

1 **The demographic and adaptive history of the African green monkey**

2 Susanne P. Pfeifer^{1,2,3}

3

4 1: School of Life Sciences, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

5 2: Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland

6 3: School of Life Sciences, Arizona State University (ASU), Tempe, AZ, United States

7 EPFL SV IBI

8 AAB 048

9 Station 15

10 CH-1015 Lausanne

11 Switzerland

12 Phone: +41 21 693 14 90

13 Email: susanne.pfeifer@asu.edu

14

15 Running title: Population genetics of African green monkeys

16

17 Keywords: demography, selection, African green monkey, vervet monkey

18

19 **Abstract**

20 Relatively little is known about the evolutionary history of the African green monkey
21 (genus *Chlorocebus*) due to the lack of sampled polymorphism data from wild
22 populations. Yet, this characterization of genetic diversity is not only critical for a better
23 understanding of their own history, but also for human biomedical research given that
24 they are one of the most widely used primate models. Here, I analyze the demographic
25 and selective history of the African green monkey, utilizing one of the most
26 comprehensive catalogs of wild genetic diversity to date, consisting of 1,795,643
27 autosomal single nucleotide polymorphisms in 25 individuals, representing all five
28 major populations: *C. a. aethiops*, *C. a. cynosurus*, *C. a. pygerythrus*, *C. a. sabaeus*, and *C. a.*
29 *tantalus*. Assuming a mutation rate of 5.9×10^{-9} per base pair per generation and a
30 generation time of 8.5 years, divergence time estimates range from 523-621kya for the
31 basal split of *C. a. aethiops* from the other four populations. Importantly, the resulting
32 tree characterizing the relationship and split-times between these populations differs
33 significantly from that presented in the original genome paper, owing to their neglect of
34 within-population variation when calculating between population-divergence. In
35 addition, I find that the demographic history of all five populations is well explained by a
36 model of population fragmentation and isolation, rather than novel colonization events.
37 Finally, utilizing these demographic models as a null, I investigate the selective history
38 of the populations, identifying candidate regions potentially related to adaptation in
39 response to pathogen exposure.

40

41 **Introduction**

42 The African green monkey (genus *Chlorocebus*), an Old World Monkey also referred to
43 as the vervet monkey, is an abundant primate inhabiting most ecological zones within
44 sub-Saharan Africa (with the exception of tropical forests and deserts (Figure 1;
45 Struhsaker 1967)). Their broad natural geographic distribution, reaching from west to
46 east Africa and from south of the Sahara to the Cape region of South Africa, together
47 with introduced populations on Cape Verde Island and the Caribbean (Groves 2001;
48 Groves 2005), makes vervets an excellent model system to investigate adaptation to
49 local environments. In addition, due to their genetic and physiological similarity to
50 humans, with whom they share a most recent common ancestor roughly 25 million
51 years ago (Mya) (Kumar and Hedges 1998), they are one of the most important non-
52 human primate models in basic and applied biomedical research, widely employed for
53 studies on development, cognition, and behavior (e.g., Fairbanks and McGuire 1988;
54 Fairbanks *et al.* 2004; van de Waal and Whiten 2012; Cramer *et al.* 2013), inflammatory,

55 infectious and metabolic disease (Rudel *et al.* 1997; Broussard *et al.* 2001; Olobo *et al.*
56 2001; Goldstein *et al.* 2006; Ma *et al.* 2014), as well as neurological disorders, in
57 particular Alzheimer's and Parkinson's disease (Lemere *et al.* 2004; Emborg 2007). In
58 contrast to rodents, which are frequently used in biomedical studies but which shared a
59 common ancestor with humans roughly 70Mya, vervet monkeys resemble humans
60 much more closely - not only in their physiology but also in their susceptibility and
61 response to infectious agents, thus making them a particularly valuable model to study
62 pathogens ranging from influenza virus to the simian immunodeficiency virus, a close
63 relative of human immunodeficiency virus.

64
65 Despite the fact that phenotypic and genetic data is publically available from both
66 managed pedigrees and feral populations, relatively little is known about the
67 evolutionary history of the vervet monkey. In fact, the taxonomy of the vervet monkey is
68 disputed and remains the topic of much scientific debate. Groves (2001; 2005) classified
69 vervets into five major species that are phenotypically and geographically distinct: *C.
70 sabaeus* (aka callithrix) inhabiting West Africa (from Senegal to the Volta river), *C.
71 tantalus* (aka tantalus) inhabiting north Central Africa (from Sudan to Ghana and
72 Kenya), *C. pygerythrus* (aka vervet) inhabiting East and Southern Africa (from the
73 eastern Rift Valley in Ethiopia, Somalia, and southern Sudan to South Africa), *C. aethiops*
74 (aka grivet) inhabiting the east of the White Nile region in Ethiopia, as well as areas in
75 Somalia from Khartoum to Mongalla, Eritrea, and Ethiopia, south to the Omo river, and
76 *C. cynosurus* (aka malbrouck) inhabiting Central and Southern Central Africa (from the
77 Albertine Rift to the Atlantic coast as well as northern Namibia and Zambia), with a
78 potentially sixth species, *C. djamdjamensis* (aka the Bale Mountain vervet), limited to
79 small mountain zones in the highland of Ethiopia (Goldstein *et al.* 2006). In contrast,
80 Grubb *et al.* (2003) classified all vervets into a single polytypic species (*Chlorocebus
81 aethiops*). Previous research indicating that vervet monkeys freely interbreed in the
82 narrow contact zones along their geographical boundaries (Detwiler *et al.* 2005;
83 Mekonnen *et al.* 2012; Haus *et al.* 2013) supports the single-species taxonomy, thus I
84 will follow the taxonomy proposed by Grubb *et al.* (2003). Furthermore, the
85 demographic history of the vervet recently inferred using whole genome sequence data
86 of single individuals from each of the major five African populations (Warren *et al.*
87 2015) is in strong disagreement with earlier work based on smaller data sets and
88 mitochondrial DNA (mtDNA) (Perelman *et al.* 2011; Guschanski *et al.* 2013; Haus *et al.*
89 2013), regarding both the inferred tree topology as well as the estimated split times.

90

91 However, this previous work was severely limited by insufficient polymorphism data
92 sampled from these populations. Fortunately, with the recent availability of a high-
93 quality reference genome (Warren *et al.* 2015), it is now feasible to directly investigate
94 the intra- and interspecific genetic diversity of these different vervet populations,
95 enabling a more accurate view of their demographic history. In this study, whole
96 genome data of five individuals from each of the major five wild African populations was
97 used to infer their demographic history providing much improved clarity to address the
98 conflicting estimates in the literature, and to perform the first naïve scan for genomic
99 targets of positive selection. Thereby, the knowledge gained from better understanding
100 the population genetics of this species may directly benefit medical research in at least
101 two ways. Firstly, by identifying candidate regions under selection, it may be possible to
102 functionally validate phenotype-genotype relationships, particularly with regards to
103 pathogen response. Secondly, understanding the extent of structure present between
104 natural populations is highly important for future genome-wide association studies - as
105 unknowingly sampling across hidden structure can result in spurious results.

106

107 **Results and Discussion**

108 Whole genome data of 25 individuals (mean coverage of 4.5X per individual) was used
109 to infer variants and haplotypes in each of the five wild African *Chlorocebus aethiops*
110 populations (see Materials and Methods for details). Across the autosomes, 1,795,643
111 variants were identified; for which 1,149,007 have an ancestral state that could be
112 unambiguously determined using rhesus macaque as an outgroup. The variants were
113 distributed appropriately across chromosomes (Table 1) and the number of variants
114 identified per sample within a population was highly consistent (Supplementary Figure
115 1). At regions with sufficient coverage, variation in single nucleotide polymorphism
116 (SNP) density was present, but no strong peaks of SNP density (indicative of a high local
117 false positive rate) were observed. Among the total number of segregating sites in the
118 five populations, 39% were shared between at least two populations and ~1% were
119 shared between all populations, with the remaining polymorphisms being private to a
120 single population (Figure 2).

121

122 **Population Structure**

123 The five populations exhibit similar levels of nucleotide diversity ($\pi_{\text{intergenic}} = 3.7\text{-}5.2 \times$
124 10^{-4}) in the non-coding, non-repetitive parts of the genome (Table 2) - on the lower end
125 of nucleotide diversity levels previously reported in other primates (Yu *et al.* 2004;
126 Fischer *et al.* 2006; Prado-Martinez *et al.* 2013). Nucleotide diversity levels, and thus

127 inferred effective population sizes (N_e), are highest in *C. a. cynosurus* and *C. a. tantalus*,
128 and lowest in *C. a. aethiops*, with intermediate levels in *C. a. pygerythrus* and *C. a.*
129 *sabaeus*. The level of genetic differentiation between populations (as assessed by
130 weighted F_{st}), ranging from 0.33 (*C. a. cynosurus* - *C. a. tantalus*) to 0.6 (*C. a. aethiops* - *C.*
131 *a. sabaeus*) (Table 2), indicates strong genetic structure between different populations,
132 overlapping the range of differentiation previously reported between different
133 chimpanzee populations (e.g., F_{st} (western chimpanzee - central chimpanzee) = 0.25-
134 0.38, and F_{st} (western chimpanzee - eastern chimpanzee) = 0.31-0.42; Becquet *et al.*
135 2007; Fischer *et al.* 2011) as well as bonobos and chimpanzees (e.g., F_{st} (bonobo - central
136 chimpanzee) = 0.49-0.54 and F_{st} (bonobo - western chimpanzee) = 0.68-0.74; Fischer *et*
137 *al.* 2006, 2011). The only exception is a weighted F_{st} of 0.16 between *C. a. cynosurus* and
138 *C. a. pygerythrus*, similar to F_{st} values observed between human populations (Rosenberg
139 *et al.* 2002). The large differentiation of *C. a. aethiops* compared with the other
140 populations provides the first line of evidence suggesting that they may represent the
141 earliest split on the tree, contrary to previous results.

142
143 To explore genetic evidence for population structure among vervets, the level of shared
144 ancestry among individuals was studied using independent-loci admixture models as
145 implemented in the software fastSTRUCTURE (Raj *et al.* 2014). fastSTRUCTURE assigns
146 individuals to a hypothesized number of populations such that the amount of linkage
147 disequilibrium across loci is minimized (see Materials and Methods). The best-fit model
148 had four ancestry components, strongly supporting the division of samples into four
149 discontinuous populations. The inferred clusters correspond well to the reported labels
150 of *C. a. aethiops*, *C. a. sabaeus*, *C. a. tantalus*, grouping *C. a. cynosurus* and *C. a. pygerythrus*
151 into the same cluster, with little evidence for admixture between the populations
152 (Supplementary Figure 2). fastSTRUCTURE also has power to assess ancestry
153 proportions for individuals with mixed ancestry, of which none were identified. In
154 addition, principal component analysis (PCA) was used to determine the levels of
155 genetic differentiation between individuals. The first principal component, explaining
156 17.4% of the variation, clearly separates the geographically most isolated population
157 with the smallest habitat range, *C. a. aethiops*, from the other four populations (Figure
158 3). *C. a. sabaeus*, *C. a. tantalus*, *C. a. cynosurus*, and *C. a. pygerythrus* are distributed along
159 the second principal component (explaining 14.5% of the variation), with no clear
160 differentiation between *C. a. cynosurus* and *C. a. pygerythrus*, both of which inhabit
161 central and south Africa. Construction of an autosomal consensus tree (*i.e.*, a tree that
162 most commonly represents the relationship between the sample locations) indicates

163 that one *C. a. pygerythrus* individual (SRR556103) falls within the *C. a. cynosurus* cluster,
164 whereas separate monophyletic groups are supported for *C. a. aethiops*, *C. a. sabaeus*,
165 and *C. a. tantalus* (Supplementary Figure 3) - consistent with the results of both the
166 fastSTRUCTURE and PCA analyses, and additionally supported by the identity-by-state
167 pattern observed for each pair of individuals (Supplementary Figure 4).

168

169 Demographic history

170 Patterns of genetic divergence were used to elucidate the evolutionary history of the
171 vervet populations. Specifically, divergence times were estimated using a molecular
172 clock based on putatively neutral fixed differences between the genomes of the
173 populations (Table 3), assuming that mutations occurred at a constant rate of 5.9×10^{-9}
174 per base pair per generation among lineages (Hernandez *et al.* 2007) and a generation
175 time of 8.5 years (Warren *et al.* 2015). The inferred topology (Figure 4) is consistent
176 with the observed levels of genetic differentiation (Table 2) as well as allele sharing
177 between the species, with evidence of incomplete lineage sorting being observed (Figure
178 2), as reported for many other primates (Patterson *et al.* 2006; Hobolth *et al.* 2007). The
179 basal split of *C. a. aethiops* from the other four populations was estimated to be 523-
180 621kya. The time estimates for $(C. a. sabaeus + C. a. tantalus)/(C. a. cynosurus + C. a.$
181 *pygerythrus)*, *C. a. sabaeus/C. a. tantalus*, and *C. a. cynosurus/C. a. pygerythrus* were 242-
182 333kya, 239kya, and 143kya, respectively. These estimated divergence times are
183 roughly consistent with the ones recently inferred using whole genome sequence data of
184 single individuals from each of the five African populations (Warren *et al.* 2015) but
185 considerably younger than those reported from mtDNA (Guschanski *et al.* 2013).
186 However, the inferred tree topology is in disagreement with results of previously
187 published studies on the topic (Perelman *et al.* 2011; Guschanski *et al.* 2013; Haus *et al.*
188 2013; Warren *et al.* 2015). Discrepancies in both divergence times and tree shapes
189 between estimates obtained from mitochondrial and nuclear genomic data have been
190 observed in many different species (Avise 1994; Funk and Omland 2003; Chan and
191 Levin 2005; Toews and Brelsford 2012) including other primates (*e.g.*, Wise *et al.* 1997;
192 Stone *et al.* 2010; Nietlisbach *et al.* 2012) - a discordance that has generally been
193 attributed to differences in the selective and demographic histories of mitochondrial
194 and nuclear DNA (such as sex-biased dispersal, different effective female than male
195 population sizes, and adaptive introgression (Toews and Brelsford 2012)), as well as
196 incomplete lineage sorting. The fourfold smaller effective population size in mtDNA
197 compared to nuclear DNA will also make it more susceptible to stochastic variation, thus
198 mtDNA may not be a good representative of ancestry and genetic diversity across the

199 entire genome. In contrast, the discrepancy in tree topology with the one recently
200 inferred by Warren *et al.* (2015) using whole genome sequence data of single
201 individuals from each of the major five African populations is likely explained by the fact
202 that the authors were unable to discern between segregating sites and fixed differences
203 in the populations, owing to the analysis of single individuals. That is, within-population
204 variation was confounded with between-population divergence. In addition, their
205 analyses were based on genome-wide variant data, potentially biasing demographic
206 inferences via both direct and linked selection (see Ewing and Jensen 2016), whereas I
207 here define putatively neutral regions for such analysis.

208
209 A standard equilibrium model without migration was the best fit to the data (Figure 5),
210 suggesting population fragmentation rather than colonization as a driver of the
211 demographic history of vervet monkeys, as well as historically stable population sizes.
212 Consistent with biases expected from utilizing a multi-sample genotype calling strategy
213 on low coverage sequencing data (Han *et al.* 2013), rare variants were under-called in
214 the data set, distorting the site frequency spectrum (SFS) to include fewer singletons
215 than doubletons. This deficit of low frequency variants results in a stronger
216 underestimation of π compared to θ_w , skewing Tajima's D towards slightly positive
217 values in all populations (Table 2). Importantly however, Han *et al.* (2013)
218 demonstrated that rank-based statistics used for genome-wide selection scans are less
219 sensitive to such biases in the inferred SFS, enabling identification even at low coverage.
220

221 Selection

222 Given the already large divergence between populations shown above, F_{st} -based scans
223 will be underpowered. Correspondingly, no significant outliers were detected when
224 such an approach was applied to the data (BayeScan; Foll and Gaggiotti 2008). Thus,
225 population-specific scans relying on site frequency spectrum-based expectations were
226 utilized. Statistical tests based on a classical selective sweep model (SweeD; Pavlidis *et*
227 *al.* 2013) suggest that 1.5-3.1% of the vervet genome may be affected by recent selective
228 sweeps (Supplementary Figure 5), when assuming a 1% false positive rate. Comparisons
229 with the extent of positive selection in other primate genomes are somewhat tenuous,
230 given the strongly differing results depending on the methodology employed (see
231 discussion in Biswas and Akey 2006; Crisci *et al.* 2012; Jensen 2014). It should further
232 be noted that the false discovery rate of selection scans may be higher than anticipated
233 (Teshima *et al.* 2006), due to the challenges in detecting the footprint of a selective
234 sweep. In particular, differentiating sweeps from other patterns of genetic background

235 variation that reflect unaccounted population history, variability in the underlying
236 mutation and recombination landscapes, as well as differing modes of selection, has
237 been challenging (see reviews of Bank *et al.* 2014; Jensen *et al.* 2016). Fortunately, the
238 demographic inference above suggests relatively stable population histories, devoid of
239 the kind of severe population size changes which have been shown to induce major
240 difficulties when conducting genomic scans (Thornton and Jensen 2007). In addition,
241 several regions have strong and consistent evidence of being targeted by positive
242 selection in multiple populations (Supplementary Table), a particularly promising result
243 given both the old split times as well as the lacking evidence for on-going migration.
244 One such region with the strong support includes the gene *DYNC1I1* (Figure 6), a known
245 target of herpes simplex virus, which interacts with the dynein motor of the nuclear
246 membrane to transport capsid-tegument structures to the nuclear pore (Ye *et al.* 2000).
247 This finding is noteworthy given the well-described infection history of vervet monkeys
248 with the herpes virus (Malherbe and Harwin 1958; Clarkson *et al.* 1967; Wertheim *et al.*
249 2014). Further, this offers an excellent candidate region for investigating host-shift
250 between humans (in which infection is associated with severe symptoms, and may be
251 fatal) and vervets (in which infections are generally asymptomatic) (Nsabimana *et al.*
252 2008).

253

254 Conclusions

255 To clarify the existing and conflicting estimates regarding the demographic history of
256 the vervet monkey and to characterize the species' adaptive history, one of the most
257 comprehensive catalogs of wild genetic diversity to date was generated, consisting of
258 1,795,643 autosomal single nucleotide polymorphisms, identified in five individuals
259 from each of the major five wild African populations. Population genetic analyses
260 confirmed strong genetic structure between the different populations, with nucleotide
261 diversity levels on the lower end of previously reported values in other primates.
262 Divergence between *C. a. aethiops* and the other four extant populations is best fit with a
263 model of population fragmentation and isolation, rather than novel colonization events,
264 beginning roughly 523-621kya. This finding is in strong disagreement with previously
265 published results based on smaller data sets, single individuals, and mitochondrial DNA
266 (Perelman *et al.* 2011; Guschnski *et al.* 2013; Haus *et al.* 2013; Warren *et al.* 2015). The
267 ability to here utilize polymorphism data for improved demographic inference, to
268 account for segregating variation when inferring divergence times, and to focus on
269 putatively neutral regions of the genome, have all contributed to this improved clarity.
270 Further, evidence of recent selective sweeps at the genomic level was detected in all

271 populations. While population-specific events are of interest, the most promising
272 candidates are likely those with evidence in multiple populations. The strongest such
273 signal contains a gene related to viral exposure, providing a valuable candidate for
274 future study concerning both host-shift and the underlying causes of symptomatic
275 infections of the herpes simplex virus.

276

277 **Materials and Methods**

278 **Read Mapping**

279 Whole-genome sequence data for 25 individuals (five individuals per population with a
280 genome-wide mean coverage of 4.5X per individual) was downloaded from SRA (*i.e.*, *C.*
281 *a. tantalus*: SRR556154, SRR556127, SRR556105, SRR556122, SRR556151; *C. a.*
282 *sabaeus*: SRR556189, SRR556192, SRR556194, SRR556191, SRR556193; *C. a.*
283 *pygerythrus*: SRR556143, SRR556109, SRR556118, SRR556116, SRR556103; *C. a.*
284 *cynosurus*: SRR556153, SRR556149, SRR556117, SRR556164, SRR556161; *C. a.*
285 *aethiops*: SRR556111, SRR556121, SRR556162, SRR556165, SRR556133). Reads were
286 aligned to the repeat-masked *Chlorocebus sabaeus* reference assembly v.1.1 (NCBI
287 GenBank accession number GCA_000409795.2), consisting of assemblies for 29
288 autosomes (mean length: 89,933,368bp), chromosome X (130,038,232bp), chromosome
289 Y (6,181,219bp), the mitochondrion (16,550bp), as well as 1,972 unplaced scaffolds
290 (mean length: 23,085bp) (Warren *et al.* 2015), and the Epstein-Barr virus genome (NCBI
291 Reference Sequence NC_007605.1) using BWA-MEM v.0.7.13 (Li and Durbin 2009).
292 Thereby, the Epstein-Barr virus assembly was included as a decoy in the read alignment
293 step to enable the absorption of reads that did not originate from vervet monkey DNA
294 (as DNA sequences of interest are often contaminated (*e.g.*, by Epstein-Barr virus,
295 frequently used in laboratories to immortalize the cell lines)). The inclusion of such a
296 decoy genome has been shown to frequently improve the accuracy of read alignments
297 by reducing false positive variant calls (see review of Pfeifer (2016)). Following the
298 Genome Analysis Toolkit (GATK) v.3.5 Best Practice (McKenna *et al.* 2010; DePristo *et al.*
299 2011; Van der Auwera *et al.* 2013), duplicates were marked using Picard v.2.1.1, before
300 conducting multiple sequence realignments with simultaneous adjustment of Base
301 Alignment Qualities (Li 2011). Next, base quality scores were recalibrated using GATK's
302 BaseRecalibrator v.3.5 together with ~500k variants from the genome-wide SNP panel
303 of the Vervet Genetic Mapping Project (Huang *et al.* 2015), downloaded from the
304 European Variant Archive (study number PRJEB7923).

305

306 **Variant Calling and Filtering**

307 Variants were called using GATK's HaplotypeCaller v.3.5, a method well suited for low
308 coverage depths averaging 4-6X per individual (Cheng *et al.* 2014), and jointly
309 genotyped using GATK's GenotypeGVCFs v.3.5. Although soft filtering methods using
310 machine learning (such as GATK's VQSR) have a better specificity at low coverage than
311 hard filtering methods (Cheng *et al.* 2014), these techniques can not readily be applied
312 in this study due to the fact that soft filtering methods require the construction of a
313 statistical model based on a set of known high-quality variant calls. Hence, these
314 methods require a large training data set of known high-quality variants in the
315 underlying model, which unfortunately is not yet available for vervet monkeys.
316 Thus, post-genotyping, the raw variant call set was limited to autosomal, bi-allelic single
317 nucleotide polymorphisms (SNPs) and conservatively filtered using the following hard
318 filtering criteria (with acronyms as defined by the GATK package), attempting to
319 minimize the number of false positives by identifying variants with characteristics
320 outside their normal distributions: SNPs were excluded if they were supported by reads
321 showing evidence of a strand bias (as estimated by Fisher's exact test (FS>60.0) or the
322 Symmetric Odds Ratio test (SOR>4.0)) or a bias in the position of alleles within the reads
323 that support them between the reference and alternate alleles (ReadPosRankSum<-8.0).
324 SNPs were also filtered out if they were supported by reads with a low read mapping
325 quality (MQ<40) or a qualitative difference between the mapping qualities of the reads
326 supporting the reference allele and those supporting the alternate allele
327 (MQRankSum<12.5). In addition, SNPs were removed from the data set if the variant
328 confidence was low (QD<2.0). Due to a frequent misalignment of reads in repetitive
329 regions of the genome, leading to an excess of heterozygous genotype calls, SNPs within
330 repeats were excluded from further analyses. In addition, SNPs showing an excess of
331 heterozygosity were removed using the '--hardy' option in VCFtools v.0.1.12b (Danecek
332 *et al.* 2011) with p<0.01. The data set was further limited to SNPs exhibiting complete
333 genotype information.
334
335 To achieve a higher specificity, a second independent variant call was performed using
336 Platypus v.0.8.1, an integrated mapping-, assembly-, and haplotype-variant caller
337 (Rimmer *et al.* 2014), and an intersection variant data set was generated.
338

339 Variant Data Set

340 The intersection data set contained 1,795,643 autosomal variants, with an average
341 transition-transversion (Ts/Tv) ratio of 2.8 (Table 1). Given the use of low coverage
342 (4.5X) sequence data in this study, genotypes were subsequently refined using the

343 software BEAGLE v4 (Browning and Browning 2007). Variants were polarized using
344 rhesus macaque as an outgroup. Using the rhesus macaque genome assembly, rheMac8,
345 consisting of 23 chromosomes as well as 284,705 scaffolds with a total size of 3.2Gb
346 (downloaded from the UCSC Genome Browser), the ancestral state of 1,149,007 variants
347 could be unambiguously determined. For each population, the number of segregating
348 sites shared between any single other population as well as all other populations was
349 recorded, together with the number of segregating sites unique to each population
350 (Figure 2). Subsequently, the data set was annotated using ANNOVAR v2016Feb01
351 (Wang *et al.* 2010) with the annotation of the vervet genome build (NCBI *Chlorocebus*
352 *sabaeus* Annotation Release 100) consisting of 29,648 genes, resulting in 22,767 exonic
353 and 577,004 intergenic SNPs.

354

355 Accessible Genome

356 Due to the fact that the application of filter criteria to the variant data set led to the
357 exclusion of a substantial fraction of genomic sites accessible to variant identification,
358 mask files, defining which genomic sites were accessible to variant discovery, were
359 created. Thereby, monomorphic sites were called and filtered using the same pipeline
360 and hard filter criteria as used for the variant sites (as described in the 'Variant Calling
361 and Filtering' section), with the exception of turning the '-all' flag in GATK's
362 GenotypeGVCFs run on to include non-variant loci. The number of autosomal
363 monomorphic sites in the reference assembly (454,322,622) was then obtained from
364 these mask files. Following the same procedure as for the polymorphic sites, the
365 ancestral state of all monomorphic sites was determined and sites were annotated,
366 resulting in 144,943,664 intergenic monomorphic sites for which the ancestral state
367 could be unambiguously determined.

368

369 Population Structure

370 A consensus tree (*i.e.*, a tree that most commonly represents the relationship between
371 the sample locations) was constructed based on autosomal variant calls with ancestral
372 allele annotation using the maximum likelihood method implemented in SNPhylo
373 v.20140701 (Lee *et al.* 2014). Gnu R's 'snpgdsLDpruning' was used with a linkage
374 disequilibrium threshold of $r^2 > 0.2$ to generate a pruned subset of SNPs from the data
375 set, where SNPs are in approximate linkage equilibrium with each other. Using this set
376 of variants, evidence of population structure was assessed using PCA (Zheng *et al.* 2012)
377 as well as an independent-loci admixture model in the software fastSTRUCTURE v.1.0
378 (Raj *et al.* 2014), detecting clusters of related individuals from multi-locus genotyping

379 data, thereby allowing individuals to have ancestry from multiple populations.
380 fastSTRUCTURE was applied to values of K (the number of clusters) between 1 and 5
381 and the best K was chosen such that it maximizes the marginal likelihood. The fraction
382 of identity-by-state (IBS) for each pair of individuals was calculated using Gnu R's
383 'snpgdsIBS'. For each population, summary statistics, namely nucleotide diversity π (Nei
384 and Li 1979) and Watterson's estimate of θ , θ_w (Watterson 1975), were calculated
385 using the libsequence package msstats v.0.3.4 (Thornton 2003) (Table 2). Both π and θ_w
386 estimate the neutral parameters $4N_e\mu$ under equilibrium conditions, where N_e is the
387 effective population size and μ is the neutral mutation rate. Effective population sizes
388 were directly estimated from the data by fixing the mutation rate μ to 5.9×10^{-9} per base
389 pair per generation (*i.e.*, the mutation rate observed in rhesus macaque, as there is no
390 direct estimate for μ available in vervet monkeys; Hernandez *et al.* 2007). Using
391 VCFtools v.0.1.12b (Danecek *et al.* 2011), Tajima's D (Tajima 1989) was calculated (in
392 10kb windows along the genome) to test for deviations from the equilibrium neutral
393 site frequency distribution, and Weir and Cockerham's fixation index F_{st} was calculated
394 between each pair of populations to assess population differentiation.

395

396 Population Divergence

397 Divergence times were estimated using a molecular clock based on putatively neutral
398 fixed differences (from the intersection data set for which the ancestral state could be
399 unambiguously determined) between the genomes of the populations (Table 3),
400 assuming that mutations occurred at a constant rate μ of 5.9×10^{-9} per base pair per
401 generation among lineages (Hernandez *et al.* 2007) and a generation time of 8.5 years
402 (Warren *et al.* 2015).

403

404 A variety of population bottleneck models were tested and their fit to the data was
405 compared to the fit of a standard equilibrium model. In the proposed bottleneck models,
406 the ancestral effective population size N_0 (varied between 10,000, 20,000, 30,000,
407 40,000, and 50,000 individuals) was reduced to levels of 10%-90% (*i.e.*, the severity of
408 the bottleneck) in 10% intervals for the last 10, 25, 50, 75, and 100 generations (*i.e.*, the
409 duration of the bottleneck). Specifically, for each model, 1,000 independent simulations
410 of 10,000bp length were performed using the coalescence simulator 'msprime'
411 (Kelleher *et al.* 2016), assuming a mutation rate μ of 5.9×10^{-9} per base pair per
412 generation (Hernandez *et al.* 2007) and a recombination rate $\rho=\mu$. For each simulation,
413 summary statistics (*i.e.*, F_{st} (Hudson *et al.* 1992) as well as Tajima's D (Tajima 1989))
414 were calculated using the libsequence package msstats v.0.3.4 (Thornton 2003) and

415 compared to the data. In addition, the fit of a 5-population standard equilibrium model
416 based on the inferred tree topology and the estimated divergence times was assessed.
417

418 **Identification of Candidate Loci under Selection**

419 The software SweeD v.3.3.2 (Pavlidis *et al.* 2013), which implements a modification of
420 the Kim and Stephan (2002) composite likelihood ratio (CLR) test as extended by
421 Nielsen *et al.* (2005), was used to detect loci putatively subject to positive selection by
422 scanning the genome for signals of hard (fixed) selective sweeps. For each population,
423 the CLR statistic was calculated from the unfolded SFS at 1,000 grid points across each
424 chromosome. Statistical thresholds were calculated following Nielsen *et al.* (2005) by
425 simulating 1,000 variant data sets under a standard equilibrium model without
426 migration and defining the threshold as the 99th percentile of the distribution of the
427 highest simulated CLR values.

428
429 In addition to SweeD, BayeScan v.2.1 (Foll and Gaggiotti 2008) was used to detect loci
430 that show evidence of selection by computing the differences in allele frequencies
431 between the different populations.

432

433 **Acknowledgements**

434 I am grateful to Jeffrey Jensen and Stefan Laurent for helpful comments and discussion.
435 Computations were performed at the Vital-IT (<http://www.vital-it.ch>) Center for high-
436 performance computing of the Swiss Institute of Bioinformatics (SIB).

437

438 **Disclosure Declaration**

439 The author declares no conflict of interest.

440

441 **References**

- 442 Avise J. 1994. Molecular Markers, Natural History and Evolution. New York: Chapman
443 and Hall.
444 Bank C, Foll M, Ferrer-Admetlla A, Ewing G, Jensen JD. 2014. Thinking too positive?
445 Revisiting current methods in population genetic selection inference. Trends
446 Genet 30 (12):540-546.
447 Becquet C, Patterson N, Stone AC, Przeworski M, Reich D. 2007. Genetic structure of
448 chimpanzee populations. PLoS Genet 3 (4):e66.
449 Biswas S, Akey JM. 2006. Genomic insights into positive selection. Trends Genet 22:437-
446.
450
451 Broussard SR, Staprans SI, White R, Whitehead EM, Feinberg MB, Allan JS. 2001. Simian
452 immunodeficiency virus replicates to high levels in naturally infected African
453 green monkeys without inducing immunologic or neurologic disease. J Virol 75

- 454 (5):2262-2275.
455 Browning SR, Browning BL. 2007. Rapid and accurate haplotype phasing and missing-
456 data inference for whole-genome association studies by use of localized
457 haplotype clustering. *Am J Hum Genet* 81 (5):1084-1097.
458 Chan KMA, Levin SA. 2005. Leaky prezygotic isolation and porous genomes: rapid
459 introgression of maternally inherited DNA. *Evolution* 59:720-729.
460 Cheng AY, Teo YY, Ong RT. 2014. Assessing single nucleotide variant detection and
461 genotype calling on whole-genome sequenced individuals. *Bioinformatics* 30
462 (12):1707-1713.
463 Clarkson MJ, Thorpe E, McCarthy K. 1967. A viral disease of captive vervet monkeys
464 (*Cercopithecus aethiops*) caused by a new herpesvirus. *Archives of Virology*
465 22:219.
466 Crisci J, Poh Y, Bean A, Simkin A, Jensen JD. 2012. Recent progress in polymorphism-
467 based population genetic inference. *J Hered* 103:287-296.
468 Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter
469 G, Marth GT, Sherry ST et al. 2011. The variant call format and
470 VCFtools. *Bioinformatics* 27 (15):2156-2158.
471 DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del
472 Angel G, Rivas MA, Hanna M et al. 2011. A framework for variation discovery and
473 genotyping using next-generation DNA sequencing data. *Nat Gen* 43 (5):491-
474 498.
475 Detwiler KM, Burrell AS, Jolly CJ. 2005. Conservation Implications of hybridization in
476 African cercopithecine monkeys. *Int J Primatol* 26 (3):661-784.
477 Emborg ME. 2007. Nonhuman primate models of Parkinson's disease. *ILAR J* 48 (4):339-
478 355.
479 Ewing GB, Jensen JD. 2016. The consequences of not accounting for background
480 selection in demographic inference. *Mol Ecol* 25 (1):135-141.
481 Ewing G, Reiff P, Jensen JD. 2015. PopPlanner: Visually constructing demographic
482 models for simulation. *Front Genet* 6:150.
483 Fairbanks LA, Jorgensen MJ, Huff A, Blau K, Hung YY, Mann JJ. 2004. Adolescent
484 impulsivity predicts adult dominance attainment in male vervet monkeys. *Am J*
485 *Primatol* 64 (1):1-17.
486 Fairbanks LA, McGuire MT. 1988. Long-term effects of early mothering behavior on
487 responsiveness to the environment in vervet monkeys. *Dev Psychobiol* 21
488 (7):711-724.
489 Fay JC, Wu CI. 2000. Hitchhiking under positive Darwinian selection. *Genetics* 155
490 (3):1405-1413.
491 Fischer A, Pollack J, Thalmann O, Nickel B, Pääbo S. 2006. Demographic history and
492 genetic differentiation in apes. *Curr Biol* 16 (11):1133-1138.
493 Fischer A, Prüfer K, Good JM, Halbwax M, Wiebe V, André C, Atencia R, Mugisha L, Ptak
494 SE, Pääbo S. 2011. Bonobos fall within the genomic variation of
495 chimpanzees. *PLoS ONE* 6 (6):e21605.
496 Foll M, Gaggiotti O. 2008. A genome-scan method to identify selected loci appropriate for
497 both dominant and codominant markers: a Bayesian perspective. *Genetics* 180
498 (2):977-993.
499 Funk DJ, Omland KE. 2003. Species level paraphyly and polyphyly: frequency, causes,
500 and consequences, with insights from animal mitochondrial DNA. *Ann Rev Ecol*
501 *Evol Syst* 34:397-423.
502 Goldstein S, Ourmanov I, Brown CR, Beer BE, Elkins WR, Plishka R, Buckler-White A,
503 Hirsch VM. 2000. Wide range of viral load in healthy African green monkeys
504 naturally infected with simian immunodeficiency virus. *J Virol* 74 (24):11744-
505 11753.
506 Groves C. 2001. Primate taxonomy. Washington, DC: Smithsonian Institution Press.

- 507 Groves C. 2005. Order Primates. In: Wilson DE, Reeder DM, editors. *Mammal Species of*
508 *the World*. Baltimore (MD): The Johns Hopkins University Press. p. 111-184.
509 Grubb R, Butynski TM, Oates JE, Bearder SK, Disotell TR, Groves CP, Struhsaker TT.
510 2003. Assessment of the diversity of African primates. *Int J Primatol* 24
511 (6):1301-1357.
512 Guschanski K, Krause J, Sawyer S, Valente LM, Bailey S, Finstermeier K, Sabin R, Gilissen
513 E, Sonet G, Nagy ZT et al. 2013. Next-generation museomics disentangles one of
514 the largest primate radiations. *Syst Biol* 62 (4):539-554.
515 Guttenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009. Inferring the joint
516 demographic history of multiple populations from multidimensional SNP
517 frequency data. *PLoS Genet* 5 (10):e1000695.
518 Han E, Sinsheimer JS, Novembre J. 2014. Characterizing bias in population genetic
519 inferences from low-coverage sequencing data. *Mol Biol Evol* 31 (3):723-735.
520 Haus T, Akom E, Agwanda B, Hofreiter M, Roos C, Zinner D. 2013. Mitochondrial
521 diversity and distribution of African green monkeys (*chlorocebus gray, 1870*).
522 *Am J Primatol* 75 (4):350-360.
523 Hernandez RD, Hubisz MJ, Wheeler DA, Smith DG, Ferguson B, Rogers J, Nazareth L,
524 Indap A, Bourquin T, McPherson J et al. 2007. Demographic histories and
525 patterns of linkage disequilibrium in Chinese and Indian rhesus
526 macaques. *Science* 316 (5822):240-243.
527 Hobolth A, Christensen OF, Mailund T, Schierup MH. 2007. Genomic relationships and
528 speciation times of human, chimpanzee, and gorilla inferred from a coalescent
529 hidden Markov model. *PLoS Genet* 3 (2):e7.
530 Huang YS, Ramensky V, Service SK, Jasinska AJ, Jung Y, Choi OW, Cantor RM, Juretic N,
531 Wasserscheid J, Kaplan JR et al. 2015. Sequencing strategies and characterization
532 of 721 vervet monkey genomes for future genetic analyses of medically relevant
533 traits. *BMC Biol* 13:41.
534 Hudson RR, Slatkin M, Maddison WP. 1992. Estimation of levels of gene flow from
535 population data. *Genetics* 132:583-589.
536 Jensen JD. 2014. On the unfounded enthusiasm for soft selective sweeps. *Nat Commun*
537 5:5281.
538 Jensen JD, Foll M, Bernatchez L. 2016. Introduction: the past, present, and future of
539 genomic scans for selection. *Mol Ecol* 25 (1): 1-4.
540 Kelleher J, Etheridge AM, McVean G. 2016. Efficient Coalescent Simulation and
541 Genealogical Analysis for Large Sample Sizes. *PLoS Comput Biol* 12
542 (5):e1004842.
543 Kim Y, Stephan W. 2002. Detecting a local signature of genetic hitchhiking along a
544 recombining chromosome. *Genetics* 160 (2):765-777.
545 Kumar S, Hedges SB. 1998. A molecular timescale for vertebrate evolution. *Nature* 392
546 (6679):917-920.
547 Lee TH, Guo H, Wang X, Kim C, Paterson AH. 2014. SNPhylo: a pipeline to construct a
548 phylogenetic tree from huge SNP data. *BMC Genomics* 15:162.
549 Lemere CA, Beierschmitt A, Iglesias M, Spooner ET, Bloom JK, Leverone JF, Zheng JB,
550 Seabrook TJ, Lourd D, Li D et al. 2004. Alzheimer's disease abeta vaccine reduces
551 central nervous system abeta levels in a non-human primate, the Caribbean
552 vervet. *Am J Pathol* 165 (1):283-297.
553 Li H. 2011. Improving SNP discovery by base alignment quality. *Bioinformatics* 27
554 (8):1157-1158.
555 Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler
556 transform. *Bioinformatics* 25 (14):1754-1760.
557 Ma D, Jasinska AJ, Feyertag F, Wijewardana V, Kristoff J, He T, Raehtz K, Schmitt CA, Jung
558 Y, Cramer JD et al. 2014. Factors associated with SIV transmission in a natural
559 African nonhuman primate host in the wild. *J Virol* 88 (10):6778-6792.
560 Malherbe H, Harwin R. 1958. Neurotropic virus in African monkeys. *Lancet* 272:530.

- 561 McFarland R, Barrett L, Boner R, Freeman NJ, Henzi SP. 2014. Behavioral flexibility of
562 vervet monkeys in response to climatic and social variability. *Am J Phys*
563 *Anthropol* 154 (3):357-364.
- 564 McGuire MT. 1974. The St. Kitts vervet (*Cercopithecus aethiops*). *J Med Primatol* 3
565 (5):285-297.
- 566 Mekonnen A, Bekele A, Fashing PJ, Lernould JM, Atickem A, Stenseth NC. 2012. Newly
567 discovered Bale monkey populations in forest fragments in southern Ethiopia:
568 Evidence of crop raiding, hybridization with grivets, and other conservation
569 threats. *Am J Primatol* 74 (5):423-432.
- 570 McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K,
571 Altshuler D, Gabriel S, Daly M et al. 2010. The Genome Analysis Toolkit: a
572 MapReduce framework for analyzing next generation DNA sequencing data.
573 *Genome Res* 20 (9):1297-1303.
- 574 Nei M, Li WH. 1979. Mathematical model for studying genetic variation in terms of
575 restriction endonucleases. *Proc Natl Acad Sci U.S.A.* 76 (10):5269-5273.
- 576 Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C. 2005. Genomic scans
577 for selective sweeps using SNP data. *Genome Res* 15 (11):1566-1575.
- 578 Nietlisbach P, Arora N, Nater A, Goossens B, Van Schaik CP, Krützen M. 2012. Heavily
579 male-biased long-distance dispersal of orang-utans (genus: *Pongo*), as revealed
580 by Y-chromosomal and mitochondrial genetic markers. *Mol Ecol* 21 (13):3173-
581 3186.
- 582 Nsabimana JM, Moutschen M, Thiry E, Meurens F. 2008. Human infection with simian
583 herpes B virus in Africa. *Sante* 18:3-8.
- 584 Olobo JO, Gicheru MM, Anjili CO. 2001. The African Green Monkey model for cutaneous
585 and visceral leishmaniasis. *Trends Parasitol* 17 (12):588-592.
- 586 Patterson N, Richter DJ, Gnerre S, Lander ES, Reich D. 2006. Genetic evidence for
587 complex speciation of humans and chimpanzees. *Nature* 441 (7097):1103-1108.
- 588 Pavlidis P, Živković D, Stamatakis A, Alachiotis N. 2013. SweeD: likelihood-based
589 detection of selective sweeps in thousands of genomes. *Mol Biol Evol* 30
590 (9):2224-2234.
- 591 Perelman P, Johnson WE, Roos C, Seuánez HN, Horvath JE, Moreira MA, Kessing B,
592 Pontius J, Roelke M, Rumpler Y et al. 2011. A molecular phylogeny of living
593 primates. *PLoS Genet* 7 (3):e1001342.
- 594 Pfeifer SP. From next-generation resequencing reads to a high-quality variant data set.
595 *Heredity*. doi 10.1038/hdy.2016.102.
- 596 Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, Veeramah KR,
597 Woerner AE, O'Connor TD, Santpere G et al. 2013. Great ape genetic diversity
598 and population history. *Nature* 499 (7459):471-475.
- 599 Raj A, Stephens M, Pritchard JK. 2014. fastSTRUCTURE: variational inference of
600 population structure in large SNP data sets. *Genetics* 197 (2):573-589.
- 601 Rimmer A, Phan H, Mathieson I, Iqbal Z, Twigg SFR, Wilkie AO, McVean G, Lunter G.
602 2014. Integrating mapping-, assembly- and haplotype-based approaches for
603 calling variants in clinical sequencing applications. *Nat Genet* 46 (8):912-918.
- 604 Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman
605 MW. 2002. Genetic structure of human populations. *Science* 298 (5602):2381-
606 2385.
- 607 Rudel LL, Haines J, Sawyer JK, Shah R, Wilson MS, Carr TP. 1997. Hepatic origin of
608 cholestryloleate in coronary artery atherosclerosis in African green monkeys.
609 Enrichment by dietary monounsaturated fat. *J Clin Invest* 100 (1):74-83.
- 610 Stone AC, Battistuzzi FU, Kubatko LS, Perry GH, Trudeau E, Lin H, Kumar S. 2010. More
611 reliable estimates of divergence times in Pan using complete mtDNA sequences
612 and accounting for population structure. *Philos Trans R Soc Lond, B, Biol Sci*, 365
613 (1556):3277-3288.
- 614 Struhsaker TT. 1967. Social structure among vervet monkeys (*Cercopithecus aethiops*).

- 615 Behaviour 29 (2):6-121.
616 Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA
617 polymorphism. Genetics 123 (3):585-595.
618 Teshima KM, Coop G, Przeworski M. 2006. How reliable are empirical genomic scans for
619 selective sweeps? Genome Res 16 (6):702-712.
620 Thornton K. 2003. Libsequence: a C++ class library for evolutionary genetic analysis.
621 Bioinformatics 19 (17):2325-2327.
622 Thornton KR, Jensen JD. 2007. Controlling the false positive rate in multi-locus genome
623 scans for selection. Genetics 175:737-750.
624 Toews DP, Brelsford A. 2012. The biogeography of mitochondrial and nuclear
625 discordance in animals. Mol Ecol 21 (16):3907-3930.
626 Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A,
627 Jordan T, Shakir K, Roazen D, Thibault J et al. 2013. From FastQ Data to High-
628 Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline.
629 Current Protoc Bioinformatics 43:11.10.1-11.10.33.
630 van de Waal E, Whiten A. 2012. Spontaneous emergence, imitation and spread of
631 alternative foraging techniques among groups of vervet monkeys. PLOS One
632 7:e47008. doi:10.1371/journal.pone.0047008.
633 Warren WC, Jasinska AJ, García-Pérez R, Svardal H, Tomlinson C, Rocchi M, Archidiacono
634 N, Capozzi O, Minx P, Montague MJ et al. 2015. The genome of the vervet
635 (Chlorocebus aethiops sabaeus). Genome Res 25 (12):1921-1933.
636 Wang K, Li M, Hakonarson H. 2010. ANNOVAR: Functional annotation of genetic variants
637 from next-generation sequencing data. Nucleic Acids Res 38 (16):e164.
638 doi:10.1093/nar/gkq603.
639 Watterson GA. 1975. On the number of segregating sites in genetical models without
640 recombination. Theor Popul Biol 7 (2):256-276.
641 Weir BS, Cockerham CC. 1984. Estimating F-Statistics for the Analysis of Population
642 Structure. Evolution 38 (6):1358-1370.
643 Wertheim JO, Smith MD, Smith DM, Scheffler K, Kosakovsky Pond SL. 2014. Evolutionary
644 Origins of Human Herpes Simplex Viruses 1 and 2. Mol Bio Evol 31:2356-2364.
645 Wise CA, Sraml M, Rubinsztein DC, Easteal S. 1997. Comparative nuclear and
646 mitochondrial genome diversity in humans and chimpanzees. Mol Biol Evol
647 14:707-716.
648 Ye GJ, Vaughan KT, Vallee RB, Roizman B. 2000. The herpes simplex virus 1 U(L)34
649 protein interacts with a cytoplasmic dynein intermediate chain and targets
650 nuclear membrane. J Virol 74 (3):1355-1363.
651 Yu N, Jensen-Seaman MI, Chemnick L, Ryder O, Li WH. 2004. Nucleotide diversity in
652 gorillas. Genetics 166 (3):1375-1383.
653 Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BC. 2012. A High-performance
654 Computing Toolset for Relatedness and Principal Component Analysis of SNP
655 Data, Bioinformatics 28 (14):3326-3328.
656

657 Table 1: Summary of SNP data.

658

659

Chromosome	Length	# SNPs	Ts/Tv	% Accessible	SNPs/kb
1	126,035,930	88,368	2.9	18.4	0.7
2	90,373,283	69,432	3.1	20.0	0.8
3	92,142,175	61,810	2.6	16.1	0.7
4	91,010,382	59,256	2.7	16.0	0.7
5	75,399,963	58,652	2.9	19.2	0.8
6	50,890,351	35,519	3.2	16.7	0.7
7	135,778,131	74,807	2.5	14.6	0.6
8	139,301,422	102,193	2.8	17.7	0.7
9	125,710,982	92,784	2.9	18.9	0.7
10	128,595,539	81,526	2.7	16.8	0.6
11	128,539,186	87,591	2.8	17.0	0.7
12	108,555,830	74,106	2.9	18.1	0.7
13	98,384,682	60,334	2.6	15.6	0.6
14	107,702,431	75,455	2.9	17.9	0.7
15	91,754,291	59,772	2.6	16.3	0.7
16	75,148,670	52,909	3.2	19.2	0.7
17	71,996,105	51,115	2.7	17.2	0.7
18	72,318,688	54,162	2.8	17.9	0.7
19	33,263,144	29,102	3.3	21.2	0.9
20	130,588,469	96,660	3.0	18.8	0.7
21	127,223,203	83,854	2.7	16.8	0.7
22	101,219,884	68,197	2.8	17.6	0.7
23	82,825,804	56,430	2.8	17.7	0.7
24	84,932,903	55,184	2.8	16.6	0.7
25	85,787,240	57,881	2.8	17.7	0.7
26	58,131,712	41,674	2.9	18.8	0.7
27	48,547,382	33,665	2.8	17.1	0.7
28	21,531,802	14,786	3.3	16.5	0.7
29	24,206,276	18,419	2.8	18.0	0.8
total	2,607,895,860	1,795,643	2.8	17.5	0.7

660

661

662 Table 2: Summaries of genetic variation.
 663

	<i>C. a. aethiops</i>	<i>C. a. cynosurus</i>	<i>C. a. pygerythrus</i>	<i>C. a. sabaeus</i>	<i>C. a. tantalus</i>
$\pi_{\text{intergenic}}$	3.7×10^{-4}	5.2×10^{-4}	5.1×10^{-4}	4.7×10^{-4}	5.2×10^{-4}
π_{exonic}	1.4×10^{-4}	2.1×10^{-4}	2.1×10^{-4}	1.9×10^{-4}	1.9×10^{-4}
$\theta_{W; \text{intergenic}}$	3.2×10^{-4}	4.7×10^{-4}	4.6×10^{-4}	4.3×10^{-4}	4.7×10^{-4}
$\theta_{W; \text{exonic}}$	1.3×10^{-4}	1.9×10^{-4}	1.9×10^{-4}	1.7×10^{-4}	1.8×10^{-4}
N_e	15,849	21,921	21,733	19,824	22,146
$D_{\text{intergenic}}$	0.27	0.21	0.23	0.19	0.23
D_{exonic}	0.06	0.06	0.07	0.06	0.06
$F_{st}(\text{aet})$	-	0.55	0.56	0.60	0.53
$F_{st}(\text{cyn})$	0.55	-	0.16	0.45	0.33
$F_{st}(\text{pyg})$	0.56	0.16	-	0.46	0.35
$F_{st}(\text{sab})$	0.60	0.45	0.46	-	0.40
$F_{st}(\text{tan})$	0.53	0.33	0.35	0.40	-

664 Nucleotide diversity π (Nei and Li, 1979) and Watterson's estimate of θ , θ_W ,
 665 (Watterson 1975) were calculated in the non-coding, non-repetitive ($\pi_{\text{intergenic}}$; $\theta_{W; \text{intergenic}}$) and exonic (π_{exonic} ; $\theta_{W; \text{exonic}}$) parts of the genome for each population (aet: *C. a. aethiops*; cyn: *C. a. cynosurus*; pyg: *C. a. pygerythrus*; sab: *C. a. sabaeus*; tan: *C. a. tantalus*)
 666 using the libsequence package msstats v.0.3.4 (Thornton 2003). Effective population
 667 sizes (N_e) were estimated from the data by fixing the mutation rate μ to 5.9×10^{-9} per
 668 base pair per generation (i.e., the mutation rate observed in rhesus macaque as there is
 669 no direct estimate for μ available in vervet monkeys; Hernandez *et al.* 2007). Using
 670 VCFtools v.0.1.12b (Danecek *et al.* 2011), Tajima's D (Tajima 1989) was calculated (in
 671 10kb windows along the genome) to test for deviations from the neutral equilibrium
 672 site frequency distribution and Weir and Cockerham's fixation index F_{st} was calculated
 673 between each pair of populations to assess population differentiation.
 674
 675
 676

677 Table 3: Putatively neutral fixed differences between the populations.

678

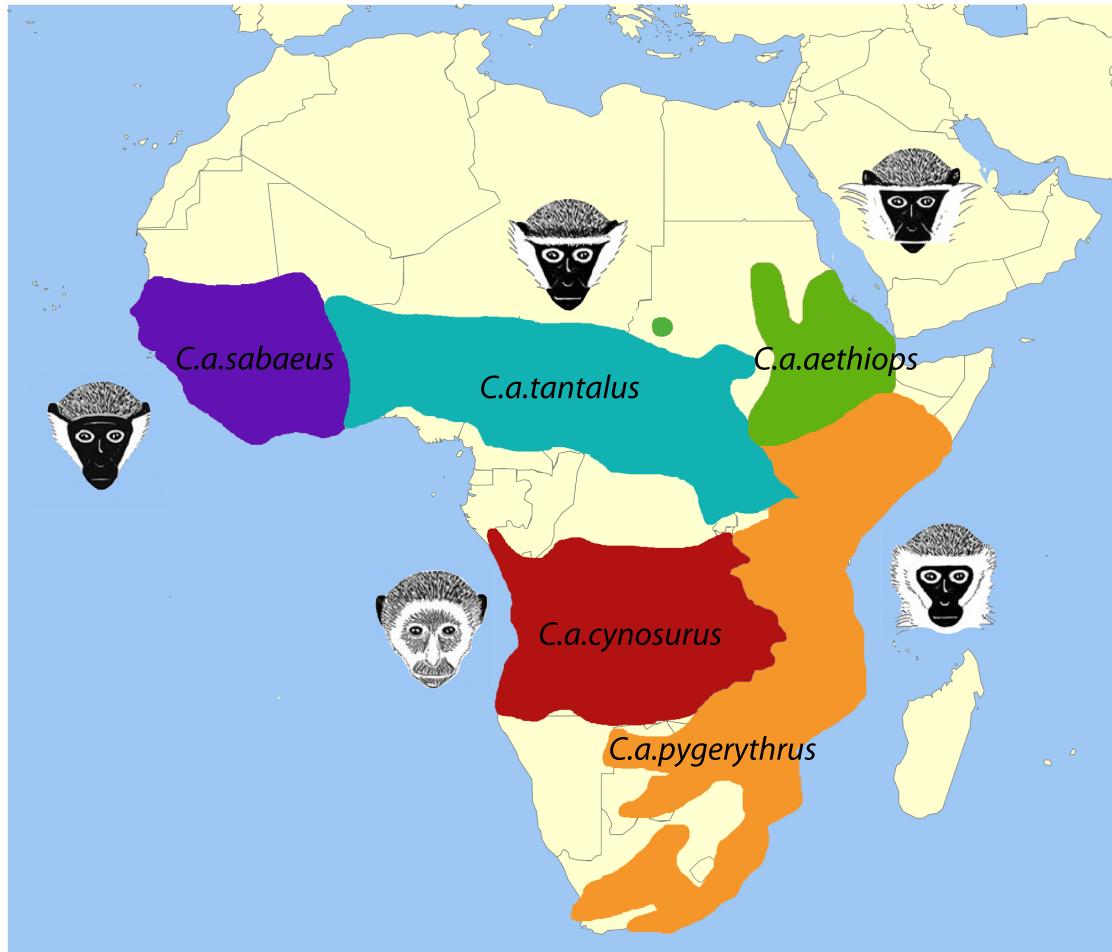
	<i>C. a. aethiops</i>	<i>C. a. cynosurus</i>	<i>C. a. pygerythrus</i>	<i>C. a. sabaeus</i>	<i>C. a. tantalus</i>
<i>C. a. aethiops</i>	-	52,841	53,239	62,716	53,180
<i>C. a. cynosurus</i>	52,841	-	14,458	32,849	24,427
<i>C. a. pygerythrus</i>	53,239	14,458	-	33,631	25,241
<i>C. a. sabaeus</i>	62,716	32,849	33,631	-	24,098
<i>C. a. tantalus</i>	53,180	24,427	25,241	24,098	-

679

680

- 681 Figure 1: Geographic distribution of the five African vervet monkey populations (red: *C.
682 a. cynosurus*; orange: *C. a. pygerythrus*; green: *C. a. aethiops*; turquoise: *C. a. tantalus*;
683 purple: *C. a. sabaeus*; adapted from Haus *et al.* 2013).
- 684
685 Figure 2: Private and shared segregating sites per population (red: *C. a. cynosurus*;
686 orange: *C. a. pygerythrus*; green: *C. a. aethiops*; turquoise: *C. a. tantalus*; purple: *C. a.
687 sabaeus*). Note that the sizes of the areas are not proportional to the magnitude of the
688 numbers.
- 689
690 Figure 3: Principal component analysis of all sampled individuals in the different
691 populations (red: *C. a. cynosurus*; orange: *C. a. pygerythrus*; green: *C. a. aethiops*;
692 turquoise: *C. a. tantalus*; purple: *C. a. sabaeus*). Data was thinned to exclude SNPs with an
693 $r^2 > 0.2$ in order to avoid a strong influence of SNP clusters in the PCA. Percentages
694 indicate the variance explain by each principle component.
- 695
696 Figure 4: Demographic history of the vervet monkey. Divergence times have been
697 estimated using a molecular clock based on putatively neutral, fixed differences between
698 the genomes of the populations (Table 3), assuming that mutations occurred at a
699 constant rate of 5.9×10^{-9} per base pair per generation among lineages (Hernandez *et al.*
700 2007) and that the generation time is 8.5 years (Warren *et al.* 2015). Effective
701 population sizes (provided at the tip of the branches) were estimated from the data by
702 fixing the mutation rate μ to 5.9×10^{-9} per base pair per generation (Hernandez *et al.*
703 2007). The figure was generated using PopPlanner (Ewing *et al.* 2015).
- 704
705 Figure 5: Fit of the data to a standard neutral equilibrium model, using the described
706 putatively neutral class of SNPs. a) Levels of genetic differentiation between populations
707 (A: *C. a. aethiops*; C: *C. a. cynosurus*; P: *C. a. pygerythrus*; S: *C. a. sabaeus*; T: *C. a. tantalus*)
708 observed in the data (red dots) are expected under this model (simulation results
709 shown as boxplots). b) Levels of nucleotide diversity in the vervet monkey populations.
710 c) Tajima's D is skewed towards positive values in all populations compared to the
711 standard equilibrium model due to a deficit of low frequency variants in the data, which
712 is an artifact from utilizing a multi-sample genotype calling strategy on low coverage
713 sequencing data (Han *et al.* 2013).
- 714
715 Figure 6: Likelihood surfaces of the CLR test calculated by SweeD (Pavlidis *et al.* 2013)
716 for chromosome 21 per population (green: *C. a. aethiops*; red: *C. a. cynosurus*; orange: *C.
717 a. pygerythrus*; purple: *C. a. sabaeus*; turquoise: *C. a. tantalus*) on a megabase (Mb) scale.
718 The dashed black horizontal line is the significance threshold of the test (based on a
719 standard equilibrium model without migration; see Materials and Methods). A selective
720 sweep near the gene *DYNC1I1* (dashed black vertical line; SweeD-score: >14.8) is
721 common to multiple populations.
- 722

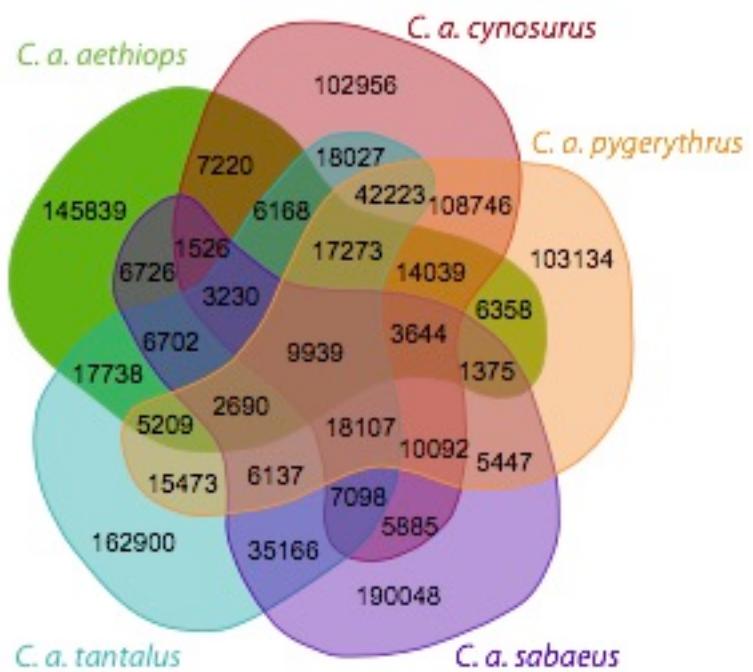
723 Figure 1
724



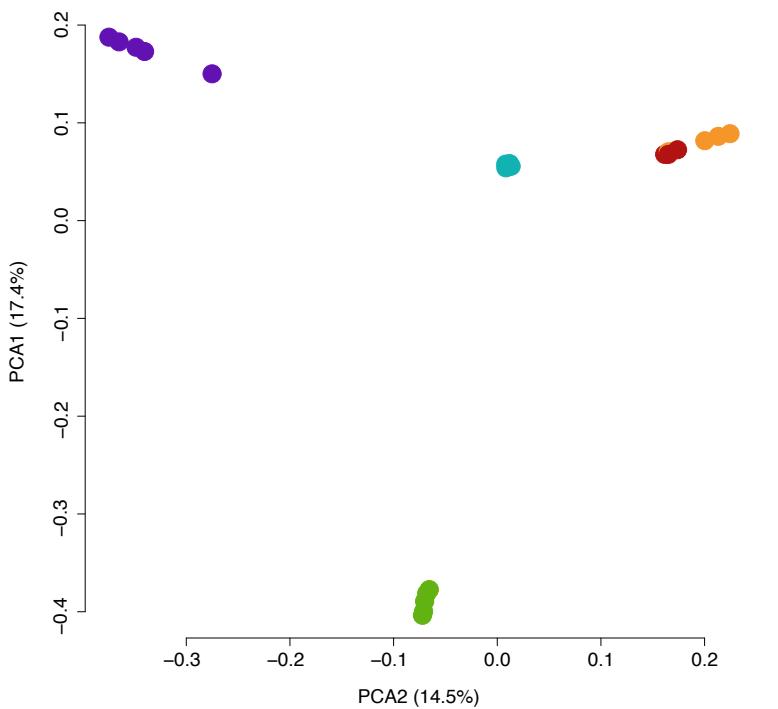
725
726

727 Figure 2
728

729
730



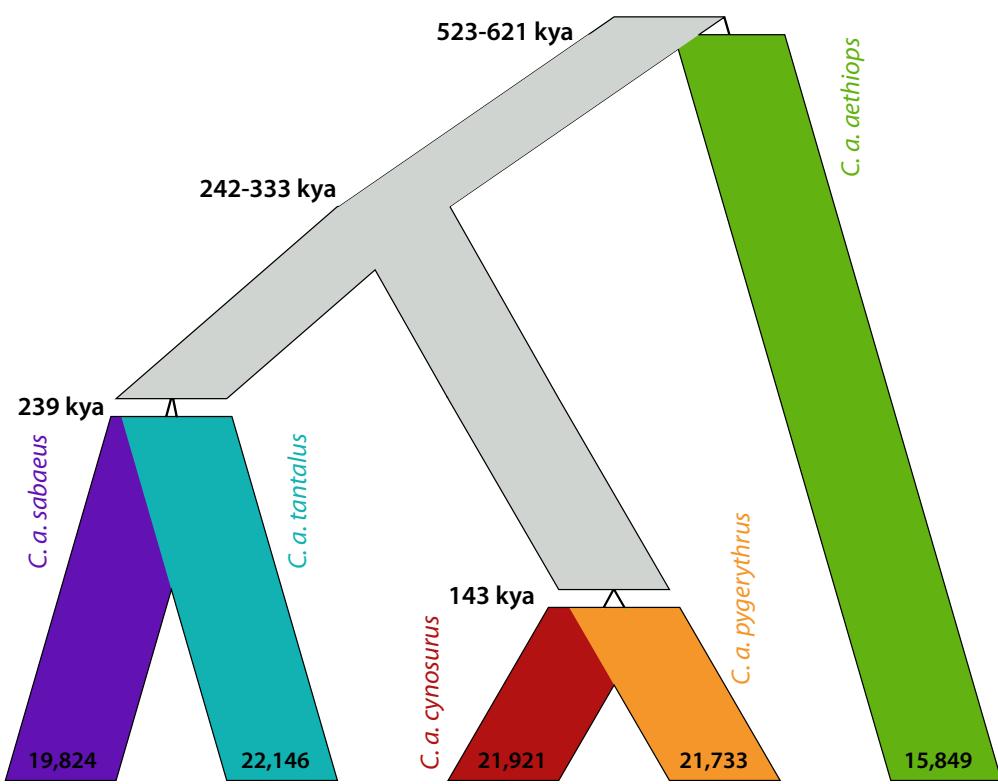
731 Figure 3
732



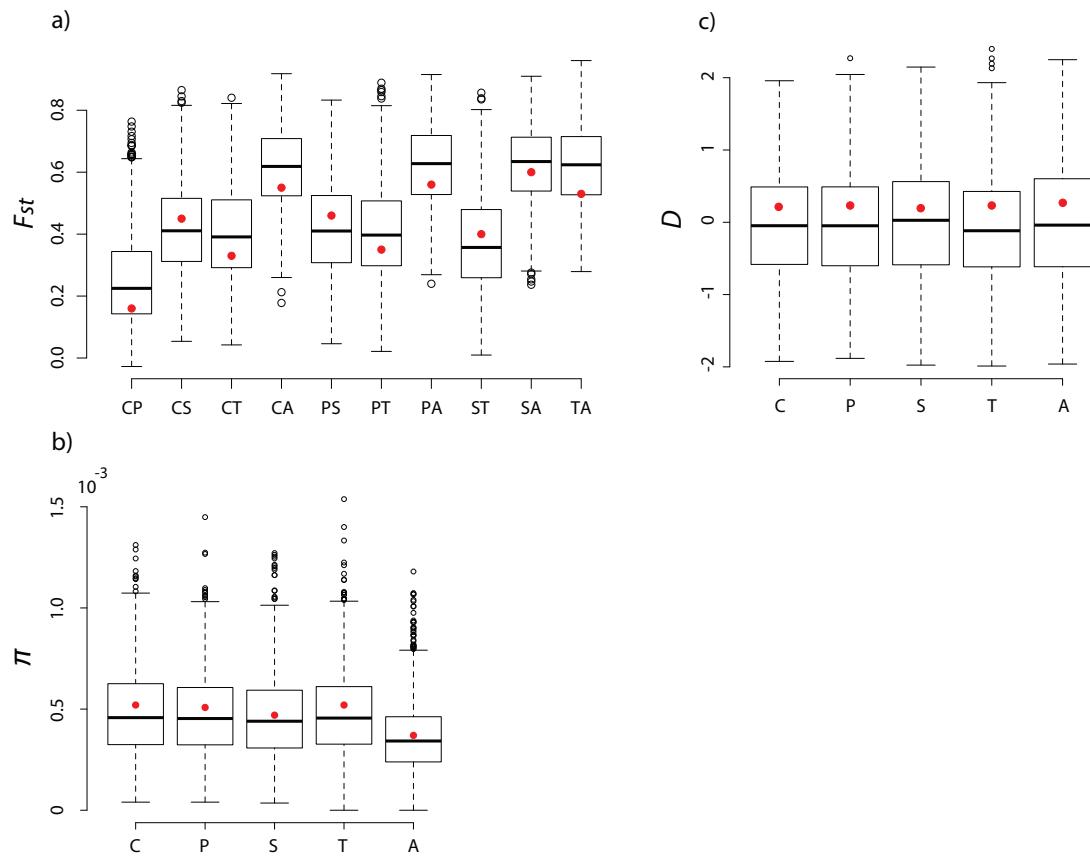
733
734

735 Figure 4
736

737
738



739 Figure 5
740



741
742

743 Figure 6
744

