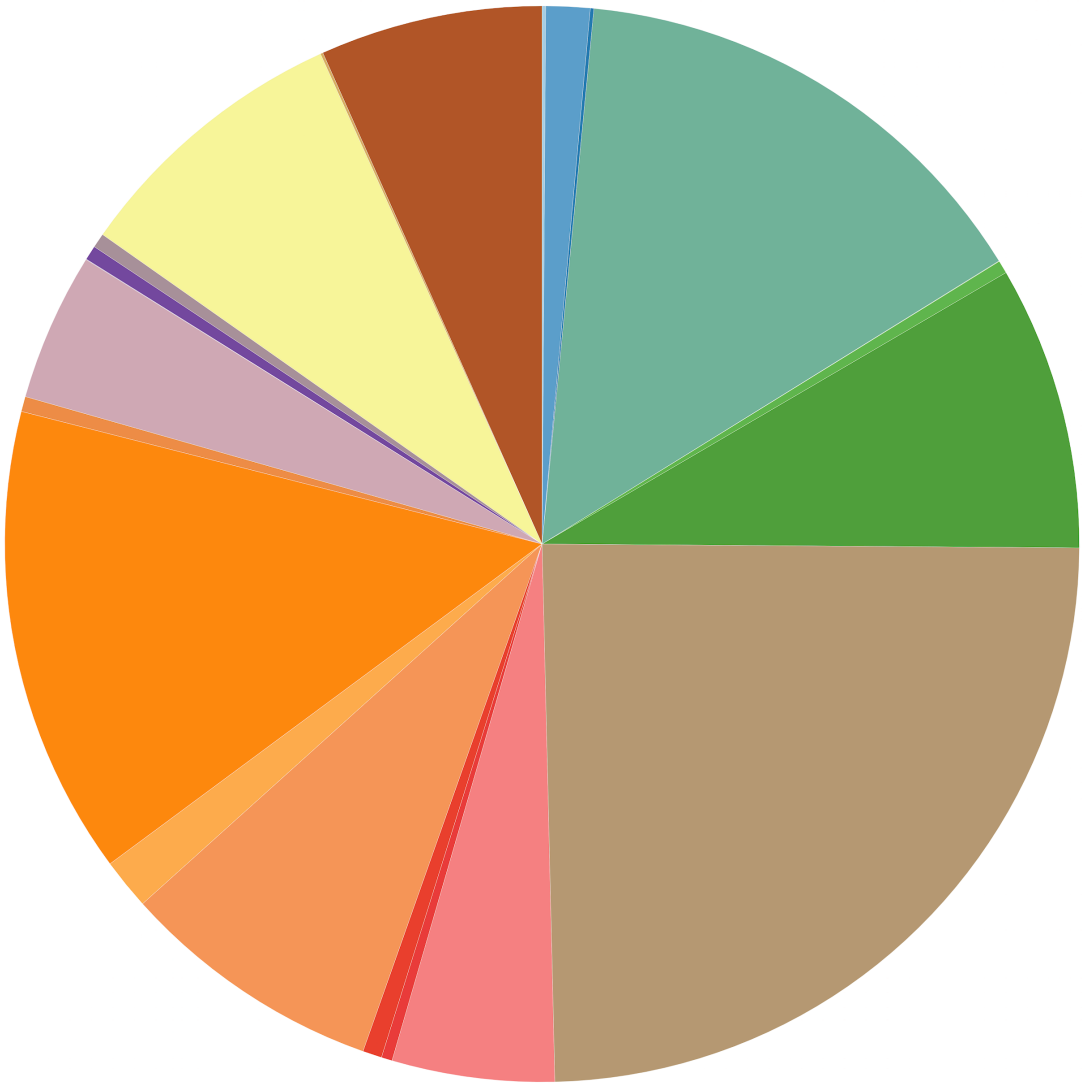


666 **Figure 3. Predicted TF binding motifs in outlier genes across all three possible gene**
667 **topologies.** Fixed differences were examined for predicted differential TF binding. (a) Genes
668 overlapping majority topo1 regions. KIF14, which overlaps baculum QTL region #1 shows unique
669 binding of CTCF for the *M. arctoides* allele. (b) Genes overlapping majority topo2 regions. CTDSPL2
670 is a sinica origin outlier gene with a f_{dM} score of -0.82. (c) Genes overlapping majority topo3 regions.
671 BMP4, which overlaps baculum QTL region #2 has two SNPs in the 3'UTR that have differential
672 predicted binding to two proteins known to interact with BMP4. Literature predicts that both of these
673 changes would increase BMP4 expression in *M. arctoides* as compared to other species. DYNC2LI1 is
674 a fascicularis origin outlier gene with a f_{dM} score of 0.79 and is known to interact with BMP4.
675

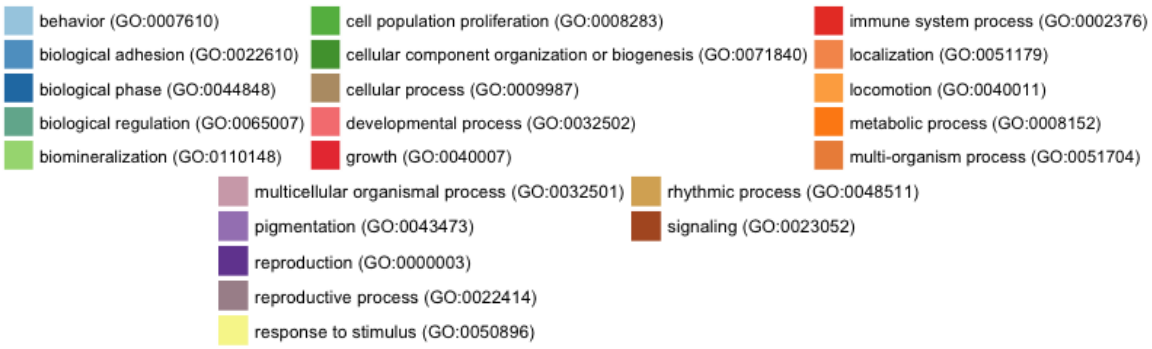


676

677 **Figure 4. Gene ontology results.** Gene ontology results from genes with majority topo1 (purple in
678 Figure 1F). Results obtained using PANTHER GO-Slim Biological Process. 4,072 IDs were used
679 resulting in a total of 6,427 process hits.
680



681



682

683

684 **Tables**

685 **Table 1. Sample information.** A list of samples used in this study with a summary of genome
686 information. Two samples were sequenced as part of this study (in bold) and the remainder were
687 downloaded from NCBI.
688

ID	species	Species Group	Coverage	No. sites VQSR Passed	No. sites heterozygous	NCBI Accession	SAMN_ID	Sex
Malaya	***M. arctoides***	**arctoides**	**20**	**6,849,957**	**3,252,204**	**SRS6488501**	**14521207**	**F**
SM1	*M. arctoides*	arctoides	32	6,848,826	3,251,892	SRR2981139	04316319	F
SM2	*M. arctoides*	arctoides	19	6,833,185	3,246,931	SRR2981140	04316320	F
BGI-CE-4	*M. fascicularis*	fascicularis	43	4,547,767	2,486,762	SRS114988	00116341	F
CR-5	*M. mulatta*	fascicularis	44	4,561,581	3,693,024	SRS115022	00113489	F
BGI	*M. mulatta*	fascicularis	10	4,531,187	3,678,452	SRS212016	00627571	M
BGI-96346	*M. mulatta*	fascicularis	4	4,334,465	3,533,580	SRS114988	00113455	M
A20	***M. assamensis***	**sinica**	**13**	**7,567,951**	**4,680,934**	**SRS6491954**	**14521608**	**F**
XH1	*M. assamensis*	sinica	49	7,669,854	4,702,055	SRR2981114	04316321	M
Tibetan-macaque-NO.3	*M. thibetana*	sinica	27	4,844,002	713,449	SRR1024051	02390221	F

689