

## Supplementary Materials for

### The comparative genomics and complex population history of *Papio* baboons

Jeffrey Rogers\*, Muthuswamy Raveendran, R. Alan Harris, Thomas Mailund, Kalle Leppälä, Georgios Athanasiadis, Mikkel Heide Schierup, Jade Cheng, Kasper Munch, Jerilyn A. Walker, Miriam K. Konkel, Vallmer Jordan, Cody J. Steely, Thomas O. Beckstrom, Christina Bergey, Andrew Burrell, Dominik Schrempf, Angela Noll, Maximillian Kothe, Gisela H. Kopp, Yue Liu, Shwetha Murali, Konstantinos Billis, Fergal J. Martin, Matthieu Muffato, Laura Cox, James Else, Todd Disotell, Donna M. Muzny, Jane Phillips-Conroy, Bronwen Aken, Evan E. Eichler, Tomas Marques-Bonet, Carolin Kosiol, Mark A. Batzer, Matthew W. Hahn, Jenny Tung, Dietmar Zinner, Christian Roos, Clifford J. Jolly, Richard A. Gibbs, Kim C. Worley, Baboon Genome Analysis Consortium

\*Corresponding author. Email: jr13@bcm.edu

Published 30 January 2019, *Sci. Adv.* **5**, eaau6947 (2019)

DOI: 10.1126/sciadv.aau6947

#### The PDF file includes:

Section S1. Rationale for baboon taxonomy and nomenclature  
Section S2. Rationale for mutation rate used in PSMC analyses  
Section S3. Sequencing and assembly of olive baboon genome  
Section S4. Annotation and gene content of the baboon genome  
Section S5. Identification of SNVs and small indels within baboon species  
Section S6. Validation of species identity within the diversity panel  
Section S7. Lineage-specific *Alu* insertion in OWMs and hominoids  
Section S8. Alternative methods for constructing phylogenetic trees  
Section S9. Identification of admixture through asymmetric allele sharing  
Section S10. Polymorphic *AluY* insertions across *Papio* species  
Section S11. Bayesian concordance analyses of gene trees devoid of coding sequences  
Section S12. CoalHMMs of admixture trees and events  
Section S13. Locus-specific phylogenetic trees for chromosomal segments containing annotated genes  
Fig. S1. Panu3.0 genome assembly process.  
Fig. S2. Workflow used in variant calling pipeline.  
Fig. S3. Details regarding SNV calls.  
Fig. S4. Maximum likelihood and Bayesian phylogenetic trees based on SNV data.  
Fig. S5. Test of phylogeny reconstruction using PoMo.  
Fig. S6. Identification of admixture using *f*-statistics.  
Fig. S7. Evidence for admixture from haplotyping sharing.  
Fig. S8. A cladogram of *Papio* individuals from the diversity panel.  
Fig. S9. Bayesian concordance analysis.  
Fig. S10. Bootstrap analysis of timing of divergence events.

Fig. S11. Confidence intervals for baboon admixture proportions.  
Fig. S12. Results for simulated admixture analysis.  
Fig. S13. Results for correction factor adjustment of admixture history.  
Fig. S14. Model used to estimate specific divergence and admixture history.  
Fig. S15. Unbiased estimates dating divergences and admixture events.  
Fig. S16. Phylogeny representing cluster 1 genic regions.  
Fig. S17. Phylogeny representing cluster 2 genic regions.  
Fig. S18. Phylogeny representing cluster 3 genic regions.  
Table S1. Assembly statistics.  
Table S2. Annotation of baboon genome assemblies.  
Table S3. DNA samples used for diversity analysis.  
Table S4. SNV variation among 15 *Papio* baboons and a gelada.  
Table S5. Full-length *AluY* insertions and lineage-specific insertions in primate genomes.  
Table S6. Effect of admixture on branch lengths measured in substitutions per site.  
Table S7. Bayes factors comparing alternate phylogenies.  
Table S8. Divergence time estimates across triplets.  
Table S9. Admixture proportion estimates across triplets.  
Legend for table S10  
References (61–78)

**Other Supplementary Material for this manuscript includes the following:**

(available at [advances.sciencemag.org/cgi/content/full/5/1/eaau6947/DC1](https://advances.sciencemag.org/cgi/content/full/5/1/eaau6947/DC1))

Table S10 (Microsoft Excel format). GO terms associated with genes falling in clusters 1 to 3 of genic regions.

# Supplementary Text

## Section S1. Rationale for baboon taxonomy and nomenclature

The taxonomy and classification of *Papio* baboons has been the subject of extensive debate and discussion for considerable time (for reviews see (9, 10, 12, 14)). Disagreements arise due to the universally shared judgment that the morphological, behavioral and genetic diversity within the genus *Papio* (excluding the gelada, genus *Theropithecus*) does not fit smoothly into any of the concepts, frameworks or theories commonly used to delineate species, most importantly the Biological Species Concept and the Phylogenetic Species Concept (4, 61-64). In keeping with current consensus among most baboon researchers (11, 12, 14, 18, 65-67), we recognize the six phenotypically distinct and readily diagnosable populations of baboons as distinct species (Main Text, Fig. 1).

As described in the main text, at least six readily distinguishable “morphotypes” of baboons occur in non-overlapping geographic distributions. These populations differ in pelage, body size, social behavior, social systems and other significant phenotypic characters widely used to define primate species. Our taxonomic approach is consistent with the Phylogenetic Species Concept (61, 64) that considers a species to be an “irreducible (basal) cluster of organisms, diagnosably distinct from other such clusters, and within which there is a parental pattern of ancestry and descent” (61).

We prefer this taxonomy over the alternative that places all baboons (excluding gelada) in a single species because external phenotypic variation is modest within the diagnosable “morphotypes” that we recognize here as species (although present in some of those species more than others). Furthermore, the transitions from one

species to another at zones of contact are generally sharp and quite restricted geographically. In addition, the degree of differentiation in their social system or body size is higher than is generally observed in polytypic species of primates (10, 11, 68).

Since the 1960s, various taxonomic schemes have been proposed (reviewed in (10, 14)), and the only conclusions that are universally accepted are that 1) the complexity of phenotypic variation and population genetic structure (including potential or actual gene flow among phenotypically distinct populations) observed among baboons defies simple Linnaean binomial nomenclature, and 2) there is discordance between the relationships among populations based on external phenotype and equivalent relationships based on genetic (mtDNA or nuclear) similarities (12, 14). Thus, we find no simple model to be entirely satisfactory in this interesting but daunting taxonomic situation. Nevertheless, some system of nomenclature that facilitates clear and unambiguous discussion about variation among the individual animals and populations under study is required. Taxonomic recognition at the level of species is thus a suitable framework for nomenclature and effective communication regarding these diverse populations.

## **Section S2. Rationale for mutation rate used in PSMC analyses**

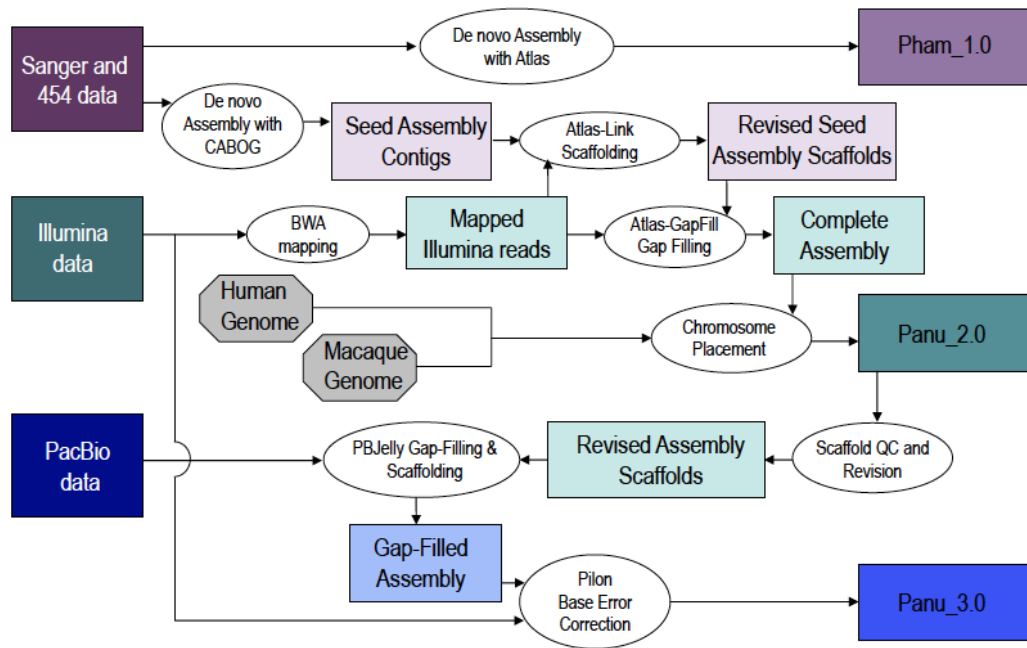
There is little information available concerning direct empirical measurements of nucleotide mutation rates in baboons or other Old World monkeys (69). Inferences can be made based on evolutionary sequence differences among primate species, but these estimates differ among studies. Estimates for the much more extensively analyzed human mutation rate using evolutionary comparisons differ significantly from estimates based on other approaches (69). Therefore, in order to obtain an estimate of

the mutation rate in *Papio*, we began with the widely cited estimate for humans of  $1.5 \times 10^{-8}$  per basepair per generation (69). This is a reasonable consensus based on a variety of studies and methods. Assuming a human generation time of 25 years, this estimate translates to a per year mutation rate of  $0.6 \times 10^{-9}$  per basepair. Prior analyses suggest that the mutation rate per year is about one-third higher in Old World monkeys than in humans (69, 70). Taking the baboon generation time (age at which the average female gives birth to her median surviving offspring) as 11 years, we obtain  $\sim 0.9 \times 10^{-8}$  per basepair per generation as a working estimate. This is clearly approximate and subject to change as improved data become available for baboons or closely related species.

We also used a second approach to estimating the baboon mutation rate. The results from the PoMo analysis of baboon phylogeny (see above) were used to calculate branch lengths for *Papio* lineages, scaled in nucleotide changes. We assumed a generation time of 11 years, and divergence of the northern clade (*P. anubis*, *P. hamadryas* and *P. papio*) from the southern clade (*P. cynocephalus*, *P. ursinus* and *P. kindae*) at 2.0 million years ago (12, 71). This method also generates an estimated mutation rate of  $0.90 \times 10^{-8}$  per basepair per generation.

### **Section S3. Sequencing and assembly of olive baboon genome**

The data types are shown on the left of Suppl. Fig. S1, and the assembly versions on the right, with processing methods and intermediate results in the middle. Suppl. Table S2 lists the assembly statistics for the later assembly versions.



**Fig. S1. Panu3.0 genome assembly process.**

**Table S1. Assembly statistics.**

	Panu_2.0	Panu_3.0
Total sequence length (bp)	2,948,397,226	2,959,356,508
Total assembly gap length (bp)	55,126,439	22,361,627
Number of scaffolds	63,250	63,234
Percentage of assembly in scaffolded contigs	96.80%	96.80%
Scaffold N50	139,646,187	140,346,614
Scaffold L50 count	9	9
Number of contigs	198,931	118,251
Contig N50	40,262	149,817
Contig L50 count	20,291	5,408
Total number of chromosomes, plasmids and mtDNA	22	22

## Section S4. Annotation and gene content of the baboon genome

**Table S2. Annotation of baboon genome assemblies.**

	NCBI		Ensembl	
	Panu_2.0	Panu_3.0	Panu_2.0	Panu_3.0
<u>Protein coding genes</u>	21,567	21,300	19,210	21,647
<u>Non-coding loci</u>	11,984	8,433	9,272	6,699

Annotation of protein-coding and non-coding genes in the olive baboon genome assemblies (Suppl. Table S2) was performed separately by NCBI (72), and Ensembl (73, 74). The number of protein-coding genes increased in the Ensembl annotations from Panu\_2.0 to Panu\_3.0, but this is due largely to changes in the analytical pipeline. The number of protein-coding genes identified by NCBI decreased from Panu\_2.0 to Panu\_3.0.

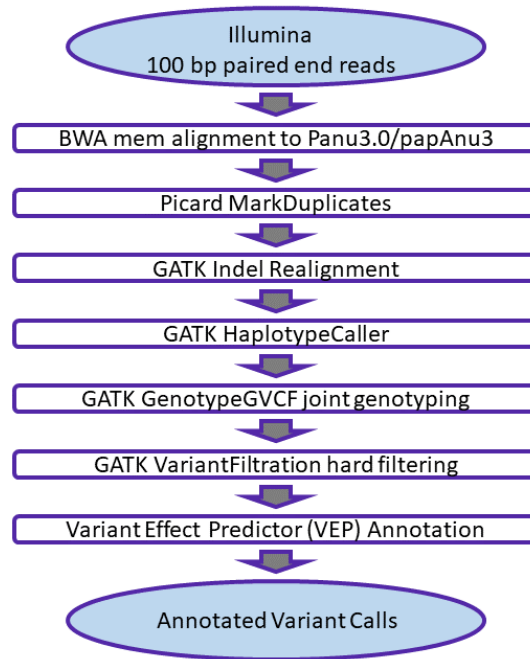
## Section S5. Identification of SNVs and small indels within baboon species

To reconstruct the history of genetic differentiation in this genus, we analyzed whole genome sequences from 15 individuals representing each of the six species as well as a gelada (*Theropithecus gelada*), a member of a closely related genus that serves as outgroup (Suppl. Table S3).

**Table S3. DNA samples used for diversity analysis.**

Species	Sample size	Depth of coverage for SNV calls, per individual	Provenance	Source
<i>Papio anubis</i>	4	22.5, 21.1, 21.6, 24.9	Wild (Aberdare region, Kenya) = 2 Captive (SNPRC) = 2	Kenya= J. Else SNPRC= J. Pecotte, K. Rice
<i>P. cynocephalus</i>	2	24.4, 20.3	Wild (Mikumi Nat. Park, Tanzania) = 2	J. Rogers
<i>P. papio</i>	2	17.5, 26.7	Captive (Southwest NPRC) = 2	J. Pecotte, K. Rice
<i>P. hamadryas</i>	2	19.8, 21.1	Wild (Awash Nat. Park, Ethiopia) = 2	J. Phillips-Conroy, C.J. Jolly
<i>P. kindae</i>	3	28.2, 23.9, 28.2	Wild (Kafue Nat. Park, Zambia) = 3	J. Phillips-Conroy, C.J. Jolly, J. Rogers
<i>P. ursinus</i>	2	21.4, 21.2	Captive (Southwest NPRC) = 2	J. Pecotte, K. Rice
<i>Theropithecus gelada</i>	1	40.5	Captive (Bronx Zoo, NY)	C.J. Jolly, Bronx Zoo



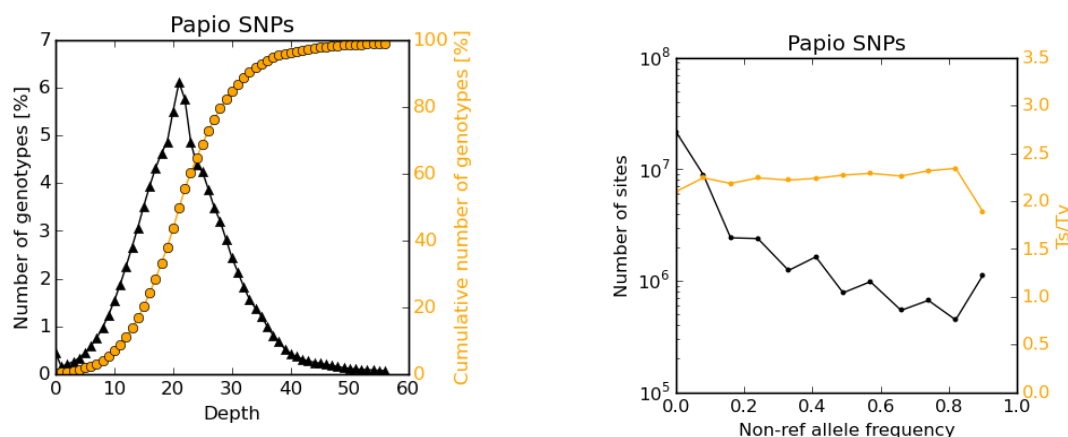


**Fig. S2. Workflow used in variant calling pipeline.**

We identified a total of 54,639,634 SNVs across this diversity panel, with 42,491,538 segregating within *Papio* (Suppl. Table S4, Suppl. Figs. S3 and S4). The average number of SNV calls (heterozygous plus homozygous non-reference) per *Papio* individual differs among species (low:  $7.44 \times 10^6$  for *P. papio*; high:  $13.69 \times 10^6$  for *P. kindae*). The SNV variants and their locations are available for visualization through the UCSC track hub accessible from <https://www.hgsc.bcm.edu/non-human-primates/baboon-genome-project>. The Ensembl baboon gene annotations for these ~42.5 million SNVs (based on Ensembl Variant Effect Predictor, VEP) generate a list of putative functional variants that includes 2577 polymorphic protein translation stop site gains relative to the reference annotation, 1857 splice donor or acceptor variants, and 133,876 missense mutations.

We next identified genome-wide insertions and deletions (indels) ranging from 1 to 60 bp across the baboon diversity panel of 15 baboons and one gelada. A total of

9,075,448 indels were identified including 7,528,451 variable within or among the *Papio* species. *Papio* indels consisted of 3,382,042 insertions, 3,317,420 deletions, and 828,989 complex sequence alterations. These indels (see the UCSC track hub at the same URL cited for SNVs above) represent 15,960,352 bp of inserted sequence and 16,052,079 bp of deleted sequence relative to the reference. We note that the total number of bases affected by these indel variants is substantially less than the total number of SNV sites identified across the same individuals.



**Fig. S3. Details regarding SNV calls. Left Panel:** Read depth across SNV genotypes for the baboon diversity panel (n=15 *Papio*). **Right Panel:** Transition-transversion ratio, non-reference allele frequency and SNV calls across the *Papio* baboon diversity panel

**Table S4. SNV variation among 15 *Papio* baboons and a gelada.**

Species	SNV Sites	Average SNV sites per Indiv	Aver. Heterozygosity	Samples	Captive or wild
<b>Northern Clade</b>					
<i>P. anubis</i> *	15,054,590	7,792,427	0.00168	4	Both
<i>P. papio</i>	8,350,713	7,442,986	0.00055	2	Captive
<i>P. hamadryas</i>	11,869,745	9,436,507	0.00149	2	Wild
<b>Southern Clade</b>					
<i>P. cynocephalus</i>	17,037,057	13,077,450	0.00230	2	Wild
<i>P. kindae</i>	21,332,911	13,690,550	0.00263	3	Wild
<i>P. ursinus</i>	12,347,275	11,052,692	0.00090	2	Captive
<b>Gelada</b>					
<i>T. gelada</i>	20,904,653	20,904,653	0.00071	1	Captive

\*Does not include reference animal

## Section S6. Validation of species identity within the diversity panel

In order to confirm the species identity of each diversity sample, we examined mtDNA sequence variation across these samples, and compared those mtDNA sequences to the results of prior analyses of mtDNA phylogeny in baboons (12, 75) using simple neighbor-joining trees. All specimens within the diversity panel fall into the expected (based on geographic origin) clades: 28697 and 28755 cluster with *P. ursinus* South; 34449, 34472 and 34474 cluster with *P. kindae*; 28547 and 30388 cluster with *P. papio*; 16066, 16098, 97074, 97124, 30877, 30977, L142 and LIV5 fall into the mixed clade containing *P. hamadryas*, *P. anubis* East and *P. cynocephalus* North.

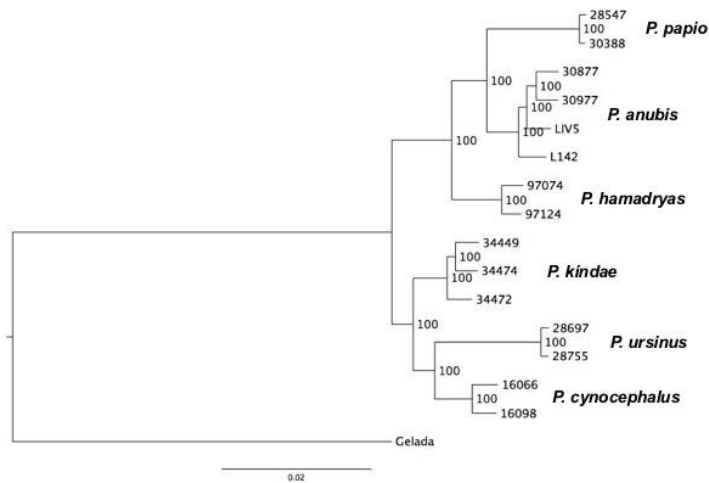
## Section S7. Lineage-specific *Alu* insertion in OWMs and hominoids

**Table S5. Full-length *AluY* insertions and lineage-specific insertions in primate genomes.**

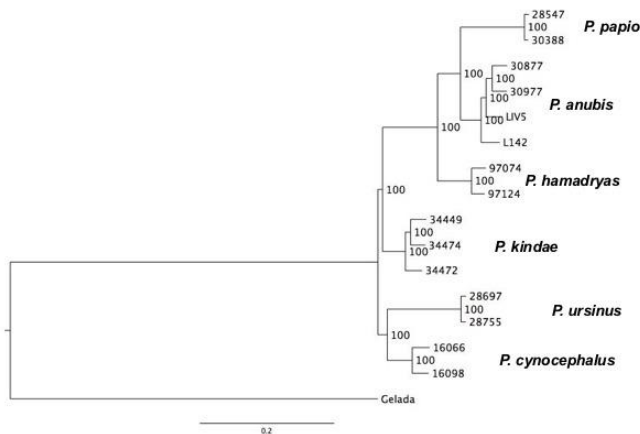
		Baboon	Rhesus macaque	Afr. Green monkey	Orangutan	Chimpanzee	Human
		Panu_3.0	Mmul8.0.1	chlSab2	<i>P. abelii</i> 2.0.2	Pan_tro 3.0	GRCh38
<b>A</b>	Total # full length <i>AluY</i>	192,889	199,894	123,121	97,140	108,931	112,768
<b>B</b>	Lineage specific subset from A	58,273	57,382	22,617	250	2,897	8,062
<b>C</b>	Divergence estimate MYA	8 (vs macaque)	8 (vs baboon)	12.5 (vs baboon)	13 (vs human)	5 (vs human)	5 (vs chimp)
<b>D</b>	Rate per MY (B/C)	7284	7173	1809	19	579	1612

## Section S8. Alternative methods for constructing phylogenetic trees

### Maximum likelihood and Bayesian trees derived from concatenated SNVs

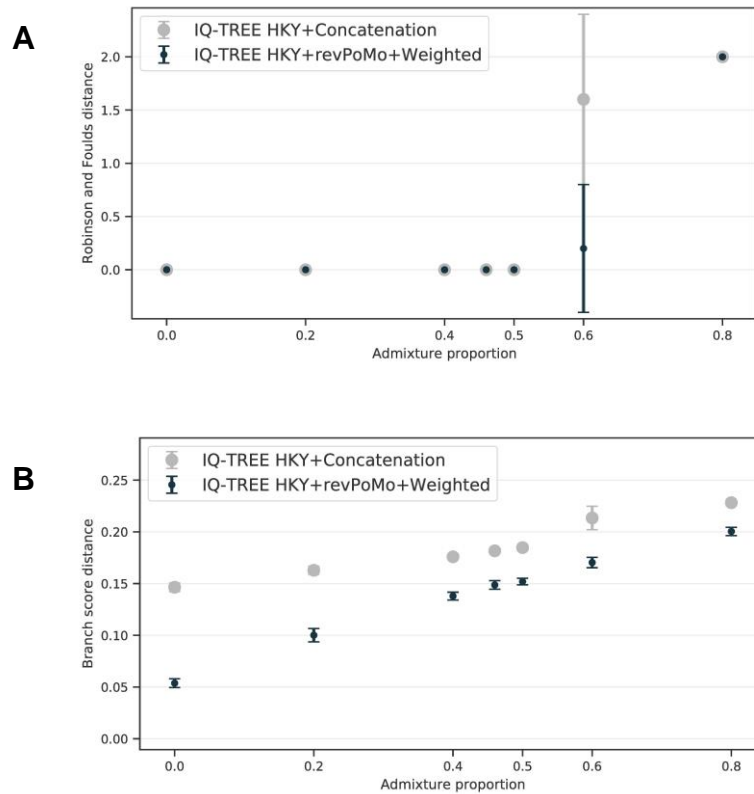


Suppl. Fig. S4:  
Panel A



Suppl. Fig. S4:  
Panel B

**Fig. S4. Maximum likelihood and Bayesian phylogenetic trees based on SNV data.** **Panel A:** Maximum likelihood (ML) tree (IQ-TREE reconstruction). **Panel B:** Bayesian tree (MrBayes reconstruction). Reconstructions are based on the best-fit model TVM+G, using SNVs identified mapping against Panu\_2.0. All nodes gained 100% support values.



**Fig. S5. Test of phylogeny reconstruction using PoMo. Panel A:** Ability of the PoMo model to correctly reconstruct phylogenetic relationships across different levels of admixture. We compared the true and inferred phylogeny for different admixture proportions: HKY model and PoMo. The Robinson-Foulds distance (76) between two phylogenies roughly corresponds to the number of topological differences between them. Error bars are standard deviations for ten replicates and only visible at admixture proportion of 0.60. **Panel B:** The tree error measured in branch score distance (77) that captures the relative difference in branch lengths between two phylogenies. For higher admixture, the error of the inferred branch lengths increases.

## Simulations of admixture effect on branch length

We also studied the effect of admixture on individual branch lengths measured in substitutions per site (Suppl. Table S6). The distance between *P. ursinus* and the root is roughly constant. On the other hand, when we increase the admixture proportion, the branch length of *P. kindae* decreases. In addition, the distance between *P. kindae* and the root decreases. This means that the divergence of *P. kindae* from *P. ursinus* is shifted towards the root. We see a ~50% relative reduction of *P. kindae* with respect to *P. ursinus*. Interestingly, the direction of the effect and its size is similar to the real data. Note that relative differences between branch length of *P. ursinus* and *P. kindae* are not due to changes in population size, mutation rates or generation length in *P. ursinus* but admixture in *P. kindae*.

**Table S6. Effect of admixture on branch lengths measured in substitutions per site.** Adm. prop.: admixture proportion;  $l_{tot}$ : total branch length of the tree;  $l_{kin}$ : distance from *P. ursinus* / *P. kindae* split to *P. kindae*;  $l_{root\_kin}$ : distance from the root of the tree to *P. kindae*;  $l_{urs}$ : distance from *P. ursinus* / *P. kindae* split to *P. ursinus*;  $l_{root\_urs}$ : distance from the root of the tree to *P. kindae*;  $\%l_{kin\_tot}$ :  $l_{kin}/l_{tot}$ ;  $\%l_{kin\_zero}$ : the percentage of  $l_{kin}$  to the value of  $l_{kin}$  without admixture;  $\%l_{kin\_urs}$ :  $l_{kin}/l_{urs}$ . The tree was rooted at its midpoint.

Admixture Proportion	0	20	40	46	50
Total Branch Length of Tree ( $l_{tot}$ )	0.00484	0.00483	0.004897	0.004882	0.00491
Distance from <i>P. ursinus</i> / <i>P. kindae</i> split to <i>P. kindae</i> ( $l_{kin}$ )	0.00042	0.00034	0.000341	0.000356	0.000364
Distance from root of tree to <i>P. kindae</i>	0.001093	0.000819	0.000661	0.000634	0.000631
Distance from <i>P. ursinus</i> / <i>P. kindae</i> split to <i>P. ursinus</i> ( $l_{urs}$ )	0.000437	0.000607	0.000761	0.000789	0.000801
Distance from root of tree to <i>P. ursinus</i>	0.00111	0.001085	0.001081	0.001069	0.001068
Percentage $l_{kin}/l_{tot}$	0.086778	0.070451	0.069701	0.072825	0.074176
Percentage $l_{kin}$ to value of $l_{kin}$ without admixture	0.272778	0.221038	0.221704	0.230939	0.236587
Percentage $l_{kin}/l_{urs}$	0.961535	0.560606	0.448342	0.450071	0.454957

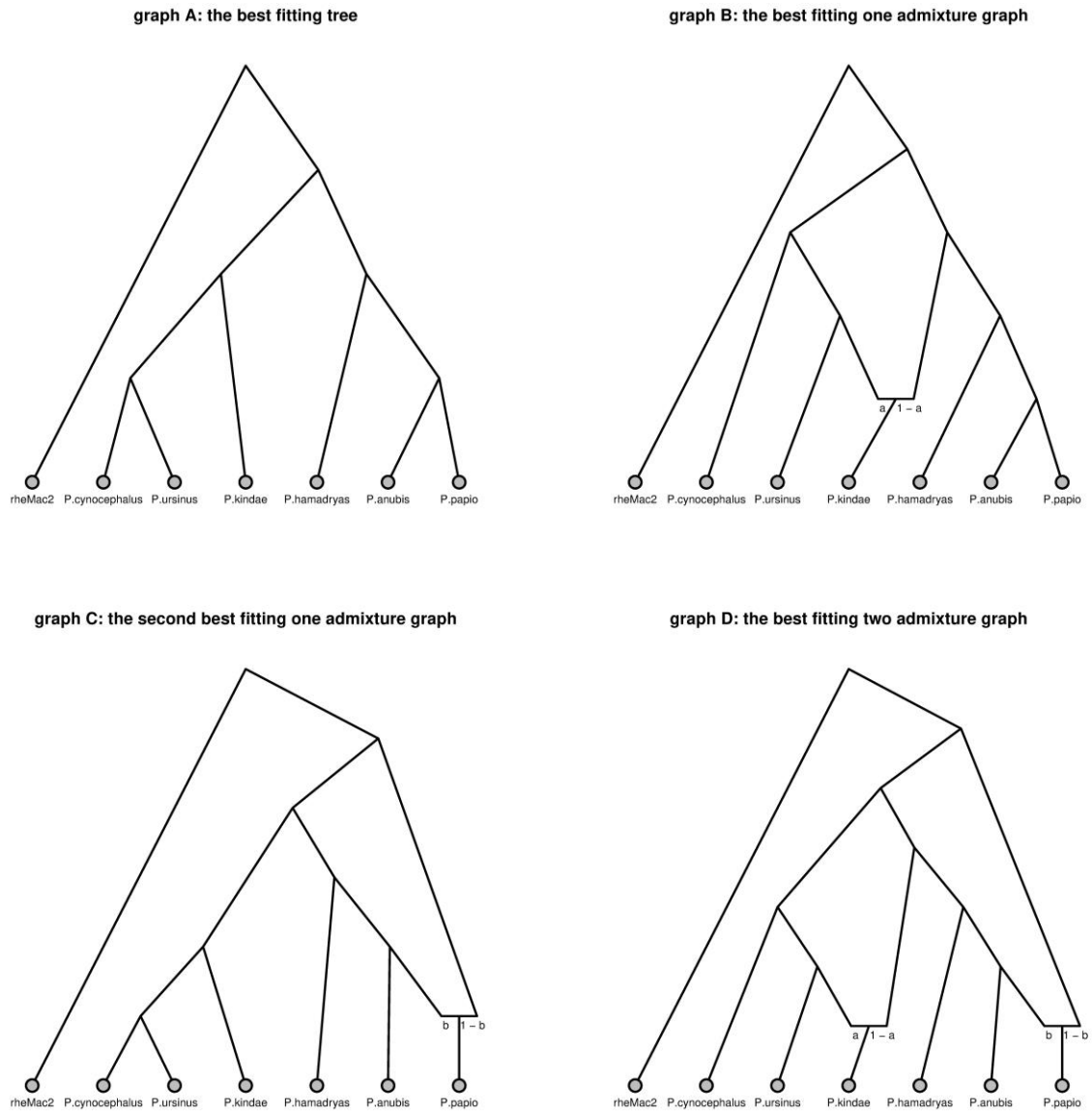
### Reconstructing admixture events using *f*-statistics

We computed the log Bayes factors comparing Graphs A, B and C in Suppl. Fig. S6, and obtained the results presented in Suppl. Table S7.

**Table S7. Bayes factors comparing alternate phylogenies.**

Comparison: Figure S6	$\log_{10}$ Bayes factor
Graph B / Graph A	8869.660
Graph C / Graph A	3937.721
Graph B / Graph C	4931.934





**Fig. S6. Identification of admixture using  $f$ -statistics.**

The Bayes factors are shown in  $\log_{10}$  so they can be read as the number of orders of magnitude by which a graph is more likely than another to which it is compared. As is evident from these results, both admixture graphs are much better fit to the data than the simple dichotomous tree (Graph A), and Graph B fits many orders of magnitude

better than Graph C. From the posterior samples from the Markov chain Monte Carlo approximation of Graph B we obtain the following estimates for the admixture proportions:

2.5%	25%	50%	75%	97.5%
0.5331256	0.5349994	0.5360654	0.5370269	0.5395091

This is slightly less than what we would obtain with the  $f_4$ -ratio estimator from Patterson *et al.* (52) where  $a = f_4(\text{rheMac2}, P.\text{cynocephalus}; P.\text{papio}, P.\text{kindae}) / f_4(\text{rheMac2}, P.\text{cynocephalus}; P.\text{papio}, P.\text{ursinus})$  is 59%.

For the second best (Graph C), the admixture proportions were estimated as

2.5%	25%	50%	75%	97.5%
0.8965939	0.8984543	0.8993151	0.9002740	0.9020158

**Fitting two admixture events.** To explore the space of graphs with two admixture events, we froze the basic topology of the graph at one of the two graphs above with a single admixture event, and exhaustively explored all possible ways of extending these to a graph with two admixture events. The search algorithm goes through each edge in the graph, splits it in two in an admixture event and tries each pair of remaining edges as the donor populations (discarding the graph if it's no longer acyclic). Thus, we did not require that one of the edges leading to the new admixture event was present in the basic topology. The best fitting graph found with this procedure combines the two admixture events found in the one admixture graphs. The graph is presented as Graph D in Suppl. Figure S6.

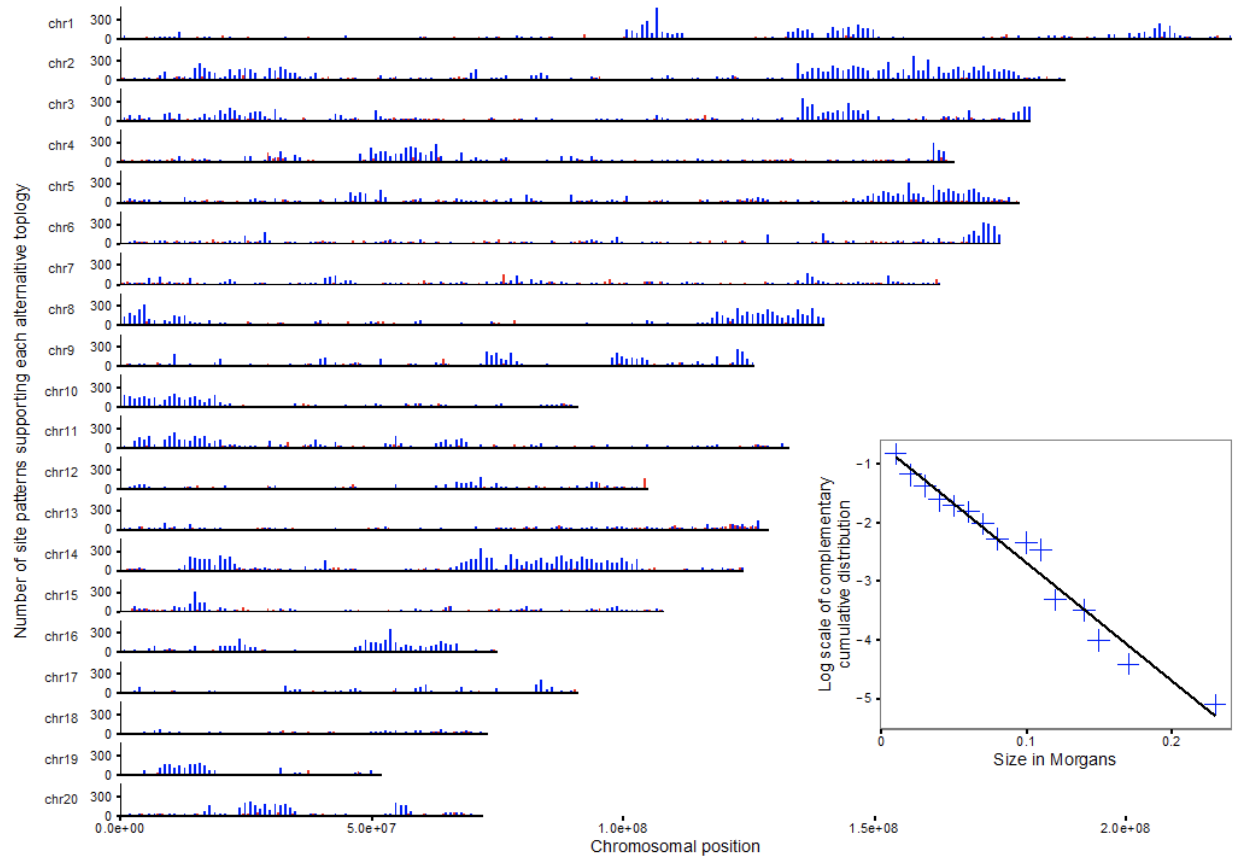
The  $\log_{10}$  Bayes factor for the two-admixture graph (D) over the best one-admixture graph (B) is 3568.586, giving very strong support to the second admixture event. The admixture proportions estimated in the two admixture graph (D) are very similar to the same proportions estimated from the single admixture graphs:

	2.5%	25%	50%	75%	97.5%
a	0.5178859	0.5201515	0.5212253	0.5224752	0.524808
b	0.8978030	0.8993201	0.8999532	0.9007638	0.902350

## Section S9. Identification of admixture through asymmetric allele sharing

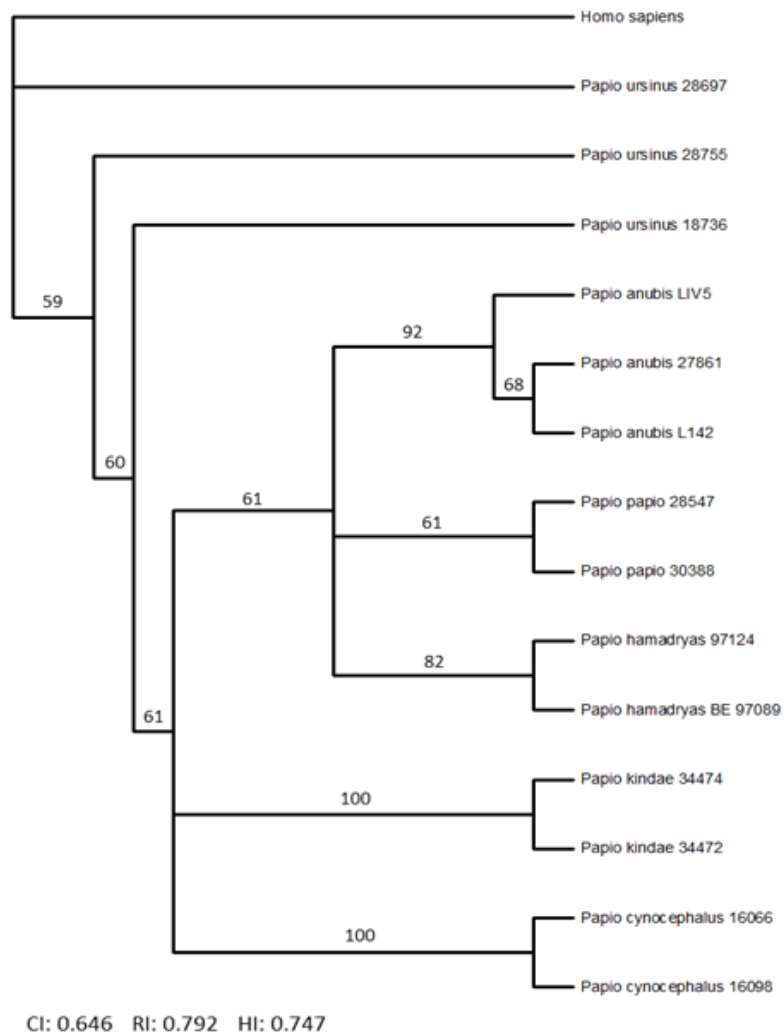
We identified genomic regions exhibiting admixture in trios that include one *P. anubis* individual from the Aberdares region (30877), a *P. cynocephalus* individual, and an individual from any other species. These regions locate to the same consecutive genomic regions irrespective of which *P. cynocephalus* individual is used in the trio, but are not found in trios using other *P. anubis* individuals. This suggests that the tracts result from recent admixture from *P. cynocephalus* into *P. anubis*. We identify 165 tracts totaling 546 megabases (Suppl. Figure S7). If we assume that the admixture is the result of a single admixture event, the distribution of tract lengths is well approximated by  $y = Ae^{-(n-1)d}$  where A is a scaling factor, d is the genetic distance (assuming 1 cM/Mb) and n is the number of generations since the admixture event. We note however, that there is now evidence that recombination in Old World monkeys may occur at substantially less than 1 cM/Mb (78), which would push the date of admixture back older in time. We fit a line to the log of the complementary cumulative distribution of tract lengths and since  $\ln(y) = -(n-1)d + \ln(A)$ , the slope of the fitted line is n-1, or one less than the number of generations. We estimate the admixture event to have

occurred 21.1 generations ago, with a confidence interval of [23.2, 19.1]. The other Aberdare region olive baboon (30977) shows a few genomic regions with excess of ((anu,cyn),pap) sites, but these are much shorter and too few to allow reliable estimation of an admixture time. Any events affecting this individual are expected to be much older, and therefore support an extended period of admixture between *P. anubis* and *P. cynocephalus*. Assuming a single admixture pulse, the proportion of the genome that is introgressed from *P. cynocephalus* represents the proportion of admixing *P. cynocephalus* in the receiving *P. anubis* population. This proportion is 21%.



**Fig. S7. Evidence for admixture from haplotyping sharing.** The number of informative nucleotide sites exhibiting a phylogeny in which cynocephalus is more similar to anubis than either is to papio ((anu,cyn),pap) are shown in blue, while the number of sites exhibiting the alternative phylogeny ((pap,cyn),anu) are shown in red. All counts are based on analysis of 1Mb windows across the autosomal chromosomes. Admixture tracts from cynocephalus into anubis are visible as regions with very strong excess of blue or ((anu,cyn),pap) sites. The predominance of blue segments over red segments indicates that ILS is unlikely to be the cause, and that admixture between cynocephalus and anubis, after anubis diverged from papio, is the best explanation. The insert figure shows the fit of the empirical and theoretical distributions of tract lengths.

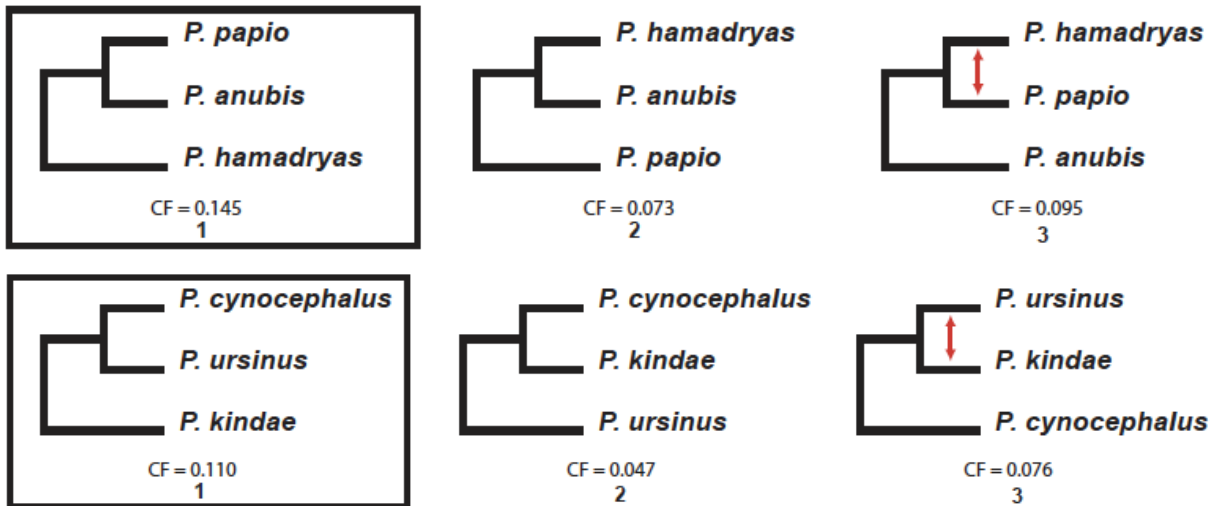
## Section S10. Polymorphic *AluY* insertions across *Papio* species



**Fig. S8. A cladogram of *Papio* individuals from the diversity panel. 494**

polymorphic *Alu* elements were used to create a Dollo parsimony tree over 10,000 bootstrap replicates. The numbers above each branch show the percent of bootstrap replicates supporting that branch.

## Section S11. Bayesian concordance analyses of gene trees devoid of coding sequences



**Fig. S9. Bayesian concordance analysis.** Bayesian concordance factors (CFs), based on genomic regions that contain no coding sequences or other features that might be subject to selection, for each possible topology within the northern clade (top row) and southern clade (bottom row). The primary concordance topologies are boxed. The CFs indicated here are calculated by dividing the number of loci exhibiting the illustrated topology by the total number of loci analyzed. The scaled CFs in main text Fig. 3 are calculated by dividing the same number of loci exhibiting the illustrated topology by the number of loci that produced clear evidence for one of the three topologies illustrated in each row of this figure. Note that the asymmetries among CFs of the less common alternative topologies (the unboxed topologies) is evidence that ILS alone is not the cause of complexity in locus topologies and CFs.

## Section S12. CoalHMMs of admixture trees and events

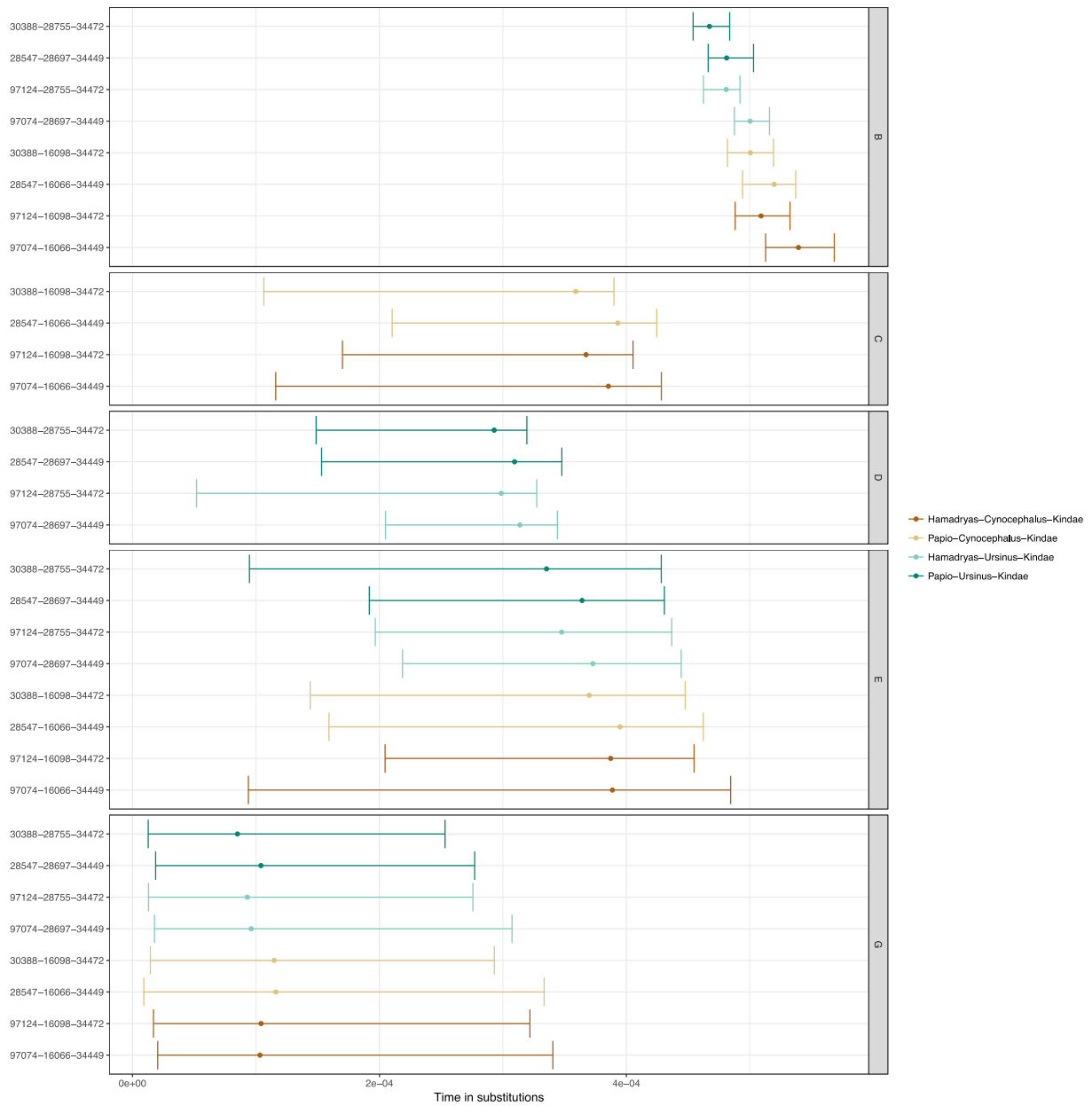
### Initial Estimates of Divergence and Admixture Dates

We consider the admixture graph inferred using  $f$ -statistics (see Fig. 4A, main text). By selecting triplets of samples we can extract sub-graphs with the topology the CoalHMM assumes and use these to estimate the admixture proportions and the timing of events. We analyzed the full autosomal genome for each triplet of samples and obtained the variance in the estimates from a blocked bootstrap with the genome split into 10 Mb blocks and 100 repetitions. Suppl. Fig. S10 shows the 95% confidence intervals obtained from the bootstrap procedure on a time-scale of substitutions per nucleotide. The figure is divided into five parts labeled to correspond to the divergence and admixture events shown in main Fig. 4A. The results are split into separate intervals for each triplet of samples we used and colored according to which triplet of species the samples are from.

As a general rule, confidence intervals are wider for more recent time estimates in this coalescent hidden Markov model, which is also apparent in Suppl. Fig. S11. Because the error bars are very wide for the four most recent events it is hard to draw definitive conclusions. But it appears the split between the northern clade and one of the ancestors of *P. kindae*, (E), and the split between *cynocephalus* and *ursinus* (C) happened close in time, and was quickly followed by the split between the *P. ursinus* lineage and the other ancestor of *P. kindae*, (D). The admixture event forming *P. kindae*, (G), appears to be substantially more recent.



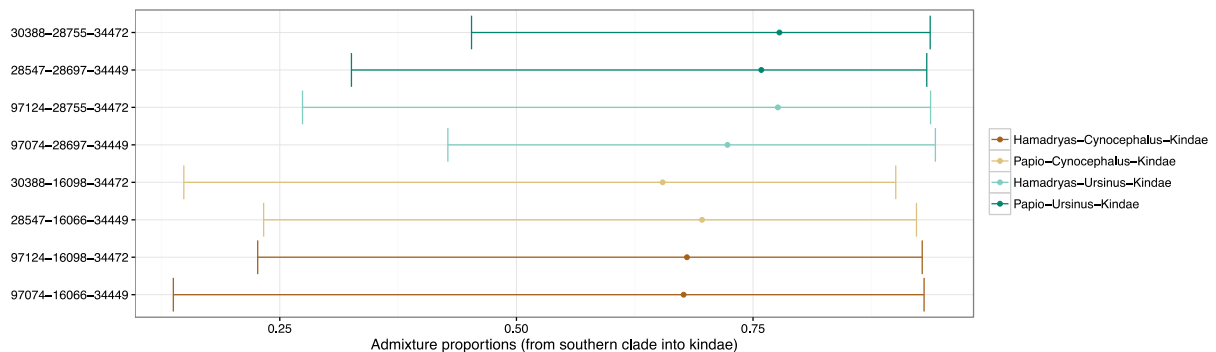
There also appears to be a gradient in divergence times depending on which triplet of species we consider, most clearly seen for the deepest split time, (B). Estimates involving *P. papio* appear to be more recent than estimates involving *P. hamadryas*, and estimates involving *P. ursinus* appear to be more recent than estimates involving *P. cynocephalus*. We do not know what is causing this gradient. Data artifacts seem unlikely since the independent sample triplets for each species triplet fall closer within species than across species. One plausible explanation could be gene flow not accounted for in the admixture graph. The gene flow into the ancestral southern lineage from the *P. papio*/*P. anubis* lineage, inferred in analyses assuming three admixture events (data not shown), might explain why *P. papio* is considered closer to *P. kindae* than is *P. hamadryas*. The more recent divergence time when *P. ursinus* is used instead of *P. cynocephalus* is not explained by the admixture graphs we have explored, but it is possible that these graphs do not capture the full complexity of the baboon phylogeny, and the difference in divergence time inferred here might hint at this. We also estimated the admixture proportions going from the *P. cynocephalus*/*P. ursinus* clade into the ancestral *P. kindae* species. Estimates are shown below in Suppl. Tables S8 and S9.



**Fig. S10. Bootstrap analysis of timing of divergence events.**

**Table S8. Divergence time estimates across triplets.**

split	species	samples	2.5%	50%	97.5%
:-----	:-----	:-----	-----:	-----:	-----:
B	Hamadryas-Cynocephalus-Kindae	97074-16066-34449	0.0005128	0.0005394	0.0005683
B	Hamadryas-Cynocephalus-Kindae	97124-16098-34472	0.0004881	0.0005090	0.0005325
B	Papio-Cynocephalus-Kindae	28547-16066-34449	0.0004941	0.0005197	0.0005371
B	Papio-Cynocephalus-Kindae	30388-16098-34472	0.0004818	0.0005004	0.0005192
B	Hamadryas-Ursinus-Kindae	97074-28697-34449	0.0004875	0.0005003	0.0005159
B	Hamadryas-Ursinus-Kindae	97124-28755-34472	0.0004625	0.0004808	0.0004920
B	Papio-Ursinus-Kindae	28547-28697-34449	0.0004662	0.0004812	0.0005030
B	Papio-Ursinus-Kindae	30388-28755-34472	0.0004541	0.0004673	0.0004837
C	Hamadryas-Cynocephalus-Kindae	97074-16066-34449	0.0001160	0.0003854	0.0004283
C	Hamadryas-Cynocephalus-Kindae	97124-16098-34472	0.0001701	0.0003673	0.0004054
C	Papio-Cynocephalus-Kindae	28547-16066-34449	0.0002103	0.0003930	0.0004246
C	Papio-Cynocephalus-Kindae	30388-16098-34472	0.0001064	0.0003590	0.0003899
D	Hamadryas-Ursinus-Kindae	97074-28697-34449	0.0002050	0.0003138	0.0003441
D	Hamadryas-Ursinus-Kindae	97124-28755-34472	0.0000518	0.0002986	0.0003274
D	Papio-Ursinus-Kindae	28547-28697-34449	0.0001532	0.0003094	0.0003477
D	Papio-Ursinus-Kindae	30388-28755-34472	0.0001488	0.0002929	0.0003194
E	Hamadryas-Cynocephalus-Kindae	97074-16066-34449	0.0000939	0.0003887	0.0004844
E	Hamadryas-Cynocephalus-Kindae	97124-16098-34472	0.0002046	0.0003873	0.0004549
E	Papio-Cynocephalus-Kindae	28547-16066-34449	0.0001591	0.0003949	0.0004623
E	Papio-Cynocephalus-Kindae	30388-16098-34472	0.0001440	0.0003699	0.0004476
E	Hamadryas-Ursinus-Kindae	97074-28697-34449	0.0002187	0.0003729	0.0004444
E	Hamadryas-Ursinus-Kindae	97124-28755-34472	0.0001967	0.0003476	0.0004367
E	Papio-Ursinus-Kindae	28547-28697-34449	0.0001918	0.0003641	0.0004308
E	Papio-Ursinus-Kindae	30388-28755-34472	0.0000947	0.0003354	0.0004283
G	Hamadryas-Cynocephalus-Kindae	97074-16066-34449	0.0000204	0.0001032	0.0003405
G	Hamadryas-Cynocephalus-Kindae	97124-16098-34472	0.0000170	0.0001041	0.0003218
G	Papio-Cynocephalus-Kindae	28547-16066-34449	0.0000092	0.0001161	0.0003334
G	Papio-Cynocephalus-Kindae	30388-16098-34472	0.0000145	0.0001148	0.0002931
G	Hamadryas-Ursinus-Kindae	97074-28697-34449	0.0000178	0.0000961	0.0003074
G	Hamadryas-Ursinus-Kindae	97124-28755-34472	0.0000130	0.0000930	0.0002757
G	Papio-Ursinus-Kindae	28547-28697-34449	0.0000188	0.0001040	0.0002772
G	Papio-Ursinus-Kindae	30388-28755-34472	0.0000128	0.0000850	0.0002531



**Fig. S11. Confidence intervals for baboon admixture proportions.**

**Table S9. Admixture proportion estimates across triplets.**

species	samples	2.5%	50%	97.5%
Hamadryas-Cynocephalus-Kindae	97074-16066-34449	0.1374797	0.6766692	0.9307072
Hamadryas-Cynocephalus-Kindae	97124-16098-34472	0.2266230	0.6801971	0.9287579
Papio-Cynocephalus-Kindae	28547-16066-34449	0.2328688	0.6961275	0.9227150
Papio-Cynocephalus-Kindae	30388-16098-34472	0.1485649	0.6545023	0.9007985
Hamadryas-Ursinus-Kindae	97074-28697-34449	0.4275969	0.7230298	0.9427228
Hamadryas-Ursinus-Kindae	97124-28755-34472	0.2740286	0.7762455	0.9376316
Papio-Ursinus-Kindae	28547-28697-34449	0.3256465	0.7586990	0.9335571
Papio-Ursinus-Kindae	30388-28755-34472	0.4525475	0.7778849	0.9372379

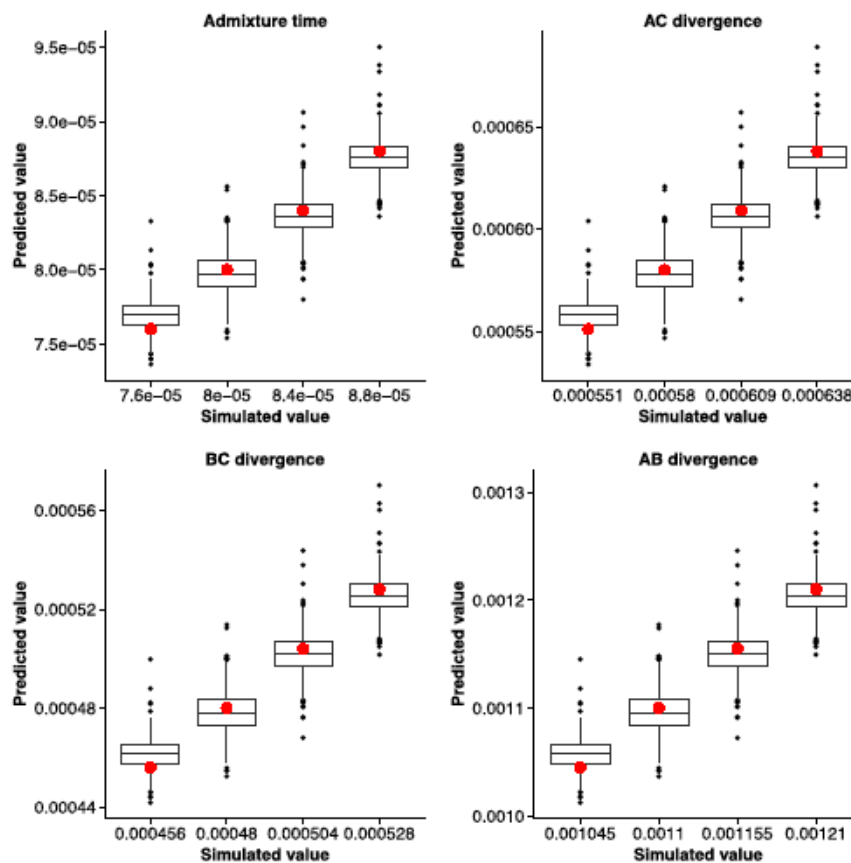
With this model (Suppl. Figure S11) the error bars are very wide, and while the error bars for all analyses overlap there do seem to be lower estimates when the analysis involve *P. cynocephalus* than when it involves *P. ursinus*. This could be an artifact of the graph topology where *P. ursinus* is more closely related to *P. kindae* than is *P. cynocephalus*. But it could also indicate gene flow not accounted for in the graph.

### Simulation test of goodness-of-fit and de-biasing estimates

Simulation experiments with the model (data not shown) have indicated that the time estimates are slightly biased and generally underestimated. To examine which parameters are likely to be biased, and by how much, we simulated data with parameters in a grid of time points around the estimated points and estimated the parameters from this simulated data. Suppl. Fig. S12 shows the results with the estimated time points (A to E) and the admixture proportions together with simulated data, where the simulated values are shown as black points and the corresponding estimated parameters as red error-bars (these error bars are wider since we used smaller data sets for the simulated data for computational reasons).



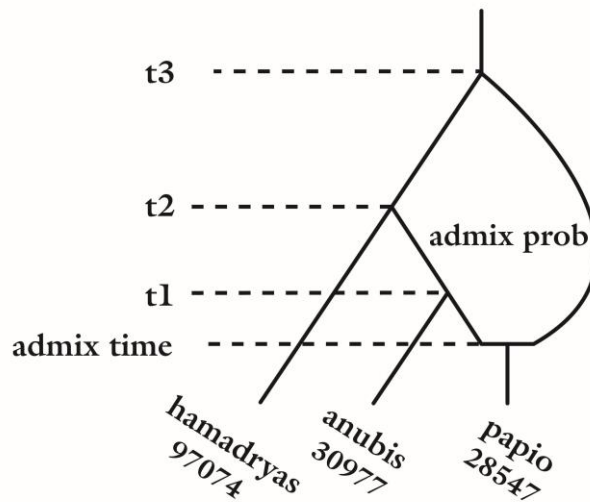
We can build a linear model that predicts the simulated value based on the estimated values in order to correct for the bias in estimates. Using all the estimated parameters as predictor variables and the “true” simulated value as the target parameter we get the following accuracy (Suppl. Fig. S13).



**Fig. S13. Results for correction factor adjustment of admixture history.** This figure below presents box plots illustrating the range of values predicted by the linear model while the red dots show the simulated, “true” value they should hit.

### Testing dates using an alternative topology

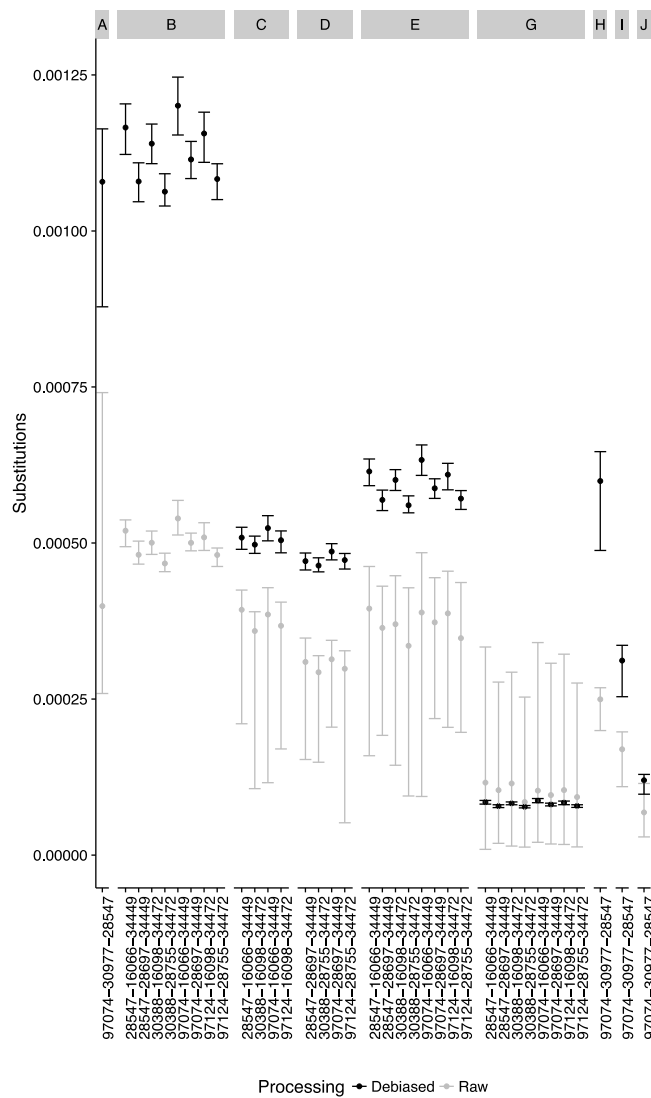
The timing for phylogenetic events A, J and K in Main Text Fig. 4A cannot be estimated using the CoalHMM topology above, so we constructed a similar model with a slightly different admixture topology to capture these events.



**Fig. S14. Model used to estimate specific divergence and admixture history.**

Using this topology (Suppl. Fig. S14), and three samples, one from each species, we estimated the timing of phylogenetic events A, J and K, and constructed linear models to correct the bias in these estimates following the same procedure as for the other triplets and original topology.

**Bias-corrected time estimates:** Applying the bias-correction to all estimates, we obtain the estimates shown in Suppl. Fig. S15 for the timing of events measured in time units of one substitution per nucleotide.

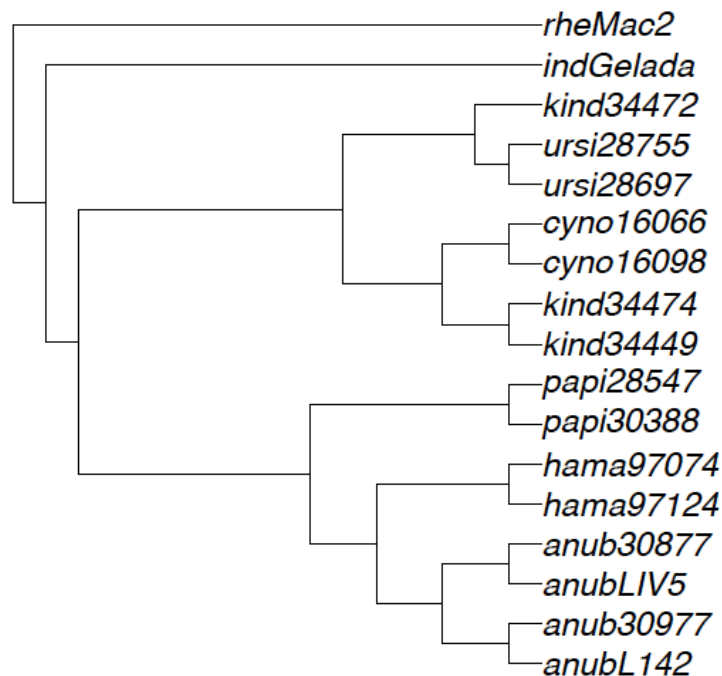


**Fig. S15. Unbiased estimates dating divergences and admixture events.**

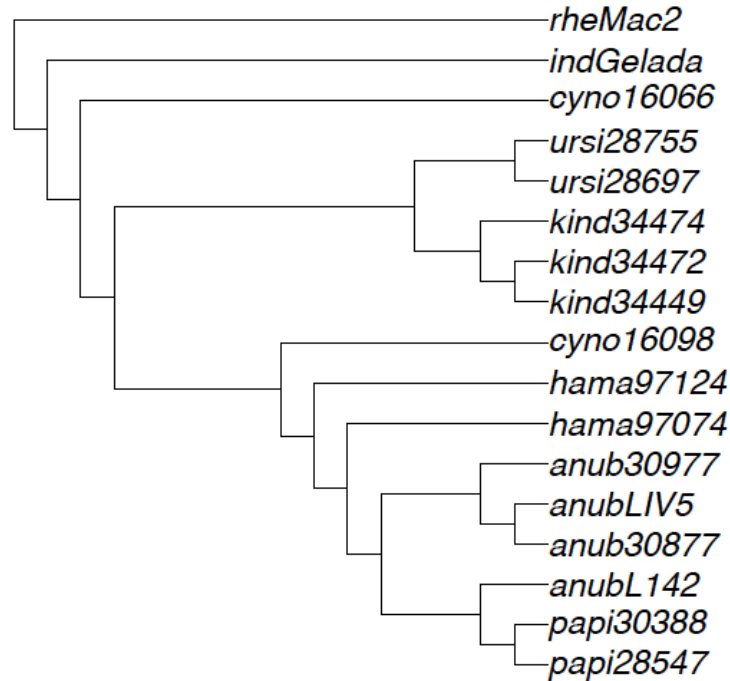


To obtain a point-estimate we take the mean of the estimates in the cases where we have several estimates for the same event and we scale the time to years in two different ways: 1) assuming a  $0.9 \times 10^{-8}$  mutation rate per generation and 11 year generations, or 2) assuming that event B occurred two million years ago.

### Section S13. Locus-specific phylogenetic trees for chromosomal segments containing annotated genes

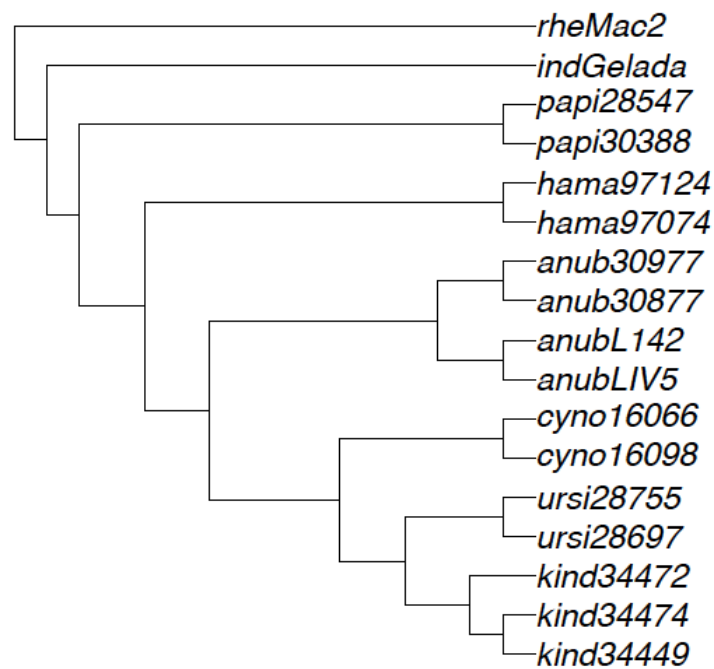


**Fig. S16. Phylogeny representing cluster 1 genic regions.** This tree presents the predominant phylogenetic relationships among the baboon diversity samples, using the 1143 genic regions from Cluster 1. Local phylogenies for individual genic regions, each containing one protein coding gene, were determined and clusters generated through PCA analysis of Euclidean distance metrics (see methods). Cluster 1 consists of 1143 genics regions for which the phylogenies closely match the species-level relationships presented in main text Figure 3.



**Fig. S17. Phylogeny representing cluster 2 genic regions.** This tree presents the predominant phylogenetic relationships among the baboon diversity samples, using the 629 genic regions (Cluster 2). Local phylogenies for individual genic regions, each containing one protein coding gene, were determined and clusters generated through PCA analysis of Euclidean distance metrics (see methods). Cluster 2 consists of 629 genic regions for which haplotypes observed in *P. cynocephalus* are either more closely related to northern clade haplotypes than to other southern clade sequences, or fall in a lineage sister to all other *Papio* haplotypes. The Cluster 2 haplotypes found in *P. cynocephalus* that are closely related to northern clade haplotypes may represent

sequences introgressed into *P. cynocephalus* through gene flow from northern clade animals, particularly *P. anubis*. The genes that fall in Cluster 2 are enriched for Gene Ontology terms “learning and memory,” “cognition,” “head development,” “brain development” and several GO categories related to reproduction (see Suppl. Table S10).



**Fig. S18. Phylogeny representing cluster 3 genic regions.** This tree presents the predominant phylogenetic relationships among the baboon diversity samples, using the 429 genic regions (Cluster 3). Local phylogenies for individual genic regions, each containing one protein coding gene, were determined and clusters generated through PCA analysis of Euclidean distance metrics (see methods). Cluster 3 consists of 429 genic regions for which haplotypes observed in *P. anubis* are more closely related to

southern clade haplotypes than to other northern clade sequences. The haplotypes found in *P. anubis* that are closely related to southern clade haplotypes may represent sequences introgressed into *P. anubis* through gene flow from southern clade animals, particularly *P. cynocephalus*. The genes in Cluster 3 are enriched for Gene Ontology terms related to the ontogenetic development of several organ systems (kidney, heart, circulatory and endocrine systems; See Suppl. Table S10 for more details).

**Table S10. GO terms associated with genes falling in clusters 1 to 3 of genic regions (see appended Excel Table)**