A curated dataset of great ape genome diversity

Sojung Han[1,2], Sepand Riyahi[1,2], Xin Huang[1,2], Martin Kuhlwilm[1,2,*]

1. Department of Evolutionary Anthropology, University of Vienna, Djerassiplatz 1, 1030 Vienna, Austria
2. Human Evolution and Archaeological Sciences (HEAS), University of Vienna, Austria

corresponding author: Martin Kuhlwilm (martin.kuhlwilm@univie.ac.at)

Abstract

Studying the genetic diversity of non-human great apes is important for research questions in evolution as well as human diversity and disease. Genomic data of the three great ape clades (*Pan*, *Gorilla*, *Pongo*) has been published across multiple studies over more than one decade. However, unlike in humans, no comprehensive dataset on great ape diversity is available, due to different scopes of the original studies. Here, we present a curated dataset of 332 high coverage (≥12-fold) whole genomes, including 198 chimpanzee, 16 bonobo, 77 gorilla and 41 orangutan individuals sequenced on the Illumina platform. By integrating data from captive individuals, we contextualize them with data from wild individuals. We discuss issues with previously published data leading to removal of individuals due to low sequencing depth, missing data, or occurrence of duplicate individuals. This resource of files in CRAM and gVCF format, as well as segregating sites per clade, will allow researchers to address questions related to human and great ape evolution and diversity in a comparative manner.

Background & Summary

Great apes have been of long-standing interest as the closest living relatives of humans. Studying their phylogenetic relationships to humans and amongst each other has been an important field of genetic and genomic research over the past few decades[1–4], fostering investigations into human uniqueness and understanding evolution and disease[5–7]. An assembly of the chimpanzee reference genome was published soon after the human reference genome[8], and by now high-quality reference genomes of all great ape species are available[9–11]. However, beyond the insights from single genome assemblies, the diversity within a clade is an important aspect to understand their evolution and their species-specific traits[12,13].

While the advent of high-throughput sequencing technologies allowed characterizing the genetic makeup of many individuals, in humans on a large scale and with high quality[14], this is not possible in great apes. Great apes are all endangered or critically endangered[15], with rapidly shrinking habitats and a high risk of extinction in the wild in the near future, and small captive populations from a limited pool of founder individuals. Hence, a limited number of individuals is available for genomic studies, and often, access to genetic diversity is only possible through non-invasive sampling[16], or from other degraded sources such as historical collections[17]. However, several key publications generated diversity data from all present-day great ape species (Fig. 1): chimpanzees (*Pan troglodytes*)[18,19], bonobos (*Pan paniscus*)[18], western gorillas (*Gorilla gorilla*)[18,20] and eastern gorillas (*Gorilla beringei*)[18,21,22], as well as Bornean, Sumatran and Tapanuli orangutans (*Pongo pygmaeus*, *Pongo abelii*, *Pongo tapanuliensis*)[18,23]. This is complemented by a number of genomes from mostly captive

individuals across these clades, published in the context of genome assemblies[9–11,24], trio sequencing for the estimation of mutation rates[25–27], functional studies[28], or large-scale studies of primate diversity[29].

In many studies of human and hominin evolution, great apes are used for comparison, either as outgroup when calculating statistics[30] or for demographic modelling[31], or with the purpose to put diversity or variation into context[5,32–36]. However, in many cases the great ape reference genomes are used for such purposes rather than diversity data, even though it is often important to put variation within species into context. For example, examining regions of homozygosity[37] or inbreeding patterns in archaic hominins are better understood by variation data from great ape species[38]. Furthermore, information on great ape and primate variation guides studies on human disease[39] and can be used to predict the effects of mutations in present-day people[40].

Given the importance of such data for human-related research questions, we suggest that it is necessary to provide a comprehensive, curated panel of published great ape genomes. Importantly, this should contain information on non-variant sites within populations, to ascertain the status of each position in the genome. A considerable number of individuals is available as short-read sequencing data (*i.e.* using the Illumina platform), while long-read sequencing data was primarily used for reference genome assemblies. Hence, here we focus on the single-nucleotide variant (SNV) diversity, providing a coherent dataset for 332 high-coverage genomes from all extant great ape species[41]. Given that great apes are very closely related to humans and each other, all data was mapped to the human reference genome (GrCH38)[19]. We provide both the mappings (in CRAM format) and whole-genome variant call files (gVCF files), as well as sets of segregating sites across clades, all of which we hope will be a useful resource for numerous studies on human and primate genomics.

Methods

Samples

We used publicly available great ape genomes published in different studies (Table 1). These entail 23 chimpanzees, 12 bonobos, 27 gorillas and 10 orangutans from a landmark study on great ape diversity[18]; 32 chimpanzees[19], 21 gorillas[20–22] and 15 orangutans[23] from subsequent population-scale studies on different wild-born individuals; and a captive panel of in total 143 chimpanzees, 4 bonobos, 29 gorillas and 16 orangutans from multiple studies with different focus[10,11,24–29,42–47]. All sequencing data was publicly available on the Sequencing Read Archive (SRA), and obtained through the European Nucleotide Archive branch for this study (Table S1). We did not consider individuals from studies reporting partial genomic data (such as chromosome 21[16] or the exome[48]) or with insufficient sequencing coverage[17,49] (below 12-fold, on average, across the genome). In some of the studies considered here, sequencing data was reported for additional individuals[18–20,23,46], which we excluded due to low average coverage, or reported evidence of cross-contamination[18]. In the case of one individual (SAMEA104361539[23]), no data was available for one sequencing run accession (ERX2240355), leading to insufficient coverage. Finally, we merged data for identical individuals published using different identifiers or in different studies in order to reach sufficient coverage (see section “Captive panel” in Technical Validation). We only considered data generated through Illumina short-read sequencing, in an effort of building a coherent dataset. We note that both long-read and short-read sequencing data was generated for some of the individuals from which the most recent genome assemblies were generated – in such cases the short-read data are included here[10,11].

	Pan		*Gorilla*		*Pongo*		
Study accession	**Chimpanzee**	**Bonobo**	**Western**	**Eastern**	**Sumatran**	**Bornean**	**Tapanuli**
PRJNA189439[18]	21	12	24	3	5	5	
PRJEB15086[19]	32						
PRJEB19688[23]					5	9	1
PRJEB3220[21]				12			
PRJEB12821[22]				6			
PRJEB60463[20]			3				
PRJNA563344[28]	2		2			1	
PRJEB59576[29]	4		19	1	1	1	
PRJEB29710[26]	4		4		3		
PRJEB5937[25]	9						
PRJDB3537[27]	3						
PRJNA369439[11]					1		
PRJNA526933[10]		1					
PRJNA986878, PRJNA986879, PRJNA976699, PRJNA976700, PRJNA976701, PRJNA602326[24]	1	3	3		2	1	
PRJNA785018[42]						1	
PRJEB1675[45]	5				5		
PRJEB21589[43]	1						
PRJNA856315[46]	77						
PRJNA635393[47]	39						
Total	**198**	**16**	**55**	**22**	**22**	**18**	**1**

Table 1. Overview of studies and numbers of individuals from each clade represented. Sequencing Read Archive (SRA) study identifiers and corresponding references are given in the first column. For metadata and per-sample identifiers see Tables S1-2.

Bioinformatic processing

Raw fastq files were downloaded using sratoolkit (https://hpc.nih.gov/apps/sratoolkit.html, version 3.0.6), and fastQC[50] was applied for initial quality assessment. Adapter trimming was performed with trimmomatic (version 0.39)[51]. Reads were then mapped to the human genome version GRCh38 (GCA_000001405.15_GRCh38_no_alt_analysis_set from the UCSC genome browser) with bwa mem (version 0.7.16a)[52], sorted with samtools (version 1.14)[53] and unmapped reads removed. Read groups were assigned with picardtools (version 2.21.4) (http://broadinstitute.github.io/picard/), and duplicated reads were marked with GATK (version 4.1.4.0)[54] MarkDuplicatesSpark. Finally, reads from all sequencing libraries for each individual were merged with samtools merge into a single CRAM file. These files are reported in the associated dataset. For two individuals, sequencing depth was more than 200-fold[27], resulting in excessive spurious heterozygous calls. For coherence of the dataset, we restricted the analysis to a subset of the raw sequencing data (only one out of two run accessions each, as reported in Table S1).

Using this merged CRAM file per individual, genotypes were called per chromosome with GATK HaplotypeCaller, using the flag '-ERC GVCF' to generate genome-wide VCF files. These files are

reported in the associated dataset. For haploid sex chromosomes in male individuals, we performed haploid genotype calling.

For downstream analyses, we created joint callsets with GATK GenomicsDBImport, and GATK GenotypeGVCFs. After creating the callset for the wild-born individuals, we added the captive panels. We report sets of segregating sites per clade (*Pan*, *Gorilla*, *Pongo*) with and without the captive panel within the PHAIDRA repository[41] (see Data Records section). Both sets are available as VCF files and in PLINK2 format after conversion using plink2[55] (version 2.00a5) with the parameters '--max-alleles 2 --snps-only --make-pgen --maf 0.00'. A permissive set contains all information on segregating sites per individual. We also obtained a more stringent set after filtering using bcftools (version 1.21)[56], retaining only bi-allelic SNPs passing a 36-basepair mappability filter[57], excluding sites outside the central 98% of the coverage distribution per individual with a minimum of 5-fold coverage per site, and removing heterozygous positions with less than 15% of reads supporting one allele. We also report a joint set of segregating sites across all species[41].

Data analysis

We estimated depth of coverage using mosdepth[58] (version 0.3.3) on the CRAM files, and used bcftools[56] (version 1.16) to generate summary statistics on the VCF files. Genetic sexing was performed using ratios of the mean coverage per chromosome, with chrX:chr1 smaller than 0.75 and chrY:chrX larger than 0.1 to determine male sex.

We performed Principal Component Analysis (PCA) using VCF2PCACluster (version 1.41)[59] on VCF files before and after filtering. For ADMIXTURE analyses, we subsampled 1,000,000 random autosomal loci from the VCF files and ran ADMIXTURE (version 1.3.0)[60]. Relatedness estimates were calculated using ngsRelate[61] (version 2.0). Runs of Homozygosity were detected using bcftools roh[62] (version 1.21) per chromosome per individual. Human contamination on captive individuals was estimated using HuConTest[63]. Subspecies assignment with f3-statistics was performed using admixtools2[64]. As outgroup, ancestral alleles were approximated by liftover of genomic coordinates to the macaque reference genome (rheMac10)[65] using rtracklayer[66] (version 1.58.0 in R version 4.2.2) and bedtools getfasta[67] (version 2.31.1). Geolocalization of captive chimpanzees was performed using rareCAGA[16] after liftover of genotypes to the human genome version hg19 with bcftools liftover[68].

We used R[69] versions 4.2.3 and 4.2.2 for plotting, with packages ggplot2[70] (versions 3.4.4 and 3.5.1), gridExtra (version 2.3; 10.32614/CRAN.package.gridExtra), dplyr[71] (version 1.1.4), tidyverse[72] (version 2.0.0), ggh4x[73] (version 0.2.8).

Data Records

The full dataset is available through PHAIDRA with the University of Vienna, under the following DOI: https://doi.org/10.25365/phaidra.514[41]. This dataset contains the CRAM files (mapped reads) for 332 individuals, as well as gVCF files (genotype calls) for all 332 individuals for autosomes and X chromosomes, as well as Y chromosomes for the male individuals. Note that gVCF files are in the intermediate format provided by GATK HaplotypeCaller, which can be used for joint genotype calling but might be genotyped if used individually. For all files, md5sums are provided in Table S7. Furthermore, joint genotype calls in VCF format are available for the three species complexes *Pan*, *Gorilla* and *Pongo*, for the full set of genotype calls as well as a filtered set. A set of joint genotype call files across all 332 individuals are available on PHAIDRA[41], as well as the EVA platform[74] under the accession PRJEB97324[75].

Technical Validation

Sequencing data

We report a curated dataset of previously published genomic data for 138 wild-born great ape individuals[18–23,25], which constitute a reference panel for population genetic studies[41]. We also included 194 captive individuals from multiple studies[10,11,18,24–29,42,43,45–47], resulting in a total dataset of 332 individuals. Only individuals with at least 12-fold average coverage across the genome were included, with a median of 23-fold and a maximum of 141-fold coverage (as obtained by mosdepth). Since in some cases the coverage of called genotypes was below this threshold, we also excluded three such individuals (SAMN02736775, SAMN01920524 and SAMEA104361528). Using the average coverage for the sex chromosomes, we report 208 (63%) of individuals as female, and 124 (37%) as male. We provide this information, alongside other summary metrics, in Table S2 and S3, as well as Fig. S1.

We obtained several quality control measures to ensure completeness of the data: average coverage per chromosome, the last position per chromosome, the numbers of non-reference records and heterozygous positions per chromosome, and the ratio of transitions to transversions per individual (Fig. S2-7). We present the genome-wide average coverage and heterozygosity in Fig. 2. Heterozygosity values recapitulate findings from previous studies when stratified by the different subspecies[3,18,19,21–23]. Furthermore, we estimated potential human contamination[63] in captive individuals. We set a threshold of 1% in order to include individuals (Fig. 4A, Table S5), leading to exclusion of some individuals (SAMN29543728, SAMN29543727, SAMN29543724, SAMN29543729)[46] with values above 1%.

Combined genotype calls of segregating sites per clade contain 117,472,161 sites for *Pan* (23,468,769 high quality sites after filtering), 80,907,619 sites for *Gorilla* (28,214,170 high quality sites after filtering), and 139,874,238 sites for *Pongo* (42,620,590 high quality sites after filtering). For downstream analyses, usually a coverage-based filtering is recommended. We report the central 98% of the coverage distribution for each individual, separately for autosomes and chromosome X in Table S4, with a lower cutoff of 5-fold coverage in cases where this value was below five. We advise the user to carefully consider additional filtering depending on their intended use of this dataset.

Population genetic validation

We performed basic population genetic characterisation of the individuals in this dataset, which allows to assess the quality of the data in the context of previous findings. First, we performed a PCA, showing the expected population clustering of all individuals within the respective clades *Pan*, *Gorilla* and *Pongo* (Fig. 3A-C; Fig. S8-12). Captive individuals are shown in grey. Notably, a PCA on the unfiltered data shows outliers for the three orangutan individuals with the highest amount of human contamination in the sequencing data (see section below; Fig. S12). We also performed clustering with ADMIXTURE (Fig. 3D; Fig. S13-15), which recapitulates known patterns of subspecies stratification in these great ape species[3,18,19,21–23].

Since we initially included all data reported in previous studies, we found several identical individuals based on relatedness estimates[61] (KING relatedness larger than 0.4). No indication of identity was given in these respective publications. This affects the orangutan individual PD_0262/ORAN23[28,29], as well as the gorilla individuals Banjo[18,26], Mimi[18,26], Mawenzi/PD_0264[26,29] and PD_0189/PD_026[29], each of which have two unique SRA biosample IDs. Furthermore, a total of 17 identical chimpanzee individuals sequenced in different studies were identified. Remarkably, the individual Donald[18] appears to have been sequenced in three independent studies (as 4x0519[47] and NS07602[46]). In most duplicate cases, the more recent study yielded high-coverage genomes (>30-fold), which we used for building this dataset. In some cases, in order to increase coverage we merged data after the additional step of inspecting heterozygosity. We report heatmaps of relatedness estimates including these

identical individuals in the Supplementary Materials (Fig. S22-23), and provide a table of the biosample IDs for duplicated individuals (Table S5). We conclude that our data is comprehensive and reflects the original data published through these studies.

Usage notes

Beyond the well-characterized datasets of wild-born great apes presented here as a reference dataset, we included 198 captive individuals from different studies. As described above, we assessed human contamination, retaining only individuals with less than 1% contamination. However, for three orangutan individuals, values close to 1% apparently still lead to false genotype calls and a shift in the PCA (Fig. S12). We conclude that quality filtering is recommended for subsequent analyses.

Furthermore, we provide an accurate assignment on the subspecies level based on f3-statistics (Table S6, Fig. S20)[76], since 24 gorilla and 131 chimpanzee individuals did not have subspecies-level information in their SRA record, as well as two orangutans which were only labelled as *Pongo*. We assigned these two orangutans as *Pongo pygmaeus*, as reported in supplementary materials of a corresponding study[77], though not in the SRA database. Most gorilla individuals are *Gorilla gorilla gorilla*, with the exception of PD_0179, a *Gorilla beringei graueri*, also reported as such only in the supplementary of the corresponding publication[29] (Fig. 4B). Among chimpanzees, we identify PD_0259 and Rogger as *Pan troglodytes schweinfurthii*, CH114 as *Pan troglodytes troglodytes*, and 88A020 as *Pan troglodytes ellioti,* while all other chimpanzees are *Pan troglodytes verus.* The individual Donald/4x051946/NS0760245 is a known subspecies hybrid[18]. Furthermore, we performed geolocalisation of the captive chimpanzees (Fig. S21), finding, for example, an approximate origin of PD_0259 in northern Democratic Republic of Congo (Fig. 4C). We also identify 16 further individuals as likely subspecies hybrids in captivity (Fig. S21, Table S6). These analyses give a meaningful context for the genomes of these individuals, as they can complement diversity datasets of their respective subspecies or local population groups.

We also estimated pairwise relatedness between individuals[61], recapitulating different degrees of background relatedness in some of the groups (Fig. S16-18), *e.g.* among bonobos (Fig. 5A) or Mountain gorillas (Fig. S17). Individuals from studies aimed at mutation rate estimates through trio sequencing[24–27] were clearly identifiable by their first-degree relationships. Furthermore, multiple first-degree relationships were determined in captive chimpanzees. Known relationships are provided in Table S2, as well as inferred first-degree relationships (based on KING relatedness larger than 0.2), allowing to exclude such individuals from downstream analyses. Finally, we estimated runs of homozygosity[62] (RoHs), a measure informative on long-term small effective population sizes, bottlenecks, and recent inbreeding[78]. We largely recapitulate previous findings[18], *e.g.* more such RoHs in bonobos than chimpanzees or more in eastern gorillas than western lowland gorillas, while the captive individuals do not seem to show a systematic increase in RoHs (Fig. 5B; Fig. S19). Metadata are presented in Supplementary Tables and Figures, with Supplementary Figure legends in Table S8.

Supplementary Tables

Table S1. Sequencing read archive (SRA) identifiers used to retrieve raw sequencing data for each individual.

Table S2. Metadata (species/subspecies, mean coverage, sex, relatedness, and other).

Table S3. Average coverage per chromosome.

Table S4. Central 98% coverage per individual.

Table S5. Reference table for duplicated individuals with their Biosample-IDs.

Table S6. Subspecies-assignment using f3-statistics, human contamination estimates for captive individuals, and geolocalization of chimpanzees.

Table S7. Md5sums for the CRAM and gVCF files in the dataset.

Supplementary Figures

Figure S1. Coverage distribution per individual, stratified by subspecies.

Figure S2. Coverage per chromosome for 22 autosomes, stratified by subspecies.

Figure S3. Coverage for sex chromosomes, stratified by subspecies and sex (chrY for males only, mitochondrial genome (chrM) for all individuals).

Figure S4. Last called position per chromosome, ensuring completeness of data, stratified by subspecies.

Figure S5. Number of non-reference genotype calls per chromosome, stratified by subspecies.

Figure S6. Number of heterozygous positions per chromosome, stratified by subspecies.

Figure S7. Transition-to-transversion ratio per chromosome, stratified by individual.

Figure S8: Principal Component Analysis for 4 PCs all *Pan* individuals, calculated on unfiltered sites.

Figure S9: Principal Component Analysis for 4 PCs all *Pan troglodytes* (chimpanzee) individuals, calculated on unfiltered sites.

Figure S10: Principal Component Analysis for 4 PCs all *Pan paniscus* (bonobo) individuals, calculated on unfiltered sites.

Figure S11: Principal Component Analysis for 4 PCs all *Gorilla* individuals, calculated on unfiltered sites.

Figure S12: Principal Component Analysis for 4 PCs all *Pongo* individuals, calculated on unfiltered sites.

Figure S13: ADMIXTURE clustering for all *Pan* individuals, for 325 k=2 to k=10, calculated on 1,000,000 randomly chosen SNVs.

Figure S14: ADMIXTURE clustering for all *Gorilla* individuals, for k=1 to k=10, calculated on 1,000,000 randomly chosen SNVs.

Figure S15: ADMIXTURE clustering for all *Pongo* individuals, for k=1 to k=10, calculated on 1,000,000 randomly chosen SNVs.

Figure S16. Relatedness between all *Pan* individuals, as determined by KING in ngsRelate. Calculated on filtered dataset.

Figure S17. Relatedness between all *Gorilla* individuals, as determined by KING in ngsRelate. Calculated on filtered dataset.

Figure S18. Relatedness between all *Pongo* individuals, as determined by KING in ngsRelate. Calculated on filtered dataset.

Figure S19. Runs of Homozygosity (RoHs) in all individuals, as estimated by bcftools roh, stratified by subspecies and partitioned into short (50,000-250,000 bp), medium (250,000-1,000,000 bp) and long (>1,000,000) RoHs.

Figure S20. Subspecies assignment based on f3-statistics.

Figure S21. Geolocalization of chimpanzees based on rareCAGA.

Figure S22. Relatedness between all *Gorilla* individuals, including duplicated individuals that were removed from the dataset, as determined by KING in ngsRelate. Calculated on an unfiltered dataset.

Figure S23. Relatedness between all *Pongo* individuals, including duplicated individuals that were removed from the dataset, as determined by KING in ngsRelate. Calculated on an unfiltered dataset.

Figure S24. Relatedness between all *Pan* individuals, including duplicated individuals that were removed from the dataset, as determined by KING in ngsRelate. Calculated on an unfiltered dataset.

Code Availability

The code used is available under https://github.com/admixVIE/Great_Ape_genomes.

Data Availability

The dataset is available at https://doi.org/10.25365/phaidra.514, and has been deposited to EVA [PRJEB97324].

Acknowledgements

We thank N. Schulmeister for revising parts of the data. This project has been funded by the Vienna Science and Technology Fund (WWTF) [10.47379/VRG20001] to M.K. S.H. was supported by the Austrian Science Fund (FWF) [10.55776/ESP546]. The computational results of this work have been achieved using supercomputer resources provided by the Vienna Scientific Cluster (VSC) and the Life Science Compute Cluster (LiSC) of the University of Vienna.

Author contributions

M.K. conceived the study, analysed data, and wrote the manuscript with feedback from all coauthors. S.H., S.R. and X.H. analysed data.

Competing interests

The authors declare no competing interests.

References

1. Kaessmann, H. & Pääbo, S. The genetical history of humans and the great apes. *J. Intern. Med.* **251**, 1–18 (2002).
2. Wall, J. D. Great ape genomics. *ILAR J.* **54**, 82–90 (2013).
3. Kuhlwilm, M. *et al.* Evolution and demography of the great apes. *Curr. Opin. Genet. Dev.* (2016) doi:10.1016/j.gde.2016.09.005.
4. Yousaf, A., Liu, J., Ye, S. & Chen, H. Current Progress in Evolutionary Comparative Genomics of Great Apes. *Front. Genet.* **12**, (2021).
5. Pollen, A. A., Kilik, U., Lowe, C. B. & Camp, J. G. Human-specific genetics: new tools to explore the molecular and cellular basis of human evolution. *Nat. Rev. Genet.* **24**, 687–711 (2023).
6. Varki, A. & Altheide, T. K. Comparing the human and chimpanzee genomes: Searching for needles in a haystack. *Genome Res.* **15**, 1746–1758 (2005).
7. Juan, D., Santpere, G., Kelley, J. L., Cornejo, O. E. & Marques-Bonet, T. Current advances in primate genomics: novel approaches for understanding evolution and disease. *Nat. Rev. Genet.* **24**, 314–331 (2023).
8. Consortium, T. C. S. and A. & The Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69–87 (2005).
9. Gordon, D. *et al.* Long-read sequence assembly of the gorilla genome. *Science (80-.).* **352**, aae0344–aae0344 (2016).
10. Mao, Y. *et al.* A high-quality bonobo genome refines the analysis of hominid evolution. *Nature* **594**, 77–81 (2021).
11. Kronenberg, Z. N. *et al.* High-resolution comparative analysis of great ape genomes. *Science (80-.).* **360**, eaar6343 (2018).
12. Zeberg, H., Jakobsson, M. & Pääbo, S. The genetic changes that shaped Neandertals, Denisovans, and modern humans. *Cell* **187**, 1047–1058 (2024).
13. Han, S., Andrés, A. M., Marques-Bonet, T. & Kuhlwilm, M. Genetic variation in Pan species is shaped by demographic history and harbors lineage-specific functions. *Genome Biol. Evol.* evz047 (2019) doi:10.1093/gbe/evz047.
14. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
15. Caldecott, J. & Miles, L. World Atlas of Great Apes and their Conservation. in *Environmental Conservation* 33, 456 (University of California Press, 2005).
16. Fontsere, C. *et al.* Population dynamics and genetic connectivity in recent chimpanzee history. *Cell Genomics* **2**, (2022).
17. van der Valk, T., Díez-del-Molino, D., Marques-Bonet, T., Guschanski, K. & Dalén, L. Historical Genomes Reveal the Genomic Consequences of Recent Population Decline in Eastern Gorillas. *Curr. Biol.* **29**, 165-170.e6 (2019).
18. Prado-Martinez, J. *et al.* Great ape genetic diversity and population history. *Nature* **499**, 471–5 (2013).
19. De Manuel, M. *et al.* Chimpanzee genomic diversity reveals ancient admixture with bonobos. *Science (80-.).* **354**, (2016).

20. Alvarez-Estape, M. *et al.* Past Connectivity but Recent Inbreeding in Cross River Gorillas Determined Using Whole Genomes from Single Hairs. *Genes* vol. 14 (2023).
21. Xue, Y. *et al.* Mountain gorilla genomes reveal the impact of long-term population decline and inbreeding. *Science (80-.).* **348**, 242–245 (2015).
22. Pawar, H. *et al.* Ghost admixture in eastern gorillas. *Nat. Ecol. Evol.* (2023) doi:10.1038/s41559-023-02145-2.
23. Nater, A. *et al.* Morphometric, Behavioral, and Genomic Evidence for a New Orangutan Species. *Curr. Biol.* **27**, 3487-3498.e10 (2017).
24. Makova, K. D. *et al.* The complete sequence and comparative analysis of ape sex chromosomes. *Nature* (2024) doi:10.1038/s41586-024-07473-2.
25. Venn, O. *et al.* Strong male bias drives germline mutation in chimpanzees. *Science (80-.).* **344**, 1272–1275 (2014).
26. Besenbacher, S., Hvilsom, C., Marques-Bonet, T., Mailund, T. & Schierup, M. H. Direct estimation of mutations in great apes reconciles phylogenetic dating. *Nat. Ecol. Evol.* **3**, 286–292 (2019).
27. Tatsumoto, S. *et al.* Direct estimation of de novo mutation rates in a chimpanzee parent-offspring trio by ultra-deep whole genome sequencing. *Sci. Rep.* **7**, 13561 (2017).
28. García-Pérez, R. *et al.* Epigenomic profiling of primate lymphoblastoid cell lines reveals the evolutionary patterns of epigenetic activities in gene regulatory architectures. *Nat. Commun.* **12**, 3116 (2021).
29. Kuderna, L. F. K. *et al.* A global catalog of whole-genome diversity from 233 primate species. *Science (80-.).* **380**, 906–913 (2023).
30. Prüfer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43–9 (2014).
31. Gronau, I., Hubisz, M. J., Gulko, B., Danko, C. G. & Siepel, A. Bayesian inference of ancient human demography from individual genome sequences. *Nat. Genet.* **43**, 1031–1034 (2011).
32. Staes, N. *et al.* FOXP2 variation in great ape populations offers insight into the evolution of communication skills. *Sci. Rep.* **7**, 16866 (2017).
33. Dennis, M. Y. *et al.* Evolution of Human-Specific Neural SRGAP2 Genes by Incomplete Segmental Duplication. *Cell* **149**, 912–922 (2012).
34. Mangan, R. J. *et al.* Adaptive sequence divergence forged new neurodevelopmental enhancers in humans. *Cell* **185**, 4587-4603.e23 (2022).
35. Weiss, C. V *et al.* The \textit{cis}-regulatory effects of modern human-specific variants. *Elife* **10**, e63713 (2021).
36. Huang, X., Kruisz, P. & Kuhlwilm, M. sstar: A Python package for detecting archaic introgression from population genetic data with S *. *Mol. Biol. Evol.* **39**, 1–6 (2022).
37. Kuhlwilm, M. *et al.* Ancient gene flow from early modern humans into Eastern Neanderthals. *Nature* **530**, 429–433 (2016).
38. Skov, L. *et al.* Genetic insights into the social organization of Neanderthals. *Nature* **610**, 519–525 (2022).
39. Richard, D. *et al.* Evolutionary Selection and Constraint on Human Knee Chondrocyte Regulation Impacts Osteoarthritis Risk. *Cell* **181**, 362-381.e28 (2020).
40. Gao, H. *et al.* The landscape of tolerated genetic variation in humans and primates. *Science (80-.).* **380**, eabn8153 (2023).
41. Kuhlwilm, M. Great ape genome diversity panel. https://doi.org/10.1101/2025.02.18.638799 (2025).
42. Shao, Y. *et al.* Phylogenomic analyses provide insights into primate evolution. *Science (80-.).* **380**, 913–924 (2023).
43. Solis-Moruno, M. *et al.* Potential damaging mutation in LRP5 from genome sequencing of the first reported chimpanzee with the Chiari malformation. *Sci. Rep.* **7**, 15224

(2017).
44. Shukla, N., Shaban, B. & Gallego Romero, I. Genetic Diversity in Chimpanzee Transcriptomics Does Not Represent Wild Populations. *Genome Biol. Evol.* **13**, evab247 (2021).
45. Gokcumen, O. *et al.* Primate genome architecture influences structural variation mechanisms and functional consequences. *Proc. Natl. Acad. Sci.* **110**, 15764–15769 (2013).
46. Bracci, A. N. *et al.* The evolution of the human DNA replication timing program. *Proc. Natl. Acad. Sci.* **120**, e2213896120 (2023).
47. Fair, B. J. *et al.* Gene expression variability in human and chimpanzee populations share common determinants. *Elife* **9**, e59929 (2020).
48. Teixeira, J. C. *et al.* Long-Term Balancing Selection in LAD1 Maintains a Missense Trans-Species Polymorphism in Humans, Chimpanzees, and Bonobos. *Mol. Biol. Evol.* **32**, 1186–1196 (2015).
49. Locke, D. P. *et al.* Comparative and demographic analysis of orang-utan genomes. *Nature* **469**, 529–33 (2011).
50. Andrews, S. FASTQC. A quality control tool for high throughput sequence data. (2010).
51. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
52. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
53. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
54. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
55. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, s13742-015-0047–8 (2015).
56. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).
57. Karimzadeh, M., Ernst, C., Kundaje, A. & Hoffman, M. M. Umap and Bismap: quantifying genome and methylome mappability. *Nucleic Acids Res.* **46**, e120–e120 (2018).
58. Pedersen, B. S. & Quinlan, A. R. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* **34**, 867–868 (2018).
59. He, W. *et al.* VCF2PCACluster: a simple, fast and memory-efficient tool for principal component analysis of tens of millions of SNPs. *BMC Bioinformatics* **25**, 173 (2024).
60. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
61. Korneliussen, T. S. & Moltke, I. NgsRelate: A software tool for estimating pairwise relatedness from next-generation sequencing data. *Bioinformatics* **31**, 4009–4011 (2015).
62. Narasimhan, V. *et al.* BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. *Bioinformatics* **32**, 1749–1751 (2016).
63. Kuhlwilm, M., Fontsere, C., Han, S., Alvarez-Estape, M. & Marques-Bonet, T. HuConTest: Testing Human Contamination in Great Ape Samples. *Genome Biol. Evol.* **13**, 2021.03.30.437753 (2021).
64. Maier, R. *et al.* On the limits of fitting complex models of population history to \textit{f}-statistics. *Elife* **12**, e85492 (2023).
65. Warren, W. C. *et al.* Sequence diversity analyses of an improved rhesus macaque genome enhance its biomedical utility. *Science (80-.).* **370**, eabc6617 (2020).
66. Lawrence, M., Gentleman, R. & Carey, V. rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics* **25**, 1841–1842 (2009).

67. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
68. Genovese, G. *et al.* BCFtools/liftover: an accurate and comprehensive tool to convert genetic variants across genome assemblies. *Bioinformatics* **40**, btae038 (2024).
69. R Core Team. R: A Language and Environment for Statistical Computing. (2015).
70. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer-Verlag New York, 2009).
71. Wickham, H., François, R., Henry, L., Müller, K. & Vaughan, D. dplyr: A Grammar of Data Manipulation. (2023).
72. Wickham, H. *et al.* Welcome to the Tidyverse. *J. Open Source Softw.* **4**, 1686 (2019).
73. van den Brand, T. ggh4x: Hacks for 'ggplot2'. (2024).
74. Cezard, T. *et al.* The European Variation Archive: a FAIR resource of genomic variation for all species. *Nucleic Acids Res.* **50**, D1216–D1220 (2022).
75. European Variation Archive. http://identifiers.org/ena.embl:PRJEB97324 (2025).
76. Maier, R. & Patterson, N. admixtools: Inferring demographic history from genetic data. (2024).
77. Ferrández-Peral, L. *et al.* Transcriptome innovations in primates revealed by single-molecule long-read sequencing. *Genome Res.* **32**, 1448–1462 (2022).
78. Kirin, M. *et al.* Genomic runs of homozygosity record population history and consanguinity. *PLoS One* **5**, 1–7 (2010).

Figure Legends

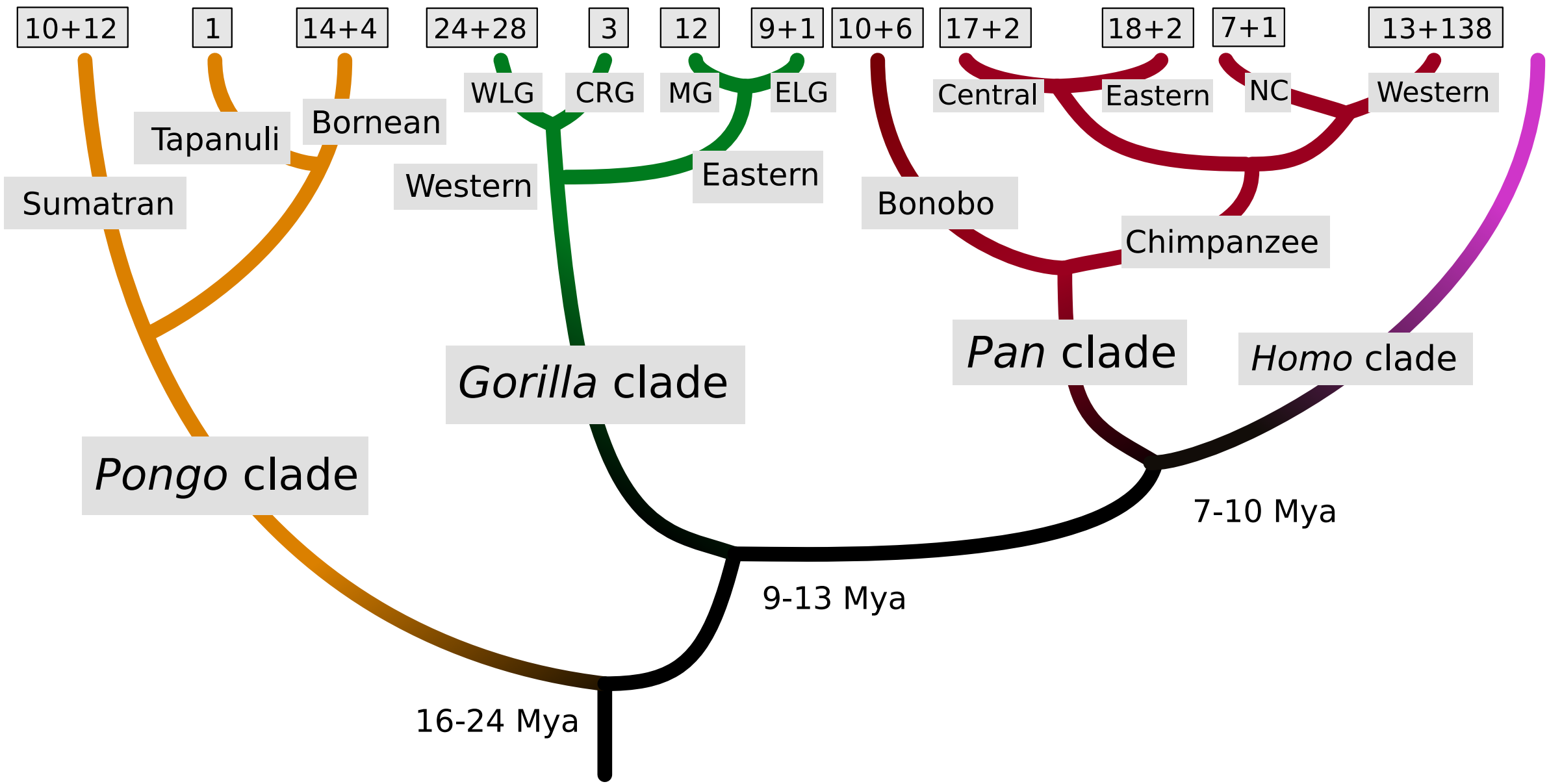
Figure 1. Phylogeny and overview of individuals from each of the clades. Orangutans (*Pongo*): Sumatran = *P. abelii*; Tapanuli = *P. tapanuliensis*; Bornean = *P. pygmaeus*. Gorillas (*Gorilla*): Western lowland gorilla (WLG) = *G. gorilla gorilla*; Cross-river gorilla (CRG) = *G. gorilla diehli*; Mountain gorilla (MG) = *G. beringei beringei*; Eastern lowland gorilla (ELG) = *G. beringei graueri*. Chimpanzees and bonobos (*Pan*): Bonobo = *P. paniscus*; Central chimpanzee = *P. troglodytes troglodytes*; Eastern chimpanzee = *P. troglodytes schweinfurthii*; Nigeria-Cameroon chimpanzee (NC) = *P. troglodytes ellioti*; Western chimpanzee = *P. troglodytes verus*. Divergence times in million years ago (Mya) from previous studies. Numbers of individuals as reference panel plus (+) captive panel.

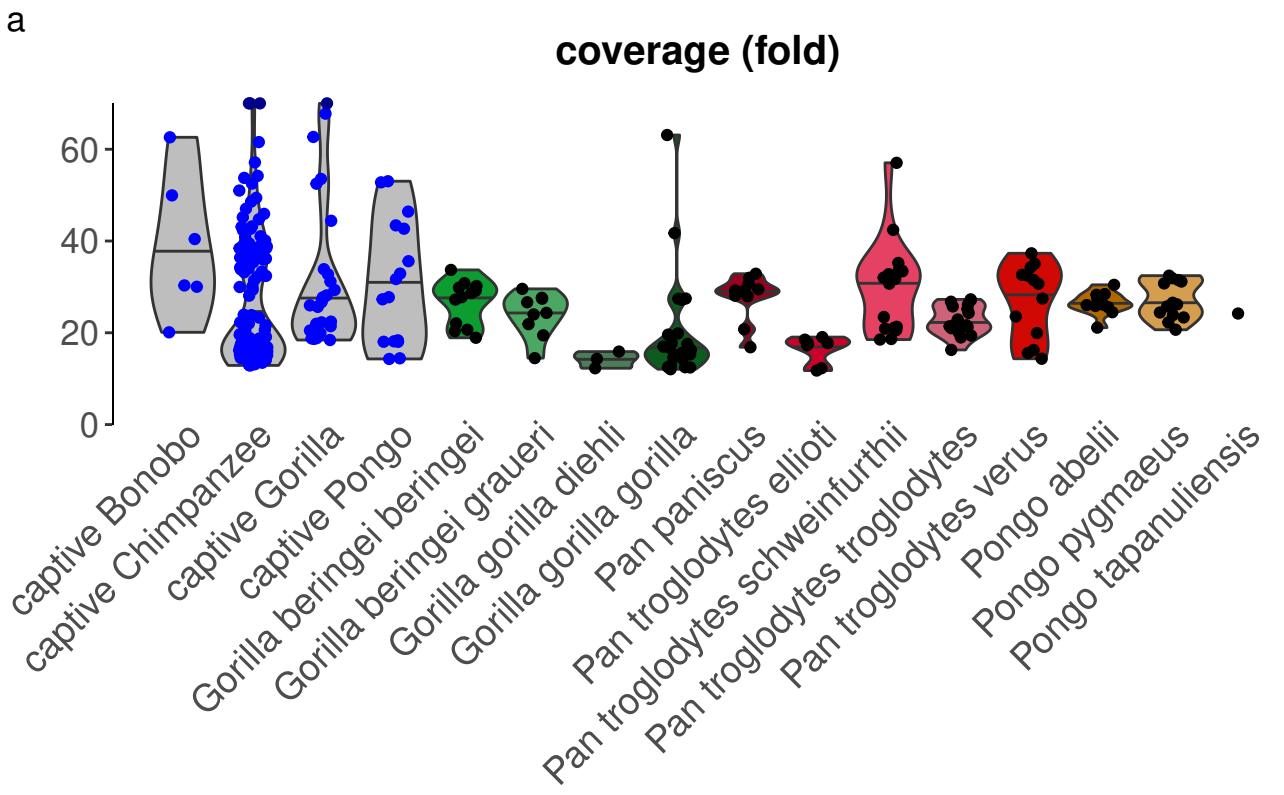
Figure 2. Distributions of coverage and heterozygosity. A) Average coverage per individual across subspecies. Note: captive *Pan* is cut at 70-fold for three samples with more than 100-fold coverage. B) Genome-wide heterozygosity per 1,000 base pairs (bp) per individual across subspecies.

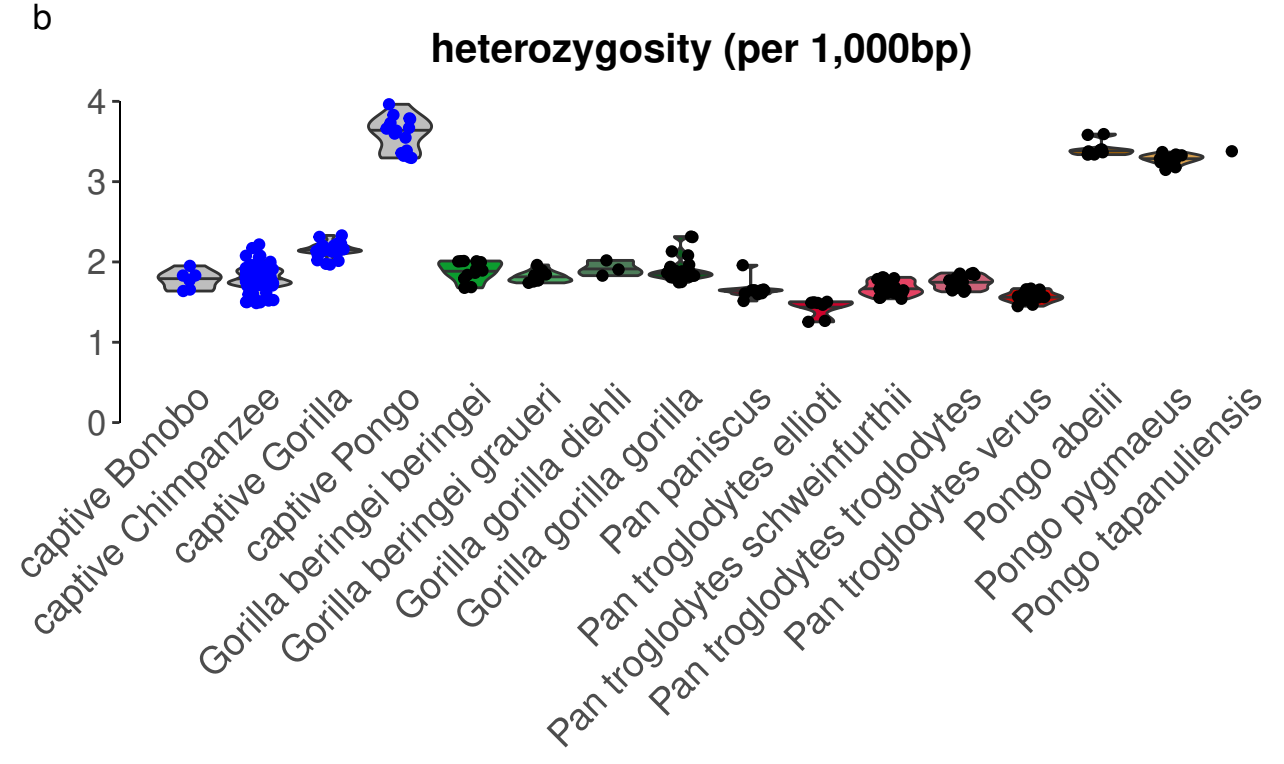
Figure 3. Basic population genetic characterisation of great ape individuals in this dataset. A) PCA for the *Pan* clade (chimpanzees and bonobos), B) PCA for the *Gorilla* clade, and C) PCA for the *Pongo* clade, all calculated on a filtered SNV callset. D) ADMIXTURE clustering for the *Pan*, *Gorilla* and *Pongo* clades (based on 1,000,000 random SNVs per clade). PPA, *Pan paniscus*; PTE, *Pan troglodytes ellioti*; PTS, *Pan troglodytes schweinfurthii*; PTT, *Pan troglodytes troglodytes*; PTV, *Pan troglodytes verus*; GBB, *Gorilla beringei beringei*; GBG, *Gorilla beringei graueri*; GGD, *Gorilla gorilla diehli*; GGG, *Gorilla gorilla gorilla*; PA, *Pongo abelii*; PP, *Pongo pygmaeus*; PT, *Pongo tapanuliensis*; CP, captive population.

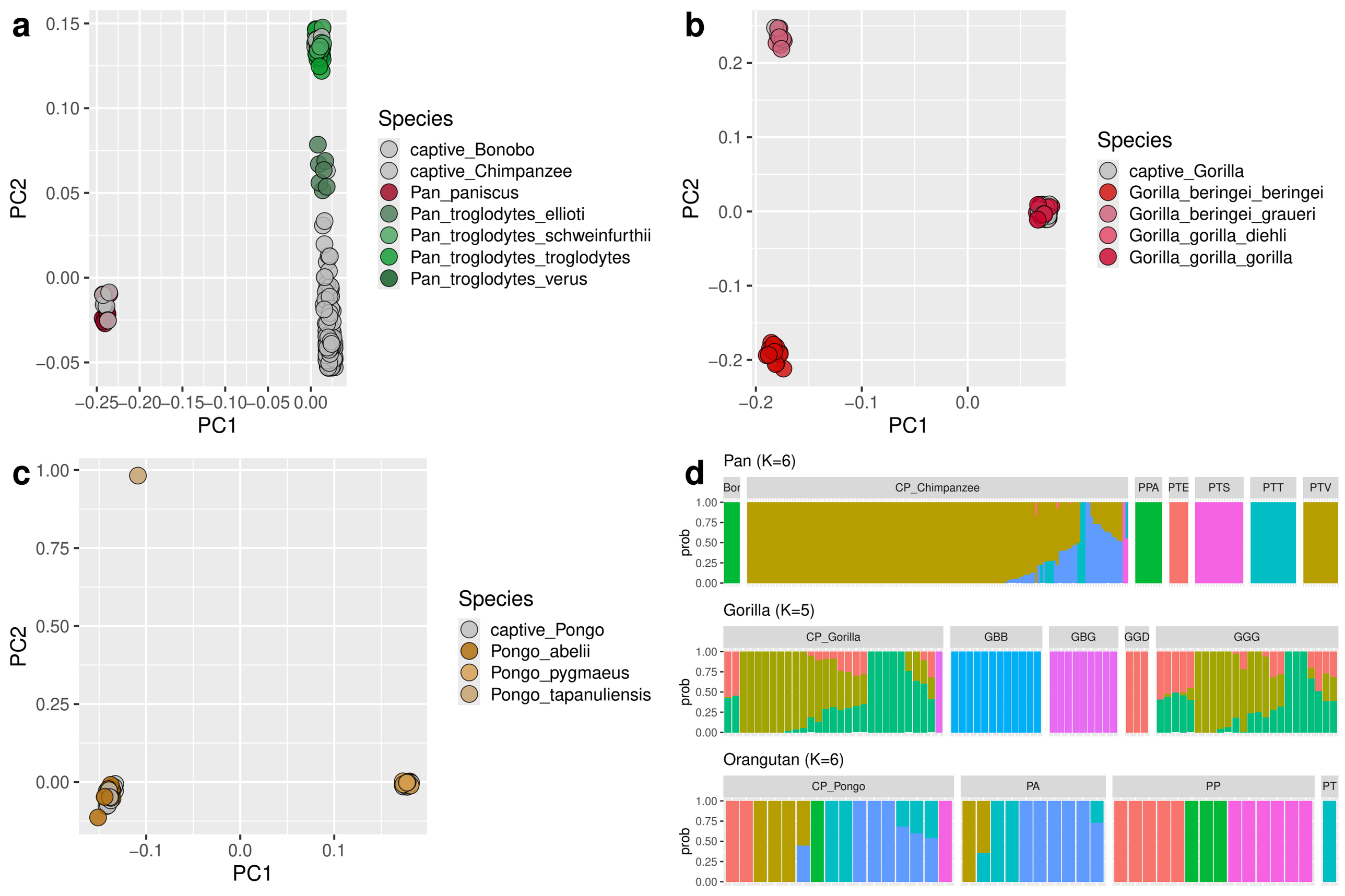
Figure 4. Characterizing the captive panel. A) Human contamination estimates in captive individuals. B) Subspecies assignment using f3-statistics for PD_0179. GBG, *Gorilla beringei beringei*; GBG, *Gorilla beringei graueri*; GGD, *Gorilla gorilla diehlii*; GGG, *Gorilla gorilla gorilla*. C) Geolocalization of a captive chimpanzee individual (PD_0259).

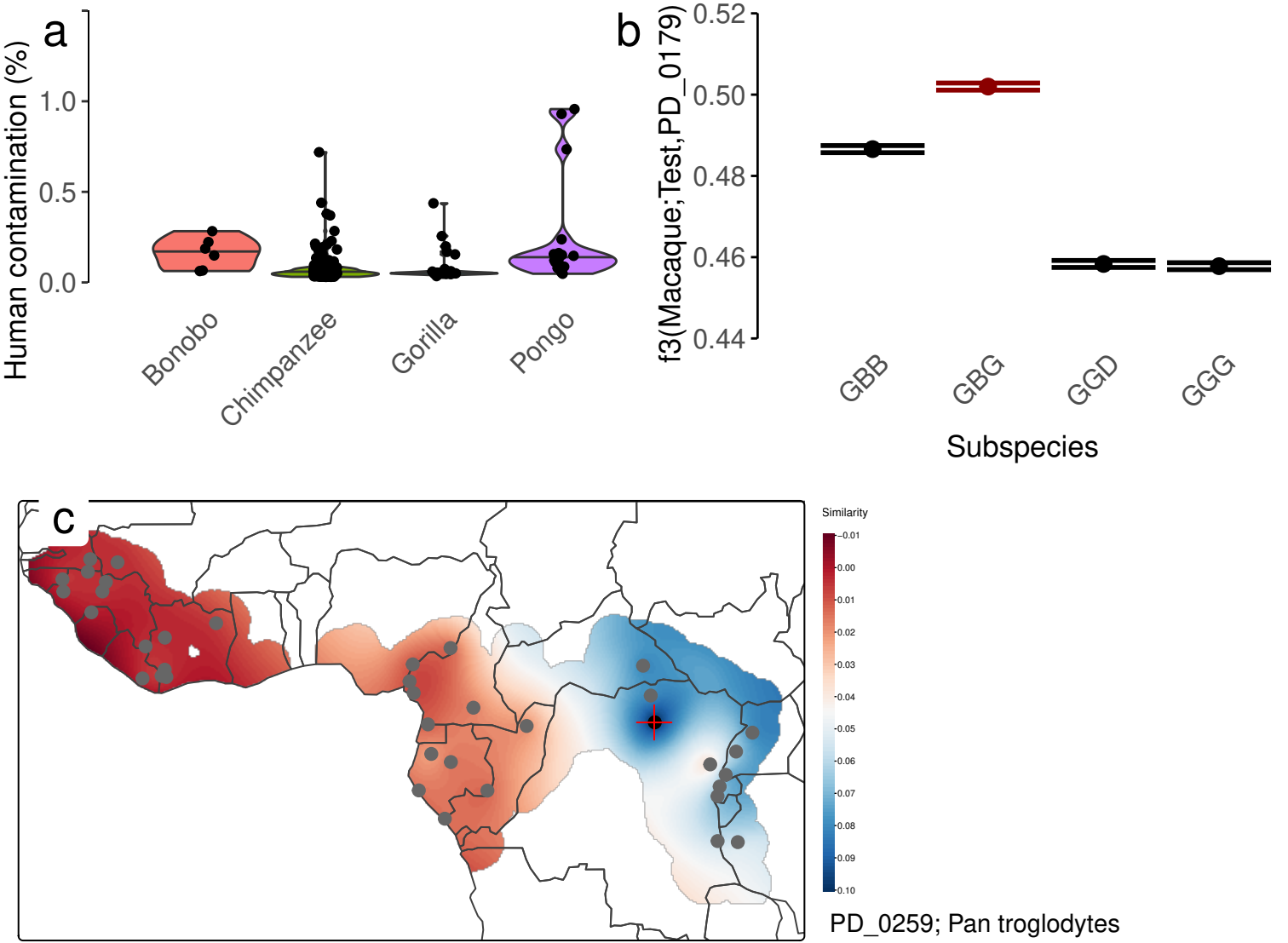
Figure 5. Kinship and runs of homozygosity. A) Relatedness among bonobos (*Pan paniscus*) as estimated by ngsRelate (KING method)[61], as an example of individual relationships. Individuals from Trio sequencing (Mhudiblu, Loretta, PR00251) are distinguishable. Plots for all clades are found in Supplementary Materials. B) Runs of homozygosity as estimated by bcftools roh[62] across all species and subspecies. For abbreviations see Table S2.


10+12
1
14+4
24+28
3
12
9+1
10+6
17+2
18+2
7+1
13+138
WLG
CRG
MG
ELG
Central
Eastern
NC
Western
Tapanuli
Bornean
Western
Eastern
Sumatran
Bonobo
Chimpanzee
Pan clade
Homo clade
Gorilla clade
Pongo clade
7-10 Mya
9-13 Mya
16-24 Mya

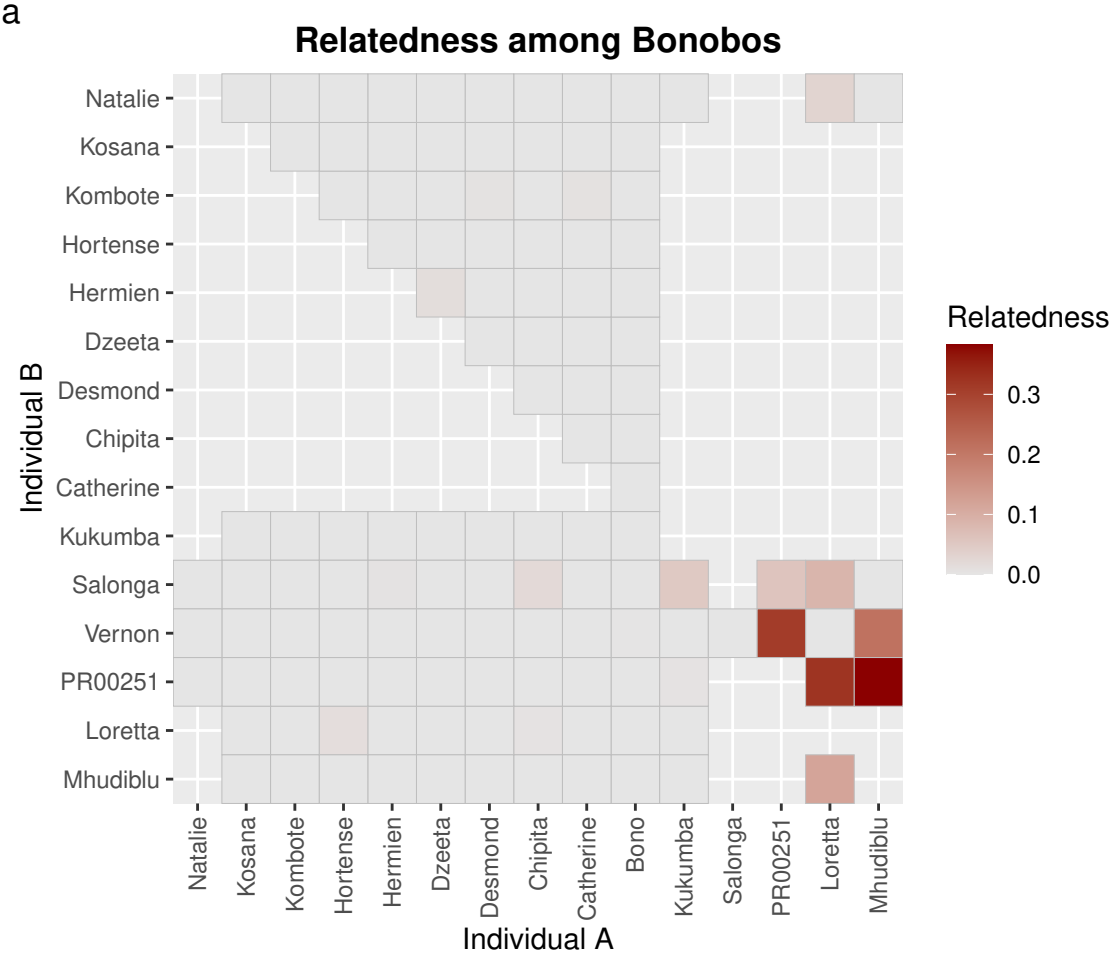

a
coverage (fold)
60
40
20
0
captive Bonobo
captive Chimpanzee
captive Gorilla
captive Pongo
Gorilla beringei beringei
Gorilla beringei graueri
Gorilla gorilla diehli
Gorilla gorilla gorilla
Pan paniscus
Pan troglodytes ellioti
Pan troglodytes schweinfurthii
Pan troglodytes troglodytes
Pan troglodytes verus
Pongo abelii
Pongo pygmaeus
Pongo tapanuliensis


b
heterozygosity (per 1,000bp)
4
3
2
1
0
captive Bonobo
captive Chimpanzee
captive Gorilla
captive Pongo
Gorilla beringei beringei
Gorilla beringei graueri
Gorilla gorilla diehli
Gorilla gorilla gorilla
Pan paniscus
Pan troglodytes ellioti
Pan troglodytes schweinfurthii
Pan troglodytes troglodytes
Pan troglodytes verus
Pongo abelii
Pongo pygmaeus
Pongo tapanuliensis

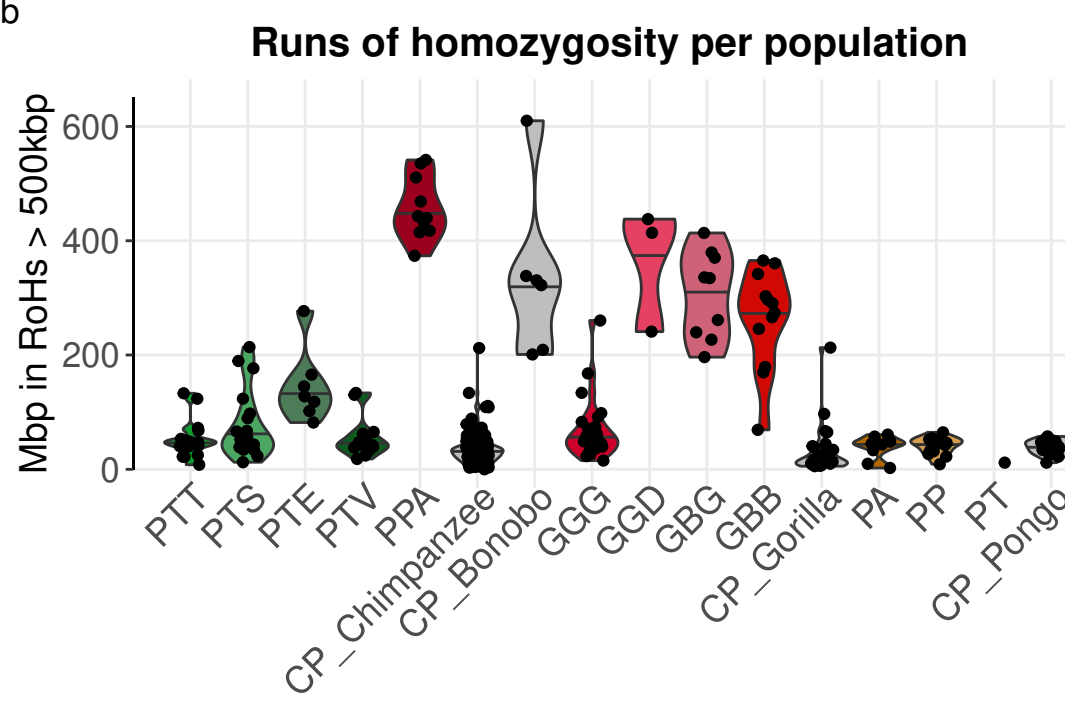

a
PC2
PC1
0.15
0.10
0.05
0.00
−0.05
−0.25 −0.20 −0.15 −0.10 −0.05 0.00
Species
captive_Bonobo
captive_Chimpanzee
Pan_paniscus
Pan_troglodytes_ellioti
Pan_troglodytes_schweinfurthii
Pan_troglodytes_troglodytes
Pan_troglodytes_verus
b
PC2
PC1
0.2
0.1
0.0
−0.1
−0.2
−0.2 −0.1 0.0
Species
captive_Gorilla
Gorilla_beringei_beringei
Gorilla_beringei_graueri
Gorilla_gorilla_diehli
Gorilla_gorilla_gorilla
c
PC2
PC1
1.00
0.75
0.50
0.25
0.00
−0.1 0.0 0.1
Species
captive_Pongo
Pongo_abelii
Pongo_pygmaeus
Pongo_tapanuliensis
d
Pan (K=6)
Bor
CP_Chimpanzee
PPA
PTE
PTS
PTT
PTV
prob
1.00
0.75
0.50
0.25
0.00
Gorilla (K=5)
CP_Gorilla
GBB
GBG
GGD
GGG
Orangutan (K=6)
CP_Pongo
PA
PP
PT


a
Human contamination (%)
1.0
0.5
0.0
Bonobo
Chimpanzee
Gorilla
Pongo
b
f3(Macaque;Test,PD_0179)
0.52
0.50
0.48
0.46
0.44
GBB
GBG
GGD
GGG
Subspecies
c
Similarity
0.01
0.00
-0.01
-0.02
-0.03
-0.04
-0.05
-0.06
-0.07
-0.08
-0.09
-0.10
PD_0259; Pan troglodytes

a


Relatedness among Bonobos
Natalie
Kosana
Kombote
Hortense
Hermien
Dzeeta
Desmond
Chipita
Catherine
Kukumba
Salonga
Vernon
PR00251
Loretta
Mhudiblu
Individual B
Natalie
Kosana
Kombote
Hortense
Hermien
Dzeeta
Desmond
Chipita
Catherine
Bono
Kukumba
Salonga
PR00251
Loretta
Mhudiblu
Individual A
Relatedness
0.3
0.2
0.1
0.0

b


Runs of homozygosity per population
Mbp in RoHs > 500kbp
600
400
200
0
PTT
PTS
PTE
PTV
PPA
CP _Chimpanzee
CP _Bonobo
GGG
GGD
GBG
GBB
CP _Gorilla
PA
PP
PT
CP _Pongo