The genome of the endangered dryas monkey provides new insights into the evolutionary history of the vervets

T. van der Valk[1], C. M. Gonda[1], J. Hart[2], T. Hart[2], K. Detwiler[3], K. Guschanski[1]

[1] Animal Ecology, Department of Ecology and Genetics, Evolutionary Biology Centre, Uppsala University, Norbyvägen 18D, 752 36, Uppsala, Sweden

[2] Lukuru Wildlife Research Foundation, Tshuapa-Lomami-Lualaba Project, Kinshasa, Democratic Republic of the Congo

[3] Departments of Anthropology and Biological Sciences, Florida Atlantic University, Boca Raton, FL 33431, USA

Corresponding authors: Tom van der Valk, tom.vandervalk@ebc.uu.se, Katerina Guschanski katerina.guschanski@ebc.uu.se

Abstract

Genomic data can be a powerful tool for inferring ecology, behaviour and conservation needs of highly elusive species, particularly when other sources of information are hard to come by. Here we focus on the dryas monkey, an endangered primate endemic to the Congo Basin with cryptic behaviour and possibly less than 250 remaining individuals. Using whole genome data we show that the dryas monkey represents a sister lineage to the vervet monkeys and has diverged from them at least 1 million years ago with additional bi-directional gene flow 590,000 – 360,000 years ago. After bonobo-chimpanzee admixture, this is the second reported case of gene flow that most likely involved crossing the Congo River, a strong dispersal barrier. As the demographic history of bonobos and dryas monkey shows similar patterns of population increase during this time period, we hypothesise that the fluvial topology of the Congo River might have been more dynamic than previously recognised. As a result of dryas monkey - vervet admixture, genes involved in resistance to the simian immunodeficiency virus (SIV) have been exchanged, possibly indicating adaptive introgression. Despite the presence of several homozygous loss-of-function mutations in genes associated with reduced sperm mobility and immunity, we find high genetic diversity and low levels of inbreeding and genetic load in the studied dryas monkey individual. This suggests that the current population carries sufficient genetic variability for the long-term survival of this species. We thus provide an example of how genomic data can directly improve our understanding of elusive species.

Introduction

The dryas monkey (*Cercopithecus dryas*) is a little-known species of guenons endemic to the Congo Basin, previously only recorded from a single location (Figure 1). The recent discovery of a second geographically distinct population has led to the elevation of its conservation status from critically endangered to endangered, however little is known about its true population size, distribution range, behaviour, ecology and evolutionary history (IUCN 2019). Based on pelage coloration, the dryas monkey was first classified as the central African representative of the Diana monkey (*Cercopithicus diana*) (Schwarz 1932). Later examinations of the few available specimens suggested that the dryas monkey should be classified as a unique *Cercopithecus* species (Colyn et al. 1991). More recently, Guschanski *et al.* 2013 described the mitochondrial genome of the dryas monkey type specimen, preserved at the Royal Museum for Central Africa (Tervuren, Belgium), providing the first genetic data for this species. The mitochondrial genome-based phylogeny places the dryas monkey within the vervet (*Chlorocebus)* genus, supporting previously suggested grouping based on similarities in feeding behavior, locomotion and cranial morphology (Kuroda et al. 1985; Jonathan Kingdon, David Happold, Thomas Butynski et al. 2013).

The vervets currently consist of six recognised species: *sabaeus, aethiops, tantalus, hilgerti, pygerythrys* and *cynosuros* (Zinner et al. 2013). They are common in savannahs and riverine forests throughout sub-Saharan Africa, as well as on several Caribbean islands, where they were introduced during the colonial times (Figure 1A). However, the dryas monkey is geographically isolated from all vervets and has numerous highly distinct morphologically characteristics (Zinner et al. 2013). As vervets are characterized by a dynamic demographic history with extensive hybridisation (Svardal et al. 2017), including female-mediated gene flow, transfer of mitochondrial haplotypes between species can result in discordance between a mitochondrial tree and the true species history. Thus, the phylogenetic placement of the dryas monkey remains uncertain.

With only two known populations and possibly fewer than 250 individuals, the dryas monkey is endangered and of significant conservation concern (Hart et al. 2019). The goal of this study was thus twofold: First, reconstruct the evolutionary and demographic history of the dryas monkey in relation to that of the vervets and second, asses the long-term genetic viability of this species. To this end, we sequenced the genome of a male dryas monkey at high coverage (33X), which represents the first genome-wide information available for this cryptic and little-known species.

2.0| Methods

2.1| Sample collection and sequencing

A dryas monkey tissue sample was obtained from an individual from the Lomami population in eastern DRC (Figure. 1A) and exported to the United States under approved country-specific permits, where DNA was extracted using the Qiaqen DNeasy Blood & Tissue kit, following the manufactures protocol. For library preparation and sequencing the DNA was sent to Uppsala University, Sweden. The Illumina TruSeq DNA PCR-free kit was used for standard library preparation and the sample was sequenced on one lane of the Illumina HiseqX platform (2x 150bp). In addition, we obtained previously published FASTQ data for all currently recognised vervet species from mainland Africa (Warren et al. 2015; Svardal et al. 2017) and divided them into two subsets: 23 *sabaeus*, 16 *aethiops*, 11 *tantalus*, 6 *hilgerti*, 16 *cynosuros* and 51 *pygerythrus* individuals sequenced on the Hiseq2000 platform at low coverage (the low-coverage dataset), and one individual of each vervet species sequenced on the Genome-Analyzer II platform at medium coverage (the medium-coverage dataset). We also obtained high coverage genomes of two rhesus macaques, to be used as outgroup in our analysis (Xue et al. 2016).

2.2| Alignment, variant detection and filtering

All FASTQ data was adapter and quality trimmed using Trimmomatic on recommend settings (Bolger et al. 2014) and then aligned against the *Chlorocebus sabaeus* reference genome (ChlSab1.1) (Warren et al. 2015) using bwa-mem (Li 2013) with default settings. Samtools was used to filter out reads below a mapping quality of 30 (phred-scale) (Li et al. 2009). Next, reads were realigned around indels using GATK IndelRealigner (DePristo et al. 2011; McKenna et al. 2010) and duplicates marked using Picard2.10.3 (https://broadinstitute.github.io/picard/). We obtained a genome wide coverage of 33X for the dryas monkey, 27X and 31X for the rhesus macaques, 1.9-6.8X for the low-coverage, and 7.4-9.8X for the medium coverage vervet genomes. Next, we called single nucleotide polymorphisms (SNPs) with GATK UnifiedCaller outputting all sites (DePristo et al. 2011; McKenna et al. 2010). Raw variant calls were then hard filtered following the GATK best practices (Van der Auwera et al. 2013). Additionally, we removed all sites below quality 30 (phred-scale), those with more than three times average genome-wide coverage across the dataset, sites for which more than 75% of samples had a non-reference allele in a heterozygous state, indels and sites within highly repetitive regions as identified from the repeatmask-track for ChlSab1.1 using vcftools (Danecek et al. 2011).

94 **2.5| Autosomal phylogeny**

95 We assessed the phylogeny of the dryas monkey in relation to the vervets using genome-wide methods
96 on either the medium or low coverage dataset. We obtained pseudo-haploid consensus sequences
97 from the medium-coverage dataset for each autosome and each species (including one rhesus
98 macaque as outgroup) using GATK-FastaAlternateReferenceMaker (Van der Auwera et al. 2013). We
99 then concatenated autosomal sequences into a multi-species alignment file. Next, non-overlapping
100 350kb genomic windows were extracted from the alignment and filtered for missing sites using PHAST
101 v1.4 (Hubisz et al. 2011). After filtering we excluded all windows with a length below 200kb resulting
102 in a final data set consisting of 3703 genomic windows (mean sequence length of 298kb ±SD 18.5kb).
103 Individual approximately-maximum-likelihood gene trees were then generated for the set of windows
104 using FastTree2 v2.1.10 (Price et al. 2010) and the GTR model of sequence evolution. Next, we
105 constructed a coalescent species tree from the obtained gene trees with ASTRAL v5.6.2 (Zhang et al.
106 2018) on default parameters. ASTRAL estimates branch length in coalescent units and uses local
107 posterior probabilities to compute branch support for the species tree topology, which gives a more
108 reliable measure of support than the standard multi-locus bootstrapping (Sayyari & Mirarab 2016).
109 Additionally, we obtained an extended majority rule consensus tree using the CONSENSE algorithm in
110 PHYLIP v3.695 (Felsenstein 2005). Finally, to explore phylogenetic conflict among the different gene
111 trees, we created consensus networks with SplitsTree v4 using different threshold values (15% 20%
112 and 25%) (Huson & Bryant 2006).

113 Next, we used the low-coverage dataset (n = 123) to assess the genetic similarity between all
114 individuals and the dryas monkey by running a Principal Components Analysis (PCA) on all filtered
115 autosomal bi-allelic sites called in all individuals using PLINK1.9 with default settings (Purcell et al.
116 2007). We estimated (sub)species divergence times using population level data by applying a clustering
117 algorithm as in (Warren et al. 2015). Next, we obtained pseudo-haploid aligned consensus sequences
118 for all 123 low-coverage vervets, the rhesus and the dryas monkey using GATK-
119 FastaAlternateReferenceMaker (Van der Auwera et al. 2013). The R package *ape* was then used to
120 calculate pairwise distances among all individuals using the Tamura and Nei 1993 model, which allows
121 for different rates of transitions and transversions, heterogeneous base frequencies, and between-site
122 variation of the substitution rate (Paradis et al. 2004; Tamura & Nei 1993). We used the R package
123 *phangorn* to construct a UPMG phylogeny form the resulting pairwise distance matrix (Schliep 2011).
124 We obtained estimates of divergence under the assumption of no gene flow as in (Warren et al. 2015),
125 calibrating the tree to the rhesus outgroup, setting the divergence time between the rhesus (*Papionini*)
126 and vervets (*Cercopithecini*) at 13.7 million years ago (Hedges et al. 2015).

127 **2.4| PSMC**

128 To infer long-term demographic history of the studied species, we used a pairwise sequentially
129 Markovian coalescent model (PSMC) (Li & Durbin 2011) based on all medium-coverage vervet
130 genomes and the high coverage dryas monkey genome. We excluding sex chromosomes, repetitive
131 regions and all sites for which read depth was less than five and higher than 100. We scaled the PSMC
132 output using a generation time of 8.5 years (Warren et al. 2015) and a mutation rate of 0.94×10^{-8} per
133 site per generation (Pfeifer 2017). Bootstrap replicates (n=100) were performed for the high-coverage
134 dryas monkey genome by splitting all chromosomal sequences into smaller segments using the splitfa
135 implementation in the PSMC software and then randomly sampling with replacement from these
136 fragments (Li & Durbin 2011). Our dryas monkey genome coverage (33X) differed strongly from that
137 of the medium-coverage vervets (7.4-9.8X). As limited coverage is known to biases PSMC results
138 (Nadachowska-Brzyska et al. 2016), we down-sampled the dryas monkey genome to a similar coverage
139 (8X) as the medium-coverage vervet genomes and repeated the PSMC analysis, allowing for qualitative
140 comparisons between species. Demographic estimates from PSMC can also be biased by admixture
141 events between divergent populations, giving a false signal of population size change (Hawks 2017).
142 We thus removed all putatively introgressed regions (see below) from the dryas monkey genome and
143 re-run the PSMC analysis.

144

145 **2.6| Mitochondrial phylogeny**

146 We *de-novo* assembled the mitochondrial genomes for the dryas monkey, the medium-coverage
147 vervets, and rhesus macaque with NOVOplasty (Dierckxsens et al. 2016) at recommend settings, using
148 a K-mer size of 39 and the *Chlorocebus sabaeus* (ChlSab1.1) mitochondrial reference genome as seed
149 sequence. We also included the previously published mitochondrial genome sequence of the dryas
150 monkey type specimen (Guschanski et al. 2013) and the olive baboon mitochondrial genome
151 (Panu_3.0) used for divergence calibration. A mitochondrial phylogeny was then obtained as in *van*
152 *der Valk* et al. 2018. Briefly, mitochondrial genomes were aligned using clustal omega with default
153 settings (Sievers et al. 2011). We partitioned the alignment into coding genes (splitting triplets into
154 1st + 2nd and 3rd position), rRNAs, tRNAs, and non-coding regions using the ChlSab1.1 annotation in
155 Geneious 10.1.2 (https://www.geneious.com). Phylogenetic tree reconstruction and divergence dating
156 was carried out with BEAST2.4.6 (Bouckaert et al. 2014), using the best fitting substitution model for
157 each partition as identified with PartitionFinder2 (Lanfear et al. 2017) and enforcing a strict molecular
158 clock. Birth-rate and clock-rate priors were set as gamma distribution with $\alpha = 0.001$ and $\beta = 1000$ as
159 recommend for mitochondrial phylogenetic analysis (Bouckaert et al. 2014). To calibrate the tree,
160 priors for the split time between rhesus and baboon were set as a log-normal distribution with

161 M = 2.44 and S = 0.095, which corresponds to a divergence time of 9.98– 13.2 Mya (95% CI) (Hedges et
162 al. 2015). The Bayesian model was run for an MCMC length of 500 million and we used Tracer 1.6
163 (Rambaut et al. 2014) to confirm run convergence and obtain probability distributions. The consensus
164 and locus-specific trees were plotted in DensiTree (Bouckaert 2010).
165
166 **2.7| Y-chromosome phylogeny**
167 The mammalian Y chromosome sequence is enriched for repeats and palindromes, and thus accurate
168 assembly from short-read data is challenging (Tomaszkiewicz et al. 2017; Kuderna et al. 2019). We
169 therefore obtained partial Y-chromosome consensus sequences using the filtered SNP calls. First, we
170 identified all male individuals in our low-coverage dataset using the ratio of X-chromosome to
171 autosomal coverage (Figure S2). Next, GATK FastaAlternateReferenceMaker was used to obtain a Y-
172 chromosome consensus sequence for each male individual using the filtered variant calls as input. We
173 masked all sites for which at least one individual showed a heterozygous call, as these represent SNP-
174 calling errors. Additionally, we masked all repetitive regions and all sites for which one or more female
175 individuals also showed a variant call, as these regions are likely enriched for SNP-errors due to
176 mismappings. A maximum-likelihood tree was then constructed in MEGAX (Kumar et al. 2018) using
177 the Tamura-Nei 1993 model (Tamura & Nei 1993), running 1000 bootstrap replicates and only
178 including sites called in all male individuals. The tree was time calibrated using the rhesus Y-
179 chromosome as outgroup, a rhesus (*Papionini*) - vervets (*Cercopithecini*) divergence time of 13.7
180 million years (Hedges et al. 2015) and assuming a uniform mutation rate.
181
182 **2.8| Gene flow**
183 We performed a model-free test of unbalanced allele sharing between the vervet individuals and the
184 dryas monkey (D-statistic) (Green et al. 2010) using all autosomal bi-allelic sites called in all the low-
185 coverage individuals and dryas monkey. Pairwise D-statistics were run for all combinations as
186 H1,H2,H3,H4 with dryas monkey as the third ingroup (H3) and the rhesus macaque as the
187 representative of the ancestral variant (H4) (we excluded sites for which the two rhesus-macaque
188 genomes were not identical). Additionally, we calculated frequency-stratified D-statistics on
189 population level as in (de Manuel et al. 2016) to obtain estimates of the direction of gene flow, again
190 using the dryas monkey genome as H3 and rhesus as H4.
191 A model-based estimate of gene flow was obtained by constructing a maximum likelihood (ML) tree
192 using TreeMix v. 1.12 (Pickrell & Pritchard 2012), accounting for linkage disequilibrium (LD) by grouping
193 sites in blocks of 1,000 SNPs (-k 1000). Based on the previous phylogenetic inferences, the dryas
194 monkey was set as root and a round of rearrangements was performed after all populations were
195 added to the tree (-global). Standard errors (-se) and bootstrap replicates (-bootstrap) were used to

196 evaluate the confidence in the inferred tree topology and the weight of migration events. After
197 constructing a maximum likelihood tree, migration events were added (-m) and iterated 50 times for
198 each value of 'm' (1-10) to check for convergence in terms of the likelihood of the model as well as the
199 explained variance following each addition of a migration event. The inferred maximum likelihood
200 trees were visualised with the in-build TreeMix R script plotting functions.
201 Additionally, we performed maximum likelihood estimation of individual ancestries using ADMIXTURE
202 (Alexander et al. 2009) based on all autosomal bi-allelic SNPs called in all individuals, filtered for MAF
203 >5% and LD pruned using plink (--indep 50 10 2). The optimal number of clusters (K) in the admixture
204 graph was identified by running 5-fold cross-validations for K 1-10.
205
206 **2.9| Identifying introgressed regions**
207 To identify putatively introgressed regions in all vervet individuals, we performed a screen for such
208 segments following a strategy outlined in (Martin et al. 2015). Briefly, in sliding windows of 10kb we
209 calculated F_d statistics (which is related to D-statistic but not subject to the same biases as D when
210 calculated in sliding windows, (Martin et al. 2015)) using all *sabaeus* individuals from Gambia as
211 ingroup (H1) (as these showed the least amount of shared derived alleles with the dryas monkey) and
212 $Dxy_{dryas\text{-}x}$, and $Dxy_{sabaeus(Gambia)\text{-}x}$, where x refers to the focal individual. Next we calculated the average
213 ratio and standard deviation for $Dxy_{dryas\text{-}x}$ / $Dxy_{sabaeus(Gambia)\text{-}x}$ across all windows. As Dxy is a measure of
214 sequence divergence, introgressed windows between the dryas monkey and non-*sabaeus* vervets are
215 expected to have a relative low $Dxy_{dryas\text{-}x}$ and relatively high $Dxy_{sabaeus(Gambia)\text{-}x}$. Windows showing an
216 excess of shared derived alleles with the dryas monkey (Z_{Fd} score > 2) and unusual low divergence
217 towards the dryas monkey ($Dxy_{Dryas\text{-}x}$ / $Dxy_{sabaeus(Gambia)\text{-}x}$) > (2·SD ± genome wide $Dxy_{Dryas\text{-}x}$ /
218 $Dxy_{sabaeus(Gambia)\text{-}x}$) were flagged as putatively introgressed.
219
220 **2.10| Gene ontology enrichment of introgressed genes**
221 Using the *Chlorocebus sabaeus* genome annotation (Warren et al. 2015) we obtained for each
222 individual all genes within putatively introgressed windows. A gene ontology enrichment was run for
223 all putatively introgressed genes fixed in all non-*sabaeus* individuals in Blast2GO using Fisher's exact
224 test (Gotz et al. 2008). Next, for all genes we obtained selection coefficient from Svardal *et al.* 2017,
225 which were calculated using a multilocus test of allele frequency differentiation (identifying regions in
226 the genome where the change in allele frequency at the locus occurred too quickly to be explained by
227 drift: *XP-CLR* selection scores) (Chen et al. 2010). Candidate genes for adaptive introgression were then
228 identified as those with high gene-selection scores and high frequency in the recipient population.

229 **2.11| Heterozygosity and inbreeding**

230 We measured genome-wide autosomal heterozygosity for all individuals with average genome
231 coverage > 3X using realSFS as implemented in ANGSD, considering only uniquely mapping reads (-
232 uniqueOnly 1) and bases with quality score above 19 (-minQ 20) (Fumagalli et al. 2013; Korneliussen
233 et al. 2014). ANGSD uses genotype-likelihoods, rather than variant calls, allowing for the incorporation
234 of statistical uncertainty in low-coverage data and shows high accuracy in estimating heterozygosity
235 for genomes above 3X coverage (Fumagalli 2013; van der Valk et al. 2019). Next, we used PLINK1.9
236 (Purcell et al. 2007) to identify stretches of the genome in complete homozygosity (runs of
237 homozygosity: ROH) for all individuals with average genome coverage > 3X. To this end, we ran sliding
238 windows of 50 SNPs on the VCF files of all included genomes, requiring at least one SNP per 50kb. In
239 each individual genome, we allowed for a maximum of one heterozygous and five missing calls per
240 window before we considered the ROH to be broken.

241

242 **2.12| Genetic load**

243 We used the variant effect predictor tool (McLaren et al. 2016) to identify loss-of-function mutations
244 (transcript ablation, splice donor variant, splice acceptor variant, stop gained, frameshift variant,
245 inframe insertion, inframe deletion, and splice region variant), missense and synonymous mutations
246 on the filtered SNP calls. As an indication of mutational load, for each individual we counted the
247 number of genes containing one or more loss-of-function and the total number of missense mutations
248 divided by the number of synonymous mutations (Fay et al. 2001). We excluded all missense mutations
249 within genes containing a loss-of-function mutation, as these are expected to behave effectively
250 neutral. Dividing by the number of synonymous mutations mitigates species-specific biases, such as
251 mapping bias due to the fact that the reference genome was derived from a *sabaeus* individual,
252 coverage differences, and mutation rate (Xue et al. 2015).

253 **3.0 | Results and discussion**

254

255 **3.1| The dynamic demographic history of the dryas monkey and vervets**

256 In this study we present the first genome sequence of the dryas monkey, and show that it is a sister
257 lineage to the vervets (Figure 1B, S1, S4B). Multiple phylogenomic approaches (MSC, pairwise
258 differences, FastTree, SplitTree, and TreeMix) unambiguously support the same tree topology, thus
259 contradicting the suggested placement within the *Chlorocebus* genus as inferred from the
260 mitochondrial data (Guschanski et al. 2013). After analysing 3703 gene trees from autosomal genomic
261 windows, we obtained a multi-species-coalescent tree with maximum support values (lpp=1.0) for all
262 nodes (Figure S1A). Our topology is consistent with the vervet phylogeny previously reported by
263 Warren *et al.* 2015. Although a majority-rule consensus tree (Figure S1B) and network analyses at
264 different threshold values (Figure S1C) showed some poorly resolved nodes within the vervet clade,
265 the position of dryas monkey as sister lineage to all vervets remains unambiguous (Figure S1).
266 The divergence times within the vervet genus inferred by us are generally more ancient (+~10% for the
267 oldest nodes, Figure. 1B) than reported by Warren *et al.* 2015, which is likely explained by our use of
268 population level data, whereas one representative genome per species was used by Warren *et al.* 2015.
269 We estimate that the dryas monkey diverged from the common ancestor of all vervets around 1 million
270 years ago, long before the vervet radiation started ca. 590 thousand years ago (Figure 1B). As we used
271 an ingroup reference for read mapping (ChlSab1.1), theses estimates could be slightly biased due to
272 decreased mapping efficiency of reads showing the non-reference allele, thereby decreasing the
273 observed genetic distance between dryas monkey and the vervets (Günther & Nettelblad 2018). Thus
274 the ~1 million years divergence time likely represent the lower boundary of the true divergence time.
275 The Y-chromosome-based phylogeny shows the same topology and divergence time estimates as the
276 autosomal trees (Figure 1C) and the Principal Components Analysis further supports that the dryas
277 monkey is genetically distinct from all currently recognised vervets (Figure S3). This stands in stark
278 contrast to the inferences based on the mitochondrial genomes, which show the dryas monkey (both
279 our dryas monkey sample and the dryas monkey type specimen) to be nested within the vervet genus
280 *Chlorocebus* (Guschanski et al. 2013) (Figure 1D).
281 The discrepancies in tree topologies derived from genomic regions with different inheritance modes
282 (autosomal, Y-chromosomal, and mitochondrial) and the known history of introgression among the
283 vervets (Svardal et al. 2017) suggest a possible role of gene flow in shaping the evolutionary history of
284 the dryas monkey. Therefore, we explored whether ancient admixture events can resolve the observed
285 phylogenetic discordance. We found that *sabaeus* individuals from Gambia share significantly fewer
286 derived alleles with the dryas monkey (D-statistic) than all other vervets (Figure 2A). As derived alleles
287 should be approximately equally frequent in all species under the scenario of incomplete lineage

288 sorting without additional gene flow (Green et al. 2010), such a pattern strongly suggest that alleles
289 were exchanged between the dryas monkey and the vervets after their separation from *Chlorocebus*
290 *sabaeus* (~590kya, Figure 1B). *Aethiops* individuals share fewer derived alleles with the dryas monkey
291 than *tantalus, hilgerti, cynosuros*, and *pygerythrus* (Figure 2A), suggesting that gene flow likely
292 occurred for an extended period of time, at least until after the separation of *aethiops* from the
293 common ancestor of the other vervets (~490kya, Figure 1B). As *tantalus, hilgerti, cynosuros* and
294 *pygerythrus* vervets share a similar amount of derived alleles with the dryas monkey, gene flow most
295 likely ended before the speciation of this group (~360kya, Figure 1B). The small observed differences
296 in the D-statistic between *tantalus, hilgerti, cynosuros* and *pygerythrus* (±14%, Figure 2A) are likely the
297 result of drift and selection due to population sizes differences among these vervet species (Svardal et
298 al. 2017). The inferred history of gene flow is also concordant with the mitochondrial phylogeny and
299 suggests that the dryas monkey mitochondrial genome was introgressed from the common ancestor
300 of all non-*sabaeus* vervets (Figure 1B and 1C). Gene flow between the common ancestor of the non-
301 *sabaeus* vervets and the dryas monkey is also supported by TreeMix and ADMIXTURE analyses, but we
302 note that these model-based estimates rely on accurate allele frequency estimates, which are absent
303 for the dryas monkey population, as it is represented by a single individual (Figure S4).
304 In contrast to the *sabaeus* individuals from Gambia, *sabaeus* vervets from Ghana also carry an excess
305 of shared derived alleles with the dryas monkey (Figure 2A). The Ghanese *sabaeus* population recently
306 hybridised with *tantalus*, so that a large proportion of their genome (~15%) is of recent *tantalus*
307 ancestry (Figure 1B) (Svardal et al. 2017). As *tantalus* individuals carry many shared derived alleles with
308 the dryas monkey, this secondary introgression event likely led to the introduction of dryas monkey
309 alleles into the Ghanese *sabaeus,* explaining the high D-statistics in this population.
310 Next, we obtained approximations of the directionality of gene flow using frequency-stratified D-
311 statistics as in (de Manuel et al. 2016). We found that the vervet populations carry derived alleles
312 shared with the dryas monkey at either low or high frequency, but few such alleles are found at
313 intermediate frequencies (Figure 2B). High frequency alleles in the donor population are more likely
314 to be introgressed during gene flow and are subsequently present at low frequency in the recipient
315 population (Kuhlwilm et al. 2016). Therefore, our observation is consistent with bi-directional gene
316 flow between the dryas monkey and the non-*sabaeus* vervets. The general direction of the gene flow
317 appears to have been dominated by the introgression from the dryas monkey into the non-*sabaeus*
318 vervets, as we observe an overall higher proportion of low-frequency shared derived alleles in the
319 vervets (Figure 2B). This difference is particularly pronounced in *aethiops*, suggesting that the gene
320 flow was primarily from the dryas monkey into the vervets before *aethiops* separated from the
321 common ancestor of *tantalus, hilgerti, cynosuros* and *pygerythrus*. After this split, gene flow likely
322 became more bi-directional with increased introgression events into the dryas monkey, as evidenced

323 by the presence of high frequency putatively introgressed alleles in *tantalus, hilgerti, cynosuros* and
324 *pygerythrus*. An alternative explanation is that the observed allele frequencies are driven by selection,
325 as introgressed alleles might be on average selected against (Juric et al. 2016). It is also noteworthy
326 that the dryas monkey caries a vervet mitochondrial genome, which must have been introduced into
327 the population through female-mediated gene flow and eventually became fixed, replacing the
328 ancestral dryas monkey mitochondrial sequence. This is supported by the clustering of the type
329 specimen and study dryas monkey individual mitochondrial genomes (Figure 1D) and their placement
330 as sister to all non-*sabaeus* vervets.
331 The putatively introgressed alleles in the *sabaeus* population from Ghana are found at intermediate
332 frequency (> 0.25 and < 0.50) (Figure 2B), which is in agreement with the indirect introduction of these
333 alleles through recent introgression from *tantalus* vervets. As the *tantalus* population carries derived
334 alleles at high and low frequency (Figure 2B), the Ghanese *sabaeus* population received a mixture of
335 both high and low frequency alleles, resulting in the observed intermediate frequency.
336 Using approaches that are relatively insensitive to demographic processes (e.g. genetic drift and
337 changes in effective population size), we obtained strong support for the presence of gene flow
338 between the dryas monkey and the vervets, but we caution that they may incorrectly infer gene flow
339 in situations with ancestral subdivision (Slatkin & Pollack 2008). However, such ancestral population
340 structure would have to persist over an extended period of time, encompassing multiple speciation
341 events. Furthermore, our inferences of gene flow are supported by the discordance between the
342 nuclear and mitochondrial phylogenies. Thus, gene flow seems to be the most parsimonious
343 explanation.
344
345 **3.2| Identifying introgressed regions and inferring their functional significance**
346 Using D_{xy} and F_d statistic (Martin et al. 2015) we identified putatively introgressed regions in sliding
347 windows of 10.000 bp for each individual. As expected under the scenario of secondary gene flow,
348 windows containing an excess of shared derived alleles with the dryas monkey have low genetic
349 divergence (D_{xy}) to the dryas monkey and high genetic divergence to *sabaeus* (Figure S5) (de Manuel
350 et al. 2016). Summing over all putatively introgressed windows, we roughly estimate that 0.4% - 0.9%
351 in Gambia *sabaeus*, 1.6 – 2.4% in *aethiops* and 2.7% - 4.8% of the genome in the other vervets shows
352 as signature of introgression with the dryas monkey. We estimate that putatively introgressed
353 haplotypes average below 10,000bp (note that we could not detect any haplotypes below the length
354 of 10,000bp with our used method) (Figure S6). The similar length of putatively introgressed
355 haplotypes in all vervet species strongly supports that gene flow occurred in the common ancestor of
356 all non-*sabaeus* species. The putatively introgressed haplotypes into the Ghanese and (to a lesser
357 extent) Gambian *sabaeus* population were later introduced during secondary gene flow with *tantalus*,

358 as an independent recent gene flow event from the dryas monkey into these sabaeus individuals would
359 have resulted in significantly longer haplotypes. However, we caution that our ability to accurately
360 identify the haplotype lengths is low given the short length of the introgressed haplotypes and a single
361 available dryas monkey genome.
362 The average length of putatively introgressed haplotypes reflects the approximate timing of admixture,
363 as haplotypes are broken apart over time due to recombination (Liang & Nielsen 2014). The average
364 length of introgressed haplotypes in the dryas monkey and the vervets is considerably shorter than
365 putatively introgressed Neanderthal haplotypes in modern *Homo sapiens* (~120kb) (Prüfer et al. 2014),
366 which hybridized 47,000 to 65,000 years ago (Sankararaman et al. 2012). This supports our inferred
367 ancient timing of gene-flow (590,000 – 360,000 years) between the dryas monkey and the vervets.
368 Interestingly, the proposed gene-flow between bonobos (*Pan paniscus*) and non-western chimpanzees
369 (*Pan troglodytes*), which have overlapping distribution range with the dryas monkey and the vervets
370 respectively, occurred around the same time period, 200,000 – 550,000 years ago. The average length
371 of the introgressed haplotypes in chimpanzees is longer (~25.000bp), likely due to the longer
372 generation time of chimpanzees compared to vervets (24.6 versus 8.5 years, (Langergraber et al. 2012;
373 Warren et al. 2015; de Manuel et al. 2016), leading to fewer recombination events in chimpanzees
374 since the admixture event.
375 It is noteworthy that both bonobo-chimpanzee and dryas monkey-vervet gene flow most likely
376 involved the crossing of the Congo River (Figure 1), previously thought to be an impenetrable barrier
377 for mammals (Colyn 1987; M. Colyn et al. 1991; Colyn & Deleporte 2004; Eriksson et al. 2004; Kennis
378 et al. 2011). As the timing of these two introgression events are highly congruent, the fluvial topology
379 of the Congo River and the geology within the Congo basin might have been more dynamic 200,000-
380 500,000 years ago than previously recognised (Beadle 1981; Stankiewicz & de Wit 2006).
381 Having identified introgressed regions, we explored if they may carry functional significance. We find
382 that genes previously identified to be under strong selection in vervets (top 10% of XP-CLR selection
383 scores (Svardal et al. 2017)) are less often introgressed (average introgression frequency 0.021) than
384 genes that did not experience strong selection (bottom 10% of XP-CLR scores; average allele frequency
385 0.029). This may be explained by weak selection against introgressed gene on average, which may be
386 deleterious in the non-host background, a pattern also observed for Neanderthal genes in *Homo*
387 *sapiens* and bonobo genes in the chimpanzee genetic background (Nye et al. 2018; Juric et al. 2016).
388 To identify genes with adaptive functions, we focused on 109 putatively introgressed genes that are
389 fixed in all non-*sabaeus* vervets. Gene ontology analysis revealed enrichment for genes related to cell
390 junction assembly and cell projection organization (Figure S7). Vervets are the natural host of the
391 simian immunodeficiency virus (SIV) and the genes under strongest selection in the vervets are related
392 to immunity against this virus (Svardal et al. 2017). Accordingly, POU2F1, AEBP2, and PDCD6IP are

393 among the fixed putatively introgressed genes in all non-*sabaeus* individuals. POU2F1 is a member of
394 the pathway involved in the formation of the HIV-1 elongation complex in humans (Sturm et al. 1993),
395 AEBP2 is a RNA polymerase II repressor (Kim et al. 2009) known to interact with viral transcription
396 (Zhou & Rana 2002; Debaisieux et al. 2012), and PDCD6IP is involved in virus budding of the human
397 immunodeficiency and other lentiviruses (Strack et al. 2003; von Schwedler et al. 2003). However, the
398 SIV resistance related genes that experienced the strongest selection in vervets (e.g. RANBP3, NFIX,
399 CD68, FXR2 and KDM6B) do not show a signal of introgression between vervets and the dryas monkey.
400 Therefore, while adaptive importance can be plausible for some of the introgressed loci, it does not
401 appear to be a strong driver for retaining particular gene classes, although adaptive function for some
402 introgressed genes in species-specific background cannot be excluded and would warrant dedicated
403 investigations.
404
405 **3.3| Genomic view on conservation of the endangered dryas monkey**
406 The dryas monkey is considered the only representative of the dryas species group (Grubb et al. 2003)
407 and is listed as endangered in the IUCN red list due to its small population size of ca. 250 individuals
408 (Hart et al. 2019). We therefore used demographic modelling and genome-wide measures of
409 heterozygosity and inbreeding to assess the long- and short-term population history of the dryas
410 monkey. Pairwise Sequential Markovian Coalescent (PSMC) analysis of the dryas monkey genome
411 revealed a dynamic evolutionary history, with a marked increase in effective population size starting
412 ca. 500,000 years ago, followed by continuous decline in the last ~200,000 years (Figure 3A). The date
413 of population size increase coincides with our estimated onset gene flow. To eliminate the possibility
414 that our PSMC inferences are driven by the increased heterozygosity due to gene flow, we removed
415 all putatively introgressed regions and repeated the PSMC analysis, which produced the same results
416 (Figure 3A). We therefore suggest that the population size increase in the dryas monkey and the
417 associated likely range expansion facilitated secondary contact between the dryas monkey and the
418 vervets.
419 As previously reported, low genomic coverage shifts the PSMC trajectory and makes inference less
420 reliable, particularly for more recent time periods (Figure 3A, (Nadachowska-Brzyska et al. 2016).
421 Therefore, to allow for demographic comparisons to the vervets, we re-ran the PSMC on the dryas
422 monkey genome down-sampled to similar coverage as the genomic data available for the vervets. This
423 analysis strongly suggest that 100,000 – 300,000 years ago the dryas monkey population was the
424 largest among all vervets, possibly ranging in the tens of thousands of individuals (Figure 3B).
425 The genetic diversity of the dryas monkey (measured as between chromosome-pair differences), a
426 proxy for the adaptive potential of a population (Lande & Shannon 1996), is high compared to that of
427 the much more abundant vervets (Figure 4A). The dryas monkey individual also shows no signs of

excessive recent inbreeding, which would manifest itself in a high fraction of the genome contained in long tracts of homozygosity (> 2.5Mb) (Figure 4B). To estimate genetic load, we identified all genes in the dryas monkey genome containing one or more loss-of-function (LoF) mutations and identified all missense mutations in genes other than those already containing LoF-mutation(s) (as such mutations likely behave neutral). We find multiple genes in the dryas monkey containing a homozygous loss-of-function mutation associated with a disease phenotype in humans (n = 27), including SEPT12, associated with reduced sperm mobility (Kuo et al. 2015) and SLAMF9, associated with reduced immunity to tapeworm infections (Cárdenas et al. 2014). However, genome-wide measures of genetic load, measured as the ratio between LoF or missense and synonymous mutations, does not show an increased genomic burden of deleterious mutations in the dryas monkey compared to the much more abundant and widely distributed vervets (Figure 4C-D). The demographic history and the genome-wide measures of genetic diversity and genetic load of the dryas monkey thus suggest that the population of this endangered primate might be larger than currently recognised and that the dryas monkey population has good chances for long-term survival, if appropriate conservation measures are implemented.

Acknowledgments

We thank the DRC government and ICCN, for facilitating sample collection. Sequencing was performed by the SNP&SEQ Technology Platform in Uppsala. The facility is part of the National Genomics Infrastructure (NGI) Sweden and Science for Life Laboratory. The SNP&SEQ Platform is also supported by the Swedish Research Council and the Knut and Alice Wallenberg Foundation. The authors acknowledge support from the Uppsala Multidisciplinary Centre for Advanced Computational Science for assistance with massively parallel sequencing and access to the UPPMAX computational infrastructure. This work was supported by FORMAS (2016-00835) to K.G. Sequence data generated in this study is available in the European nucleotide archive under accession number PRJEB32105.

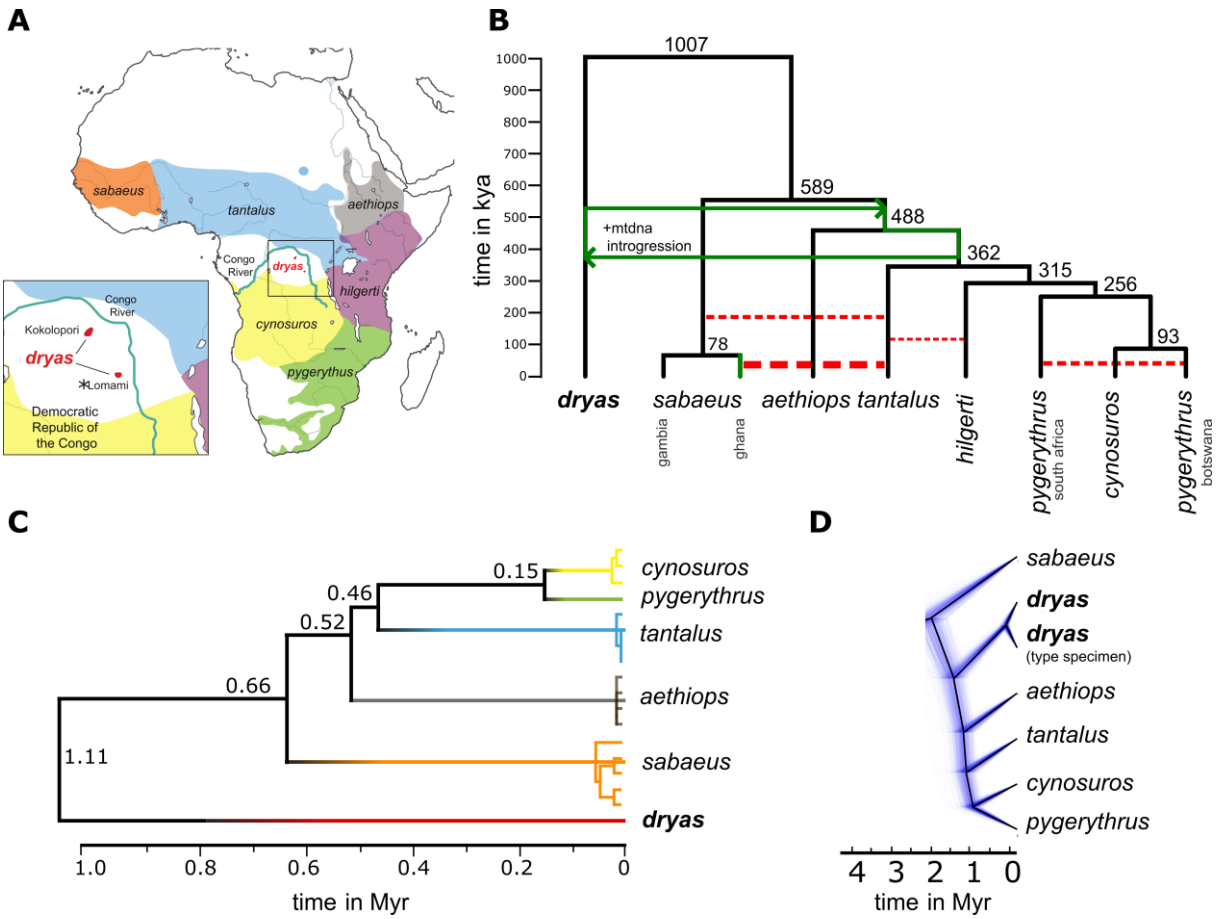


Figure 1. Phylogeny of the dryas monkey and the vervets. (A) Distribution ranges of vervets. *Cercopithecus dryas* is currently known from only two isolated populations: The Kokolopori-Wamba and the Lomami National Park. The sequenced dryas monkey individual was sampled from the Lomami population as indicated by the asterisk. **(B)** The autosomal consensus phylogeny of the dryas monkey and the vervets. The tree topology is supported by multiple analyses (see Methods, Figure S1). Red dotted lines depict previously reported admixture events between the different vervet species (Svardal et al. 2017), with the line width corresponding to the relative strength of admixture. Green arrows show the best-supported timing of admixture between the dryas monkey and the vervets. S*abaeus* individuals from Ghana have a higher proportion of shared derived alleles with the dryas monkey than s*abaeus* individuals from Gambia, which is the result of a strong secondary admixture pulse between *tantalus* and the Ghana *sabaeus* population (Svardal et al. 2017). (**C**) Maximum likelihood Y-chromosome phylogeny, calibrated using the rhesus macaque Y-chromosome as outgroup. (**D**) Mitochondrial phylogeny based on *de-novo* assembled and the previously published dryas monkey type-specimen mitochondrial genomes, rooted with the rhesus macaque (not shown). Solid line depicts the consensus tree, whereas thin purple lines show the posterior distribution of species-trees recovered in BEAST.

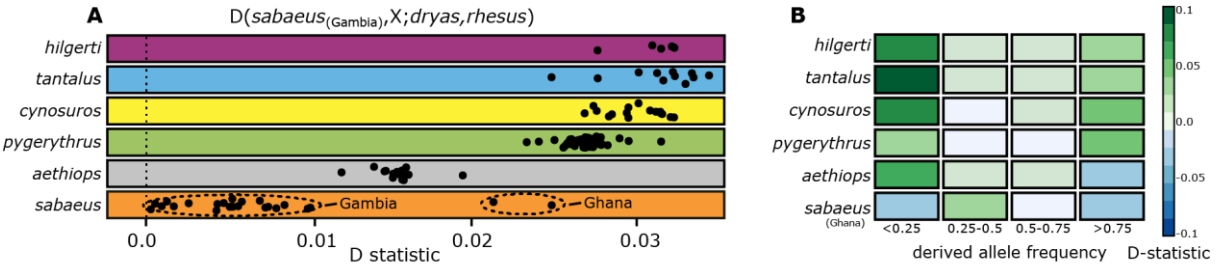


Figure 2. Genome-wide statistics support gene flow between the dryas monkey and the vervets (**A**) Pairwise D-statistics for all individuals using *Chlorocebus sabaeus*$_{\text{Gambia}}$ with the least amount of shared derived alleles to the dryas monkey as ingroup. (**B**) D-statistics stratified by derived allele frequency for each species, using all *Chlorocebus sabaeus*$_{\text{Gambia}}$ individuals as ingroup.

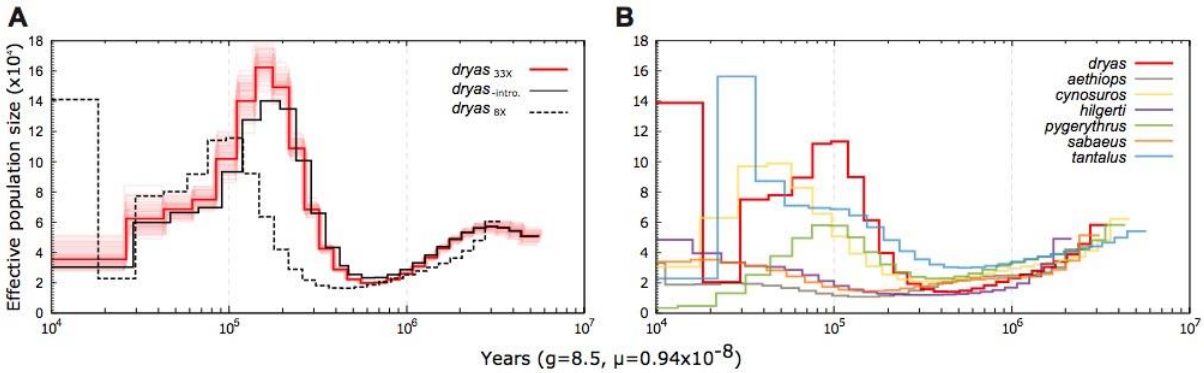


Figure 3. Demographic history of the dryas monkey and the vervets (**A**) PSMC analyses for the repeat-filtered dryas monkey genome (red), after removing putatively introgressed regions (black), and after downsampling to 8x coverage (dotted line). As the curve is strongly shifted at lower coverage, the downsampled genome was used for between-species comparison in (B). (B) PSMC analysis of medium coverage genomes (7.4-9.8X) for all vervets and the downsampled dryas monkey genome (8X).

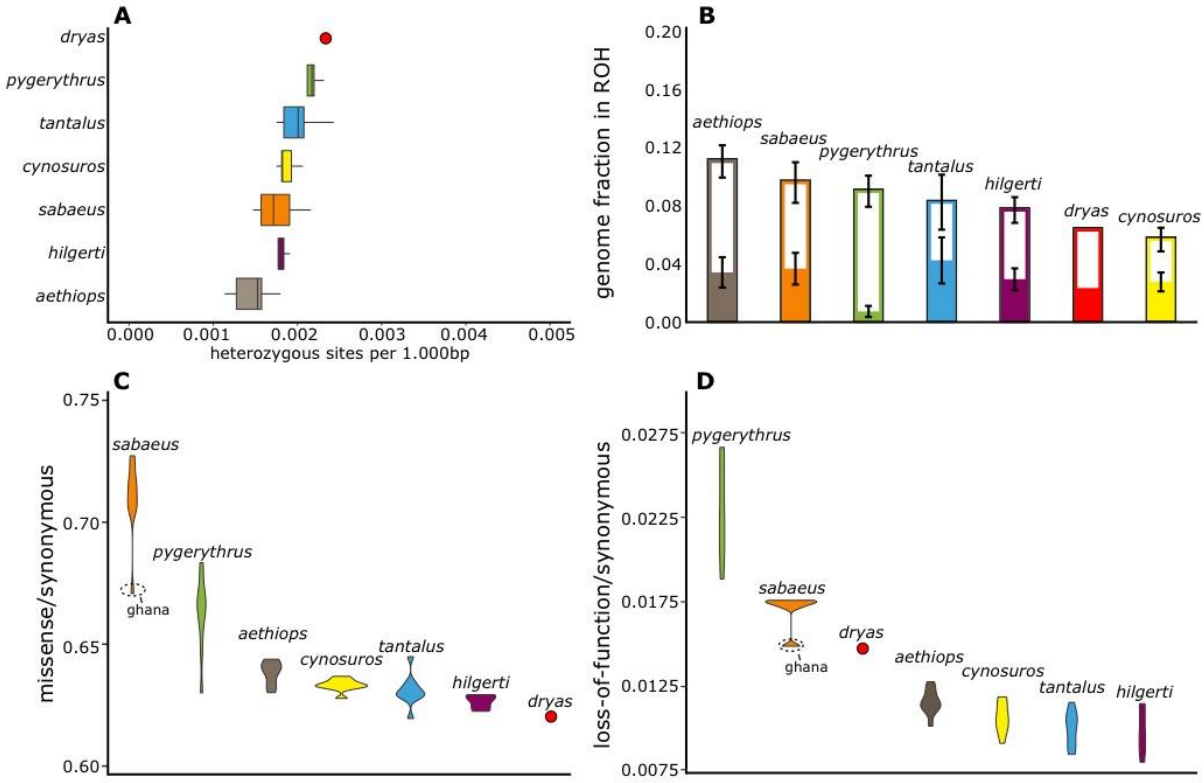


Figure 4. Genome-wide measures of genetic diversity and load in the dryas monkey (**A**) Average number of heterozygous sites per 1000 base pairs. (**B**) Fraction of the genome in runs of homozygosity (ROH) above 100kb (open bars) and fraction of the genome in ROH > 2.5MB (solid bars). (**C**) Ratio of missense to synonymous mutations, excluding all missense mutations within genes containing one or more LOF mutations. (**D**) Ratio of LOF to synonymous mutations, counting genes with more than one LOF mutation only once.

References

Alexander, D.H., Novembre, J. & Lange, K., 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome research*, 19(9), pp.1655–64.

Van der Auwera, G.A. et al., 2013. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. In *Current Protocols in Bioinformatics*. Hoboken, NJ, USA: John Wiley & Sons, Inc., p. 11.10.1-11.10.33.

Beadle, L., 1981. *The inland waters of tropical Africa: an introduction to tropical limnology* 2nd editio., London/New York.

Bolger, A.M., Lohse, M. & Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)*, 30(15), pp.2114–20.

Bouckaert, R. et al., 2014. BEAST 2: A Software Platform for Bayesian Evolutionary Analysis A. Prlic, ed. *PLoS Computational Biology*, 10(4), p.e1003537.

Bouckaert, R.R., 2010. DensiTree: Making sense of sets of phylogenetic trees. *Bioinformatics*, 26(10), pp.1372–1373.

Cárdenas, G. et al., 2014. Neurocysticercosis: the effectiveness of the cysticidal treatment could be influenced by the host immunity. *Medical Microbiology and Immunology*, 203(6), pp.373–381.

Chen, H., Patterson, N. & Reich, D., 2010. Population differentiation as a test for selective sweeps. *Genome Research*, 20(3), pp.393–402.

Colyn, Gautier-Hion, A. & van den Audenaerde, T., 1991. Cercopithecus dryas Schwarz 1932 and C. salongo Thys van den Audenaerde 1977 are the same species with an age-related coat pattern. *Folia Primatologica*, 56(3), pp.167–170.

Colyn, M., 1987. Les primates des forêts ombrophiles de la cuvette du Zaïre: interprétations zoogéographiques des modèles de distribution. *Revue De Zoologie Africaine*, 101, pp.183–196.

Colyn, M. & Deleporte, P., 2004. Biogeographic Analysis of Central African Forest Guenons. In *The Guenons: Diversity and Adaptation in African Monkeys*. Boston: Kluwer Academic Publishers, pp. 61–78.

Colyn, M., Gautier-Hion, A. & Verheyen, W., 1991. A Re-Appraisal of Palaeoenvironmental History in Central Africa: Evidence for a Major Fluvial Refuge in the Zaire Basin. *Journal of Biogeography*, 18(4), p.403.

518 Danecek, P. et al., 2011. The variant call format and VCFtools. *Bioinformatics*, 27(15), pp.2156–2158.

519 Debaisieux, S. et al., 2012. The Ins and Outs of HIV-1 Tat. *Traffic*, 13(3), pp.355–363.

520 DePristo, M.A. et al., 2011. A framework for variation discovery and genotyping using next-
521 generation DNA sequencing data. *Nature Genetics*, 43(5), pp.491–498.

522 Dierckxsens, N., Mardulyn, P. & Smits, G., 2016. NOVOPlasty: De novo assembly of organelle
523 genomes from whole genome data. *Nucleic Acids Research*, 45(4), p.gkw955.

524 Eriksson, J. et al., 2004. Rivers influence the population genetic structure of bonobos (Pan paniscus).
525 *Molecular Ecology*, 13(11), pp.3425–3435.

526 Fay, J.C., Wyckoff, G.J. & Wu, C.I., 2001. Positive and negative selection on the human genome.
527 *Genetics*, 158(3), pp.1227–34.

528 Felsenstein, J., 2005. PHYLIP (Phylogeny Inference Package).

529 Fumagalli, M., 2013. Assessing the effect of sequencing depth and sample size in population genetics
530 inferences L. Orlando, ed. *PLoS ONE*, 8(11), p.e79667.

531 Fumagalli, M. et al., 2013. Quantifying population genetic differentiation from next-generation
532 sequencing data. *Genetics*, 195(3), pp.979–992.

533 Gotz, S. et al., 2008. High-throughput functional annotation and data mining with the Blast2GO suite.
534 *Nucleic Acids Research*, 36(10), pp.3420–3435.

535 Green, R.E. et al., 2010. A draft sequence of the Neandertal genome. *Science (New York, N.Y.)*,
536 328(5979), pp.710–722.

537 Grubb, P. et al., 2003. Assessment of the Diversity of African Primates. *International Journal of*
538 *Primatology*, 24(6), pp.1301–1357.

539 Günther, T. & Nettelblad, C., 2018. The presence and impact of reference bias on population genomic
540 studies of prehistoric human populations.

541 Guschanski, K. et al., 2013. Next-generation museomics disentangles one of the largest primate
542 radiations. *Systematic Biology*, 62(4), pp.539–554.

543 Hart, J.A. et al., 2019. Cercopithecus dryas. The IUCN Red List of Threatened Species.

544 Hawks, J., 2017. Introgression Makes Waves in Inferred Histories of Effective Population Size. *Human*
545 *Biology*, 89(1), pp.67–80.

546 Hedges, S.B. et al., 2015. Tree of life reveals clock-like speciation and diversification. *Molecular*
547 *Biology and Evolution*, 32(4), pp.835–845.

548 Hubisz, M.J., Pollard, K.S. & Siepel, A., 2011. PHAST and RPHAST: phylogenetic analysis with
549 space/time models. *Briefings in Bioinformatics*, 12(1), pp.41–51.

550 Huson, D.H. & Bryant, D., 2006. Application of Phylogenetic Networks in Evolutionary Studies.
551 *Molecular Biology and Evolution*, 23(2), pp.254–267.

552 IUCN, 2019. The IUCN Red List of Threatened Species. Version 2019.1. , (March).

553 Jonathan Kingdon, David Happold, Thomas Butynski, M.H., Happold, M. & Kalina, J., 2013. *Mammals*
554 *of Africa* 1st editio.,

555 Juric, I., Aeschbacher, S. & Coop, G., 2016. The Strength of Selection against Neanderthal
556 Introgression D. Reich, ed. *PLOS Genetics*, 12(11), p.e1006340.

557 Kennis, J. et al., 2011. The impact of the Congo River and its tributaries on the rodent genus Praomys:
558 Speciation origin or range expansion limit? *Zoological Journal of the Linnean Society*, 163(3),
559 pp.983–1002.

560 Kim, H., Kang, K. & Kim, J., 2009. AEBP2 as a potential targeting protein for Polycomb Repression
561 Complex PRC2. *Nucleic acids research*, 37(9), pp.2940–50.

562 Korneliussen, T.S., Albrechtsen, A. & Nielsen, R., 2014. ANGSD: Analysis of Next Generation
563 Sequencing Data. *BMC Bioinformatics*, 15(1), p.356.

564 Kuderna, L.F.K. et al., 2019. Selective single molecule sequencing and assembly of a human Y
565 chromosome of African origin. *Nature Communications*, 10(1), p.4.

566 Kuhlwilm, M. et al., 2016. Ancient gene flow from early modern humans into Eastern Neanderthals.
567 *Nature*, 530(7591), pp.429–433.

568 Kumar, S. et al., 2018. MEGA X: Molecular Evolutionary Genetics Analysis across Computing
569 Platforms F. U. Battistuzzi, ed. *Molecular Biology and Evolution*, 35(6), pp.1547–1549.

570 Kuo, Y.-C. et al., 2015. SEPT12 orchestrates the formation of mammalian sperm annulus by organizing
571 core octameric complexes with other SEPT proteins. *Journal of Cell Science*, 128(5), pp.923–934.

572 Kuroda, S., Kano, T. & Muhindo, K., 1985. Further information on the new monkey
573 species,Cercopithecus salongo Thys van den Audenaerde, 1977. *Primates*, 26(3), pp.325–333.

574 Lande, R. & Shannon, S., 1996. The Role of Genetic Variation in Adaptation and Population

575 Persistence in a Changing Environment. *Evolution*, 50(1), p.434.

576 Lanfear, R. et al., 2017. Partitionfinder 2: New methods for selecting partitioned models of evolution
577 for molecular and morphological phylogenetic analyses. *Molecular Biology and Evolution*, 34(3),
578 pp.772–773.

579 Langergraber, K.E. et al., 2012. Generation times in wild chimpanzees and gorillas suggest earlier
580 divergence times in great ape and human evolution. *Proceedings of the National Academy of*
581 *Sciences of the United States of America*, 109(39), pp.15716–21.

582 Li, H., 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.

583 Li, H. et al., 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16),
584 pp.2078–2079.

585 Li, H. & Durbin, R., 2011. Inference of human population history from individual whole-genome
586 sequences. *Nature*, 475(7357), pp.493–496.

587 Liang, M. & Nielsen, R., 2014. The Lengths of Admixture Tracts. *Genetics*, 197(3), pp.953–967.

588 de Manuel, M. et al., 2016. Chimpanzee genomic diversity reveals ancient admixture with bonobos.
589 *Science*, 354(6311).

590 Martin, S.H., Davey, J.W. & Jiggins, C.D., 2015. Evaluating the use of ABBA-BABA statistics to locate
591 introgressed loci. *Molecular Biology and Evolution*, 32(1), pp.244–257.

592 McKenna, A. et al., 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-
593 generation DNA sequencing data. *Genome research*, 20(9), pp.1297–303.

594 McLaren, W. et al., 2016. The Ensembl Variant Effect Predictor. *Genome Biology*, 17(1), p.122.

595 Nadachowska-Brzyska, K. et al., 2016. PSMC analysis of effective population sizes in molecular
596 ecology and its application to black-and-white Ficedula flycatchers. *Molecular ecology*, 25(5),
597 pp.1058–72.

598 Nye, J. et al., 2018. Selection in the Introgressed Regions of the Chimpanzee Genome. *Genome*
599 *biology and evolution*, 10(4), pp.1132–1138.

600 Paradis, E., Claude, J. & Strimmer, K., 2004. APE: Analyses of Phylogenetics and Evolution in R
601 language. *Bioinformatics*, 20(2), pp.289–290.

602 Pfeifer, S.P., 2017. The Demographic and Adaptive History of the African Green Monkey. *Molecular*
603 *Biology and Evolution*, 34(5), pp.1055–1065.

604 Pickrell, J.K. & Pritchard, J.K., 2012. Inference of Population Splits and Mixtures from Genome-Wide
605 Allele Frequency Data H. Tang, ed. *PLoS Genetics*, 8(11), p.e1002967.

606 Price, M.N., Dehal, P.S. & Arkin, A.P., 2010. FastTree 2 – Approximately Maximum-Likelihood Trees
607 for Large Alignments A. F. Y. Poon, ed. *PLoS ONE*, 5(3), p.e9490.

608 Prüfer, K. et al., 2014. The complete genome sequence of a Neanderthal from the Altai Mountains.
609 *Nature*, 505(7481), pp.43–49.

610 Purcell, S. et al., 2007. PLINK: A Tool Set for Whole-Genome Association and Population-Based
611 Linkage Analyses. *The American Journal of Human Genetics*, 81(3), pp.559–575.

612 Rambaut, A. et al., 2014. Tracer v1.6.

613 Sankararaman, S. et al., 2012. The date of interbreeding between Neandertals and modern humans.
614 *PLoS genetics*, 8(10), p.e1002947.

615 Sayyari, E. & Mirarab, S., 2016. Fast Coalescent-Based Computation of Local Branch Support from
616 Quartet Frequencies. *Molecular Biology and Evolution*, 33(7), pp.1654–1668.

617 Schliep, K.P., 2011. phangorn: phylogenetic analysis in R. *Bioinformatics*, 27(4), pp.592–593.

618 Schwarz, E., 1932. Der Vertreter der Diana-Meerkatze in Zentral-Afrika. *Rev Zoologique Botanique*
619 *Africaine*, (21), pp.251–254.

620 von Schwedler, U.K. et al., 2003. The protein network of HIV budding. *Cell*, 114(6), pp.701–13.

621 Sievers, F. et al., 2011. Fast, scalable generation of high-quality protein multiple sequence alignments
622 using Clustal Omega. *Molecular systems biology*, 7(1), p.539.

623 Slatkin, M. & Pollack, J.L., 2008. Subdivision in an ancestral species creates asymmetry in gene trees.
624 *Molecular biology and evolution*, 25(10), pp.2241–6.

625 Stankiewicz, J. & de Wit, M.J., 2006. A proposed drainage evolution model for Central Africa—Did the
626 Congo flow east? *Journal of African Earth Sciences*, 44(1), pp.75–84.

627 Strack, B. et al., 2003. AIP1/ALIX is a binding partner for HIV-1 p6 and EIAV p9 functioning in virus
628 budding. *Cell*, 114(6), pp.689–99.

629 Sturm, R.A. et al., 1993. Chromosomal Structure and Expression of the Human OTF1 Locus Encoding
630 the Oct-1 Protein. *Genomics*, 16(2), pp.333–341.

631 Svardal, H. et al., 2017. Ancient hybridization and strong adaptation to viruses across African vervet
632 monkey populations. *Nature Genetics*, 49(12), pp.1705–1713.

633 Tamura, K. & Nei, M., 1993. Estimation of the number of nucleotide substitutions in the control
634 region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution*,
635 10(3), pp.512–26.

636 Tomaszkiewicz, M., Medvedev, P. & Makova, K.D., 2017. Y and W Chromosome Assemblies:
637 Approaches and Discoveries. *Trends in Genetics*, 33(4), pp.266–282.

638 van der Valk, T. et al., 2019. Historical Genomes Reveal the Genomic Consequences of Recent
639 Population Decline in Eastern Gorillas. *Current Biology*, 29(1), p.165–170.e6.

640 van der Valk, T. et al., 2018. Significant loss of mitochondrial diversity within the last century due to
641 extinction of peripheral populations in eastern gorillas. *Scientific Reports*, 8(1), p.6551.

642 Warren, W.C. et al., 2015. The genome of the vervet (Chlorocebus æthiops sabæus). *Genome*
643 *Research*, 25(12), pp.1921–1933.

644 Xue, C. et al., 2016. The population genomics of rhesus macaques (Macaca mulatta) based on whole-
645 genome sequences. *Genome research*, 26(12), pp.1651–1662.

646 Xue, Y. et al., 2015. Mountain gorilla genomes reveal the impact of long-term population decline and
647 inbreeding. *Science (New York, N.Y.)*, 348(6231), pp.242–5.

648 Zhang, C. et al., 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved
649 gene trees. *BMC Bioinformatics*, 19(S6), p.153.

650 Zhou, C. & Rana, T.M., 2002. A bimolecular mechanism of HIV-1 Tat protein interaction with RNA
651 polymerase II transcription elongation complexes. *Journal of molecular biology*, 320(5), pp.925–
652 42.

653 Zinner, D. et al., 2013. *Family Cercopithecidae (Old World Monkeys). In Handbook of the Mammals of*
654 *the World: Primates (pp. 550–753). Lynx Edicions.*,

655

Supplementary figures

The genome of the endangered dryas monkey provides new insights into the evolutionary history of the vervets

T. van der Valk, C. M. Gonda, J. Hart, T. Hart, K. Detwiler, K. Guschanski

Corresponding authors: Tom van der Valk, tom.vandervalk@ebc.uu.se, Katerina Guschanski katerina.guschanski@ebc.uu.se

Content **Page**

Figure S1	2
Figure S2	3
Figure S3	4
Figure S4	5
Figure S5	6
Figure S6	7
Figure S7	8

Figure S1. Phylogeny of the *dryas* and the vervets. (A) MSC-based species trees generated by ASTRAL using 3703 autosomal genomic windows. The tree was rooted with the rhesus macaque. Branch lengths are given in coalescent units and are an indicator of gene-tree discordance. The normalized quartet score of this topology is 0.87. Asterisk symbols at nodes indicate maximum local quartet support posterior probabilities (lpp=1.0) and *q1* values displayed at each node show the percentage of quartets in all gene trees that agree with that branch topology. **(B)** Majority-rule consensus tree obtained from 3703 autosomal gene-trees using CONSENSE in PHYLIP v3.695. The number above each branch shows the total number of trees out of all 3703 gene trees supporting the given branch and the number below corresponds to its percentage. **(C)** SplitsTree consensus networks using 3703 gene trees at median thresholds of 25%, 20% and 15%.

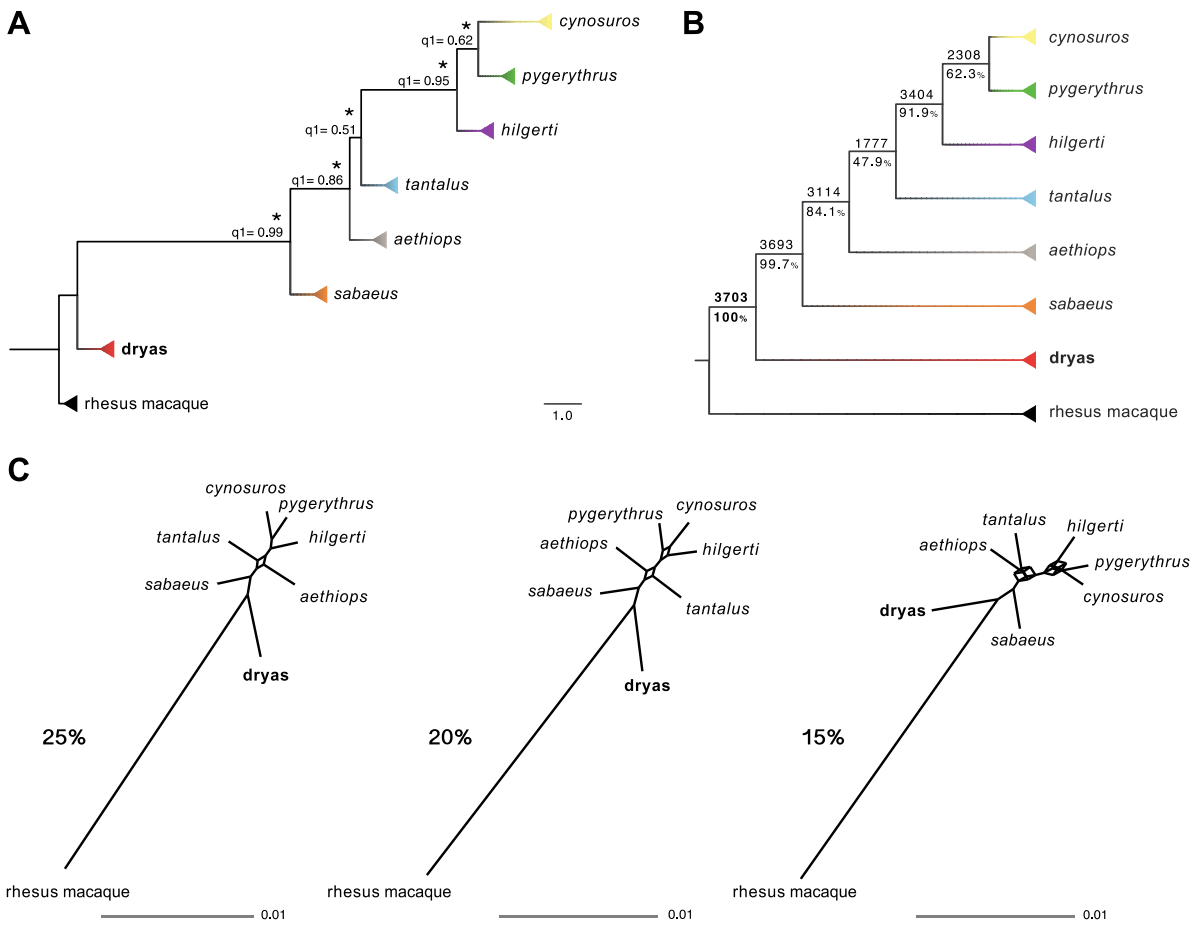


Figure S2. Sexing a subset of individuals based on autosomal versus X chromosome coverage.

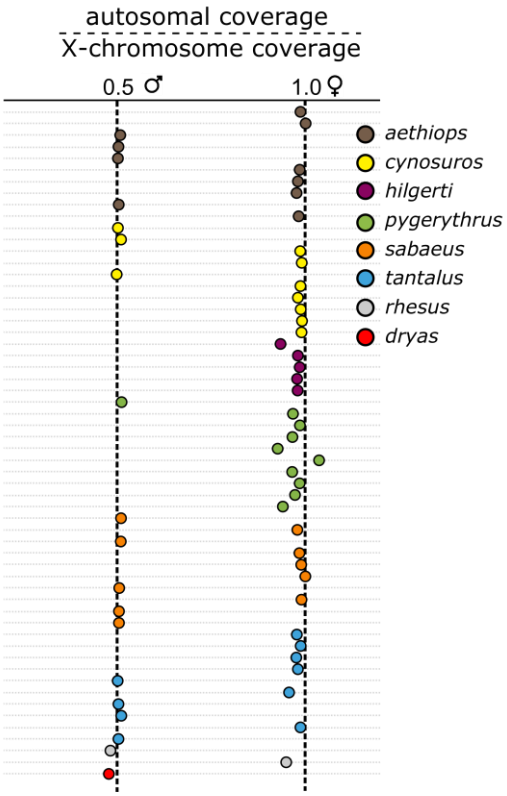


Figure S3. Genome wide Principal Components Analysis on all quality filtered bi-allelic sites.

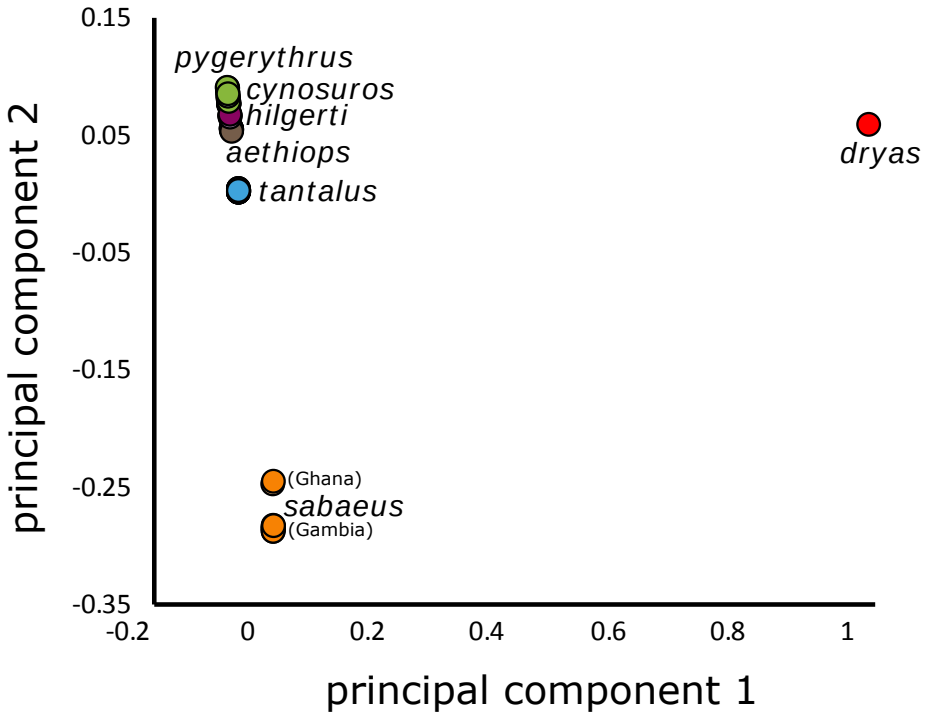


Figure S4. Model-based estimates of gene-flow. (A) Whole genome ADMIXTURE analysis for 9 predefined clusters based on all quality filtered autosomal bi-allelic SNPs. Note the magnified upper part of the bar for the dryas monkey, which shows similar proportions of all non-*sabaeus* vervets in the dryas monkey genome **(B)** TreeMix analysis for five migration events (m=5). Population allele frequencies are separated by species and geographic origin. **(C)** Likelihood support values for TreeMix models with 0 to 10 migration events respectively. After modelling five migration events, the model likelihood does not increase any further.

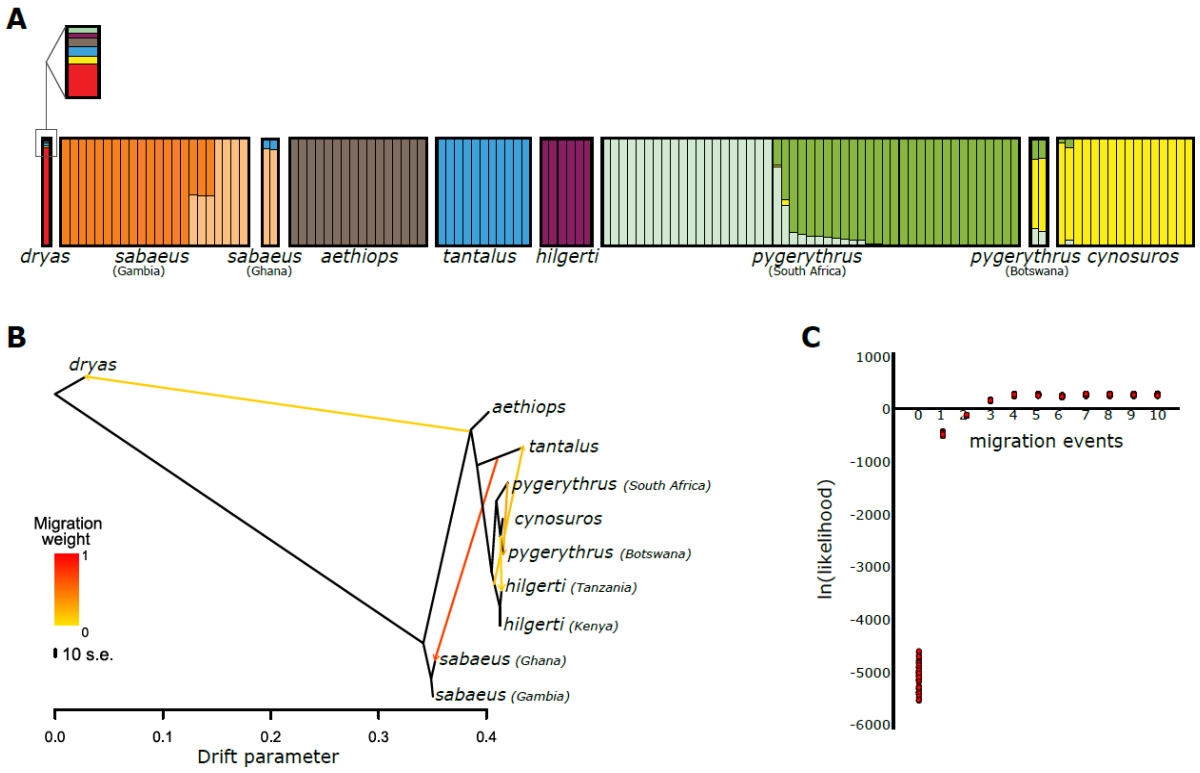


Figure S5. Dxy statistic stratified by Fd. Errors bars show ±3SE. Window size = 10kb. Windows with high F_d statistic (the excess of shared derived alleles with the dryas monkey over that with *sabaeus*) have on average unusual large genetic distance (D_{xy}) to *sabaeus* and unusual low genetic distance to the dryas monkey. Note that the vast majority of windows is at F_d ~ 0 (>90%). $D_{xy\text{-}sabaeus}$ at windows with F_d = 0 (putative non-introgressed windows) is generally around 0.007 whereas $D_{xy\text{-}dryas}$ at these windows ~0.012. This is in agreement with the obtained divergence time estimates (Figure 1B) (e.g. genetic distance of the non-*sabaeus* vervets to the dryas monkey is on average 1.75 - 2.25 times larger than non-*sabeus* vervets to *sabaeus*). At high F_d, $D_{xy\text{-}dryas}$ is around ½-⅓ that of F_d = 0, suggesting that these windows diverged on average half to one third as long ago as the other windows (thus ~500.000 to 330.000 years ago as the most likely time period of introgression, if we assume the divergence of the dryas monkey to the vervets to be ~1 million years ago, as indicated by our analyses).

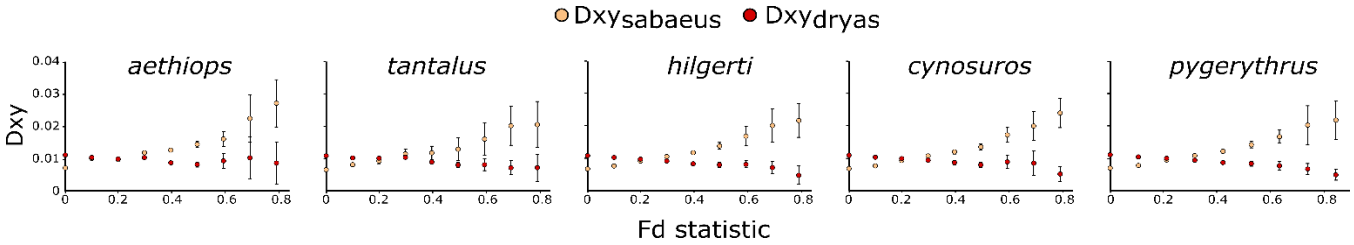


Figure S6. Length distribution of putatively introgressed haplotypes. Windows are short on average for all species. *Sabaeus* individuals from Gambia have the lowest number of putatively introgressed windows, which were likely introduced into this species via secondary gene-flow with *tantalus*. Ghana individuals have a much higher number of these windows, as a result of the recent strong secondary admixture with *tantalus*.

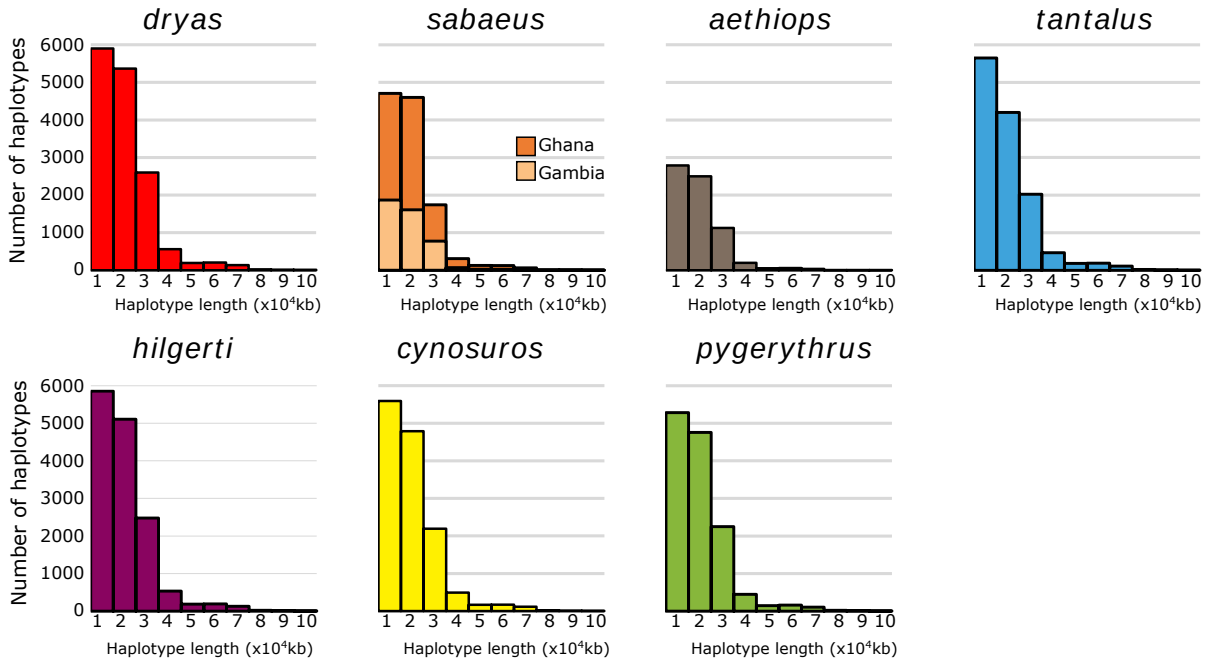
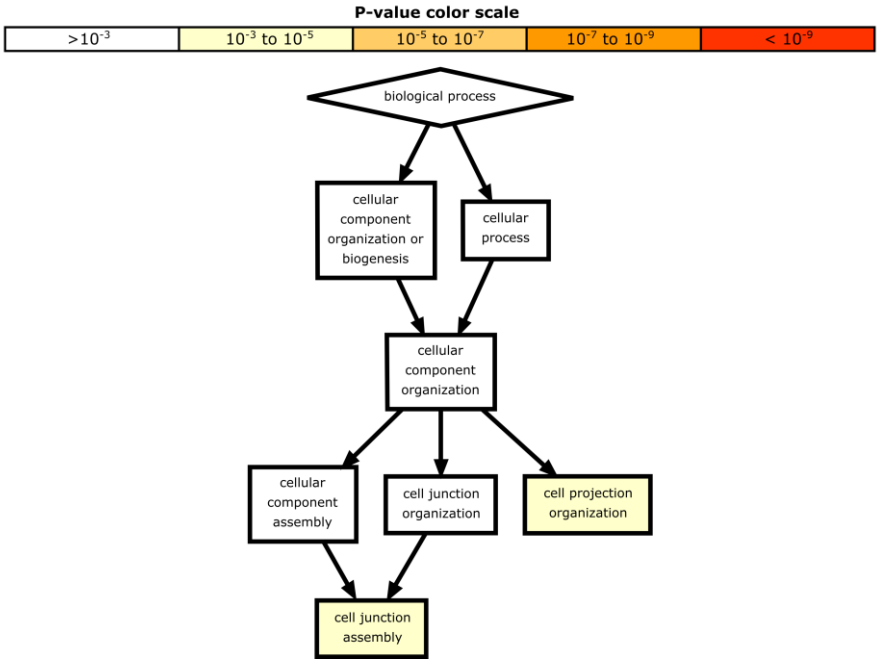


Figure S7. **Gene-ontology enrichment for all putatively introgressed genes fixed in all non-*sabaeus* individuals**. Top panel shows enriched GO-terms and the genes associated with these terms. Bottom panel shows the enrichment network.

GO-term	Description	P-value	FDR q-value	Enrichment (N,B,n,b)	Genes
GO:0034329	cell junction assembly	3.1E-4	1	6.56 (15912,160,91,6)	CDH13 - cadherin 13, h-cadherin (heart) CDH12 - cadherin 12, type 2 (n-cadherin) LAMA3 - laminin, alpha 3 PDCD6IP - programmed cell death 6 interacting protein PTPRK - protein tyrosine phosphate, receptor type, k APC - adenomatous polyposis coli
GO:0030030	cell projection organization	9.32E-4	1	2.86 (15912,733,91,12)	TNFRSF21 - tumor necrosis factor receptor superfamily, member 21 MEF2A - myocyte enhancer factor 2a ABL1 - c-abl oncogene 1, non-receptor tyrosine kinase PTPRZ1 - protein tyrosine phosphatase, receptor-type, z polypeptide 1 WTIP - wilms tumor 1 interacting protein CDH13 - cadherin 13, h-cadherin (heart) PRKG1 - protein kinase, cgmp-dependent, type i NEK1 - nima-related kinase 1 CTNNA2 - catenin (cadherin-associated protein), alpha 2 PTPRK - protein tyrosine phosphatase, receptor type, k PKHD1 - polycystic kidney and hepatic disease 1 (autosomal recessive) GPM6A - glycoprotein m6a