doi:10.1038/nature12228

Great ape genetic diversity and population history
Supplementary Material

Section 1: Sample collection and sequencing

A total of 88 individuals (79 non-human and 9 human great apes) were collected and sequenced in this study encompassing all the species and subspecies of great apes except for mountain gorillas (*Gorilla beringei beringei*). These populations are summarized as follows: 9 humans (*Homo sapiens*) sampled from the HGDP panel[1]; 13 bonobos without known geographical origin (*Pan paniscus*); 25 chimpanzees covering from west to east Africa (10 *Pan troglodytes ellioti*, 6 *Pan troglodytes schweinfurthii*, 4 *Pan troglodytes troglodytes*, 4 *Pan troglodytes verus*, and 1 chimpanzee hybrid); 31 gorillas from Rwanda, Cameroon and Congo (3 *Gorilla beringei graueri*, 1 *Gorilla gorilla diehli,* and 27 *Gorilla gorilla gorilla*); and 10 Sumatran and Bornean orangutans (5 *Pongo abelii* and 5 *Pongo pygmaeus*). As the aim of this study was to assess the genomic variation among natural populations of great apes, our sampling criteria maximized wild-born individuals (77%) or the first generation of captive individuals (23%). Moreover, the samples were mostly obtained from blood with the exception of at least three samples coming from low passage cell lines (**Table S1**). All samples were sequenced on an Illumina sequencing platform (HiSeq 2000) with data production at four different sequencing centers; samples were collected under the supervision of ethical committees and CITES permissions were obtained when necessary.

Section 2: Mapping and SNP calling

Javier Prado-Martinez, Peter H. Sudmant, Jeffrey M. Kidd, Joanna L. Kelley, Dorina Twigg, Carlos D. Bustamante, Evan E. Eichler, Tomas Marques-Bonet

The goal of this study was to explore the wide spectrum of diversity in great ape populations and, thus, many of the analyses must be provided against a single reference genome. As the human reference is the best annotated primate genome, most analyses were performed with mappings to the human reference NCBI Build 36. We were aware that this could introduce biases, so we also mapped and called variants against the available nonhuman primate reference genomes. Analyses particularly sensitive to mapping were provided against these references.

2.1. Human reference mappings

Genomes were mapped to the human reference assembly NCBI Build 36 (UCSC hg18) using the BWA mapping software[2]. Read qualities were first converted/scaled to Sanger format then BWA paired-end mapping was performed using the BWA *aln* and *sampe* tools. All reads were mapped using the *aln* trim parameter *–q 15* and nonhuman genomes were additionally mapped with the increased edit distance parameter of *–n 0.01.* Pairing was performed using the *sampe* tool and limiting the maximum occurrences of a read for pairing to 1000 using the *–o 1000* option.

SNP calling was performed using the Genome Analysis Toolkit (GATK) software (version 1.4)[3]. First, samples combined by species were realigned around putative indels. SNP calling was then performed on the combined individuals for each species. SNPs were finally filtered if they met any of the following criteria:

```
DP < (mean_read_depth/8.0) || DP > (mean_read_depth*3)
QUAL < 33
FS > 26.0
-sites within 5 bp of a reported indel-
MQ < 25
MQ0 >= 4 && ((MQ0 / (1.0 * DP)) > 0.1)
```

To ensure contamination (**Suppl. Section 4**) would not contribute to any reported heterozygous calls, we applied an allele balance (AB) filter that theoretically removed the 10% of heterozygous calls with the most skewed allele balance assuming a binomial distribution of reads with *p=0.5* for the A or B alleles. Genome-wide assessments of heterozygous counts were corrected by a factor of *1/0.9*. Finally, sites overlapping predicted segmental duplications, as predicted by mrsFAST-based read-depth counts, were filtered.

We next identified all those base pairs called as a *reference* or *variant* versus those that were *uncallable* as a result of filtering thresholds (as defined by the above noted filtering parameters) to delineate the fraction of the genome that was *callable* in any particular species. We thus generated VCFs with all bases annotated and applied the filters described above. These so-called *uncallable* bases and segmentally duplicated base pairs were then combined into a mask that defined the *callable* versus the *uncallable* fraction of the genome. Ancestry informative markers, or AIMs, (**Table 1**) were defined as specific, fixed variants at the subspecies level.

2.2. Species reference mapping

Mapping to species reference assemblies

We assessed the limitations of calling variants based on mapping primate sequences to the human genome reference by undertaking an independent alignment and variant discovery process that utilized the chimpanzee[4], gorilla[5], and orangutan[6] reference genomes. The specifics of the analysis are described below, but each species was processed using the same basic pipeline.

For each species, Illumina reads were mapped to the corresponding species reference assembly using BWA[2] version 0.5.9 with dynamic quality trimming (-q 15) and default read alignment identity thresholds (-n 0.04). Paired-end placements were identified using BWA *sampe* (with -o 1000). We performed empirical base quality score recalibration for each sequencing lane using the GATK[3,7] version 1.2-65 and identified duplicate read pairs from each library using Picard version 1.62 (http://picard.sourceforge.net/). For each species, we performed indel realignment using GATK jointly across all samples and produced a preliminary SNP callset using the GATK Unified Genotyper.

From the resulting set of candidate SNPs, we identified a high-quality set of variants using the variant quality score recalibration (VQSR) procedure implemented in GATK[3]. This procedure utilizes a training set of known variant positions to define the characteristics of high-quality calls (based on a joint analysis of criteria such as total depth, mapping quality, strand balance, etc.) and identifies a subset of the candidate SNP set that meets the resulting criteria. Using the VQSR methodology, we selected a set of SNP positions such that 99% of the SNP positions in the training set were retained. For humans, the training set for the VQSR procedure is typically derived from sets of positions known to be variant based on SNP genotyping arrays, the HapMap Project, or the 1000 Genomes Project. Such a resource is not available for the nonhuman primates considered in this paper. Instead, we created set of training SNP positions that are independent of Illumina short-read data based on capillary sequence traces obtained for each species from the NCBI trace archive based on the whole-genome shotgun (WGS) reads produced for the species reference assemblies. We mapped capillary reads using ssaha2[8] and called SNPs using the neighborhood quality score criteria implemented in ssahaSNP.

To limit the impact of segmental duplications and copy number variants in the nonhuman primate species, we further constrained the training set to those SNPs with a unique mapping to the human genome (hg18, based on the UCSC liftOver program) and that did not overlap with duplications identified in humans, chimpanzees, gorillas, or orangutans.

All analyses were limited to the autosomes. The VQSR step utilized the following parameters:
-resource:capillary,known=false,training=true,truth=true,prior=12.0
-an QD -an HaplotypeScore -an MQRankSum -an ReadPosRankSum -an MQ -an FS -an DP

After identifying sites such that 99% of the training set positions were retained, we created phased and imputed individual genotypes based on the Unified Genotyper output using BEAGLE[3,9] version 3.3.2. For some analyses, we masked out individual heterozygous genotypes that failed an AB filter with a two-tailed binomial p-value less than 0.05.

Gorilla

For gorilla, we used the gorGor3 gorilla genome assembly available as part of Ensembl release 62. We created a VQSR training set using 8,308,425 gorilla capillary WGS reads obtained from the NCBI trace archive:

SPECIES_CODE = 'GORILLA GORILLA' and TRACE_TYPE_CODE = 'WGS'

and limited the training set to 1,539,968 SNP positions with a unique liftOver to the human genome (hg18, NCBI Build 36) that did not overlap with any segmental duplications.

Chimpanzee and Bonobo

For chimpanzee and bonobo, we used the panTro-2.1.4 assembly obtained from Ensembl release 65. We created a VQSR training set using 20,596,701 chimpanzee capillary WGS reads obtained from the NCBI trace archive:

CENTER_NAME = 'BI' and SPECIES_CODE = 'PAN TROGLODYTES' and CENTER_PROJECT = 'G591'
and
SPECIES_CODE = 'PAN TROGLODYTES' and CENTER_NAME = 'WUGSC' and strategy = 'WGS'

and limited the training set to 1,287,455 SNP positions with a unique liftOver to the human genome (hg18, NCBI Build 36) that did not overlap with any segmental duplications.

Orangutan

For orangutan, we used the ponAbe2 assembly obtained from the UCSC genome browser. We created a VQSR training set using 26,569,515 orangutan capillary WGS reads obtained from the NCBI trace archive:

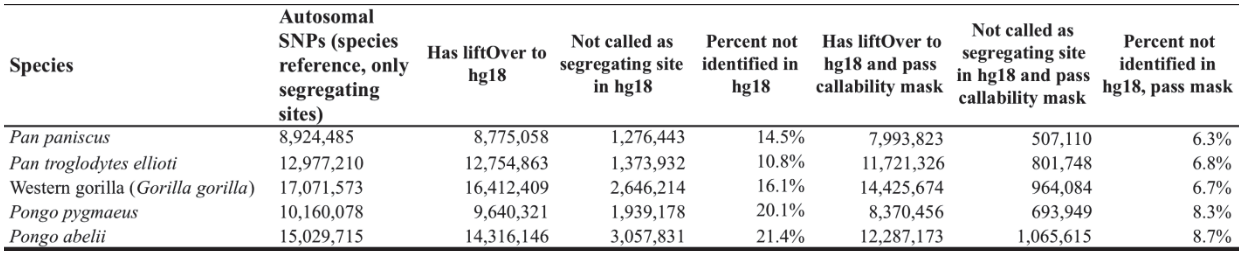
SPECIES_CODE = 'PONGO ABELII' and TRACE_TYPE_CODE = 'WGS'

and limited the training set to 2,706,869 SNP positions with a unique liftOver to the human genome (hg18, NCBI Build 36) that did not overlap with any segmental duplications.

Comparison to hg18 SNP positions

We used the species reference SNP callset to estimate the number of SNPs missed due to mapping to the human reference, which is diverged from each nonhuman primate species. In this analysis, we did not consider SNPs as missing because of lineage-specific deletions, which result in sequences that are absent from the human genome reference.

To avoid confounding calls due to fixed differences, we limited analysis to autosomal variants that were identified as polymorphic among the analyzed samples. Only the positions of segregating sites, not individual genotypes or allele frequencies, were considered. First, we identified all segregating sites in the species reference mappings that liftOver to hg18 (**Suppl. Table 2.2.1**). We find that 10.8%–21% of the segregating sites identified from the species reference mappings with the VQSR procedure are not called as variable based on mapping to hg18. However, the hg18 callset includes hard filters as well as an explicit mask of regions of the genome that are not callable. When we limit the analysis only to the regions of hg18 where reliable calls are reported for each species, we find that 6.3%–8.7% of the segregating sites identified in the species reference mappings are not found to be polymorphic based on the hg18 mappings. We note that orangutan, which has the highest divergence from human, shows the highest rate of missing SNPs. The SNPs identified from the species reference mappings contain a mixture of true and false positive sites, and we thus take the values in **Suppl. Table 2.2.1** as an upper bound on the rate of missing variation. Overall, however, we estimate that less than 9% of variable positions are missed when consideration is limited to the portion of the hg18 genome reference where reliable calls can be made. In the opposite direction, around 15% of the sites are called in the human mappings and not in the species-specific mappings.



Species	Autosomal SNPs (species reference, only segregating sites)	Has liftOver to hg18	Not called as segregating site in hg18	Percent not identified in hg18	Has liftOver to hg18 and pass callability mask	Not called as segregating site in hg18 and pass callability mask	Percent not identified in hg18, pass mask
Pan paniscus	8,924,485	8,775,058	1,276,443	14.5%	7,993,823	507,110	6.3%
Pan troglodytes ellioti	12,977,210	12,754,863	1,373,932	10.8%	11,721,326	801,748	6.8%
Western gorilla (*Gorilla gorilla*)	17,071,573	16,412,409	2,646,214	16.1%	14,425,674	964,084	6.7%
Pongo pygmaeus	10,160,078	9,640,321	1,939,178	20.1%	8,370,456	693,949	8.3%
Pongo abelii	15,029,715	14,316,146	3,057,831	21.4%	12,287,173	1,065,615	8.7%

Suppl. Table 2.2.1 – *Identification of segregating sites called based on mapping to the species references that are not called based on mapping to the human genome reference.*

Species	Autosomal SNPs (hg18, only segregating sites)	Has liftOver to species reference (autosomes)	Not called as segregating site in species reference	Percent not identified in species reference
Pan paniscus	8,950,002	8,615,793	1,128,532	13.1%
Pan troglodytes ellioti	13,715,319	13,162,422	1,795,182	13.6%
Western gorilla (*Gorilla gorilla*)	17,217,951	16,279,942	2,548,820	15.7%
Pongo pygmaeus	10,321,213	9,540,296	1,893,039	19.8%
Pongo abelii	14,543,573	13,475,311	2,284,908	17.0%

Suppl. Table 2.2.2 – *Identification of segregating sites called based on mapping to hg18 that are not called based on mapping to the species references.*

2.3. Ancestral allele calls and variant orientation

Asger Hobolth, Marta Mele, Anders E. Halager, Thomas Mailund

To call ancestral alleles and orient variants present in the extant groups into ancestral and derived alleles, we employ Felsenstein's pruning algorithm to compute probability distributions for alleles at the inner nodes in the phylogeny and weigh these with population frequencies.

For each possible allele at the tips of the phylogeny, we built a table of the posterior probabilities at the inner nodes. Given allele counts for a site, we then compute a weighted average of posterior probabilities from this table and use this as the basis for the ancestral allele call algorithm. To avoid that varying number of calls have a large effect on the weighted posteriors, we use a pseudo count for the weighting, and we group the two *Gorilla gorilla* subspecies together as the Cross River gorilla would weigh too much otherwise.

Allele calling algorithm

The first step after computing the weighted posteriors is to classify inner nodes as either polymorphic or monomorphic.

For this, we first assume that an inner node can have at most two different alleles and then consider the two alleles with the highest and second highest posterior probability. We use the second highest posterior probability to determine if we consider the node monomorphic or polymorphic; if the second highest probability falls below a threshold, we consider the node monomorphic for the highest probability allele, and otherwise we consider it polymorphic for the two most probable alleles.

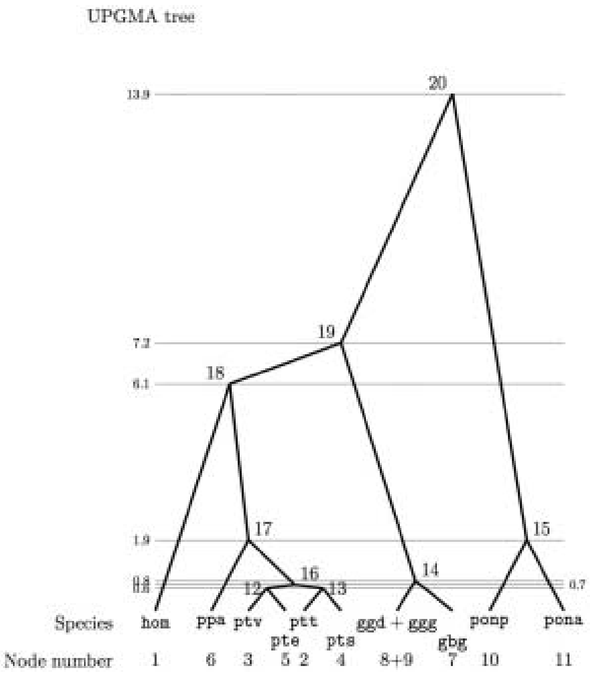
The threshold used in the monomorphic/polymorphic classification depends on the depth of the node in the tree and was determined by careful inspection of outputs from the algorithm with different choices of thresholds. The nodes within common chimpanzees (inner nodes 12, 13, and 16) have a threshold of 1%, the within-genera nodes (inner nodes 14, 15, and 17) have a threshold of 5%, and the inter-genera nodes (inner nodes 18, 19, and 20) have a threshold of 10% (**Suppl. Figure 2.3.1**).

One exception to using the posterior probabilities to determine if a node is polymorphic is when both children of a node are called as monomorphic with the same allele. In this case, we always call the parent as monomorphic.

This, generally, calls the alleles at inner nodes, but the thresholds are chosen so they are likely biased to call monomorphic nodes as polymorphic rather than the other way around, since this ensures that the allele orientation algorithm (described below) will be conservative.

Orienting polymorphisms

To orient polymorphisms, we search up the tree from each polymorphic leaf. Ancestral nodes with the same polymorphism are just stepped over. If we reach a monomorphism, the allele in this node can either be one of the alleles in the polymorphism, in which case we call that allele as ancestral, or it can be a third allele, in which case we again cannot orient the polymorphism. If the search up the tree reaches an ancestral node with a different polymorphism, we will not be able to orient the polymorphism and we give up. If we reach the root of the tree without seeing a monomorphism, we also cannot orient the polymorphism and we give up. For the root of the tree, if African and Asian apes have different alleles, we orient the root using the macaque genome.



Suppl. Figure 2.3.1 – *UPGMA phylogenetic tree. To provide the proper orientation and classification of the internal nodes, distances and different weights were used.*

Section 3: Validation

Peter H. Sudmant, Javier Prado-Martinez, Carl Baker, Maika Malig, Jessica Hernandez-Rodriguez, James C. Mullikin, Tomas Marques-Bonet, Evan E. Eichler

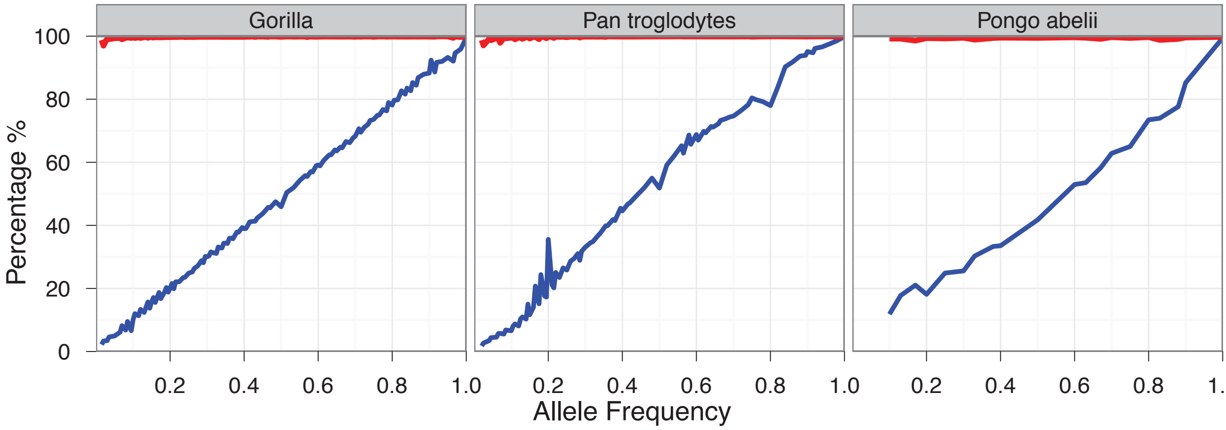
3.1. Reference fixed sites

The first evaluation of our variation was made comparing to the available reference genome assemblies[4–6]. We downloaded the genome alignments to the human NCBI Build 36 reference and the nonhuman primate genomes from the UCSC database and retrieved all annotated variants. As these alignments include many short and low-quality read alignments that lead to an excess of variation between the assemblies, we only considered those alignments longer than 1 Kbp (**Suppl. Table 3.1**). It is important to note that these variants do not only correspond to fixed events between the lineages and a significant proportion of these variants are segregating within the populations of great apes.

Reference Genome	Callable (1 Kbp alignments)	Reference SNV (compared to hg18)	Variants per bp	Overlap in Fixed Variants	Concordance in Fixed Variants
Pan troglodytes (Pantro2)	2,559,497,801 bp	34,178,305	0.0134	99.86%	99.89%
Gorilla gorilla (Gorgor3)	2,530,410,350 bp	42,055,936	0.0166	99.5%	99.87%
Pongo abelii (PonAbe2)	2,471,794,228 bp	84,759,693	0.0343	99.44%	99.76%

Suppl. Table 3.1 – *Summary table of the variants found in the reference alignments compared to the human reference genome (hg18). The overlap with the fixed variants corresponds to the percentage of our variants that are also found in the reference alignments. The concordance is the rate of variants that overlap with the reference and have the same variant allele.*

As expected, given that the references mostly consist of a single individual, the intersection of variants called from our dataset and those in the reference genomes increases as a function of the allele frequency (**Suppl. Figure 3.1**), reaching the highest overlap at fixed variants. The percentage of variants concordant with our callset varies between 99.44%–99.86% (orangutan-chimpanzee), commensurate with what we would expect given the divergence with respect to the human reference genome. We note that despite the fact that the percentage of sites in our dataset intersecting variants in the reference genomes varies widely depending on the allele frequency, the allele concordance in the variants that are found from both sources is over 99% irrespectively of the allele frequency.



Suppl. Figure 3.1 – *Comparison of the allele frequency to the percentage of overlap (blue) and the allele concordance (red) with the reference alignments. The overlap between the variants present in the reference genomes (Gorgor3, Pantro2, and Ponabe2) with respect to the variation in the different populations increases as a function of the frequency of the variants in the species. In contrast, the variants that overlap in both sources have a high concordance of the alleles that are found, meaning that when a variant is present in both the reference and our sequencing data they agree in the derived allele found. The increase in the percent of overlap in chimpanzees around 20% corresponds to the Pan troglodytes verus samples, corresponding to the subspecies used in the chimpanzee assembly.*

3.2. Validation of segregating variants by Sanger

We also performed Sanger capillary sequencing on a subset of ~480 random variants (**Suppl. Table 3.2**). Two individuals from each species were selected to be tested for a random subset of heterozygous and homozygous sites (80% heterozygotes and 20% homozygotes). Both forward and reverse strands were then sequenced. Though we obtained low validations rates for the two individuals Abe and Tzambo, likely due to poor DNA quality, our overall genotype concordance was >96%.

			Correct Calls			Incorrect Calls				
Species	**Sample**	**Sites**	**Het**	**Hom**	**Hom Ref**	**Het**	**Hom**	**Hom Ref**	**%correct**	**%correct variant only**
Chimp	Koto	61	33	17	8	1	1	1	95.1	96.0
Chimp	Vincent	68	40	9	17	1	1	0	97.1	95.9
Gorilla	Abe	54	23	16	6	8	0	1	83.3	79.5
Gorilla	Tzambo	59	34	13	6	4	2	0	89.8	87.2
Gorilla	Kokamo	77	34	13	28	1	0	1	97.4	97.9
Gorilla	Azizi	34	28	6	0	1	0	0	97.1	97.1
Orangutan	Dunja	27	12	7	6	1	1	0	92.6	89.5
Orangutan	Napoleon	23	10	7	6	0	0	0	100.0	100.0
Bonobo	Dzeeta	38	18	10	10	0	0	0	100.0	100.0
Bonobo	Hermien	38	19	10	7	2	0	0	94.7	93.1
Total/Median		**479**	**251**	**108**	**94**	**19**	**5**	**3**	**96.1**	**96.0**

Suppl. Table 3.2 – *We performed 479 Sanger sequencing experiments to validate our Illumina sequencing-based SNP calls. Overall, our median per individual concordance at variant sites was 96% confirming our SNP calls to be of high quality.*

3.3. HGDP SNP arrays

The human samples used in this study are part of the HGDP panel and were previously analyzed using Illumina 650Y SNP arrays (http://hagsc.org/hgdp/files.html). This resource allows us to compare the quality of our variant calls genome-wide. This is especially useful in the characterization of loss of heterozygotes as a consequence of the AB filters (**Suppl. Section 2.2**) applied in the contamination correction. This filter should theoretically remove 10% of the heterozygous calls, but with the SNP arrays we can assess the true overall effect of this filter. We find a general reduction of ~12%, slightly higher than the theoretical estimation, but this may also account for false heterozygous calls with skewed allele balance and additionally the HGDP samples assessed in this study were of lower coverage than the nonhuman primates, making them more susceptible to AB filtering. We also estimated the total sensitivity of both homozygous and heterozygous positions varying from the reference. Before AB filtering, we obtained a sensitivity of 99.5% compared to the SNP array while this number dropped to ~93% after the filter. Finally, we determined there to be a >99.5% overall genotype concordance in our callset. (**Suppl. Table 3.3**)

		Before Allele Balance Filter			After Allele Balance Filter		
Sample	Coverage	Heterozygous Concordance	Non-reference Sensitivity	Genotype Concordance	Heterozygous Concordance	Non-reference Sensitivity	Genotype Concordance
San HGDP01029	28.56	99.48	99.58	99.88	85.16	92.24	99.92
Han HGDP00778	21.91	99.52	99.59	99.89	87.78	93.01	99.94
French HGDP00521	21.07	99.50	99.57	99.86	88.03	92.61	99.91
Mandenka HGDP01284	19.85	99.40	99.54	99.70	87.34	92.34	99.85
Sardinian HGDP00665	19.68	99.43	99.55	99.83	87.93	92.61	99.88
Dai HGDP01307	16.82	99.23	99.49	99.77	87.48	92.86	99.83
Karitiana HGDP00998	15.56	98.91	99.44	99.62	86.88	93.68	99.78
Papuan HGDP00542	15.03	98.84	99.42	99.64	87.27	93.83	99.75
Mbuti HGDP00456	14.30	98.72	99.35	99.50	87.70	93.43	99.63
All	19.20	99.23	99.50	99.74	87.29	92.96	99.83

Suppl. Table 3.3 – *Effect of allele balance in the human samples based on SNP array comparison.*

3.4. Heterozygous variants in Clint

To provide further validation in the heterozygous calls genome-wide in a nonhuman primate, we made use of the chimpanzee reference WGS data (Clint)[4]. We downloaded all the data produced from WGS sequencing (ftp://ftp.ncbi.nih.gov/pub/TraceDB/pan_troglodytes/) and mapped to it to Pantro2 with the following parameters:

```
ssahaSNP des_qual 23 maxSNPs/1 Kbp 45 maxDepth 10 Qne 15 Nne 6 maxNdiff 2
```

Among the data used for this assembly, the primary donor was Clint; however, other chimpanzees contributed, most notably Donald. We filtered out the reads from Donald and infer the variants using the coverage and allele frequency criteria; a variant was only considered if the region was covered between 6 and 10 NQS (neighborhood quality standard) with an allele frequency ranging between 0.3 and 0.7 in order to call heterozygous variants. Variants called against the chimpanzee genome were lifted over to the human reference genome were discarded if they did not fall within the callable fraction determined for chimpanzee Illumina sequence mapping. In total, 428,022 heterozygous variants were identified from this procedure.

We compared these heterozygous positions to the Illumina calls before and after the AB filter and obtained similar performances to those with the HGDP SNP arrays. Before this filter the rate of validation was 94.3%, a number that appears low, but two factors may influence this rate: first, the liftOver process (which may produce some misalignments across species) and second, some of the variants may be derived from related sequencing projects and not necessarily from the individual Clint. After the AB filtering, the concordance in heterozygotes dropped to 86.0%, a change similar to that observed with the HGDP SNP arrays.

3.5. Indel validation

The quality of the indels (**Suppl. Section 12**) was first assessed by comparing indel variants to reference genome alignments, as previously with the fixed SNVs. Analyses were limited to 1 Kbp alignments between the references. As the alignment algorithms can place the indels with some variation around a small region, we screened for an indel of the same length and within 20 bp of the prediction. With this method, we were able to validate computationally ~97% of the variants that are predicted to damage human gene models. We further validated a random subset of 111 indels with Sanger capillary sequencing and confirmed 110 of 111 events (99.1%), representing a very low FDR (<1%) (**Suppl. Table 3.4**).

Method	Sample	Sites	Incorrect Calls	Correct Calls	%Correct
Sanger Validations	Chimpanzee	16	0	16	100.00
	Bonobo	20	0	20	100.00
	Gorilla	23	1	22	95.65
	Orangutan	52	0	52	100.00
	All	111	1	110	99.10
Reference Alignments	Chimpanzee	732	14	718	98.09
	Gorilla	813	41	772	94.96
	Orangutan	1584	45	1539	97.16
	All	3129	100	3029	96.80

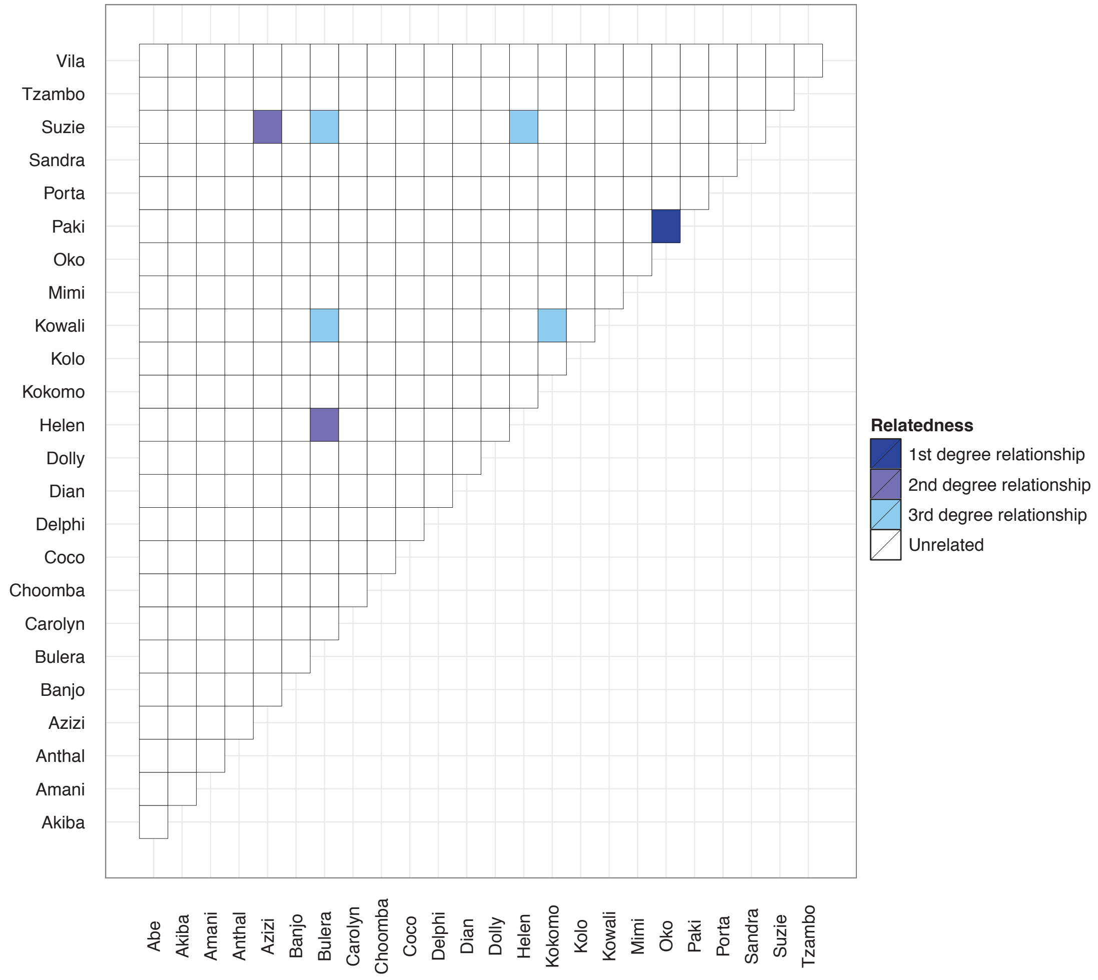
Suppl. Table 3.4 – *Indel validation rates with Sanger sequencing validation and comparison with the reference alignments.*

Section 4: Kinship and contamination

Javier Prado-Martinez, Peter H. Sudmant, Jeffrey M. Kidd, Joanna L. Kelley, Evan E. Eichler, Tomas Marques-Bonet

4.1. Kinship among samples

Some individuals sequenced in this study are offspring of wild-caught individuals that were bred in zoos (**Table S1**). The software KING was used to estimate kinship coefficients between all the samples[10]. As expected, no hidden relatedness was identified within any of the chimpanzees, orangutans, humans, or bonobos sequenced because of the criteria of selection. Among the gorillas, the relationship between Helen and Bulera was correctly identified (**Suppl. Figure 4.1.1**) and additionally we identified a 2nd degree relationship between Azizi and Suzie, a 1st degree relationship between Paki and Oko, and 3rd degree relationships between Kowali, Bulera and Kokamo. We revised the gorilla studbook relationships with our findings and discarded the related individuals Bulera, Kowali, Suzie and Oko from all population genetic analyses.



Suppl. Figure 4.1.1 – *The degree of relatedness is plotted between all pairs of Western lowland gorillas as estimated from the coefficient of kinship.*

4.2. Contamination assessment

During the SNP quality checks we detected a significant difference in the amount of singletons in some samples when comparing the human mappings and the species-specific mappings. The larger divergence between mitochondrial genomes (4% between bonobo and chimpanzee and 14% between bonobo and orangutan mitochondria) combined with the higher mtDNA coverage allowed us to study the extent contamination. We applied two different methods to determine inter-species contamination and intra-species contamination.

For interspecific contamination we mapped all WGS data to all available mitochondrion sequences of all great apes (human, bonobo, Western chimpanzee, Western gorilla, and Bornean and Sumatran orangutans) and we recorded the unique best-quality mappings to each mitochondrion. Then, we computed the ratio of coverage between the endogenous mtDNA and the contaminant sample that could be translated into the percentage of contamination (**Suppl. Table 4.2.1**).

			Inter-species Contamination				Intra-species Contamination		
Sample	**Sample Autosomic Coverage**	**Ratio mtDNA/ Autosomic**	**Sample mtDNA Coverage**	**Contami-nation From**	**Contaminated Coverage**	**% of mtDNA Contamination**	**Contamination from**	**Fraction of Overlap**	**% of mtDNA Contamination**
Gbg-A929_Kaisi*	41.3	0.56	23	Ppa	7	23.333			
Ggg-A932_Mimi	38.7	27.80	1076	Ppa	1	0.093	Gbg-A929_Kaisi	0.833	1.52
Ggg-A934_Delphi	40.1	38.25	1534	Ppa	1	0.065			
Ppa-A927_Salonga	28.7	55.64	1597	None	0	0.000	Ppa-Catherine	0.918	2.21
Ppa-A928Kumbuka	43.2	35.60	1538	Ggg	10	0.646	Ppa-Catherine	0.846	2.36
Pts-100040_Andromeda	23.2	10.26	238	Ppa	2	0.833			
Pts-A911_Kidongo	49.8	2.27	113	Ppa	1	0.877			
Pts-A912_Nakuu	46.4	7.67	356	Ppa	1	0.280			
Ptt-A957_Vaillant	35	3.57	125	Ppa	1	0.787	Ptv-A956_Jimmie	0.818	2.51
				Pab	1	0.787			
Ptt-A958_Doris	39.4	2.46	97	Pab	1	1.020			
Ptv-9730_Donald	21.731	38.01	826	None	0	0.000	Ptv-9668_Bosco	0.800	1.78
Ptv-A956_Jimmie	31.7	27.51	872	Pab	8	0.909			
Ptv-X00100_Koby	39.3	118.51	4657.5	Ggg	6	0.129			
Pab-A947_Elsi	39.8	56.51	2249	Ppa	2	0.089	Pab-A949_Dunja	0.918	2.07
Pab-A948_Kiki	34.1	69.65	2375	Ppy	1	0.042			
Pab-A949_Dunja	41.1	96.23	3955	Ppa	10	0.252			
Ppy-A940_Temmy	29.2	76.99	2248	Pab	1	0.044			
Ppy-A941_Sari	32.3	35.54	1148	Pab	1	0.087			
Ppy-A943_Tilda	37.7	75.41	2843	Pab	38	1.319			
Ppy-A944Napoleon	36.8	65.52	2411	Ppa	1	0.040			
				Pab	29	1.188			

Suppl. Table 4.2.1 – *Summary of low contaminated samples used in the study.*

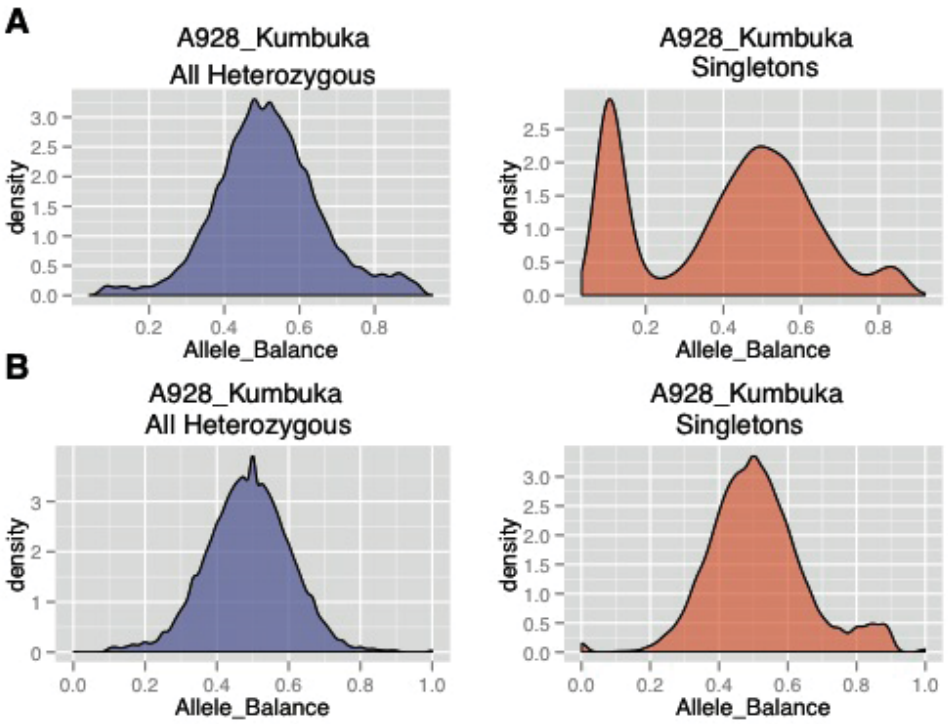
**Gbe-Kaisi: This sample show only 23X of mitochondrial coverage because it was mapped against the Western lowland gorilla mtDNA. As a result, the estimation of the contamination seems very high because only 7X of mitochondrial coverage is found.*

Assuming that the Eastern gorillas have a similar proportion of mitochondrial coverage as the Western lowland gorillas (~1000X), this estimation would be 0.7%.

To assess intraspecific contamination, we first called the mitochondrion variation to the specific references supported by the majority of reads (>90%) and the variants that had low frequency (between <50% of t2he reads). We considered the former variants as real mutations and the latter could be considered as contaminants, heteroplasmy, as well as sequencing errors. We then intersected variants with low support in one species to variants with high support in other species, which may be indicative of cross-contamination (**Suppl. Table 4.2.1**). This analysis is more ambiguous and presents biases, such as a reference effect (that can mask contamination from samples that do not have variants with respect to the reference), and possible heteroplasmy in those samples is not considered, so this number should be considered an overestimate. After this analysis, only a few samples showed traces of cross-contamination. We removed samples with higher levels of contamination (>2%), although there was a high heterogeneity in mitochondrial coverage due to different sample sources (cell lines had a higher mitochondrial coverage than blood samples) and due to the reference effect that lowered the coverage as a result of stringent mapping parameters.

We also used the autosomal portion of the genome to study the proportion of alleles present from each copy in heterozygous calls. We observed that samples with traces of contamination appear to have skewed allele balance distributions compared to the rest of the samples, where we expect a normal distribution centered on 50%. Indeed, the samples with inter-species contamination had a skewed distribution mostly in singletons (**Suppl. Figure 4.2.1**).

We corrected this problem by applying filters to the allele distribution of heterozygous variants. We applied a 0.1 two-tailed probability filter to the expected binomial distribution as a function of coverage. This conservative filter reduced the peaks in the tails of the distributions as a result of the elimination of variants with skewed allele balance, a large proportion of which indicative of contamination. However, a proportion of real heterozygous variants have been affected by this strict filtering (~10%) (**Suppl. Figure 4.2.1**).



Suppl. Figure 4.2.1 – *Effect of the filters on allele balance distribution in the heterozygous calls. This sample (Kumbuka - Bonobo), with traces of gorilla in the sequencing, shows an increase of singletons with skewed allele balance (A). These peaks are significantly reduced with the application of the allele balance (AB) filter (B).*

Section 5: Divergence

Peter H. Sudmant, Javier Prado-Martinez, Tomas Marques-Bonet, Evan E. Eichler

We initially explored the genetic relationships between individuals assessed in our study by constructing phylogenies based on the autosomal variant calls and *de novo* assembled mitochondrial genomes (**Suppl. Section 7 – Mitochondrial reconstruction**). The autosomal phylogenies were constructed from the consensus of genetic distance based neighbor-joining trees from all 10 Mbp subsegments of the genome. Consensus trees were also calculated for 1, 5, and 20 Mbp windows with little change in the overall topology of the tree. Divergence was estimated between all pairs of individuals using only sites callable among all species, defined as: (2**homs*+*hets*)/(2*callable fraction of the genome), where *homs* refers to the number of homozygous bases differing between the two individuals and *hets* the number of heterozygous sites. Divergence between two populations was thus computed as the mean pairwise divergence between all pairs of individuals between the two populations.

Homo sapiens non-African	*Homo sapiens African*	*Pan troglodytes ellioti*	*Pan troglodytes schweinfurthii*	*Pan troglodytes troglodytes*	*Pan troglodytes verus*
0.0009	0.0011	0.0121	0.0121	0.0120	0.0121
Pan paniscus	*Gorilla gorilla gorilla*	*Gorilla gorilla diehli*	*Gorilla beringei graueri*	*Pongo abelii*	*Pongo pygmaeus*
0.0122	0.0157	0.0156	0.0158	0.0303	0.0305

Suppl. Table 5.1 – *Mean divergence estimates between populations sequenced in this study and the human reference genome*.

	Western Gorilla	Bonobo	Eastern Chimpanzee	Bornean Orangutan	Sumatran Orangutan	Western Chimpanzee	Non-African Human	Nigeria-Cameroon Chimpanzee	Central Chimpanzee	African Human	Eastern Gorilla
Bonobo	0.0138										
Eastern Chimpanzee	0.0138	0.0035									
Bornean Orangutan	0.0287	0.0288	0.0288								
Sumatran Orangutan	0.0286	0.0287	0.0287	0.0032							
Western Chimpanzee	0.0138	0.0036	0.0019	0.0288	0.0287						
Non-African Human	0.0141	0.0117	0.0117	0.0291	0.0290	0.0116					
Nigeria-Cameroon Chimpanzee	0.0138	0.0036	0.0018	0.0288	0.0287	0.0017	0.0117				
Central Chimpanzee	0.0137	0.0035	0.0017	0.0287	0.0286	0.0019	0.0116	0.0018			
African Human	0.0141	0.0117	0.0116	0.0291	0.0290	0.0116	0.00095	0.0117	0.0116		
Eastern Gorilla	0.0020	0.0138	0.0138	0.0287	0.0286	0.0138	0.0141	0.0138	0.0137	0.0140	
Cross River Gorilla	0.0016	0.0137	0.0137	0.0286	0.0286	0.0137	0.0140	0.0138	0.0136	0.0140	0.00199

Suppl. Table 5.2 – *Pairwise species and subpopulation genetic distances.*

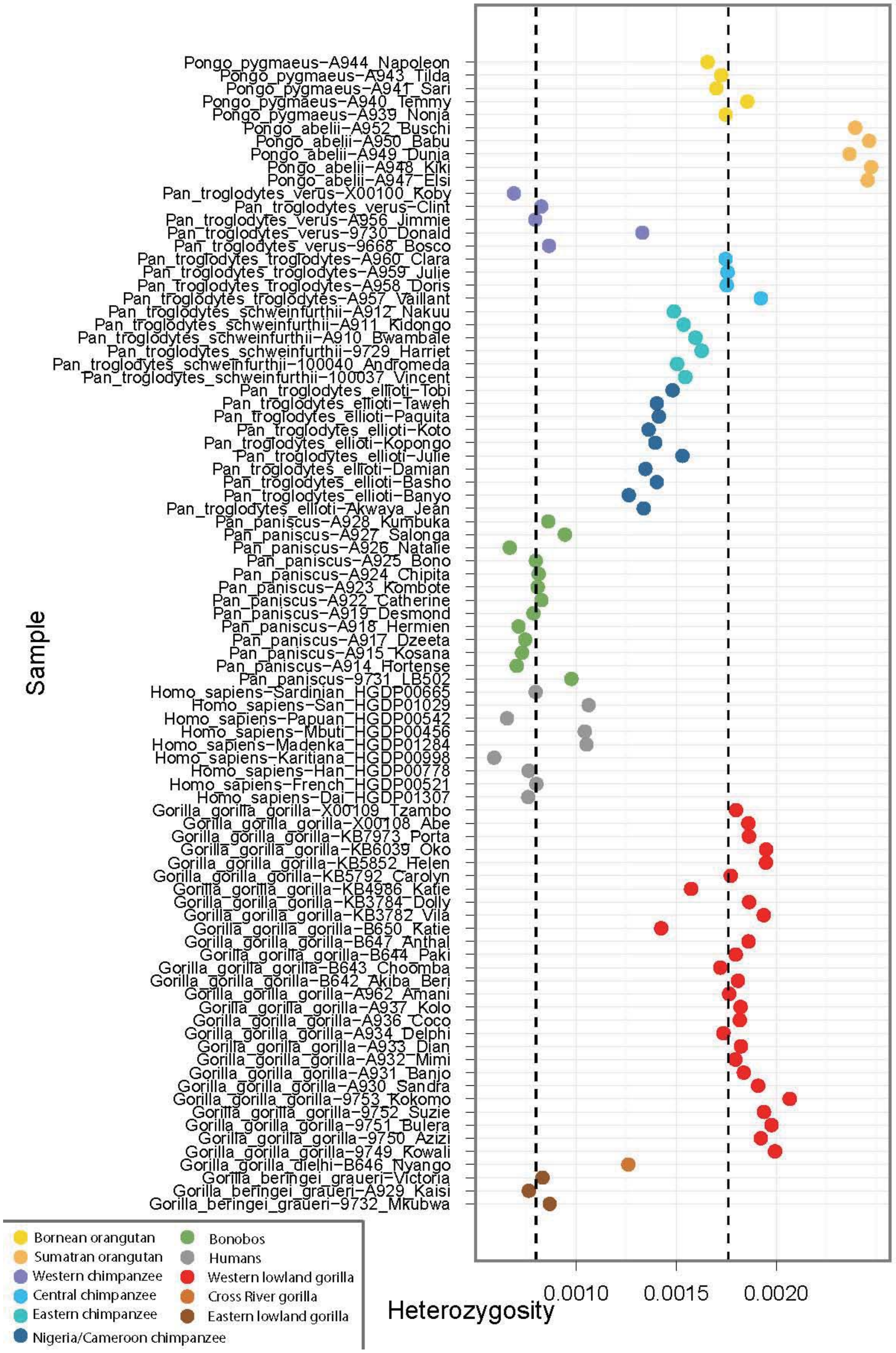
As expected, both mitochondrial and autosomal cladograms showed 100% bootstrap support for all known species relationships; however, subspecies relationships and additional population substructures varied between the two trees. Among chimpanzees, Nigeria-Cameroon and the Western individuals strikingly formed two distinct, high-confidence clades that cluster together. This result is supported by both the mitochondrial and autosomal cladograms, which contract previous reports that may have been biased due to the lack of informative markers targeted[11]. Central and Eastern chimpanzee populations each form clades that cluster together separately from the Nigeria-Cameroon and Western chimpanzees. This split for each of these populations is supported by 72% of autosomal trees, however, and 100% of mitochondrial trees place Eastern chimpanzees into a subclade of Central chimpanzees supporting these populations as being closely related.

Section 6: Heterozygosity

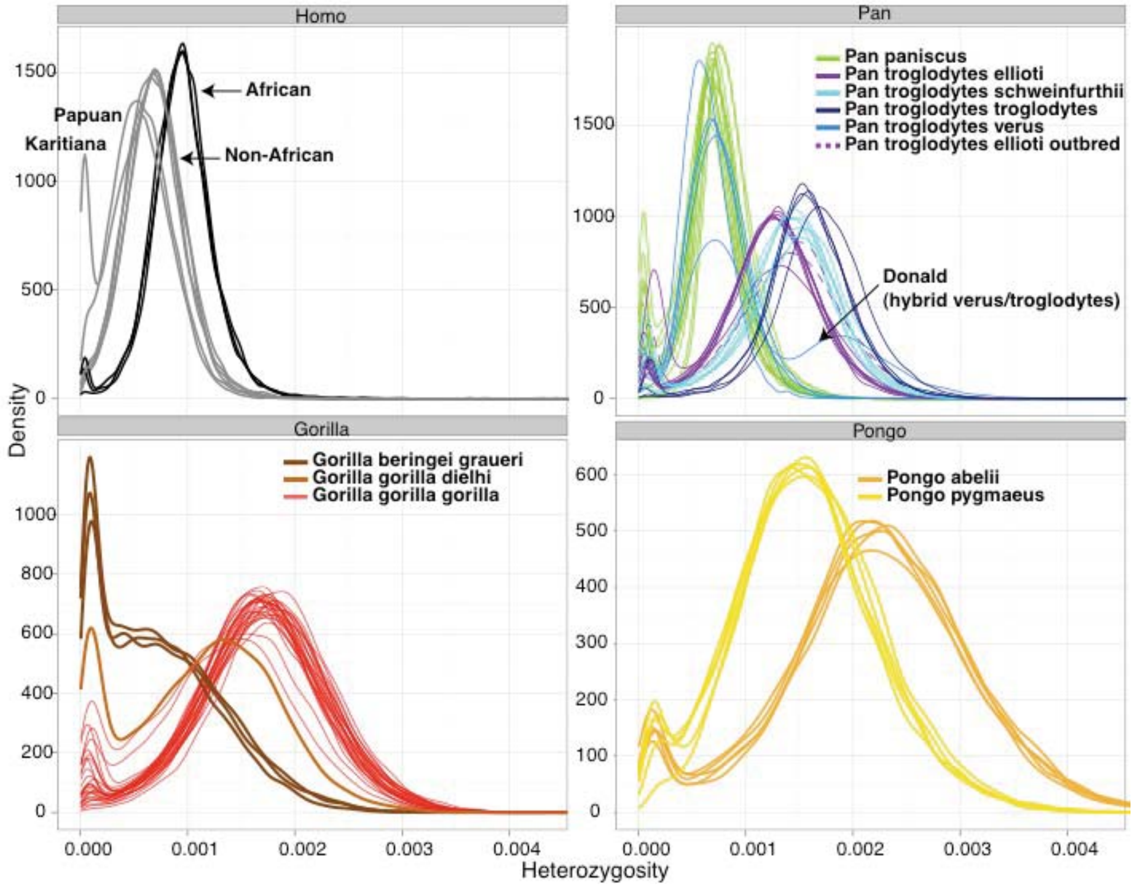
Peter H. Sudmant, Javier Prado-Martinez, Tomas Marques-Bonet, Evan E. Eichler

To assess the genetic diversity within and between great apes, we analyzed the distribution of heterozygosities in each of the species and subpopulations targeted in our study (**Figure 1b**). Relationships between the population groups were inferred by constructing a neighbor-joining tree based on inter-population divergence estimates between species and subpopulations (**Suppl. Table 5.2**). Heterozygosity estimates were computed for each individual (**Suppl. Figure 6.1**) and then combined into species/population distributions. The four related Western lowland gorilla individuals (Bulera, Kowali, Suzie and Oko) were discarded from this analysis in addition to the admixed individual Donald. We find a fourfold range in genome-wide nucleotide diversities for different hominid species and subpopulations. They range from ~6e-4 in highly bottlenecked human Karitiana and Papuan individuals to ~25e-4 in Sumatran orangutans. Non-African humans, Eastern lowland gorillas, bonobos, and Western chimpanzees all demonstrate very similar and comparatively low heterozygosities (~8e-4). Each of the chimpanzee, gorilla and orangutan genera contain high-diversity subpopulations with heterozygosities approximately twice that of these low diversity populations. Populations of intermediate diversity are also present in the gorilla and chimpanzee populations. The orangutan species demonstrates the highest overall heterozygosity of all the great apes.

A more in-depth analysis of the distribution of heterozygosity in 1 Mbp sliding windows (200 Kbp overlap) across the genome in each individual (**Suppl. Figure S6.2**) revealed a striking homogeneity in the distribution of heterozygosity in all the populations assessed, with the exception of humans. The complex demography of the human population has resulted in distinct genome-wide distributions of diversity in different populations. For example, African populations, Asian and European populations, and inbred individuals each show distinct distributions.



Suppl. Figure 6.1 – *Individual heterozygosity estimates corrected for the AB filter. The dotted lines represent heterozygosity estimated by the Chimpanzee Sequencing Consortium for Western and Central chimpanzees, respectively.*



Suppl. Figure 6.2 – *Distributions of heterozygosity of 1 Mbp windows with 200 Kbp overlap for each individual and grouped by the four genera targeted in this study. There is an enrichment of windows with low or no heterozygosity that point to events of recent inbreeding.*

Section 7: Mitochondrial reconstruction

Belen Lorente-Galdos, Gabriel Santpere, Marc Dabad, Tomas Marques-Bonet

We assembled the mitochondrial genome (mtDNA) of the samples with paired-end reads with length between 94 and 114 bp: 9 *Homo sapiens*, 5 *Pan troglodytes ellioti,* 6 *Pan troglodytes schweinfurthii,* 4 *Pan troglodytes troglodytes,* 5 *Pan troglodytes verus,* 13 *Pan paniscus,* 27 *Gorilla gorilla gorilla,* 1 *Gorilla gorilla diehli,* 2 *Gorilla beringei graueri,* 5 *Pongo abelii* and 5 *Pongo pygmaeus*. All the mtDNA were reconstructed from WGS data only (**Suppl. Table 7.1**).

	#	Read length	Median coverage per genome	Minimum coverage	Maximum coverage
Homo sapiens	9	94	22.65	16.44	35.52
Pan troglodytes ellioti	5	100	12.11	11.14	13.21
Pan troglodytes schweinfurthii	6	100	34.57	13.22	48.26
Pan troglodytes troglodytes	4	100	31.32	24.83	38.17
Pan troglodytes verus	5	100	21.05	17.21	38.09
Pan paniscus	13	100,101	32.15	11.61	46.65
Gorilla gorilla gorilla	27	100,101	21.53	12.31	38.88
Gorilla gorilla diehli	1	100	23.05	23.05	23.05
Gorilla beringei graueri	2	100,101	25.92	18.34	33.50
Pongo abelii	5	100	38.53	33.04	39.85
Pongo pygmaeus	5	100	31.60	28.24	36.55

Suppl. Table 7.1 – *Summary per species of the initial WGS reads used to reconstruct the mtDNA.*
**Coverage is computed relative to the 3 Gbp of the human assembly.*

For each sample, we captured reads from the mitochondrial genome by mapping the raw data to previously published mitochondrial assemblies of the corresponding species (**Suppl. Table 7.2**). In a second round of mapping, in order to increase the number of captured reads at the extremes of the assemblies and take advantage of the circularity of mtDNA, we aligned the reads to a modified sequence assembly, changing the origin of the reference assembly at the middle of the mtDNA in the databases (8 Kbp from the start). The mapping was carried out using mrFAST[12] with paired-end mode and 6% of divergence. We removed low-quality reads when at least one of both paired-ends had a median Phred quality score lower than 32 (**Suppl. Table 7.3**).

Species	Accession Code	Length	Length without D-loop
Homo sapiens	NCBI36.1	16,571	15446
Pan troglodytes verus	PanTro2	16,554	15441
Pan paniscus	NC_001644.1	16,563	15442
Gorilla gorilla gorilla	NC_011120.1	16,412	15448
Pongo abelii	X97707.1	16,499	15483
Pongo pygmaeus	NC_001646.1	16,389	15472

Suppl. Table 7.2 – *Mitochondrial assemblies used to capture mitochondrial reads. Notice that some control regions are not totally represented for some species.*

	Map to reference			**Map to reference with modified origin**		
	Median coverage per genome	**Minimum coverage**	**Maximum coverage**	**Median coverage per genome**	**Minimum coverage**	**Maximum coverage**
Homo sapiens	9,261.48	5,634.85	28,714.16	9,240.06	5,637.32	28,675.46
Pan troglodytes ellioti	937.89	873.75	1,252.12	933.10	865.04	1,248.50
Pan troglodytes schweinfurthii	2,076.92	546.73	2,969.60	2,059.71	542.49	2,944.84
Pan troglodytes troglodytes	980.72	503.21	1,726.18	968.19	496.28	1,712.90
Pan troglodytes verus	1,170.62	834.48	8,449.28	1,163.05	826.79	8,442.76
Pan paniscus	1,293.91	779.67	4,146.29	1,274.47	769.53	4,072.05
Gorilla gorilla gorilla	1,663.86	534.28	18,350.96	1,643.94	528.09	18,086.83
Gorilla gorilla diehli	1,069.95	1,069.95	1,069.95	1,057.71	1,057.71	1,057.71
Gorilla beringei graueri	2,277.75	953.12	3,602.38	2,247.87	938.80	3,556.95
Pongo abelii	1,325.77	1,186.36	8,877.87	1,307.68	1,170.10	8,777.60
Pongo pygmaeus	1,823.21	925.90	2,344.22	1,831.11	930.65	2,344.79
All	**1,620.45**	**503.21**	**28,714.16**	**1,597.90**	**496.28**	**28,675.46**

Suppl. Table 7.3 – *Mitochondrial coverage relative to the length of the mitochondrial reference of the corresponding species.*

We then constructed contigs for mtDNA using Hapsembler[13] (-p Illumina -t 4 -d no --PHRED_OFFSET 33 --MIN_CONTIG_SIZE 1000 –EPSILON 0.05). These contigs were finally oriented via local alignments to the corresponding reference assembly (using BLAST[14]) and then joined (using mafft[15]) incorporating N's in the existing gaps.

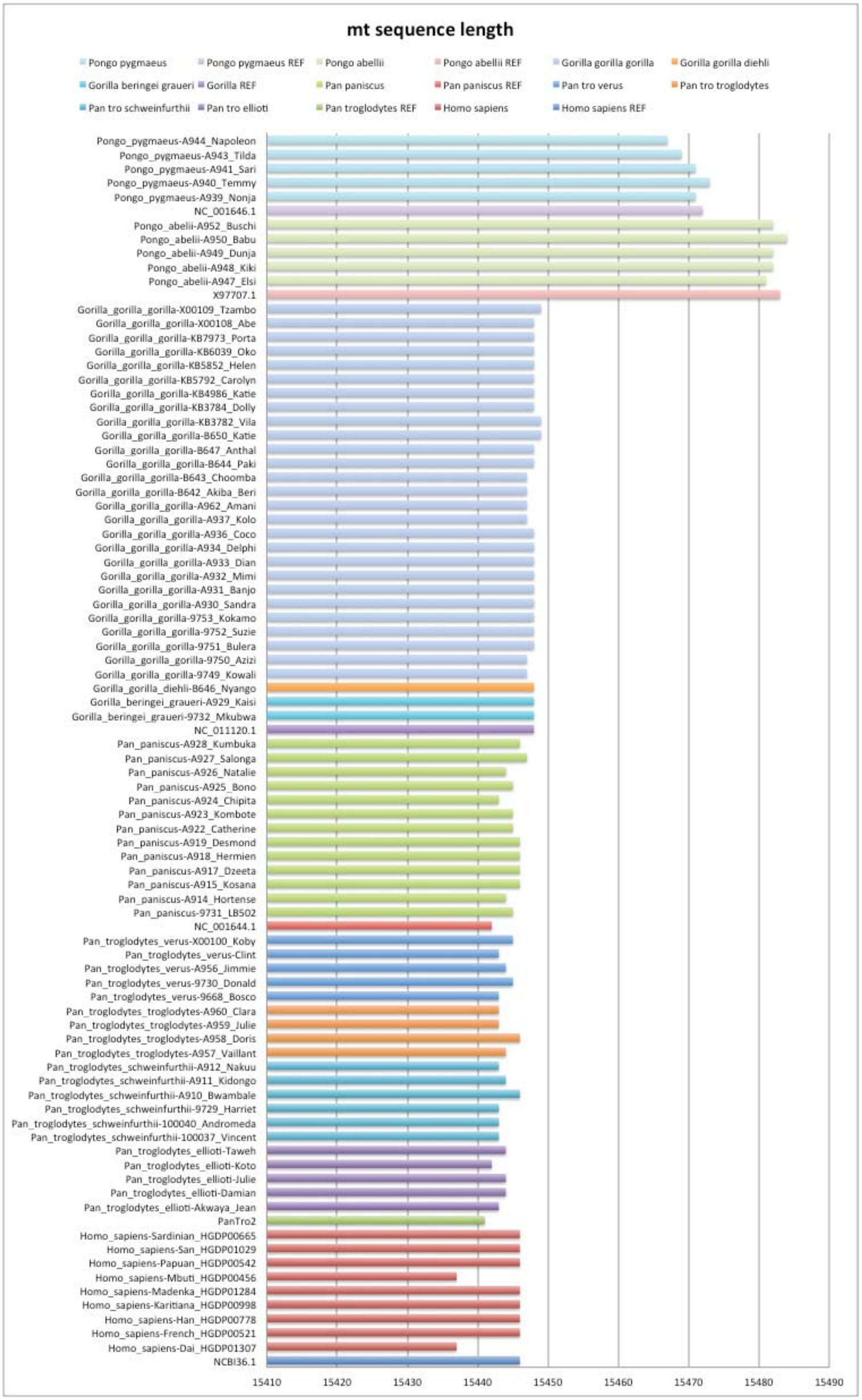
We reduced the final coverage to 350X so the assembler improved its efficiency; the highly variable D-loop region of the mitochondria was also eliminated. To compensate the random representation caused by this reduction and also to reduce the potential problem of *numts* into our reconstructions, we recreated a random reduction of coverage to 350X replicated five times. We did this step twice, once per each reference assembly (standard and changing the origin of the assembly). Thus, we created 10 mitochondrial assemblies per individual. The consensus sequence resulting from the 10 assemblies is the final mitochondrial assembly per sample.

A phylogenetic tree from the final assemblies was created using RAxML (parameters -m GTRGAMMA -# 1000 -n T1 -T 8, for deducing the best tree and, -T 8 -n result -# 1000 -x 12345 -p 12345 -m GTRGAMMA, for calculating bootstrap values).

We were able to reconstruct the mitochondrial sequence of the 82 samples studied, with no gaps outside the D-loop. Except for one individual, the resampling was carried out by reducing coverage to 350X. For the orangutan A949_Dunja, we considered the sequence obtained with the resampling done to 300X. The length of the sequences we obtained is shown in **Suppl. Figure 7.1**.

The mitochondrial sequences of 14 of our samples (4 *Pan troglodytes troglodytes*, 7 *Pan paniscus*, and 3 *Gorilla gorilla gorilla*) were independently obtained via long-range PCR and Illumina sequencing[16] (Hvilsom et al. in prep). We compared the sequences obtained through both methods (**Suppl. Table 7.4**).

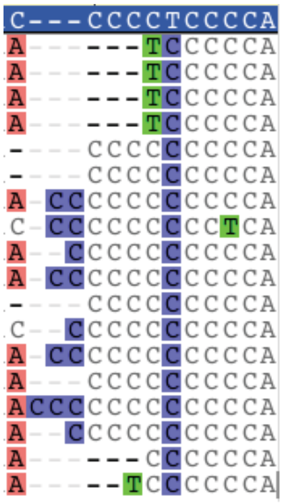
All sequences analyzed show a high level of identity. The differences of these samples correspond only to three regions (~379-389, ~7963-7965, ~8450-8471, starting from the end of the D-loop). The first region is shown in **Suppl. Figure 7.2**, and it is a complex region with homopolymers that complicates its correct identification by any of the two methods.


mt sequence length
Pongo pygmaeus
Pongo pygmaeus REF
Pongo abellii
Pongo abellii REF
Gorilla gorilla gorilla
Gorilla gorilla diehli
Gorilla beringei graueri
Gorilla REF
Pan paniscus
Pan paniscus REF
Pan tro verus
Pan tro troglodytes
Pan tro schweinfurthii
Pan tro ellioti
Pan troglodytes REF
Homo sapiens
Homo sapiens REF
Pongo_pygmaeus-A944_Napoleon
Pongo_pygmaeus-A943_Tilda
Pongo_pygmaeus-A941_Sari
Pongo_pygmaeus-A940_Temmy
Pongo_pygmaeus-A939_Nonja
NC_001646.1
Pongo_abelii-A952_Buschi
Pongo_abelii-A950_Babu
Pongo_abelii-A949_Dunja
Pongo_abelii-A948_Kiki
Pongo_abelii-A947_Elsi
X97707.1
Gorilla_gorilla_gorilla-X00109_Tzambo
Gorilla_gorilla_gorilla-X00108_Abe
Gorilla_gorilla_gorilla-KB7973_Porta
Gorilla_gorilla_gorilla-KB6039_Oko
Gorilla_gorilla_gorilla-KB5852_Helen
Gorilla_gorilla_gorilla-KB5792_Carolyn
Gorilla_gorilla_gorilla-KB4986_Katie
Gorilla_gorilla_gorilla-KB3784_Dolly
Gorilla_gorilla_gorilla-KB3782_Vila
Gorilla_gorilla_gorilla-B650_Katie
Gorilla_gorilla_gorilla-B647_Anthal
Gorilla_gorilla_gorilla-B644_Paki
Gorilla_gorilla_gorilla-B643_Choomba
Gorilla_gorilla_gorilla-B642_Akiba_Beri
Gorilla_gorilla_gorilla-A962_Amani
Gorilla_gorilla_gorilla-A937_Kolo
Gorilla_gorilla_gorilla-A936_Coco
Gorilla_gorilla_gorilla-A934_Delphi
Gorilla_gorilla_gorilla-A933_Dian
Gorilla_gorilla_gorilla-A932_Mimi
Gorilla_gorilla_gorilla-A931_Banjo
Gorilla_gorilla_gorilla-A930_Sandra
Gorilla_gorilla_gorilla-9753_Kokamo
Gorilla_gorilla_gorilla-9752_Suzie
Gorilla_gorilla_gorilla-9751_Bulera
Gorilla_gorilla_gorilla-9750_Azizi
Gorilla_gorilla_gorilla-9749_Kowali
Gorilla_gorilla_diehli-B646_Nyango
Gorilla_beringei_graueri-A929_Kaisi
Gorilla_beringei_graueri-9732_Mkubwa
NC_011120.1
Pan_paniscus-A928_Kumbuka
Pan_paniscus-A927_Salonga
Pan_paniscus-A926_Natalie
Pan_paniscus-A925_Bono
Pan_paniscus-A924_Chipita
Pan_paniscus-A923_Kombote
Pan_paniscus-A922_Catherine
Pan_paniscus-A919_Desmond
Pan_paniscus-A918_Hermien
Pan_paniscus-A917_Dzeeta
Pan_paniscus-A915_Kosana
Pan_paniscus-A914_Hortense
Pan_paniscus-9731_LB502
NC_001644.1
Pan_troglodytes_verus-X00100_Koby
Pan_troglodytes_verus-Clint
Pan_troglodytes_verus-A956_Jimmie
Pan_troglodytes_verus-9730_Donald
Pan_troglodytes_verus-9668_Bosco
Pan_troglodytes_troglodytes-A960_Clara
Pan_troglodytes_troglodytes-A959_Julie
Pan_troglodytes_troglodytes-A958_Doris
Pan_troglodytes_troglodytes-A957_Vaillant
Pan_troglodytes_schweinfurthii-A912_Nakuu
Pan_troglodytes_schweinfurthii-A911_Kidongo
Pan_troglodytes_schweinfurthii-A910_Bwambale
Pan_troglodytes_schweinfurthii-9729_Harriet
Pan_troglodytes_schweinfurthii-100040_Andromeda
Pan_troglodytes_schweinfurthii-100037_Vincent
Pan_troglodytes_ellioti-Taweh
Pan_troglodytes_ellioti-Koto
Pan_troglodytes_ellioti-Julie
Pan_troglodytes_ellioti-Damian
Pan_troglodytes_ellioti-Akwaya_Jean
PanTro2
Homo_sapiens-Sardinian_HGDP00665
Homo_sapiens-San_HGDP01029
Homo_sapiens-Papuan_HGDP00542
Homo_sapiens-Mbuti_HGDP00456
Homo_sapiens-Madenka_HGDP01284
Homo_sapiens-Karitiana_HGDP00998
Homo_sapiens-Han_HGDP00778
Homo_sapiens-French_HGDP00521
Homo_sapiens-Dai_HGDP01307
NCBI36.1
15410
15420
15430
15440
15450
15460
15470
15480
15490

Suppl. Figure 7.1 – *Lengths of the sequences without D-loop. The lengths of the references used for mapping are also shown.*

Sample-ID *	Sample-ID **	Length *	Length **	Matches	Mismatches	Indels	# bp in indels
Ptt_Vaillant	A957_Vaillant	15444	15444	15444	0	0	0
Ptt_Doris	A958_Doris	15447	15446	15446	0	1	1
Ptt_Julie	A959_Julie	15443	15443	15443	0	0	0
Ptt_Clara	A960_Clara	15449	15443	15443	0	1	6
Ppa_Hortense	A914_Hortense	15444	15444	15443	1	0	0
Ppa_Kosana	A915_Kosana	15448	15446	15445	1	1	2
Ppa_Dzeeta	A917_Dzeeta	15447	15446	15446	0	1	1
Ppa_Hermien	A918_Hermien	15448	15446	15445	1	1	2
Ppa_Desmond	A919_Desmond	15448	15446	15445	1	1	2
Ppa_Natalie	A926_Natalie	15444	15444	15444	1	0	0
ppa_Kumbuka	A928_Kumbuka	15446	15446	15446	1	0	0
Ggg_Mimi	A932_Mimi	15448	15448	15443	5	0	0
Ggg_Dian	A933_Dian	15448	15448	15443	5	0	0
Ggg_Amani	A962_Amani	15447	15447	15442	5	0	0

Suppl. Table 7.4 – *Comparison of the sequences obtained via two independent methods. *Long-range PCR plus Illumina sequencing. ** Assemblies in this study from Illumina data.*



```
C---CCCCTCCCCA
A------TCCCCCA
A------TCCCCCA
A------TCCCCCA
A------TCCCCCA
----CCCCCCCCCA
----CCCCCCCCCA
A-CCCCCCCCCCCA
C-CCCCCCCCCTCA
A--CCCCCCCCCCA
A-CCCCCCCCCCCA
----CCCCCCCCCA
C--CCCCCCCCCCA
A-CCCCCCCCCCCA
A---CCCCCCCCCA
ACCCCCCCCCCCCA
A--CCCCCCCCCCA
A------CCCCCCA
A-----TCCCCCCA
```

Suppl. Figure 7.2 – *Alignment of the sequences obtained through long-PCR and Illumina sequencing. In this region, approximately at 380 bp from the D-loop, many of these sequences differ from the ones obtained via our method.*

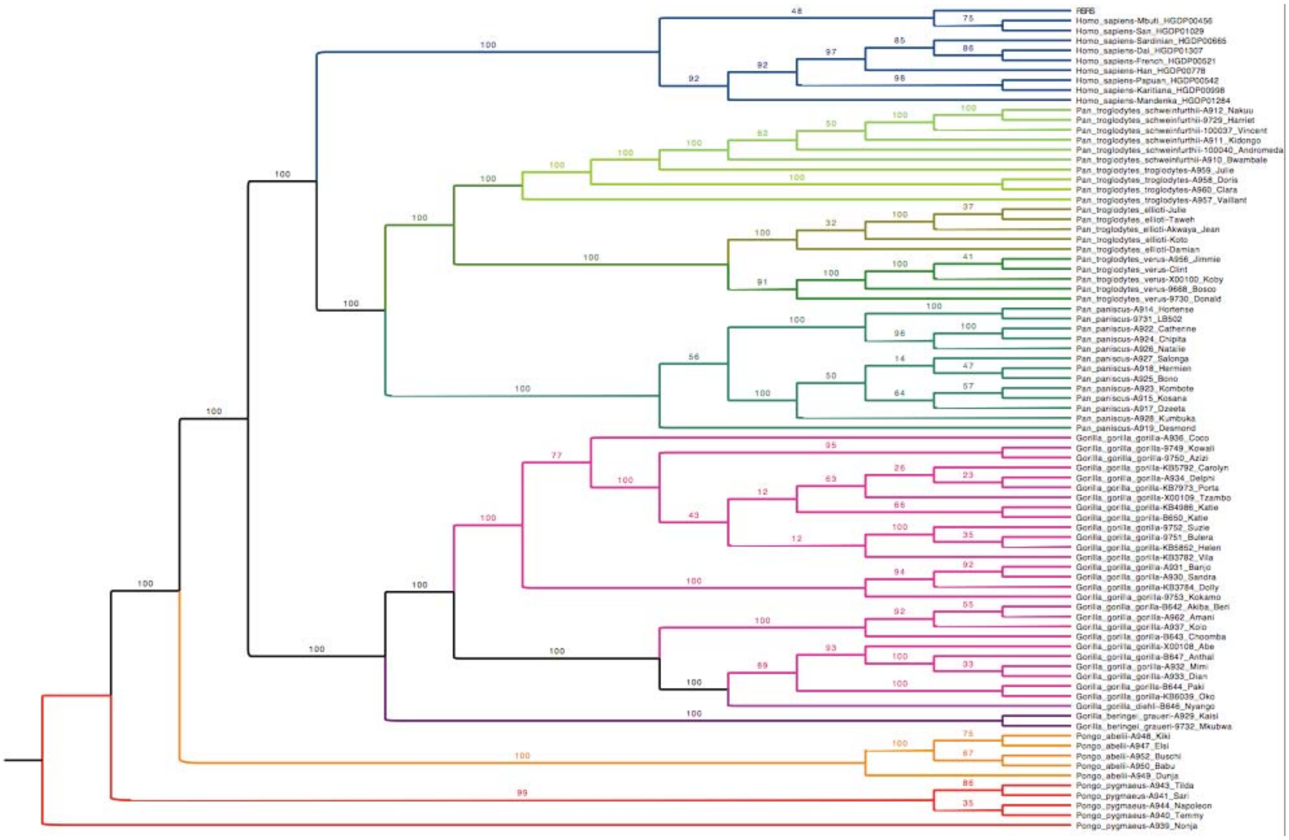
Human haplogroups

For the human samples, we identified the variants relative to a high-quality sequence[17] (**Suppl. Table 7.5**). The haplogroups for these sequences match the expected population they belong to. All common variants that define each haplogroup are found in our sequences.

Population	Sample ID	Haplogroup	Variants
Dai	HGDP01307	B4a1c4	709A, 769G, 825T, 1018G, 2758G, 2885T, 3594C, 4104A, 4312C, 5465C, 7146A, 7256C, 7521G, 8281-8289d, 8468C, 8655C, 8701A, 9123A, 9540T, 10238C, 10398A, 10664C, 10688G, 10810T, 10873T, 10915T, 11914G, 12705C, 12904G, 13105A, 13276A, 13506C, 13650C
French	HGDP00521	T1a	709A, 769G, 825T, 1018G, 1888A, 2758G, 2885T, 3394C, 3594C, 4104A, 4216C, 4312C, 4639C, 4917G, 7146A, 7256C, 7521G, 8468C, 8655C, 8697A, 8701A, 9540T, 10398A, 10463C, 10664C, 10688G, 10810T, 10873T, 10915T, 11251G, 11914G, 12633A, 12705C, 13105A, 13276A, 13368A, 13506C, 13650C, 14905A, 15452A, 15607G, 15928A
Han	HGDP00778	A5b	663G, 769G, 825T, 961C, 1018G, 1709A, 1736G, 2758G, 2885T, 3594C, 4104A, 4248C, 4312C, 4316G, 4824G, 7146A, 7256C, 7521G, 8468C, 8563G, 8655C, 8701A, 8794T, 9540T, 10398A, 10664C, 10688G, 10810T, 10873T, 10915T, 11536T, 11914G, 13105A, 13276A, 13506C, 13650C, 13999T
Karitiana	HGDP00998	D1	769G, 825T, 1018G, 1821G, 2092T, 2758G, 2885T, 3010A, 3591A, 3594C, 4104A, 4312C, 4883T, 5178A, 7146A, 7256C, 7521G, 8414T, 8468C, 8655C, 10118C, 10400T, 10664C, 10688G, 10810T, 10915T, 11914G, 11928G, 12732C, 13105A, 13276A, 13506C, 13650C, 14256C, 14668T, 14783C, 15043A, 15301A
Mandenka	HGDP01284	L2c3a	680C, 709A, 825T, 1442A, 2332T, 2416C, 2589G, 2758G, 2885T, 3200A, 4312C, 5255T, 6521T, 7146A, 7624A, 8206A, 8468C, 8655C, 8733C, 9221G, 10115C, 10664C, 10688G, 10810T, 10915T, 11914G, 11944C, 12236A, 13105A, 13276A, 13506C, 13590A, 13928C, 13958C, 15077A, 15110A, 15217A, 15301A, 15849T
Mbuti	HGDP00456	L0a2b	1048T, 2245G, 3372C, 3516A, 4586C, 5147A, 5231A, 5237A, 5442C, 5460A, 5603T, 5711G, 6185C, 6257A, 8281-8289d, 8428T, 8460G, 8566G, 9042T, 9347G, 9755A, 9818T, 10589A, 11172G, 11176A, 11269T, 11641G, 12007A, 12172G, 12720G, 13281C, 14308C, 15136T, 15431A
Papuan	HGDP00542	Q3a	769G, 825T, 1018G, 2758G, 2768G, 2885T, 3594C, 4104A, 4117C, 4312C, 4335T, 5843G, 7146A, 7256C, 7521G, 8468C, 8578T, 8655C, 8790A, 10400T, 10664C, 10688G, 10810T, 10915T, 11260C, 11914G, 12940A, 13105A, 13276A, 13500C, 13506C, 13650C, 14783C, 15043A, 15172A, 15301A
San	HGDP01029	L0d1b1	719A, 1048T, 1438A, 2706A, 3438A, 3516A, 3618C, 3756G, 4232C, 5029C, 5442C, 6185C, 6266G, 6815C, 7283C, 8113A, 8152A, 8251A, 8383C, 8937C, 9042T, 9347G, 9755A, 10589A, 12007A, 12121C, 12720G, 13759A, 14315T, 14659T, 15466A, 15692G, 15930A, 15941C
Sardinian	HGDP00665	H3u	769G, 825T, 1018G, 2706A, 2758G, 2885T, 3594C, 4104A, 4312C, 6776C, 7028C, 7146A, 7256C, 7521G, 8468C, 8655C, 8701A, 9540T, 9966A, 10398A, 10664C, 10688G, 10810T, 10873T, 10915T, 11719G, 11914G, 12705C, 13105A, 13276A, 13506C, 13650C, 14766C, 15315T

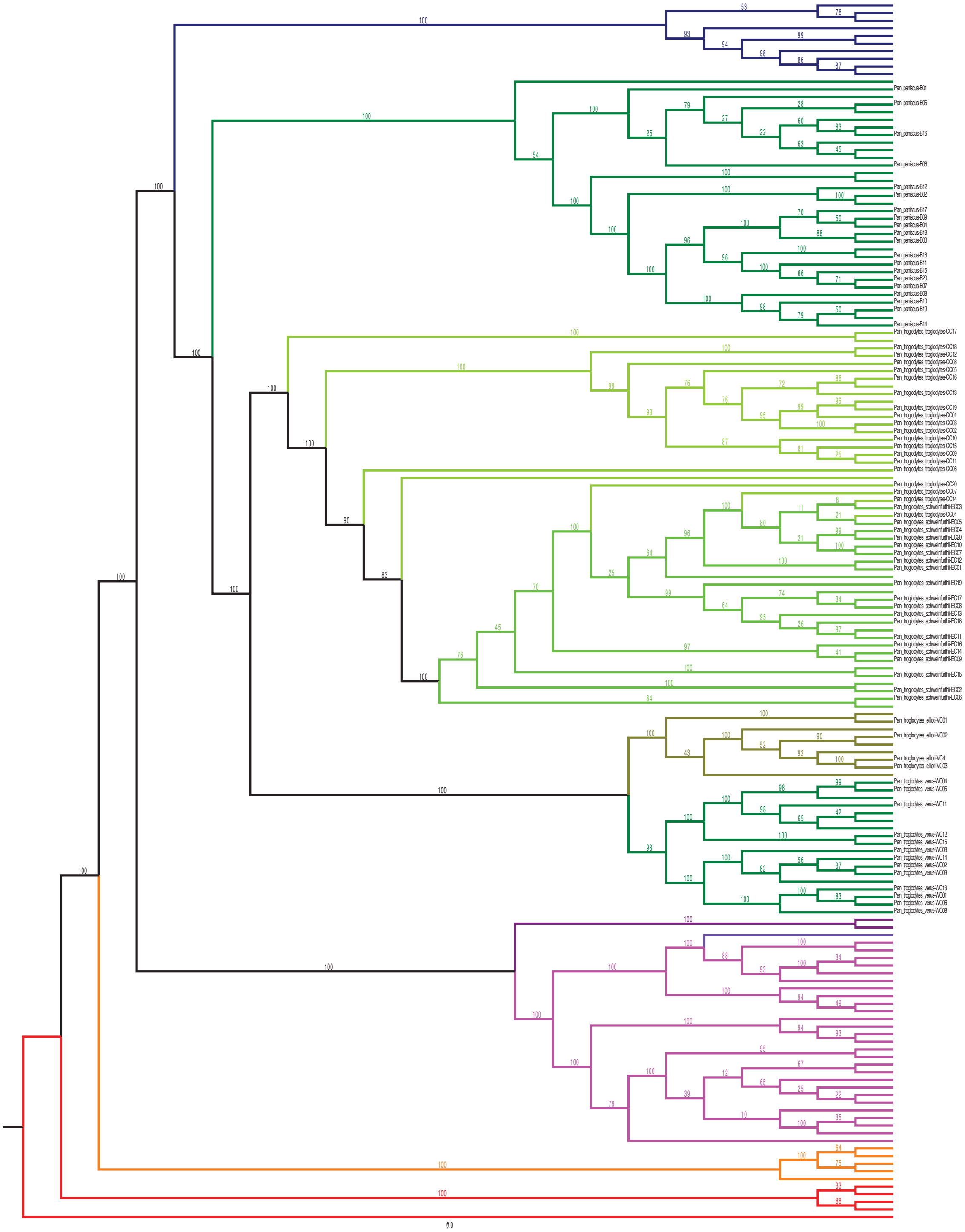
Suppl. Table 7.5 – *Human haplogroups. The variants are given in coordinates of a high-quality human reference*[17].

The phylogenetic tree of our sequences is shown in **Suppl. Figure 7.3**. All sequences were ultimately aligned to the human reference[17], which is also included in the tree.


48
75
100
85
97
86
92
92
98
100
100
50
63
100
100
100
100
100
100
100
37
100
32
100
100
100
41
91
100
100
100
100
96
56
14
47
50
100
100
64
57
100
95
77
26
23
100
63
12
66
43
100
100
35
12
94
92
100
100
92
55
100
93
100
100
33
89
100
100
100
75
100
67
100
86
99
35
RSRS
Homo_sapiens-Mbuti_HGDP00456
Homo_sapiens-San_HGDP01029
Homo_sapiens-Sardinian_HGDP00665
Homo_sapiens-Dai_HGDP01307
Homo_sapiens-French_HGDP00521
Homo_sapiens-Han_HGDP00778
Homo_sapiens-Papuan_HGDP00542
Homo_sapiens-Karitiana_HGDP00998
Homo_sapiens-Mandenka_HGDP01284
Pan_troglodytes_schweinfurthii-A912_Nakuu
Pan_troglodytes_schweinfurthii-9729_Harriet
Pan_troglodytes_schweinfurthii-100037_Vincent
Pan_troglodytes_schweinfurthii-A911_Kidongo
Pan_troglodytes_schweinfurthii-100040_Andromeda
Pan_troglodytes_schweinfurthii-A910_Bwambale
Pan_troglodytes_troglodytes-A959_Julie
Pan_troglodytes_troglodytes-A958_Doris
Pan_troglodytes_troglodytes-A960_Clara
Pan_troglodytes_troglodytes-A957_Vaillant
Pan_troglodytes_ellioti-Julie
Pan_troglodytes_ellioti-Taweh
Pan_troglodytes_ellioti-Akwaya_Jean
Pan_troglodytes_ellioti-Koto
Pan_troglodytes_ellioti-Damian
Pan_troglodytes_verus-A956_Jimmie
Pan_troglodytes_verus-Clint
Pan_troglodytes_verus-X00100_Koby
Pan_troglodytes_verus-9668_Bosco
Pan_troglodytes_verus-9730_Donald
Pan_paniscus-A914_Hortense
Pan_paniscus-9731_LB502
Pan_paniscus-A922_Catherine
Pan_paniscus-A924_Chipita
Pan_paniscus-A926_Natalie
Pan_paniscus-A927_Salonga
Pan_paniscus-A918_Hermien
Pan_paniscus-A925_Bono
Pan_paniscus-A923_Kombote
Pan_paniscus-A915_Kosana
Pan_paniscus-A917_Dzeeta
Pan_paniscus-A928_Kumbuka
Pan_paniscus-A919_Desmond
Gorilla_gorilla_gorilla-A936_Coco
Gorilla_gorilla_gorilla-9749_Kowali
Gorilla_gorilla_gorilla-9750_Azizi
Gorilla_gorilla_gorilla-KB5792_Carolyn
Gorilla_gorilla_gorilla-A934_Delphi
Gorilla_gorilla_gorilla-KB7973_Porta
Gorilla_gorilla_gorilla-X00109_Tzambo
Gorilla_gorilla_gorilla-KB4986_Katie
Gorilla_gorilla_gorilla-B650_Katie
Gorilla_gorilla_gorilla-9752_Suzie
Gorilla_gorilla_gorilla-9751_Bulera
Gorilla_gorilla_gorilla-KB5852_Helen
Gorilla_gorilla_gorilla-KB3782_Vila
Gorilla_gorilla_gorilla-A931_Banjo
Gorilla_gorilla_gorilla-A930_Sandra
Gorilla_gorilla_gorilla-KB3784_Dolly
Gorilla_gorilla_gorilla-9753_Kokamo
Gorilla_gorilla_gorilla-B642_Akiba_Beri
Gorilla_gorilla_gorilla-A962_Amani
Gorilla_gorilla_gorilla-A937_Kolo
Gorilla_gorilla_gorilla-B643_Choomba
Gorilla_gorilla_gorilla-X00108_Abe
Gorilla_gorilla_gorilla-B647_Anthal
Gorilla_gorilla_gorilla-A932_Mimi
Gorilla_gorilla_gorilla-A933_Dian
Gorilla_gorilla_gorilla-B644_Paki
Gorilla_gorilla_gorilla-KB6039_Oko
Gorilla_gorilla_diehl-B646_Nyango
Gorilla_beringei_graueri-A929_Kaisi
Gorilla_beringei_graueri-9732_Mkubwa
Pongo_abelii-A948_Kiki
Pongo_abelii-A947_Elsi
Pongo_abelii-A952_Buschi
Pongo_abelii-A950_Babu
Pongo_abelii-A949_Dunja
Pongo_pygmaeus-A943_Tilda
Pongo_pygmaeus-A941_Sari
Pongo_pygmaeus-A944_Napoleon
Pongo_pygmaeus-A940_Temmy
Pongo_pygmaeus-A939_Nonja

Suppl. Figure 7.3 – *Phylogenetic tree of 82 samples and the human reference.*

Another 77 individuals (20 *Pan troglodytes troglodytes*, 20 *Pan troglodytes schweinfurthii*, 4 *Pan troglodytes ellioti*, 13 *Pan troglodytes verus*, and 20 *Pan paniscus*) from a previous publication[16] have been included in the global phylogenetic tree to ensure the results of our classification (**Suppl. Figure 7.4**).


Pan_paniscus-B01
Pan_paniscus-B05
Pan_paniscus-B16
Pan_paniscus-B06
Pan_paniscus-B12
Pan_paniscus-B02
Pan_paniscus-B17
Pan_paniscus-B09
Pan_paniscus-B04
Pan_paniscus-B13
Pan_paniscus-B03
Pan_paniscus-B18
Pan_paniscus-B11
Pan_paniscus-B15
Pan_paniscus-B20
Pan_paniscus-B07
Pan_paniscus-B08
Pan_paniscus-B10
Pan_paniscus-B19
Pan_paniscus-B14
Pan_troglodytes_troglodytes-CC17
Pan_troglodytes_troglodytes-CC18
Pan_troglodytes_troglodytes-CC12
Pan_troglodytes_troglodytes-CC08
Pan_troglodytes_troglodytes-CC05
Pan_troglodytes_troglodytes-CC16
Pan_troglodytes_troglodytes-CC13
Pan_troglodytes_troglodytes-CC19
Pan_troglodytes_troglodytes-CC01
Pan_troglodytes_troglodytes-CC03
Pan_troglodytes_troglodytes-CC02
Pan_troglodytes_troglodytes-CC10
Pan_troglodytes_troglodytes-CC15
Pan_troglodytes_troglodytes-CC09
Pan_troglodytes_troglodytes-CC11
Pan_troglodytes_troglodytes-CC06
Pan_troglodytes_troglodytes-CC20
Pan_troglodytes_troglodytes-CC07
Pan_troglodytes_troglodytes-CC14
Pan_troglodytes_schweinfurthii-EC03
Pan_troglodytes_troglodytes-CC04
Pan_troglodytes_schweinfurthii-EC05
Pan_troglodytes_schweinfurthii-EC04
Pan_troglodytes_schweinfurthii-EC20
Pan_troglodytes_schweinfurthii-EC10
Pan_troglodytes_schweinfurthii-EC07
Pan_troglodytes_schweinfurthii-EC12
Pan_troglodytes_schweinfurthii-EC01
Pan_troglodytes_schweinfurthii-EC19
Pan_troglodytes_schweinfurthii-EC17
Pan_troglodytes_schweinfurthii-EC08
Pan_troglodytes_schweinfurthii-EC13
Pan_troglodytes_schweinfurthii-EC18
Pan_troglodytes_schweinfurthii-EC11
Pan_troglodytes_schweinfurthii-EC16
Pan_troglodytes_schweinfurthii-EC14
Pan_troglodytes_schweinfurthii-EC09
Pan_troglodytes_schweinfurthii-EC15
Pan_troglodytes_schweinfurthii-EC02
Pan_troglodytes_schweinfurthii-EC06
Pan_troglodytes_ellioti-VC01
Pan_troglodytes_ellioti-VC02
Pan_troglodytes_ellioti-VC4
Pan_troglodytes_ellioti-VC03
Pan_troglodytes_verus-WC04
Pan_troglodytes_verus-WC05
Pan_troglodytes_verus-WC11
Pan_troglodytes_verus-WC12
Pan_troglodytes_verus-WC15
Pan_troglodytes_verus-WC03
Pan_troglodytes_verus-WC14
Pan_troglodytes_verus-WC02
Pan_troglodytes_verus-WC09
Pan_troglodytes_verus-WC13
Pan_troglodytes_verus-WC01
Pan_troglodytes_verus-WC06
Pan_troglodytes_verus-WC08
0.0

Suppl. Figure 7.4 – *Phylogenetic tree of our 82 samples, 77 from a previous study*[16] *and the human reference. Only the 77 added individuals are labeled in this tree.*

Section 8: Population structure

8.1. Genetic structure analysis

Timothy D. O'Connor, Joshua Akey

For all groups we used the following settings from PLINK[18]:
PLINK –geno 0.1 –maf MAF –indep-pairwise 50 5 0.1

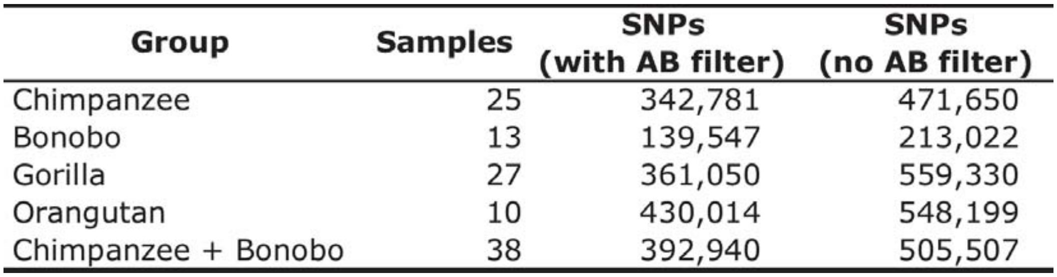
In the case of *Homo sapiens*, *Pongo,* and *Pan paniscus*, we used MAF = 0.06 to remove singletons, otherwise we used 0.05 as the MAF threshold. Additionally, we removed all SNVs from the X chromosome to make the resulting data autosomal only. This resulted in SNV counts of 96,473 (*Homo sapiens*); 330,941 (*Gorilla* combined; with Katies 361049); 139,547 (*Pan paniscus*); 342,781 (*Pan troglodytes*); and 271,889 (*Pongo*).

Using this filtered data, we performed a Frappe analysis[19]. Frappe is a maximum likelihood method that finds a preset number of clusters of structure (K) in genetic data. For the expectation-maximization algorithm utilized by Frappe we used a maximum of 50,000 iterations and a likelihood increase termination threshold of 0.001 per step, the latter being the primary criterion for termination with these data. For each value of K (2 to 8) we ran the analysis 10 times and selected the highest log likelihood. All visualization was done in the R software environment (**Figure 1**).

8.2. Statistical support of population structure

Jeffrey M. Kidd, Joanna L. Kelley

We based our analysis on an "LD-thinned" set of polymorphic sites identified in each group of samples. We first limited analysis to the indicated samples, then pruned for minor allele frequency, rate of missing genotypes, and LD using PLINK[18] (--geno 0.1 --maf 0.05 and --indep-pairwise 50 5 0.1). We only considered autosomal SNPs and removed variants on random or unassigned chromosomes. The dataset we used had an AB filter applied to heterozygous sites. The AB filter introduces a bias against high frequency events in the final dataset, since higher frequency alleles have more heterozygous genotypes, which can potentially be filtered out. Using this set of variants (**Suppl. Table 8.2.1**), we assessed the evidence of population structure based on PCA (using the smartpca program in EIGENSOFT[20] and associated Tracy–Widom statistics) and ADMIXTURE[21] with 10-fold cross-validation[22]. ADMIXTURE analysis was performed five times using independent random seeds for each interrogated value of K. We conducted analysis on SNP sets with and without the AB filter and obtained concordant results. We therefore focus on the results obtained with the AB filter.

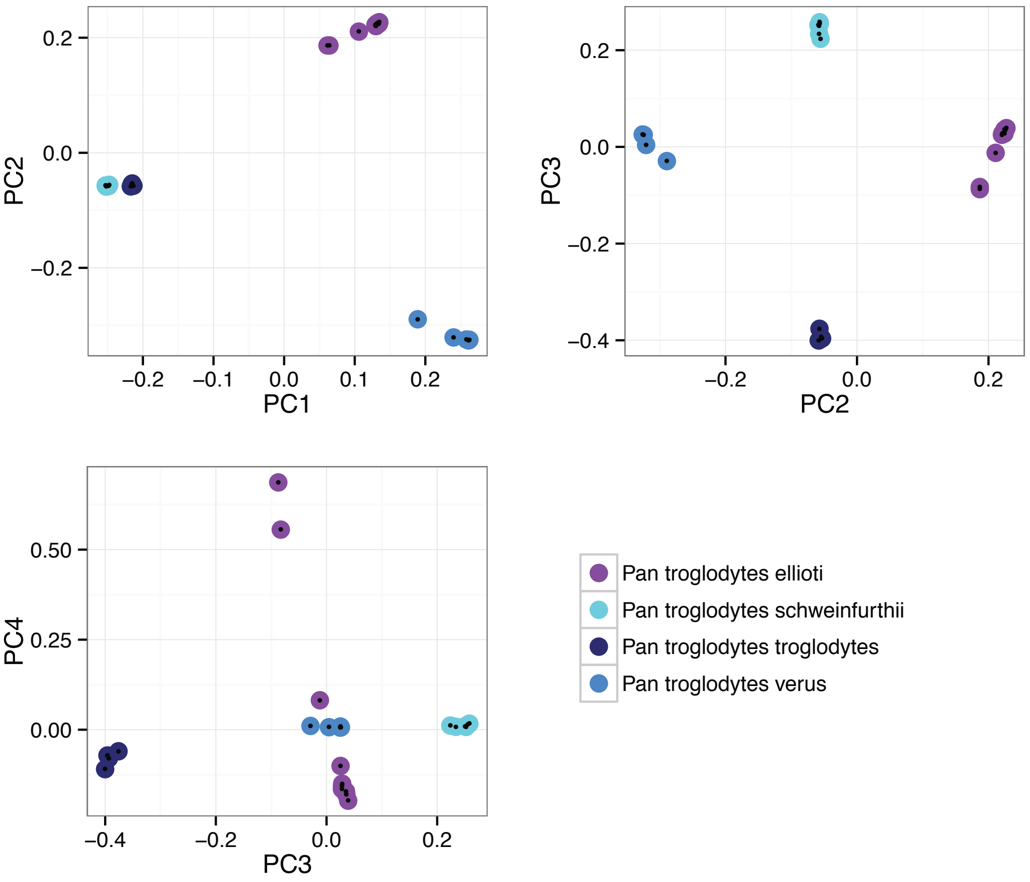


Group	Samples	SNPs (with AB filter)	SNPs (no AB filter)
Chimpanzee	25	342,781	471,650
Bonobo	13	139,547	213,022
Gorilla	27	361,050	559,330
Orangutan	10	430,014	548,199
Chimpanzee + Bonobo	38	392,940	505,507

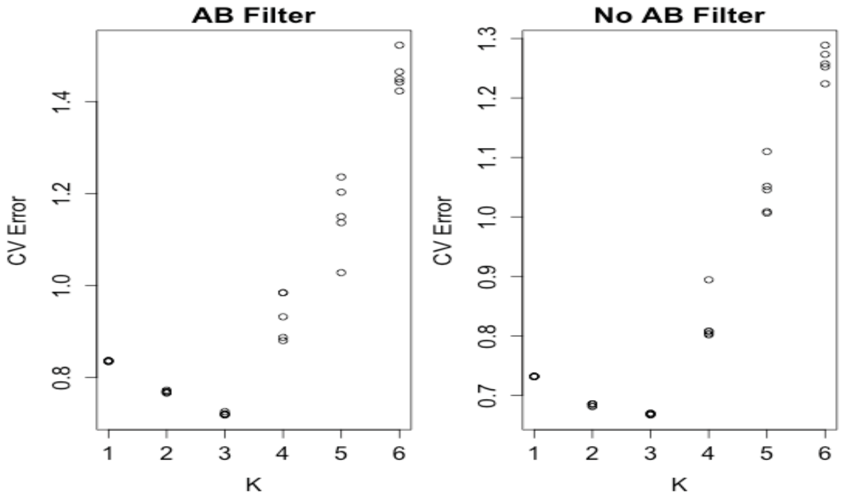
Suppl. Table 8.2.1 – *Summary of SNPs used for analysis. The combined chimpanzee and bonobo callset was constructed based on the individual species calls using the merged mask of the callable genome.*

Chimpanzee

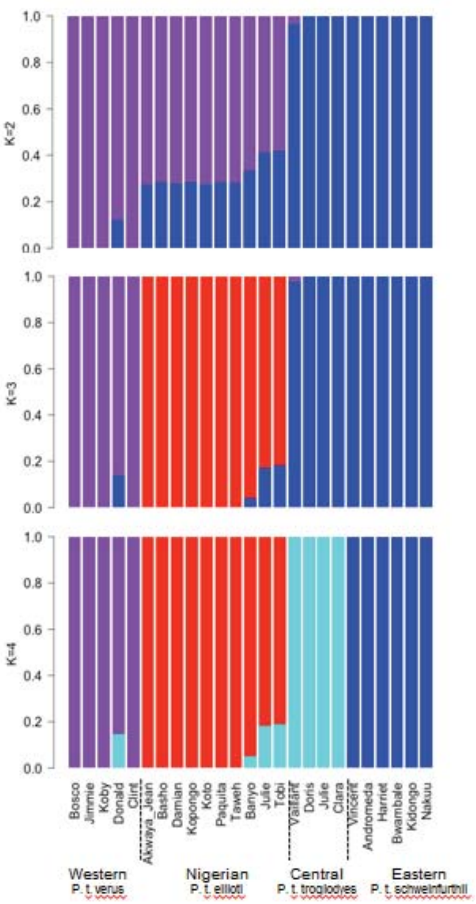
PCA of SNP genotypes identified three significant principal components. The first PC (20.8% of the variance, p = 0.0023) separates the Central and Eastern chimpanzees from the Nigeria-Cameroon and Western chimpanzees. The second PC (16.6% of the variance, p = 1.9e-07) separates the Nigeria-Cameroon and Western chimpanzees while the third PC (5.9% of the variance, p = 3.405e-06) separates the Central and Eastern chimpanzees (**Suppl. Figure 8.2.1**). Cross-validation analysis using ADMIXTURE offers support for three clusters (**Suppl. Figure 8.2.2**). Central and Eastern chimpanzees are separated at K = 4, a value not supported by the cross-validation metric. (**Suppl. Figure 8.2.3**)



Suppl. Figure 8.2.1 – *PC analysis of chimpanzees.*



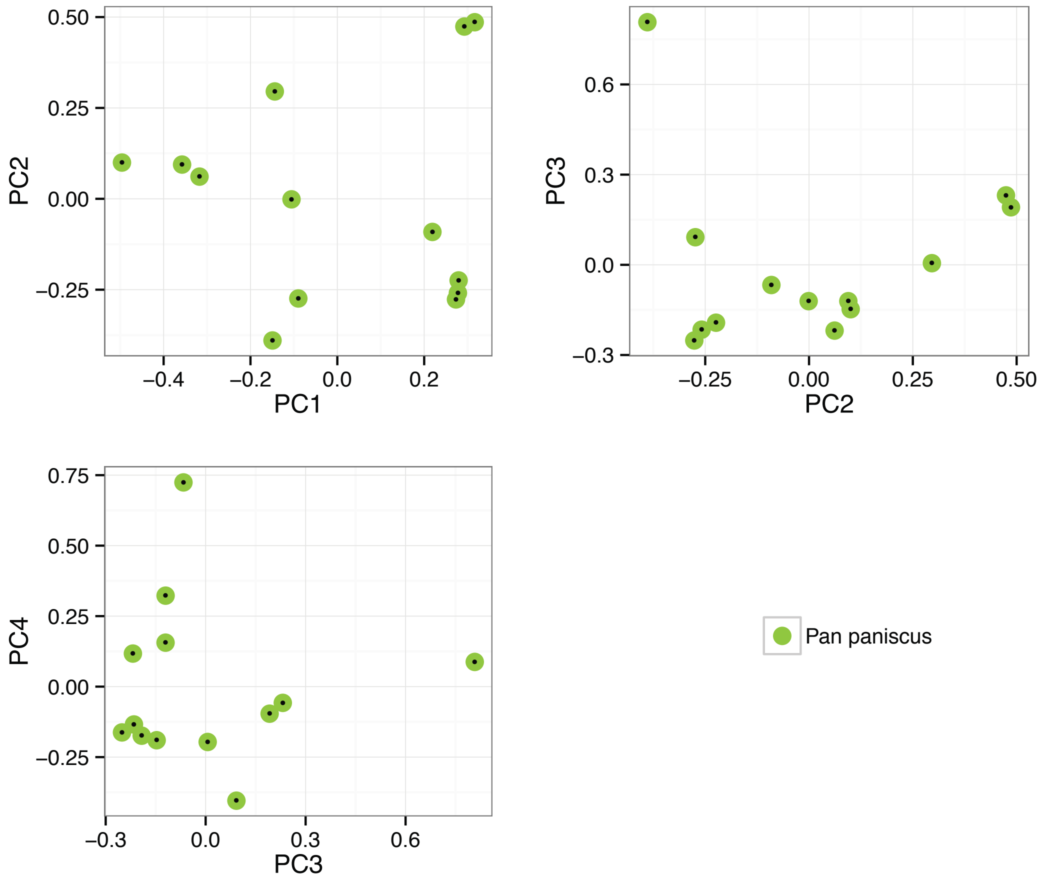
Suppl. Figure 8.2.2 – *ADMIXTURE cross-validation errors for chimpanzees.*



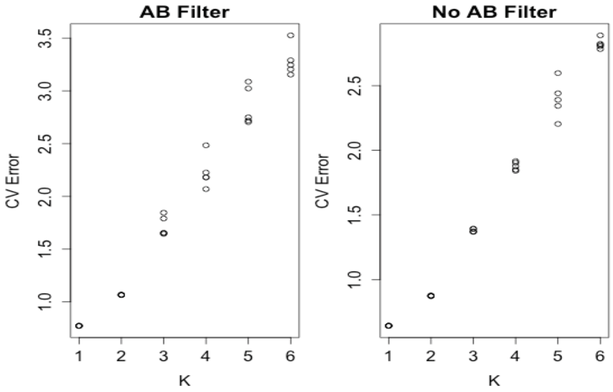
Suppl. Figure 8.2.3 – *ADMIXTURE cluster-membership for chimpanzees. The assignments at K = 4 are not supported by the cross-validation metric*.

Bonobo

PCA of SNP genotypes for the bonobos identifies a single significant component (14.3% of the variance p = 0.000393) along which the 13 bonobos are arrayed (**Suppl. Figure 8.2.4**). The second PC is not significant (p = 0.09). ADMIXTURE cross-validation indicates that K = 1 has the best support (**Suppl. Figure 8.2.5**).



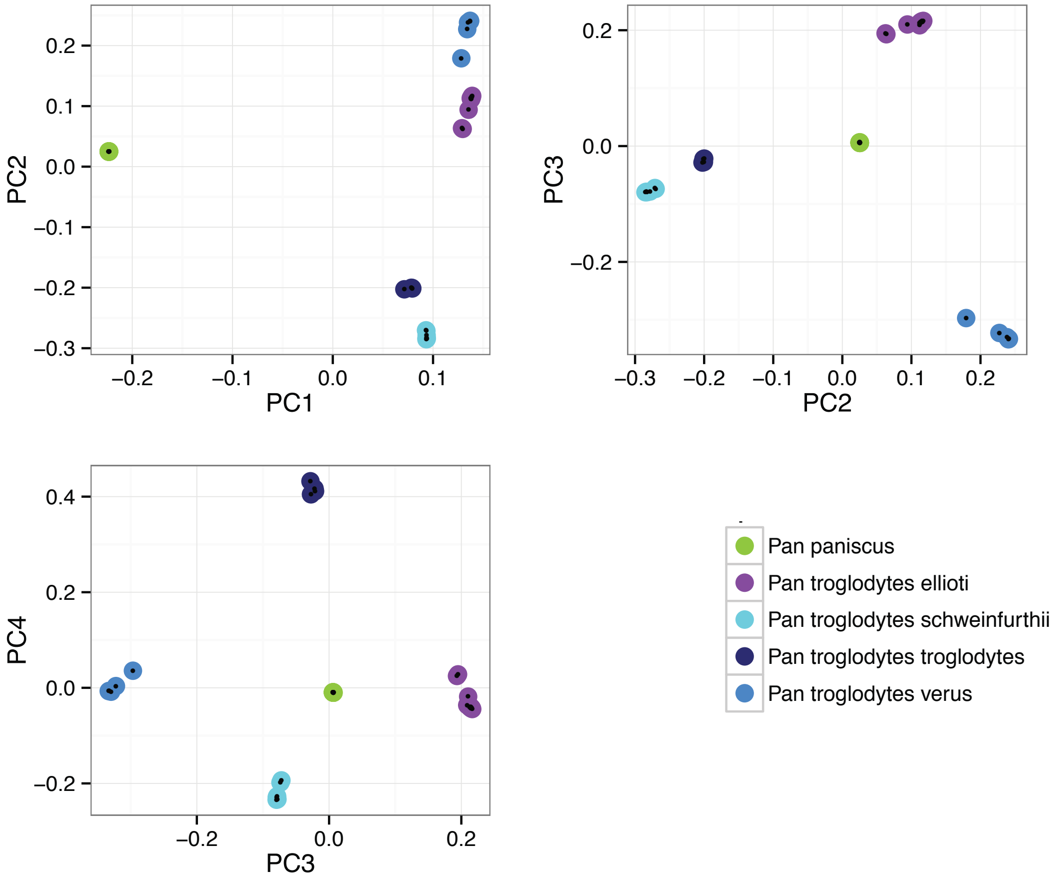
Suppl. Figure 8.2.4 – *PCA of bonobo genotypes.*



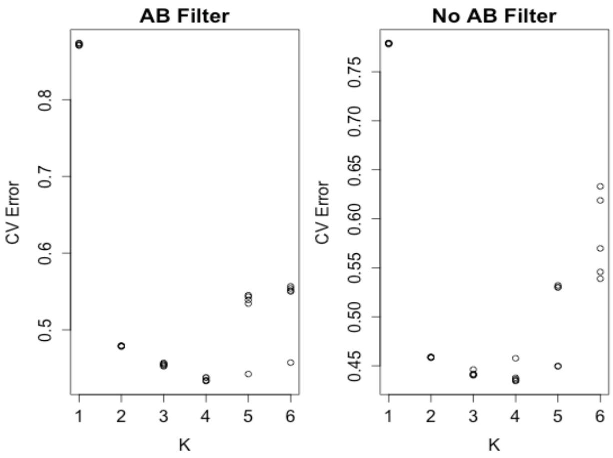
Suppl. Figure 8.2.5 – *ADMIXTURE cross-validation errors for bonobos.*

Chimpanzee and Bonobo

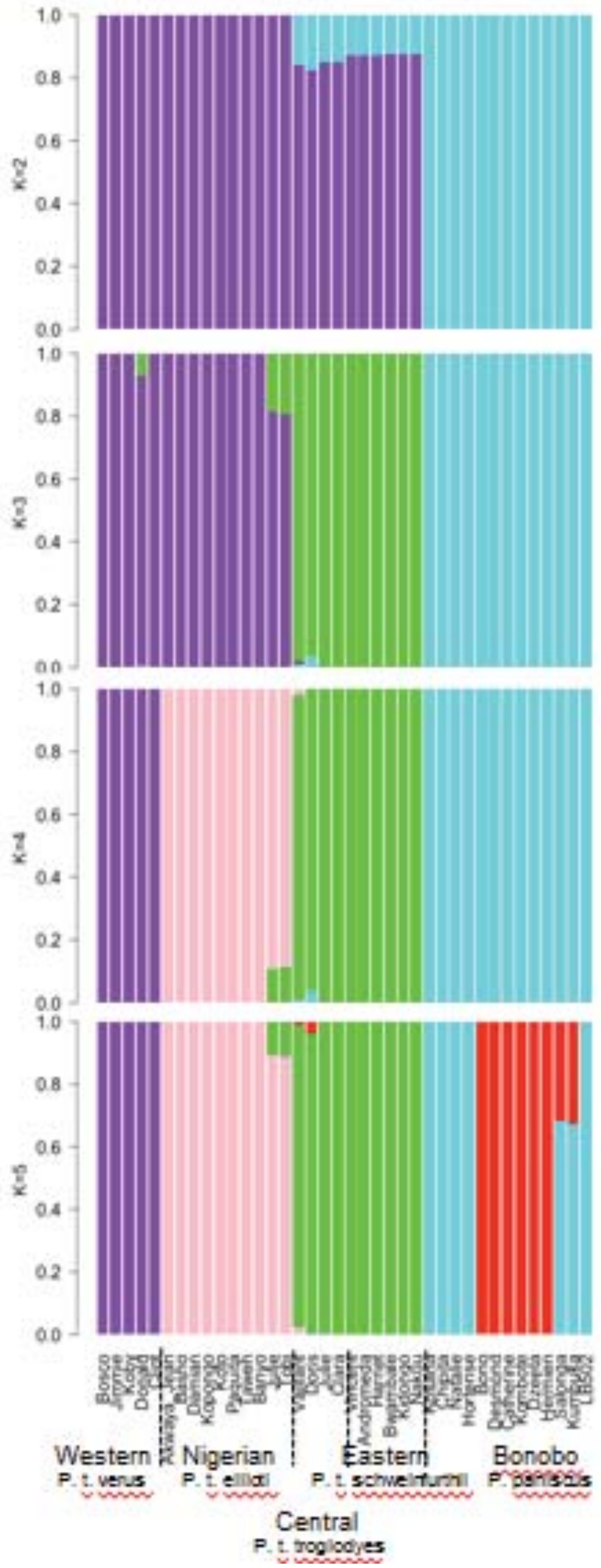
We also performed a joint analysis of the combined chimpanzee and bonobo data. In this sample set, we find that the first four PCs are significant (p = 0.00142, p = 7.93e-06, p = 1.67e-5, and p = 0.000169). PC1 accounts for 43.2% of the total variance and separates bonobos from chimpanzees while the higher PCs separate out the chimpanzees (**Suppl. Figure 8.2.6**). ADMIXTURE cross-validation shows the lowest assignment error at K = 4 (**Suppl. Figure 8.2.7**). At K = 2, chimpanzees and bonobos are separated, with the Central and Eastern chimpanzees showing ~12.5% component of shared ancestry with bonobos. At higher K, the Central and Eastern chimpanzees pull out as their own component. Interestingly, K = 5, which has a worse cross-validation error than K = 4, separates bonobos into two groups rather than dividing the Central and Eastern chimpanzee populations. (**Suppl. Figure 8.2.8**)



Suppl. Figure 8.2.6 – *PC analysis of combined chimpanzee and bonobo data.*



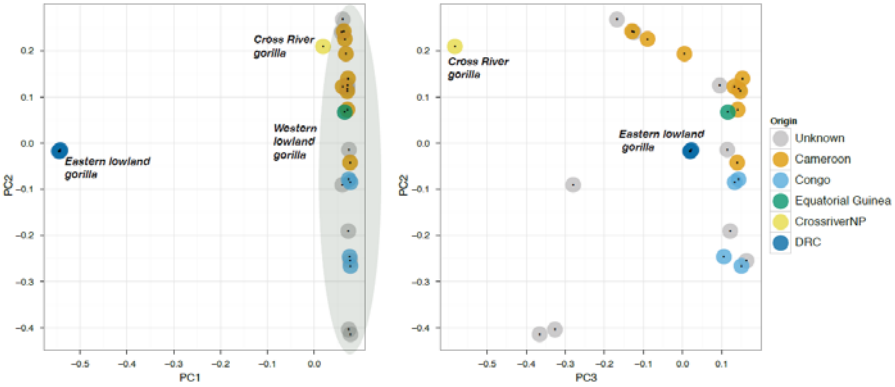
Suppl. Figure 8.2.7 – *Admixture cross-validation errors for combined chimpanzee and bonobo.*

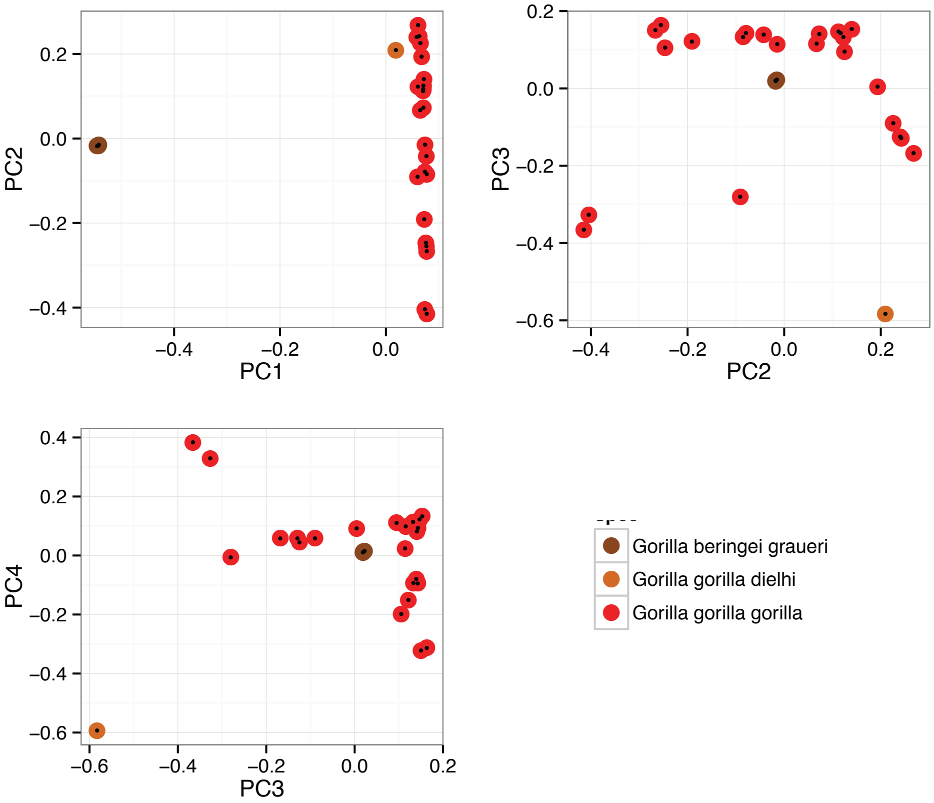

K=2
K=3
K=4
K=5
1.0
0.8
0.6
0.4
0.2
0.0
Bosco
Jimmie
Koby
Donald
Clint
Akwaya_Jean
Basho
Damian
Kopongo
Koto
Paquita
Taweh
Banyo
Julie
Tobi
Vaillant
Doris
Julie
Clara
Vincent
Andromeda
Harriet
Bwambale
Kidongo
Nakuu
Alfred
Chipita
Natalie
Hortense
Bono
Desmond
Catherine
Kombote
Dzeeta
Hermien
Salonga
Kumbuka
LB502
Western
P. t. verus
Nigerian
P. t. ellioti
Eastern
P. t. schweinfurthii
Bonobo
P. paniscus
Central
P. t. troglodyes

Suppl. Figure 8.2.8 – *Admixture cluster-membership for combined chimpanzees and bonobos. The assignments at K = 5 are not supported by the cross-validation metric.*

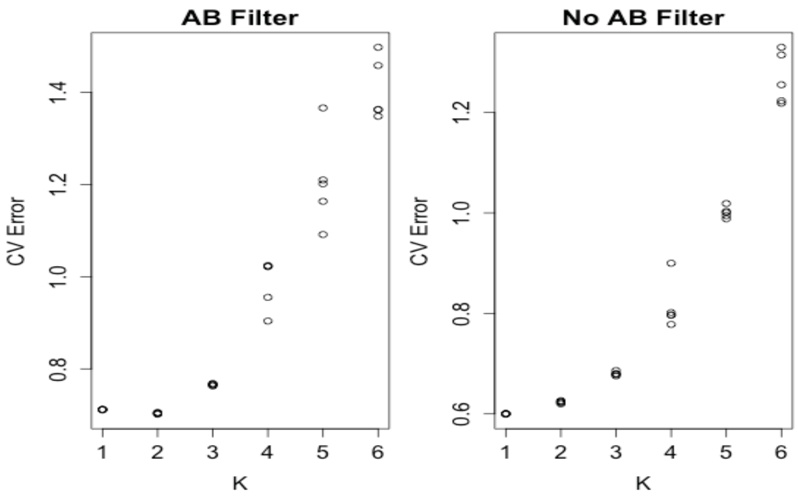
Gorilla

Analysis of the gorilla genotypes identifies two PCs that are significant. The first component (13.5% of the variance, p = 8.45e-12) separates the Eastern lowland from the Western lowland gorillas (**Suppl. Figure 8.2.9**). The Eastern lowland and the Cross River gorilla are arrayed out along the second component (6.6% of the variance, p = 2.64e-07). Cross-validation indicates that K = 2 has the lowest error rate, although we note that without the AB filter, cross-validation suggests that K = 1 is the best fit. At K = 2, Eastern and Western lowland gorillas are separated, with the single Cross River sample showing some evidence of a minor component shared with the Eastern gorillas (**Suppl. Figure 8.2.10**), which can also be inferred by the position of this sample along PC1 (**Suppl. Figure 8.2.9**).

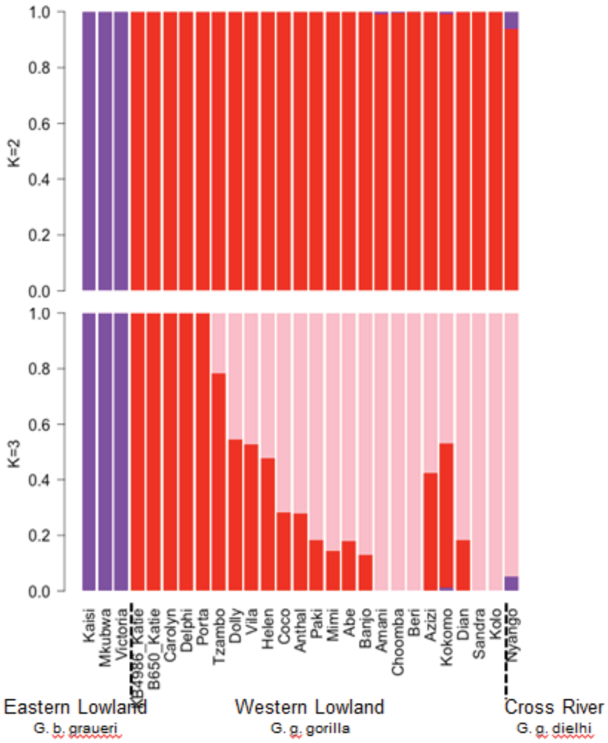




Suppl. Figure 8.2.9 – *PCA of gorillas showing the origin of each sample. Notice that PC2 tends to separate samples according to the origin of gorillas.*



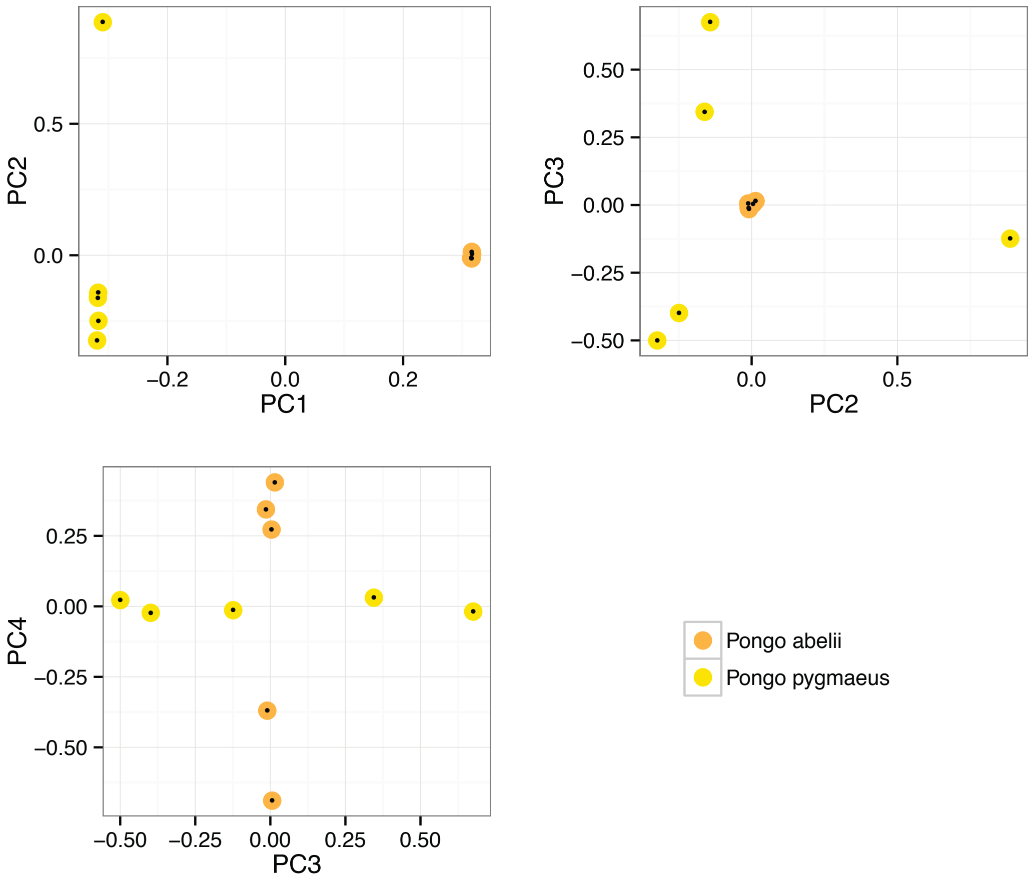
Suppl. Figure 8.2.10 – *Cross-validation results for gorillas.*



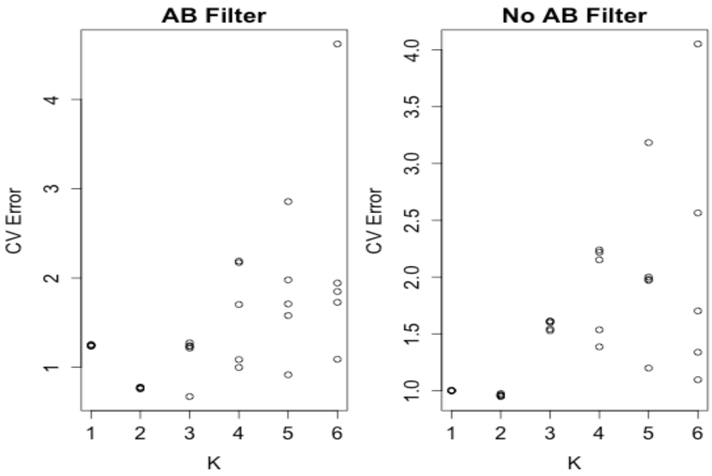
Suppl. Figure 8.2.11 – *ADMIXTURE cluster-membership for gorillas. The cross-validation metric does not support the selection of K = 3.*

Orangutan

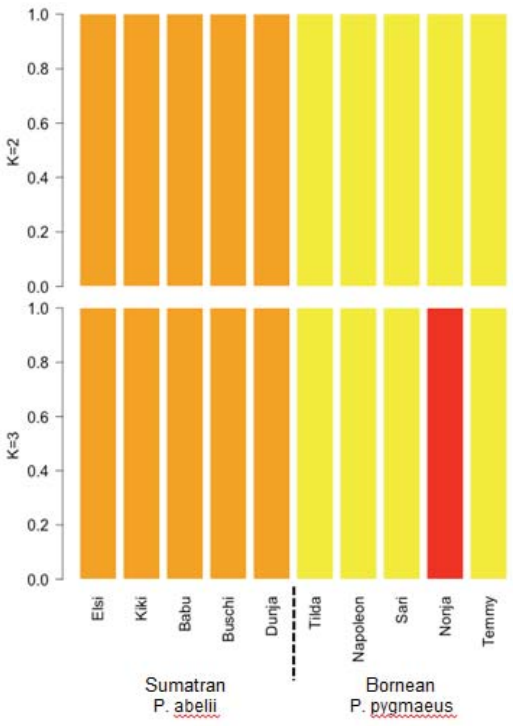
For the orangutans, only the first PC is significant. It accounts for 72% of the total variance and separates the Sumatran and Bornean samples. Most ADMIXTURE runs show K = 2 as having the lowest cross-validation error (**Suppl. Figure 8.2.12**), but a single run supports K = 3 with a single Bornean individual pulling out as a separate component (**Suppl. Figures 8.2.13 and 8.2.14**).



Suppl. Figure 8.2.12 – *PC analysis of orangutans.*



Suppl. Figure 8.2.13 – *Cross-validation results for orangutans.*



Suppl. Figure 8.2.14 – *ADMIXTURE cluster-memberships for orangutans. The cross-validation metric supports K = 3 in only one of five runs.*

8.3. Admixture deconvolution

Jeffrey M. Kidd, Joanna L. Kelley, Heng Li

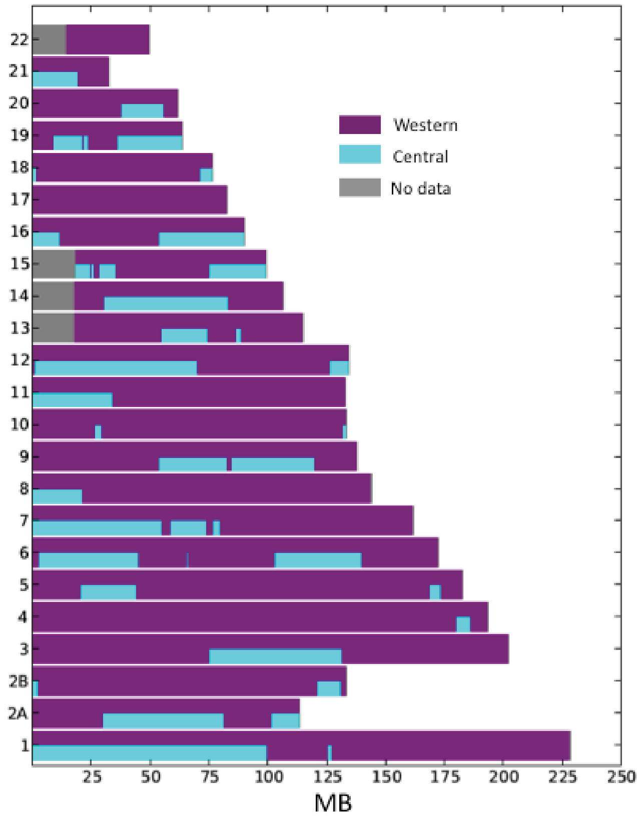
Analysis with frappe and ADMIXTURE indicated that Donald is a hybrid with ancestry from Western and Central chimpanzees and identified three Nigeria-Cameroon chimpanzees (Banyo, Julie, and Tobi) as having an unusual ancestry component. To investigate this further, we performed admixture deconvolution to infer the ancestral origin of each position along the genome of these samples. We based this analysis on the autosomal SNP calls identified based on mapping to the chimpanzee reference genome and used HAPMIX v2[23] along with the genetic map previously inferred for Western chimpanzees[24]. We ran HAPMIX in diploid mode, with the mutation parameter increased by a factor of 10 from the default value, as recommended for use on resequencing data instead of genotypes from SNP arrays.

For Donald, as source populations we used haplotypes from the four Western and four Central chimpanzees. We filtered the output from HAPMIX to require that tracts be at least 1 cM long and contain 1,000 SNPs in order for an ancestry switch to occur. Overall, we assign 15.5% of the genome to Central ancestry, and note that one complete set of chromosome homologs is inferred to be entirely Western in origin.

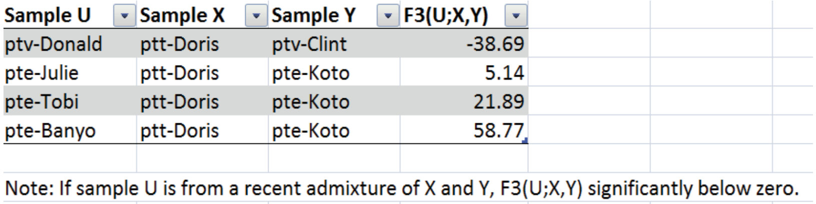
Min 1,000 SNPs and 1 cM			
Ancestry	Tracts	Bp	FractionBp
Central (P. t. troglodytes)	41	863,907,777	15.5%
Western (P.t. verus)	71	4,694,834,609	84.5%

Suppl. Table 8.3.1 – *Local ancestry assignments for Donald based on HAPMIX.*

We attempted the same analysis for Banyo, Julie, and Tobi using as reference panels the seven remaining Nigeria-Cameroon chimpanzees and either the four Central chimpanzees alone or a combined set of 10 Central and Eastern chimpanzees. We were unable to identify ancestry tracks, suggesting that these three samples are not the result of a recent admixture with Central chimpanzees.



Suppl. Figure 8.3.1 – *Ancestry painting for Donald.*



Sample U	Sample X	Sample Y	F3(U;X,Y)
ptv-Donald	ptt-Doris	ptv-Clint	-38.69
pte-Julie	ptt-Doris	pte-Koto	5.14
pte-Tobi	ptt-Doris	pte-Koto	21.89
pte-Banyo	ptt-Doris	pte-Koto	58.77

Note: If sample U is from a recent admixture of X and Y, F3(U;X,Y) significantly below zero.

Suppl. Figure 8.3.2 – *Results from the F3 statistic (Patterson et al. 2012).*

8.4. AIMs

Javier Prado-Martinez, Gabriel Santpere, Peter H. Sudmant, Jessica Hernandez-Rodriguez, Belen Lorente, Irene Hernado-Herraez, Arcadi Navarro, Evan E. Eichler, Tomas Marques-Bonet

Ancestry informative markers (AIMs) are variants that exhibit large allele-frequency differences between populations. These variants can accurately predict the population ancestry even when using a small subset of markers and can be extremely useful in conservation genetics and breeding programs of great apes. This resource may be a starting point towards the study of population genetics at a genome-wide level and from it we have retrieved the variants with a major degree of differentiation between the populations within each genus. Given the small sample sizes we could obtain from specific populations, we performed further analysis on whether these predicted AIMs are informative.

Expectation of AIMs FDR

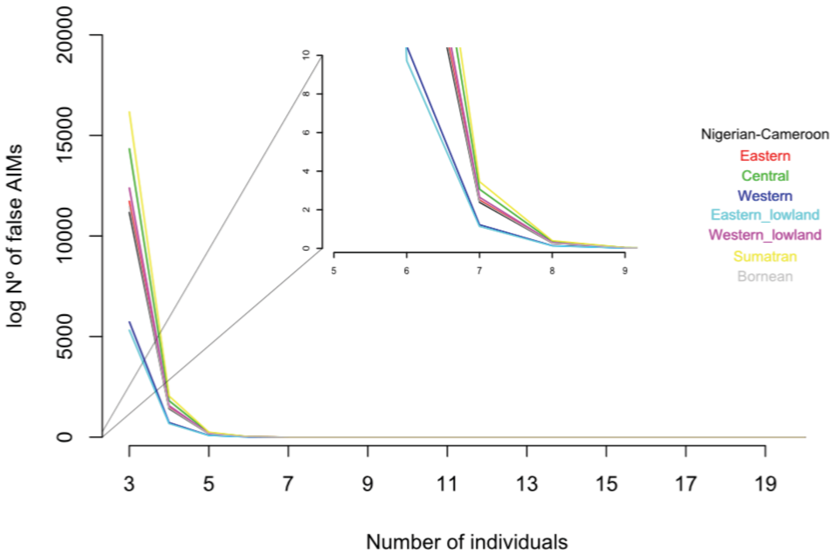
To predict the fixation of our AIMs, they were defined as fixed alleles in a particular species, while all other species being compared carry the other allele. That accounts for divergence between species but, given a low sample size, a number of segregating sites may mimic fixed differences. We estimated how many SNPs could be expected to resemble fixed differences with different sample sizes and diversity values. Given a segregating site (derived alleles), we calculated the probability of extracting one homozygous individual, considering Hardy-Weinberg equilibrium, from a neutral unfolded site frequency spectrum of 46 alleles, the maximum sample size used in this study. We use this probability to calculate the binomial probability of finding all sampled individuals of a given species homozygous for the derived allele in this site. Finally, the number of expected false AIMs in sampled individuals from one species is the product of all segregating sites identified in these individuals by the calculated probability of finding one false AIM.

Species	Sample size	Prob(All-HZ)	SS	AIMs Observed*	AIMs Expected	% false AIMs
Nigeria-Cameroon	10	2.9E-10	12,605,585	1,941	0	0
Eastern	6	1.9E-06	11,264,879	1,117	21	2
Central	**4**	**1.5E-04**	**11,820,858**	427	1,814	425
Western	4	1.5E-04	4,729,933	136,061	726	1
Eastern lowland	3	1.4E-03	3,866,117	278,190	5,330	2
Western lowland	23	1.2E-22	17,314,403	3,009	0	0
Sumatran	5	1.7E-05	14,543,573	446,133	248	0
Bornean	5	1.7E-05	10,321,213	733,535	176	0

Suppl. Table 8.4.1 – *Expected proportion of false AIMs as the product of the probability of extracting all individuals homozygous for one allele (Prob(All-HZ)) (for a given species sample size) by the number of segregating sites. *Derived alleles oriented with hg18.*

We found that for species with a very low sample size but higher diversity, such as Central chimpanzees, the number of expected AIMs, under neutrality assumptions, surpasses the number AIMs reported (**Suppl. Table 8.4.1.**).

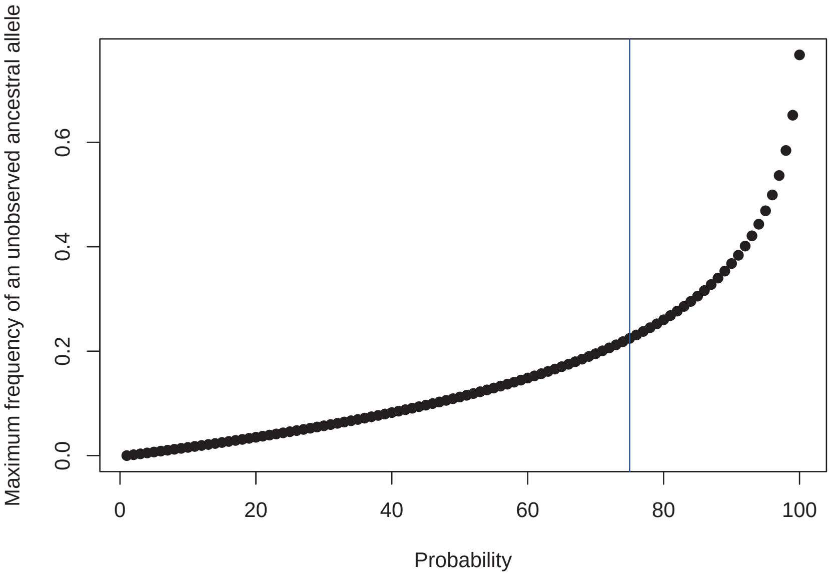
We have also projected the number of expected false AIMs sampling a different number of individuals for each species, scaling the number of segregating sites with the harmonic proportion. For all species, by sampling more than seven individuals, again considering a neutral site frequency spectra (SFS), no false AIMs are expected (**Suppl. Figure 8.4.1.**).



Suppl. Figure 8.4.1 – *Expected proportion of false AIMs as the product of the probability of extracting all individuals homozygous for one allele (Prob(All-HZ)) (for a given species sample size) by the number of segregating sites. *Derived alleles oriented with hg18.*

Given a number of sampled alleles, all bearing the derived allele, it is also possible to estimate which is the maximum frequency in the population that could achieve the ancestral allele to be unobserved in our sampling. Considering a 95% probability, the highest possible allele frequency (F) for an unobserved ancestral allele could be estimated as $1\text{-}e^{-Fn} = 0.95$, according to the Poisson distribution. With our minimum sample size (n) of three individuals (6 chromosomes), then F≈50%. If we consider a probability of 99%, F grows to ≈77%. That means a derived allele could be in a frequency in the population as low as 23% and still create false AIMs if we only sample three individuals. However, it also means that 95% of the time, the allele frequency of the derived allele in the population for these false AIMs is higher than 50%.

In fact, in the 75% percentile, the derived allele frequency increases to 77.4% (**Suppl. Figure 8.4.2**), which would not invalidate these segregating sites completely to be used as AIMs, especially if we considered combinations of them.



Suppl. Figure 8.4.2 – *Quantile distribution of the maximum frequency that can be achieved by a possible unobserved ancestral allele at site with an apparent AIM. The blue line represents the 75% quantile.*

AIM validation

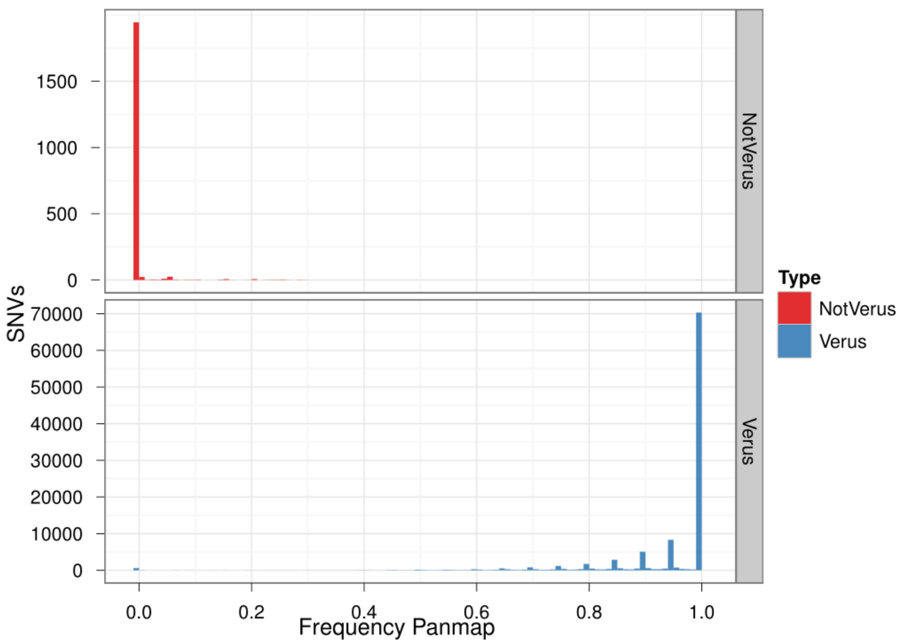
To complement this approach, we have also genotyped a set of 22 AIMs (~6 from each subspecies) in seven new individuals of known ancestry (1 *P.t.troglodytes*, 3 *P.t.schweinfurthii* and 3 *P.t.verus*) and a wild-born individual with unknown origin, using Sanger capillary sequencing. We genotyped 171 positions (5 sequences failed), 44 of which were tested for AIMs specific to the group and 127 were used as a negative controls (AIMs not specific). 96.9% of the negative controls have been correctly assigned (123/127). All but four sequences confirmed the presence of these alleles among these populations but they are mostly polymorphic in *P.t.schweinfurthii* and *P.t.troglodytes* due to the larger variation in these subspecies. Most of the specific AIMs are validated in *P.t.verus* and were determined as fixed variants in this group. As expected, the power of the AIM approach increased with the combination of them. Just as with 22 AIMs, all the individuals were correctly classified to their subspecies and we were able to determine the sample with unknown origin as a *P.t.schweinfurthii* (**Suppl. Table 8.4.2**).

			Allele		TROGLODYTES	SCHWEINFURTHII			VERUS			Unknown
			REF	AIM	**Noemie**	**Jac**	**Kina**	**Washu**	**Alice**	**Benita**	**Mike**	**Bihati**
AIMs	VERUS	chr1:156246124	G	A	-	-	-	+/-	+	+	+	+/-
		chr2:83764348	T	A	-	-	-	-	+	+	+	-
		chr12:78540784	G	A	-	-	-	-	+	+	+	-
		chr14:41791060	T	A		-	-	-	+	+	+	-
		chr13:35136820	T	A	-	-	-	-	+	+	+	-
		chr3:183884588	C	T	-	-	-	-	+	+	+	-
	SCHWEINFURTHII	chr6:118685644	A	G	-	+	+	+	-	-	-	+
		chr10:93452416	T	C	-	+	+/-	+/-	-	-	-	+
		chr1:183152149	T	A		+	+/-	-	-	-	-	+
		chr10:93382776	G	A	-	+	+	+/-	-	-	-	+
		chr6:73412868	A	G	-	+/-	+/-	-	-	-	-	+/-
	ELLIOTI	chr12:122387054	A	G	-	-	-	-	-	-	-	-
		chr8:111236243	C	G	-	-	-	-	-	-	-	-
		chr12:54999690	G	A	-	-	+/-	-	-	-	-	-
		chr1:179913228	G	C	-	-	-	-	-	-	-	-
		chr2:108857399	T	C	-		-	-	-	-	-	
	TROGLODYTES	chr8:139513779	G	A	-	-	-	-	-	-	-	-
		chr11:114695783	A	T	+	-	-		-	-	-	-
		chr9:271906	A	G	+	-	-	-	-	-	-	-
		chr3:34545483	G	T	-	-	-	-	-	+/-	-	-
		chr22:35007755	T	C	+/-	-	-	-	-	-	-	-
		chr14:94709629	A	T	+/-	-	-	-	-	-	-	-
Distances	Distance P.t.verus 22 AIMS				16	21	20	15	0	1	0	20
	Distance P.t.schweinfurthii 22 AIMS				14	1	4	7	22	23	22	2
	Distance P.t.ellioti 22 AIMS				16	17	16	15	22	23	22	18
	Distance P.t.troglodytes 22 AIMS				6	21	20	15	24	23	24	22

Suppl. Table 8.4.2 – *Sanger capillary genotyping of 22 predicted AIMs in eight new individuals. The global distance to the combination of AIMs for each subspecies (bottom) allows us to unambiguously classify each individual to the proper population. The allele REF corresponds to the allele for which we are not predicting to be at high frequency in the population we are testing, i.e., a "+" in the table always correspond to the AIM column. Blank spaces denote that the sequencing did not work.*

Whole-genome validation

Finally, we have also computationally genotyped for our set of AIMs the 10 chimpanzee genomes analyzed in a previous work (Auton et al. Science 2012). All the individuals but one are wild born and they belong to *Pan troglodytes verus* subspecies. Most of the non-verus AIMs (~95%) are not present in these individuals and the remaining 5% are present at very low frequencies. The *verus*-specific AIMs are mostly fixed (~70% of them) and the others are found at high frequencies, i.e., ~90% of these AIMs are found with an allele frequency (AF) higher than 0.8 in these samples. (**Suppl. Table 8.4.3**).



Suppl. Table 8.4.3 – *Frequency of P.t.verus AIMs and non-P.t.verus AIMs in the 10 chimpanzee individuals from (Auton et al Science 2012). Almost all the non-verus AIMs are absent or at low frequency (top panel) whereas the verus-specific AIMs (bottom panel) are fixed (70%) or at high frequency.*

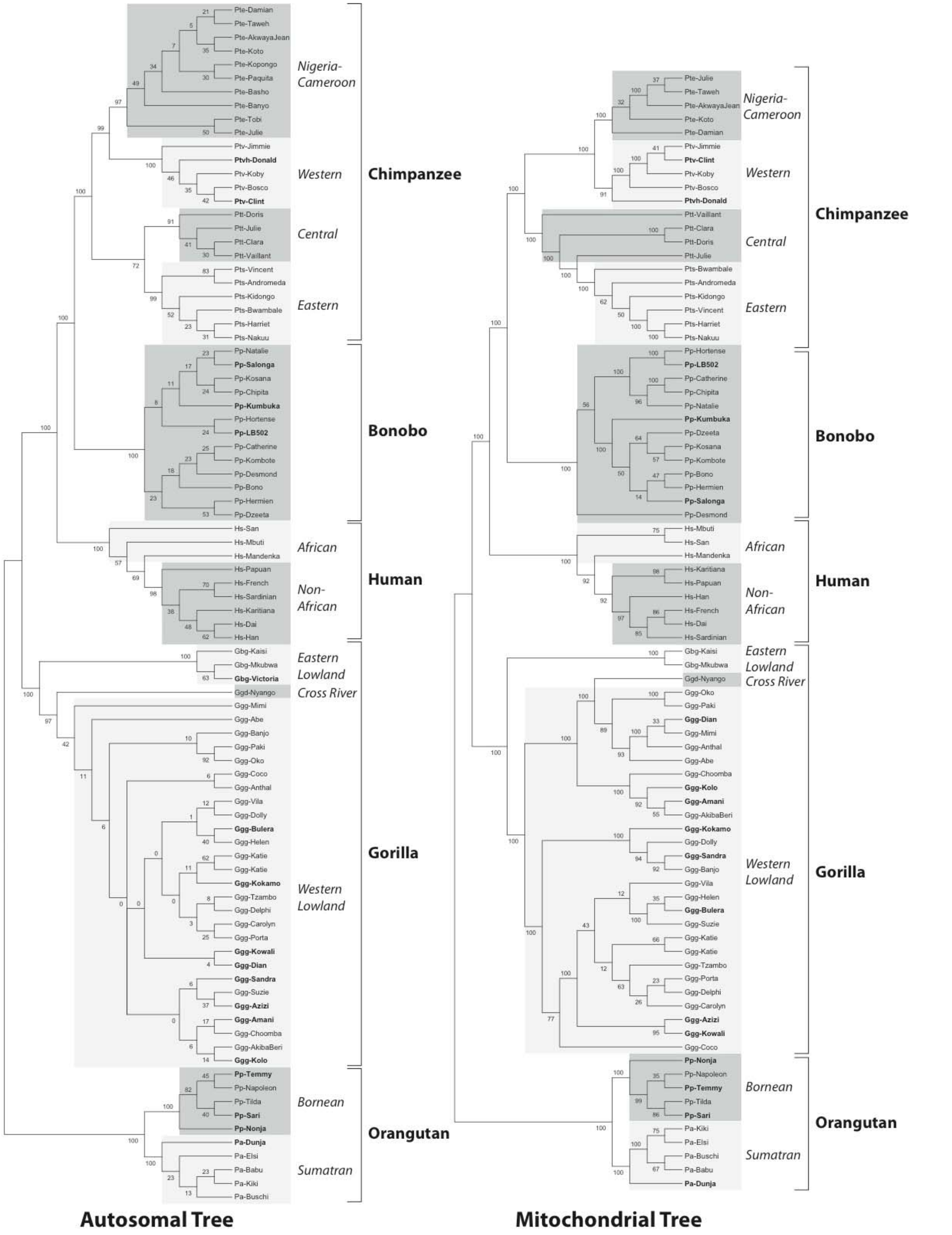
Nevertheless, overfitting is definitely an important consideration in light of our small sample size and this effect is most pronounced for subspecies with the greatest diversity. While our AIMs provide a starting point towards further characterization of unknown samples (especially for Nigeria-Cameroon and Western chimpanzees), we have minimized discussion of conservation and the use of these markers until more experimental validation emerges.

8.5. Phylogeny

Peter H. Sudmant, Belen Lorente-Galdos, Evan E. Eichler

We created an autosomal phylogeny based on a consensus of neighbor-joining trees (n=560) constructed from non-overlapping 5 Mbp blocks across the genome, whereas mitochondrial phylogeny was assessed using a maximum likelihood method from *de novo* assembled mitochondrial genomes (Section 7). The autosomal tree shows separate monophyletic groupings for each species/subspecies designation (**Suppl. Figure 8.5.1**) and support a split of extant chimpanzees into two groups. Nigeria-Cameroon and Western chimpanzees form a monophyletic clade (>97% of all autosomal trees). Central and Eastern chimpanzees form a second group (72% of the genomic trees separate these subspecies). The topology of mitochondrial phylogeny supports this bipartite division with the exception that Eastern chimpanzee and Central chimpanzees are grouped into a single clade. This difference may be consistent with the evolutionary stochasticity of a single locus and or sex-specific processes associated with the maternal transmission of the mitochondria. The single Cross River gorilla genome also shows some evidence of genetic differentiation from the Western lowland gorillas—although only 42% of the autosomal trees suggest it as a potential outgroup of all Western lowland gorillas and the mtDNA tree groups it with other Western gorillas. There is evidence of additional Western lowland gorilla substructure based on the tree topology and additional analyses (see PCA and PSMC). As expected, orangutan species cluster into two distinct monophyletic clades with complete support.

Suppl. Figure 8.5.1 – Phylogeny. *(Below) Cladograms constructed from autosomal and mitochondrial sequence illustrate the relationships between the individuals sequenced in this study. The autosomal tree is constructed by neighbor-joining trees on genetic distance estimates of 5 Mbp autosomal segments. Confidences represent the proportion of autosomal segments supporting the topology. The mitochondrial tree is derived by maximum likelihood from assembled mitochondrial haplotypes. Bootstrap confidence values are displayed for clades with >50% confidence. In bold we highlight captive individuals.*


Pte-Damian
Pte-Taweh
Pte-AkwayaJean
Pte-Koto
Pte-Kopongo
Pte-Paquita
Pte-Basho
Pte-Banyo
Pte-Tobi
Pte-Julie
Nigeria-Cameroon
Ptv-Jimmie
Ptvh-Donald
Ptv-Koby
Ptv-Bosco
Ptv-Clint
Western
Ptt-Doris
Ptt-Julie
Ptt-Clara
Ptt-Vaillant
Central
Pts-Vincent
Pts-Andromeda
Pts-Kidongo
Pts-Bwambale
Pts-Harriet
Pts-Nakuu
Eastern
Chimpanzee
Pp-Natalie
Pp-Salonga
Pp-Kosana
Pp-Chipita
Pp-Kumbuka
Pp-Hortense
Pp-LB502
Pp-Catherine
Pp-Kombote
Pp-Desmond
Pp-Bono
Pp-Hermien
Pp-Dzeeta
Bonobo
Hs-San
Hs-Mbuti
Hs-Mandenka
African
Hs-Papuan
Hs-French
Hs-Sardinian
Hs-Karitiana
Hs-Dai
Hs-Han
Non-African
Human
Gbg-Kaisi
Gbg-Mkubwa
Gbg-Victoria
Eastern Lowland
Ggd-Nyango
Cross River
Ggg-Mimi
Ggg-Abe
Ggg-Banjo
Ggg-Paki
Ggg-Oko
Ggg-Coco
Ggg-Anthal
Ggg-Vila
Ggg-Dolly
Ggg-Bulera
Ggg-Helen
Ggg-Katie
Ggg-Katie
Ggg-Kokamo
Ggg-Tzambo
Ggg-Delphi
Ggg-Carolyn
Ggg-Porta
Ggg-Kowali
Ggg-Dian
Ggg-Sandra
Ggg-Suzie
Ggg-Azizi
Ggg-Amani
Ggg-Choomba
Ggg-AkibaBeri
Ggg-Kolo
Western Lowland
Gorilla
Pp-Temmy
Pp-Napoleon
Pp-Tilda
Pp-Sari
Pp-Nonja
Bornean
Pa-Dunja
Pa-Elsi
Pa-Babu
Pa-Kiki
Pa-Buschi
Sumatran
Orangutan
Autosomal Tree
Pte-Julie
Pte-Taweh
Pte-AkwayaJean
Pte-Koto
Pte-Damian
Nigeria-Cameroon
Ptv-Jimmie
Ptv-Clint
Ptv-Koby
Ptv-Bosco
Ptvh-Donald
Western
Ptt-Vaillant
Ptt-Clara
Ptt-Doris
Ptt-Julie
Central
Pts-Bwambale
Pts-Andromeda
Pts-Kidongo
Pts-Vincent
Pts-Harriet
Pts-Nakuu
Eastern
Chimpanzee
Pp-Hortense
Pp-LB502
Pp-Catherine
Pp-Chipita
Pp-Natalie
Pp-Kumbuka
Pp-Dzeeta
Pp-Kosana
Pp-Kombote
Pp-Bono
Pp-Hermien
Pp-Salonga
Pp-Desmond
Bonobo
Hs-Mbuti
Hs-San
Hs-Mandenka
African
Hs-Karitiana
Hs-Papuan
Hs-Han
Hs-French
Hs-Dai
Hs-Sardinian
Non-African
Human
Gbg-Kaisi
Gbg-Mkubwa
Eastern Lowland
Ggd-Nyango
Cross River
Ggg-Oko
Ggg-Paki
Ggg-Dian
Ggg-Mimi
Ggg-Anthal
Ggg-Abe
Ggg-Choomba
Ggg-Kolo
Ggg-Amani
Ggg-AkibaBeri
Ggg-Kokamo
Ggg-Dolly
Ggg-Sandra
Ggg-Banjo
Ggg-Vila
Ggg-Helen
Ggg-Bulera
Ggg-Suzie
Ggg-Katie
Ggg-Katie
Ggg-Tzambo
Ggg-Porta
Ggg-Delphi
Ggg-Carolyn
Ggg-Azizi
Ggg-Kowali
Ggg-Coco
Western Lowland
Gorilla
Pp-Nonja
Pp-Napoleon
Pp-Temmy
Pp-Tilda
Pp-Sari
Bornean
Pa-Kiki
Pa-Elsi
Pa-Buschi
Pa-Babu
Pa-Dunja
Sumatran
Orangutan
Mitochondrial Tree

8.6. Cross River gorilla

From a phylogenetic perspective, the Cross River gorilla genome is located at the base of all Western lowland gorillas. 42% of the autosomal trees suggest it as a potential outgroup of all Western lowland gorillas. In contrast, the mtDNA tree positions the Cross River individual within the diversity of the other Western gorillas. From the point of view of PCA, all the Western lowland gorillas are distributed along a gradient on the PC2 axis. The Cross River gorilla is more similar to the Cameroon Western lowland gorillas. However, it is the third component that clearly separates the Cross River gorilla from most Western gorillas. However, using ancestry (STRUCTURE), a separate signal is not observed for the Cross River gorilla subspecies until we establish six clusters (K=6) suggesting that the signal that separates the Cross River gorilla from the Western lowland gorillas is not clearly distinguished considering the genetic diversity in Western lowland gorillas. From the point of view of population dynamics, we infer a smaller effective population size for the Cross River gorilla when compared with the Western lowland gorillas beginning ~100 kya. Evidence of historical gene flow between Eastern lowland and Cross River gorillas is observed according to the D-statistic, ABC and divergence time estimators. Finally, we find that the Cross River sample shows reduced heterozygosity as a result of long and thus recent runs of homozygosity events as corresponds from a very limited number of individuals (300 individuals (IUCN Red List of Threatened Species)).

Section 9: Inbreeding

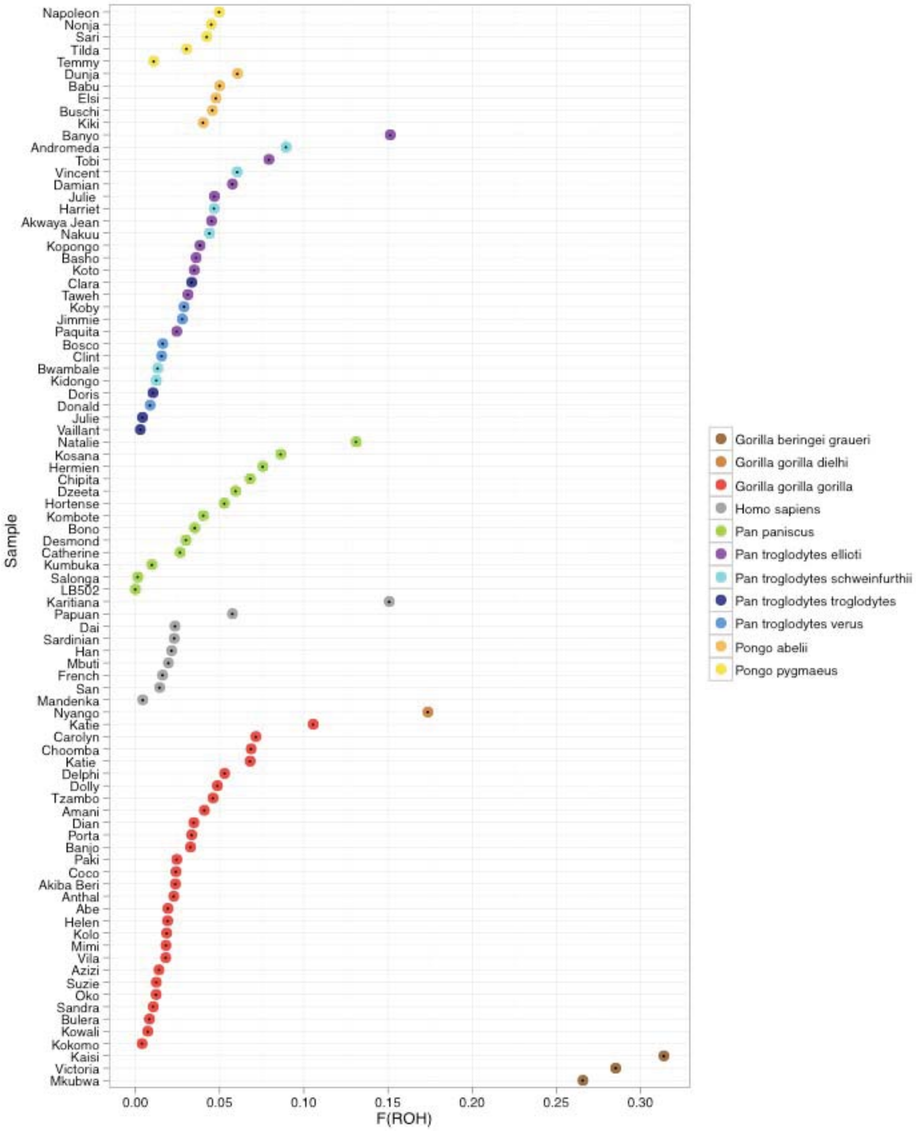
Javier Prado-Martinez, Krishna Veeramah, Tomas Marques-Bonet

To determine the amount of inbreeding, we calculated the genome-wide heterozygosity in windows of 1 Mbp with 200 Kbp sliding windows. We detected an excess of windows with very low heterozygosity in the density plots (**Suppl. Figure 6.2**) pointing to some extent of inbreeding. In order to estimate the cutoff values for inbreeding coefficients, we calculated a different local minima for different species since this value is variable depending in the heterozygosity and divergence to the human genome:

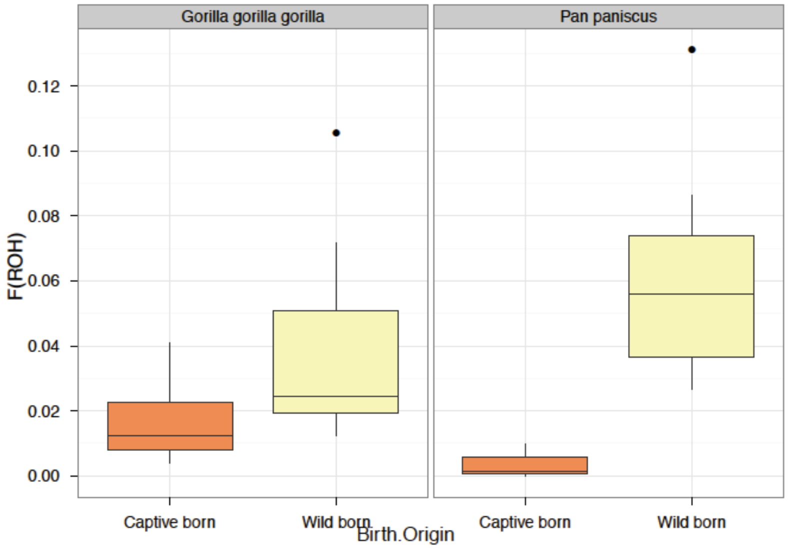
- Human, bonobo and Western chimpanzee: 0.00015
- Eastern, Central and Nigeria-Cameroon chimpanzees: 0.00025
- Gorillas and Bornean orangutan: 0.0003
- Sumatran orangutan: 0.0004

We then clustered together the neighboring regions to account for runs of homozygosity (ROH). From there we could estimate the percentage of the genome that is autozygous (F_{ROH}), which is a good measure of inbreeding as previously calculated[25]. We chose 1 Mbp as a threshold to consider a region as autozygous according to previous estimations[26], which identified that regions smaller than 0.5 Mbp are the result of background relatedness, while tracts larger than 1.6 Mbp are evidence of recent parental relatedness.

In general we observe a degree of background relatedness among all the groups studied with more recent inbreeding events in almost all the populations (**Suppl. Figure 9.1**). In humans around 2.5% of the genome shows a certain degree of autozygosity while the Karitiana and Papuan samples show a high degree of inbreeding due to the smaller population sizes among these groups. Bonobos and Nigeria-Cameroon chimpanzees show the highest amount of autozygosity among the *Pan* genus and can be explained by the patchy distribution and small population respectively. Lower levels of inbreeding are shown in Western and Central chimpanzees, but the small sample sizes in these populations may not reflect the actual trend of inbreeding. Among gorillas and bonobos, the captive-born have lower autozygosity than wild-born gorillas. We tested this trend with a Wilcoxon rank-sum test and the difference between the groups was significant in both species (p-values = 0.0067 and 0.0035, respectively) (**Suppl. Figure 9.2**). Moreover, the gorilla samples that have been geographically located in the Congo seem to have a higher amount of inbreeding than the rest of the gorillas. Eastern and Cross River gorillas show strikingly high levels of inbreeding as a possible result of population decline and habitat fragmentation[27]. Both orangutan populations show a moderate degree of inbreeding, comparable to those in Nigeria-Cameroon chimpanzees and bonobos, suggesting that the habitat loss in the Borneo and Sumatra islands may be having an effect in the random mating among this species.



Suppl. Figure 9.1 – *Summary of inbreeding coefficients (F_{ROH}) clustered by subspecies and ordered by inbreeding rank.*



Suppl. Figure 9.2 – *Comparison of inbreeding between captive and wild individuals.*

Section 10: Loss-of-function variants

Javier Prado-Martinez, Peter H. Sudmant, Maika Malig, Carl Baker, Belen Lorente-Galdos, Marcos Fernandez-Callejo, Can Alkan, Evan E. Eichler, Arcadi Navarro, Tomas Marques-Bonet

Loss-of-function SNVs

To characterize mutations in the coding sequence of the human gene models, we annotated the variants with the ANNOVAR software[28] with the RefSeq human gene models. In order to assign the precise lineage where the mutations occurred, we clustered the species sharing these mutations and assigned mutations to the human lineage in case all the species but human carry the mutation and to the human-Pan branch in case the mutations are shared between gorilla and orangutan species.

In total we obtained 806 stop-gain/stop-loss mutations (**Suppl. Figure 10.6, Table S4**). We compared the extant studies of premature termination codons and found 91 were previously reported[5,29–32]. We compared our predictions with the EPO alignment between human-chimpanzee-gorilla-orangutan and found 63% (506) of the variants were supported by all the assemblies. We performed 151 capillary Sanger validations for which 150 were correctly predicted (99.3%) (**Table S2**).

Indel variants

Frameshift mutations may account for a large proportion of gene disruptions, doubling the number of premature termination codons (PTCs) compared to SNVs in previous studies[29]. Starting from the mappings to the human reference assembly (hg18) and with the reads previously realigned with GATK in a multi-sample fashion (see **Suppl. Section 2.1**), we used GATK Unified Genotyper to produce an initial set of indel candidates. Then we applied the following filters for indels:

- QD < 2.0, ReadPosRankSum < -20.0, FS > 200.0
- Variants overlapping segmental duplications and tandem repeats (TRF from UCSC).

We finally removed indels clustering within 10 bp to remove possible artifacts on problematic regions. We focused on indels in coding regions and performed quality controls on this subset. We first assessed the distribution of sizes, expecting triplet multiplicity as a result of purifying selection to preserve the reading frame dividing our variants in fixed and polymorphic to account for selective pressure (**Suppl. Figure 10.1**). Interestingly, if we divide the indels into fixed and segregating, we can observe how purifying selection is stronger in fixed indels, given that most of the variants are triplet multiple. See **Suppl. Section 3.5** for the validations on indels.

Gene large deletions

Larger lineage-specific deletions were identified by discordant read-pair analysis using the VariationHunter software[33]. The approach that followed was similar to the procedure used by Ventura et al[34]. Calling was performed on a per-individual basis and the resulting individual callsets were merged within species by 50% reciprocal overlap criteria. All calls were then genotyped by read depth to confirm the deletions were indeed fixed among the individuals assessed in this study. Deletion events were further confirmed by a custom designed array comparative genomic hybridization (aCGH). A total of 30 aCGH experiments were performed using all nonhuman primate species and subspecies against the human reference NA12878. In total 374/382 (97.9%) aCGH experiments successfully confirmed the lineage-specific deletion event. Excluding aCGH experiments with fewer than two probes, 100% of experiments successfully validated. We identified a total of 96 lineage-specific deletions partially or completely overlapping coding exons.

We thus screened all mutations leading to loss-of-function events using the human gene models (RefSeq) and classified them into their evolutionary context within the great ape lineage (**Suppl. Figure 10.6**). For the first time we can provide a complete picture of events leading to significant changes in genes at different branches. In total we detected 1,982 fixed loss-of-function events in 1,481 genes. (**Table S4).**

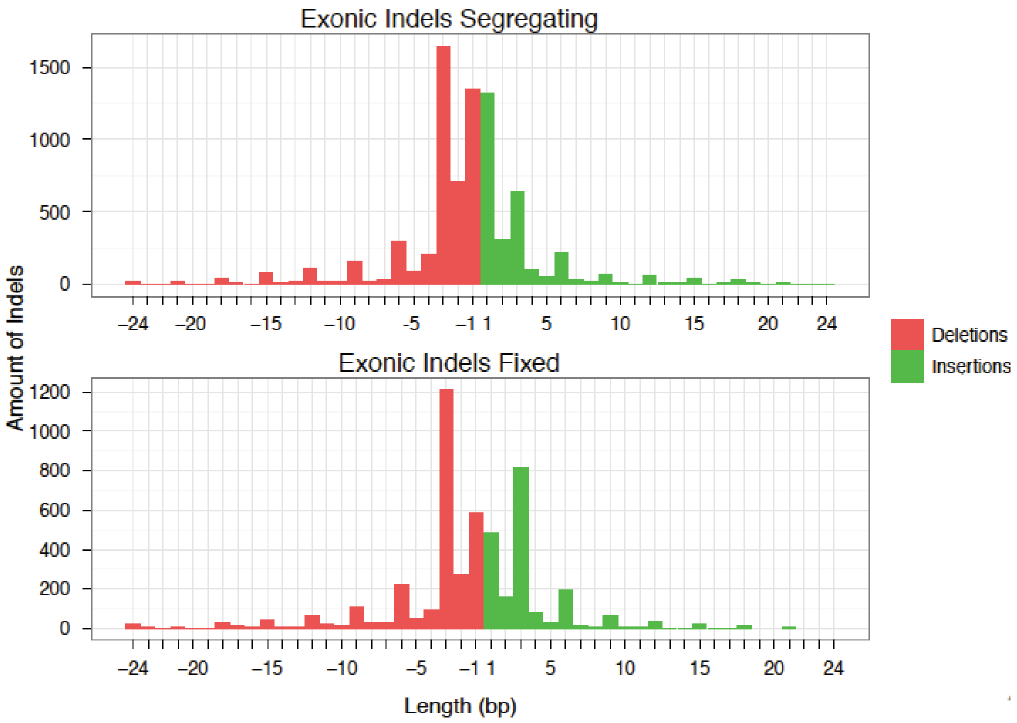
We then studied the position along the gene of these events. We divided the relative position of these mutations in bins of 5% of the human gene model and we plotted the amount of variants leading to frameshift/in-frame mutations (**Suppl. Figure 10.7, Suppl. Figures 10.2 and 10.3**). We detected an enrichment of disrupting events in the beginning and the end of the genes. These regions of the gene have been associated with a lower selective constraint given that the proportion of functional domains is smaller[29].

The genes we are reporting in **Table S4** may account for a significant number of important genes during great ape evolution and the gain/loss or modification of these proteins may have had a crucial impact during the phenotypic differentiation between the different lineages. Further work may be needed to perform functional characterization of the effect of these mutations, but this characterization of events is a step towards a better understanding of recent speciation of the human and great ape lineages.

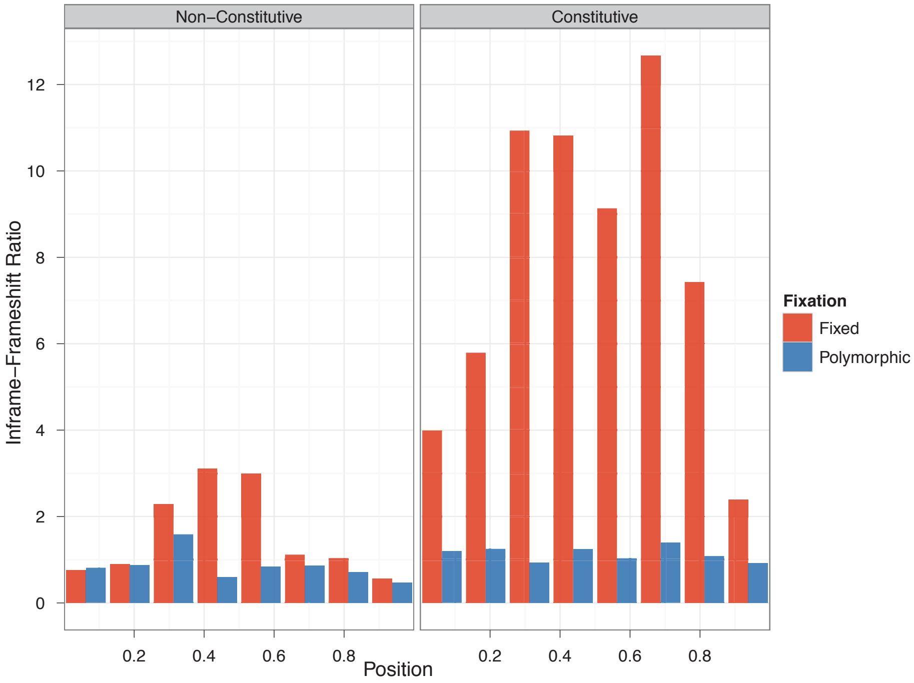
Less is more hypothesis

After the appearance of the less is more hypothesis[35], it has been long debated whether gene losses have an important role in the evolution. This has been hypothesized as being a very important engine of evolutionary change in the human lineage, based on several adaptations and phenotypes by using this mechanism[36]. For this reason we have tested this hypothesis along the hominid phylogeny. We studied the number of events detected in the different branches and compared them to the evolutionary distance as a function of the genetic divergence. **Suppl. Figure 10.4** shows the correlation between genetic distance and the amount of gene losses.

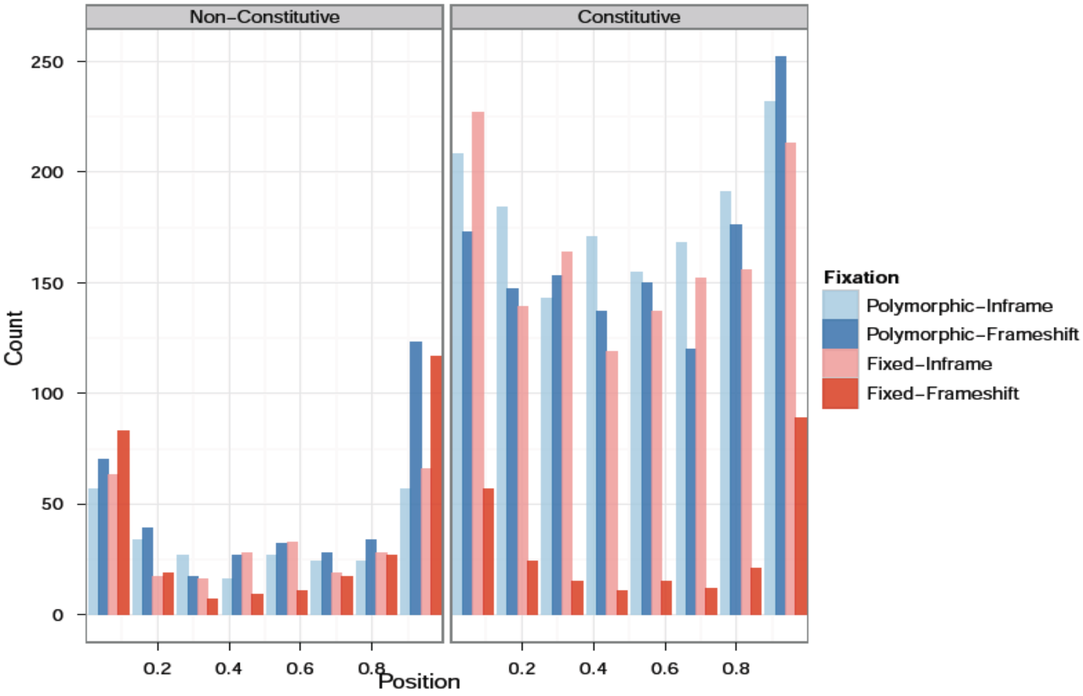
The main problem of this analysis is that we rely on the human gene models. This introduces a bias given that the pseudogenes that are specific to the human lineage are underrepresented and this can lead to an underestimation of the number of gene losses in the human lineage. We accounted for this including 67 human-specific pseudogenes from Wang et al[36]. We performed the same correlation accounting for the branches with more than 0.3% of divergence (Pongo-Homininae, Gorilla, Pan and Human branches), and we did not observe an excess of loss-of-function events during the human lineage (**Suppl. Figure 10.5**). We used a Maximum Likelihood Ratio test, to test whether the human branch has an excess of loss-of-function events or if all four lineages follow a single rate. We found that a single rate (39.7 losses per mut/Kbp) fits better these data.



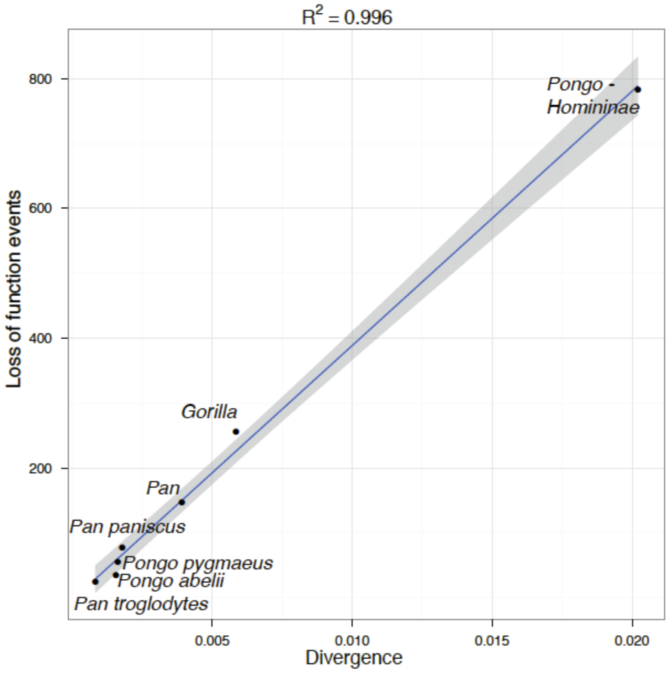
Suppl. Figure 10.1 – *Indel size distribution in coding regions. The effect of selection is stronger in fixed events; this is noticed by the larger proportion of events maintaining the reading frame compared to disruptive frameshift mutations.*



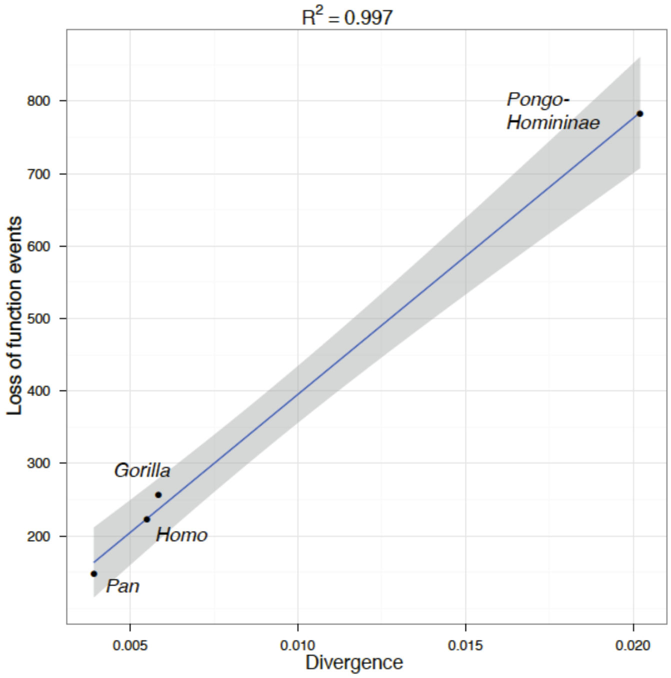
Suppl. Figure 10.2 – *Distribution of in-frame/frameshift ratio along the gene in 1:1 orthologous genes in primates. We consider constitutive the exons that appear in all the isoforms of the gene. Fixed variants are in red and polymorphic in blue.*



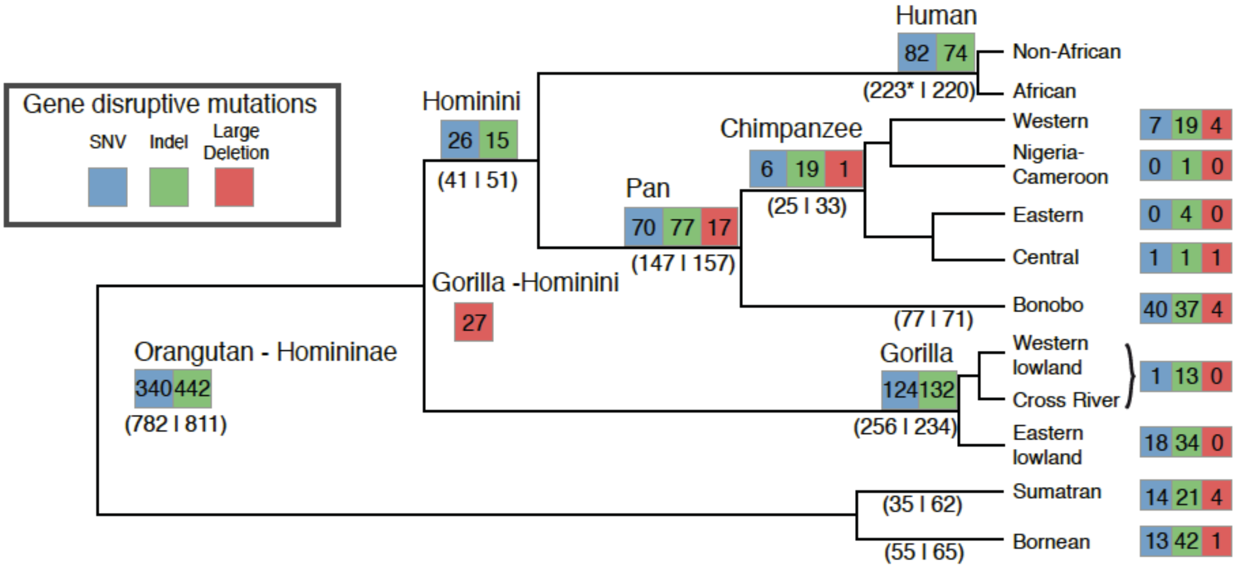
Suppl. Figure 10.3 – *Distribution of frameshift mutations across the gene positions. Both polymorphic and fixed mutations show an increase towards 5' and 3' of the gene models (1:1 orthologous used).*



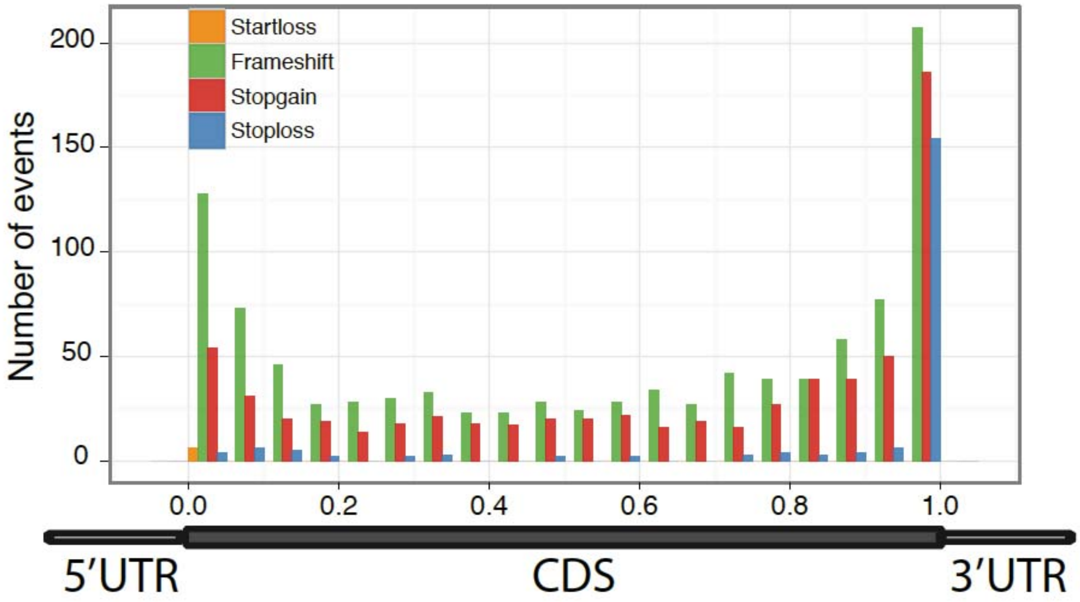
Suppl. Figure 10.4 – *Correlation between amount of gene loss events and genetic distance. The amount of loss-of-function events appears to appear at the same rate as the SNV divergence between species.*



Suppl. Figure 10.5 – *Correlation between amount of gene loss events and genetic distance in the branches larger than 0.003. We account for the Homo branch correcting for 67 non-processed pseudogenes. The human lineage does not appear to have accumulated an excess of loss-of-function events.*



Suppl. Figure 10.6 – *Mutations resulting in stop codon gains and losses fixed through great ape evolution are superimposed onto the great ape phylogeny. The numbers on each branch correspond to fixed substitutions, indels, or large deletions resulting in disruptions of the coding sequence. To correct for the human reference, mutations seen in all species, except human, were assigned to the human branch. Numbers in parenthesis show the observed and expected number of events per branch.*



Suppl. Figure 10.7 – *Genic position of disruptive events. Distributions are shown for startloss, frameshift, stopgain, and stoploss events in 5% bins throughout the gene. Both the beginning and end of the genes show an excess of loss-of-function mutations. Stoplosses in the middle of the gene are only predicted in the human and the hominini branches and indicate gene extensions in these lineages.*

We formally tested whether the rate of fixation of LoFs in the lineage of the great apes is 1) similar to that of substitutions per Kbp; and 2) shows any signs of being different in the human internal branch vs. the internal branches other great apes.

We conducted two different analyses. First, we followed the approach of Marques-Bonet *et al.*[37] and built a Likelihood Ratio Test framework that tests whether seven independent rates explain our observations significantly better than a single one. Secondly, we used a frequentist approach and presented rates in terms of observed vs. expected.

As a unit of time, we use, for each branch, either the number of substitutions per Kbp obtained from the same individuals or an estimate in units of Myrs of the length of each branch. The following table indicates the units of indel and time in the branches being tested. We use internal branches to avoid several problems, overfitting amongst them.

	Indels	**SNVs**	**Divergence*** Kbp	**Time**
Terminal abelii	21	14	1.55	0.88
Terminal Bonobo	37	40	1.78	1.3
Internal Chimpanzees	19	6	0.81	0.8
Internal Gorilla	132	124	5.84	8.06
Internal *Pongo*	442	340	20.21	15.92
Internal *Pan*	77	70	3.91	3.8
Terminal pygmaeus	42	13	1.62	0.88

To perform an LRT, we first obtained maximum-likelihood estimates for two different models. The simplest one assumes a single rate of accumulation of indels everywhere and the other one assumes that every branch has its own rate. Afterwards, we perform an LTR between the two models. We use 6 degrees of freedom since the second model has 6 more parameters.

Tabulated below are the results of the estimates and the p-value of the LOFS test that can be performed with the two units of time, Myrs or substitutions per Kbp.

LOFS / Myrs			
	Model 1 (all identical rate)	Model 2 (Seven different rates)	LTR p-value
One vs. Seven rates	λ = 43.52 LoFs/ Myrs	λ_{abe} = 39.77 λ_{bon} = 59.23 λ_{chi} = 31.25 λ_{gor} = 31.76 λ_{ora} = 49.12 λ_{pan}= 38.68 λ_{pyg} = 62.50	1.25×10^{-10}

LOFS / NumSubstperKb			
	Model 1 (all identical rate)	**Model 2 (Seven different rates)**	**LTR p-value**
One vs. Seven rates	λ = 38.55 LoFs/ NumSubstperKb	λ_{abe} = 22.58 λ_{bon} = 43.26 λ_{chi} = 30.86 λ_{gor} = 43.83 λ_{ora} = 38.69 λ_{pan}= 37.60 λ_{pyg} = 33.95	0.00344822

Using both units of time, several rates explain things better. Naturally, all the ML estimated rates in each branch are equivalent to their observed rate. Admittedly, however, we do have some overfitting.

Now we want to see if in any particular branch the observed rate is larger or smaller than the expected rate (considering as the expected the overall rate). We indicate the difference and the p-values corresponding to a test assuming a Poisson distribution of events per branch. This would be equivalent to asking: To which branches do we owe a larger deviation from a common expectation?

LOFS / NumSubstperKb		
Branches	O/E ratio	p-value
P.abelii	0.59	0.00275
P.paniscus	1.12	0.20972

Internal chimpanzees	0.80	0.09373
Internal gorilla	1.14	0.20972
Internal orangs	1.00	0.50753
Internal Pan	0.98	0.44321
P. pygmaeus	0.88	0.21053

Human branch

We applied the same test for human vs. the other three internal branches.

	Divergence* Kbp	LOFS
Human	5.55	223
GGO	5.84	256
Pan	3.91	147
Orang	20.21	782

The overall expected rate would be λ = 39.71 LOFS/ NumSubstperKb, which is not significantly different than the other four different rates (p-value 0.7349). In terms of O/E, it is again self-explanatory.

O/E approach		
LOFS / NumSubstperKb		
Branches	O/E ratio	p-value
Human terminal	1.02	0.43983
Shared with gorilla	1.10	0.26817
Shared with orang	0.97	0.43395
Shared with Pan	0.95	0.37174

RNAseq validation

To provide further evidence of whether these variants are found on expressed genes and carry the same mutations previously predicted on genomic data, we used RNAseq data[38]. We mapped these data to the hg18 allowing up to an indel size of 8 using TopHat splice junction mapper with the human gene models. Then we called variants using SAMtools roughly, without applying further filtering. We analyzed only the LoF events where the region is covered by least five reads to have enough support in the SNP and indel calling. We obtained >97% of validation in all LoF predicted with genomic data (**Suppl. Table 10.1**). This validation rate is concordant with the previous validations using Sanger capillary sequencing.

	SNV (%)	SNV (#)	INDEL (%)	INDEL (#)
Bonobo	100%	108 of 108	97.40%	75 of 77
Chimpanzee	98.90%	90 of 91	97.33%	73 of 75
Gorilla	100%	122 of 122	98.89%	95 of 96
Orangutan	97.84%	136 of 139	97.43%	76 of 78

Suppl. Table 10.1 – *Validation of fixed LoF mutations in RNAseq data from Brawand et al. where coverage (>4X) allow us to genotype according to transcriptome expression data.*

Section 11: X versus autosomes

August Woerner, Krishna R Veeramah, Michael F Hammer

For both autosomes and the X chromosome, genic regions for hg18 were defined based on the set-union (to take into account overlapping and alternative transcripts, etc.) of all genes defined in the RefSeq Gene Collection. 20 Kbp loci were then identified as previously reported[39]. In brief, a central nongenic locus was first found whose 5' and 3' tips are maximally distant, in genetic units, from the nearest genes. Successive non-overlapping 20 Kbp loci were identified walking towards genic boundaries from this central locus in both the 5' and 3' directions until the gene boundaries were reached. Genetic distances were defined based on the fine-scale recombination map of Hinch et al.[9] estimated using ancestry switches detected in African Americans (http://www.well.ox.ac.uk/~anjali/AAmap/). Loci in the pseudo-autosomal region of the X chromosome were excluded from the analysis.

Nucleotide diversity/divergence (π/D) was calculated for each locus for each subspecies. Species-specific masks were applied to the loci and thus not all loci contained 20 Kbp of callable sequence for this calculation and the total sequence considered may change between species. Loci with less than 5 Kbp of callable sequence were dropped from the analysis. Note that 20 Kbp was found to be of sufficient size to estimate π/D at a reasonable level of accuracy (i.e., contain sufficient segregating sites) but still represent a relatively confined interval of genetic distance.

For a given subspecies, D was taken as the average divergence of all individuals relative to the ancestral node of the primate phylogeny defined for this dataset (see section on ancestral allele calls). For each nucleotide position considered, an ancestral allele was chosen randomly, weighted by the relative probabilities of the four alleles at the ancestral node. For nucleotide positions segregating with two alleles in a subspecies where one of the alleles was the same as the ancestral node, a divergence value of 1 was assigned to all individuals with the derived allele. For nucleotide positions segregating where neither allele was the same as the ancestral node, we assumed that two mutations had occurred and that one mutation occurred on the background of the other (rather than occurring independently on the background of the ancestral allele). The first mutation to occur was chosen randomly, weighted by the relative allele frequencies (i.e., the major allele is more likely to have mutated first). Individuals inferred to have two changes from the ancestral allele were assigned a divergence value of 2. As the sample sizes were fairly restrictive, male X chromosomes were used in the estimate of π and D. This was achieved by assigning male genotypes to the most-frequent allele for SNPs on the X chromosome.

For each subspecies, loci were grouped into six bins ([0–0.05], [0.05–0.1], [0.1–0.2], [0.2–0.4], [0.4–0.8], [0.8–2.0]) of increase cM distance from the nearest genic regions. Mean π/D was then calculated for all loci within each bin. Bins were of increasing interval size (0.05cM in the first bin to 1.2cM in the final bin) to account for the number of loci available being reduced moving further away from genes, as described in Gottipati et al[40]. 95% Confidence intervals (CIs) for each bin were calculated using standard bootstrapping of loci, though we note that this will not fully take into account the interdependence between neighboring loci in the same bin and thus are likely somewhat anti-conservative. The ratio of X to autosomal diversity at each of the bin positions was calculated by dividing the mean π_X/D_X by the mean π_A/D_A.

Data quality filtering

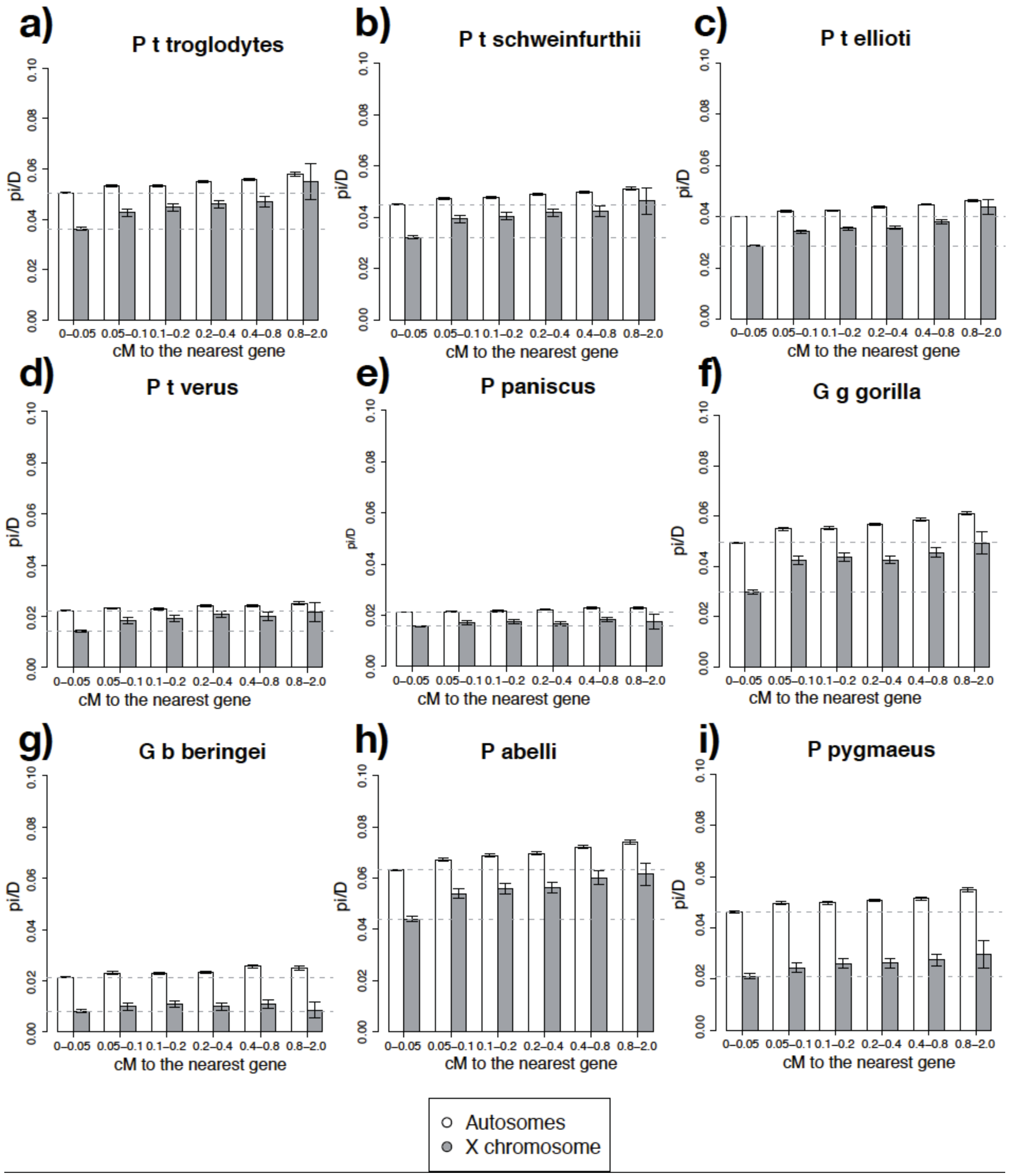
For the initial analysis, no coverage filters were applied to the data and for male samples haploid calls on the X chromosome were based on the allele with highest read depth (AD field from GATK Unified Genotyper). We also explored applying 5X coverage filters on the autosomes and X chromosomes and found no noticeable different in results without filters.

Contrasting autosome and X chromosome variation can inform us about population demography, sex-specific behaviors, and selective constraint. Therefore, we examined the level of both autosomal and X chromosomal diversity as a function of genetic distance from genes within each great ape subspecies. To control for variation in mutation rate among loci, diversity was divided by divergence from a common ancestral primate node. Consistent with previous studies in humans[39–41], both autosomal and X chromosome diversity increase when moving away from genes in nearly all subspecies (**Suppl. Figure 11.1**). In addition, X chromosome diversity generally increases at a faster rate, as seen previously in humans. Close to genes, X/A diversity is lower than the expected neutral value of 0.75 (**Suppl. Figure 11.1**), suggesting that the effect of purifying or positive selection is greater on the X chromosome. Such an effect is expected if novel mutations tend to be partially recessive because they are more quickly exposed to selection in hemizygous males. In contrast, X/A diversity is usually greater than 0.75 at regions far from genes. Given that these regions are expected to be less affected by selection, this pattern of diversity is consistent with an increased variance in male reproductive success for seven of the nine subspecies considered here.

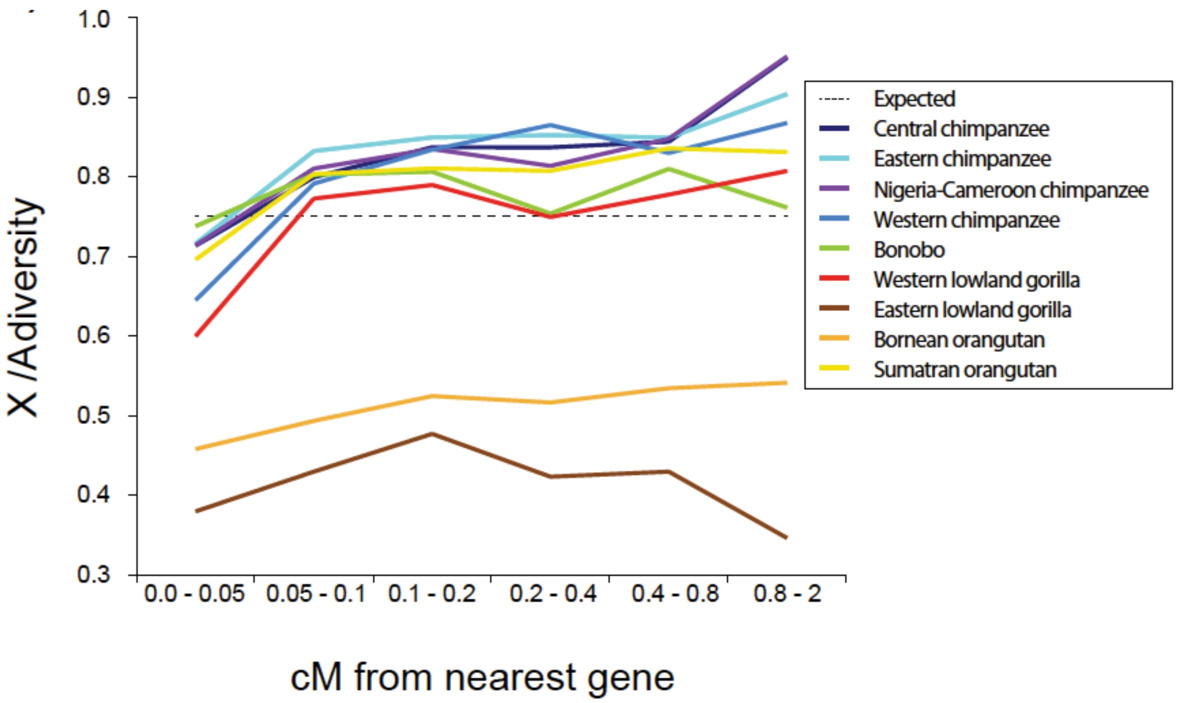
The overall rate of increase of X/A diversity is similar across all four chimpanzee subspecies despite a wide range of N_e estimates, suggesting similar distributions of fitness effects among species. Central, Eastern, and Nigeria-Cameroon chimpanzees show very similar levels of X/A diversity close to genes. Interestingly, the subspecies with the lowest effective population size, Western chimpanzees, exhibits substantially lower X/A diversity near genes (the CIs of Western chimpanzee X/A diversity do not overlap with any of the other three subspecies). Demographic effects are unlikely to be the cause of this as X/A diversity levels quickly recover as one moves away from genes. One possible explanation is that Western chimpanzees have experienced stronger selection on the X chromosome relative to the autosomes.

Patterns of X/A diversity are similar in Western gorillas and Sumatran orangutans. However, both Bornean orangutans and Eastern gorillas demonstrate substantially reduced levels of X/A diversity compared with expectations, regardless of distance from genes (e.g., mean values are between 0.54 and 0.35, respectively) (**Suppl. Figure 11.1.a, g, i**). This is consistent with theoretical predictions for a recent reduction in population size[42]. Another possible contributing factor is male-specific migration into these populations from other subspecies following their divergence, which would increase the number of breeding males (a similar hypothesis has been proposed in humans to explain the reduction in X/A diversity in non-Africans relative to Africans[43]).

Interestingly, bonobos exhibit little or no evidence of reduced variation close to genes on either the X chromosome or the autosomes (**Suppl. Figure 11.1.e**). Prüfer et al.[30] also noted a poor correlation between bonobo diversity and regions defined in humans to be under different levels of background selection. This suggests that levels of selective constraint might be reduced in bonobos compared with most great apes, which would be consistent with the finding of a higher ratio of deleterious nonsynonymous to synonymous mutations (**Suppl. Figure 11.2**). Alternative (and perhaps less likely) explanations include a dramatic shift in rates of recombination or position of hotspots on the bonobo lineage, or complex demographic factors that effectively erased signals of positive or purifying selection in the bonobo genome.



Suppl. Figure 11.1 – *Mean estimates of π/D on the autosomes and X chromosome as a function of distance from genic regions for all primate subspecies.*



Suppl. Figure 11.2 – *Levels of X/A diversity as a function of distance from genes. The dotted line represents the expected X/A ratio under neutrality.*

Section 12: Recent demography

Estimating gene flow in chimpanzees

Whether gene flow has occurred between populations that are genetically distinct is major question for those interested in the demographic history of populations[44,45]. Gene flow can be estimated in different ways—for instance, by evaluating a model where effective population size, population divergence, and gene flow occur simultaneously and are parameterized. Models derived from a rich body of population genetic theory such as those based on the backwards-in-time coalescent (e.g., isolation-migration (IM) methods[46] or similar Approximate Bayesian Computation (ABC) frameworks[47]) and the forward-in-time diffusion approximation[48,49] present elegant and powerful approaches for obtaining actual estimates for these parameters. However, as in all explicit modeling inference, the investigator must guard against the curse of dimensionality and balance the number of parameters estimated (and thus the complexity of the model) with inferential power; this balance is particularly important to consider in the case of ABC methods where extremely complicated models can be proposed based on simulations[50] (IM and diffusion approximation models are somewhat already limited by certain analytical constraints). Both oversimplified and overcomplicated models can produce parameter estimates that lack accuracy, precision, or both and the effect is not always predictable or easily quantifiable. Even the relatively simple scenario of gene flow between two populations after divergence can be affected by factors such as variable strength of migration over time, the time migration began and ended after population divergence, and the presence of asymmetric migration in one direction.

Therefore, methods that can infer gene flow without invoking a particular model of demography are attractive alternatives, especially in scenarios where there is very little information about the underlying demography to appropriately parameterize a model with any confidence. An example of such a method that is particularly applicable to whole-genome data is the D-statistic[51], which is based on a relatively simple summary of allele sharing between three populations of interest and an outgroup. For example, when two populations are known to show a close phylogenetic relationship compared to a more distantly diverged third, the D-statistic can produce compelling evidence of unbalanced gene flow between the external population and the two internal populations (e.g., evidenced by the Neanderthal introgression into non-Africans[52]). However, while the D-statistic can provide evidence that such gene flow may have occurred, the quantification and timing of this gene flow is not possible or extremely difficult to infer without invoking explicit models (e.g., how much gene flow has there been and was the gene flow recent, old, continuous, or instantaneous). The interpretation of a D-statistic usually requires the assumption of some underlying model of divergence (e.g., the D-statistic result in Neanderthals can also be explained assuming a model of ancient population structure in Africa[52,53]).

We additionally examine gene flow using TreeMix[54], which estimates population splits and migrations between populations using a graph. The method works by first building a bifurcating tree to represent the relationships among populations. Populations that are a poor fit to the inferred tree are then identified, and gene flow events involving such populations are incorporated to improve the fit of the model to the observed data. This statistical framework allows us to estimate the presence and amount of gene flow between divergent populations with the caveat that inferred migrations is limited by the number of populations considered (in our case, we can detect only one migration).

Previous studies on chimpanzee gene flow

Our analysis of genetic variation in chimpanzees demonstrated the presence of distinct populations that correspond to the four known subspecies. The major patterns of genetic differentiation between the four subspecies can be parameterized by a model involving a series of population divergence events beginning approximately ~500,000 years ago (**Figure 2**). However, a number of papers examining autosomal loci in chimpanzees[46,55–57] have demonstrated that it is also important to consider gene flow occurring subsequent to population divergence to fully explain patterns of genetic variation amongst chimpanzees. All four previous studies invoked a coalescent-based modeling approach that assumed some topology of population splitting with subsequent gene flow. Becquet and Prezorwski examined Central, Western and Eastern chimpanzees as pairs of populations with symmetric migration and identified the strongest signal of gene flow between Western and Eastern chimpanzees. Both Hey[46] and Wegmann and Exoffier[57] examined the same populations in a single analysis, but while the former allowed asymmetric migration between all pairs of extant and ancestral populations, the latter restricted asymmetric migration to Central and Western chimpanzees (based on the results of Won and Hey) and did not allow migration between Eastern and Western chimpanzees. The more parameterized model of Hey found migration into Eastern chimpanzees from both neighboring Central chimpanzees as well the more geographically distant Western chimpanzees and also found posteriors with non-zero peaks for all other pairwise comparisons except from Eastern to Western chimpanzees, though some results were dependent on the particular priors applied. Even more interesting was the identification of statistically significant migration from Western chimpanzees into the ancestors of Central and Eastern chimpanzees. Wegmann and Excoffier, who used an ABC approach with a more limited parameterization of migration, also estimated population growth parameters within a single framework. Therefore, the estimates of migration were less nuanced in this study, though they did identify strong asymmetric migration from Western to Central chimpanzees.

It is important to note, however, that all these previous studies were based on limited amounts of sequence data and did not incorporate Nigeria-Cameroon chimpanzee population genetic data. Though the phylogeny for bonobos and Western, Central and Eastern common chimpanzees is well established, there is still uncertainty regarding their relationship to Nigeria-Cameroon chimpanzees (*P.t. ellioti*)[58]. Thus, we examined the relationship among all four chimpanzee subspecies by inferring the pattern of sequence divergence using classical phylogenetic methods as well as population divergence using a coalescent-based ABC analysis (as the former does not always reflect the latter[59]). Regional neighbor-joining trees and a maximum-likelihood tree, estimated from allele frequencies, show that Nigeria-Cameroon chimpanzees and Western chimpanzees form a clade at the sequence divergence level. This topology is also supported at the population divergence level in our ABC analysis as well as a pairwise PSMC divergence analysis (**Figure 3**). Though we note that the inclusion of a relatively simple model of symmetrical migration between all pairs of populations complicates this inference and it becomes harder to discriminate this balanced topology to the unbalanced topology previously inferred from microsatellite data[58] with Western chimpanzees as an outgroup to the other three chimpanzee subspecies. This suggests that a simple model of divergence with isolation cannot fully explain our whole-genome data and indicates the presence of complex patterns of post-divergence migration and admixture as suggested by the previous studies described above.

The strongest signal we identified of symmetric migration between adjacent populations from the ABC analyses was between the Eastern and Central chimpanzees, as previously observed in an analysis using a similarly specified model of symmetrical migration[56]. A second signal of migration is also observed in the other parapatric (**Figure 1**) comparison involving Central and Nigeria-Cameroon chimpanzees (the latter population of which was not included in Becquet and Prezworski 2007).

However, as described above, two previous analyses have identified asymmetrical gene flow patterns amongst chimpanzees[46,57] that our ABC analysis is not parameterized to infer.

Given the best-supported topology of chimpanzees inferred above, we applied the D-statistic to test whether unequal levels of gene flow have occurred between an out group subspecies and two subspecies that have more recently diverged. Consistent with the observation of Hey (2010), this analysis shows that Western chimpanzees are genetically closer to Eastern than to Central chimpanzees (D(H,W;E,C)>16SD). Yet, an even larger D-statistic was found that is suggestive of gene flow between Nigeria-Cameroon chimpanzees and Eastern chimpanzees, while TreeMix (which is only able to model the strongest gene flow event) also identified this signal ($P=2 \times 10^{-300}$) and orientated the event from Nigeria-Cameroon into Eastern chimpanzees. Finally, we also find that Eastern and Central chimpanzees are both closer to Nigeria-Cameroon than to Western chimpanzees (D(H,E;W,N)>25SD, D(H,C;W,N)>17SD).

As noted by Hey (2010)[46], direct migration from Western to Eastern chimpanzees seems geographically unlikely (today they are separated by 3,000 km), but indirect gene flow through an intermediary population such as Nigeria-Cameroon chimpanzees, as hinted by our analyses, may provide a more plausible mechanism to explain previous inferences of gene flow (though more complicated scenarios involving ancestral migration likely also contribute). Clearly the inclusion of Nigeria-Cameroon samples will be key in future studies that focus in more depth on teasing apart migration and admixture patterns amongst chimpanzees.

12.1. D-statistic

Heng Li

To formally test unbalanced gene flows between species, we performed a D-statistic test. For four haploid sequences U, V, X and Y, a site is classified as BABA if at the site U=X≠Y=V, or classified as ABBA if U=Y≠V=X. Define:

$$D'(U,V;X,Y) = \frac{\#BABA - \#ABBA}{\#BABA + \#ABBA}$$

The D-statistic equals the ratio of D' to its standard deviation estimated with block jackknife. A positive D indicates that the genetic distance between U and X is larger than between V and Y, while a negative D indicates that the distance between U and Y is larger than V and X. In particular, if U is a known outgroup of other samples, D(U,V;X,Y)>0 if V is genetically closer to Y, while D(U,V;X,Y)<0 if V is genetically closer to X. The D-statistic provides a formal and model-free test for testing unbalanced gene flows between the (U,V) clade and the (X,Y) clade.

Suppl. Table 12.1.1 shows the D-statistics between different subspecies. Let d(X,Y) be the genetic distance between two populations X and Y. The table suggests that the following inequalities: d(pte-Koto,pts-Nakuu)<d(pte-Koto,ptt-Vaillent)<d(ptv-Clint,ptt-Vaillent) and d(pte-Koto,pts-Nakuu)<d(ptv-Clint,pts-Nakuu)<d(ptv-Clint,ptt-Vaillent). Combining the two inequalities we know that d(pte-Koto,pts-Nakuu) is the closest pair between the two chimpanzee clades, which is also identified by TreeMix. Interestingly, the smallest D-statistic among common chimpanzees is achieved for D(ptv,pte;pts,ptt). A parsimonious scenario that is consistent with an insignificant D(ptv,pte;pts,ptt) but significant other D values would be that a branch of ancestral population of Western and Nigeria-Cameroon later admixed into Eastern, while a branch of ancestral population of Eastern and Central admixed into Nigeria-Cameroon chimpanzee. Because these two gene flows do not affect the balance of the (ptv,pte) and (pts,ptt) clades, they will result in an insignificant D(ptv,pte;pts,ptt).

Sample U	Sample V	Sample X	Sample Y	$D'(U,V;X,Y)$	D
Human	pp-Dzeeta	ptv-Clint	pte-Koto	0.53%	1.24
		ptv-Clint	ptt-Doris	0.83%	2.11
		ptv-Clint	pts-Kidongo	-0.57%	-1.43
		pte-Koto	ptt-Doris	0.43%	1.14
		pte-Koto	pts-Kidongo	-0.85%	-2.19
		ptt-Doris	pts-Nakuu	-1.05%	-3.00
Human	ptv-Clint	pts-Nakuu	ptt-Vaillent	-5.33%	-16.35
	ptv-Koto	pts-Nakuu	ptt-Vaillent	-6.61%	-20.15
	pts-Nakuu	ptv-Clint	pte-Koto	9.06%	25.72
	ptt-Vaillent	ptv-Clint	pte-Koto	7.34%	21.83
ptv-Clint	pte-Koto	pts-Nakuu	ptt-Vaillent	-2.54%	-6.66
ptv-Clint	pte-Julie	pts-Nakuu	ptt-Vaillent	0.50%	1.30
ptv-Clint	pte-Tobi	pts-Nakuu	ptt-Vaillent	0.83%	2.08
ptv-Clint	pte-Banyo	pts-Nakuu	ptt-Vaillent	-0.92%	-2.27
Human	gbg-Mkubwa	ggg-Delphi	ggd-Nyango	2.82%	8.23

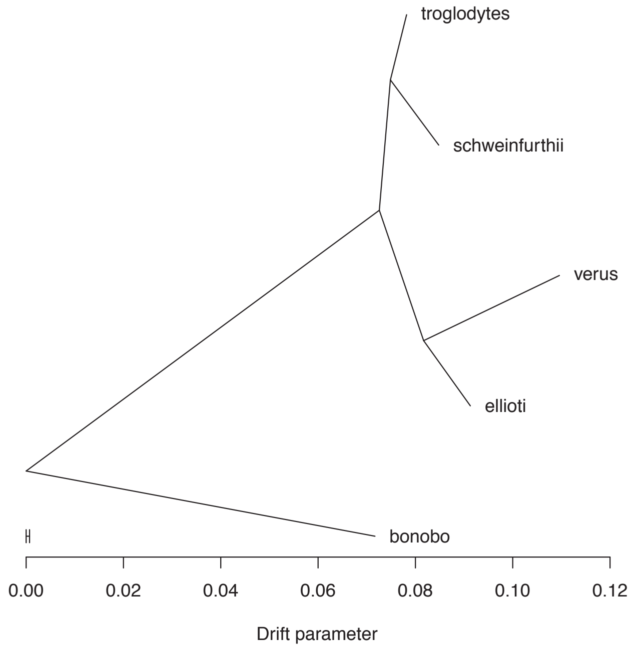
Suppl. Table 12.1.1 – *D-statistic for four haploid sequences, A, B, X and O, a site is classified as ABBA if at the site base A = O B = X, or classified as BABA if B = O= A = X. Define D'(A, B, X; O)== (#ABBA −#BABA)/(#ABBA +#BABA). D(A, B, X; O) equals the ratio of the mean of D' to its standard deviation, estimated by block Jack-knife. A positive D value indicates that sample B is genetically closer to X, while a negative vale indicates A closer to X.*

12.2 TreeMix

Joanna L. Kelley, Jeffrey M. Kidd

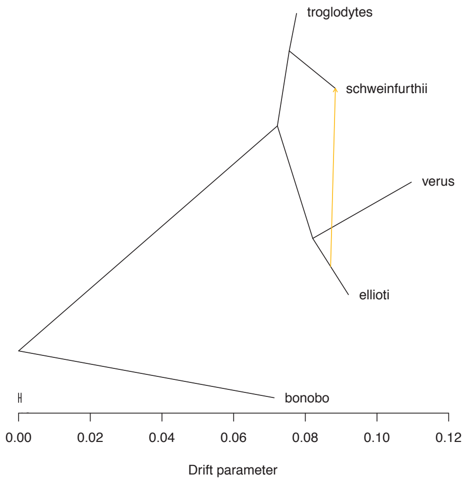
To provide a more complete picture among chimpanzee demographic history, we applied a method to infer population splits and gene flow events, TreeMix[54]. We based our analysis on a set of ~5 million polymorphic sites that were randomly selected from the total set of sites using PLINK[18] (--thin 0.31). We used the SNP calls derived from mapping to the human genome reference (NCBI Build 36). We only considered autosomal SNPs, and removed variants on random or unassigned chromosomes.

We first assessed the tree topology using TreeMix with the randomly thinned data we infer a ML tree and residual fit (**Suppl. Figure 12.2.1**) with all 38 individuals. The tree model explains 99.92% of the variance in the data.



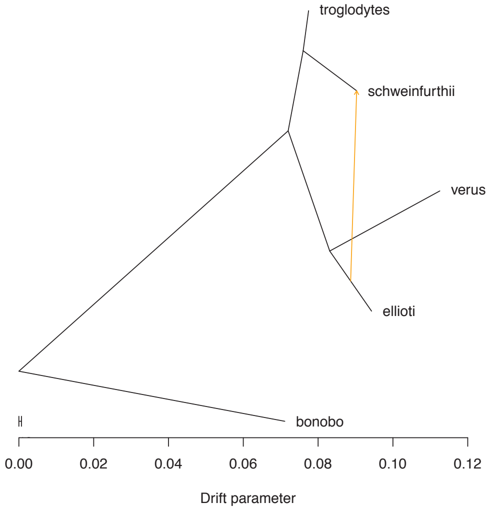
Suppl. Figure 12.2.1 – *ML tree inferred with TreeMix. Five million randomly selected SNPs without regard to MAF or LD. LD grouping with 5,000 SNPs per bin.*

Using the ML tree as the user specified tree, we use TreeMix to infer one migration event (**Suppl. Figure 12.2.2**). The model with one migration event explains 99.98% of the variation. The migration event is inferred from the *P.t.ellioti* branch to *P.t.schweinfurthii*. The TreeMix migration event weight for the *P.t.ellioti* ancestry into the *P.t.schweinfurthii* population is 9.2% ± 0.2% standard errors ($P<2.22x10^{-308}$), using a block jackknife to obtain standard errors and P-values (with 5,000 sites in each block).



Suppl. Figure 12.2.2 – *ML tree with one migration event. It shows a significant migration* *between ellioti and schweinfurthii populations.*

The results are robust to removing individuals with potential admixture (**Suppl. Figure 12.2.3**). 34 individuals remain after removing those with evidence of admixture. The tree without migration explains 99.93% of the variance. With one migration event, 99.98% of the variance is explained. The migration estimate is 13.6% ± 0.6% (P<2.22x10-308), using a block jackknife to obtain standard errors and P-values (with 5,000 sites in each block). This provides further evidence in addition to the D-statistic to suggest gene flow between Nigeria-Cameroon and Eastern chimpanzees.



Suppl. Figure 12.2.3 – *ML tree with one migration event for non-admixed individuals. The gene flow between ellioti and schweinfurthii is still maintained.*

12.3 Approximate Bayesian Computation analysis of chimpanzee demography

Krishna R Veeramah, August Woerner, Michael F Hammer

Methods

In order to characterize *Pan troglodytes* demographic history at the whole-genome level we developed an Approximate Bayesian Computation (ABC) approach based on coalescent simulations[60] amenable to handling large scale data (tens of millions of bases for multiple individuals). The aim of this analysis was primarily to identify the appropriate topology describing *p.t.* subspecies divergence but the general framework also allowed us to estimate parameters such effective population sizes, population divergence times and migration rates.

Sequence data

Performing ABC in the context of the data assembled in this study is complicated by the following factors; mapping of chimpanzees to a human reference (and thus no usable recombination map), a lack of information on haplotype phase and potential intra and inter species contamination (and the subsequent application of an allele balance filter resulting in a ~10% loss in heterozygosity). Therefore we chose to perform the analysis on a subset of ~37,000 1 Kbp loci previously identified for inference of population genetic parameters in humans (and that show good synteny with chimpanzees)[61] and that had at least 90% of bases called in all individuals with no or little evidence of contamination (after applying filters previously described as well as the masks described in **Section 2**) This resulted in sequence data from 3 *p.t.t*, 3 *p.t.s*, 7 *p.t.e* and 2 *p.t.v* individuals at 10,008 loci, with no allele balance filter applied. Ancestral alleles at segregating sites were identified by orientating against the bonobo and human reference genomes. False heterozygotes caused by high error rates in next generation sequencing could present a problem in demographic inference with regard to singletons[62]. However, as the average coverage for these chimpanzee genomes was high and the total sequence data considered relatively low (~10MB per individual), we do not consider this to be a major issue.

Mutation rate estimates

Mutation rates for loci were estimated using previously aligned human data[61] as previously described[62], with a Jukes Cantor correction applied to sequence divergence estimates and the assumption of a human chimpanzee split of 6 million years, 25 years per generation in humans, and an ancestral population size of 83,000 (estimated by CoalHMM analysis in **Section 13**). The mean mutation rate across loci was 1.08×10^{-8} per base per generation (stdev = 5.6×10^{-9}), in line with recent estimates of the human mutation rate[63,64]. While the absolute estimates of mutation rates at these loci may not be directly applicable to chimpanzees, the relative rates are likely to be captured amongst loci, at which point the inferential problem becomes one of scaling parameter estimates appropriately.

Summary statistics

We computed the following summary statistics to describe the data for every pair of populations: number of shared polymorphisms, number of private polymorphisms in each populations and the number of private fixed sites in each population. These statistics are known to contain substantial information about population demography[65] and are utilized in the program MIMAR which has previously been used to estimate Chimpanzee demography under an isolation-migration model[56]. These statistics are particularly useful as they are ambiguous to the requirement of haplotype inference. We use the mean and variance of these summary statistics across all loci to describe the data (unlike MIMAR where these summaries are used to calculate a likelihood for each locus individually, which is computationally intensive for the amount of data considered here). Other summary statistics that might traditionally be considered useful for demographic inference such as Tajima's D were not utilized due to the low sample size for some subspecies considered in the analysis. Therefore our method is unlikely to capture the inference of parameters such as exponential growth rates for population size and thus are not considered here (though they have been estimated previously[57]).

Demographic models

We consider two main subspecies topologies when constructing models of *p.t* demography (**Suppl. Figure 12.3.1**). The first is an unbalanced topology with *p.t.v* representing the earliest diverged subspecies that has been described previously[58] and is consistent with the PCA, Frappe (**Figure 1 and Section 8**) and diploid PSMC analysis (**Figure 3**). The second is a balanced topology, with *p.t.v* and *p.t.e* forming a distinct clade from *p.t.t* and *p.t.s*, and matches the neighbor-joining tree and sequence divergence patterns and was also inferred by the haploid PSMC analysis. However the latter analysis is problematic when coalescence times overlap speciation time, as would almost certainly be the case for the time frame of *p.t* subspecies divergence, with populations with the largest N_e being underestimated the most.

Parameters (and associated priors) describing both topologies are indicated in **Suppl. Figure 12.3.1, 12.3.4** and **Suppl. Table 12.3.1**. Prior distributions are motivated by Wegmann and Excoffier[57] and all are uniform distributed on a log10 (x) scale. The following classes of parameters are considered: effective population size (N), time of population divergence (t) and, for later iterations of the analysis, number of migrants per generation (M=Nm, with m equal to the migration rate). As priors for absolute times of population divergence would likely overlap, in order to obtain flat priors we considered the time of an internal branching event with respect to the more recent branching event above in the topology, a divergence scheme used in the likelihood-based method MCMCcoal[66].

ABC analysis

ABC analysis was performed using two different regression adjustments depending on their application. When estimating model parameters we utilized ABCtoolbox[67], which implements a general linear model (GLM) adjustment[68] on retained simulations. To maximize sufficiency but limit dimensionality, the full set of summary statistics was transformed into partial least squares (PLS) components[47] and we used the change in Root Mean Square Error (RMSE) to guide the choice of number components. These PLS components were then used to estimate parameters. When performing model choice ABCtoolbox can be used to find the marginal density of each model in order to calculate a Bayes Factor. However there are concerns about biases resulting from a lack of summary statistic sufficiency when applying Bayes Factors in ABC[69]. Therefore we used the logistic regression (LR) method previously described[70] to perform model choice using an adapted version of the R function calmod.r as well a more naïve method (the direct method, DM) of the proportion of retained simulations from each model[71]. We used the Kruskal-Wallis-based ranking method described in Veeramah et al[72] to identify the set of and number summary statistics most relevant to model choice. Simulations were performed using a version of ms[73] adapted for Python to allow fast, parallel processing. Individual locus mutation rates were incorporated into theta and any sites missing in the real data (including non-segregating sites) were also masked in simulated data. 1% of simulations were retained for the GLM (parameter estimation) and LR (model choice) adjustments. Principal component analysis (PCA) was used for comparing the multidimensional distribution of summary statistics using the ‘‘prcomp’’ function in R.

Results

Models without migration

We examined the relative probabilities of two likely branching models involving *p.t.* Model 1 results in *p.t.v* branching off earliest from the *p.t.* lineage followed by *p.t.e* (unbalanced model, **Suppl. Figure 12.3.1A**), while Model 2 involves the ancestors of *p.t.v* and *p.t.e* branching off together (balanced model, **Suppl. Figure 12.3.1B**). Initially, in order to reduce the number of parameters examined, no migration was considered in these models. Unlike Model 1 (which is restricted by the divergence order), the divergence time T2 in Model 2 was chosen without regard to T1 and was free to be larger or smaller.

PCA visually demonstrated a good multidimensional fit between the observed summary statistic data and simulated data generated under both models (**Suppl. Figure 12.3.2**). We then identified the set of summary statistics that best distinguished the two models based on simulated data, and were able to identify the correct model in 97% of cases using LR and 82% using DM[71]. Using this tuned set of summary statistics we obtained a LR and DM posterior probability (*P*) for Model 2 of 96% and 64% respectively. Based on these two estimates of the

posterior probability there was a 100% (LR) and 80% (DM) probability respectively of Model 2 being the true model using the method of Fagundes et al.[16]
(i.e., Pr(P_{Md2}=0.96|Md2)/[Pr(P_{Md2}=0.96|Md2)+ Pr(P_{Md2}=0.96|Md1)]).

Summary statistics were then transformed into PLS components in order to infer parameters from Model 2. 10 PLS components were used to infer parameters based on a total of 300K simulations. A p-value for the fit of the GLM based on the fraction of retained simulations with a smaller or equal likelihood to the observed data was 0.558, indicating a good fit of the local adjustment to the observed data. In addition the 95% CIs appeared relatively reliable from simulated data, with 94-97% of known true parameter values falling within them from 1000 simulated pseudo-observed sets. Posterior probabilities are shown in **Suppl. Table 12.3.2** and posterior distributions visualized in **Suppl. Figure 12.3.3**. GLM-fitted posterior distributions generally showed good peakedness and were congruent with the raw retained distributions.

The estimates of N_e in present day populations, at least with regard to the order of magnitude, were compatible with other estimates described in this and other studies, with *p.t.t* highest and *p.t.v* lowest. The estimates of ancestral N_es (N_{T1}, N_{T2} and N_{anc}) were also compatible with previous work and the PSMC analysis (**Figure 3**), which suggests a decrease of N_e backwards in time along the p.t. lineage, with a relatively small N_e before all subspecies started diverging. The divergence time estimates suggested an older split for the *p.t.e*/*p.t.v* clade than the *p.t.s*/*p.t.t* clade by about 80K years (albeit with large CIs), which would be compatible with the current geographic distribution and the large genetic distance we observe between *p.t.e* and *p.t.v* via PCA analysis as well as being consistent with **Figure 8.5.1 and Figure 2**. The timing of the *p.t.s*/*p.t.t* split (3-36K and 1-30K generations with and without singletons respectively) was also in line with that of Wegmann and Excoffier [2] (8-25K generations) (we obtained wider CIs, in part as we allow a larger prior while Wegmann and Excoffier hit the limit of their prior within their 95% CI) while our additive median estimate of N_{T1}, N_{T2} and N_{anc} of 27K generations (30K without singletons) was also compatible with Wegmann and Excoffier's [2] estimate of the split of *p.t.v* from *p.t.t*/*p.t.s* (16-47K generations, 300K-940K years).

Models with symmetrical migration

Inference of migration parameters between Chimpanzee subspecies has previously been performed by Wegmann and Excoffier[57] and Hey[46]. Though our low sample sizes limit our power to infer these parameters with great confidence we updated our divergence models to included symmetric migration between all pairs of present day and ancestral populations (an additional 9 migration parameters to the original models). By adding migration between ancestral populations it was necessary to further parameterize the balanced model by assuming one branching event occurred before another. Therefore Model 2 with migration was split into two balanced models with migration, Model 4A, where *p.t.t* and *p.t.s* diverged most recently and Model 4B, where *p.t.e* and *p.t.v* diverged most recently. Model 1 with symmetric migration was named Model 3 (**Suppl. Figure 12.3.4**).

PCA of the two original models against the three migration models showed a substantially better fit of the latter to the observed data (**Suppl. Figure 12.3.4**). This was confirmed in a model choice analysis using the DM, which consistently showed a much higher posterior probability for each migration model against its non-migration model counterpart (P>0.88-0.97) (**Suppl. Table 12.3.3**) (Model 2 was re-simulated as two models, 2A and 2B, to match the branching order restrictions of Models 4A and 4B respectively). The average power to distinguish migration from non-migration models using the DM was 86% (min 75%, max =100%) based on simulated data. Therefore migration appears to be an important factor to consider when inferring Chimpanzee demography. Though generally in good agreement with the DM estimates, the LR method on one occasion gave unusual results, which, having manual examined the data, is likely due to poor fit of the logistic regression model because of the lack of representation of the non-migration model in the retained dataset, and thus we discarded results using LR in this instance.

However, simulations demonstrated that there was almost no power to distinguish between the three migration models. When comparing Models 4A and 4B there was at least some (but not particularly strong) evidence that Model 4A (i.e., a more recent *p.t.t*/*p.t.s* split) is the most likely with a posterior probability of 57% and probability that this is the correct model of 63% given the posterior probability. This is consistent with the estimates of T1 being slightly more recent than T2 in Model 2 and the general patterns of diversity in the whole-genome data and previous work. In addition, when we attempted parameter inference for Model 4B, T1 was much older than for Models 3 and 4A while the posterior distribution for T2 for this model was extremely flat and non-informative unlike the other two models, indicating the method may be finding it difficult to identify a good divergence time for the second divergence event because of an incorrect branching order.

Models 3 and 4A were almost completely indistinguishable, with the probability of choosing the correct model being 50% (i.e., no better than random chance). This is perhaps not surprising as the likely recent population divergence events coupled with subsequent migration makes this a ‘hard’ phylogenetic problem that may need more samples and/or larger or different loci to resolve (for example STRs to distinguish recent and ancient migrations). Therefore we consider the topology of Chimpanzee divergence with regard to *p.t.e* somewhat uncertain with both unbalanced and balanced models with migration fitting the data equally well. As a consequence we report parameter estimates for both models (**Suppl. Tables 12.3.4** and **12.3.5** and **Suppl. Figure 12.3.6** and **12.3.7**).

Despite not being able to distinguish the correct divergence topology there is still substantial information about the demographic processes connecting the *p.t* subspecies when conducting parameter inference with migration regardless of the model chosen. P-values examining the fit of the GLM to the observed data were again good (>0.69) while 95% CIs were generally reliable (though somewhat more noisy than for the non-migration model, not surprising given the increase in the number of parameters). N_e estimates were still in line with the non-migration models but median divergence times became older and the CIs more diffuse, as would be expected from adding migration after population divergence events. However, more interestingly, despite the low sample size historical migration was detected with relatively peaked posteriors regardless of which model of divergence topology was examined. Substantial migration was observed between *p.t.t* and *p.t.s,* as would be expected by their overlapping geographic ranges but migration was also detected between *p.t.e* and *p.t.t* (i.e., parapatric populations). There is also some indication of migration between the *p.t.t*/*p.t.s* ancestors and *p.t.v*, which is consistent with the finding of Hey[46].

In order to not over-parameterize the model we did not consider asymmetrical migration rates between pairs of populations as in Hey[46] or population growth like in Wegmann and Excoffier[57]. However, this will be an interesting question to tackle in the future in a similar ABC framework using larger numbers of whole genomes and longer loci (which will require mapping to the chimpanzee reference genome and a recombination map).

Parameter	Min	Max	Min^a	Max^a	Distribution	Md1	Md2	Md3	Md4A	Md4B
log(N1)	3.0	5.4	1,000	251,189	Uniform	X	X	X	X	X
log(N2)	3.0	5.4	1,000	251,189	Uniform	X	X	X	X	X
log(N3)	3.0	5.4	1,000	251,189	Uniform	X	X	X	X	X
log(N4)	3.0	5.4	1,000	251,189	Uniform	X	X	X	X	X
log(N_{T1})	3.0	5.4	1,000	251,189	Uniform	X	X	X	X	X
log(N_{T2})	3.0	5.4	1,000	251,189	Uniform	X	X	X	X	X
log(Nanc)	3.0	5.4	1,000	251,189	Uniform	X	X	X	X	X
log(T1)	3.0	5.0	20,000	2,000,000	Uniform	X	X	X	X	X
log(T2)	3.0	5.0	20,000	2,000,000	Uniform	X	X	X	X	X
log(T3)	3.0	5.0	20,000	2,000,000	Uniform	X	X	X	X	X
log($M_{1\text{-}2}$)	-2.0	1.0	0.01	10	Uniform			X	X	X
log($M_{1\text{-}3}$)	-2.0	1.0	0.01	10	Uniform			X	X	X
log($M_{1\text{-}4}$)	-2.0	1.0	0.01	10	Uniform			X	X	X
log($M_{2\text{-}3}$)	-2.0	1.0	0.01	10	Uniform			X	X	X
log($M_{2\text{-}4}$)	-2.0	1.0	0.01	10	Uniform			X	X	X
log($M_{3\text{-}4}$)	-2.0	1.0	0.01	10	Uniform			X	X	X
log($M_{1,2\text{-}3}$)	-2.0	1.0	0.01	10	Uniform			X	X	
log($M_{1,2\text{-}4}$)	-2.0	1.0	0.01	10	Uniform			X	X	
log($M_{1\text{-}3,4}$)	-2.0	1.0	0.01	10	Uniform					X
log($M_{2\text{-}3,4}$)	-2.0	1.0	0.01	10	Uniform					X
log($M_{1\text{-}2\text{-}3,4}$)	-2.0	1.0	0.01	10	Uniform			X		
log($M_{1\text{-}2,3\text{-}4}$)	-2.0	1.0	0.01	10	Uniform				X	X

Note. ^{a}Values converted from log10 scale to real world estimates, with divergence time assuming a 20 year generation time. N1=*p.t.s*, N2=*p.t.t*, N3=*p.t.e*, N4=*p.t.v*

Suppl. Table 12.3.1 - *Priors for the various ABC models (X marks where prior is relevant).*

Parameter	HDPI 95% fit[a]	Posterior Estimation[b]				Real World Estimates[c]		
		Mode	Median	HDPI 95		Median	95% CI	
				Lower	Upper		Lower	Upper
log(N1)	0.97	4.62	4.54	3.73	5.26	34,532	5,315	181,238
log(N2)	0.97	5.34	5.15	4.62	5.40	140,443	41,611	251,177
log(N3)	0.96	4.53	4.48	3.73	5.18	30,319	5,315	150,998
log(N4)	0.96	4.21	4.12	3.18	4.93	13,212	1,523	84,475
log(N_{T1})	0.95	5.13	5.02	4.43	5.40	104,342	27,029	251,189
log(N_{T2})	0.95	4.68	4.67	4.02	5.37	46,739	10,496	235,055
log(N_{anc})	0.94	4.42	4.42	4.37	4.47	26,459	23,668	29,693
log(T1)	0.95	4.01	4.01	3.42	4.56	205,334	52,899	725,421
log(T2)	0.95	4.19	4.15	3.60	4.62	281,203	79,365	825,353
log(T3)	0.96	3.38	3.51	3.00	4.15	65,232	20,000	284,564

Note. [a]A metric demonstrating how often known simulated values (n=1000) fell within the calculated 95% CI, which gives a guide to the reliability of these CI's for real data. [b]Calculated using 10PLS components, 300K simulations and retaining 1%. [c]Values converted from log10 scale to real world estimates, with divergence time assuming a 20-year generation time.

Suppl. Table 12.3.2 - *Posterior estimates for Model 2.*

Best Fit Model (BFM)	Alternative Model	Probability of BFM (*Pr*)	Power from 1000 sims	Pr(BFM true | *Pr*)
Md 3	Md 1	96%	98%	98%
Md 4A	Md 2A	97%	88%	99%
Md 4B	Md 2B	87%	82%	82%
Md 4A	Md 4B	57%	62%	63%
Md 4B	Md 3	54%	54%	57%

Suppl. Table 12.3.3 - *Posterior probabilities comparing various combinations of model.*

Parameter	HDPI 95% fit[a]	Posterior Estimation[b]				Real World Estimates[c]		
		Mode	Median	HDPI 95		Median	95% CI	
				Lower	Upper		Lower	Upper
log(N1)	0.99	3.70	3.96	3.00	4.97	9,220	1,000	93,319
log(N2)	0.93	5.07	4.88	4.01	5.40	76,268	10,210	251,189
log(N3)	0.99	3.78	3.82	3.05	4.60	6,584	1,123	39,372
log(N4)	1.00	3.64	3.65	3.14	4.15	4,423	1,394	14,150
log(N_{T1})	0.91	4.91	4.44	3.30	5.40	27,234	1,975	251,189
log(N_{T2})	0.94	4.17	4.27	3.17	5.35	18,490	1,481	224,880
log(N_{anc})	0.96	4.09	4.02	3.09	4.85	10,543	1,234	71,162
log(T1)	0.95	4.62	4.35	3.31	5.00	450,837	41,243	2,000,000
log(T2)	0.92	4.67	4.25	3.23	5.00	352,128	34,298	2,000,000
log(T3)	0.93	4.61	4.14	3.19	5.00	278,535	30,990	2,000,000
log(M_{1-2})	0.98	0.08	-0.07	-1.41	1.00	0.85	0.04	10.00
log(M_{1-3})	0.91	-1.42	-1.22	-2.00	-0.13	0.06	0.01	0.74
log(M_{1-4})	0.98	-1.72	-1.50	-2.00	-0.68	0.03	0.01	0.21
log(M_{2-3})	0.96	-0.72	-0.77	-1.98	0.38	0.17	0.01	2.37
log(M_{2-4})	0.92	-1.54	-1.35	-2.00	-0.50	0.04	0.01	0.32
log(M_{3-4})	0.94	-1.04	-1.05	-1.71	-0.40	0.09	0.02	0.40
log($M_{1,2-3}$)	0.91	-0.75	-0.63	-1.96	0.77	0.23	0.01	5.83
log($M_{1,2-4}$)	0.94	-1.41	-0.93	-2.00	0.61	0.12	0.01	4.07
log($M_{1-2-3,4}$)	0.90	-1.17	-0.51	-1.90	0.86	0.31	0.01	7.18

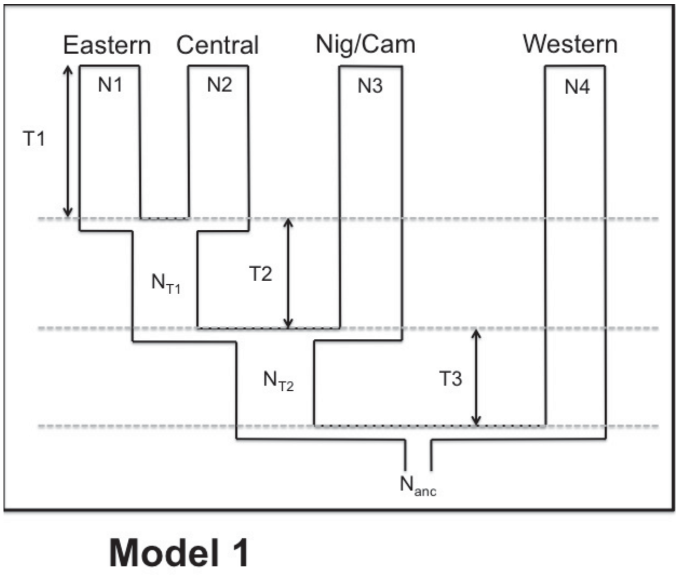
Suppl. Table 12.3.4 - *Posterior estimates for Model 3.*

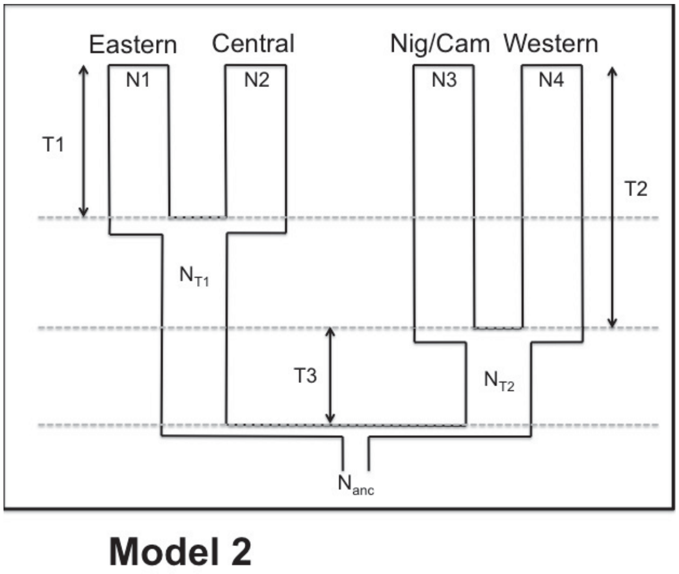
Note. [a]A metric demonstrating how often known simulated values (n=1000) fell within the calculated 95% CI, which gives a guide to the reliability of these CI's for real data. [b]Calculated using 10PLS components, 1M simulations and retaining 1%. [c]Values converted from log10 scale to real world estimates, with divergence time assuming a 20-year generation time.

Parameter	HDPI 95% fit[a]	Posterior Estimation[b]				Real World Estimates[c]		
		Mode	Median	HDPI 95		Median	95% CI	
				Lower	Upper		Lower	Upper
log($N1$)	0.94	4.01	4.02	3.03	5.02	10,528	1,069	104,814
log($N2$)	0.98	5.06	4.86	4.00	5.40	72,001	9,986	251,189
log($N3$)	0.98	3.67	3.78	3.01	4.59	6,033	1,022	38,511
log($N4$)	0.99	3.76	3.76	3.11	4.39	5,710	1,283	24,739
log(N_{T1})	0.93	4.71	4.48	3.31	5.40	30,241	2,041	251,189
log(N_{T2})	0.91	4.60	4.21	3.08	5.28	16,275	1,200	188,395
log(N_{anc})	0.96	4.02	3.97	3.12	4.72	9,240	1,319	52,786
log($T1$)	0.98	4.52	4.25	3.26	5.00	358,517	36,416	2,000,000
log($T2$)	0.95	4.60	4.19	3.21	5.00	309,557	32,452	2,000,000
log($T3$)	0.87	4.64	4.18	3.20	5.00	301,655	31,567	2,000,000
log($M_{1\text{-}2}$)	0.97	0.11	-0.11	-1.47	1.00	0.77	0.03	10.00
log($M_{1\text{-}3}$)	0.96	-1.50	-1.12	-2.00	0.03	0.08	0.01	1.06
log($M_{1\text{-}4}$)	0.93	-1.71	-1.49	-2.00	-0.68	0.03	0.01	0.21
log($M_{2\text{-}3}$)	0.97	-0.65	-0.72	-1.98	0.44	0.19	0.01	2.76
log($M_{2\text{-}4}$)	0.94	-1.57	-1.35	-2.00	-0.44	0.05	0.01	0.36
log($M_{3\text{-}4}$)	0.97	-1.29	-1.26	-1.95	-0.58	0.06	0.01	0.26
log($M_{1,2\text{-}3}$)	0.90	-1.31	-0.61	-1.96	0.77	0.25	0.01	5.95
log($M_{1,2\text{-}4}$)	0.92	-1.41	-0.92	-2.00	0.57	0.12	0.01	3.72
log($M_{1\text{-}2,3\text{-}4}$)	0.90	-1.22	-0.54	-1.88	0.89	0.29	0.01	7.74

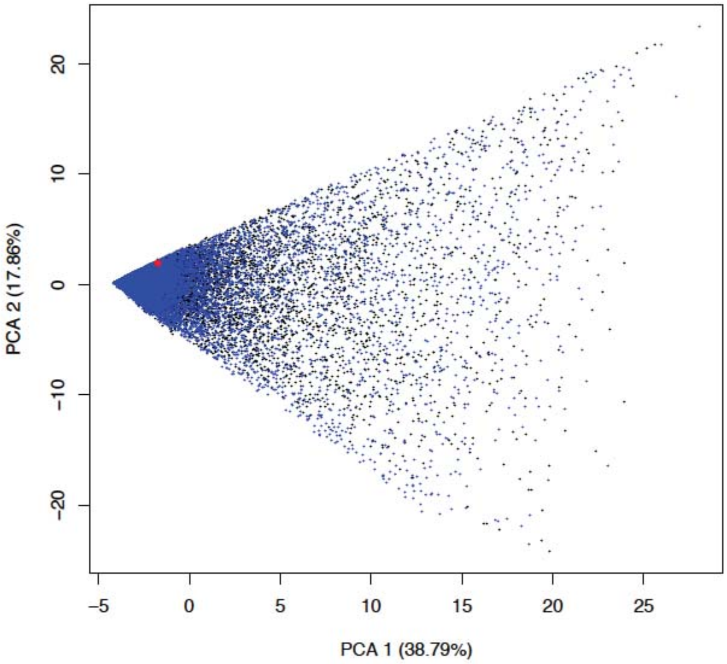
Suppl. Table 12.3.5 - *Posterior estimates for Model 4A.*

Note. [a]A metric demonstrating how often known simulated values (n=1000) fell within the calculated 95% CI, which gives a guide to the reliability of these CI's for real data. [b]Calculated using 10PLS components, 1M simulations and retaining 1%. [c]Values converted from log10 scale to real world estimates, with divergence time assuming a 20-year generation time.

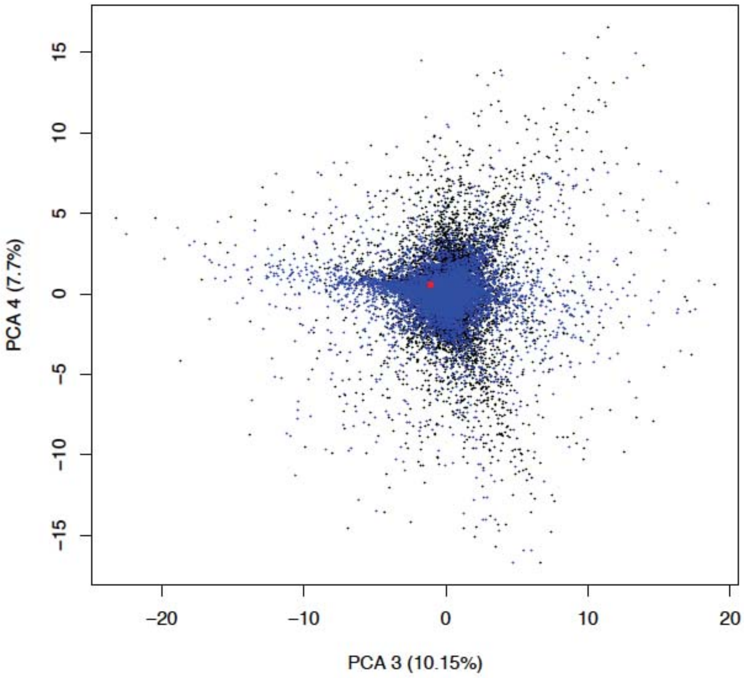




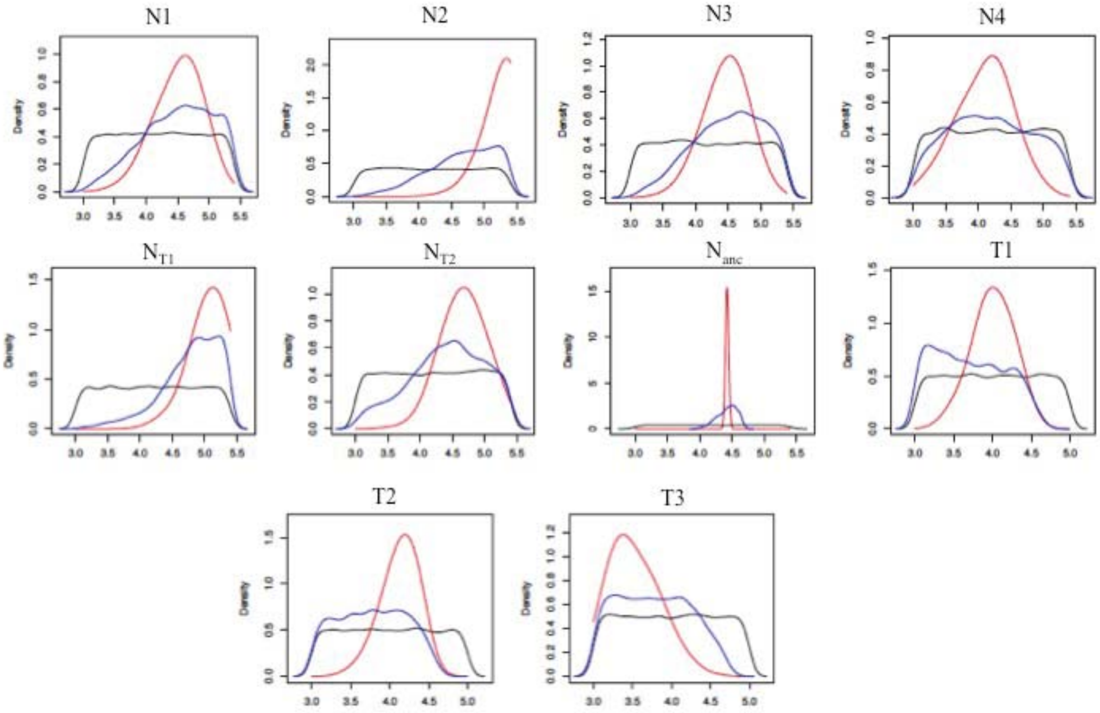
Suppl. Figure 12.3.1 - *Demographic Models 1 and 2 for P.t divergence without migration.*



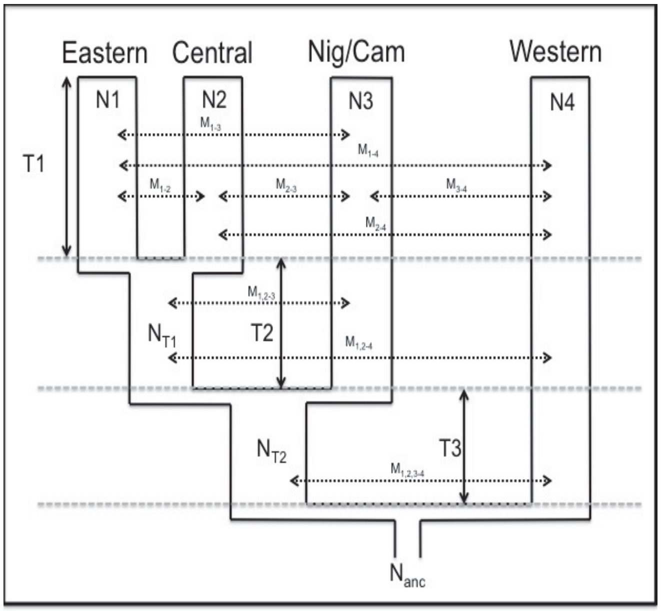
Suppl. Figure 12.3.2A - *PCA 1 and 2 of summary statistics for Model 1 (black), Model 2 (blue) and observed data (red).*



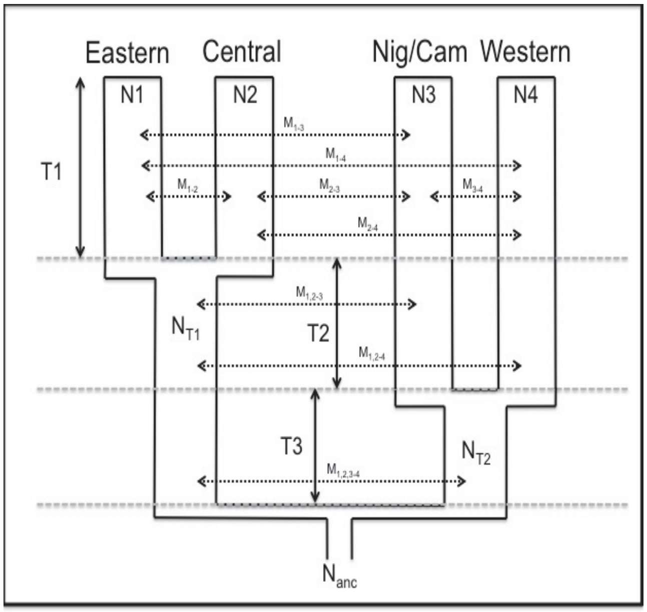
Suppl. Figure 12.3.2B - *PCA 3 and 4 of summary statistics for Model 1 (black), Model 2 (blue) and observed data (red).*



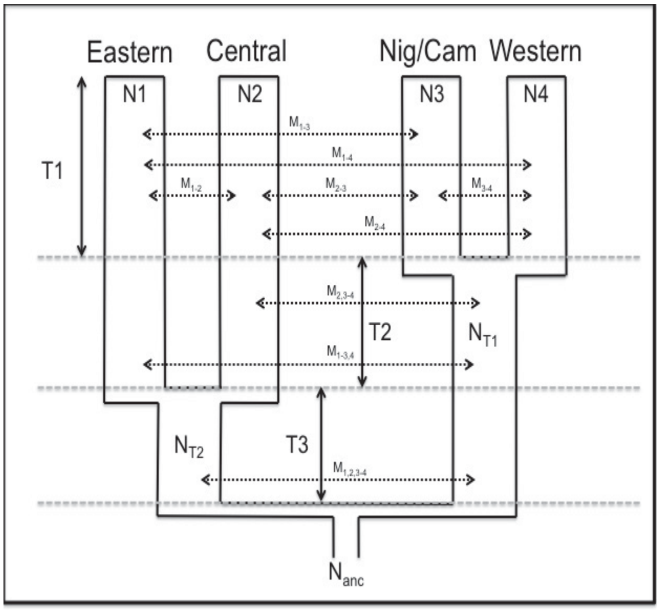
Suppl. Figure 12.3.3 - *Prior (black), retained (blue) and GLM adjusted posterior distributions under Model 2.*



Model 3

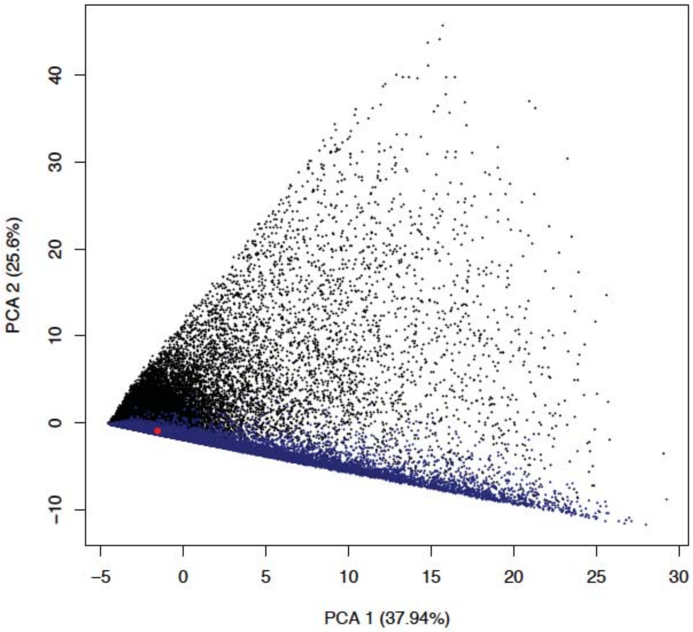


Model 4A

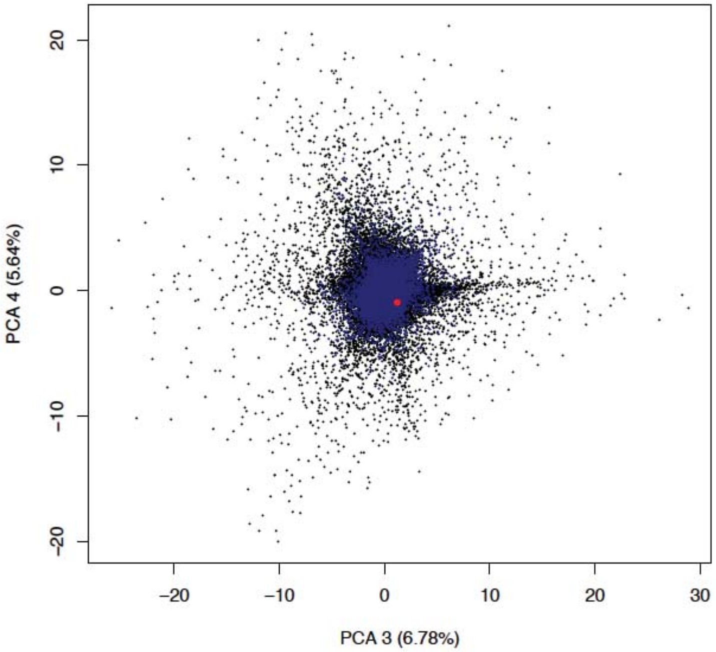


Model 4B

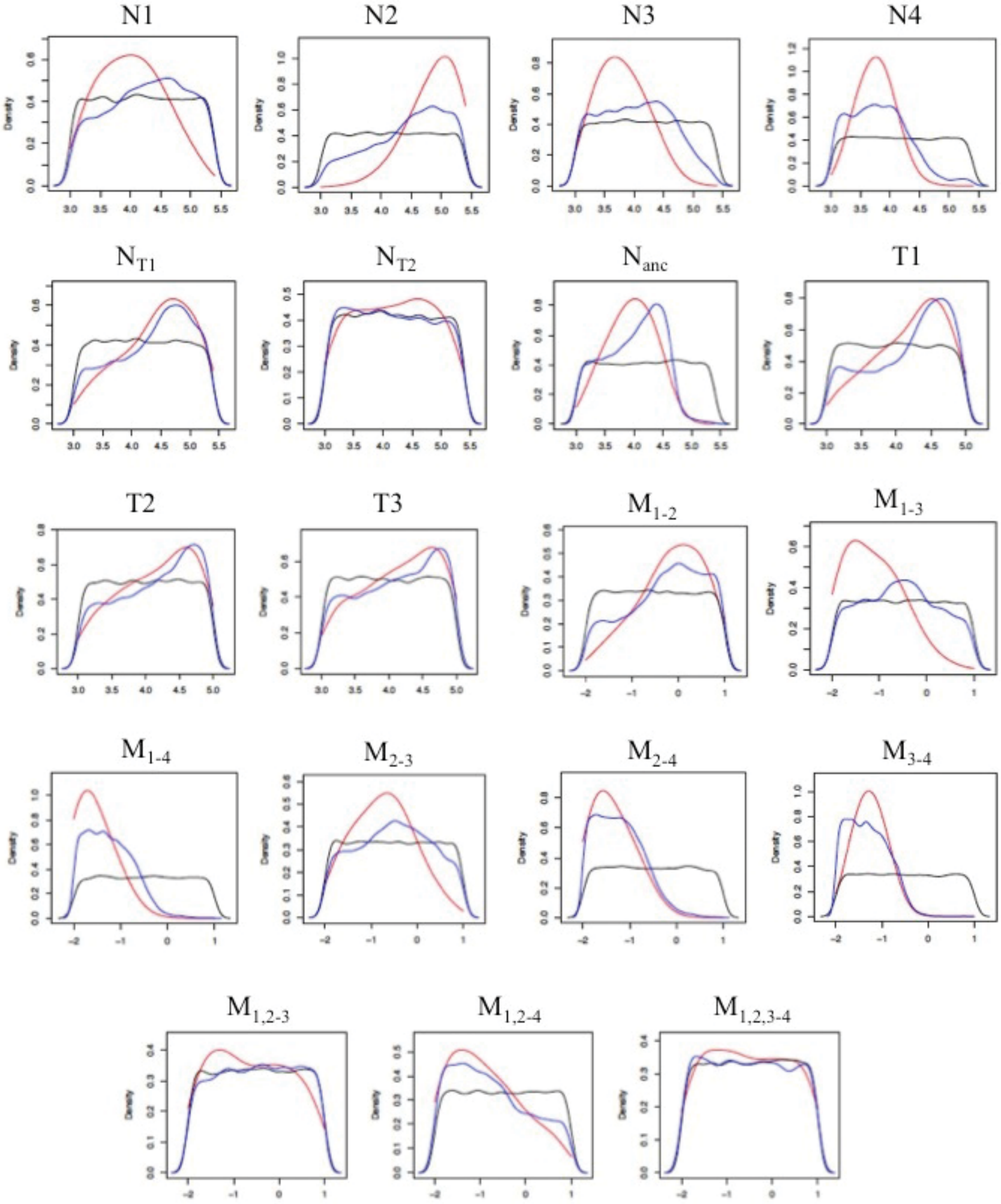
Suppl. Figure 12.3.4 - *Demographic Models 3, 4A and 4B for P.t divergence with symmetric migration.*



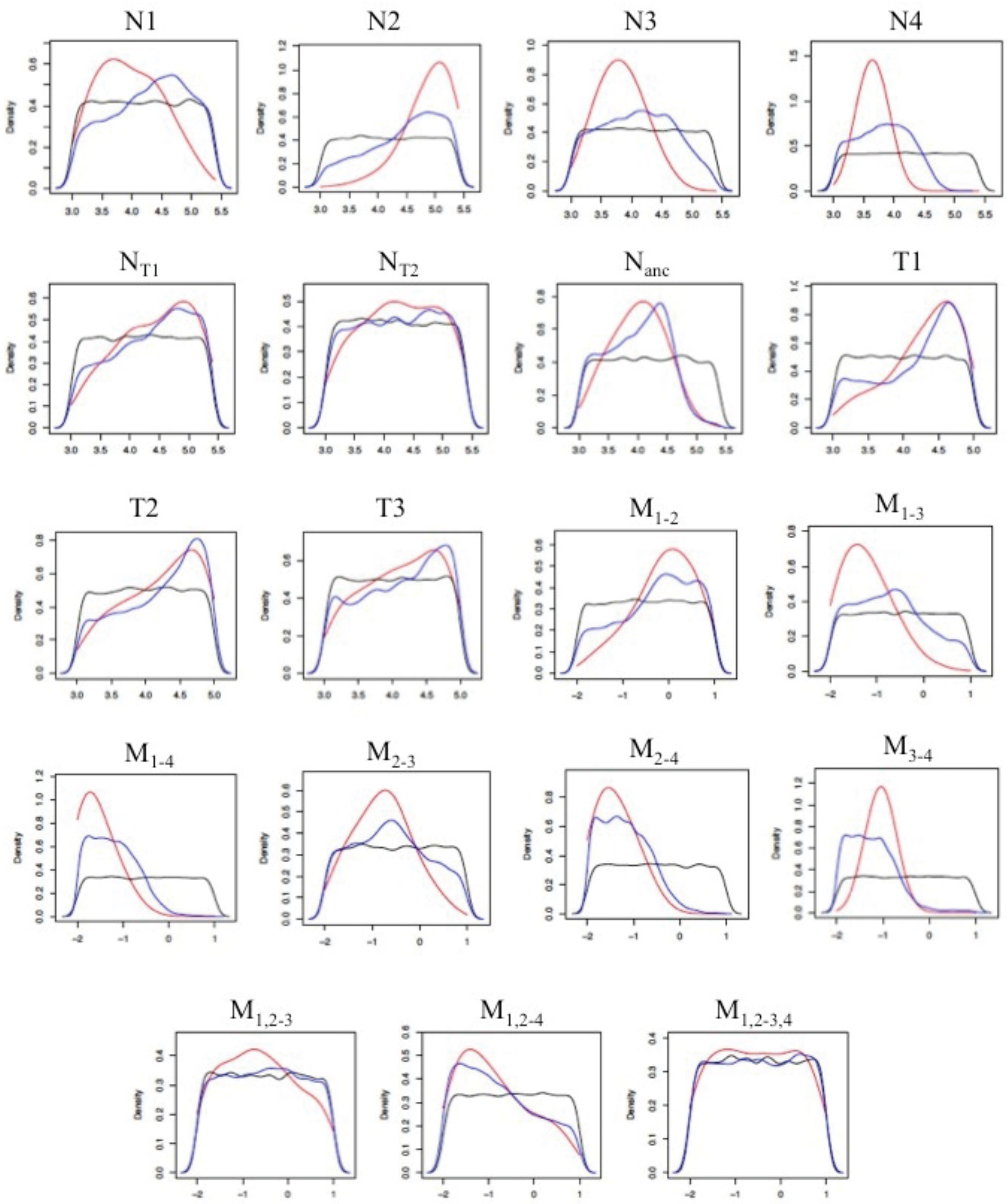
Suppl. Figure 12.3.5A - *PCA 1 and 2 of summary statistics for Models 1 and 2 (black), Model 3, 4A and 4B (blue) and observed data (red).*



Suppl. Figure 12.3.5B - *PCA 3 and 4 of summary statistics for Models 1 and 2 (black), Model 3, 4A and 4B (blue) and observed data (red).*



Suppl. Figure 12.3.6 - *Prior (black), retained (blue) and GLM adjusted posterior distributions under Model 3.*



Suppl. Figure 12.3.7 - *Prior (black), retained (blue) and GLM adjusted posterior distributions under Model 4A.*

12.4 PSMC

Heng Li, Jeffrey M. Kidd, Joanna L. Kelley, Carlos D. Bustamante, David Reich

Methods

1.1 Calling consensus sequence

We aligned great ape short reads to the hg18 human genome with BWA[2]. We called the consensus using SAMtools[74][1]. For the PSMC analysis[75], we selected samples from each subspecies (**Suppl. Table 12.4.1**) such that 1) each sample has relatively high read depth in each subspecies; 2) each sample is known to have a low contamination level and without evident hybridization according to the principle component analysis (PCA); and 3) if allowed, at least three samples, including one male sample, are chosen from each subspecies. A few samples were dropped after the PSMC analysis due to excessively large inferred population size, which is typically an indication of poor consensus calling or contamination.

1.2 Scaling population parameters

We use d, the number of substitutions per base between a pair of sequences (i.e., pairwise sequence divergence), to measure time, and use θ, the scaled mutation rate, to measure the effective population size. The advantage of such scaling is that both d and θ can be directly inferred from sequence data without using any additional scaling parameters that cannot be determined by a coalescent model. When we know the generation time g and the mutation rate per base per generation μ from other sources, $d/(2\mu/g)$ gives the time in years and $\theta/(4\mu)$ gives the effective population size. For primates, μ/g is typically ranged from 10^{-9} to 0.5×10^{-9} per base per year. This value may differ slightly across species and might have been changed over the past 20 million years[76].

1.3 Inferring population size history

We inferred the historical population size with PSMC[22]. We measured the variance of the estimate by bootstrapping: we selected one sample from each subspecies (**Suppl. Table 12.4.1**), split its consensus into 10 Mbp segments, randomly resampled about 300 segments with replacement, and then rerun PSMC on the resampled segments. We repeated the procedure 100 times. The fluctuation of the 100 rounds of inferences suggests the variance.

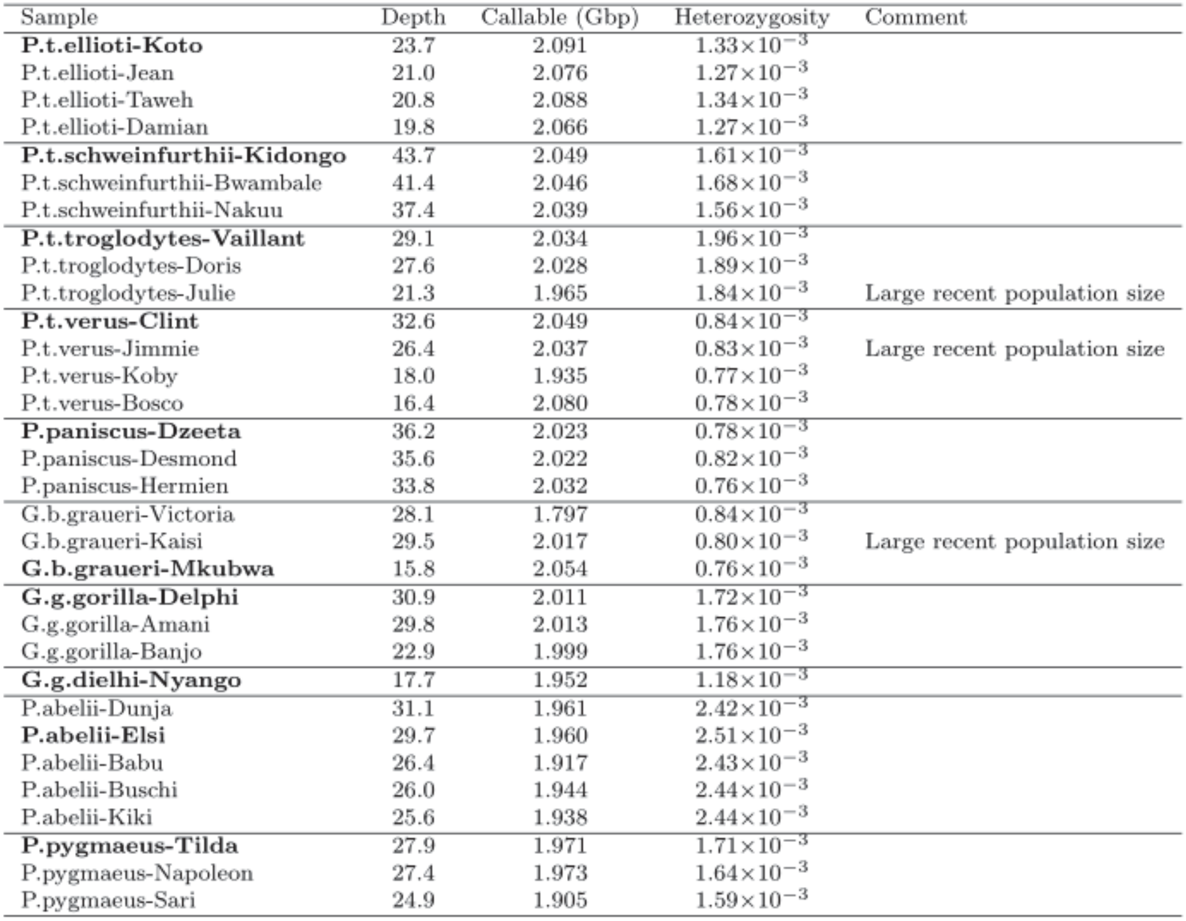
[1] Command line: “SAMtools mpileup -Euf ref.fa aln.bam | bcftools view -c -| vcfutils.pl vcf2fq -d min depth -D max depth” where min depth is set to 1/3 of the average read depth and max depth set to twice of the average.

[2] Command line: “psmc -N25 -t15 -r5 -p "4+25*2+4+6" -o result.psmc cns.psmcfa”.

1.4 Inferring divergence time

When we plot the PSMC inferences of two subspecies together, the time point where the two historical population sizes diverge approximates the divergence time. However, although this approach is intuitive and works apparently well, it has several problems. Firstly, the plot does not provide quantified time. Telling where sizes of two subspecies diverge is not always obvious and, at times, subjective. Secondly, it is possible that the two subspecies had the same size after the split, which will lead to an underestimate. Thirdly, there might be considerable gene flows between the two subspecies after the initial split. Differentiation in population sizes may not correspond to the final split. At last, PSMC has a known artifact where it may smooth out sudden size changes and push back the divergence time (**Suppl. Figure 12.4.1**). The PSMC plot only gives us a qualitative sense of the divergence time.

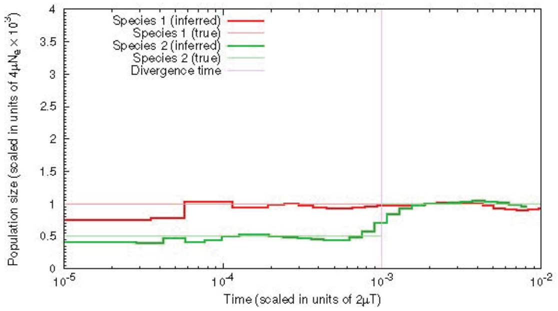
A second approach to infer divergence time is to hybridize two haploid sequences from each species and then run PSMC on the pseudo-diploid sequence. The time point where the inferred population goes to infinity corresponds to the divergence time. As the samples we use are diploid, we randomly select an allele at a heterozygous site to derive a haploid sequence. This approximation works well if the speciation time is much deeper than the average the most recent common ancestor (TMRCA) of each species such that most heterozygotes arose after the speciation. This method gives a stronger signal of speciation than the first method, but it does not quantify the time either.



Sample	Depth	Callable (Gbp)	Heterozygosity	Comment
P.t.ellioti-Koto	23.7	2.091	1.33×10^{-3}	
P.t.ellioti-Jean	21.0	2.076	1.27×10^{-3}	
P.t.ellioti-Taweh	20.8	2.088	1.34×10^{-3}	
P.t.ellioti-Damian	19.8	2.066	1.27×10^{-3}	
P.t.schweinfurthii-Kidongo	43.7	2.049	1.61×10^{-3}	
P.t.schweinfurthii-Bwambale	41.4	2.046	1.68×10^{-3}	
P.t.schweinfurthii-Nakuu	37.4	2.039	1.56×10^{-3}	
P.t.troglodytes-Vaillant	29.1	2.034	1.96×10^{-3}	
P.t.troglodytes-Doris	27.6	2.028	1.89×10^{-3}	
P.t.troglodytes-Julie	21.3	1.965	1.84×10^{-3}	Large recent population size
P.t.verus-Clint	32.6	2.049	0.84×10^{-3}	
P.t.verus-Jimmie	26.4	2.037	0.83×10^{-3}	Large recent population size
P.t.verus-Koby	18.0	1.935	0.77×10^{-3}	
P.t.verus-Bosco	16.4	2.080	0.78×10^{-3}	
P.paniscus-Dzeeta	36.2	2.023	0.78×10^{-3}	
P.paniscus-Desmond	35.6	2.022	0.82×10^{-3}	
P.paniscus-Hermien	33.8	2.032	0.76×10^{-3}	
G.b.graueri-Victoria	28.1	1.797	0.84×10^{-3}	
G.b.graueri-Kaisi	29.5	2.017	0.80×10^{-3}	Large recent population size
G.b.graueri-Mkubwa	15.8	2.054	0.76×10^{-3}	
G.g.gorilla-Delphi	30.9	2.011	1.72×10^{-3}	
G.g.gorilla-Amani	29.8	2.013	1.76×10^{-3}	
G.g.gorilla-Banjo	22.9	1.999	1.76×10^{-3}	
G.g.dielhi-Nyango	17.7	1.952	1.18×10^{-3}	
P.abelii-Dunja	31.1	1.961	2.42×10^{-3}	
P.abelii-Elsi	29.7	1.960	2.51×10^{-3}	
P.abelii-Babu	26.4	1.917	2.43×10^{-3}	
P.abelii-Buschi	26.0	1.944	2.44×10^{-3}	
P.abelii-Kiki	25.6	1.938	2.44×10^{-3}	
P.pygmaeus-Tilda	27.9	1.971	1.71×10^{-3}	
P.pygmaeus-Napoleon	27.4	1.973	1.64×10^{-3}	
P.pygmaeus-Sari	24.9	1.905	1.59×10^{-3}	

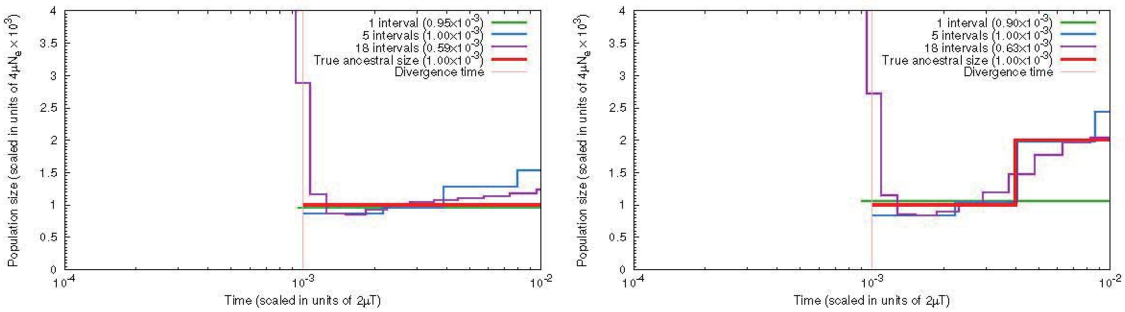
Suppl. Table 12.4.1 – *Samples used for PSMC analysis. The average read depth (second column) is estimated at HapMap3 sites. The third column gives the number of sites in hg18 where a genotype can be called confidently. A site is masked as 'uncalled' if at the site: 1) the read depth is more than twice or less than one-third of the average depth; 2) the site is within 5 bp around a predicted short indel; 3) the root-mean-square mapping quality is below 10; 4) the estimated consensus quality is below 30; and 5) fewer than 18 out of 35 overlapping 35-mer from hg18 can be mapped elsewhere with zero or one mismatch. Heterozygosity is estimated in callable regions only. PSMC bootstrapping is applied to samples in the bold font face*.

We will propose a third approach to quantify divergence time with a small modification to the original PSMC model; we assume no coalescences after divergence. The divergence time is just another parameter of the PSMC model that can be estimated together with population sizes. A caveat is that divergence has a similar effect to infinite population size. When we use many small time intervals, PSMC will be confused by the two scenarios and underestimate the time.



Suppl. Figure 12.4.1 – *PSMC inference given sudden population size changes. In simulation, species 1 keeps a constant population size, while species 2 has halved its population size immediately after the divergence at x = 10^{-3}. The ms command line in use is: "ms4500-t1000-r300 1000000 -I2 22 -en011-en02 0.5 -ej 0.5 2 1 -eN 0.500001 1".*

Our temporary solution is to use fewer size parameters to force the population size around divergence to be small. This is not an ideal solution, but it seems to work on simulated data. In **Suppl. Figure 12.4.2**, we estimate the divergence time by fitting the ancestral population with 1, 5 or 18 time intervals[3]. With many intervals, PSMC infers excessively large population size around the speciation. Although this is also an indication of speciation, the time is underestimated. Using fewer intervals gives accurate estimates, especially when PSMC can estimate the ancestral population size changes well.



Suppl. Figure 12.4.2 – *Estimating divergence time on simulated data. Numbers in parentheses give the inferred divergence time. The thick red line shows the simulated ancestral population size.*

[3] The '-p' parameter used by psmc is set to '40', '20+4*5' and '6+17*2', respectively.

Results

2.1 Population size history and divergence time

Figure 3 shows the population size history for human and primates. Overall, samples from the same subspecies/population agree well with each other and the fluctuation between samples largely falls within the variance of a single sample. The ancestral population sizes of different species before speciation also match well. These are expected.

Figure 3 already hints at the divergence time between subspecies, but as we discussed earlier, the time is not quantitative and may be subjected to artifacts. To quantify the divergence time, we applied PSMC with divergence time as an extra parameter. **Suppl. Table 12.4.2** shows the inferred divergence time ds, the ancestral population size θs before the speciation. The standard deviation estimated by bootstrapping for the *P.t.verus*-Clint and *P.t.schweinfurthii*-Kidongo sample pair is 0.012×10^{-3}. The consistency between sample pairs from same species pairs also suggests a small variance predicted by the model.

It is worth noting that for a constant ancestral population with a clean speciation, we would expect dg = ds + θs[4]. This is not true in the table because: 1) we fit the ancestral population using a piece-wise constant function with five intervals; 2) segmental duplications in great apes are shared between samples and will inflate dg; and 3) a real history deviating from a simple speciation model will break the equality. If we simulate a constant ancestral population size and fit the ancestral population with one time interval, dg = ds + θs approximately stands.

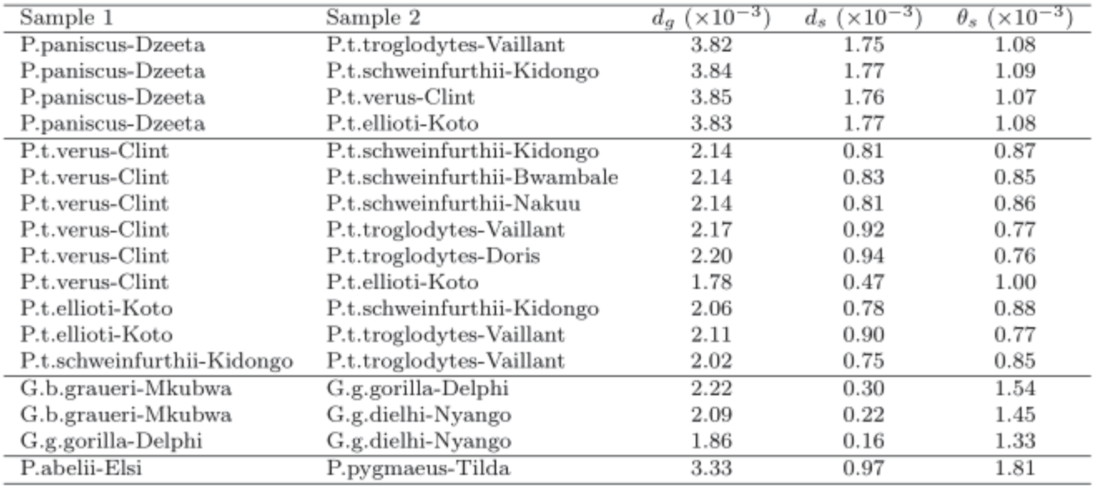
[4] For a population with a constant effective population size Ns, the average coalescent time between two sequences from the population is 2Ns generations. If two species diverged Ts generations ago, the average coalescent time between two sequences from each species is Tg = Ts +2Ns. Multiplying 2μ to each side of the equation yields dg = ds + θs.

For most sample pairs, **Suppl. Table 12.4.2** broadly agrees with **Figure 3**, with a few exceptions. Firstly and most strikingly, P.t.verus and P.t.ellioti diverged more recently than what we see from **Figure 3**. This implies considerable genetic exchanges between the two subspecies after the initial split. Secondly, P.t.verus and P.t.ellioti seem to be closer to P.t.schweinfurthii than to P.t.troglodytes, but this is not obvious from **Figure 3**. This observation might suggest more gene flow between Western and Eastern chimpanzees than between Western and Central several hundred thousand years ago. Thirdly, PSMC predicts G.b.graueri to be closer to G.g.diehli than to G.g.gorilla. Similarly this may imply unbalanced gene flows.

To formally test unbalanced gene flows between species, we performed a D-statistic test (**Suppl. Table 12.4.3**). This test also suggests that Western chimpanzees are genetically closer to Eastern chimpanzees than to Central and that Eastern lowland gorillas are closer to Cross River gorillas than to Western lowland gorillas. The test prefers to put Western and Nigeria-Cameroon chimpanzees in one clade, though it also implies the latter is closer to Central and Eastern chimpanzees.

2.2 Comparison to the previous studies

When we compare the divergence time from different studies, a major complication is time scaling. There are typically two approaches to time scaling. The first is to assume a fixed mutation rate μ and generation time g and to use the two parameters to scale time to years.

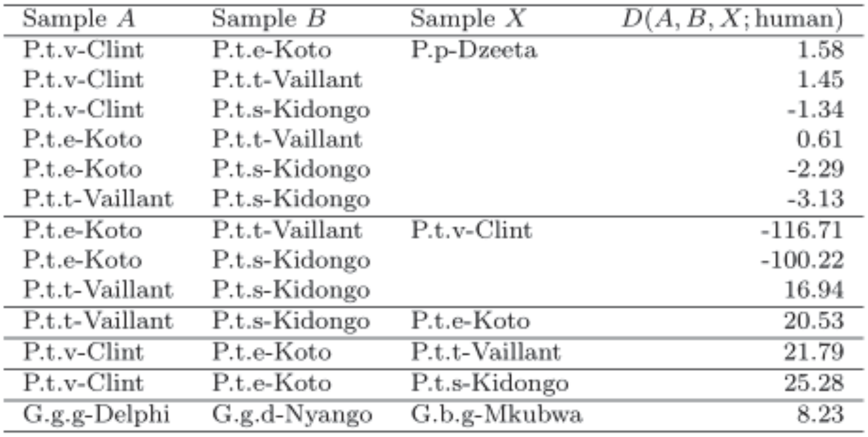


Sample 1	Sample 2	d_g $(\times 10^{-3})$	d_s $(\times 10^{-3})$	θ_s $(\times 10^{-3})$
P.paniscus-Dzeeta	P.t.troglodytes-Vaillant	3.82	1.75	1.08
P.paniscus-Dzeeta	P.t.schweinfurthii-Kidongo	3.84	1.77	1.09
P.paniscus-Dzeeta	P.t.verus-Clint	3.85	1.76	1.07
P.paniscus-Dzeeta	P.t.ellioti-Koto	3.83	1.77	1.08
P.t.verus-Clint	P.t.schweinfurthii-Kidongo	2.14	0.81	0.87
P.t.verus-Clint	P.t.schweinfurthii-Bwambale	2.14	0.83	0.85
P.t.verus-Clint	P.t.schweinfurthii-Nakuu	2.14	0.81	0.86
P.t.verus-Clint	P.t.troglodytes-Vaillant	2.17	0.92	0.77
P.t.verus-Clint	P.t.troglodytes-Doris	2.20	0.94	0.76
P.t.verus-Clint	P.t.ellioti-Koto	1.78	0.47	1.00
P.t.ellioti-Koto	P.t.schweinfurthii-Kidongo	2.06	0.78	0.88
P.t.ellioti-Koto	P.t.troglodytes-Vaillant	2.11	0.90	0.77
P.t.schweinfurthii-Kidongo	P.t.troglodytes-Vaillant	2.02	0.75	0.85
G.b.graueri-Mkubwa	G.g.gorilla-Delphi	2.22	0.30	1.54
G.b.graueri-Mkubwa	G.g.dielhi-Nyango	2.09	0.22	1.45
G.g.gorilla-Delphi	G.g.dielhi-Nyango	1.86	0.16	1.33
P.abelii-Elsi	P.pygmaeus-Tilda	3.33	0.97	1.81

Suppl. Table 12.4.2 – *Inferred speciation time and ancestral population size. For each sample, a pseudo-haploid sequence was derived by choosing a random allele at a heterozygote. dg is the average sequence divergence between a pair of pseudo-haploid sequences from two samples, excluding uncalled regions in either sample. ds is the PSMC inferred speciation time and θs is the inferred scaled mutation rate (proportional to the effective population size) right before the* *speciation. ds and θs were estimated from a pseudo-diploid sequence generated by hybridizing two haploid sequences from different samples. Five population size parameters were used to fit the ancestral population size history.*

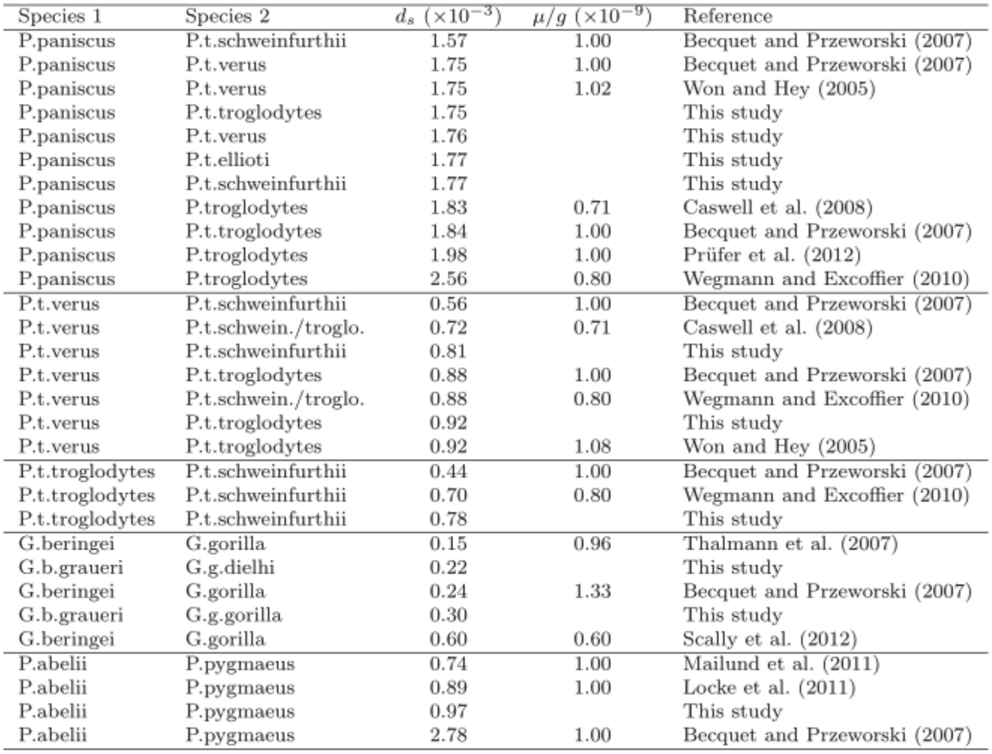
For time scaled this way, it is easy to convert the time back to sequence divergence ds. The second approach to time scaling is to scale the inferred time by the human-chimpanzee sequence divergence. If we assume the average human-chimpanzee genetic divergence to be 7 million years ago (mya) or so, we can derive time in years. The latter method aims to account for variable mutation rates in different regions. However, due to the large ancestral population size, the human-chimpanzee sequence divergence may vary greatly around 7 mya. The effectiveness of the second approach might be debatable. In addition, for the second approach, it is still possible to scale time back to sequence divergence ds if the divergence between human and chimpanzee is given.

A second complication is that when the speciation is not clean, it may not be straightforward to precisely define the divergence time. A population model may fold other factors, such as present and ancestral population sizes, migration, and structure into divergence time. For example, a model without considering migration will prefer more recent divergence time in comparison to an isolation-migration model. The divergence time estimate from different studies may not be strictly comparable. **Suppl. Table 12.4.4** shows the speciation time between pairs of subspecies in the previous studies[55–57,77–79]. Our PSMC estimates tend to be close to the majority.



Sample A	Sample B	Sample X	$D(A, B, X; \text{human})$
P.t.v-Clint	P.t.e-Koto	P.p-Dzeeta	1.58
P.t.v-Clint	P.t.t-Vaillant		1.45
P.t.v-Clint	P.t.s-Kidongo		-1.34
P.t.e-Koto	P.t.t-Vaillant		0.61
P.t.e-Koto	P.t.s-Kidongo		-2.29
P.t.t-Vaillant	P.t.s-Kidongo		-3.13
P.t.e-Koto	P.t.t-Vaillant	P.t.v-Clint	-116.71
P.t.e-Koto	P.t.s-Kidongo		-100.22
P.t.t-Vaillant	P.t.s-Kidongo		16.94
P.t.t-Vaillant	P.t.s-Kidongo	P.t.e-Koto	20.53
P.t.v-Clint	P.t.e-Koto	P.t.t-Vaillant	21.79
P.t.v-Clint	P.t.e-Koto	P.t.s-Kidongo	25.28
G.g.g-Delphi	G.g.d-Nyango	G.b.g-Mkubwa	8.23

Suppl. Table 12.4.3 – *D-statistic for four haploid sequences, A, B, X and O, a site is classified as ABBA if at the site base A = O B = X, or classified as BABA if B = O= A = X. Define D′(A, B, X; O)== (#ABBA −#BABA)/(#ABBA +#BABA). D(A, B, X; O) equals the ratio of the mean of D′ to its standard deviation, estimated by block Jack-knife. A positive D value indicates that sample B is* *genetically closer to X, while a negative vale indicates A closer to X.*

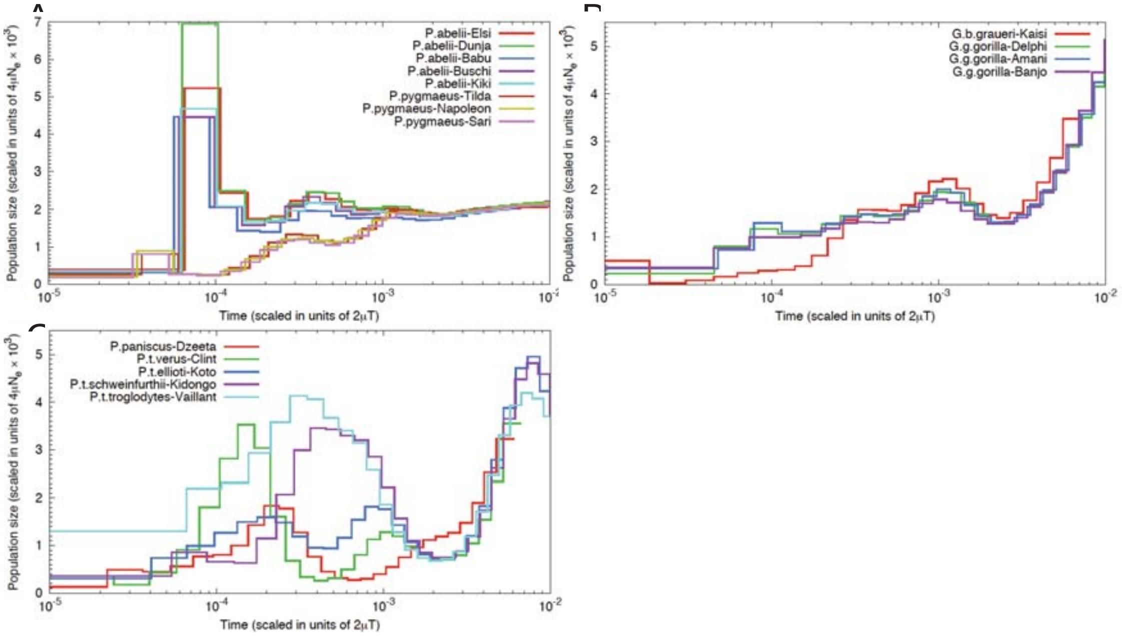


Species 1	Species 2	d_s ($\times10^{-3}$)	μ/g ($\times10^{-9}$)	Reference
P.paniscus	P.t.schweinfurthii	1.57	1.00	Becquet and Przeworski (2007)
P.paniscus	P.t.verus	1.75	1.00	Becquet and Przeworski (2007)
P.paniscus	P.t.verus	1.75	1.02	Won and Hey (2005)
P.paniscus	P.t.troglodytes	1.75		This study
P.paniscus	P.t.verus	1.76		This study
P.paniscus	P.t.ellioti	1.77		This study
P.paniscus	P.t.schweinfurthii	1.77		This study
P.paniscus	P.troglodytes	1.83	0.71	Caswell et al. (2008)
P.paniscus	P.t.troglodytes	1.84	1.00	Becquet and Przeworski (2007)
P.paniscus	P.troglodytes	1.98	1.00	Prüfer et al. (2012)
P.paniscus	P.troglodytes	2.56	0.80	Wegmann and Excoffier (2010)
P.t.verus	P.t.schweinfurthii	0.56	1.00	Becquet and Przeworski (2007)
P.t.verus	P.t.schwein./troglo.	0.72	0.71	Caswell et al. (2008)
P.t.verus	P.t.schweinfurthii	0.81		This study
P.t.verus	P.t.troglodytes	0.88	1.00	Becquet and Przeworski (2007)
P.t.verus	P.t.schwein./troglo.	0.88	0.80	Wegmann and Excoffier (2010)
P.t.verus	P.t.troglodytes	0.92		This study
P.t.verus	P.t.troglodytes	0.92	1.08	Won and Hey (2005)
P.t.troglodytes	P.t.schweinfurthii	0.44	1.00	Becquet and Przeworski (2007)
P.t.troglodytes	P.t.schweinfurthii	0.70	0.80	Wegmann and Excoffier (2010)
P.t.troglodytes	P.t.schweinfurthii	0.78		This study
G.beringei	G.gorilla	0.15	0.96	Thalmann et al. (2007)
G.b.graueri	G.g.dielhi	0.22		This study
G.beringei	G.gorilla	0.24	1.33	Becquet and Przeworski (2007)
G.b.graueri	G.g.gorilla	0.30		This study
G.beringei	G.gorilla	0.60	0.60	Scally et al. (2012)
P.abelii	P.pygmaeus	0.74	1.00	Mailund et al. (2011)
P.abelii	P.pygmaeus	0.89	1.00	Locke et al. (2011)
P.abelii	P.pygmaeus	0.97		This study
P.abelii	P.pygmaeus	2.78	1.00	Becquet and Przeworski (2007)

Suppl. Table 12.4.4 – *Speciation time in the literature. The speciation time in all the previous studies is converted to sequence divergence ds where possible. μ/g is the per-base per-year mutation rate used in the corresponding study. dHC is the average human-chimpanzee sequence divergence.*

Comparison of PSMC Results

To assess the effect of reference sequence divergence on our analysis, we applied the PSMC method using the mappings to each species reference. The mappings were processed using BWA, Picard, and GATK as described above. We made diploid consensus sequence calls using SAMtools and ran PSMC as previously described (with psmc -N 25 -t 15 -r 5 -p "4+25*2+4+6"). To avoid potential artifacts, we limited analysis to a subset of samples with high coverage and showing low evidence of potential contamination. These results recapitulate the basic patterns observed from analysis using hg18 (see **Suppl. Figure 12.4.3**) both in terms of relative effective population sizes among subspecies and apparent times in which demographic trajectories diverge. We note, however, that for the orangutans, mapping to the orangutan genome shows a qualitative difference in the inferred effective population sizes for *Pongo abelii*.



Suppl. Figure 12.4.3 – *PSMC analysis based on mappings to species reference genome assemblies. We limited analysis to a subset of samples with high coverage and low evidence of contamination including eight orangutans (A), four gorillas (B), and four chimpanzees and one bonobo (C).*

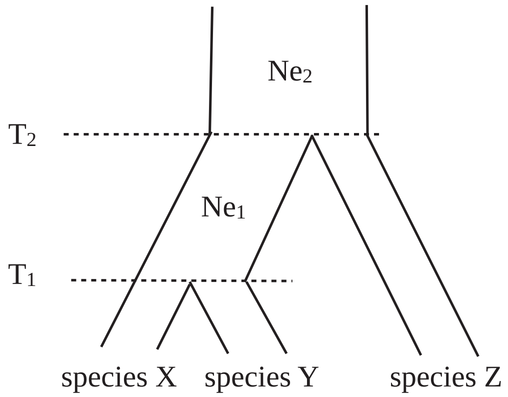
Section 13: Analysis of demography and incomplete lineage sorting (ILS)

13.1 Incomplete lineage sorting (ILS)

Kasper Munch, Thomas Mailund, Mikkel H. Schierup

The coalescent hidden Markov model (CoalHMM)

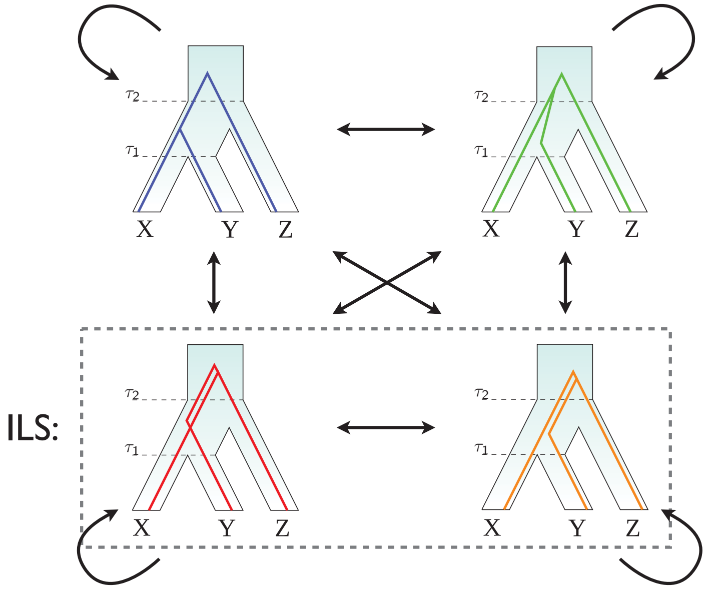
A consequence of ILS is that segments of a genomic alignment have evolutionary relationships different from the species tree. The CoalHMM framework[80] allows for inference of population genetic parameters and patterns of ILS and the model is based on a hidden Markov chain with hidden states representing gene trees with separate topologies and separate coalescent times. The model is thus able to represent incomplete lineage sorting along a genomic alignment. The model applied is a three species isolation model (**Suppl. Figure 13.1**) with the following demographic parameters: two ancestral population sizes, Ne_1, Ne_2, and two speciation times, T_1, T_2. These parameters are all scaled with the substitution rate.



Suppl. Figure 13.1 – *Isolation model used in the analysis. T_1: speciation time of species X and Y. T_2: speciation time of species Y and Z. Ne_1: effective population size of the population size ancestral to species X and Y. Ne_2: effective population size of the ancestor to all three species.*

The CoalHMM operates with four different trees connecting three species: species X and Y may find a common ancestor in their ancestral population (**Suppl. Figure 13.2** top left) or in the population ancestral to all three species (**Suppl. Figure 13.2** top right), and species Z may find a common ancestor with either X or Y in the population ancestral to all three species (**Suppl. Figure 13.2** bottom left and right).

The model is applied in turn to genomic alignments of four species of apes of which one only serves as outgroup. Individuals used are listed in **Suppl. Table 13.1** and combinations of individuals used in each analysis are listed in **Suppl. Table 13.2**.



Suppl. Figure 13.2 – *The four hidden states in the HMM. The four states correspond to the four different trees describing the ancestry of an alignment column. Arrows indicate possible transitions.*

P. tro. tro. I	*Pan troglodytes_troglodytes* A959 Julie
P. tro. tro. II	*Pan troglodytes_troglodytes* A960 Clara
P. tro. sch.	*Pan troglodytes_schweinfurthii* 9729 Harriet
P. tro. ver.	*Pan troglodytes_verus* 9668 Bosco
P. tro. eli.	*Pan troglodytes_ellioti* Koto
P. pan.	*Pan paniscus* A915 Kosana
H. sap.	*Homo sapiens* San HGDP01029
G. gor. gor.	*Gorilla gorilla gorilla* A933 Dian
G. gor. ber.	*Gorilla beringei_graueri* Victoria
P. pan. abe.	*Pongo abelii* A950 Babu

Suppl. Table 13.1 – *Key for individuals included in analyses.*

1	P. tro. tro. I	P. pan	H. sap.	P. pan. abe.
2	P. tro. tro. II,	P. pan	H. sap.	P. pan. abe.
3	P. tro. sch.,	P. pan.	H. sap.	P. pan. abe.
4	P. tro. ver.,	P. pan	H. sap.	P. pan. abe.
5	P. tro. eli.,	P. pan	H. sap.	P. pan. abe.
6	P. tro. tro. I	H. sap.	G. gor. gor	P. pan. abe.

7	P. tro. tro. I	H. sap.	G. gor. ber.	P. pan. abe.

Suppl. Table 13.2 – *Combinations of individuals used in analyses.*

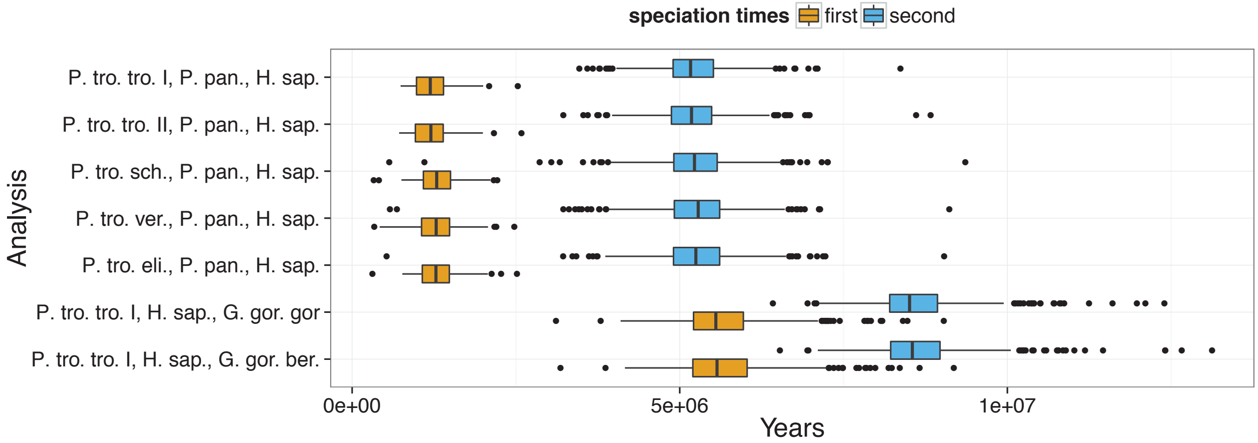
Preparation of input alignments

To generate input alignments we map all called SNPs to the human reference genome (NCBI Build 36). Bases not called are substituted with N to indicate missing data. This results in a genomic sequence in human coordinates for each individual. These sequences are further masked using the RepeatMasker track from the UCSC genome browser. Genomic sequences from four individuals from different species are chosen. The implicit alignment that these form is filtered removing consecutive runs of more than 100 alignment columns of all N characters splitting alignment accordingly. To aid the computation, an upper bound on the length of consecutive alignment of 100,000 bases is imposed and the alignment is split accordingly. An individual CoalHMM analysis is performed on ~1 Mbp of such alignment to estimate demographic model parameters and proportions of ILS.

Estimates of model parameters

The distributions of speciation times estimated on individual 1 Mbp alignments are shown in **Suppl. Figure 13.3**. Estimates for individual species combinations are listed in **Suppl. Table 13.3-13.6** along with confidence intervals and standard error of the mean. CoalHMM measures time in mutations rather than in years but the estimated model parameters can be rescaled to years using a per-year mutation rate and a generation time. Here we have used a per-year mutation rate of 0.6e-9 per site[64] and a generation time of 25 years. All parameters scale linearly with mutation rate and population sizes scale linearly with generation time, making it straightforward to obtain split times and population sizes from an alternative choice of rescaling.

We found a general agreement between analyses when including different subspecies of chimpanzee and gorilla, but also when including chimpanzee, human and gorilla compared to chimpanzee, bonobo and human. The chimpanzee-bonobo speciation time is slightly lower, but significantly, for those analyses including Central chimpanzee than for the analyses including Nigeria-Cameroon, Western or Eastern chimpanzee. The p-value of a Kruskal-Wallis test on estimates from all analyses is 3.4e-12, whereas it is 0.94 and 0.76 when performing the test on the analyses including and excluding Central chimpanzee, respectively. This would be in line with a notion that bonobos split from a population ancestral to Central chimpanzees.



Suppl. Figure 13.3 – *Distributions of estimated split times (five outliers not shown).*

Chimpanzee subsp.	Split time	1.96 * SE of mean
P. tro. eli.	1,285,736	23,507
P. tro. ver.	1,284,132	23,564
P. tro. sch.	1,294,088	23,681
P. tro. tro. II	1,204,658	23,539
P. tro. tro. I	1,203,596	23,622

Suppl. Table 13.3 – *Chimpanzee-bonobo speciation time estimated from chimpanzee, bonobo, human, and orangutan analyses.*

Chimpanzee subsp.	Split time	1.96 * SE of mean
P. tro. eli.	5,244,082	53,789
P. tro. ver.	5,262,635	55,058
P. tro. sch.	5,235,109	54,378
P. tro. tro. II	5,205,198	49,153
P. tro. tro. I	5,201,148	46,426

Suppl. Table 13.4 – *Chimpanzee-human speciation time estimated from chimpanzee, bonobo, human, and orangutan analyses.*

Gorilla subsp.		
G. gor. ber.	5,684,228	62,533
G. gor. gor.	5,648,467	61,730

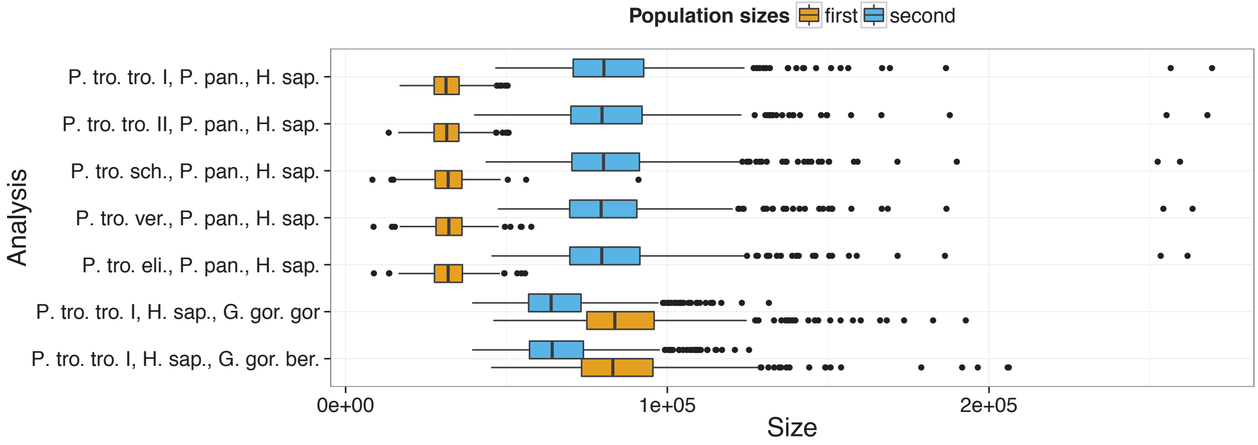
Suppl. Table 13.5 – *Chimpanzee-human speciation time estimated from chimpanzee, human, gorilla, orangutan analyses.*

Gorilla subsp.	Split time	1.96 * SE of mean
G. gor. ber.	8,669,901	65,141
G. gor. gor.	8,644,655	62,319

Suppl. Table 13.6 – *Chimpanzee-gorilla speciation time estimated from chimpanzee, human, gorilla, and orangutan analyses.*

Ancestral population sizes

The distributions of estimated ancestral population sizes on individual 1 Mbp of alignment are shown in **Suppl. Figure 13.4**. Estimates for individual species combinations are listed in **Suppl. Tables 13.7-13.10** along with standard error of the mean.



Suppl. Figure 13.4 – *Distributions of estimated effective population sizes.*

Chimpanzee subsp.	Ne	1.96 * SE of mean
P. tro. eli.	32,000	525
P. tro. ver.	32,189	515
P. tro. sch.	32,036	542
P. tro. tro. II	31,532	501
P. tro. tro. I	31,573	477

Suppl. Table 13.7 – *Chimpanzee-bonobo population size estimated from chimpanzee, bonobo, human, and orangutan analyses.*

Chimpanzee subsp.	Ne	1.96 * SE of mean
P. tro. eli.	83,133	1,868
P. tro. ver.	83,015	1,798
P. tro. sch.	83,413	1,843
P. tro. tro. II	83,400	1,860
P. tro. tro. I	84,032	1,858

Suppl. Table 13.8 – *Chimpanzee-Human population size estimated from chimpanzee, bonobo, human, and orangutan analyses.*

Gorilla subsp.	Ne	1.96 * SE of mean
G. gor. ber.	86,230	1,804
G. gor. gor.	86,967	1,747

Suppl. Table 13.9 – *Chimpanzee-Human population size estimated from chimpanzee, human, gorilla, orangutan analyses*.

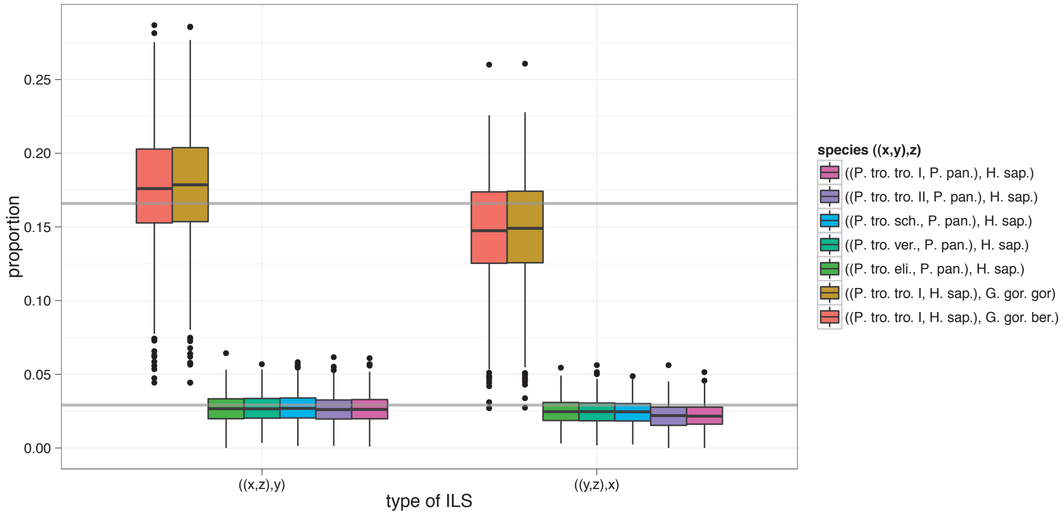
Gorilla subsp.	Ne	1.96 * SE of mean
G. gor. ber.	67,123	1,274
G. gor. gor.	66,747	1,267

Suppl. Table 13.10 – *Chimpanzee-Gorilla population size estimated from chimpanzee, human, gorilla, orangutan analyses*.

Proportions of ILS

The most likely hidden state is assigned to each column of analyzed alignment using posterior decoding of the optimized hidden Markov model (HMM). The proportion of each type of ILS is calculated in 1 Mbp windows in human reference coordinates. Only windows covering at least 500 Kbp of alignment are included. The distribution of inferred proportions of ILS is shown in **Suppl. Figure 13.5** with summary statistics listed in **Suppl. Table 13.11**. The two types of ILS, ((x,z),y) and ((y,x),x), should be in equal proportions. In the two chimpanzee-human-gorilla analyses, however, the estimated proportion of ILS where chimpanzee and gorilla coalesce first is larger than the proportion of ILS with human and gorilla coalescing first. CoalHMM assumes an ultrametric tree and a violation of this assumption will produce such an effect—in this case if the human branch is longer. SNPs are called by mapping reads to the human reference genome and this may potentially result in higher sensitivity in calling human SNP, thus producing the observed effect.

The estimated proportions agree well with the theoretically expected proportion that is readily calculated from rescaled estimates as exp[-(T2-T1)/(25*2*Ne1)]. This yields 2.9% for chimpanzee-bonobo-human ILS assuming a chimpanzee-bonobo speciation at 1.3 mya, a chimpanzee-human speciation at 5.2 mya, and a chimpanzee-bonobo ancestral population size of 32,000. The proportion of chimpanzee-human-gorilla ILS is expected to be 16.6% assuming a chimpanzee-human speciation at 5.6 mya, a human-gorilla speciation at 8.6 mya, and a chimpanzee-human ancestral population size of 86,000. Note that the theoretically expected proportions of ILS are not dependent on the mutation rate and generation time used for rescaling. The expected proportions are shown in **Suppl. Figure 13.5** as gray lines.



Suppl. Figure 13.5 – *Distributions of the proportion of ILS in 1 Mbp windows in human reference coordinates.*

Species: ((x,y),z)	% ((x,z),y)	% ((y,z),x)
((P. tro. eli., P. pan), H. sap)	2.7 -/+ 0.001	2.5 -/+ 0.001
((P. tro. ver., P. pan), H. sap)	2.7 -/+ 0.001	2.5 -/+ 0.001
((P. tro. sch., P. pan), H. sap)	2.7 -/+ 0.001	2.5 -/+ 0.001
((P. tro. tro. I,P. pan), H. sap)	2.7 -/+ 0.001	2.2 -/+ 0.001
((P. tro. tro. II, P. pan), H. sap)	2.7 -/+ 0.001	2.2 -/+ 0.001
((P. tro. tro. I, H. sap.), G. gor. ber.)	17.5 -/+ 0.004	14.7 -/+ 0.004
((P. tro. tro. I, H. sap.), G. gor. gor.)	17.6 -/+ 0.004	14.8 -/+ 0.004

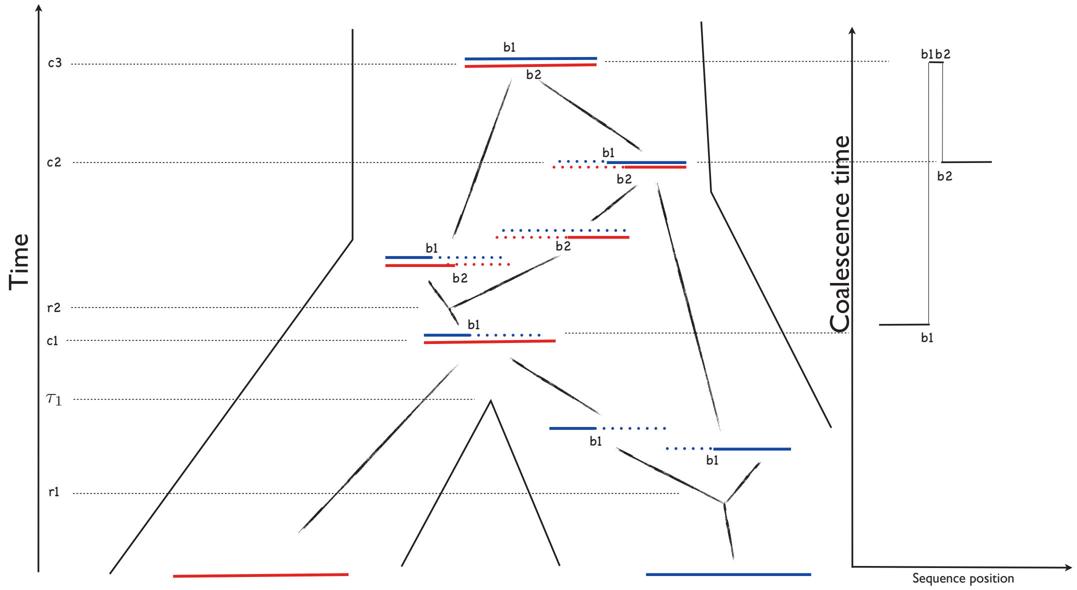
Suppl. Table 13.11 – *Proportion of ILS in 1 Mbp windows in human reference coordinates. Confidence interval is calculated as -/+ 1.96 * standard error of mean.*

13.2 Isolation model CoalHMM

Kasper Munch, Thomas Mailund, Anders E. Halager, Mikkel H. Schierup

Model

The coalescence with recombination is a model of the local genealogy along a sample of genomes. For two haploid genomes, it models how the time to TMRCA changes as one scans along a genome alignment.



Suppl. Figure 13.6 – *Example ancestral recombination graph for a sequence segment from two different species. Recombinations can occur at any point in time but coalescence can only occur once in the common ancestor, i.e., further back in time than the split tau1. Right graph shows TMRCA along the sequence.*

The distribution of TMRCA, and the distribution of segment lengths sharing TMRCA, is determined by the split time between the populations/species the genomes are taken from, the effective population sizes in the two populations and the ancestral population, and the recombination rate along the genomes (**Suppl. Figure 13.6**). By modeling the changes in TMRCA along the genome alignment in an HMM, we can thus infer these parameters[79].

Data preparation

We obtained genomes from each individual as in **section 12.** We then translated these diploid genomes into haploid genomes by, for each heterozygotic site, picking an allele at random. This means that the phase of the haploid genomes changes as we scan along the genome, but if the level of shared polymorphism between species is low (see simulations[79] supplemental text

S1), it does not change TMRCA between the genomes, which is what the method uses for its inference.

We then further filtered the data by running a sliding window along the genome, with a window size of 1 Kbp and a jump of 100 bp, removing windows with more than 10% missing data, since we found that regions with missing data display a greater variance in divergence than regions without missing data. We analyzed the genome in chunks of 10 Mbp of alignment, with the parameters of the HMM reinitiated whenever a gap in the alignment has been introduced by the data filtering.

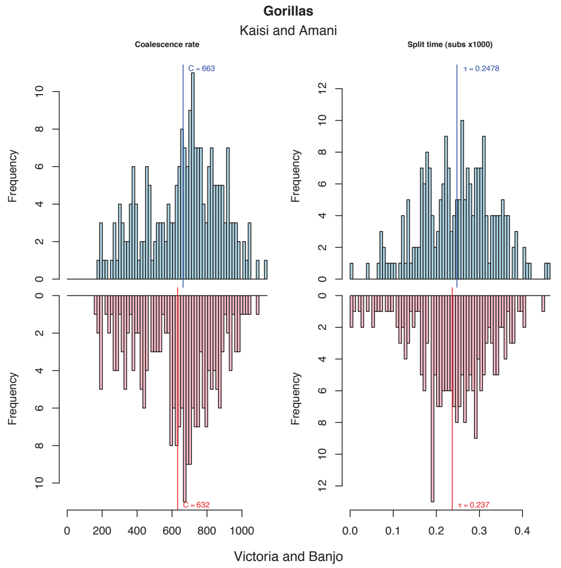
Parameter inference

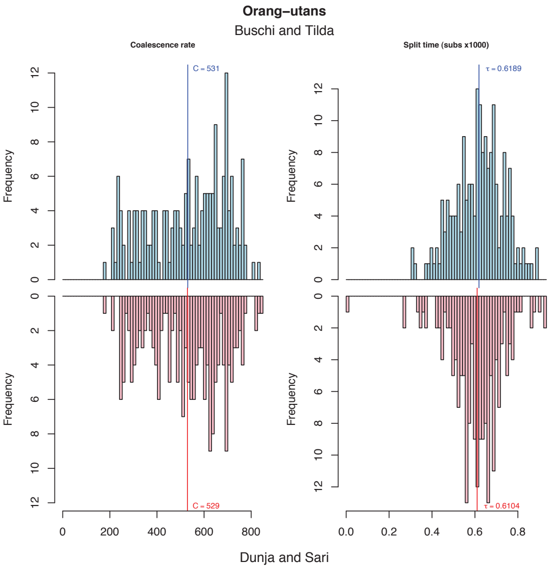
We applied the method[79] with three parameters: a coalescence rate C, a split time tau, and a recombination rate R, all measured in a time scale of substitutions per base pair. This means that the model assumes that the effective population size in the two extant species is the same as in the ancestral population, but as previously shown[81], this means that the model will estimate the ancestral coalescence rate rather than the extant rates, since this is the most important rate for the likelihood of the model.

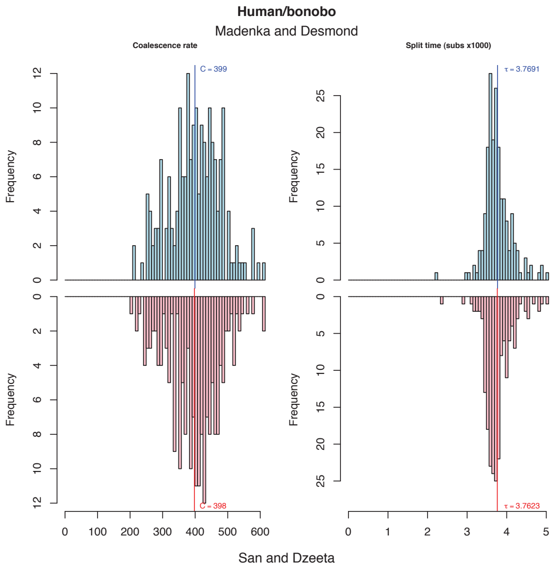
If the mutation rate per year, u, is known, and the generation time, g, is known, these parameters can be rescaled since tau/u is then the split time in years, 2/C = 4N u g, and R/(u g) is the recombination rate per base pair per generation.

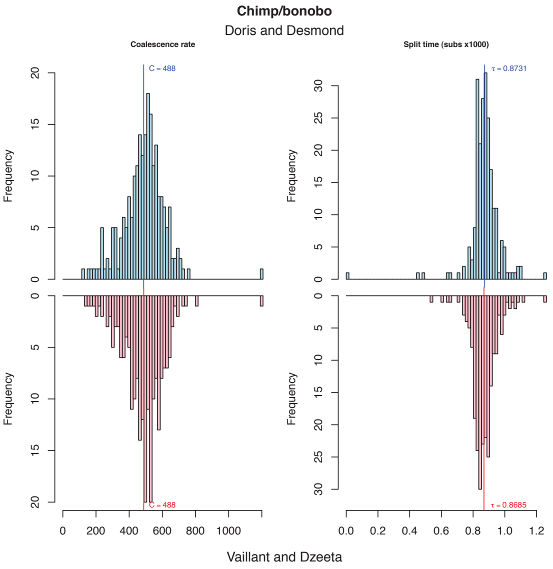
Results

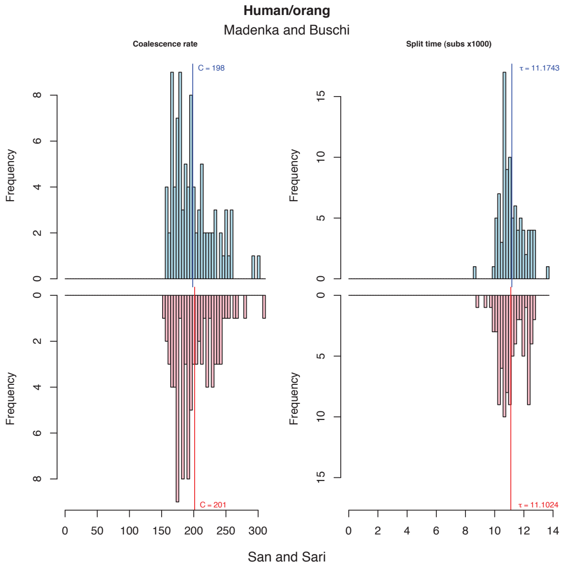
With the filtering described above, >95% of 10 Mbp fragments converge to reasonable values except for the two most ancient split events, i.e., between human and gorilla and between human and orangutan. In these cases we have removed obvious outliers and report on the remaining fragments.

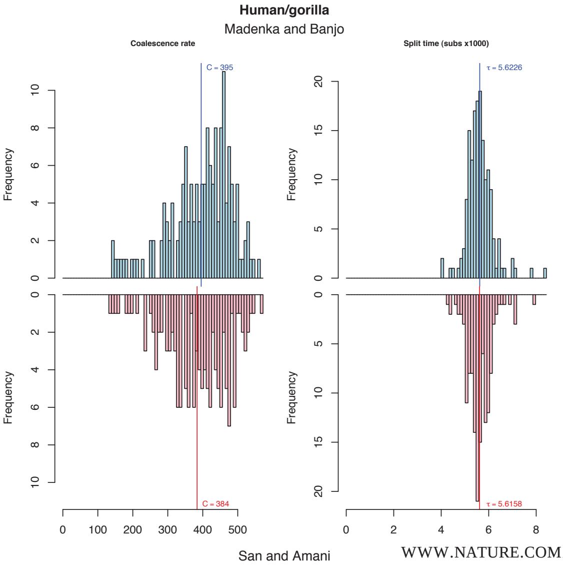

Gorillas
Kaisi and Amani
Coalescence rate
Split time (subs x1000)
C = 663
τ = 0.2478
Frequency
0
2
4
6
8
10
12
C = 632
τ = 0.237
0
200
400
600
800
1000
0.0
0.1
0.2
0.3
0.4
Victoria and Banjo


Orang–utans
Buschi and Tilda
Coalescence rate
Split time (subs x1000)
C = 531
τ = 0.6189
Frequency
0
2
4
6
8
10
12
C = 529
τ = 0.6104
0
200
400
600
800
0.0
0.2
0.4
0.6
0.8
Dunja and Sari


Human/bonobo
Madenka and Desmond
Coalescence rate
Split time (subs x1000)
C = 399
τ = 3.7691
Frequency
0
2
4
6
8
10
12
0
5
10
15
20
25
C = 398
τ = 3.7623
0
100
200
300
400
500
600
0
1
2
3
4
5
San and Dzeeta


Chimp/bonobo
Doris and Desmond
Coalescence rate
Split time (subs x1000)
C = 488
τ = 0.8731
Frequency
0
5
10
15
20
25
30
C = 488
τ = 0.8685
0
200
400
600
800
1000
0.0
0.2
0.4
0.6
0.8
1.0
1.2
Vaillant and Dzeeta


Human/orang
Madenka and Buschi
Coalescence rate
Split time (subs x1000)
C = 198
τ = 11.1743
Frequency
0
2
4
6
8
0
5
10
15
C = 201
τ = 11.1024
0
50
100
150
200
250
300
0
2
4
6
8
10
12
14
San and Sari


Human/gorilla
Madenka and Banjo
Coalescence rate
Split time (subs x1000)
C = 395
τ = 5.6226
Frequency
0
2
4
6
8
10
0
5
10
15
20
C = 384
τ = 5.6158
0
100
200
300
400
500
0
2
4
6
8
San and Amani

Suppl. Figure 13.7 – *Coalescence rates C=1/4Nu and split times (x 1e3) for different species comparisons, with two different sets of individuals compared in each case (blue and red). A. Eastern and Western gorilla, B. Bornean and Sumatran orangutan, C. Chimpanzee and bonobo, D. Human and bonobo, E. Human and gorilla, F. Human and orangutan.*

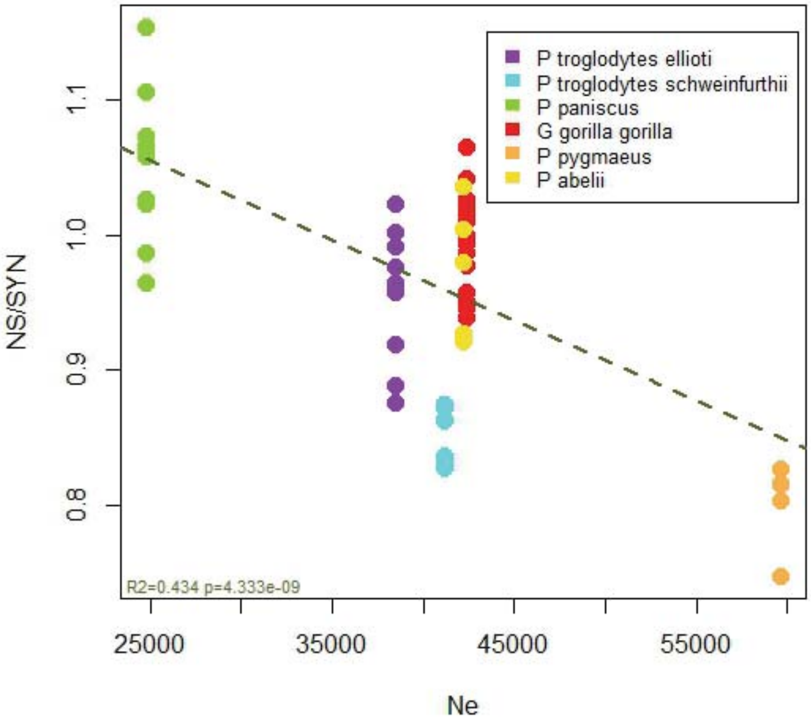
From the results of **Suppl. Figure 13.7**, we have calculated the mean (with standard error of the mean, SEM) for each comparison (**Suppl. Table 13.12.** These numbers can be translated into years and population sizes if we make assumptions about the generation time and the mutation rate per year. We show results assuming a generation time of 25 years (20 years for the comparison of gorillas) and three different mutation rates, either 1e-9 or 0.6e-9 per year or 1.1e-8 per generation.

						u=1e-9		u=0.6e-9		u g =1.1e10-8	
Species	Individuals	mean C	SEM	mean T * 1e3	SEM	Ne	Split kya	Ne	Split kya	Ne	Split kya
Gorillas	Victoria and Bajo	632.2	14.4	0.237	0.006	39547	237	65912	395	71904	431
	Kaisi and Amani	663.2	14.4	0.248	0.006	37695	248	62825	413	68537	451
Orangutans	Buschi and Tilda	530.7	12.1	0.619	0.008	37685	619	62808	1031	85647	1407
	Dunja and Sari	529.3	11.8	0.610	0.009	37783	610	62972	1017	85870	1387
Chimps and bonobos	Doris and Desmond	487.6	8.5	0.873	0.007	41014	873	68356	1455	93213	1984
	Vaillant and Dzeeta	487.9	8.3	0.869	0.005	40995	869	68326	1448	93171	1974
Human and chimps	Madenka and Doris	403.0	5.5	3.781	0.023	49628	3781	82713	6301	112791	8593
	San and Vaillant	398.9	5.5	3.749	0.024	50132	3749	83553	6249	113936	8521
Human and bonobos	Madenka and Desmond	399.2	5.6	3.769	0.025	50097	3769	83495	6282	113857	8566
	San and Dzeeta	397.6	5.6	3.762	0.024	50307	3762	83846	6270	114335	8551
Human and gorillas	Madenka and Bajo	395.5	7.2	5.623	0.043	50569	5623	84282	9371	114931	12779
	San and Amani	383.8	7.8	5.616	0.045	52111	5616	86851	9360	118433	12763
Human and orangutans	Madenka and Buschi	198.4	3.2	11.174	0.086	100802	11174	168003	18624	229096	25396
	San and Sari	201.3	3.4	11.102	0.092	99340	11102	165567	18504	225773	25233

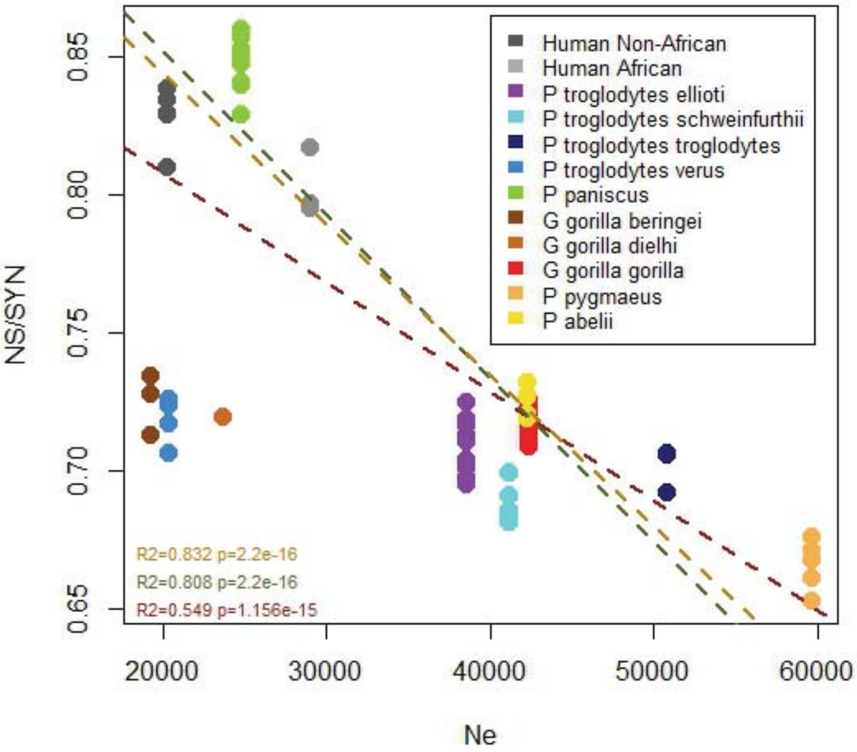
Suppl. Table 13.12 – *Coalescence rates C=1/4Nu and split times (x 1e3) for different species comparisons with standard errors of the mean (SEM). Effective population sizes of ancestral species and split times are shown for two different mutation rates per year and a generation time of 25 years (20 years when comparing the two gorillas). Mutation rate calibrations are 1e-9 and 0.6e-9 per year and 1.1e-8 per generation.*

Section 14: Nonsynonymous to synonymous variants

The proportion of rare nonsynonymous to rare synonymous variants per individual correlates with *Ne* (**Suppl. Figure 14.1**), as seen when all the variants are considered regardless of their frequency. This observation is due to the major efficiency of natural selection to remove detrimental variants at higher *Ne*[82]. It could be predicted that this effect should be stronger when only rare variants are considered, since rare variants are enriched for functional variants that negative selection keeps at low frequencies. However, the correlation described for the rare variants is not stronger to that when all frequency classes are considered together (**Suppl. Figures 14.1** and **14.2**). This weaker correlation may be originated by a greater dispersion of the rare nonsynonymous to synonymous variants ratios among the individuals in the same population because of the smaller number of mutations, as well as to the fact that common variants also contribute to the described differences across species.



Suppl. Figures 14.1 – *Effective population site versus the ratio of rare nonsynonymous to rare synonymous mutations per individual (MAF ≤ 0.1). Only populations with N ≥ 5 are included in this analysis.*



Suppl. Figures 14.2 – *Effective population site versus the ratio of nonsynonymous to synonymous mutations per individual. Linear regression lines and R^2 values are shown in brown when all the samples are considered, in green when only the same populations included in the adaptive selection test (Section 14) are considered, and in yellow when Pan troglodytes verus, Gorilla gorilla beringei and Gorilla gorilla diehli are excluded.*

References

1. Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222–6 (2012).

2. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–60 (2009).

3. DePristo, M. a *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics* **43**, 491–8 (2011).

4. Consortium, C. S. and A. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69–87 (2005).

5. Scally, A. *et al.* Insights into hominid evolution from the gorilla genome sequence. *Nature* **483**, 169–175 (2012).

6. Locke, D. P. *et al.* Comparative and demographic analysis of orang-utan genomes. *Nature* **469**, 529–533 (2011).

7. Mckenna, A. *et al.* The Genome Analysis Toolkit : A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**, 1297–1303 (2010).

8. Ning, Z., Cox, A. J. & Mullikin, J. C. SSAHA: a fast search method for large DNA databases. *Genome Research* **11**, 1725–1729 (2001).

9. Browning, B. L. & Yu, Z. Simultaneous Genotype Calling and Haplotype Phasing Improves Genotype Accuracy and Reduces False-Positive Associations for Genome-wide Association Studies. *The American Journal of Human Genetics* **85**, 847–861 (2009).

10. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics (Oxford, England)* **26**, 2867–73 (2010).

11. Bowden, R. *et al.* Genomic Tools for Evolution and Conservation in the Chimpanzee : Pan troglodytes ellioti Is a Genetically Distinct Population. *PLoS Genetics* **8**, 1–10 (2012).

12. Alkan, C. *et al.* Personalized copy number and segmental duplication maps using next-generation sequencing. *Nature genetics* **41**, 1061–1067 (2009).

13. Donmez, N. & Brudno, M. Hapsembler : An Assembler for Highly Polymorphic Genomes. *Research in Computational Molecular Biology* 38–52 (2011).doi:10.1007/978-3-642-20036-6_5

14. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403–410 (1990).

15. Katoh, K., Kuma, K., Toh, H. & Miyata, T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Research* **33**, 511–518 (2005).

16. Fischer, A. *et al.* Bonobos Fall within the Genomic Variation of Chimpanzees. *PLoS ONE* **6**, e21605 (2011).

17. Behar, D. M. *et al.* A "Copernican" Reassessment of the Human Mitochondrial DNA Tree from its Root. *The American Journal of Human Genetics* **90**, 675–684 (2012).

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

19. Tang, H., Peng, J., Wang, P. & Risch, N. J. Estimation of individual admixture: Analytical and study design considerations. *Genetic epidemiology* **28**, 289–301 (2005).

20. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS genetics* **2**, e190 (2006).

21. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research* **19**, 1655–1664 (2009).

22. Alexander, D. H. & Lange, K. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics* **12**, 246 (2011).

23. Price, A. L. *et al.* Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genetics* **5**, e1000519 (2009).

24. Auton, A. *et al.* A Fine-Scale Chimpanzee Genetic Map from Population Sequencing. *Science* **336**, 193–198 (2012).

25. Mcquillan, R. *et al.* Runs of Homozygosity in European Populations. *American journal of human genetics* **83**, 359–372 (2008).

26. Pemberton, T. J. *et al.* Genomic patterns of homozygosity in worldwide human populations. *American journal of human genetics* **91**, 275–92 (2012).

27. Liengola, H. J. and & I Post-Conflict Inventory of Kahuzi-Biega National Park,. *Gorilla Journal* **30**, 3–5 (2005).

28. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research* **38**, e164 (2010).

29. Wetterbom, A., Gyllensten, U., Cavelier, L. & Bergström, T. F. Genome-wide analysis of chimpanzee genes with premature termination codons. *BMC genomics* **10**, 56 (2009).

30. Prüfer, K. *et al.* The bonobo genome compared with the chimpanzee and human genomes. *Nature* 1–6 (2012).doi:10.1038/nature11128

31. Hahn, Y. & Lee, B. Human-specific nonsense mutations identified by genome sequence comparisons. *Human genetics* **119**, 169–78 (2006).

32. Hahn, Y. & Lee, B. Identification of nine human-specific frameshift mutations by comparative analysis of the human and the chimpanzee genome sequences. *Bioinformatics* **21 Suppl 1**, i186–94 (2005).

33. Hormozdiari, F., Alkan, C., Eichler, E. E. & Sahinalp, S. C. Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Research* **19**, 1270–1278 (2009).

34. Ventura, M. *et al.* Gorilla genome structural variation reveals evolutionary parallelisms with chimpanzee. *Genome Research* **21**, 1640–1649 (2011).

35. Olson, M. V When less is more: gene loss as an engine of evolutionary change. *American Journal of Human Genetics* **64**, 18–23 (1999).

36. Wang, X., Grus, W. E. & Zhang, J. Gene Losses during Human Origins. *PLoS Biology* **4**, e52 (2006).

37. Marques-Bonet, T. *et al.* A burst of segmental duplications in the genome of the African great ape ancestor. *Nature* **457**, 877–81 (2009).

38. Brawand, D. *et al.* The evolution of gene expression levels in mammalian organs. *Nature* **478**, 343–8 (2011).

39. Hammer, M. F. *et al.* The ratio of human X chromosome to autosome diversity is positively correlated with genetic distance from genes. *Nature genetics* **42**, 830–1 (2010).

40. Gottipati, S., Arbiza, L., Siepel, A., Clark, A. G. & Keinan, A. Analyses of X-linked and autosomal genetic variation in population-scale whole genome sequencing. *Nature genetics* **43**, 741–3 (2011).

41. Hernandez, R. D. *et al.* Classic Selective Sweeps Were Rare in Recent Human Evolution. *Science* **331**, 920–924 (2011).

42. Pool, J. E. & Nielsen, R. Population size changes reshape genomic patterns of diversity. *Evolution: International Journal of Organic Evolution* **61**, 3001–3006 (2007).

43. Keinan, A. & Reich, D. Can a sex-biased human demography account for the reduced effective population size of chromosome X in non-Africans? *Molecular Biology and Evolution* **27**, 2312–2321 (2010).

44. Wakeley, J. Distinguishing migration from isolation using the variance of pairwise differences. *Theoretical Population Biology* **49**, 369–386 (1996).

45. Nielsen, R. & Wakeley, J. Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics* **158**, 885–96 (2001).

46. Hey, J. The divergence of chimpanzee species and subspecies as revealed in multipopulation isolation-with-migration analyses. *Molecular biology and evolution* **27**, 921–33 (2010).

47. Wegmann, D., Leuenberger, C. & Excoffier, L. Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. *Genetics* **182**, 1207–18 (2009).

48. Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H. & Bustamante, C. D. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS genetics* **5**, e1000695 (2009).

49. Lukic, S. & Hey, J. Demographic inference using spectral methods on SNP data, with an analysis of the human out-of-Africa expansion. *Genetics* **192**, 619–39 (2012).

50. Aeschbacher, S., Futschik, A. & Beaumont, M. A. Approximate Bayesian computation for modular inference problems with many parameters: the example of migration rates. *Molecular ecology* **22**, 987–1002 (2013).

51. Durand, E. Y., Patterson, N., Reich, D. & Slatkin, M. Testing for ancient admixture between closely related populations. *Molecular biology and evolution* **28**, 2239–52 (2011).

52. Green, R. R. E. *et al.* A Draft Sequence of the Neandertal Genome. *Science (New York, N.Y.)* **328**, 710–722 (2010).

53. Eriksson, A. & Manica, A. Effect of ancient population structure on the degree of polymorphism shared between modern human populations and ancient hominins. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 13956–60 (2012).

54. Pickrell, J. K. & Pritchard, J. K. Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. *PLoS Genetics* **8**, e1002967 (2012).

55. Won, Y.-J. & Hey, J. Divergence population genetics of chimpanzees. *Molecular Biology and Evolution* **22**, 297–307 (2005).

56. Becquet, C. & Przeworski, M. A new approach to estimate parameters of speciation models with application to apes. *Genome Research* **17**, 1505–1519 (2007).

57. Wegmann, D. & Excoffier, L. Bayesian inference of the demographic history of chimpanzees. *Molecular Biology and Evolution* **27**, 1425–1435 (2010).

58. Gonder, M. K. *et al.* Evidence from Cameroon reveals differences in the genetic structure and histories of chimpanzee populations. *PNAS* **108**, 4766–4771 (2011).

59. Degnan, J. H. & Rosenberg, N. A. Discordance of species trees with their most likely gene trees. *PLoS genetics* **2**, e68 (2006).

60. Beaumont, M. A., Zhang, W. & Balding, D. J. Approximate Bayesian computation in population genetics. *Genetics* **162**, 2025–2035 (2002).

61. Gronau, I., Hubisz, M. J., Gulko, B., Danko, C. G. & Siepel, A. Bayesian inference of ancient human demography from individual genome sequences. *Nature genetics* **43**, 1031–4 (2011).

62. Achaz, G. Testing for neutrality in samples with sequencing errors. *Genetics* **179**, 1409–24 (2008).

63. Roach, J. C. *et al.* Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science (New York, N.Y.)* **328**, 636–9 (2010).

64. Kong, A. *et al.* Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**, 471–5 (2012).

65. Wakeley, J. & Hey, J. Estimating ancestral population parameters. *Genetics* **145**, 847–55 (1997).

66. Rannala, B. & Yang, Z. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics* **164**, 1645–56 (2003).

67. Wegmann, D., Leuenberger, C., Neuenschwander, S. & Excoffier, L. ABCtoolbox: a versatile toolkit for approximate Bayesian computations. *BMC bioinformatics* **11**, 116 (2010).

68. Leuenberger, C. & Wegmann, D. Bayesian computation and model selection without likelihoods. *Genetics* **184**, 243–52 (2010).

69. Robert, C. P., Cornuet, J.-M., Marin, J.-M. & Pillai, N. S. Lack of confidence in approximate Bayesian computation model choice. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 15112–7 (2011).

70. Fagundes, N. J. R. *et al.* Statistical evaluation of alternative models of human evolution. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 17614–9 (2007).

71. Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A. & Feldman, M. W. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular biology and evolution* **16**, 1791–8 (1999).

72. Veeramah, K. R. *et al.* An early divergence of KhoeSan ancestors from those of other modern humans is supported by an ABC-based analysis of autosomal resequencing data. *Molecular biology and evolution* **29**, 617–30 (2012).

73. Hudson, R. R. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics (Oxford, England)* **18**, 337–8 (2002).

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

75. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–6 (2011).

76. Scally, A. & Durbin, R. Revising the human mutation rate: implications for understanding human evolution. *Nature reviews. Genetics* **13**, 745–53 (2012).

77. Caswell, J. L. *et al.* Analysis of chimpanzee history based on genome sequence alignments. *PLoS genetics* **4**, e1000057 (2008).

78. Thalmann, O., Fischer, A., Lankester, F., Pääbo, S. & Vigilant, L. The complex evolutionary history of gorillas: insights from genomic data. *Molecular Biology and Evolution* **24**, 146–158 (2007).

79. Mailund, T. *et al.* A New Isolation with Migration Model along Complete Genomes Infers Very Different Divergence Processes among Closely Related Great Ape Species. *PLoS genetics* **8**, e1003125 (2012).

80. Hobolth, A., Christensen, O. F., Mailund, T. & Schierup, M. H. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. *PLoS genetics* **3**, e7 (2007).

81. Mailund, T., Dutheil, J. Y., Hobolth, A., Lunter, G. & Schierup, M. H. Estimating Divergence Time and Ancestral Effective Population Size of Bornean and Sumatran Orangutan Subspecies Using a Coalescent Hidden Markov Model. *PLoS Genetics* **7**, 15 (2011).

82. Kimura, M., Maruyama, T. & Crow, J. F. the Mutation Load in Small Populations. *Genetics* **48**, 1303–12 (1963).