

RESEARCH

Open Access



# Genomic variation in baboons from central Mozambique unveils complex evolutionary relationships with other *Papio* species

Cindy Santander<sup>1,2\*</sup>, Ludovica Molinaro<sup>3</sup>, Giacomo Mutti<sup>4,5</sup>, Felipe I. Martínez<sup>6</sup>, Jacinto Mathe<sup>7</sup>, Maria Joana Ferreira da Silva<sup>8,9,10</sup>, Matteo Caldon<sup>5</sup>, Gonzalo Oteo-García<sup>5</sup>, Vera Aldeias<sup>11</sup>, Will Archer<sup>12</sup>, Marion Bamford<sup>13</sup>, Dora Biro<sup>2</sup>, René Bobe<sup>7,14</sup>, David R. Braun<sup>15</sup>, Philippa Hammond<sup>7</sup>, Tina Lüdecke<sup>7,16</sup>, Maria José Pinto<sup>17</sup>, Luis Meira Paulo<sup>17</sup>, Marc Stalmans<sup>18</sup>, Frederico Tátá Regala<sup>11</sup>, Francesco Bertolini<sup>19</sup>, Ida Moltke<sup>1</sup>, Alessandro Raveane<sup>19</sup>, Luca Pagani<sup>3,20</sup>, Susana Carvalho<sup>7,14\*†</sup> and Cristian Capelli<sup>2,5\*†</sup>

## Abstract

**Background:** Gorongosa National Park in Mozambique hosts a large population of baboons, numbering over 200 troops. Gorongosa baboons have been tentatively identified as part of *Papio ursinus* on the basis of previous limited morphological analysis and a handful of mitochondrial DNA sequences. However, a recent morphological and morphometric analysis of Gorongosa baboons pinpointed the occurrence of several traits intermediate between *P. ursinus* and *P. cynocephalus*, leaving open the possibility of past and/or ongoing gene flow in the baboon population of Gorongosa National Park. In order to investigate the evolutionary history of baboons in Gorongosa, we generated high and low coverage whole genome sequence data of Gorongosa baboons and compared it to available *Papio* genomes.

**Results:** We confirmed that *P. ursinus* is the species closest to Gorongosa baboons. However, the Gorongosa baboon genomes share more derived alleles with *P. cynocephalus* than *P. ursinus* does, but no recent gene flow between *P. ursinus* and *P. cynocephalus* was detected when available *Papio* genomes were analyzed. Our results, based on the analysis of autosomal, mitochondrial and Y chromosome data, suggest complex, possibly male-biased, gene flow between Gorongosa baboons and *P. cynocephalus*, hinting to direct or indirect contributions from baboons belonging to the “northern” *Papio* clade, and signal the presence of population structure within *P. ursinus*.

**Conclusions:** The analysis of genome data generated from baboon samples collected in central Mozambique highlighted a complex set of evolutionary relationships with other baboons. Our results provided new insights in the population dynamics that have shaped baboon diversity.

**Keywords:** Evolutionary genetics, Primate genomics, *Papio*, Population genomics

## Background

Baboons (genus *Papio*) are the primate genus beside humans with the largest distribution across Africa [1]. Over the last two million years they came to occupy a variety of different habitats, from the gallery forests of western Africa all the way to the rocky Cape Peninsula in southern Africa, extending their presence also outside

\*Correspondence: cindyg.santander@gmail.com; susana.carvalho@anthro.ox.ac.uk; cristian.capelli@unipr.it

†Susana Carvalho and Cristian Capelli are co-senior author.

<sup>1</sup> Department of Biology, University of Copenhagen, Copenhagen, Denmark

<sup>2</sup> Department of Zoology, University of Oxford, Oxford, UK

<sup>7</sup> School of Anthropology, University of Oxford, Oxford, UK

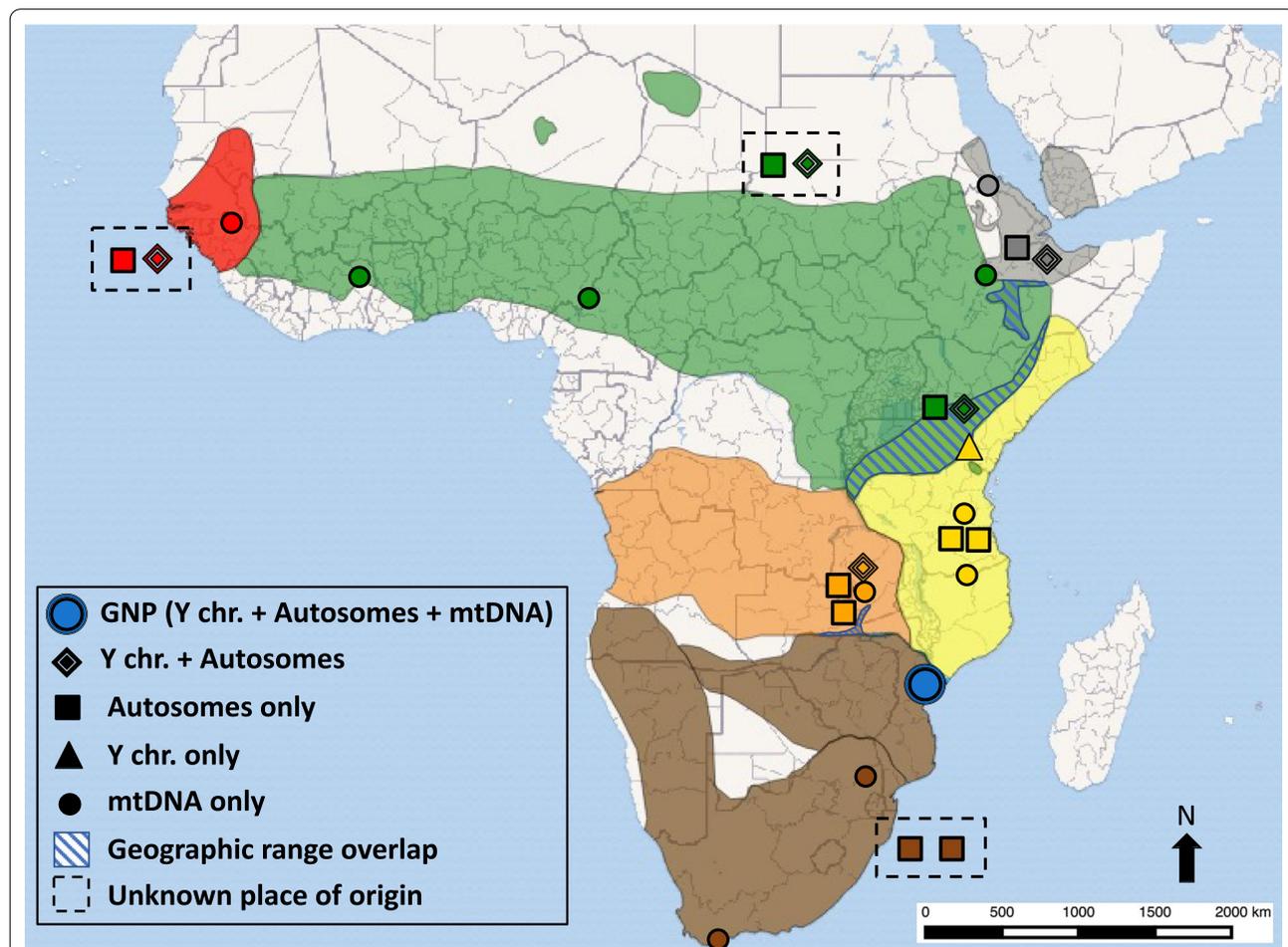
Full list of author information is available at the end of the article



Africa into the southwestern part of the Arabian Peninsula [2, 3]. Members of this genus are characterized by relatively extensive morphological and behavioral variation, such that their taxonomic classification is debated [2–4]. Most of the disagreement stems from what species concept is used, as their parapatric distribution, ecological flexibility and interfertility challenge many of the taxonomic criteria used for species classification [5–9]. A consensus has been recently reached supporting six extant baboon species (morphotypes, sensu Jolly et al. [9]) which are phylogenetically grouped in two geographic clades, one northern including *Papio papio*, *P. anubis* and *P. hamadryas*, and one southern comprising *P. cynocephalus*, *P. ursinus*, and *P. kindae* [4, 7, 10–13] (Fig. 1). *Papio* genetic diversity appears to be further characterized by a complex history of ancient

and more recent gene flow, to include a deep admixture event for the origin of *P. kindae*, the contribution from a now extinct lineage to *P. papio*, and historical contacts between *P. cynocephalus* and *P. anubis* [11].

Admixture among different baboon species has been reported numerous times and the existence of several hybrid zones across the sub-saharan African distribution of the genus has been suggested [10, 13–16]. MtDNA data and microsatellite analysis have suggested contacts between populations of *P. hamadryas*, *P. anubis* and *P. cynocephalus* in Eastern Africa, and between *P. cynocephalus* and *P. ursinus*, as well as and *P. ursinus* and *P. kindae* in Southern Africa [4, 13, 14]. More recently, genome data provided some indication of the spatial extent of *P. anubis* introgression in *P. cynocephalus* populations in a contact zone in Eastern Africa [15] and hybridization



**Fig. 1** Distribution of *Papio* species and location of samples analyzed. Ychr.: Y chromosome sequence data; mtDNA: mitochondrial DNA genome. Shapes refer to the inset legend. Samples whose original place of origin is unknown are enclosed in a dashed box. Color scheme of *Papio* species distribution is as follows: *P. papio* = red, *P. anubis* = green, *P. kindae* = orange, *P. cynocephalus* = yellow, *P. ursinus* = brown, and GNP = blue. (Information on sample provenance collected from Additional file 2: Table S3 in Rogers et al. [11], Table 1 in Zinner et al. [10] and Appendix S1 in Wall et al. [15])

between *P. kindae* and *P. ursinus* in the Kafue River Valley of Zambia [17].

Interestingly, despite female philopatry in several *Papio* species and the observation in the field of male-biased admixture dynamics [3, 4, 15], little molecular work has been conducted on Y chromosome variation in *Papio*. Phylogenetic studies including *Papio* species have surveyed only a few thousands bp of Y chromosome sequence, while population-based studies have screened a handful of Y chromosome specific SNPs and STRs [18–21]. As such, virtually no large Y chromosome datasets have been systematically investigated for population and evolutionary studies. The analysis of Y chromosome sequences hundreds of kilo-bases (kb) in length retrieved from genomic data is therefore expected to provide not only a high resolution male-specific insight of *Papio* phylogeny but also to give access to hundreds of Y chromosome markers useful for population studies addressing sex-biased interactions [4], as shown for great apes including humans [22].

Overall, gene-exchange appears to be common among neighboring groups of baboons; however, the extent and the dynamics of these events are yet to be fully appreciated within *Papio* and the occurrence of gene flow should be tested in a more systematic way across other potential areas of contact between different species [10]. Understanding these processes at the genome level is of particular relevance to reconstruct the demographic and evolutionary dynamics that have shaped the diversity of behavior, social structure and phenotype seen in baboons, and has significant implications for interpreting the evolutionary history of other species, extant and extinct, including hominins [2, 3].

Two species of baboons, *P. cynocephalus* and *P. ursinus*, are present in Mozambique in south-eastern Africa (Fig. 1), whose distribution has been inferred by a combination of molecular and morphological data [4, 12, 23]. The Zambezi River, a major hydrological feature which runs east to west through central Mozambique, has been suggested as the natural barrier dividing the regions of occupation of the two species [10, 24]. Baboons south of the Zambezi have been assigned to the subspecies *P. ursinus griseipes* (i.e. grayfoot chacma baboons) one of the three morphotypes identified within *P. ursinus* [25]. The Zambezi has been suggested as a biogeographical barrier for other species as well, for example Blue wildebeest (*Connochaetes taurinus*) and sable antelope (*Hippotragus niger*), but it is not yet fully understood how permeable to migration this barrier has been over time [26–28].

As part of a project assessing primate adaptations and evolution in the southern African Rift System [29, 30] we recently investigated the morphological variation of baboons located approximately 100 km south of the

Zambezi River, within Gorongosa National Park (Fig. 1) [23]. The park hosts a large population of baboons, which comprises more than 200 groups [31] which have been previously assigned to *P. ursinus griseipes*, in line with the clustering of Gorongosa mitochondrial DNA with lineages sampled in the northern area of *P. ursinus* distribution [10, 25]. However, cranial morphometric variation placed Gorongosa baboons in between *P. ursinus* and *P. cynocephalus*, closer to *P. ursinus*, but with several phenotypes observed in Gorongosa described as a mosaic of features of the two species [23].

Given the potential for baboons in Gorongosa National Park having experienced interspecies admixture, we posit to investigate their evolutionary history, including testing for evidence of gene flow, using a genomic approach. We generated low and high coverage genome sequences of two baboons from Gorongosa National Park, and compared them to available genomes from other baboon species revealing a complex evolutionary relationship between baboons in Gorongosa, and other members of the genus *Papio*. We complemented the whole genome analysis with the investigation of uniparental markers, exploring whole mitogenome data and Y chromosome sequences several hundreds of thousand nucleotides in length, providing female and male specific insights into the phylogenetic relationships of Gorongosa baboons within the genus *Papio*.

## Results

### Whole genome sequence of baboon samples from Gorongosa National Park, Mozambique

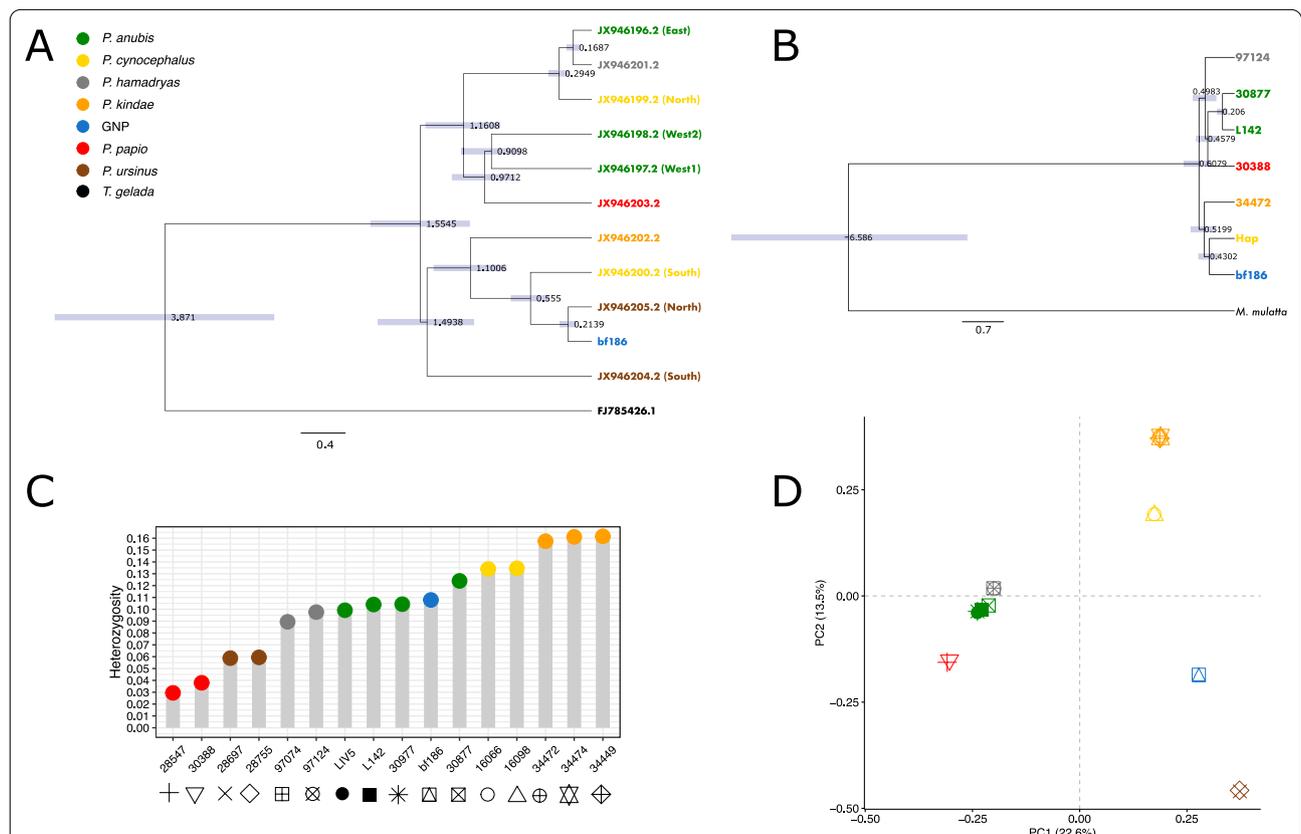
We generated the whole genome sequence of a baboon individual sampled in Gorongosa National Park, in central Mozambique (sample GNP). The median coverage across the genome was 36.7X. The X chromosome was reported having a coverage of 18X, approximately half of the coverage reported for chromosome 20 (35X), consistent with the ratios observed and corresponding information on sex (Additional file 2: Table S1) for baboon samples included in Rogers et al. [11]. This led us to conclude that the sample was originally collected from a male individual. Out of the more than 36 million variants identified across the whole genomic dataset analyzed in this study (GNP and [11]), 1,488,924 variants were found to be private to the GNP baboon. We filtered variants to remove singletons, retaining a final set of 36,538,410 biallelic polymorphisms.

We additionally generated whole genome sequence data from a blood sample also collected from a Gorongosa baboon (sample BB1). The quality of the DNA was such that the mean coverage across the genome was only 0.03X. The reads were mapped to the Panu\_3.0 reference genome through BWA-MEM (v. 0.7.17-r1188).

**Mitochondrial phylogenetics**

As part of the genome sequencing, we also retrieved the whole mitogenome of the GNP baboon. GNP mtDNA sequence matched the few hundred base pairs previously analyzed in samples collected in Gorongosa [13, 25]. To fully explore the mitochondrial variation, we assembled a reference dataset of *Papio* mitogenomes and reconstructed the phylogenetic relationship of our sample and other baboon mitogenomes [10] (Figs. 1; 2A). *P. ursinus* mitogenomes are located in two distinct parts of the phylogenetic tree, one comprising samples from the southern part of the *P. ursinus* range and one clustering the samples further north [10, 21]. The GNP mitogenome clustered with the northern *P. ursinus* samples, in a clade also including *P. kindae* and southern *P. cynocephalus*. Moreover, we dated the different nodes in the tree, obtaining dates in line with previous

results [10]. The Time to the Most Recent Common Ancestor (TMRCA) for the GNP/Northern *P. ursinus* mtDNA clade was 0.21 Mya (confidence interval: 0.14–0.29) (Additional file 2: Table S3). The TMRCA of the next node, extended to include the mtDNA of a southern *P. cynocephalus* sample, was dated to 0.55 Mya (0.38–0.73), while the clade including all *P. ursinus* mtDNA lineages dated to 1.49 Mya (1.06–1.94) (Additional file 2: Table S3). We also generated a tree using Zinner et al. [10] dataset and all of the mitogenomes recovered from whole genome sequence data (GNP and [11]) (Additional file 1: Fig. S2). The tree showed a topology identical to the one presented in Fig. 2A; TMRCA were broadly similar, the CIs being larger but in line with the estimates generated by Mathieson et al. [21] using the same whole genome data from Rogers et al. [11] to extract mitogenomes.



**Fig. 2** GNP baboon genomic variation and *Papio* diversity. **A** Phylogeny of *Papio* mitogenomes using *Theropithecus gelada* as an outgroup; codes refer to the specimen identifiers used in Zinner et al. [10]. GNP: baboon sample collected in Gorongosa National Park (code: bf146); colors indicate the different species listed in the legend. All nodes have a posterior probability of 1. Divergence times are reported (confidence intervals shown as gray boxes; values in Tables S2). **B** Y chromosome *Papio* phylogeny based on six concatenated genes (see main text); codes as in Additional file 2: Table S1; Hap refers to a *P. cynocephalus* male sample from Wall et al. [15]; colors refer to legend in panel A. All nodes have a posterior probability of 1. Divergence times are reported (confidence intervals shown as gray boxes; values in Tables S3). *Macaca mulatta* was used as an outgroup. **C** Heterozygosity estimates across *Papio*; colors as legend in panel A; sample codes as reported in Additional file 2: Table S1. **D** Principal component analysis (PCA) of *Papio* autosomal genomic data; codes and colors as in panel legend in panel A

### Y chromosome analysis

We extended our genomic analysis to include sequences from six Y chromosome genes: *SRY*, *DDX3Y*, *KDM5D*, *ZFY*, *UTY*, *USP9Y*. Approximately 456 kb were recovered for each of the seven baboon males available to us: the GNP sample, the five males present in the Baboon Genome Diversity Panel [11] (Additional file 2: Table S1; Fig. 1; see Methods section) and a previously published genome of a male *P. cynocephalus* individual [15]. *P. ursinus* male whole genome sequence data was not available. We aligned the *Papio* sequences including *M. mulatta* Y chromosome data as an outgroup and retrieved a concatenated section of Y chromosome 425 kb in length. A total of 2,098 variable positions were found across *Papio* Y chromosomes, of which 267 variants were detected only on the GNP Y chromosome (Additional file 2: Table S4). We used the retrieved sequences to generate a phylogenetic tree (Fig. 2B), the first exploring Y chromosome variation over hundreds of thousands of nucleotides across *Papio* species. The topology of the tree supported a separation of Y chromosome baboon lineages in two clades, one comprising only northern baboon species and the other only southern baboon individuals. A similar separation was reported when genomic and mtDNA data were analyzed [10, 11] (Fig. 2A). The closest Y chromosome lineage for the GNP sample belongs to *P. cynocephalus* from Amboseli National Park in southern Kenya [15]. Interestingly, the mtDNA lineages of Amboseli *P. cynocephalus* have been shown to cluster within the northern clade of the mitochondrial phylogeny [13]. The divergence time for the GNP/*P. cynocephalus* clade was dated to 0.43 Mya (C.I. 0.26–0.63 Mya), and the TMRCA of the whole southern baboons' clade (*P. cynocephalus*, *P. kindae* and GNP) was dated to 0.52 Mya (0.32–0.75 Mya). The TMRCA for the whole of the Y chromosome *Papio* lineages here included was dated to 0.61 Mya (0.38–0.87 Mya) (Additional file 2: Table S5).

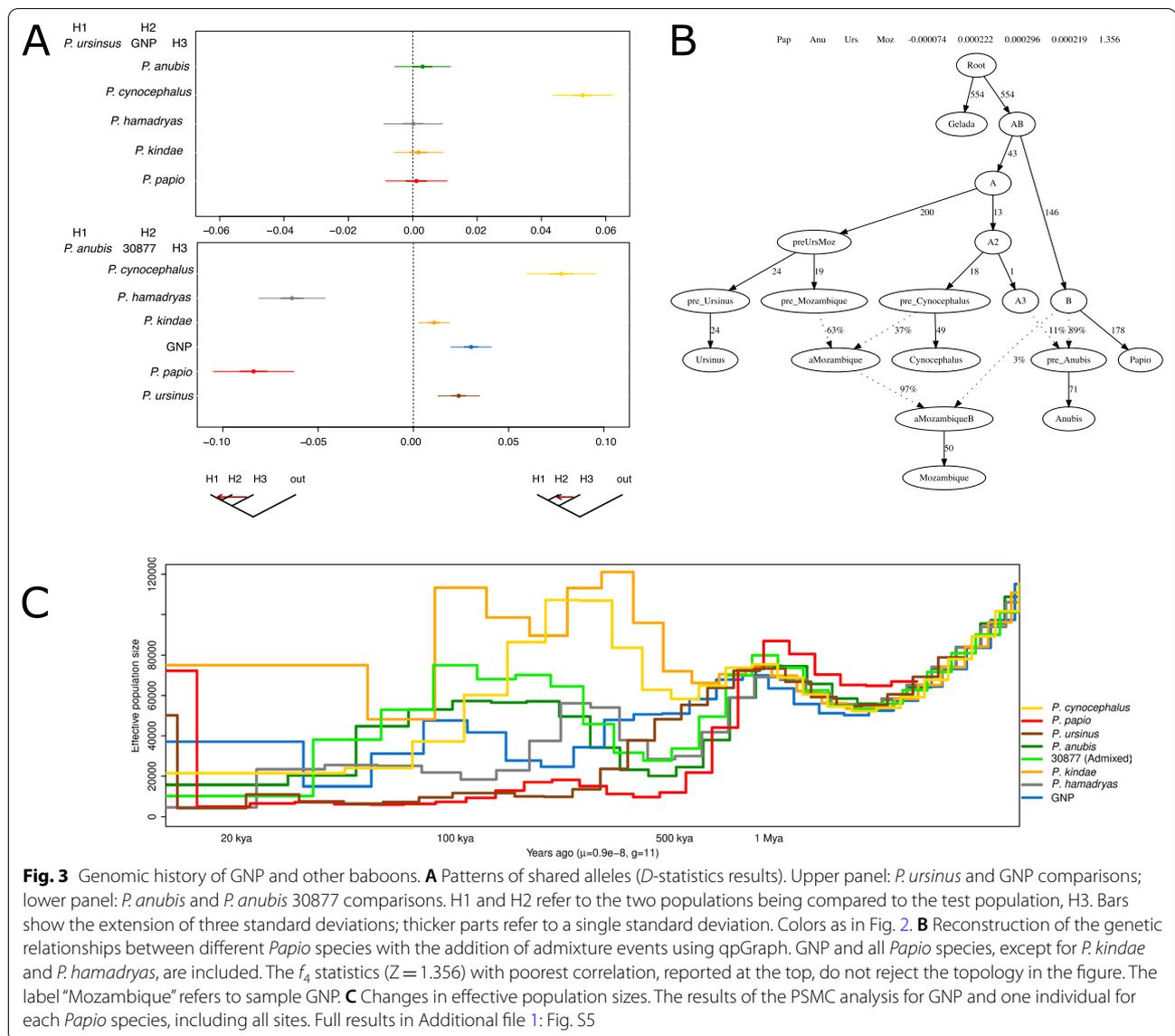
### Population genomics of Gorongosa baboons

The GNP sample was compared to Rogers et al. [11] baboon genome data using the genomic variants called as described in Material and Methods. The heterozygosity for each genome was estimated as the fraction of the total number of variable sites being heterozygous in a given individual (Fig. 2C; Additional file 2: Table S1). Baboons sampled in captivity (*P. ursinus* and *P. papio*) were characterized by lower SNP heterozygosity than individuals collected in the field, as previously reported [11]. The estimate for the GNP sample at 0.12 is in the middle of the heterozygosity range for other baboons (0.03–0.16), slightly lower than values reported for *P. cynocephalus* (0.13) but twice as much of that observed in *P. ursinus* (0.06) (Fig. 2C). Principal Components Analysis (PCA)

was implemented to visualize the relationship of the GNP baboon with other *Papio* species (Fig. 2D). PC1 separated baboons along the same North–South division observed in the mtDNA (with the exception of *P. cynocephalus* being paraphyletic) and Y-chromosome analyses and previously reported for genomic data [11]. PC2 further separated southern baboons in two groups: *P. kindae*/*P. cynocephalus* and *P. ursinus*/GNP, the GNP sample located in between *P. cynocephalus* and *P. ursinus*, closer to the latter. The low coverage sample BB1 was closest to GNP when projected on the PCA, the two separated to a certain extent along PC1. A similar behavior was observed when GNP was downsampled to a coverage comparable to that of BB1 (i.e. GNP\_ds) and projected on the PCA, suggesting that the departure of BB1 from GNP is affected mostly by the degree of genome coverage more than the occurrence of significant differences between the two Gorongosa samples (Additional file 1: Fig. S1A).

The genus *Papio* is characterized by a rich history of admixture between species [11, 15, 17] and Gorongosa baboons display a mosaic of phenotypic features [23]. We therefore tested if the genomic variation of the GNP sample could be modeled as the result of an admixture event involving different baboon species. As a control, we ran each of the analyses performed on the GNP sample on sample 30877, an individual known to be admixed between *P. anubis* and *P. cynocephalus* [11] (Supplementary Information).

We used the  $f_3$  statistics in the form of  $f_3$  (GNP; X, Y), where X and Y represented different baboon species [11]. None of the tested pairs provided significant ( $Z \leq -3$ )  $f_3$  results for GNP (Additional file 2: Table S6). We then investigated to what extent GNP and *P. ursinus* were similarly related to other *Papio* species. We implemented the  $D$ -statistics in the form of  $D(T, \text{gelada}, X, \text{GNP}, P. \text{ursinus})$ , X representing individuals belonging to other *Papio* species. GNP resulted significantly enriched in *P. cynocephalus* alleles relative to *P. ursinus* ( $Z = 5.31$  and  $Z = 5.37$ , when compared to the two *P. ursinus* samples, 28755 and 28697, respectively) (Fig. 3A). The Gorongosa sample BB1 was not significantly different from GNP when the two were tested in the form  $D(T, \text{gelada}, P. \text{cynocephalus}, \text{GNP}, \text{BB1})$  (Additional file 1: Fig. S1B). Notably, both showed significant differences when compared to *P. ursinus* for the number of shared alleles with *P. cynocephalus* (Additional file 1: Fig. S1). Testing GNP vs. other species did not generate significant departures from zero; however, when *P. anubis* baboons were included in the analysis, the  $D$  value diverged from zero more than others, even if not significantly. We quantified the amount of *P. cynocephalus* ancestry present in the GNP genome using the  $f_4$  ratio alpha in the form  $f_4(\text{gelada}, \text{cynocephalus}, \text{GNP}, \text{ursinus})/f_4(\text{gelada},$



*cynocephalus*; *cynocephalus*, *ursinus*). The estimated value of alpha was 8% (Additional file 2: Table S7). On the contrary, no significant contribution of *P. cynocephalus* was detected in GNP when the  $f_4$  ratio was calculated using X chromosome SNPs (Additional file 2: Table S7). We also attempted local ancestry deconvolution of the *P. ursinus*/*P. cynocephalus* components in GNP using ELAI but failed to retrieve any *P. cynocephalus* contribution (Additional file 2: Table S8; Supplementary Information). This outcome is in line with the non-significant  $f_3$  admixture tests but contrasts with the  $f_4$  ratio results. Comparatively, known admixed baboon 30877 gave significant values for both  $f_3$  and  $D$ -statistics when set-up similarly to GNP tests (i.e. *P. anubis* in place of *P. ursinus*), in

addition to an  $f_4$  ratio alpha that yielded approximately a *P. cynocephalus* contribution of 11% (Additional file 2: Table S7). Local ancestry analysis conducted using *P. anubis* individuals as recipients and *P. cynocephalus*/other *P. anubis* as donors consistently identified *P. cynocephalus* ancestry within 30877, but a smaller fraction was detected when 30877 was the only *P. anubis* sample included as target (11% and 7% respectively; Additional file 2: Table S8, Supplementary Information). *P. cynocephalus* contribution to the *P. anubis* individuals here analyzed was dated 10–20 generations ago using Local Ancestry and linkage-based approaches (Additional file 1: Fig. S7, Additional file 2: Table S9; Supplementary Information). No substantial differences were observed when

a recombination map was included in the analysis (Additional file 2: Table S8; Additional file 1: Fig. S8). Interestingly, the heterozygosity of 30877 decreased when SNPs included in regions assigned as having a *P. cynocephalus* origin were excluded, reaching values in line with other *P. anubis* samples (Additional file 2: Table S2; Supplementary Information).

To gather additional insights on the evolutionary relationships of GNP with other baboons we explored different tree-topologies using qpGraph (Fig. 3B; Additional file 1: Fig. S3). We initially tested the tree proposed by Rogers et al. [11]. Major features of this tree are a subdivision in northern and southern baboons, *P. kindae* modeled as the result of the admixture between northern and southern baboons, *P. papio* identified as the recipient of gene-flow from an unknown *Papio* lineage and *T. gelada* assigned as an outgroup to *Papio* variation. This proposed topology was not supported, in particular the use of *T. gelada* as a full outgroup to *Papio* (Additional file 1: Fig. S3). We therefore started from a simpler topology and subsequently added different lineages, including admixture events (Supplementary Information). The addition of GNP to the tree required the inclusion of contributions from sources related to both *P. cynocephalus* and a proto northern baboon population (~37% and ~3%, respectively). The inclusion of only one of the two contributions resulted in unsupported topologies (data not shown). QpGraph analysis supported the inclusion of *P. anubis* only when an admixture event involving 11% contribution from *P. cynocephalus* was added (Fig. 3B). Contacts between *P. cynocephalus* and GNP/*P. anubis* were also supported by Treemix analysis (Additional file 1: Fig. S4). The most supported trees in Treemix included either one or five migration edges between different branches. All trees reported a migration edge between the *P. ursinus* lineage and *P. kindae*. The additional edges connected *T. gelada* with *P. hamadryas* and *P. papio*, and *P. cynocephalus* with *P. anubis* and GNP (Additional file 1: Fig. S4).

Finally, we attempted to reconstruct the changes in population size across time of the Gorongosa population. We examined the distribution of coalescent times across the genome of the GNP sample and all the other genomes here included [11] using the PSMC software. A selection of the PSMC runs (one individual per species) is included in Fig. 3C, together with the results of the analysis of the GNP baboon (full results in Additional file 1: Fig. S5). These PSMC runs replicate results already published in addition to providing a direct comparison for the GNP sample which was analyzed here for the first time [11]. The GNP genome showed fluctuations in population sizes between 500 and 100 kya. At 50 kya, GNP baboon population size was similar to other *Papio* populations in

the wild, but higher than the one reported by the two species sampled in captivity, *P. ursinus* and *P. papio*. Interestingly, population sizes for GNP and *P. ursinus* started to diverge at the time when they began to decrease, around 500 kya (Fig. 3C). Notably admixture events can lead to fluctuations in effective population size, sometimes denoted in PSMC curves as large humps or, in more complex scenarios, as increasing and decreasing humps over time [32, 33]. Fluctuations in size between 500 and 100 kya are present in both the ancestral populations of GNP and 30877, markedly so when compared to *P. ursinus* and *P. anubis*, respectively (Fig. 3C; Additional file 1: Fig. S5). Nevertheless, ancient population structure cannot be ruled out for the GNP baboon when also taking into account the discordant results obtained from the admixture tests performed [34, 35].

The PSMC analysis of GNP mapped to Panubis1.0 was consistent with results observed using genomic data mapped to Panu\_3.0. (Additional file 1: Fig. S6).

## Discussion

### Gorongosa National Park baboons in the context of *Papio* variation

The recent publication of whole genome sequences of samples representative of the six different baboon species has provided the context for investigating the variation recorded in other baboon populations [11]. We used this resource to investigate the genomic variation recovered from two baboons in Gorongosa National Park, Mozambique (GNP and BB1, high and low coverage, respectively). Southeast Africa is a region that has been only marginally investigated so far in relation to the local baboon genetic variation [4, 13, 17, 25]. Autosomal and mtDNA data were consistent in suggesting a closer affinity for Gorongosa baboons to *P. ursinus* (Fig. 2; Additional file 1: Fig. S1). The extensive dataset available for *Papio* mitogenomes provided further resolution within *P. ursinus*, and confirmed previous results placing GNP mitochondrial lineage closer to baboon from the northern part of *P. ursinus* range (Fig. 2A; Zinner et al. [10]). The TMRCA for GNP and the closest *P. ursinus* mtDNA lineage was dated to around 200 kya, while the earliest coalescent event with *P. cynocephalus* was more than 500 kya. We also extended our phylogenetic analysis of uniparental markers to the Y chromosome, analyzing more than 400 kb of the non-recombining part of the Y chromosome. No *P. ursinus* Y chromosomes were available for comparison and the closest Y chromosome lineage to GNP belonged to a *P. cynocephalus* individual from Amboseli National Park, southern Kenya [10, 13–16], with a TMRCA dated to 0.38 Mya (Fig. 2B). Y chromosome and mtDNA results are broadly in line in identifying Gorongosa baboons as belonging to “the southern

baboons' lineage" described in Zinner et al. [13], with similar dates for the mtDNA divergence time with *P. cynocephalus*. However, the mtDNA dates for the overall antiquity of southern baboons, with or without *P. ursinus* lineages from the south (1.10 Mya and 1.49 Mya respectively; Additional file 2: Table S2), were much older than the estimate obtained for the Y chromosome for the GNP/*P. kindae*/*P. cynocephalus* clade (0.52 Mya; Fig. 2B). Such discrepancies could be related to our limited survey of Y chromosome variation across baboons, in particular the absence of Y chromosome types belonging to *P. ursinus* sampled further south. But we note that this is also the case when the overall antiquity of the *Papio* clade was considered (0.61 Mya and 1.55 Mya, for the Y chromosome and the mtDNA, respectively) and limited sampling alone might not be enough to explain these results. Male-biased dispersal and polygynous mating patterns have been proposed as ways to explain Y chromosome TMRCA significantly younger than mtDNA ones observed in gorillas (genus *Gorilla*) and orangutans (genus *Pongo*) [22]. A refined Y chromosome phylogeny generated by including a wider and more informative dataset will make it possible to investigate the impact of such behaviors on the genetic variation of *Papio*.

#### The genomic history of Gorongosa baboons

Gorongosa baboons have been historically defined as belonging to *P. u. griseipes*, because the geographical location of Gorongosa falls within its range of distribution (i.e. IUCN, [36]). However, the taxonomic assignment of baboons from the Lower and Middle Zambezi region (Malawi, eastern Zambia, and north-central Mozambique) are not yet fully understood [9]. Recent morphological/craniometric analyses reported a mix of *ursinus/cynocephalus* features in this population [23].

Despite *P. ursinus* being the *Papio* species closer to Gorongosa baboons, the genomes of the Gorongosa specimens were different from *P. ursinus* samples. GNP and *P. ursinus* had well differentiated mtDNA lineages (Fig. 2A), presented different degrees of variation (Fig. 2C), were separated along PC2 (Fig. 2D; Additional file 1: Fig. S1), and were characterized by different demographic histories (Fig. 3C). The two groups also differed in their affinity with other *Papio* lineages, as the GNP genome resulted closer to *P. cynocephalus* than *P. ursinus* (Fig. 3A) and showed a potential contribution from an additional source (Fig. 3B). BB1 results based on low coverage data were consistent with those of GNP (Additional file 1: Fig. S1). Contrary to *P. anubis* 30877,  $f_3$  and ELAI results did not support a simple recent admixture scenario for the increased affinity of GNP to *P. cynocephalus*. Combined, the  $f_3$ ,  $f_4$  ratio, ELAI and  $D$ -statistics suggest a complex relationship between GNP and *P.*

*cynocephalus*. Male-mediated introgression of *P. ursinus* into *P. cynocephalus* (the "nuclear swamping" scenario) has been suggested to explain the relatedness between the mtDNA lineages present in northern *P. ursinus* and southern *P. cynocephalus* [37], a hypothesis compatible with our genome-wide results. Interestingly, the GNP X chromosome, differently from the rest of the genome, does not show a closer affinity to *P. cynocephalus*, suggesting that contacts might have been male mediated. However, if indeed male-biased gene flow was responsible for the observed autosomal/X imbalance, introgression would have been operating, from *P. cynocephalus* to *P. ursinus*, opposite to what was previously proposed. Male hybrid sterility is an alternative explanation for rapid depletion of introgressed X chromosomes. Both have been proposed to explain similar patterns observed in modern humans in relation to Neanderthal introgression [38], but infertility has not been reported across the different *Papio* hybrid zones identified so far [2]. One of the topologies supported by qpGraph (Fig. 3B; Additional file 1: Fig. S3) suggests the occurrence of subdivision within *P. ursinus* and contacts between the GNP lineage and a lineage related to *P. cynocephalus*. An origin of *P. cynocephalus* from a population of proto *P. ursinus* closer to Gorongosa baboons, and subsequent gene flow among *P. ursinus* populations, including GNP complements our findings. A similar scenario, labeled as ancient population structure, has been reported to inflate  $D$  results and was initially suggested to explain the increase in shared variants between non-African populations and Neanderthals [39]. Both scenarios (gene flow and ancient population structure) assume the presence of deep population subdivision within *P. ursinus*, which is supported by previous surveys of *P. ursinus* mtDNA variation [25]. The deep TMRCA of GNP/*P. ursinus* mtDNA clade and the PSMC results suggest a separation (complete or partial) between the tested *P. ursinus* and GNP genome dating to hundreds of thousands of years ago. We also note here that *P. cynocephalus* from the north (as the ones available here) might not represent an appropriate proxy of the true source population and the non-significant  $f_3$  results might have been driven by the occurrence of population structure in *P. cynocephalus*, which is suggested by mtDNA data [13] (Fig. 2A), and supported by the identification of multiple morphotypes in yellow baboons [9]. Similarly, drift operating on the three populations since the event of admixture might have affected the  $f_3$  results [40].

The indication of additional contributions to GNP related to proto northern baboons is more difficult to interpret given our current limited understanding of baboon genomic diversity. We also note that the evidence suggesting this contact is based on only one set

of analyses (qpGraph). One possible explanation might be that northern baboon variants entered Gorongosa via interactions with Kinda baboons. *P. kindae* have been suggested as originating from an ancient admixture event between northern and southern baboons. However, neither the  $f_3$  results nor the  $D$ -statistics support a direct GNP/*P. kindae* admixture (Additional file 2: Table S6, Fig. 3A). Alternatively, northern baboon ancestry might have entered the Gorongosa population via *P. ursinus* populations mixed with *P. kindae*. Notably, *kindae/ursinus* admixture has been detected in Zambia [4, 17]. Unexpected variation might have also originated from more exotic sources, including different Papionini. Intergenera interactions have been highlighted for *P. cynocephalus/R. kipunji* using mtDNA data, and *Papio/Theropithecus* interactions have been reported in captivity and directly observed in the wild, and suggested on the basis of transposable elements detected in both genera [41–44], the latter compatible with our qpGraph and Treemix results (Additional file 1: Figs. S3, S4). The characterization of these occurrences and the relevance for Gorongosa baboons are definitely worth additional investigation, which will be enabled by the addition of other Gorongosa genomes in the future.

Finally, we note that the lack of significant reduction in genetic variation reported for the GNP genome (Fig. 2C) is of interest in the context of the decrease in size of local populations reported for large mammals after the Mozambican Civil War from 1976 to 1992 [31]. In combination with the steady increase in the number of troops since 1994 and an effective population size similar to other *Papio* species in their more recent evolutionary history (Fig. 3C), this makes the morphological and genetic variation of Gorongosa baboons a valuable resource to investigate the selection-based dynamics that have shaped local diversity. Notably, the curves for captive samples of *P. ursinus* and *P. papio* are indicative of a long-term reduction in population size, dating back to hundreds of thousands of years. This opens up the possibility that the low degree of heterozygosity displayed by these individuals might not simply be the result of their captive status but instead reflect a more general reduction in diversity in the populations in the wild from which they descend. It is relevant here to note that mtDNA analysis suggested an overall reduction in variation for *P. papio* [45, 46].

#### Implications for our understanding of the evolutionary history of the genus *Homo*

The results gathered from the analysis of Y chromosome, mitochondrial and genome data have highlighted the presence of complex dynamics across *Papio* populations in southern Africa. The timeline of these events spans

the last few hundreds of thousands of years, a time when phenotypically differentiated *Homo* populations were present in the same area. It is tempting to speculate on the significance of the parallels between *Homo* and *Papio* in the region: the occurrence of morphologically different *Homo* populations in southern Africa more than 230 kya (e.g. *H. naledi*, Florisbad and Kabwe) [47, 48] and the genomic differentiation reported for GNP and other *P. ursinus* individuals. More data is necessary to explore the full extent of the genomic variation present across *P. ursinus* subpopulations. Nevertheless, investigating how morphology and genomics map onto each other across *P. ursinus* morphotypes and providing a chronological context for their relationships is expected to help develop an interpretative framework for the phylogenetic significance of the phenotypic diversity observed in *Homo*. In this context the observation of shared polymorphisms between GNP/BB1 and *P. cynocephalus* is particularly enlightening as it suggests that relationships extending further north might have been also present among *Homo* groups, much in line with the African metapopulation model that has been proposed for the origin of our own species [49]. The further integration of climatic records is expected to highlight the role played by ecological changes in shaping events of drift, gene flow, isolation and extinction in the region in both *Papio* and *Homo* [50].

#### Conclusions

Gorongosa National Park in southeastern Africa lies within the region occupied by *P. ursinus griseipes* (gray-foot baboons), which is one of three morphotypes identified within the *P. ursinus* morphological variation [9]. Our results provided evidence for increased allele sharing between baboons in Gorongosa and *P. cynocephalus*, not related to recent gene flow. A mosaic of *P. ursinus/P. cynocephalus* features have been reported in Gorongosa baboons, but to which extent *P. cynocephalus* genetic variants affected their morphology remains unclear [23]. Similarly, if and how much the occurrence of *P. cynocephalus* variants extends into other grayfoot baboon populations and into other *P. ursinus* morphotypes is still unknown. Additional Gorongosa genomes collected from morphologically characterized individuals and the sampling across different populations of *P. ursinus* are expected to clarify the dynamics that shaped the phenotypic variation and the evolutionary relationships of populations and species of baboons in southern Africa. Finally, extensive genomic survey of *Papio* diversity, including Y chromosome, will provide additional insights on the extent of within-species population structure and interspecies sex-biased gene-exchange, contributing to the development of an interpretive framework for

the phylogenetic significance of the phenotypic diversity observed in other genera in the region, including *Homo* [2].

## Materials and methods

### Sample collection and DNA extraction

DNA to generate high coverage whole genome sequence was obtained from a muscle tissue sample collected post-mortem as a result of an infanticide that took place in Gorongosa, Mozambique, in 2017. The infanticide occurred in the early afternoon in Chitengo Safari Camp (coordinates lat:  $-18.97727$ ; long:  $34.35140$ ). A sample of muscle tissue was collected ca. four hours after the death of the animal. As in other baboons of similar age in Gorongosa [23], the infant's pelage was black, but no other morphological details were recorded, as the body was left in the open after sample collection to continue monitoring the spatial dynamics of the members of the troop around the body. By evening, the carcass was taken by a genet (*Genetta genetta*).

We additionally collected a blood sample in 2019 from a sedated female adult individual belonging to a troop different from the one roaming in and near Chitengo to which the infant (Sample GNP) belonged (coordinates of the collected sample: lat  $-18.9051644$ ; long  $34.3760742$ ; ca. 8 km north of Chitengo). The animal was sedated to be collared as part of the Paleo Primate Project Gorongosa (PPPG), to track the movement of the troop.

DNA was initially extracted from tissue and blood at the Molecular Genetics laboratory in Chitengo using DNeasy Blood and Tissue kit (Qiagen), with minor improvements from the manufacturer's protocol (e.g. overnight digestion of tissue sample using proteinase K). The DNA extracts were further purified using RNase at the Research Centre in Biodiversity and Genetic Resources (CIBIO, Vairão, Portugal). Extracted DNA was quantified with a Qubit fluorometer using the Qubit dsDNA BR Assay kit (Invitrogen) following manufacturer's instructions. Genome sequencing was performed by Edinburgh Genomics (Edinburgh, UK). The tissue sample from the infanticide is here referred to as GNP (code: bf186). The blood sample coagulated after collection and as a result only a small amount of DNA could be recovered; this sample is here referred to as BB1.

### DNA sequencing and genome mapping

Genome sequencing was performed by Edinburgh Genomics (Edinburgh, UK). The GNP sample was sequenced to an average depth of  $36.8\times$  using Illumina HiSeq X platform (150 bp paired-end reads). The integrity of the sequenced sample yielded a total of 121.4 Gb of data, with 98.8% of reads mapped to the Panu\_3.0 assembly. We aligned our Illumina reads using BWA-MEM

0.7.17-r1188 [51] to baboon reference Panu\_3.0 assembly and generated a BAM file which was used for further downstream analyses. Duplicate reads were marked and removed using Picard MarkDuplicates version 2.8.1 [52]. A total of 100 Mb sequence data of the BB1 genomic DNA was generated using a MiSeq instrument, which resulted in a 0.03X coverage once mapped to Panu\_3.0. Given the limited amount of data, BB1 was only used for Principal Components Analysis and *D*-statistics, to provide further insights in the genomic variation present in Gorongosa baboons, as described below.

After we started to analyze the GNP sample, an additional baboon reference genome became available (Panubis1.0 [53]). Structural differences between the Panu\_3.0 and Panubis1.0 assemblies have been highlighted. We explored the impact a different assembly might have on our results by re-mapping the GNP genomic data to Panubis1.0 and using it to run PSMC analysis on GNP only (see below). We want to note here, that a systematic evaluation of how two assemblies might differ when used for population genomics investigations is out of the scope of this manuscript. We therefore limited this comparison to a subset of analyses that are potentially affected by differences in the assemblage of contiguous genomic sections, such as PSMC.

### Sex determination

The injuries suffered by the GNP individual and the very young age prevented a morphological identification of the sex of the animal. We therefore attempted to sex it by investigating the genome coverage across chromosomes. As males comprised only one X chromosome, the average X chromosome coverage should be approximately half of that for the autosomes, which are present in pairs. The sex of the individual was established by comparing the average coverage reported for sex chromosome X and the coverage obtained for autosomal chromosome 20. We validated this approach after Rogers et al. [11]; we also included the three samples for whom the sex was not provided in the 2019 paper (codes: 28547, 30877, 30977). Additionally, the sex of the GNP baboon was confirmed using a molecular sex determination protocol that simultaneously amplifies via Polymerase Chain Reaction (PCR) fragments of the amelogenin X gene and the SRY Y-linked gene [54] (data not shown).

Sample BB1 was taken from a female baboon while she was sedated to place a radio-collar on her.

### Mitochondrial DNA analysis

We first downloaded the complete mitochondrial genomes originally published by Zinner et al. [10]. We then mapped the reads of the GNP individual, as well as the reads of each baboon from the Baboon Genome

Diversity Panel [11], to the mitochondrial reference for *Papio anubis* found in the papAnu2 assembly (NC\_020006.2). To obtain a consensus mitogenome from whole genome sequence data we used ANGSD [55]. A fasta file was created for each sample, where consensus sequences were called per individual using the highest effective depth. A minimum base quality of 30 and a minimum of 10 reads were used to consider a base for the consensus.

NGS consensus mitogenomes and Zinner et al. [10] mitogenomes were concatenated and aligned using the multiple sequence aligner, MUSCLE 3.8.31 [56], an algorithm best suited for nucleotide datasets. The resulting alignment of 16,569 bp was then further filtered using GBlocks 0.91b [57] to remove any indels and gaps unsuitable for phylogenetic analyses. The mtDNA alignment, with both whole genome sequence consensus mitogenomes and Zinner et al. [10] mitogenomes, was 16,402 bp in length. We used jModelTest2 [58] and IQ-TREE [59] to determine the best substitution model for our alignment based on the corrected Akaike inference criterion (AICc) and Bayesian inference criterion (BIC). Both AICc and BIC gave TrN+I+G as the best substitution model when using jModelTest2 and similarly, IQ-TREE gave TN+F+R3 as the best fit (i.e., TrN using a FreeRates model instead of a gamma distribution). We produced the best Maximum Likelihood (ML) tree with support values from 10,000 bootstraps. Similarly, we also produced an mtDNA alignment using only our GNP mitogenome and Zinner et al. [10] mitogenomes. For this alignment jModelTest2 and IQ-TREE both recommended GTR as the best substitution model.

Divergence dates for the mitogenomes were estimated using BEAST 2.6.3 [60]. A relaxed lognormal clock model was chosen to model lineage variation and birth–death process prior branching rates. Using the jModelTest2 recommendation we used a TrN+I+G site model for the dataset including Rogers et al. [11] mitogenomes while for the dataset using only Zinner et al. [10] mitogenomes and our GNP mitogenome we used a GTR site model. We provided a fossil-based calibration point for the *Theropithecus-Papio* node at 4 Mya (CI: 0.5 Mya) based on previous work [10]. Four replicates were performed for 25 million generations, tree and parameter sampling taking place every 1000 generations. We used Tracer 1.7.1v [61] to check whether each replicate, with a 10% burn-in, resulted in convergence of all parameters across generations. These sampling distributions were then combined using 25% burn-in with LogCombiner. A consensus tree was produced using TreeAnnotator where node heights were summarized to reflect posterior median node heights. Lastly, the consensus tree was visualized using FigTree v1.4.4 (<http://tree.bio.ed.ac.uk/>

[software/figtree](http://tree.bio.ed.ac.uk/software/figtree)). Two trees were produced: one including all the mitogenome sequences recovered from whole genome sequences (GNP and Rogers et al. [11]) plus the Zinner et al. [10] dataset, and one with only the Zinner et al. [10] dataset and the GNP mitogenome here assembled.

### Y chromosome analysis

The Baboon Genome Diversity Panel analyzed in Rogers et al. [11] included data for four males, one each from *P. anubis*, *P. papio*, *P. hamadryas* and *P. kindae*. We identified an additional male sample in the Rogers et al. [11] dataset by comparing the coverage of the X chromosome and chromosome 20 (sample 30877, *P. anubis*; Additional file 2: Table S1). To complement this set, we further included the genomic data of a *P. cynocephalus* male presented in Wall et al. [15], together with the genomic sequence of the male GNP baboon we generated, and the *M. mulatta* (NC\_027914.1) Y chromosome reference sequence. We recovered data from six Y-specific single-copy genes (*SRY*, *DDX3Y*, *KDM5D*, *ZFY*, *UTY*, *USP9Y*) selected as described in the Supplementary Information. Preliminary phylogenetic trees were built using the ML algorithm and 100 bootstrap replicates. The best substitution model was evaluated with the “Find best DNA/Protein Models” function available in MEGAX [62]. The best substitution model, both in terms of BIC and AICc, was the GTR+G+I. The reconstruction of the divergence tree, based upon the best substitution model above identified, and molecular dating were performed with BEAST v1.10.4 [63]. The model chosen was an uncorrelated lognormal relaxed clock model with a Yule speciation model. The starting tree was built with the Unweighted Pair Group Method with Arithmetic Mean (UPGMA). One calibration point was applied as normal prior with a mean of 7 Mya, and a standard deviation of 1 Mya to the node containing all Papionini species [64]. The Markov chain Monte Carlo was run for 100 million generations sampling every 1000 generations. The log file resulting from the analysis was imported in Tracer v1.7.1 [61] with a burn-in of 30% to assess convergence of all parameters by examination of ESS values. Finally, the consensus tree was produced using TreeAnnotator and visualized with FigTree v1.4.4 (<http://tree.bio.ed.ac.uk/software/figtree>).

### Genome comparisons

#### Variant calling

Following GATK best practices, we called variants for the autosomal chromosomes using GATK version 4.1.8.1 [65]. Indels were realigned using IndelRealigner and followed by HaplotypeCaller which was used to generate gVCFs for our sample. Initially, we performed joint

genotype calling with 16 previously published baboons representing six species of *Papio* (nine individuals collected in the wild and seven from captive colonies) [11]. Joint calling was done as reported by Rogers et al. [11]. Filtered GATK variants were used in all the genome-based analyses, excluding the PSMC analysis where variants were determined using individual-based genotype calling with bcftools [66]. PSMC requires a consensus sequence score (FQ) alongside the consensus sequences for each individual; a calculation only provided by bcftools at this time. For consistency, we also replicated the genome-based analyses (PCA, *D*-statistics, heterozygosity) using these individual-based calls, the results agreeing with the ones based on the joint-calling (data not shown). All variants were filtered using the following hard-filters (SNPs: “QD<2.0 || FS>60.0 || MQ<40.0 || MQRankSum<-12.5 || ReadPosRankSum<-8.0”; Indels: “QD<2.0 || FS>200.0 || ReadPosRankSum<-20.0”). We also filtered to exclude indels and multiple nucleotide polymorphisms providing a total of 56,030,625 SNPs (GATK). Only sites with biallelic calls were kept. Sites with a minor allele frequency (MAF) below 0.05 were removed. After filtering, a final set of 36,538,410 autosomal SNPs was used for further statistical analyses. Similarly, variants were called for the X chromosome where, after filtering, a set of 1,413,976 SNPs was obtained. Variant calling for BB1 was done using GATK v4.2.4.1 *HaplotypeCaller*, which generated a set of 3,716 SNPs in the final gVCF.

### Heterozygosity

Using PLINK [67], we estimated heterozygosity as the proportion of heterozygous loci per-individual within each respective variant dataset (joint-called and individual-based calling). For comparison with previously reported results [11] we have also estimated the heterozygosity as the number of polymorphic sites over the total length of the ungapped autosomal scaffolds (Chr. 1–20) (Additional file 2: Table S2).

### Principal components analysis (PCA)

In order to determine where the GNP sample fits within the population structure of baboons, we performed PCA on the 16 *Papio* individuals using PLINK [67] on both joint-called (GATK) and individual-based called (bcftools) variants.

We performed principal component analysis of low coverage data using the smartpca function implemented in EIGENSOFT [68] software 8.0.0. Specifically, we projected sample BB1, GNP and *P. ursinus* 28697 downsampled genotypes onto the PCA inferred from the 16

*Papio* individuals and GNP. The downsampling was done through Picard v.2.26.4 *DownsampleSam*, performing 10 replicates for GNP and *P. ursinus* 28697. The options -P 0.0003 and -R were used.

### TreeMix

To better understand the topology and gene flow between the *Papio* species, we used TreeMix 1.13 [69] to determine the population tree using our GATK called variants and using PLINK to calculate the allele count per population. We ran TreeMix with migration edges that ranged from 0 to 10. For each migration edge, we performed 100 independent runs with bootstrap replicates using 500 SNPs per block. To evaluate the optimal number of migration edges in our population tree we used OptM [70]. We observed that with a single migration edge, more than 99.8% of the variance observed could be explained. Based on  $\Delta m$ , the second most optimal number of migration edges was 5. The first time we observe any migration edge going into our GNP sample is at 4. For each population tree presented here we plotted the TreeMix replicate with the highest global likelihood after a convergence where the top 5 out of every 100 runs had similar likelihoods ( $\pm 50$ ).

### PSMC

We inferred the effective population size for each baboon across time using PSMC. The recommended procedure to run PSMC requires generating diploid sequences per individual with a FQ score using bcftools [66]. Sites with a coverage below 10 and greater than three times the genome-wide coverage per individual were excluded as well as sites called heterozygous with less than three reads of support. PSMC was applied with default settings. We used a mutation rate of  $0.9 \times 10^{-8}$  and a generation time of 11 years as estimated by Rogers et al. [11]. Additionally, we performed the same procedure excluding repetitive regions as indicated by RepeatMasker [71]. The analysis was repeated for the GNP sample when mapped to Panubis1.0.

### D-statistics

To test for differences in allele sharing between populations ( $H_1$ ,  $H_2$ ) relative to different baboon species ( $H_3$ ) we performed *D*-statistics (ABBA-BABA) using POPSTATS [72]. The Z-scores were calculated using the weighted jack-knife procedure where a 5 Mb block weighted by the number of loci was used to estimate the standard error [73, 74].

Given the much lower coverage of BB1, we also performed *D*-statistics (ABBA-BABA) using ANGSD [55].

We used the option `-doAbbababa 1`, an approach that samples a random base at each position of a BAM file to estimate the counts of ABBA and BABA sites between H1, H2, and H3 and is therefore particularly suitable for low coverage samples. Z-scores were calculated based on a jackknife procedure [74] for blocks of data that were 5 Mb in size, performed through the R-script available on the ANGSD github website (<https://github.com/ANGSD/angsd/tree/master/R>). We initially tested if BB1 and GNP shared a similar number of alleles with the two available *P. cynocephalus* samples (16,066 and 16,098; Additional file 2: Table S1). We then independently compared BB1 and GNP to each of the two available *P. ursinus* samples (28697 and 28755, Additional file 2: Table S1) to test for differences in the number of shared alleles with the two *P. cynocephalus* genomes.

### $f_3$ statistics

The presence of admixture events was tested using the  $f_3$ -admixture statistics via qp3Pop as implemented in AdmixTools version 6.0 [40]. We performed the test in the form  $f_3(\text{Pop1, Pop2, Target})$ , where the target was either the GNP genome or one *P. anubis* individual, and Pop1 and Pop2 were a combination of available populations.

### $f_4$ ratio (autosomal and X chromosome data)

We performed the  $f_4$ -statistics using the AdmixTools qpDstat program, with the setting `f4mode: YES` [40].  $f_4$ -statistics measures the amount of shared drift in a defined four population tree topology [40, 73]. Significant deviations of the  $f_4$ -statistics from 0 indicate that the tree topology does not fit the data and suggest higher shared genetic drift between clusters than expected. It is possible to leverage on these properties of the  $f_4$ -statistics to estimate the global proportions of an ancestry in an admixed population, through the  $f_4$ -ratio test [73].

We use the format  $f_4(A, \text{Target}, B_1, \text{gelada})$  and  $f_4(A, B_2, B_1, \text{gelada})$  to infer the numerator and the denominator of the ratio, respectively. In turn, the quotient indicates the proportion of gene flow from  $B_1$  to Target. We maintained the same  $f_4$ -statistics format when focusing on the X chromosome, considering only one individual per population to ensure matching numbers of X chromosomes across samples. Given the inclusion of both males and females for the X chromosome analysis, for the male samples we opted to perform the  $f_4$ -statistic in a hybrid form, thus computing the statistic between individuals of different sex.

### qpGraph

To visualize the relationships and admixture proportions between the studied samples we used the qpGraph package included in AdmixTools [40]. Given a defined topology, where the target populations are set as terminal leaves and internal nodes are pseudo-populations, the tool computes  $f_2$ ,  $f_3$ , and  $f_4$ -statistics to evaluate the amount of genetic drift between the target groups. QpGraph returns the defined topology where branch lengths and admixture proportions are estimated from the computed F-statistics, along with the worst  $f_4$ -statistic Z-score. In this study, we consider the defined topology supported when the worst  $f_4$ -statistics Z-score is  $|Z| < 3$ . We set the option `outgroup: NULL` for all tree-like models, except when we replicated the topology presented in [11], where we set *T. gelada* as the formal outgroup of the graph.

### Local ancestry (ELAI)

Local Ancestry Inferences (LAI) identify the ancestral fragments in an admixed genome and assign them to the putative sources of the admixture event. After statistically phasing the samples using SHAPEIT [75], we performed LAI with ELAI, a tool based on a two layers Hidden Markov Model able to detect haplotype structure between the source populations as well as within each source. This feature allows ELAI to handle delicate scenarios where the ancestry segments are short, and where there is a grade of haplotype variability within the source samples [76]. ELAI is shown to be able to deconvolute target individuals even when only a few source samples are available [77], therefore proving to be an ideal tool for this study. We ran ELAI [76] by averaging 10 independent runs, each of them characterized by 20 steps of inference optimization, as recommended by ELAI developers. We indicated 15 admixture generations (`-mg`), two upper clusters (`-C 2`) and eight lower clusters (`-c 8`). Downstream analyses were performed after removing all SNPs that could not be assigned to one ancestry with a mean local ancestry dosage less than 0.8. We perform the analysis of three different sets of test genomes: (i) the GNP sample, considering as putative source populations *P. cynocephalus* and *P. ursinus*; (ii) the 30,877 sample from *P. anubis*, as a mixture of *P. cynocephalus* and *P. papio*; (iii) all four individuals of the *P. anubis* population, selecting as putative sources *P. cynocephalus* and *P. papio*.

Statistical phasing of samples for LAI and dating ancestry switches was performed with SHAPEIT [75] in two ways. One phasing was performed using a uniform recombination rate of  $1 \times 10^{-8}$  per base pair

per generation [11] while another was done using the recombination map generated from pedigreed captive baboons [78]. The recombination map produced using a block penalty value of 5 was used as those were the rates focused on in the Robinson et al. [78] study. Recombination rates for the Panu\_3.0 assembly from the Panu\_2.0 assembly were generated using the liftOver chain provided by Robinson et al. Furthermore, regions which were found to be erroneously assembled in Panu\_3.0 [53] were masked for both aforementioned analyzes.

### Dating ancestry switches

We harvest ELAI results to calculate the number of generations occurring since the admixture event on the basis of the number of ancestry switches occurring along chromosomes, following the equation reported in Johnson et al. [79]. The equation considers two types of recombination events to count the number of ancestry switches: within the haploid genome (consisting in a genome having a single copy of each chromosome) and between two haplotypes of opposite ancestry. The expected number of recombination events in a haploid genome is computed as  $0.01 \times TL$ , with  $T$  being the number of generations and  $L$  the total genome length. The recombination events occurring between two haplotypes of opposite ancestry are indicated as  $2 \times a(1 - a)$ , where  $a$  stands for the genome-wide ancestry proportion.

Thus, the expected number of ancestry switches in a diploid genome is:

$$B = (2 \times 2 \times 0.01) \times TL \times a(1 - a)$$

We computed the admixture event  $T$  from the observed  $B$  (ancestry switches) after [79]. The equation was applied on haploid genomes (after phasing), therefore resulting in:

$$T = B / (2 \times 0.01) \times L \times a(1 - a)$$

### ALDER and MALDER

We compute weighted LD statistics to date the admixture events by running both ALDER [80] and MALDER [81] on the *P. anubis* dataset, considering different bin sizes (0.0005, 0.0001 and 0.001). All other parameters were set as default. The availability of a single genome prevented the implementation of these analyses on the GNP sample.

### Abbreviations

TMRCAs: Time to most recent common ancestor; mtDNA: Mitochondrial DNA; NGS: Next generation sequencing; CI: Confidence interval; PCA: Principal component analysis; LD: Linkage disequilibrium; LAI: Local ancestry inference; Mya: Million years ago; kya: Thousand years ago; UPGMA: Unweighted pair group

method with arithmetic mean; WGS: Whole genome sequencing; PSMC: Pair-wise Sequentially Markovian Coalescent; MAF: Minor allele frequency.

### Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12862-022-01999-7>.

**Additional file 1.** Distribution, location, and heterozygosity of samples (Figure S1). mtDNA phylogenetics (Figure S2). Y chromosome data & phylogenetics. qpGraph (Figure S3). TreeMix (Figure S4). PSMC (Figures S5–S6). Characterizing *P. cynocephalus* ancestry in *P. anubis* (Figures S7–S8).

**Additional file 2. Table S1.** Samples information, including genomic coverage for Chromosome X and Chromosome 20. **Table S2.** Impact of *P. cynocephalus* ancestry on *P. anubis* 30877 heterozygosity. **Table S3.** TMRCAs and C.I. for mtDNA tree in figure 1. **Table S4.** Number of differences between Y chromosome sequences analysed in figure 1b. **Table S5.** TMRCAs and C.I. for Y chromosome tree in figure 1b. **Table S6.** f3 results. **Table S7.** f4 ratio alpha estimates. **Table S8.** Local Ancestry estimations (Global Proportions obtained with cutoff of 0.8). **Table S9.** ALDER and MALDER results (All combinations of *Papio* species were tested as PopA/PopB; only significant results are reported here).

### Acknowledgements

We would like to thank the following institutions for the support given in the completion of this project: Gorongosa National Park; the Boise Fund Trust (CS, JM, CC); The Gorongosa Restoration Project, The Leverhulme Trust, The National Geographic, John Fell Fund Oxford (SC); St Hugh's College (CC, SC) and the Association for Cancer Research AIRC (AR, FB). MJFS worked under a FCT contract (CEECIND/01937/2017). GM would like to thank Toni Gabaldon and Giulio Pavesi for their support. FB would like to thank the Italian Ministry of Health for their support. The work in GNP is only possible due to the visionary approach of Greg Carr and the dedicated staff from the park, guided by Dr. Mateus Mutemba and Pedro Muagara. We are very grateful to all the 'fiscais' (park rangers), the students, and colleagues across all the institutions who have been extremely supportive of the Paleo-Primate-Project Gorongosa. This work has benefited from the framework of the COMP-HUB Initiative, funded by the 'Departments of Excellence' program of the Italian Ministry for Education, University and Research (MIUR, 2018-2022). The authors would also like to acknowledge the use of the University of Oxford Advanced Research Computing (ARC) facility (<https://doi.org/10.5281/zenodo.22558>) and the HPC (High Performance Computing) facility of the University of Parma.

### Author contributions

FIM, MJFS, SC and CC initially conceived the project; FIM and JM collected the sample; MJFS performed the initial molecular work; CS, LM, GM, MC and GOG performed data analysis under the supervision of IM, AR, LP, and CC. VA, ZA, WA, MB, DB, RB, DRB, PH, TL, MP, LMP, MS, FT were all involved in the work in GNP that enabled the collection of the sample. CS and CC wrote the manuscript with contributions from all co-authors.

### Funding

CS, JM, and CC were supported by The Boise Trust Fund of the University of Oxford. MJFS was supported by FCT (ref: CEECIND/01937/2017).

### Availability of data and material

The genomic data of samples GNP and BB1, supporting the conclusions of this article, are available under the ENA accession study number PRJEB52124. The GitHub repository for this paper can be found here: <https://github.com/santaci/GNPbaboons>.

### Declarations

#### Ethics approval and consent to participate

The PPPG team has been granted research permits to collect biological materials in Gorongosa National Park by the Department of the Scientific Services

of the Park. Permits PNG/DSci/C80/2017 (Granted to C. Capelli, F. Martinez and MJ Ferreira da Silva in 2017), PNG/DSci/C117/2018 (granted to F. Martinez and MJ Ferreira da Silva in 2018) and PNG/DSci/C142/2019 (Granted to F. Hammond, D. Biro and S. Carvalho in 2019). The collection of the single blood sample occurred while the individual was sedated for collaring (PNG/DSci/C147/2019, granted to L.C. Lewis-Bevan, D. Biro and S. Carvalho in 2019). Tissue and blood samples were not exported.

#### Consent for publication

All authors have provided consent for the publication of this paper.

#### Competing interests

The authors declare that they have no competing interests.

#### Author details

<sup>1</sup>Department of Biology, University of Copenhagen, Copenhagen, Denmark. <sup>2</sup>Department of Zoology, University of Oxford, Oxford, UK. <sup>3</sup>Estonian Biocentre, Institute of Genomics, University of Tartu, Tartu, Estonia. <sup>4</sup>Department of Biosciences, University of Milan, Milan, Italy. <sup>5</sup>Department of Chemistry, Life Sciences and Environmental Sustainability, University of Parma, Parma, Italy. <sup>6</sup>Escuela de Antropología, Facultad de Ciencias Sociales, Pontificia Universidad Católica de Chile, Santiago, Chile. <sup>7</sup>School of Anthropology, University of Oxford, Oxford, UK. <sup>8</sup>BIOPOLIS Program in Genomics, Biodiversity and Land Planning, CIBIO, Campus de Vairão, Vairão, Portugal. <sup>9</sup>CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, InBIO Laboratório Associado, Campus de Vairão, Universidade do Porto, Vairão, Portugal. <sup>10</sup>ONE - Organisms and Environment Group, School of Biosciences, Cardiff University, Sir Martin Evans Building, Cardiff, UK. <sup>11</sup>Interdisciplinary Center for Archaeology and Evolution of Human Behavior (ICAEHB), Universidade do Algarve, Faro, Portugal. <sup>12</sup>Department of Archaeology, National Museum, Bloemfontein, South Africa. <sup>13</sup>Evolutionary Studies Institute, University of the Witwatersrand, Johannesburg, South Africa. <sup>14</sup>Gorongosa National Park, Sofala, Mozambique. <sup>15</sup>Center for the Advanced Study of Human Paleobiology, George Washington University, Washington, USA. <sup>16</sup>Emmy Noether Group for Hominin Meat Consumption, Max Planck Institute for Chemistry, Mainz, Germany. <sup>17</sup>AESDA – Associação de Estudos Subterrâneos e Defesa do Ambiente, Lisbon, Portugal. <sup>18</sup>Department of Scientific Services, Gorongosa National Park, Chitengo, Sofala Province, Mozambique. <sup>19</sup>Laboratory of Hematology-Oncology, European Institute of Oncology IRCCS, Milan, Italy. <sup>20</sup>Department of Biology, University of Padua, Padua, Italy.

Received: 10 November 2021 Accepted: 28 March 2022

Published online: 11 April 2022

#### References

- Barrett L, Henzi SP. Baboons. *Curr Biol*. 2008;18:R404–6.
- Jolly CJ. A proper study for mankind: analogies from the Papionin monkeys and their implications for human evolution. *Am J Phys Anthropol*. 2001;33:177–204.
- Fischer J, Higham JP, Alberts SC, Barrett L, Beehner JC, Bergman TJ, et al. Insights into the evolution of social systems and species from baboon studies. *Elife*. 2019;8:1.
- Jolly CJ, Burrell AS, Phillips-Conroy JE, Bergy C, Rogers J. Kinda baboons (*Papio* kindae) and grayfoot chacma baboons (*P. ursinus* griseipes) hybridize in the Kafue river valley, Zambia. *Am J Primatol*. 2011;73:291–303.
- Cracraft J. Species concepts and speciation analysis. *Curr Ornithol*. 1983;18:159–87.
- Mayr E. The growth of biological thought: diversity, evolution, and inheritance. Cambridge: Harvard University Press; 1982.
- Frost SR, Marcus LF, Bookstein FL, Reddy DP, Delson E. Cranial allometry, phylogeography, and systematics of large-bodied papionins (primates: *Cercopithecinae*) inferred from geometric morphometric analysis of landmark data. *Anat Rec A Discov Mol Cell Evol Biol*. 2003;275:1048–72.
- Singleton M, Seitelman BC, Krecioch JR, Frost SR. Cranial sexual dimorphism in the Kinda baboon (*Papio hamadryas kindae*). *Am J Phys Anthropol*. 2017;164:665–78.
- Jolly CJ. Species, subspecies, and baboon systematics. In: Kimbel WH, Martin LB, editors. *Species, species concepts and primate evolution*. Boston: Springer; 1993. p. 67–107.
- Zinner D, Wertheimer J, Liedigk R, Groeneveld LF, Roos C. Baboon phylogeny as inferred from complete mitochondrial genomes. *Am J Phys Anthropol*. 2013;150:133–40.
- Rogers J, Raveendran M, Harris RA, Mailund T, Leppälä K, Athanasiadis G, et al. The comparative genomics and complex population history of baboons. *Sci Adv*. 2019;5:eaau6947.
- Burrell AS. *Phylogenetics and population genetics of central African baboons*. New York: New York University; 2009.
- Zinner D, Groeneveld LF, Keller C, Roos C. Mitochondrial phylogeography of baboons (*Papio* spp.)—indication for introgressive hybridization? *BMC Evol Biol*. 2009;9:1–15.
- Charpentier MJE, Fontaine MC, Chelou E, Renoult JP, Jenkins T, Benoit L, et al. Genetic structure in a dynamic baboon hybrid zone corroborates behavioural observations in a hybrid population. *Mol Ecol*. 2012;21:715–31.
- Wall JD, Schleich SA, Alberts SC, Cox LA, Snyder-Mackler N, Nevoen KA, et al. Genomewide ancestry and divergence patterns from low-coverage sequencing data reveal a complex history of admixture in wild baboons. *Mol Ecol*. 2016;25:3469–83.
- Wango TL, Musiega D, Mundia CN, Altmann J, Alberts SC, Tung J. Climate and land cover analysis suggest no strong ecological barriers to gene flow in a natural baboon hybrid zone. *Int J Primatol*. 2019;40:53–70.
- Chiou KL, Bergy CM, Burrell AS, Disotell TR, Rogers J, Jolly CJ, et al. Genome-wide ancestry and introgression in a Zambian baboon hybrid zone. *Mol Ecol*. 2021;30(8):1907–20.
- Lawson Handley LJ, Hammond RL, Emaesi G, Reber A, Perrin N. Low Y chromosome variation in Saudi-Arabian hamadryas baboons (*Papio hamadryas hamadryas*). *Heredity*. 2006;96:298–303.
- Tosi AJ, Disotell TR, Morales JC, Melnick DJ. Cercopithecine Y-chromosome data provide a test of competing morphological evolutionary hypotheses. *Mol Phylogenet Evol*. 2003;27:510–21.
- Zinner D, Arnold ML, Roos C. Is the new primate genus *rungecebus* a baboon? *PLoS ONE*. 2009;4:e4859.
- Mathieson I, Abascal F, Vinner L, Skoglund P, Pomilla C, Mitchell P, et al. An ancient baboon genome demonstrates long-term population continuity in Southern Africa. *Genome Biol Evol*. 2020;12:407–12.
- Hallast P, Maisano Delser P, Batini C, Zadik D, Rocchi M, Schempp W, et al. Great ape Y Chromosome and mitochondrial DNA phylogenies reflect subspecies structure and patterns of mating and dispersal. *Genome Res*. 2016;26:427–39.
- Martinez FI, Capelli C, Ferreira da Silva MJ, Aldeias V, Alemseged Z, Archer W, et al. A missing piece of the *Papio* puzzle: Gorongosa baboon phenostructure and intrageneric relationships. *J Hum Evol*. 2019;130:1–20.
- Kingdon J. *The kingdon field guide to African mammals*. London: Academic Press; 1997.
- Sithaldeen R, Bishop JM, Ackermann RR. Mitochondrial DNA analysis reveals Plio-Pleistocene diversification within the chacma baboon. *Mol Phylogenet Evol*. 2009;53:1042–8.
- Van Steenberge MW, Vanhove MPM, Chocha Manda A, Larmuseau MHD, Swart BL, Khang'Mate F, et al. Unravelling the evolution of Africa's drainage basins through a widespread freshwater fish, the African sharp-toothed catfish (*Clarias gariepinus*). *J Biogeogr*. 2020;47:1739–54.
- Tinley KL. Montane to mangrove—framework of the Gorongosa ecosystem. Nelspruit: Hamilton-Fynch; 2020.
- McDonough MM, Šumbera R, Mazoch V, Ferguson AW, Phillips CD, Bryja J. Multilocus phylogeography of a widespread savanna-woodland-adapted rodent reveals the influence of Pleistocene geomorphology and climate change in Africa's Zambezi region. *Mol Ecol*. 2015;24:5248–66.
- Bobé R, Martínez FI, Carvalho S. Primate adaptations and evolution in the Southern African Rift Valley. *Evol Anthropol*. 2020;29:94–101.
- Habermann JM, Alberti M, Aldeias V, Alemseged Z, Archer W, Bamford M, et al. Gorongosa by the sea: First Miocene fossil sites from the Urema Rift, central Mozambique, and their coastal paleoenvironmental and paleoecological contexts. *Palaeogeogr Palaeoclimatol Palaeoecol*. 2019;514:723–38.
- Stalmans ME, Massad TJ, Peel MJS, Tarnita CE, Pringle RM. War-induced collapse and asymmetric recovery of large-mammal populations in Gorongosa National Park Mozambique. *PLoS ONE*. 2019;14:e0212864.

32. Li H, Durbin R. Inference of human population history from individual whole-genome sequences. *Nature*. 2011;475:493–6.
33. Schiffels S, Durbin R. Inferring human population size and separation history from multiple genome sequences. *Nat Genet*. 2014;46:919–25.
34. Chikhi L, Rodríguez W, Grusea S, Santos P, Boitard S, Mazet O. The IICR (inverse instantaneous coalescence rate) as a summary of genomic diversity: insights into demographic inference and model choice. *Heredity*. 2017;120:13–24.
35. Rodríguez W, Mazet O, Grusea S, Boitard S, Chikhi L. The IICR and the non-stationary structured coalescent: demographic inference with arbitrary changes in population structure. *Heredity*. 2018. <https://doi.org/10.1101/341750>.
36. Riashna Sithaldeen (University of Cape Town). IUCN Red List of Threatened Species: *Papio ursinus*. IUCN Red List of Threatened Species. 2019.
37. Zinner D, Arnold ML, Roos C. The strange blood: natural hybridization in primates. *Evol Anthropol*. 2011;20:96–103.
38. Sankararaman S, Mallick S, Dannemann M, Prüfer K, Kelso J, Pääbo S, et al. The genomic landscape of Neanderthal ancestry in present-day humans. *Nature*. 2014;507:354–7.
39. Eriksson A, Manica A. Effect of ancient population structure on the degree of polymorphism shared between modern human populations and ancient hominins. *Proc Natl Acad Sci USA*. 2012;109:13956–60.
40. Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, et al. Ancient admixture in human history. *Genetics*. 2012;192:1065–93.
41. Zinner D, Chuma IS, Knaut S, Roos C. Inverted intergeneric introgression between critically endangered kipunjis and yellow baboons in two disjunct populations. *Biol Lett*. 2018;14:729.
42. Walker JA, Jordan VE, Storer JM, Steely CJ, Gonzalez-Quiroga P, et al. Alu insertion polymorphisms shared by *Papio* baboons and *Theropithecus gelada* reveal an intertwined common ancestry. *Mob DNA*. 2019;10:1–13.
43. Jolly CJ, Woolley-Barker T, Beyene S, Disotell TR, Phillips-Conroy JE. Intergeneric hybrid baboons. *Int J Primatol*. 1997;18:597–627.
44. Dunbar RIM, Dunbar P. On hybridization between *Theropithecus gelada* and *Papio anubis* in the wild. *J Hum Evol*. 1974;3:187–92.
45. Kopp GH, Ferreira-da-Silva MJ, Fischer J, Brito JC, Regnaut S, Roos C, et al. The influence of social systems on patterns of mitochondrial DNA variation in baboons. *Int J Primatol*. 2014;35:210–25.
46. da Silva MJF, da Silva MJF, Godinho R, Casanova C, Minhós T, Sá R, et al. Assessing the impact of hunting pressure on population structure of Guinea baboons (*Papio papio*) in Guinea-Bissau. *Conserv Genet*. 2014;15:1339–55.
47. Grün R, Brink JS, Spooner NA, Taylor L, Stringer CB, Franciscus RG, et al. Direct dating of Florisbad hominid. *Nature*. 1996;382:500–1.
48. Dirks PH, Roberts EM, Hilbert-Wolf H, Kramers JD, Hawks J, Dosseto A, et al. The age of and associated sediments in the Rising Star Cave South Africa. *Elife*. 2017;6:24231.
49. Scerri EML, Thomas MG, Manica A, Gunz P, Stock JT, Stringer C, et al. Did our species evolve in subdivided populations across Africa, and why does it matter? *Trends Ecol Evol*. 2018;33:582–94.
50. Stone OML, Herries AIR, Brink JS, Laffan SW. The chacma baboon (*Papio ursinus*) through time: a model of potential core habitat regions during a glacial–interglacial cycle. *Evol Ecol*. 2016;30:755–82.
51. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*. 2009;25:1754–60.
52. Broad Institute. Picard Tools 2.8.1. 2018. <http://broadinstitute.github.io/picard/>. Accessed 11 Dec 2020.
53. Batra SS, Levy-Sakin M, Robinson J, Guillory J, Durinck S, Vilgalys TP, et al. Accurate assembly of the olive baboon (*Papio anubis*) genome using long-read and Hi-C data. *Gigascience*. 2020;9:134.
54. Di Fiore A. A rapid genetic method for sex assignment in non-human primates. *Conserv Genet*. 2005;6:1053–8.
55. Korneliusson TS, Albrechtsen A, Nielsen R. ANGSD: analysis of next generation sequencing data. *BMC Bioinformatics*. 2014;15:356.
56. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 2004;32:1792–7.
57. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol*. 2000;17:540–52.
58. Darriba D, Taboada GL, Doallo R, Posada D. jModelTest 2: more models, new heuristics and parallel computing. *Nat Methods*. 2012;9:772.
59. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol Biol Evol*. 2020;37:1530–4.
60. Bouckaert R, Vaughan TG, Barido-Sottani J, Duchêne S, Fourment M, Gavryushkina A, et al. BEAST 2.5: an advanced software platform for Bayesian evolutionary analysis. *PLoS Comput Biol*. 2019;15:e1006650.
61. Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. Posterior summarization in Bayesian phylogenetics using tracer 1.7. *Syst Biol*. 2018;67:901–4.
62. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol Biol Evol*. 2018;35:1547–9.
63. Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol*. 2018;4:vey016.
64. Perelman P, Johnson WE, Roos C, Seuánez HN, Horvath JE, Moreira MAM, et al. A molecular phylogeny of living primates. *PLoS Genet*. 2011;7:e1001342.
65. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernysky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20:1297–303.
66. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetic parameter estimation from sequencing data. *Bioinformatics*. 2011;27:2987–93.
67. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007;81:559–75.
68. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. *PLoS Genet*. 2006;2:e190.
69. Pickrell JK, Pritchard JK. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet*. 2012;8:e1002967.
70. Fitak RR. optM: an R package to optimize the number of migration edges using threshold models. 2019.
71. Smit, AFA, Hubble, R & Green, P. RepeatMasker. 1996. <http://www.repeatmasker.org>. Accessed 14 Jan 2021.
72. Skoglund P, Mallick S, Bortolini MC, Chennagiri N, Hünemeier T, Petzl-Erler ML, et al. Genetic evidence for two founding populations of the Americas. *Nature*. 2015;525:104–8.
73. Reich D, Thangaraj K, Patterson N, Price AL, Singh L. Reconstructing Indian population history. *Nature*. 2009;461:489–94.
74. Busing FMTA, Meijer E, Leeden RVD. Delete-m Jackknife for unequal m. *Stat Comput*. 1999;9:3–8.
75. Delaneau O, Zagury J-F, Marchini J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods*. 2013;10:5–6.
76. Guan Y. Detecting structure of haplotypes and local ancestry. *Genetics*. 2014;196:625–42.
77. Molinaro L, Marnetto D, Mondal M, Ongaro L, Yelmen B, Lawson DJ, et al. A chromosome-painting-based pipeline to infer local ancestry under limited source availability. *Genome Biol Evol*. 2021;13:1–8.
78. Robinson JA, Belsare S, Birnbaum S, Newman DE, Chan J, Glenn JP, et al. Analysis of 100 high-coverage genomes from a pedigreed captive baboon colony. *Genome Res*. 2019;29:848–56.
79. Johnson NA, Coram MA, Shriver MD, Romieu I, Barsh GS, London SJ, et al. Ancestral components of admixed genomes in a Mexican cohort. *PLoS Genet*. 2011;7:e1002410.
80. Loh P-R, Lipson M, Patterson N, Moorjani P, Pickrell JK, Reich D, et al. Inferring admixture histories of human populations using linkage disequilibrium. *Genetics*. 2013;193:1233–54.
81. Pickrell JK, Patterson N, Loh P-R, Lipson M, Berger B, Stoneking M, et al. Ancient west Eurasian ancestry in southern and eastern Africa. *Proc Natl Acad Sci USA*. 2014;111:2632–7.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.