www.sciencemag.org/content/348/6231/242/suppl/DC1

Supplementary Materials for

Mountain gorilla genomes reveal the impact of long-term population decline and inbreeding

Yali Xue, Javier Prado-Martinez, Peter H. Sudmant, Vagheesh Narasimhan, Qasim Ayub, Michal Szpak, Peter Frandsen, Yuan Chen, Bryndis Yngvadottir, David N. Cooper, Marc de Manuel, Jessica Hernandez-Rodriguez, Irene Lobon, Hans R. Siegismund, Luca Pagani, Michael A. Quail, Christina Hvilsom, Antoine Mudakikwa, Evan E. Eichler, Michael R. Cranfield, Tomas Marques-Bonet, Chris Tyler-Smith,* Aylwyn Scally*

*Corresponding author. E-mail: cts@sanger.ac.uk (C.T.-S.); aos21@cam.ac.uk (A.S.)

Published 10 April 2015, *Science* **348**, 242 (2015)
DOI: 10.1126/science.aaa3952

This PDF file includes:

Other supplementary material for this manuscript includes the following:

Data files S1 and S2

Materials and Methods

Samples and sequencing

Details of the samples collected and sequenced in this project are given in table S1. All sequenced samples were collected by the Mountain Gorilla Veterinary Project (www.gorilladoctors.org) from wild-born individuals, and were transferred in compliance with current legislation (CITES). DNA was extracted from blood in each case, and sequencing was carried out on Illumina HiSeq 2000 sequencing machines using standard library preparation protocols.

The mountain gorillas sampled here include two individuals originating on the DRC side of the Virunga mountain range, and five individuals from three different groups on the Rwandan side (Pablo, Titus and Umubano) whose home ranges span the region between Mount Karisimbi and Mount Bisoke (fig. S1). This region is home to over a dozen gorilla groups estimated to comprise in total about one third of gorilla individuals in the Virunga region as a whole (*3*). The DRC individuals were rescued separately as infants from poachers within the DRC; their precise origins within the *Parc National des Virunga* are unknown. Our analyses revealed no clear signals correlating with these distinctions or groupings, and in particular no overall difference between the DRC and Rwandan samples.

These considerations suggest that the samples we have sequenced plausibly represent genetic variation and ancestry in the Virunga population at large. However it remains the case that further sampling could change the picture presented here to some degree, for example with respect to relatedness and chromosomal sequence sharing. We note also that a substantial fraction (approximately 40%) of surviving mountain gorillas inhabit the Bwindi Impenetrable Forest region in Uganda, 30 km north of Virunga and were not sampled in this study. Thus further sampling elsewhere in Virunga and from the Bwindi population would be valuable to complement the findings reported here.

Additional data: In addition, previously published whole-genome data for three further eastern lowland gorillas, 27 western lowland gorillas and one Cross River gorilla were used in some analyses. These samples are listed in table S2; further details are given in Prado-Martinez *et al*. (2013) (*12*). Data for some further samples were used in mtDNA and Y-chromosomal analyses; details in Supplementary sections 6 and 7.

Alignment and variant calling

Except where otherwise stated, the analyses in this study were based on an alignment of sequence data from all samples listed in tables S1 and S2 to the gorilla genome reference sequence gorGor3.1 (http://www.ensembl.org/Gorilla_gorilla). Alignment was carried out using BWA-MEM v0.7.5a-r405 (*26*) with default parameters. After PCR duplicate removal using PICARD v1.91(1451) (http://sourceforge.net/projects/picard), single nucleotide polymorphisms (SNPs) were called using FREEBAYES v0.9.14 (*27*) using the following parameters: `--standard-filters --no-population-priors -p 2 --report-genotype-likelihood-max --standard-gls --prob-contamination 0.05`

We then constructed a mappability mask, to identify positions in the reference genome where variants could not be confidently called. We began by applying the mappability module in GEM (*28*), which uses a *k*-mer approach to identify genomic regions that are duplicated and therefore likely to be problematic for SNP calling. We retrieved duplications with up to 4 mismatches. Then we identified the regions of the genome where all samples had between 3-fold and 100-fold depth of aligned read coverage, and excluded any site outside the intersection of all such regions. Conservatively, we excluded polymorphic sites which were not biallelic or where QUAL < 30. Finally, to avoid false positive heterozygous calls due to mapping errors, we excluded any site where more than 80% of the samples had non-reference alleles in a heterozygous state. After excluding these filtered positions the total callable genome length was 1,649,453,084 sites.

Larger copy-number variants were investigated using a separate analysis pipeline (Supplementary section 5).

Evaluation of SNP error rates. The presence of long homozygous stretches in sampled individuals (Supplementary section 9) enables us to estimate the false positive rate in our heterozygous variant calls. Within each individual a homozygous region longer than 10 Mb is expected to correspond to a region in which the two chromosomal copies are separated by a small number of meioses, and are thus identical by descent (IBD) over most or all of their length. Heterozygous positions present in such regions may either be due to errors in sequencing and variant calling or to genuine *de novo* mutations occurring on the lineage separating the two chromosomes. Since we expect very few *de novo* mutations in each IBD region, by examining long homozygous regions across all individuals we obtain false positive error rate estimates of 0.00090% per base pair and 0.082% per called variant.

Principal component analysis

A principal component analysis (PCA) of genetic variation was carried out to detect inherent structure within and between the populations sampled. Three separate analyses were performed: one including all samples from both gorilla species; one focusing only on samples from the eastern species (including mountain gorillas) and one on the western species.

In each case, all variants segregating within the populations concerned were drawn from the filtered VCF file (Supplementary section 2) and used as initial input. Additionally, individuals found to be very closely related to another on the basis of shared sequence analysis (Supplementary section 10) were also removed. For the two within-species analyses, the resulting variant sites were pruned to remove all pairs in linkage disequilibrium (LD) with an R^2 correlation coefficient of 0.5 or greater. In the eastern species PCA this stringent threshold left 21,921 sites, and in the western species 51,393 sites. The LD pruning procedure was not carried out for the all-sample PCA across both species, to avoid excluding sites corresponding to fixed differences between the species; as a result a much larger number of sites, 11,780,090, were included. Genetic variation at the remaining sites was passed to the principal component decomposition step, and figures 1A, 1B and S1 show samples projected onto the first two principal

components for each of the three analyses. Both the LD pruning and principal component decomposition calculations were carried out using PLINK v.1.9 (https://www.cog-genomics.org/plink2).

In the western species, the projection in the first two components is suggestive of substructure within the western lowland population, and correlates with what limited information exists on the geographical origins of these samples (table S2).

Admixture analysis

Another view of structure within the populations sampled is provided by the analysis of shared genetic components implemented in the ADMIXTURE program (*29*). To carry out this analysis we began with one million sites sampled randomly from the filtered VCF file (Supplementary section 2) and reduced this number by applying an LD pruning filter using PLINK with a threshold R^2 correlation of 0.5. Some individuals were found to be in closely related pairs in an analysis of relatedness (Supplementary Section 10), and to avoid bias one of each such pair was excluded here (from the mountain gorillas we excluded Maisha and Umurimo). The resulting 439,586 SNPs were passed to ADMIXTURE, and the output for several values of the clustering parameter k is shown in fig. S3, with cross-validation (CV) errors for each k shown in fig. S4.

Overall this analysis agrees very well with the results of PCA (Supplementary section 3). CV error is lowest for $k = 2$, reflecting the subdivision into eastern and western gorilla species. At $k = 3$ some substructure within the western gorilla population is seen. At $k = 4$ the separation between eastern lowland and mountain gorillas appears, and at $k = 5$ the substructure within the eastern lowland population seen in fig. 1B is replicated. Higher values of k increase the CV error and introduce further structure into the western species. (Interestingly, unlike in fig. S2 the Cross River individual Nyango is not clearly separated from other western gorillas at any value of k.)

Copy number variation

To assess copy number variation among gorilla species and populations we employed a read-depth based approach, described in previous studies (*30*, *31*). Reads from each genome were divided into their 36bp constituents and mapped to the human reference genome (Build 37, hg19) and to the gorilla genome (gorGor3.1) using the mrsFAST aligner (*32*). The human genome was used as a reference to assess putative genic events using GENCODE v.19 gene annotations. Read depth profiles were corrected for local GC content and copy number was estimated in overlapping 500bp windows of unmasked sequence, slid across the genome at 100bp intervals. In addition to the 7 mountain gorillas and 6 eastern lowland gorillas sequenced in this study (table S1), we also analyzed 11 previously published western lowland gorillas (*12*).

We began by performing a quality control analysis of the 24 genomes analyzed, particularly focusing on the relationship between read-depth and GC-content. We assessed the genome-wide proportion of 500bp windows estimated as diploid as a function of GC-content in each individual (fig. S5). From this analysis, 5 western lowland gorillas and a single mountain gorilla (Turimaso) were identified as exhibiting aberrant GC-associated read-depth profiles, and were discarded from further analysis. The

remaining 18 individuals (6 mountain, 6 eastern lowland and 6 western lowland gorillas, table S4) were then assessed for copy number variation.

To analyze lineage-specific fixed events and their potential impact on genes we first analyzed the read-depth profiles of individuals mapped to the human reference genome. We classified lineage-specific deletions as fixed homozygous losses along a specific branch, lineage-specific duplications as fixed (copy 4) duplications along a specific branch while other branches remain diploid, and expansions as increases in copy number along a particular branch though all branches exhibit higher copy number than 2 (Auxiliary file S1).

As expected, the number of fixed events broadly reflected the overall topology of the phylogeny (table S5); however, eastern lowland and mountain gorillas exhibit a marked excess of shared fixed deletion events in addition to an excess of lineage-specific deletions compared to western lowland gorillas. The shared deletion events are potentially the result of a bottleneck prior to the divergence of the two populations while the lineage-specific events suggest that there has been prolonged independent lack of diversity in both eastern lowland and mountain gorillas.

We next assessed these fixed events for their intersection with putative functional elements utilizing GENCODE annotations of coding and non-coding transcripts (http://www.gencodegenes.org/) (table S6). Again, while the overall tree topology is supported, genic deletions in particular are more prevalent in eastern lowland and mountain gorillas. Along the shared mountain and eastern lowland gorilla lineage, fixed losses are observed in exon 5 of *CEACAM1* – a cell adhesion gene of the carcinoembryonic antigen family, in two linc-RNAs and in the processed transcript *LINC00343*. Mountain gorillas exhibit lineage-specific loss of the olfactory receptor *OR7A5*, miRNA *AC107020.1* and an antisense transcript *RP11-231L11.1*. Mountain gorillas also exhibit a lineage-specific expansion of exon10 of the *HERC2* gene (fig. S6), mutations in which have been shown to affect nearby OCA2 expression in humans and are associated with hair and eye color (*33*).

We next sought to assess the CNV diversity of gorilla species by identifying polymorphic structural variants. We specifically assayed for bi-allelic deletions (copy number 0,1,2) segregating in gorilla genomes using the gorGor3.1 read-depth profiles and digital comparative genomic hybridization (dCGH) to detect CNVs among individuals (*31*). Briefly, dCGH involves comparing the windowed copy number estimates pair-wise between each of the 18 assessed individuals, by calculating the $\log_2$-ratio between their copy number estimates. An algorithm based on scale space filtering is then employed to identify putative CNVs which are genotyped using an EM-fitted Gaussian mixture model. Finally, CNV calls were filtered to remove sites with poorly distinguishable genotypes resulting in a conservative, but confident dataset.

We identified 1083 CNVs among the 18 gorillas assessed, 384 of which were specific to western lowland gorillas, compared to 46 and 51 specific to eastern lowland and mountain gorillas respectively. Treating these CNVs as bi-allelic markers we were able to construct a tree of the relationships between assessed individuals (fig. S7) clearly distinguishing eastern lowland and mountain gorillas as a single clade from western lowland gorillas.

Similarly, these genotypes can be used to perform principal components analysis (fig. S8) which yielded a similar clustering with PC1 (32.8% of variance) separating

eastern lowland and mountain gorillas from western lowland and PC2 (16.0% of variance) distinguishing eastern lowland and mountain gorillas.

Finally, we sought to assess the relative levels of CNV diversity among different gorilla populations by comparing the number of homozygous and heterozygous CNVs identified per individual among the different populations (fig. S9). Strikingly, western lowland gorillas exhibited on average 1.91-fold more heterozygous sites than eastern lowland and mountain gorillas ($p = 8.858e-7$; t-test). Conversely, mountain and eastern lowland gorillas exhibit 2.8-fold more homozygous CNVs than western lowland gorillas ($p = 1.185e-6$; t-test). No significant difference was observed in the number of heterozygous or homozygous sites between eastern lowland and mountain gorillas.

mtDNA analysis

Previous studies of mountain gorilla mtDNA have featured noninvasive (e.g. fecal) sample collections, including up to more than 100 samples (*10*, *34*). However they were restricted to analyses of small mtDNA pieces, as DNA extracted from environmental samples is highly degraded and only short fragments can be amplified reliably. Moreover, the variety of translocated sequences of mtDNA in the nuclear genome (NUMT) present in the gorilla nuclear genome, highly similar to organellar mtDNA, makes analysis using standard approaches very difficult (*35*). Since the current study involved a substantial collection of blood samples taken from eastern gorillas including both mountain and lowland subspecies, it allowed amplification and sequencing of the complete mtDNA using a long-range PCR approach, avoiding amplification of nuclear insertions.

Samples and sequencing: Thirty gorilla DNA samples were used, including 14 western lowland gorillas, 8 eastern mountain gorillas and 8 eastern lowland gorillas (table S10). The western lowland gorilla DNA samples were obtained from individuals previously sequenced in Scally *et al.* (2013) (*36*). Murphy's DNA was extracted from lymphoblast cell line, whereas all other samples were derived from blood. Note that the 8 mountain gorillas samples included one individual, Ishema (female), who was not selected for whole genome sequencing, being the mother of Imfura (male). We include this sample here as a check on the analysis and sequence reconstruction process, since we expect Ishema's mtDNA sequence to match that of Imfura (apart from any *de novo* mutations which may be present).

Primers: Universal primers designed to work across many taxa (*35*) were used to amplify two long-range overlapping segments (A and B, around 9.2 kb and 10 kb respectively) that together cover the entire mitochondrial DNA molecule and avoid amplification of NUMTs. Table S9 lists the characteristics of the primer pairs. Sequences of the conserved primers were compared with the reference mtDNA sequence of western lowland gorilla (GenBank accession number: NC_011120.1). Four mismatches were found for the 12So universal primer and one difference in the Cytbf primer sequence when compared to the reference (table S8). To improve amplification efficiency, the 12So primer was redesigned to match the reference sequence (see Results, below). Two alternative variants of the COII28for primer (shifted 1bp down-/upstream the sequence) were tested to avoid potential mismatches at the 3′–end of a primer and to select an optimally performing primer combination. From the screened primer pairs, the

combinations COII28for-GgmtB02 and GgmtB01F-GgmtB01R tended to work best for segment B amplification.
Another three gorilla-specific primer pairs were designed to amplify ~1000 bp of the control region in 3 short (~450 bp) overlapping fragments for validation purposes (table S9). They were designed using Primer3 online tool (release 1.1.4) and the reference sequence (GenBank accession number: NC_011120.1).

DNA amplification: Two separate long-range PCR amplifications were performed on each DNA sample. The final amplification reaction contained 1x High Fidelity PCR Buffer (Invitrogen), 2 mM MgSO4, 0.2 mM each dNTP, 0.45 U Platinum® Taq DNA Polymerase High Fidelity (Invitrogen), 0.2 μM each of the forward and reverse primer (table S7) and approximately 10ng template DNA in a 20 μl total volume. A master mix of reagents was prepared to minimize pipetting error. DNA was amplified with initial denaturation at 94°C for 30 s followed by 35 cycles of denaturation at 94°C for 30 s, annealing at 60°C for 30 s and elongation at 68°C for 11 min (following Thalmann et al. (2004) (35)). After cycling, reactions were maintained at 4°C. Short control region fragments were amplified from long-range segments (A or B) using ~0.25 μl of the long-range PCR product and above described conditions with the exception of the elongation step (72°C for 30 sec). Negative controls containing only PCR reagents were also included to control for potential contamination. PCR blanks never yielded products. In order to check the fragment size and quality of amplification, 4 μl of the PCR products were analyzed by 0.8% agarose gel (containing ethidium bromide), electrophoresis lasting approximately 1.5 hour (80 Volts). DNA ladder (HyperLadder I (Bioline) and/or HighMass DNA Ladder (Invitrogen) for long-range products or Low DNA Mass Ladder (Life Technologies) for D-loop fragments) were run alongside the samples. After separation, bands were visualized under UV transillumination (GeneGenius Bio Imaging System and GeneSnap software (Syngene))

Library preparation and sequencing: Around 200-500 ng of each long-range fragment A and B were pooled (equimolar quantities) and purified using Agencourt AMPure XP purification. SPRI beads were added to the sample in the ratio of 1.5:1, then the subsequent steps were performed according to the manufacturer's instructions. Libraries were prepared according to methods described in (*37*).

DNA libraries were sequenced using an Illumina MiSeq sequencer with 150 bp paired-end reads. Each eastern gorilla sample was sequenced twice for validation purposes. Furthermore, ends of both A and B segments were sequenced using capillary electrophoresis and reverse primers (same as for amplification) for a few samples to validate SNPs in the coding region. Similarly, amplified control region fragments were purified and sequenced two times for each individual using Sanger capillary sequencing to validate substitutions in the hypervariable region.

Sequence assembly and validation: Next-generation paired-end reads were assembled *de novo* using Velvet v.1.1.06 (*38*). Assembly was performed three times for each sample using random subset of paired reads (~10%-20%) and maximum k-mer size (99) due to high coverage, and usually resulted in 1 or 2 overlapping contigs covering the whole mitochondrial molecule. The assembled sequences were aligned using SEQMAN v.11.2.1 and the consensus sequence resulting from 3 or 6 assemblies for each individual

(for western and eastern gorillas, respectively, as eastern gorilla samples were sequenced twice) were taken as the final sequence. Furthermore, reads were also mapped to the newly generated eastern gorilla genome sequence using BWA software and SNPs were called using samtools mpileup (http://samtools.sourceforge.net). Calling results were congruent with *de novo* assembly results.

To further validate the sequences obtained, we compared our MiSeq results to the capillary sequencing reads. Around 800–1000 bp of control region were sequenced twice from each individual using capillary sequencing. In addition, around 500 bp of *COX2* and 200 bp of *CYTB* coding regions were similarly sequenced for 6 samples (Itebero, Mukisi, Maisha, Zirikana, Fubu, Kwan; see table S10). Electropherograms were aligned to *de novo* assemblies using MEGA 5.10 (*39*). Capillary sequencing results perfectly matched *de novo* assemblies after exclusion of ambiguous single base pair insertions and deletions around homopolymer stretches (which were removed from downstream analyses). All polymorphic positions covered by capillary sequencing were confirmed, including a 19-nucleotide indel in the control region differentiating between western and eastern gorillas. Finally, our validation check included comparison of mother-son sequences (Ishema-Imfura) obtained in the study, which yielded matching haplotypes as expected.

Sequence analysis: Mitochondrial SNPs and indels across all gorilla individuals sequenced in this study and one external sequence (GenBank; table S11) were reported from the multiple alignment of whole mtDNA sequences using SeqMan Pro v. 11.2.1. Each variable site was annotated with its type (SNP or indel), location in a gene, and consequences for protein sequence according to the reference (GenBank accession number NC_011120.1) (Auxiliary file 1).

After filtering out variation in homopolymeric stretches, there were 820 variable sites within the set of 27 gorilla mtDNA sequences (26 sequenced here plus the reference sequence). We found an extremely low level of genetic diversity in mountain gorillas, with only 3 distinct haplotypes differing by only 1-3 mutations in the entire sample. By contrast, there were 13 polymorphic sites in the eastern lowland subspecies and 100 SNPs differentiating mountain and lowland eastern gorillas. Nine non-synonymous mutations were exclusively fixed in mountain gorillas (Auxiliary file 1: 1 in *ND2*; 1 in *COX2*; 2 in *ND4*; 1 in *ND5* and 4 in *CYTB*), while eastern lowland gorillas exhibited six substitutions of this class (1 in *COX2*, 1 in *ATP8*, 2 in *ATP6*, 1 in *ND4* and 1 in *ND5*). Genetic variation was higher in western lowland gorillas than in the other subspecies (257 variable sites), consistent with published literature (*34*, *40*, *41*).

Phylogenetic analysis: Phylogenetic calculations were based on complete mtDNA sequences, excluding homopolymeric sites, performed in MEGA 5.10 and BEAST v. 1.8.0 (http://beast.bio.ed.ac.uk, (*42*)). Two additional sequences external to this project were also included in the analysis, namely the human Revised Cambridge Reference Sequence and a western lowland gorilla reference sequence (table S11).

We inferred an ultrametric tree using a strict clock model of rate variation among branches and a constant population size coalescent as a tree prior. A General Time Reversible sequence substitution model with 4 gamma distributed rate categories (GTR+G) proved to be the best fit to our data according to the corrected Akaike Information Criterion (AICc as implemented in MEGA5). Two independent MCMC runs

were performed with 30,000,000 iterations each, sampling every 1,000 steps. Both independent runs were combined using LOGCOMBINER v.1.8.0 (http://beast.bio.ed.ac.uk/logcombiner) to increase the effective sample size of the analysis, and the first 6,000,000 iterations (6000 trees) were discarded as burn-in on each run.

The resulting tree topology (fig. S10, also shown in fig. 1E) shows the deepest separation between western and eastern gorillas, in agreement with previous studies (*9*, *10*, *34*, *40*) and with the accepted species taxonomy for *Gorilla*. (Note however that the genetic tree at a single locus need not match the species phylogeny, due to phenomena such as incomplete lineage sorting and genetic admixture.) We note also that, as expected, the mother-son pairing Ishema and Imfura are grouped together in this tree.

Additional mtDNA reconstruction from whole genome sequencing data: To extend our mtDNA phylogeny we also reconstructed mitochondrial genomes from whole genome sequencing (WGS) data comprising 31 *Gorilla gorilla gorilla*, 9 *Gorilla gorilla graueri* and 7 *Gorilla Gorilla berengei* individuals, and 1 *Gorilla gorilla diehli* individual. This dataset comprised all the samples listed in tables S1 and S2, plus the four additional western lowland gorillas listed in table S12, previously sequenced as part of the Great Ape Genome Project (http://biologiaevolutiva.org/greatape/).

In each case our reconstruction used a two-stage read capture and assembly approach, as follows:

Stage 1. For each sample we began by creating a sample-specific modified mtDNA sequence. This involved an iterative mapping, variant calling and reference correction procedure, starting from a western lowland gorilla mtDNA reference sequence (gi|195952353|ref|NC_011120.1) and ending when no variants were called. This strategy enhanced WGS read capture in highly variable regions due to the fact that the resulting modified reference sequence represented the sample mitochondrial genome more accurately than the initial reference.

Stage 2. In order to increase the number of captured reads at each end of the reference sequence and to account for the circularity of the mitochondrial genome, we also aligned WGS reads to a further-modified mtDNA sequence in which the start was shifted 8 kbp towards the middle of the genome. Taking the whole set of reads captured in this way, we removed low-quality read pairs in which at least one end pair had a median Phred quality score lower than 32. We then randomly sub-sampled reads to have at most 150-fold depth of coverage, and used HAPSEMBLER v.1.1 (http://compbio.cs.toronto.edu/hapsembler/hapsembler.html) to construct contigs from the remaining reads. This random sub-sampling and contig construction was performed 20 times on both the standard and the shifted-origin reference sample-specific reference sequences. In this way we compensated for the random representation of reads that could lead to problems such as the incorporation of nuclear mitochondrial sequences ‘numts’ into the assembly. Thus we created 40 mitochondrial assemblies per sample, and as our final mitochondrial sequence for each individual we took the consensus sequence of the 40 assemblies.

Y chromosome analysis

Since the gorilla reference sequence is based on DNA from a female individual (*16*), it does not contain a Y-chromosomal reference sequence. We therefore made use of the human reference sequence and developed an analysis based on unique homologous regions of the human Y chromosome.

To begin with we mapped sequence reads from all 7 mountain gorillas in table S1, 6 eastern lowland gorillas (Itebero (F), Tumani (F), Pinga (F), Ntabwoba (M), Dunia (F) and Serufuli (F)) and 2 western gorillas (Banjo (M) and Coco (F)) (thus including both males and females) to the human genome reference sequence (hs37d5, also used by the 1000 Genomes Project (http://www.1000genomes.org)).

We called consensus bases at all sites in this alignment, regardless of whether or not they were variable between these samples, using samtools mpileup. We then carried out a filtering step, excluding the following sites:

1. All sites called in one or more female samples with coverage of more than 10x – this excludes regions with autosomal or X chromosome homology

2. All sites outside the 'unique' region of human Y chromosome as defined in Wei *et al.* (2013) (*43*), which excludes regions duplicated on Y.

We ended up with 1,905,874 sites in total, of which 630 were SNPs variable within the set of five males. Of these, 457 sites represented eastern versus western species differences; 84 mountain vs eastern lowland differences; and just 2 sites differing within mountain gorillas (Auxiliary file 1).

As with the mtDNA analysis we inferred an ultrametric tree based on these data using BEAST, shown in fig. 1E.

mtDNA and chrY tree dates: Both the mtDNA and Y-chromosomal trees are presented in fig. 1 with internal node heights given in units of substitutions per base pair. In order to convert these to dates and obtain estimates of the TMRCA for these loci across the *Gorilla* genus as a whole and within particular subclades, we require estimates of the corresponding mtDNA and Y-chromosomal substitution rates in gorillas. Since these are not currently known, we use the rates in humans for these chromosomes as a best guess.

For mtDNA, using a substitution rate of 1.665×10^{-8} per nucleotide per year estimated for the whole human mtDNA molecule (*44*) we obtain a surprisingly ancient TMRCA for the whole *Gorilla* mtDNA tree of 1.34 Myr. By contrast, on Y using the directly measured Y mutation rate in humans of 1.0×10^{-9} bp^{-1} yr^{-1} (*45*) yields a corresponding TMRCA of 142 kyr (fig. S12). Although there is no strong expectation for these dates to coincide, such an ancient TMRCA for mtDNA is surprising given the demographic history inferred from autosomal genomic data, wherein the two gorilla species appear to have begun to diverge only a few hundred thousand years ago (Supplementary section 11). Some possible contributing factors for this discrepancy are:

a) The true mtDNA mutation rate may be higher in gorillas than in humans.
b) The mutation rate may be accurate for longer time scales but many of the differences counted in this analysis may be deleterious mutations and hence ultimately removed by purifying selection, an effect which has not been allowed for in this calculation (*44*).

c) Gene flow within and between gorilla populations during speciation may have been very strongly male biased (a phenomenon which has been observed in western lowland gorillas (*46*)), which may lead to an older apparent demographic timescale for mitochondrial ancestry.
d) Assembly and/or alignment errors in the mtDNA reconstruction have inflated branch lengths.

Linkage disequilibrium

We looked at genome-wide patterns of linkage disequilibrium (LD) in each of the sampled populations of gorilla. For comparison, we also included two human populations with contrasting demographic histories, namely the Yoruba population of Ibadan, Nigeria (YRI) and Utah Residents with Northern and Western European ancestry (CEU). To avoid biases in terms of sample size differences, we selected six unrelated individuals at random from each population (table S14).

Initial filtering was done using VCFTOOLS v. 0.1.12 (http://vcftools.sourceforge.net/), to exclude sites with minor allele frequencies below 0.15 and above 0.85 as well as any sites with missing data. Using PLINK v.2, each chromosome was thinned to include only a random 50% of all filtered sites, as some chromosomes otherwise would contain too many SNPs for calculation of LD at the desired depth. (Pilot analyses on short chromosomes showed that this did not have any noticeable influence on results.) After filtering there remained approximately 2 million SNPs per individual. LD was calculated as the mean R^2 correlation coefficient between pairs of SNPs, using the R BIOCONDUCTOR package snpMatrix. Calculations of LD as a function of distance along the chromosome were done with a pairwise calculation depth of 2000 SNPs and a maximum distance of 2 Mb, and averaged over 10 kbp windows. Each chromosome was analyzed individually and then the results combined to obtain an overall LD decay profile for each population.

As seen in fig. 2A, the three gorilla populations have different patterns of LD decay, with mountain gorillas showing a much slower rate of decay and more extensive LD over larger distances than the other populations analyzed. For example at a distance of 2Mb, R^2 in mountain gorillas is 0.41, compared to 0.22 in eastern lowland gorillas, 0.16 in western lowland gorillas, and 0.17 in African (YRI) and European (CEU) humans. A particularly striking example of a long LD block in mountain gorillas on chromosome 4 is shown in fig. S13.

By contrast, LD decays most rapidly in western lowland gorillas and is similar to that of the YRI population. These patterns reflect contrasting demographic histories in the populations studied, including population bottlenecks and inbreeding, which contribute to different rates of LD breakdown by recombination.

Within-individual regions of homozygosity

We used a hidden Markov model (HMM) to identify tracts of homozygosity (absence of inter-chromosomal variation) due to inbreeding within the sampled individuals. The HMM is applied to genetic variation data (in VCF format) for the population containing the sample, with positions in the chain corresponding to segregating sites in the population. The two hidden states represent homozygosity (H)

and non-homozygosity (N) within the sample, and genotypes within the sample are represented by RR for a homozygous site matching the reference, RA for a heterozygous site and AA for a homozygous non-reference site. Thus H tracts can only include RR and AA sites, whereas N tracts can include sites of any genotype.

Emission probabilities correspond to a Hardy-Weinberg model with inbreeding, and thus for any site i are determined by the minor allele frequency f_i at that site (excluding non-biallelic sites) and likelihoods of observed alignment data for possible genotypes in the sample (given by the variant calling algorithm described above):

$$P(D_i \mid X_i = H) = (1 - f_i)P(D_i \mid RR) + f_i P(D_i \mid AA)$$
$$P(D_i \mid X_i = N) = (1 - f_i)^2 P(D_i \mid RR) + 2f_i(1 - f_i)P(D_i \mid RA) + f_i^2 P(D_i \mid AA)$$

where D_i represents the data (*i.e.* aligned reads) and X_i is the homozygosity (hidden) state at site i.

Transition probabilities in the HMM incorporate the likelihood of a recombination event since the last site, given by $l_i\rho$, where l_i is the physical distance from the previous site and ρ is the recombination rate, (assumed constant here and equal to 1×10^{-8} per base pair). This is then multiplied by conditional probabilities for transitions between states given a recombination event:

$$P(X_{i+1} = N \mid X_i = N) = (1 - \rho l_{i+1})(1 - p_{NH})$$
$$P(X_{i+1} = H \mid X_i = N) = \rho l_{i+1} p_{NH}$$
$$P(X_{i+1} = H \mid X_i = H) = \rho l_{i+1} p_{HN}$$
$$P(X_{i+1} = H \mid X_i = H) = (1 - \rho l_{i+1})(1 - p_{HN})$$

The parameters p_{NH} and p_{HN} were learnt from the data using a Viterbi training scheme (also known as the segmental k-means algorithm), with the initial probabilities of being in the N or H state at the start of each chromosome set to be equal. The resulting state assignments given by the Viterbi sequence with the optimal parameters comprised our inferred homozygous and non-homozygous tracts (fig. S14).

The algorithms used here for homozygosity inference are implemented as part of the samtools suite of tools for sequence analysis:
https://github.com/samtools/bcftools/tree/RoH.

Between-individual sequence identity

In order to identify genetic sequences shared between individuals we need to resolve (phase) the distinct haplotypes within each sample. This was done using BEAGLE v.4 (http://faculty.washington.edu/browning/beagle/b4.html), applied to the all-sample set of filtered variant calls, which uses a HMM framework to infer phased haplotypes for each sample and regions shared identically by descent (IBD) between individuals. The sequence sharing between samples inferred by BEAGLE is shown in fig. S16.

To assess the accuracy of IBD inference we carried out the following imputation cross validation procedure across all samples. Taking each sample in turn, we constructed a phased sample-specific reference panel by excluding it from the set of phased haplotype calls generated for the whole data set. Then, for each of a set of 1000 variants on chromosome 20, we excluded that site from the sample data and imputed its genotype using the sample-specific reference panel (again with BEAGLE v.4). We then compared the set of imputed genotypes with the genotypes for each site in the sample data.

We can summarize the results of cross validation by calculating the correlation between imputed and known allele dosages across all samples within each population. Imputation accuracy was highest in mountain gorillas, with $R^2 = 0.97$, followed by eastern lowland gorillas ($R^2 = 0.96$), and lower in western gorillas ($R^2 = 0.85$).

As a further check on the robustness of these results, we repeated the phasing and IBD identification procedure using an alternative methodology as implemented in the program SLRP (*48*). Beagle has been found in previous studies to perform less well in highly inbred or isolated population samples (*48*, *49*). However, we found that levels of total shared IBD between individuals obtained by SLRP were very similar to those inferred by BEAGLE, such that the resulting tree topology was identical, and suggesting that the results reported here are not substantially affected by methodological biases.

Several individuals share a particularly high fraction of genomic sequence relative to the population background level, suggesting first- or second-degree kinship, notably Zirikana and Umurimo, and to a slightly lesser extent Kaboko and Maisha, and the eastern lowland gorillas Tumani and Dunia. These findings can be compared with what is known from field studies about the relatedness of the individuals sampled. In particular, Umurimo and Zirikana are known to be related as mother and child respectively. Both Kaboko and Maisha were rescued from poachers who had taken them from the DRC side of the Virunga massif, but at different times (Maisha in 2004 and Kaboko in 2007), and little else is known about their relatedness. Similarly, Tumani and Dunia were also rescued as infants from poachers, again on two separate occasions, and their original location is not known.

The mean amount of sharing between populations (summarized in table S16) is in good agreement with the species taxonomy and the results of PCA, except that interestingly the Cross River individual does not cluster apart from other western individuals in this analysis.

The sub-clustering of eastern lowland gorilla samples seen in Fig. 1D is also evident here, with individuals in one cluster having a 10% higher mean level of sequence sharing than those in the other.

Within mountain gorillas, mean sequence sharing between samples from the DRC and Rwandan side of the Virunga mountain chain is 33.0% as a percentage of chromosome length, similar to that between different groups on the Rwandan side (33.6%). (Sampled groups are listed in Table S1.) Mean sequence sharing across all samples excluding Umurimo-Zirikana (but including other within-group comparisons) is 33.9% (Table S16). This suggests that within-group comparisons are not substantially biasing our estimates of relatedness and shared genetic sequence in Virunga mountain gorillas, notwithstanding the possibility that further sampling could reveal hitherto undetected genetic structure.

Relatedness analysis: As a further independent analysis of relatedness between the sampled individuals, we used the KING program (*50*) which infers degrees of relatedness based on a pairwise comparison of SNP data for individuals. We found 5 pairs of mountain gorilla individuals that are closely related: one first-degree relationship, two second-degree and two third-degree relationships (fig. S17). Indeed all the mountain gorilla individuals apart from Turimaso show some degree of kinship with at least one other individual By comparison in eastern lowland gorillas, only two pairs show close

relatedness, at second and third degree. These results are also in good agreement with the IBD analysis presented above, with all inferred first- and second-degree relationships corresponding to individuals clustered adjacently by IBD in fig. S16.

Inference of ancestral effective population size (PSMC)

The Pairwise Sequential Markovian Coalescent (PSMC) (*20*) model enables the inference of ancestral effective population size (N_e) using information from inter-chromosomal genetic differences within a single individual. We applied this model to each sample in our data set to investigate demographic history in mountain gorillas and their sister taxa.

Autosomal sequences: PSMC was first applied to the full autosomal sequence for each individual. Input in `psmcfa` format was generated from the all-sample VCF variation data described above, using 100 bp bins and accounting for uncallable sites (described in Supplementary section 2) as required by the software usage specification. PSMC was run with the command `psmc -p 4+25*2+4+6`, these temporal binning parameters were chosen based on consistency between individuals within the same population, but a reasonably broad family of similar parameters was found to give similar results.

Results were scaled using an assumed mutation rate of 1.2×10^{-8} per bp per generation (i.e. the same rate as has been measured in both humans and chimpanzees (*51*, *52*)) and a generation time of 19.3 years (*53*), and are shown in fig. 3A.

Variation between individuals within the same population provides a measure of the methodological noise, and we see that this is small except for in the most recent time step. Therefore we can be confident that the large-scale features of this plot, which are consistent within populations and even (at older timescales) across all three populations, are a faithful reflection of variation in ancestral N_e in gorillas. Nevertheless there are some factors which can lead to biases in PSMC inference. For example an apparent increase in N_e over a particular time period may be due to a genuine rise in census population size, but may also reflect the effects of population substructure or more complex demographic phenomena such as admixture from one population into another (*20*, *54*). Similarly, a very low recent N_e and recent inbreeding, such as is seen in both eastern gorilla populations, may lead to an overestimate of N_e at older time periods.

Thus while it seems likely that the sharp drop in population size amongst eastern gorillas from around 100,000 years ago reflects a genuine decline and a divergence from the western population, it is not clear whether the apparent earlier divergence of the inferred curves around 1 million years (Myr) ago reflects a genuine divergence or is artefactual.

The brief increase in inferred mountain gorilla N_e around 10 kyr ago may correspond to a genuine rise in population, perhaps associated with post-glacial montane forest recovery (*55*), or may alternatively be due to population substructure as noted above, perhaps relating to the incipient divergence of the two eastern subspecies at that time

Note that to the extent that the true values of gorilla mutation rate and/or generation time differ over the time periods considered from those assumed here, the resulting inferred timescale would be scaled accordingly. Thus for example, if the true rate were

half of that assumed here, our timescale and effective population size estimates would be doubled.

Male X-chromosomal sequences: We explored the divergence of gorilla species and subspecies more directly by repeating the PSMC analysis using pseudo-diploid sequences constructed from pairs of (haploid) male X chromosomes in our data. To understand the logic of this analysis, consider the case where two populations have diverged at some time t_d in the past, with no subsequent gene flow between them. In a comparison of two male X chromosomes, one from each population, we would expect to find no loci coalescing at any time more recently than t_d. In PSMC's inference this would manifest as an effectively infinite inferred N_e during this period. Prior to t_d, the N_e inferred from such a comparison should match that inferred from analysing the (diploid) X chromosome of a female individual from either population.

In more complex cases where the populations diverged gradually with ongoing gene flow, perhaps even via a third intermediate population, the signal of inferred N_e from any cross-population haplotype comparison is difficult to interpret. However even in such cases, we can interpret the point where N_e becomes infinite as a time after which detectible gene flow (by whatever route) ceases.

For this analysis, ten male X-chromosomal sequences were available in total: 3 mountain gorilla, 3 eastern lowland and 4 western lowland gorilla males, allowing 31 different cross-population comparisons in total. Input to PSMC in `psmcfa` format was generated from each of these pseudo-diploid sequences exactly as if they were ordinary diploid chromosomes, using 100 bp bins and accounting for uncallable sites as above. Inference was run separately on each of the 31 possible cross-population chromosomal comparisons. In addition, three combined cross-population analyses were carried out, in which PSMC inference was run on a concatenation of all the relevant pairwise combinations for each cross-population comparison into a single input file. The mutation rate on X is expected to be lower than that on autosomes due to the fact that germline mutations are more frequent in males. We have assumed a male-female mutation rate ratio of 3, corresponding to a factor 5/6 reduction on X compared to the autosomes.

The results of all these analyses are shown in fig. 3B. The clearest signal is the cessation of gene flow between both eastern subspecies and western lowland gorillas around 10–20 kyr ago, indicated by the sudden and extreme increase in inferred N_e in Gbb-Ggg and Gbg-Ggg comparisons at that time step. No equivalent signal is seen in comparisons between the two eastern subspecies, suggesting that gene flow has continued between them until more recently than this analysis is able to detect (and possibly until the present day).

Code used for managing SMC-based demographic inference is available at https://github.com/aylwyn/smcrun.git.

Gene flow (*D*-statistic) analysis

To investigate evidence of ancestral gene flow between the gorilla populations sampled here, we applied a sequence based *D*-statistic test, also referred to as the ABBA-BABA test, as follows. If G1, G2 and G3 denote 3 gorilla populations, the test evaluates whether the observed sequence data are consistent with a simple tree (((G1, G2), G3),

human) with no gene flow between G3 and either G1 or G2 (or any related populations) after their divergence. There are several different definitions of the *D*-statistic in the literature (e.g. (*56*, *57*)); we employed the test as described and implemented in ADMIXTOOLS (*58*), where we first define

N_i = p(BABA) – p(ABBA) = $(w - x)(y - z)$

M_i = p(BABA) + p(ABBA) = $(w + x - 2wx)(y + z - 2yz)$

where p(ABBA) is the probability of the G1 having the same allele as the outgroup (human) while G2 and G3 share a different allele (ABBA sites), and p(BABA) is the probability that G2 has the same allele as the outgroup and G1 and G3 share a different allele (BABA sites). The variables *w, x, y* and *z* represent the corresponding allele frequencies at each site in the four populations. Then the *D*-statistic is defined as the ratio of numerator and denominator, each summed over many SNPs

$D = \Sigma_i N_i / \Sigma_i M_i$

As usually applied, G1 and G2 are sister taxa and the statistic tests for asymmetry in the number of derived alleles either shares with G3 (e.g. due to historical or present-day gene flow). In the absence of such asymmetry, and assuming no issues of contamination or differential error rates, $D = 0$. We assess the significance of any deviation from 0 using a Z-score based on jackknife resampling with a 5 Mb block size (as implemented in ADMIXTOOLS).

In this case, we can use D-statistics to examine asymmetry between subspecies with regard to gene flow between the eastern and western gorilla species. In the results, shown in table S17, we see no significant difference between the two western subspecies (the top two rows of the table). There is on the other hand marginal evidence for more east-west gene flow involving mountain gorillas than eastern lowland gorillas. However given the relative geographical position of the two eastern subspecies, their particular demography and the apparently recent timescale of their divergence, this has to be regarded as a very tentative finding.

Protein coding variant consequences

We assigned consequences to each of the SNP variants called in our data using the Ensembl Variant Effect Predictor (VEP) (*59*), which predicts protein sequence consequences for each transcript associated with a given variant. (Annotation of genes and transcripts in the gorGor3.1 assembly was drawn from the Ensembl database v. 76 (www.ensembl.org).) We then applied VEP's filtering script with the --pick option, which selects one consequence per variant based on transcript canonical status, biotype and length, as well as the severity of the consequence for the resulting protein sequence (http://www.ensembl.org/info/docs/tools/vep/script/vep_options.html).

We classified coding sequence variants based on the following sequence ontology terms as assigned by VEP:

- *LoF variants*: transcript_ablation, splice_donor_variant, splice_acceptor_variant, stop_gained, frameshift_variant, inframe_insertion, inframe_deletion, splice_region_variant
- *Missense variants*: missense_variant
- *Synonymous variants*: synonymous_variant

Selection and purging of deleterious mutations

In order to investigate the effects of low diversity and inbreeding on selection in gorilla populations, we carried out a comparative analysis of mutations at functional loci, looking at different categories of genomic site as identified by annotation (Supplementary section 13).

Relative excess of missense and deficit of LoF variants in eastern gorillas compared to western lowland gorillas: We began by calculating a statistic which compares two populations, given a particular category of sites, in terms of the number of derived alleles found at sites within that category in one population rather than the other. Specifically, at each site i we write the observed derived allele frequency in population A as $f^{A}_{i} = d^{A}_{i} / n^{A}_{i}$, where n^{A}_{i} is the total number of alleles called there in population A and d^{A}_{i} is the number of derived alleles called. Similarly we define f^{B}_{i} in population B. Then if C is a particular category of protein-coding sites and I a set of intergenic sites, we define

$$L_{\mathrm{A,B}}(C) = \frac{\sum_{i \in C} f_i^A \left(1 - f_i^B\right)}{\sum_{j \in I} f_j^A \left(1 - f_j^B\right)}$$

The denominator (summing over putatively neutral intergenic sites) mitigates population-specific biases, such as mapping bias due to the fact that the reference genome was derived from a western gorilla individual.

We then define the following ratio:

$R_{\mathrm{A/B}}(\mathrm{C}) = L_{\mathrm{A,B}}(\mathrm{C}) / L_{\mathrm{B,A}}(\mathrm{C})$

as a measure of the relative number of derived alleles found more often in population A compared to population B. Estimates of the variance in $R_{\mathrm{A/B}}(\mathrm{C})$ were obtained using 100 block jackknifes on the set of sites in C.

Using the annotation of coding sequence genetic variants described in Supplementary section 13, we then compared the numbers of variants in each gorilla population as measured by the $R_{\mathrm{A/B}}(\mathrm{C})$ statistic in two categories: 1) variants giving rise to missense or non-synonymous changes in the protein coding sequence, and 2) variants causing loss of function (LoF). We expect the latter category to be more enriched for severely deleterious mutations than the former.

Figure 4 shows $R_{\mathrm{A/B}}$(missense) and $R_{\mathrm{A/B}}$(LoF) for comparisons between the two eastern gorilla subspecies and western lowland gorillas. In both eastern subspecies the burden of moderately deleterious (missense and non-synonymous) mutations is higher than in western lowland gorillas, consistent with selection having been less efficient and drift stronger effect in these smaller populations. However, the opposite is true of variation at LoF sites, and we surmise that this is due to the purging effect of very low effective population size and inbreeding.

No deficit of homozygous LoF genotypes in mountain gorillas: Many severely deleterious mutations may be recessive and are therefore only exposed to selection when present in the homozygous state within an individual. In an outbred population we

therefore expect to find a relative paucity of homozygous genotypes of the derived allele at LoF sites, whereas in an inbred population from which such alleles have already been purged we expect a much reduced deficit or even no such deficit. To investigate this, in each subspecies we counted n_{hom}, the number of LoF segregating sites where at least one individual is homozygous for the derived allele. Across all allele frequencies we found 87 such sites in mountain gorillas, 84 in eastern lowland gorillas and 143 in western lowland gorillas. However these numbers are likely to be influenced by several factors, including the differing numbers of samples in each population. To normalise for these factors and compare with the expectation under neutrality, we also counted n_{hom} in each of 10,000 random samples of segregating synonymous sites, ensuring that each sample had a distribution of allele frequencies matching that found in the segregating LoF sites.

Figure 4 shows the resulting distributions and LoF site counts, with all values for each subspecies normalised by the median of the sample distribution. A substantial deficit in LoF sites with homozygous derived allele genotypes is seen in western lowland gorillas, compared to the much reduced deficit found in eastern lowland gorillas and mountain gorillas. Treating the comparison as a test of whether LoF sites differ from synonymous sites in this regard for each subspecies, we assessed significance by the proportion of synonymous samples in which n_{hom} is smaller than the observed LoF value. The lack of significant difference in mountain gorillas reflects the fact that many severely deleterious LoF variants have already been purged from that population, compared to the strongly significant deficit seen in western lowland gorillas.

No LoF rate reduction in homozygous tracts in mountain and eastern lowland gorillas: A related signature of the purging effect of inbreeding on recessive deleterious LoF mutations can be seen in comparing the rates at which such mutations occur in homozygous and non-homozygous tracts within individuals. Since individuals in this study were either adults or mature juveniles when sampled, recessive LoF mutations with a deleterious effect on viability or survival in early infancy should be less common in homozygous tracts, where they are exposed, than elsewhere in the genome. Once again, to control for sample numbers, demographic factors and differing amounts of homozygosity within individuals, we normalise the rates of LoF variant sites in homozygous and heterozygous portions of the genome by the rates of synonymous variant sites in the same regions.

Results for each individual in all three subspecies are again shown in fig. 4. We assess significance of the difference between relative rates of LoF variants in the homozygous and non-homozygous portions of the genome using a Kolmogorov-Smirnov test. Consistent with the previous indicators, we see a signifcant reduction in relative LoF variant rate in homozygous tracts within western lowland gorillas, suggesting that they retain a population of severely deleterious but recessive mutations, whereas in both eastern lowland gorillas there is no significant difference between the rates in homozygous and non-homozygous tracts, suggesting that such mutations have already been purged.

Loss of function and missense variants in mountain gorillas

A total of 234 homozygous LoF genotypes (Auxiliary file 1) were observed in 241 protein coding genes in at least one of the gorillas in our dataset. Restricting the analysis to the 7 mountain gorillas sequenced, 95 homozygous LoF genotypes were found in at least one individual, affecting 99 protein coding genes. Sixty-seven of these genes had annotation in Ensembl based on evidence projected from the human genome, while the remainder were novel Ensembl predictions. Of note was a homozygous LoF variant in *ABCA12*, in which knock-out mutations have been associated with severe dermatological conditions in both humans and mice, with variable penetrance reported for other variants in this gene (*60*). Also Semenogelin II (*SEMG2*), a protein involved in the formation of sperm coagulum, in which the homozygous LoF genotype is fixed (i.e. present in all samples) in mountain gorillas.

There were a total of 113 genes with fixed missense differences between eastern lowland and mountain gorillas and 100 of these could be matched with a one-to-one human ortholog. Of these, 92 were found in the manually curated Ingenuity Pathway Analysis (IPA, QIAGEN Redwood City, www.qiagen.com/ingenuity) database. A core analysis did not reveal any genome-wide disease or functional enrichment, but several interesting disease categories such as morphology of nervous system, blood cells and immunoglobulin quantity were near the top of the list (fig. S17).

High-altitude adaptation: We did not find any fixed differences in genes previously associated with high altitude adaptation (*61*, *62*). (Several genes, such as the globin gene clusters and *EPAS1,* with well-characterized altitude association in humans, were not annotated in the gorilla reference genome, so syntenic regions in the gorGor3.1 assembly were examined in these cases.) We also examined patterns of linkage disequilibrium (LD) in 12 of the 18 loci (table S18) consistently associated with altitude adaptation in humans and animals, and observed unusually high LD in *VHL*, *VEGFA*, *ANGPT1* and *EPB42*. However it is difficult to attribute these signals to selection given the presence of large homozygous genomic tracts in mountain gorillas. Nevertheless *EPB42*, a gene affecting red blood cell morphology, has several interesting features suggestive of it being a candidate for altitude adaptation in mountain gorillas (fig. S19). Besides extensive surrounding LD the gene has 15 fixed differences between mountain and eastern lowland gorillas, including a non-synonymous change, and 10/14 intronic differences are fixed for the alternative allele in mountain gorillas. Although not reported to be associated with high altitude adaptation, EPB42 forms part of an interaction network with most of the proteins linked to adaptation at high altitude as determined by the STRING v.9.05 database (http://string-db.org/), which records known and predicted direct and indirect protein interactions derived from multiple sources and organisms (fig. S20). *EPB42* regulates the shape and mechanical properties of red blood cells and may thus contribute to adaptation to mountain living, perhaps by enabling the cells to store increased amounts of hemoglobin. It is thus an interesting target for future physiological studies on mountain gorilla blood samples.

Overlap with disease associated variants: We compared genetic variation in gorillas with variants previously associated with disease in humans by mapping gorilla sequence data to the human genome and comparing against the Human Gene Mutation Database Professional 2013.2 version (HGMD® http://www.hgmd.org/). For this

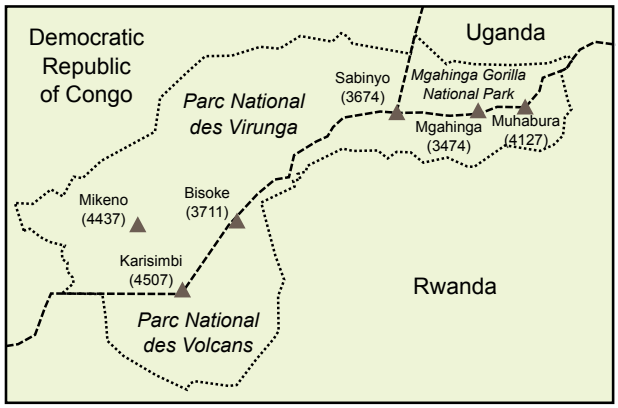
analysis we focussed on the mountain and eastern lowland gorillas sequenced here and two western individuals (Coco and Banjo) for comparison. Twenty-three gorilla variants corresponded to human disease-causing variants in 21 genes, with two variants each in the genes *F13A1* and *SLC26A4* (Auxiliary file 1). Pathway analysis using IPA (http://www.qiagen.com/ingenuity) revealed a significant enrichment for genes associated with blood coagulation (fig. S21).

Nine of these variants were fixed for the disease allele in all gorilla samples examined, and one associated with *TNNT2* was present at moderate frequency (0.33) in mountain gorillas, and was homozygous in one individual (Kaboko). This Ala28Val variant in exon 5 of *TNNT2* is associated with increased left ventricular thickness and cardiomyopathy in humans and has so far been reported in 4 patients. It is found in the heterozygous state in 4 out of 4,300 apparently healthy European Americans from the NHLBI Exome Sequencing Project and one TSI sample in the 1,000 Genomes populations (Phase 1). Clinvar (http://www.ncbi.nlm.nih.gov/clinvar/) labels it as likely benign, because it lies in the part of *TNNT2* that is poorly conserved and the valine residue has also been found in other species. Kaboko died aged 9 from an intestinal infection in 2012, but there was post-mortem evidence of myofiber hypertrophy in his heart, raising the possibility that he might have developed cardiomyopathy had he lived longer.

Two additional variants were fixed for the human disease allele (specifically long QT syndrome (*KCNE1*) and arrhythmogenic right ventricular dysplasia/cardiomyopathy (*PKP2*)) in 14 gorilla samples (6 mountain, 6 eastern lowland and 2 western lowland).

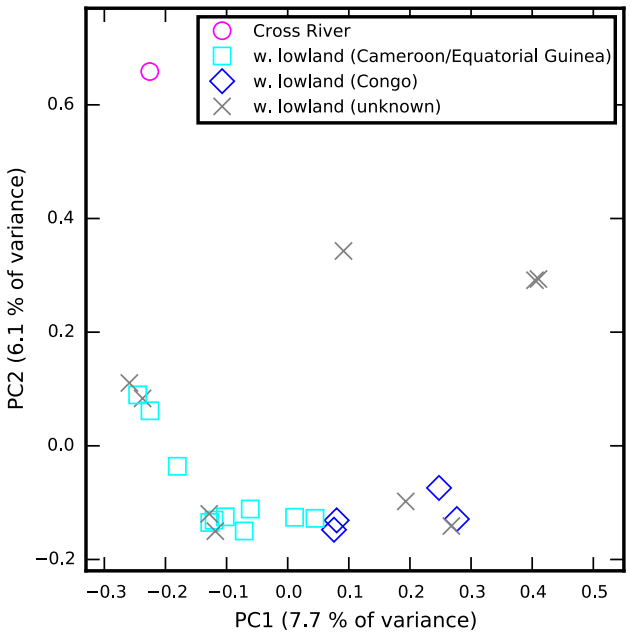
Assuming the disease associations of these variants are valid in humans, it may nevertheless be that compensatory molecular changes or differing environmental conditions mitigate their effects in gorillas.

Fig. S1.



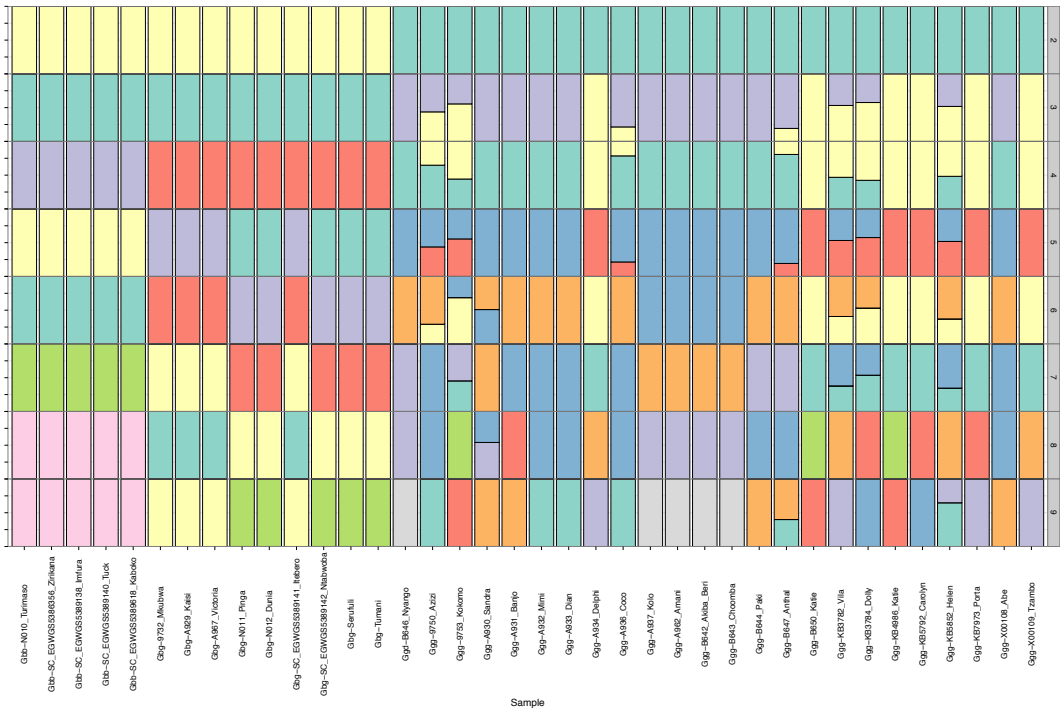
The Virunga volcanoes region. Mountain altitudes in meters above sea level.

Fig. S2



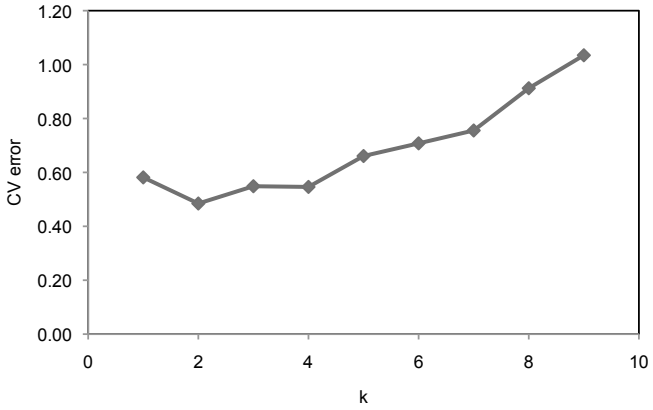
PCA plot of western lowland and Cross River gorilla SNP data, also showing available information on sample geographical origins.

Fig. S3



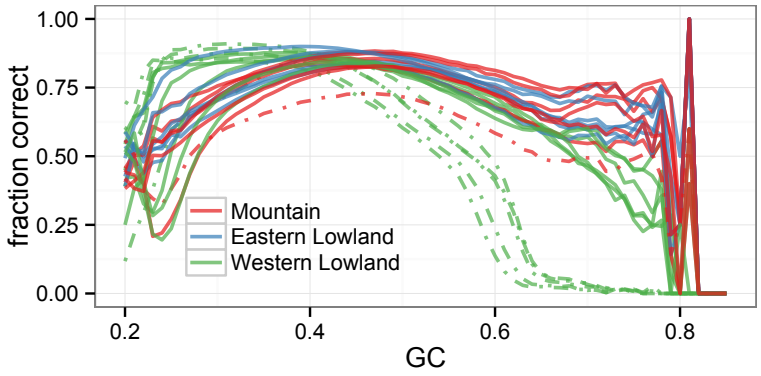
ADMIXTURE analysis for SNP data from all gorilla subspecies.

Fig. S4



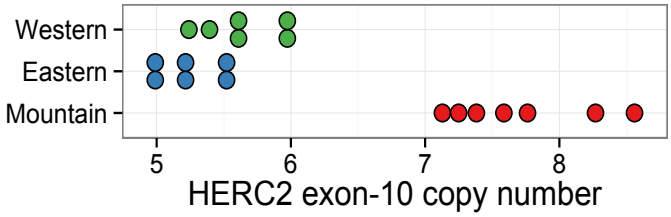
Cross-validation (CV) error for varying values of *k* in the ADMIXTURE analysis.

Fig. S5



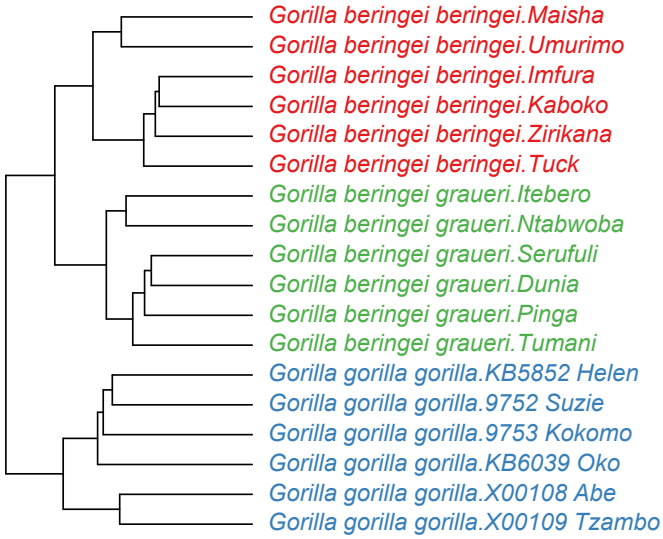
Genome-wide fraction of 500bp windows estimated as diploid as a function of GC-content. Dotted lines indicate outlier individuals which were discarded from further analysis.

Fig. S6



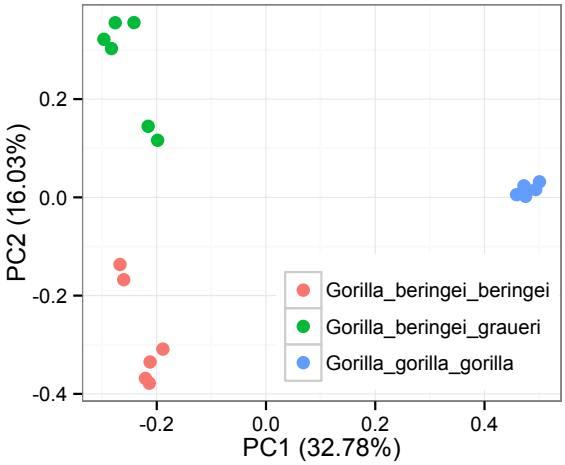
Copy number of *HERC2* exon 10 in western lowland, eastern lowland and mountain gorillas.

Fig. S7



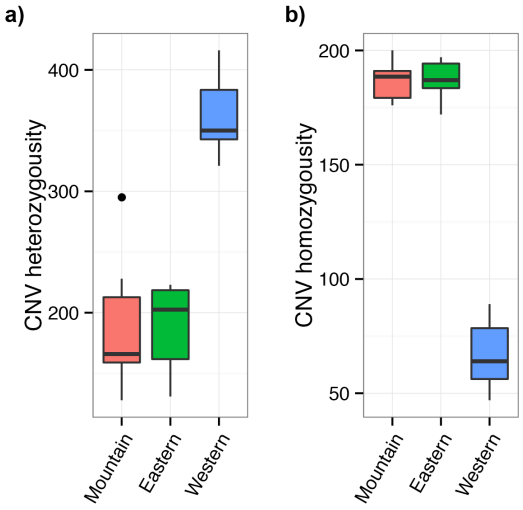
Dendrogram of relationships between gorillas based on hierarchical clustering of CNV genotypes.

Fig. S8



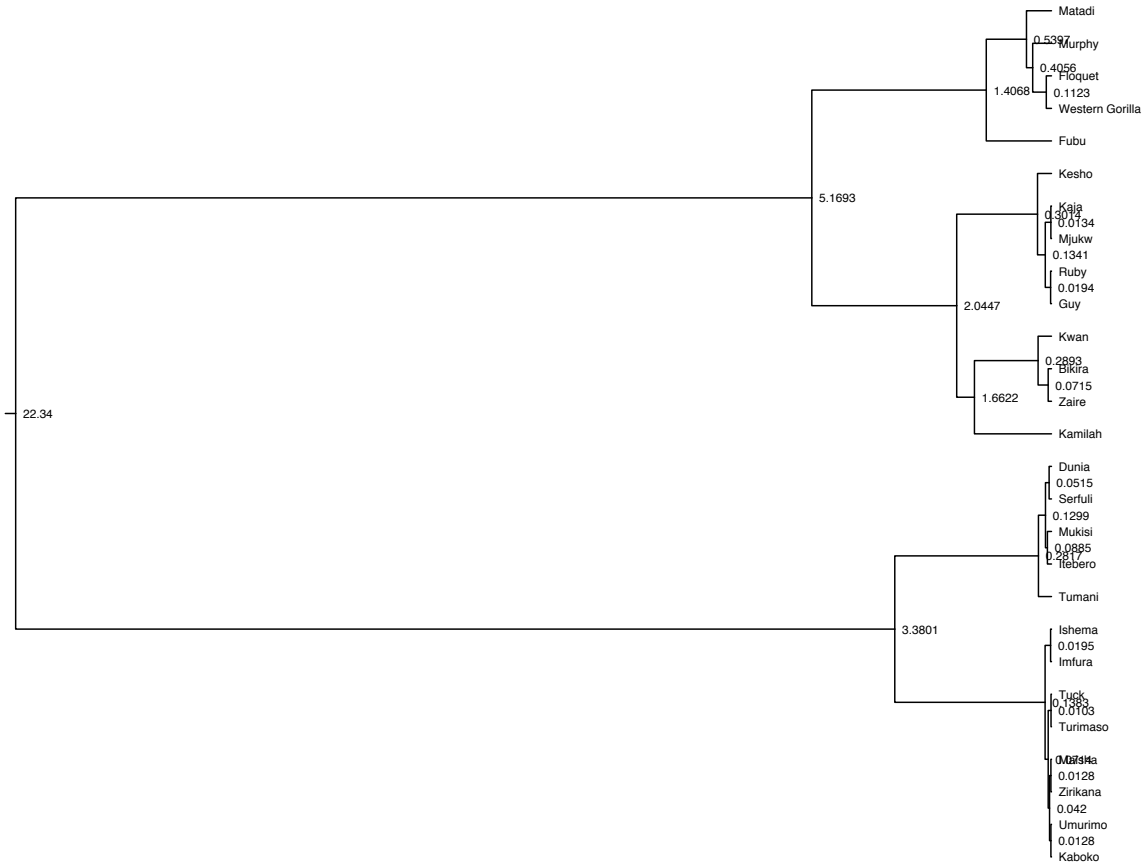
Principal component analysis of gorilla CNVs.

Fig. S9



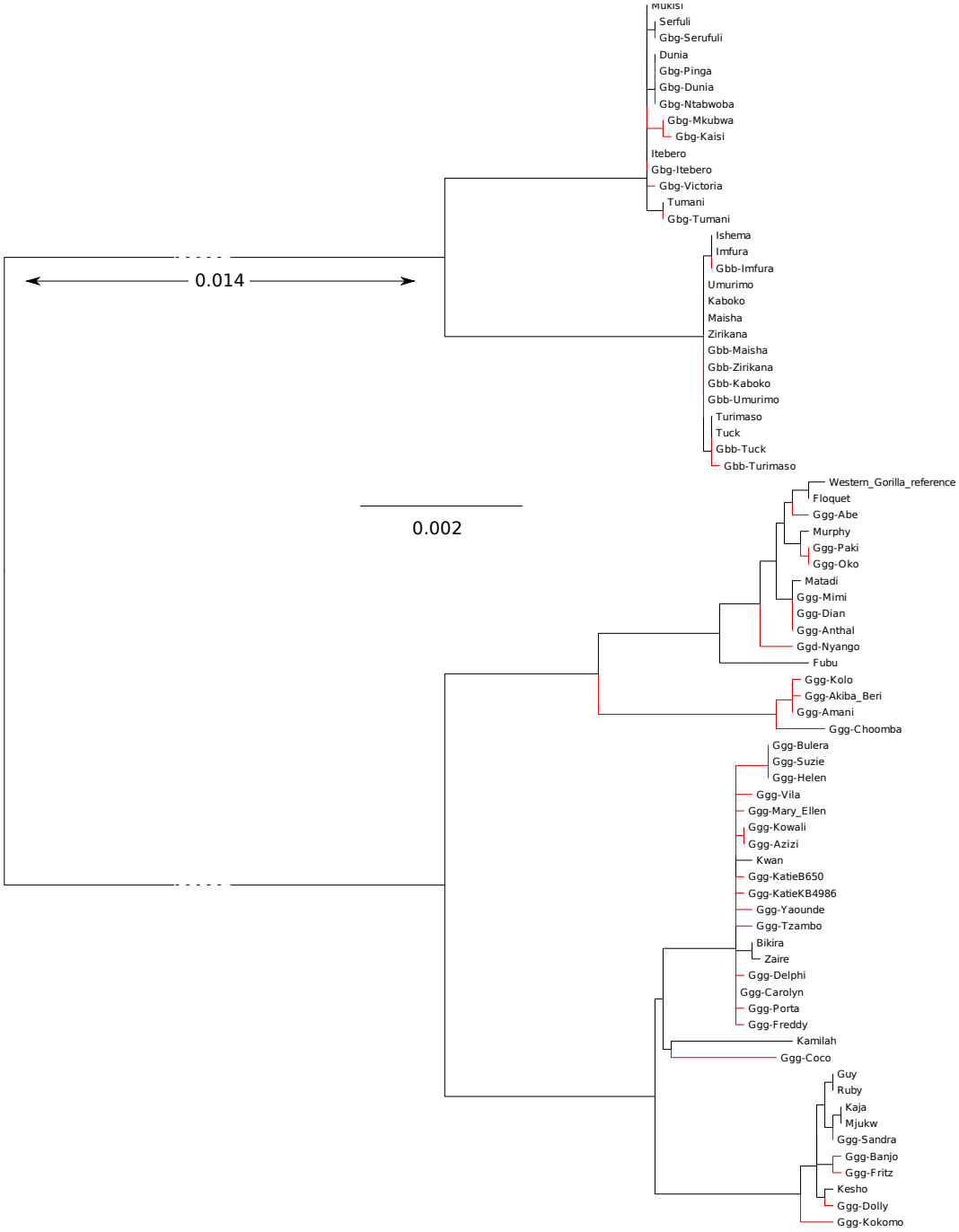
CNV heterozygosity and homozygosity. Boxplots of a), the number of heterozygous CNVs per individual and b) the number of homozygous CNVs per individual grouped by population.

Fig. S10



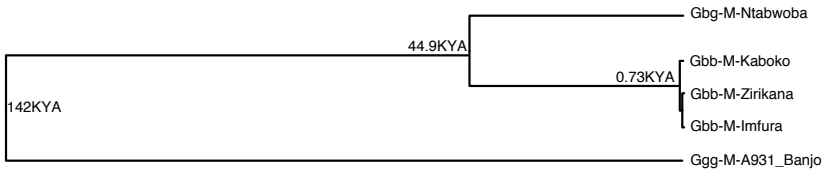
Ultrametric tree of gorilla mtDNA sequences assembled from long-range PCR sequencing. Node heights, annotated in units of substitutions per site, indicate the estimated sequence divergence between taxa.

Fig. S11



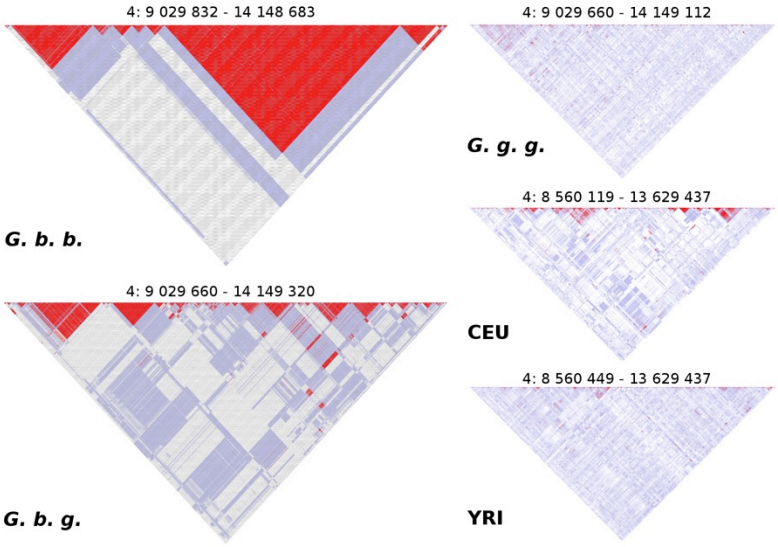
Phylogenetic neighbour-joining tree of mtDNA sequences reconstructed from both WGS and long-range PCR. WGS mtDNA branches are red and long-range PCR mtDNA are black; 10 samples were reconstructed using both methods, and are grouped together by the tree construction algorithm. Branch length scale indicated is in substitutions per bp.

Fig. S12



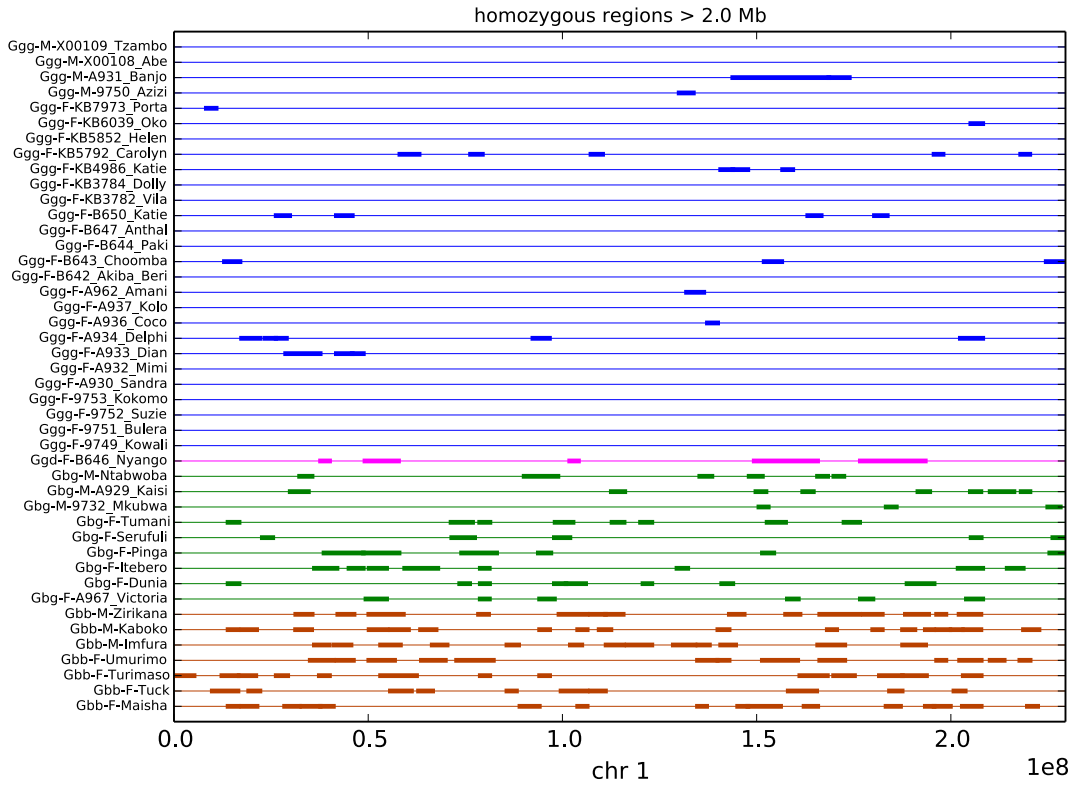
Phylogeny of gorilla Y-chromosomal sequences. Nodes dated assuming the human mutation rate of 1.0×10^{-9} bp^{-1} yr^{-1}.

Fig. S13



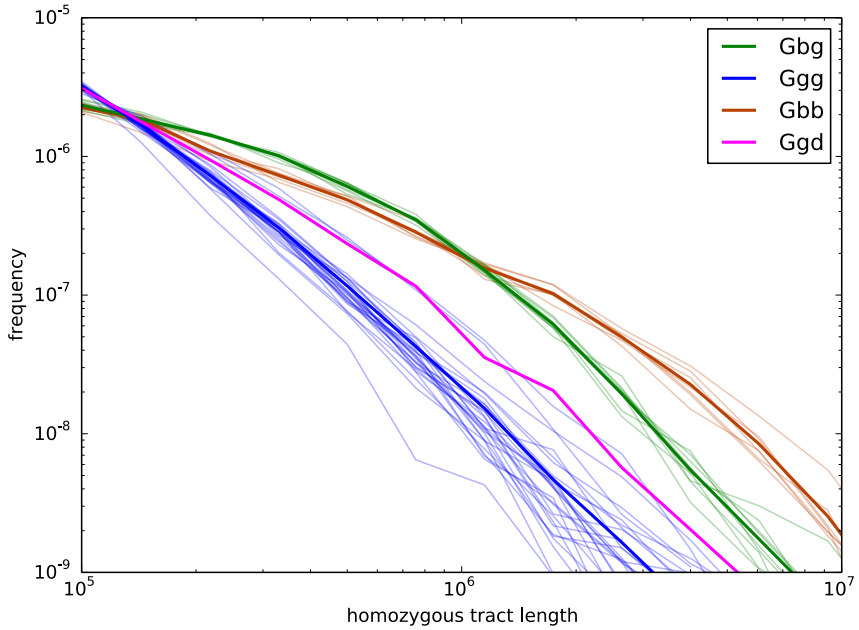
Example of LD blocks on chromosome 4. LD (here represented as D' (*47*)) calculated in windows of 5 kbp at a pairwise depth of 5000 SNPs, with red indicating D'=1, in a region where large blocks are present in mountain gorillas and in the syntenic region of the human genome. Human population data from the 1000 Genomes Project phase 1 v.3 20101123 release: CEU: Utah residents with northern and western European ancestry; YRI: Human Yoruba in Ibadan, Nigeria.

Fig. S14



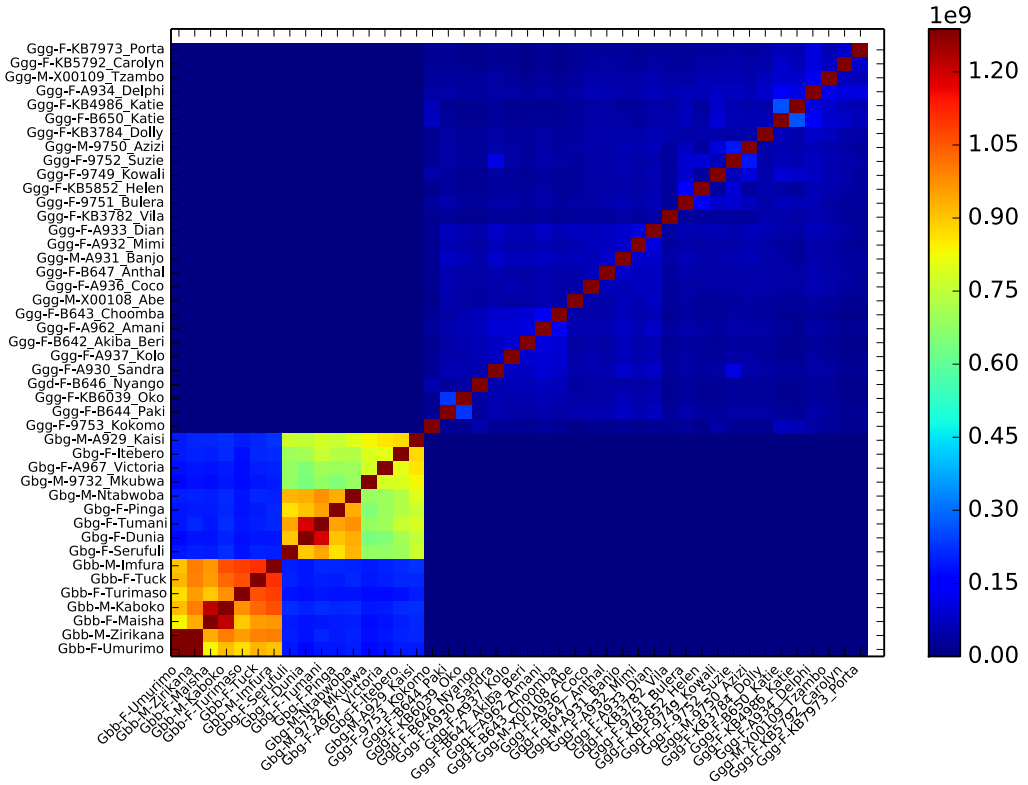
Example of homozygous regions longer than 2 Mb inferred on chromosome 1 for each sample.

Fig. S15



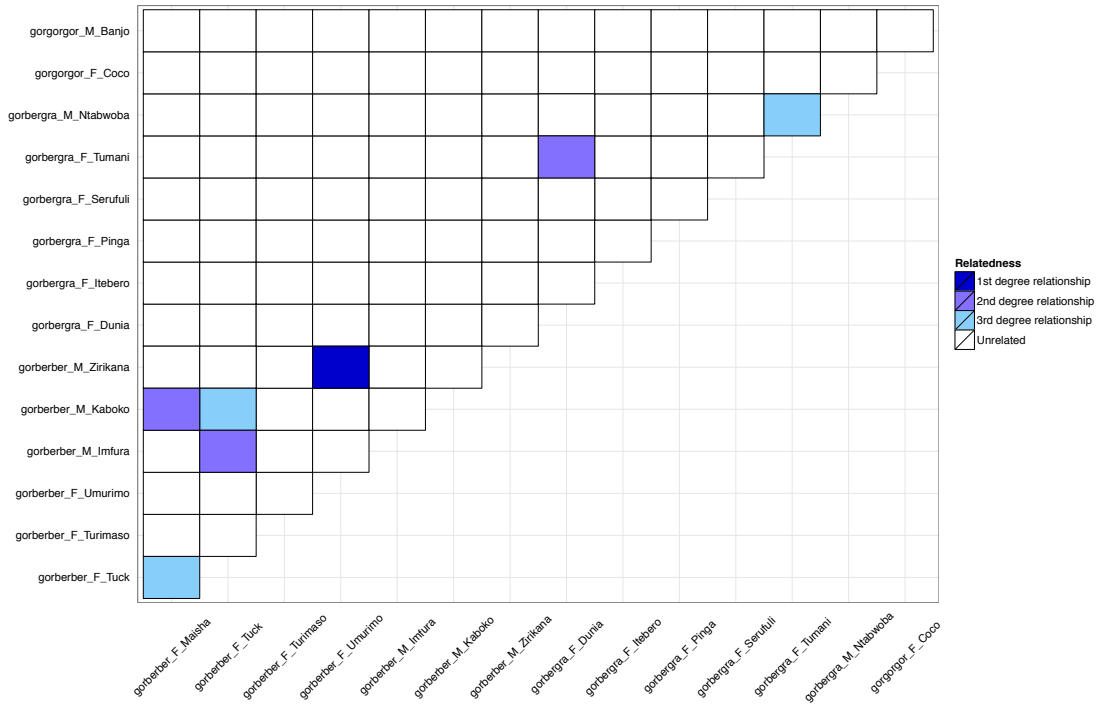
Distribution of homozygous tract lengths in sequenced samples. Thicker lines show whole-population distributions.

Fig. S16



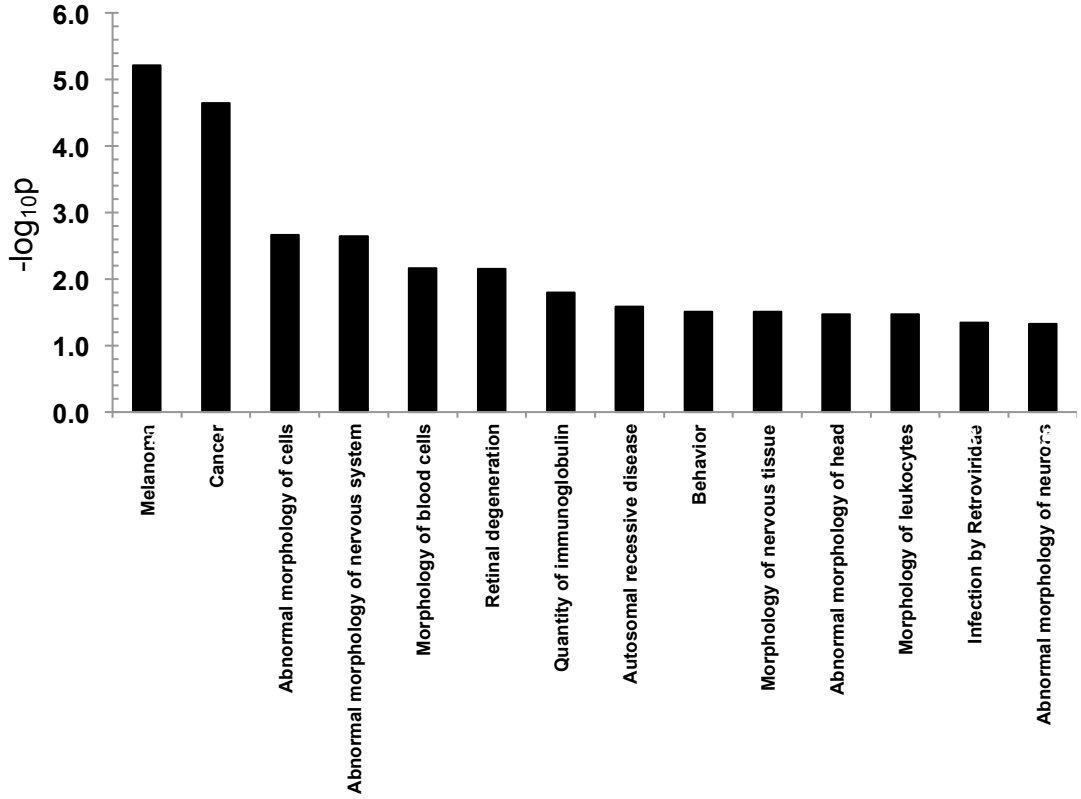
Total genome sequence shared between haplotypes in different individuals. Colours indicate the total length in base pairs of shared sequence between individuals, considering only one (haploid) set of chromosomes in each individual.

Fig. S17



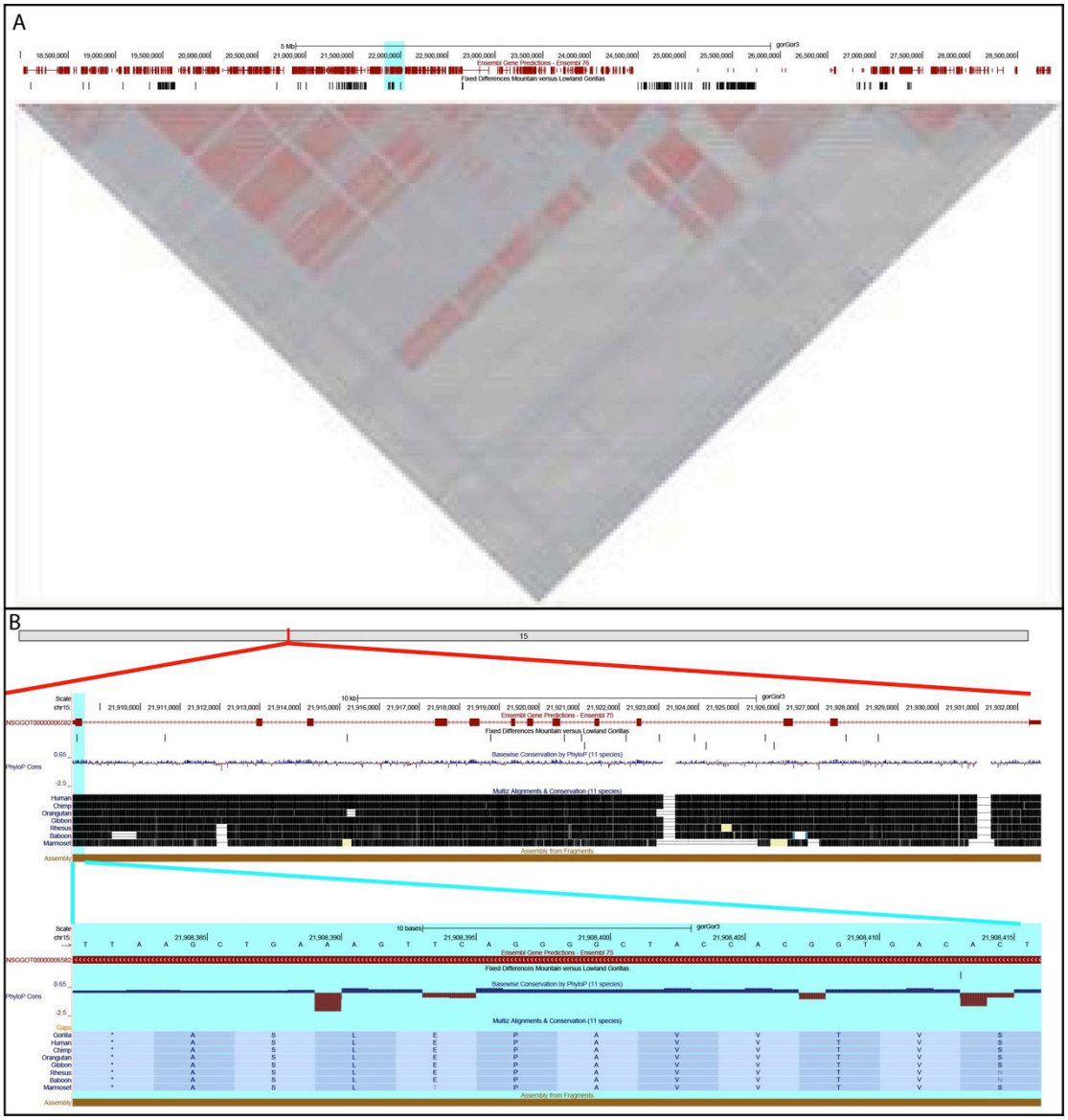
Estimated relatedness between individuals.

Fig. S18



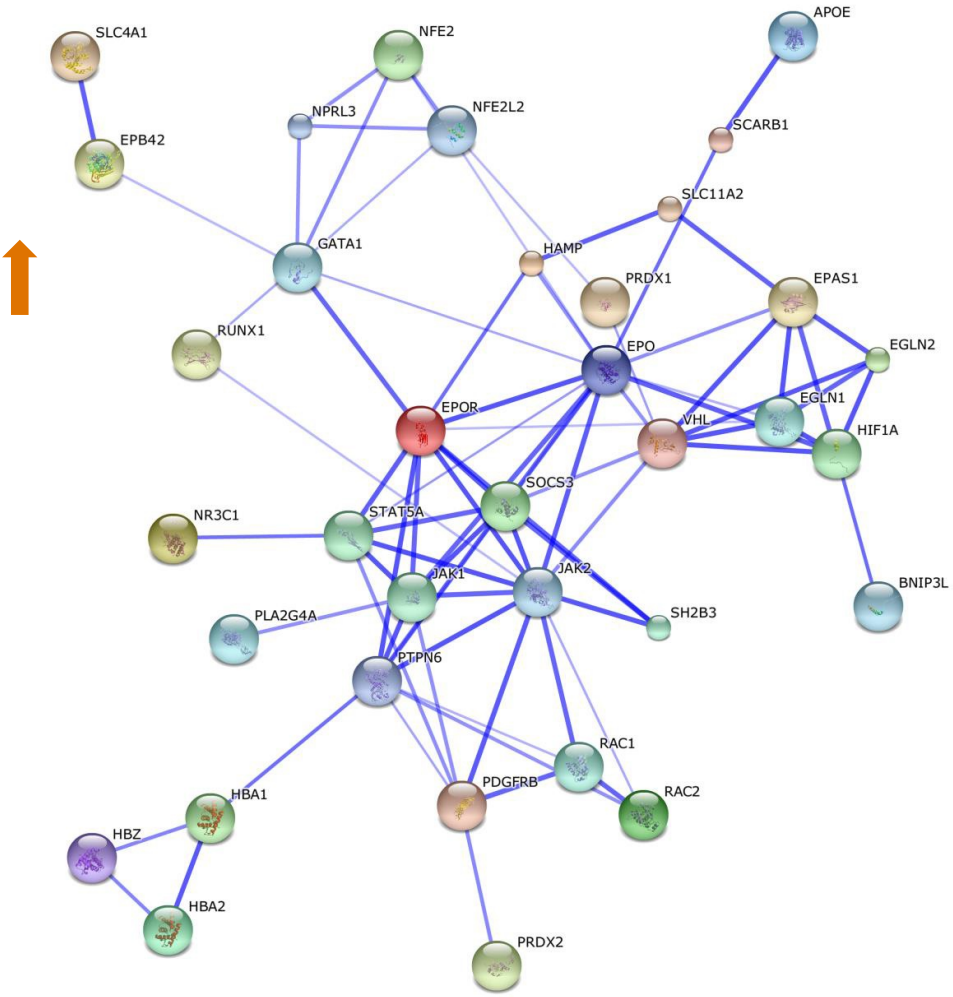
IPA core analysis of genes with fixed differences between lowland and mountain gorillas. The core analysis collates direct and indirect interactions which have been experimentally observed in human, mouse and rat models and reports pathways based on diseases and disorders, molecular and cellular functions and physiological system development and function. Significance of the association between the dataset and the pathways was tabulated by estimating a ratio between the number of genes from the dataset that met the expression value cut off that map to the pathway, and the total number of molecules present in the pathway using Fisher‘s exact test. Results are shown for pathways with at least 5 genes, and numbers of genes in each category are shown on the bars. None of the pathways are significant after applying a strict Bonferroni correction cut-off based on all coding genes in gorilla.

Fig. S19



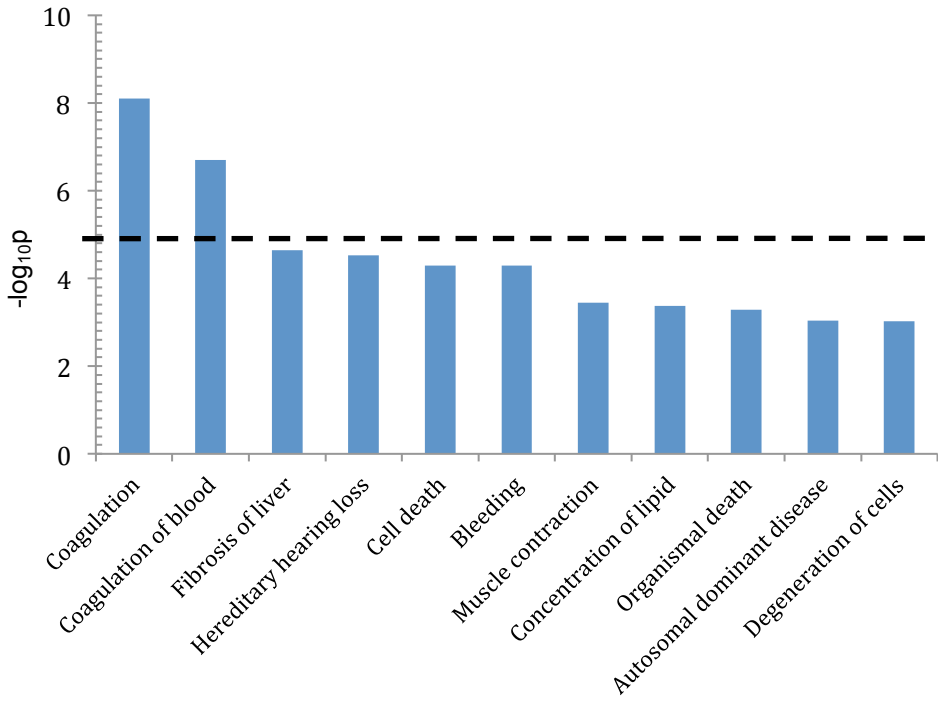
Evidence for adaptive changes at the *EPB42* locus. A. Linkage disequilibrium in an 11 Mb region on gorilla (gorGor3.1) chromosome 15. *EPB42* highlighted in blue; red in the LD plot indicates D' = 1. B. Region on chromosome 15 (chr15:21,908,311-21,932,582) showing the fixed missense variant at position 21908413 (C > T) in *EPB42* (region highlighted in blue). The change results in substitution of valine to isoleucine in the last exon of *EPB42* (Gtc > Atc). This is not the only fixed change observed in this gene as there are 14 additional fixed differences between the eastern lowland and mountain gorillas shown in the track below the gene's exonic- intronic boundaries.

Fig. S20



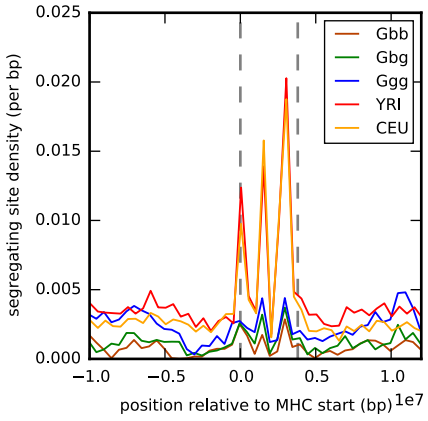
STRING interaction network of EPB42 and proteins linked to high altitude adaptation. EPB42 is indicated by the arrow. Network connections are based on high-confidence experimental evidence, with line thickness indicating the strength of evidence.

Fig. S21



IPA core analysis of genes containing human disease-associated variants. Information on disease association is taken from the Human Gene Mutation Database. Fisher's Exact Test was used to calculate p-values; the dashed line represents the threshold for significance after applying a strict Bonferroni correction using a cut-off based on all coding genes in the gorilla (20,962 in gorGor 3.1).

Fig. S22



Gorilla and human genetic diversity at the MHC. Density of segregating sites in the 20 Mb surrounding the major histocompatibility (MHC) locus on chromosome 6, as measured using 500 kbp windows in six individuals from each of three gorilla subspecies (Gbb: mountain gorillas; Gbg: eastern lowland gorillas; Ggg: western lowland gorillas), six YRI and six CEU human individuals. Gorilla individuals are those listed in table S4; human individuals are NA18486, NA18489, NA18498, NA18501, NA18504, NA18507, NA06984, NA06989, NA07000, NA07048, NA07056, NA07357 for which data from the 1000 Genomes Project Phase 1 was used. Chromosomal position is given in terms of position upstream (negative) or downstream (positive) of the MHC start (29.7 Mb along chr6 in the human reference genome, 30.8 Mb in gorilla).

Table S1.

Samples sequenced in this study. Species abbreviations: Gbb – *Gorilla beringei beringei* (mountain gorilla); Gbg – *Gorilla beringei graueri* (eastern lowland gorilla). DRC: Democratic Republic of Congo. ENA: European Nucleotide Archive (http://www.ebi.ac.uk/ena).

Species	Sex	Name	Birth year	Geographic origin	Median insert size (bp)	Read length (bp)	Sequenced Gbp	Sequencing centre	ENA accession No.
Gbb	F	Maisha	2001	Virunga National Park, DRC	313	101	97.98	UWash	ERS168410
Gbb	F	Tuck	1972	Rwanda (Titus group)	353	100	106.64	WTSI	ERS168204
Gbb	F	Turimaso	2003	Rwanda (Pablo group)	315	100	96.37	CNAG	ERS525618
Gbb	F	Umurimo	1989	Rwanda (Umubano group)	292	101	80.04	UWash	ERS525617
Gbb	M	Imfura	2008	Rwanda (Pablo group)	351	100	108.78	WTSI	ERS168207
Gbb	M	Kaboko	2003	Virunga National Park, DRC	345	100	120.01	WTSI	ERS168410
Gbb	M	Zirikana	2007	Rwanda (Umubano group)	340	100	102.81	WTSI	ERS168174
Gbg	F	Dunia	2004	Walikale, DRC	289	100	96.25	CNAG	ERS525621
Gbg	F	Itebero	2002	Parc National de Kahuzi-Biega, DRC	329	100	96.71	WTSI	ERS168205
Gbg	F	Pinga	2000	DRC	301	100	77.88	CNAG	ERS525620
Gbg	F	Serufuli	2002	DRC	282	101	107.72	UWash	ERS525622
Gbg	F	Tumani	2006	Walikale, DRC	311	101	93.19	UWash	ERS525619
Gbg	M	Ntabwoba	2000	Walikale, DRC	343	100	93.18	WTSI	ERS168206

Table S2.

Additional samples used in whole genome analyses. Sequence data and full sample details were previously published in Prado-Martinez et al. (2013) (12). Species abbreviations: Gbb – Gorilla beringei beringei (mountain gorilla); Gbg – Gorilla beringei graueri (eastern lowland gorilla); Ggg – Gorilla gorilla gorilla (western lowland gorilla); Ggd – Gorilla gorilla diehli (Cross River gorilla).

Species	Sex	Name	Geographic origin	Born	Studbook #
Gbg	M	M'kubwa	DRC (Tulakwa)	Wild born	9907
Gbg	M	Kaisi	DRC (Walikale)	Wild born	9909
Gbg	F	Victoria	DRC	Captive born	9919
Ggd	F	Nyango	West Africa	Wild born	9941
Ggg	F	Kowali	Unknown	Captive born	663
Ggg	F	Bulera	Cameroon	Captive born	1120
Ggg	M	Azizi	Cameroon	Captive born	1459
Ggg	F	Suzie	Unknown	Wild born	636
Ggg	F	Kokomo	Unknown	Captive born	1049
Ggg	M	Banjo	Cameroon	Wild born	255
Ggg	F	Mimi	Cameroon	Wild born	241
Ggg	F	Dian	Cameroon	Captive born	1091
Ggg	F	Delphi	Congo	Wild born	230
Ggg	F	Sandra	Cameroon	Captive born	969
Ggg	F	Coco	Equatorial Guinea	Wild born	1351
Ggg	F	Kolo	Cameroon	Captive born	936
Ggg	F	Amani	Unknown (captive born)	Captive born	899
Ggg	M	Abe	Unknown	Wild born	52
Ggg	M	Tzambo	Unknown	Wild born	440
Ggg	F	Akiba Beri	Cameroon	Wild born	1926
Ggg	F	Choomba	West Africa	Wild born	180
Ggg	F	Paki	West Africa	Wild born	191
Ggg	F	Anthal	West Africa	Wild born	1930
Ggg	F	Katie	West Africa	Wild born	498
Ggg	F	Carolyn	Congo	Wild born	3
Ggg	F	Porta	Unknown	Wild born	64
Ggg	F	Vila	Congo	Wild born	80
Ggg	F	Helen	Cameroon	Wild born	96
Ggg	F	Oko	Unknown	Wild born	192
Ggg	F	Dolly	Congo	Wild born	195
Ggg	F	Katie	Unknown	Wild born	498

Table S3.

Summary statistics for SNP sites. All sites in the all-sample filtered VCF file (sites which segregate across all samples) are included here.

	Gorilla beringei beringei	*Gorilla beringei graueri*	*Gorilla gorilla gorilla*	*Gorilla gorilla diehli*	All
Samples	7	9	27	1	44
Mean depth	26.8	23.5	17.2	15.3	20.0
Fixed sites differing from gorGor3.1 [a]	2,071,591	1,858,680	41,483	1,955,859	36,683
Fixed sites matching gorGor3.1 [b]	6,918,149	6,588,591	1,755,172	8,199,333	0
Polymorphic sites [c]	2,790,350	3,332,819	9,983,435	1,624,898	11,743,407
Mean heterozygosity per sample per kbp [d]	0.647885	0.642765	1.438033	0.985048	1.150145
S [e]	0.00169	0.00202	0.00605	0.00099	0.00714
Θ_w [f]	0.000652	0.000714	0.001555	0.000985	0.001633

a Variants fixed within the indicated population or group (column heading) which differ from the reference genome. b SNP sites at which all samples in the indicated population match the reference genome. c Sites segregating within the indicated population. d Mean number of heterozygous positions per sample divided by the callable genome length. e Number of segregating sites divided by the callable genome length. f Watterson's theta estimator (S divided by the harmonic number of samples).

Table S4

Individuals assessed for copy number variation

Gorilla beringei beringei	*Gorilla beringei graueri*	*Gorilla gorilla gorilla*
Imfura	Dunia	9752_Suzie
Maisha	Itebero	9753_Kokomo
Tuck	Ntabwoba	KB6039_Oko
Umurimo	Pinga	X00108_Abe
Zirikana	Serufuli	X00109_Tzambo
Kaboko	Tumani	KB5852_Helen

Table S5

Lineage specific fixed copy number variants.

Lineage	deletions	duplications	expansions
Western lowland	14 (85.1 kbp)	66 (221.8 kbp)	457 (1779.3 kbp)
Eastern lowland	24 (112.1 kbp)	14 (53.7 kbp)	58 (485.8 kbp)
Mountain	18 (59.3 kbp)	4 (18.1 kbp)	58 (330.8 kbp)
Mountain + eastern lowland	57 (340.0 kbp)	29 (112.8 kbp)	428 (1961.5 kbp)

Table S6

Lineage-specific fixed genic copy number variants.

Lineage	Genic deletions	Genic duplications	Genic expansions
Western lowland	3	4	172
Eastern lowland	2	1	28
Mountain	3	3	27
Mountain and eastern lowland	4	16	98

Table S7

Characteristics of long-range primer pairs.

Segm.	F/R[1]	Primer name	Forward primer sequence 5′→3′	Gene
A	F	Cytbf*	CACGAAACAGGATCAAATAACCC	CYTB
A	R	COIIrev592*	TGGTTGGCTCCACAGATTTC	COX2
B	F	COII28for*	AAGACGCTACTTCTCCTATCATAGA	COX2
B	F	GgmtB01F	CAAGACGCTACTTCTCCTATCATAG	COX2
B	F	GgmtB02F	AGACGCTACTTCTCCTATCATAGAA	COX2
B	R	12So*	GTCGATTATAGGACAGGTTCCTCTA	rRNA-12S
B	R	GgmtB01R	TCGATTACAGGACAGGCTCCTCTA	rRNA-12S
B	R	GgmtB02R	TATCGATTACAGGACAGGCTCCTCTA	rRNA-12S

[1] F/R: forward or reverse primer. * Universal primers published in (Thalmann *et al*. 2004).

Table S8

Overview of primer–template mismatches.

	12So sequence 5′→3′	Cytbf sequence 5′→3′
Ref. seq.*	ATCGATTACAGAACAGGCTCCTCTA	CACGAAACAGGATCAAACAACCC
Primer seq.	G-------T--G-----T-------	----------------T-----

* NCBI Reference Sequence: NC_011120.1

Table S9

Characteristics of control region gorilla-specific primer pairs.

Fragment	F/R*	Primer name	Forward primer sequence 5′→3′	Gene
1	F	GD01F	ACCATCAGCACCCAAAGCTA	D-loop
1	R	GD01R	CAGATGCCGGATACAGTTCA	D-loop
2	F	GD02F	ACACCATCCTCCGTGAAATC	D-loop
2	R	GD02R	TTTTGGTGTGAAGGGTGGTT	D-loop
3	F	GD03F	GCACCACATGTCGCAGTATC	D-loop
3	R	GD03R	CCCGTCGAAACATTTTCAGT	D-loop

* F/R: forward or reverse primer.

Table S10

Samples examined in this study and amplicons obtained.

Individual	Species*	Sex	Segment A	Segment B
BiKira	Ggg	Female	×	×
Effie	Ggg	Female		
Floquet	Ggg	Male	×	×
Fubu	Ggg	Male	×	×
Guy	Ggg	Male	×	×
Kaja	Ggg	Female	×	×
Kamilah	Ggg	Female	×	×
Kesho	Ggg	Male	×	×
Kwan	Ggg	Male	×	×
Matadi	Ggg	Male	×	×
Mjukuu	Ggg	Female	×	×
Murphy	Ggg	Male	×	×
Ruby	Ggg	Female	×	×
Zaire	Ggg	Female	×	×
Imfura	Gbb	Male	×	×
Ishema	Gbb	Female	×	×
Kaboko	Gbb	Male	×	×
Maisha	Gbb	Female	×	×
Tuck	Gbb	Female	×	×
Turimaso	Gbb	Female	×	×
Umurimo	Gbb	Female	×	×
Zirikana	Gbb	Male	×	×
Dunia	Gbg	Female	×	×
Itebero	Gbg	Female	×	×
Mukisi	Gbg	Male	×	×
Ntabwoba	Gbg	Male	×	
Pinga	Gbg	Female	×	
Serufuli	Gbg	Female	×	×
Tumani	Gbg	Female	×	×
Victoria	Gbg	Female	×	(×)

× indicates at least one successful amplification and sequencing; Victoria fragment B failed sequencing. * Gorilla subspecies: Ggg, western lowland gorilla; Gbb, mountain gorilla; Gbg, eastern lowland gorilla.

Table S11

GenBank mtDNA sequences used in evolutionary analyses.

GenBank accession number	Species
NC_011120.1*	*Gorilla gorilla gorilla*
NC_012920.1**	*Homo sapiens*

* Used as gorilla reference sequence in this study. ** Revised Cambridge Reference Sequence used as an outgroup.

Table S12

Additional samples used for WGS mitochondrial genome reconstruction.

Species	Name	Sex
Gorilla gorilla gorilla	Fritz	M
Gorilla gorilla gorilla	Yaounde	M
Gorilla gorilla gorilla	Mary Ellen	F
Gorilla gorilla gorilla	Freddy	M

Table S13

Comparison of WGS mtDNA reconstruction to experimentally-validated long-range PCR mtDNA amplification.

Name	Species	whole mtDNA	D-loop
Dunia	Gbg	99.99 %	99.80 %
Imfura	Gbb	99.99 %	99.90 %
Itebero	Gbg	99.99 %	99.80 %
Kaboko	Gbb	99.98 %	99.90 %
Maisha	Gbb	99.99 %	99.80 %
Tuck	Gbb	99.99 %	99.90 %
Tumani	Gbg	99.98 %	99.69 %
Turimaso	Gbb	99.99 %	99.80 %
Umurimo	Gbb	99.98 %	99.69 %
Zirikana	Gbb	99.98 %	99.70 %
Average		99.99 %	99.80 %

In each case the few differences between the two methods (3 sites per mtDNA) are in the D-loop.

Table S14

Samples included in LD analysis.

Population	Individuals
G. b. beringei	Maisha, Turimaso, Zirikana, Imfura, Tuck, Kaboko, Umurimo
G. b. graueri	Mkubwa, Kaisi, Victoria, Pinga, Dunia, Itebero, Tumani
G. g. gorilla	Banjo, Mimi, Delphi, Coco, Choomba, Paki, Vila
human CEU*	NA11932, NA11933, NA12046, NA12282, NA12283, NA12340, NA12341
human YRI§	NA18501, NA18502, NA18504, NA18505, NA18507, NA18522, NA18523

* Human Utah Residents with Northern and Western European ancestry. § Human Yoruba in Ibadan, Nigera. CEU and YRI samples are all from the 1000 Genomes Project phase 1 v.3 20101123 release.

Table S15

Mean total genomic fraction shared between chromosomes in the same individual for different gorilla species and subspecies.

Gorilla gorilla gorilla	13.8%
Gorilla gorilla diehli	34.2%
Gorilla beringei graueri	38.4%
Gorilla beringei beringei	34.5%

Table S16

Mean chromosomal sequence sharing (as a percentage of chromosome length) between and within different gorilla populations, excluding sharing between Zirikana and Umurimo.

	Gbb	Gbg	Ggg	Ggd
Gbb	33.9%	6.75%	0.005%	0.015%
Gbg		27.5%	0.006%	0.009%
Ggg			1.7%	1.4%

Table S17

D-statistic estimates from gorilla population SNP data, using human as an outgroup.

G1	G2	G3	D-statistic	Z-score
Ggg	Ggd	Gbb	–0.0191	–1.964
Ggg	Ggd	Gbg	–0.0145	–1.531
Gbb	Gbg	Ggd	0.0275	2.183
Gbb	Gbg	Ggg	0.0210	2.135

Table S18

High-confidence candidate genes for altitude adaptation in humans.

			gorGor3.1			
***Hs* Ensembl Gene ID**	**HGNC**	**Ensembl Gene ID**	**Chr**	**Strand**	**Gene Start**	**Gene End**
ENSG00000135766	*EGLN1*	ENSGGOG00000024267	1	-1	211562474	211621707
ENSG00000198626	*RYR2*	ENSGGOG00000012582	1	1	217665358	218153888
ENSG00000116016	*EPAS1*	Not annotated				
ENSG00000134086	*VHL*	ENSGGOG00000010456	3	1	10470360	10481554
ENSG00000150630	*VEGFC*	Not annotated				
ENSG00000078401	*EDN1*	ENSGGOG00000013300	6	1	12812470	12819367
ENSG00000112715	*VEGFA*	ENSGGOG00000012714	6	1	45082785	45099070
ENSG00000130427	*EPO*	Not annotated				
ENSG00000164867	*NOS3*	ENSGGOG00000016825	7	1	149442787	149467320
ENSG00000154188	*ANGPT1*	ENSGGOG00000023090	8	-1	106603128	106853411
ENSG00000244734	*HBB*	Not annotated				
ENSG00000173511	*VEGFB*	ENSGGOG00000012193	11	1	61047250	61050586
ENSG00000129521	*EGLN3*	ENSGGOG00000014900	14	-1	15150371	15177077
ENSG00000100644	*HIF1A*	ENSGGOG00000003957	14	1	42668724	42719850
ENSG00000206172	*HBA1*	Not annotated				
ENSG00000159640	*ACE*	Not annotated				
ENSG00000187266	*EPOR*	ENSGGOG00000022182	19	-1	11596799	11603371
ENSG00000269858	*EGLN2*	ENSGGOG00000014955	19	1	38229848	38240860

Additional Data table S1 (separate file)

Catalogs of mtDNA variation, Y variation, lineage-specific CNV events, homozygous LoFs and disease-associated variants.

Additional Data table S2 (separate file)

Catalog of gorilla ancestral informative markers

References and Notes

1. G. B. Schaller, *The Year of the Gorilla* (Univ. of Chicago Press, Chicago, 2010), p. 7.

2. M. Robbins *et al*., *Gorilla beringei* ssp. *beringei* (listing in IUCN Red List of Threatened Species, 2008); www.iucnredlist.org/details/39999/0.

3. M. Gray, J. Roy, L. Vigilant, K. Fawcett, A. Basabose, M. Cranfield, P. Uwingeli, I. Mburanumwe, E. Kagoda, M. M. Robbins, Genetic census reveals increased but uneven growth of a critically endangered mountain gorilla population. *Biol. Conserv.* **158**, 230–238 (2013). doi:10.1016/j.biocon.2012.09.018

4. K. Guschanski, L. Vigilant, A. McNeilage, M. Gray, E. Kagoda, M. M. Robbins, Counting elusive animals: Comparing field and genetic census of the entire mountain gorilla population of Bwindi Impenetrable National Park, Uganda. *Biol. Conserv.* **142**, 290–300 (2009). doi:10.1016/j.biocon.2008.10.024

5. A. H. Harcourt, Is the gorilla a threatened species? How should we judge? *Biol. Conserv.* **75**, 165–176 (1996). doi:10.1016/0006-3207(95)00059-3

6. A. W. Weber, A. Vedder, Population dynamics of the virunga gorillas: 1959–1978. *Biol. Conserv.* **26**, 341–366 (1983). doi:10.1016/0006-3207(83)90096-4

7. P. J. Le Gouar, D. Vallet, L. David, M. Bermejo, S. Gatti, F. Levréro, E. J. Petit, N. Ménard, How Ebola impacts genetics of Western lowland gorilla populations. *PLOS ONE* **4**, e8375 (2009). Medline doi:10.1371/journal.pone.0008375

8. IUCN, “Ebola outbreak highlights critical links between biodiversity loss and human health, says IUCN’s Wildlife Health Specialist Group” (2014); www.iucn.org/?18439.

9. K. J. Garner, O. A. Ryder, Mitochondrial DNA diversity in gorillas. *Mol. Phylogenet. Evol.* **6**, 39–48 (1996). Medline doi:10.1006/mpev.1996.0056

10. M. I. Jensen-Seaman, K. K. Kidd, Mitochondrial DNA variation and biogeography of eastern gorillas. *Mol. Ecol.* **10**, 2241–2247 (2001). Medline

11. J. Roy, M. Arandjelovic, B. J. Bradley, K. Guschanski, C. R. Stephens, D. Bucknell, H. Cirhuza, C. Kusamba, J. C. Kyungu, V. Smith, M. M. Robbins, L. Vigilant, Recent

divergences and size decreases of eastern gorilla populations. *Biol. Lett.* **10**, 2014.0811 (2014). doi:10.1098/rsbl.2014.0811

12. J. Prado-Martinez, P. H. Sudmant, J. M. Kidd, H. Li, J. L. Kelley, B. Lorente-Galdos, K. R. Veeramah, A. E. Woerner, T. D. O'Connor, G. Santpere, A. Cagan, C. Theunert, F. Casals, H. Laayouni, K. Munch, A. Hobolth, A. E. Halager, M. Malig, J. Hernandez-Rodriguez, I. Hernando-Herraez, K. Prüfer, M. Pybus, L. Johnstone, M. Lachmann, C. Alkan, D. Twigg, N. Petit, C. Baker, F. Hormozdiari, M. Fernandez-Callejo, M. Dabad, M. L. Wilson, L. Stevison, C. Camprubí, T. Carvalho, A. Ruiz-Herrera, L. Vives, M. Mele, T. Abello, I. Kondova, R. E. Bontrop, A. Pusey, F. Lankester, J. A. Kiyang, R. A. Bergl, E. Lonsdorf, S. Myers, M. Ventura, P. Gagneux, D. Comas, H. Siegismund, J. Blanc, L. Agueda-Calpena, M. Gut, L. Fulton, S. A. Tishkoff, J. C. Mullikin, R. K. Wilson, I. G. Gut, M. K. Gonder, O. A. Ryder, B. H. Hahn, A. Navarro, J. M. Akey, J. Bertranpetit, D. Reich, T. Mailund, M. H. Schierup, C. Hvilsom, A. M. Andrés, J. D. Wall, C. D. Bustamante, M. F. Hammer, E. E. Eichler, T. Marques-Bonet, Great ape genetic diversity and population history. *Nature* **499**, 471–475 (2013). Medline doi:10.1038/nature12228

13. D. Fossey, *Gorillas in the Mist* (Harcourt Brace, New York, 1983), p. 72.

14. A. Routh, J. Sleeman, *in Proceedings of the British Veterinary Zoological Society* (Howletts and Port Lympne Wild Animal Parks, Kent, 14–15 June 1997), pp. 22–25.

15. See supplementary materials on *Science* Online.

16. A. Scally, J. Y. Dutheil, L. W. Hillier, G. E. Jordan, I. Goodhead, J. Herrero, A. Hobolth, T. Lappalainen, T. Mailund, T. Marques-Bonet, S. McCarthy, S. H. Montgomery, P. C. Schwalie, Y. A. Tang, M. C. Ward, Y. Xue, B. Yngvadottir, C. Alkan, L. N. Andersen, Q. Ayub, E. V. Ball, K. Beal, B. J. Bradley, Y. Chen, C. M. Clee, S. Fitzgerald, T. A. Graves, Y. Gu, P. Heath, A. Heger, E. Karakoc, A. Kolb-Kokocinski, G. K. Laird, G. Lunter, S. Meader, M. Mort, J. C. Mullikin, K. Munch, T. D. O'Connor, A. D. Phillips, J. Prado-Martinez, A. S. Rogers, S. Sajjadian, D. Schmidt, K. Shaw, J. T. Simpson, P. D. Stenson, D. J. Turner, L. Vigilant, A. J. Vilella, W. Whitener, B. Zhu, D. N. Cooper, P. de Jong, E. T. Dermitzakis, E. E. Eichler, P. Flicek, N. Goldman, N. I. Mundy, Z. Ning, D. T. Odom, C. P. Ponting, M. A. Quail, O. A. Ryder, S. M. Searle, W. C. Warren, R. K.

Wilson, M. H. Schierup, J. Rogers, C. Tyler-Smith, R. Durbin, Insights into hominid evolution from the gorilla genome sequence. *Nature* **483**, 169–175 (2012). Medline doi:10.1038/nature10842

17. C. Alkan, J. M. Kidd, T. Marques-Bonet, G. Aksay, F. Antonacci, F. Hormozdiari, J. O. Kitzman, C. Baker, M. Malig, O. Mutlu, S. C. Sahinalp, R. A. Gibbs, E. E. Eichler, Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat. Genet.* **41**, 1061–1067 (2009). Medline doi:10.1038/ng.437

18. T. J. Pemberton, D. Absher, M. W. Feldman, R. M. Myers, N. A. Rosenberg, J. Z. Li, Genomic patterns of homozygosity in worldwide human populations. *Am. J. Hum. Genet.* **91**, 275–292 (2012). Medline doi:10.1016/j.ajhg.2012.06.014

19. K. Prüfer, F. Racimo, N. Patterson, F. Jay, S. Sankararaman, S. Sawyer, A. Heinze, G. Renaud, P. H. Sudmant, C. de Filippo, H. Li, S. Mallick, M. Dannemann, Q. Fu, M. Kircher, M. Kuhlwilm, M. Lachmann, M. Meyer, M. Ongyerth, M. Siebauer, C. Theunert, A. Tandon, P. Moorjani, J. Pickrell, J. C. Mullikin, S. H. Vohr, R. E. Green, I. Hellmann, P. L. Johnson, H. Blanche, H. Cann, J. O. Kitzman, J. Shendure, E. E. Eichler, E. S. Lein, T. E. Bakken, L. V. Golovanova, V. B. Doronichev, M. V. Shunkov, A. P. Derevianko, B. Viola, M. Slatkin, D. Reich, J. Kelso, S. Pääbo, The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43–49 (2014). Medline doi:10.1038/nature12886

20. H. Li, R. Durbin, Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496 (2011). Medline doi:10.1038/nature10231

21. D. Anhuf, M.-P. Ledru, H. Behling, F. W. Da Cruz Jr., R. C. Cordeiro, T. Van der Hammen, I. Karmann, J. A. Marengo, P. E. De Oliveira, L. Pessenda, A. Siffedine, A. L. Albuquerque, P. L. Da Silva Dias, Paleo-environmental change in Amazonian and African rainforest during the LGM. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **239**, 510–527 (2006). doi:10.1016/j.palaeo.2006.01.017

22. S. Glémin, How are deleterious mutations purged? Drift versus nonrandom mating. *Evolution* **57**, 2678–2687 (2003). Medline doi:10.1111/j.0014-3820.2003.tb01512.x

23. D. P. Watts, Composition and variability of mountain gorilla diets in the Central Virungas. *Am. J. Primatol.* **7**, 323–356 (1984). doi:10.1002/ajp.1350070403

24. R. McManamon, L. Lowenstine, in *Fowler's Zoo and Wild Animal Medicine* (Elsevier, St. Louis, vol. 7, 2012), pp. 408–415.

25. A. Pusey, M. Wolf, Inbreeding avoidance in animals. *Trends Ecol. Evol.* **11**, 201–206 (1996). doi:10.1016/0169-5347(96)10028-8

26. H. Li, http://arxiv.org/abs/1303.3997 (2013).

27. E. Garrison, G. Marth, http://arxiv.org/abs/1207.3907 (2012).

28. T. Derrien, J. Estellé, S. Marco Sola, D. G. Knowles, E. Raineri, R. Guigó, P. Ribeca, Fast computation and applications of genome mappability. *PLOS ONE* **7**, e30377 (2012). Medline doi:10.1371/journal.pone.0030377

29. D. H. Alexander, J. Novembre, K. Lange, Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009). Medline doi:10.1101/gr.094052.109

30. P. H. Sudmant, J. O. Kitzman, F. Antonacci, C. Alkan, M. Malig, A. Tsalenko, N. Sampas, L. Bruhn, J. Shendure, E. E. Eichler, Diversity of human copy number variation and multicopy genes. *Science* **330**, 641–646 (2010). Medline doi:10.1126/science.1197005

31. P. H. Sudmant, J. Huddleston, C. R. Catacchio, M. Malig, L. W. Hillier, C. Baker, K. Mohajeri, I. Kondova, R. E. Bontrop, S. Persengiev, F. Antonacci, M. Ventura, J. Prado-Martinez, T. Marques-Bonet, E. E. Eichler, Evolution and diversity of copy number variation in the great ape lineage. *Genome Res.* **23**, 1373–1382 (2013). Medline doi:10.1101/gr.158543.113

32. F. Hach, F. Hormozdiari, C. Alkan, F. Hormozdiari, I. Birol, E. E. Eichler, S. C. Sahinalp, mrsFAST: A cache-oblivious algorithm for short-read mapping. *Nat. Methods* **7**, 576–577 (2010). Medline doi:10.1038/nmeth0810-576

33. H. Eiberg, J. Troelsen, M. Nielsen, A. Mikkelsen, J. Mengel-From, K. W. Kjaer, L. Hansen, Blue eye color in humans may be caused by a perfectly associated founder mutation in a

regulatory element located within the HERC2 gene inhibiting OCA2 expression. *Hum. Genet.* **123**, 177–187 (2008). Medline doi:10.1007/s00439-007-0460-x

34. N. M. Anthony, M. Johnson-Bawe, K. Jeffery, S. L. Clifford, K. A. Abernethy, C. E. Tutin, S. A. Lahm, L. J. White, J. F. Utley, E. J. Wickings, M. W. Bruford, The role of Pleistocene refugia and rivers in shaping gorilla genetic diversity in central Africa. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 20432–20436 (2007). Medline doi:10.1073/pnas.0704816105

35. O. Thalmann, J. Hebler, H. N. Poinar, S. Pääbo, L. Vigilant, Unreliable mtDNA data due to nuclear insertions: A cautionary tale from analysis of humans and other great apes. *Mol. Ecol.* **13**, 321–335 (2004). Medline doi:10.1046/j.1365-294X.2003.02070.x

36. A. Scally, B. Yngvadottir, Y. Xue, Q. Ayub, R. Durbin, C. Tyler-Smith, A genome-wide survey of genetic variation in gorillas using reduced representation sequencing. *PLOS ONE* **8**, e65066 (2013). Medline doi:10.1371/journal.pone.0065066

37. M. A. Quail, T. D. Otto, Y. Gu, S. R. Harris, T. F. Skelly, J. A. McQuillan, H. P. Swerdlow, S. O. Oyola, Optimal enzymes for amplifying sequencing libraries. *Nat. Methods* **9**, 10–11 (2012). Medline doi:10.1038/nmeth.1814

38. D. R. Zerbino, E. Birney, Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821–829 (2008). Medline doi:10.1101/gr.074492.107

39. K. Tamura, D. Peterson, N. Peterson, G. Stecher, M. Nei, S. Kumar, MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* **28**, 2731–2739 (2011). Medline doi:10.1093/molbev/msr121

40. S. L. Clifford, N. M. Anthony, M. Bawe-Johnson, K. A. Abernethy, C. E. Tutin, L. J. White, M. Bermejo, M. L. Goldsmith, K. McFarland, K. J. Jeffery, M. W. Bruford, E. J. Wickings, Mitochondrial DNA phylogeography of western lowland gorillas (*Gorilla gorilla gorilla*). *Mol. Ecol.* **13**, 1551–1565 (2004). Medline doi:10.1111/j.1365-294X.2004.02140.x

41. R. Noda, C. G. Kim, O. Takenaka, R. E. Ferrell, T. Tanoue, I. Hayasaka, S. Ueda, T. Ishida, N. Saitou, Mitochondrial 16S rRNA sequence diversity of hominoids. *J. Hered.* **92**, 490–496 (2001). Medline doi:10.1093/jhered/92.6.490

42. A. J. Drummond, A. Rambaut, BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7**, 214 (2007). doi:10.1186/1471-2148-7-214

43. W. Wei, Q. Ayub, Y. Chen, S. McCarthy, Y. Hou, I. Carbone, Y. Xue, C. Tyler-Smith, A calibrated human Y-chromosomal phylogeny based on resequencing. *Genome Res.* **23**, 388–395 (2013). Medline doi:10.1101/gr.143198.112

44. P. Soares, L. Ermini, N. Thomson, M. Mormina, T. Rito, A. Röhl, A. Salas, S. Oppenheimer, V. Macaulay, M. B. Richards, Correcting for purifying selection: An improved human mitochondrial molecular clock. *Am. J. Hum. Genet.* **84**, 740–759 (2009). Medline doi:10.1016/j.ajhg.2009.05.001

45. Y. Xue, Q. Wang, Q. Long, B. L. Ng, H. Swerdlow, J. Burton, C. Skuce, R. Taylor, Z. Abdellah, Y. Zhao, D. G. MacArthur, M. A. Quail, N. P. Carter, H. Yang, C. Tyler-Smith; Asan, Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree. *Curr. Biol.* **19**, 1453–1457 (2009). Medline doi:10.1016/j.cub.2009.07.032

46. M. I. Douadi, S. Gatti, F. Levrero, G. Duhamel, M. Bermejo, D. Vallet, N. Menard, E. J. Petit, Sex-biased dispersal in western lowland gorillas (*Gorilla gorilla gorilla*). *Mol. Ecol.* **16**, 2247–2259 (2007). Medline doi:10.1111/j.1365-294X.2007.03286.x

47. M. Slatkin, Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. *Nat. Rev. Genet.* **9**, 477–485 (2008). Medline doi:10.1038/nrg2361

48. K. Palin, H. Campbell, A. F. Wright, J. F. Wilson, R. Durbin, Identity-by-descent-based phasing and imputation in founder populations using graphical models. *Genet. Epidemiol.* **35**, 853–860 (2011). Medline doi:10.1002/gepi.20635

49. J. O'Connell, D. Gurdasani, O. Delaneau, N. Pirastu, S. Ulivi, M. Cocca, M. Traglia, J. Huang, J. E. Huffman, I. Rudan, R. McQuillan, R. M. Fraser, H. Campbell, O. Polasek, G. Asiki, K. Ekoru, C. Hayward, A. F. Wright, V. Vitart, P. Navarro, J. F. Zagury, J. F. Wilson, D. Toniolo, P. Gasparini, N. Soranzo, M. S. Sandhu, J. Marchini, A general approach for haplotype phasing across the full spectrum of relatedness. *PLOS Genet.* **10**, e1004234 (2014). Medline

50. A. Manichaikul, J. C. Mychaleckyj, S. S. Rich, K. Daly, M. Sale, W. M. Chen, Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010). Medline doi:10.1093/bioinformatics/btq559

51. A. Scally, R. Durbin, Revising the human mutation rate: Implications for understanding human evolution. *Nat. Rev. Genet.* **13**, 745–753 (2012). Medline doi:10.1038/nrg3295

52. O. Venn, I. Turner, I. Mathieson, N. De Groot, R. Bontrop, G. McVean, Strong male bias drives germline mutation in chimpanzees. *Science* **344**, 1272–1275 (2014). doi:10.1126/science.344.6189.1272

53. K. E. Langergraber, K. Prüfer, C. Rowney, C. Boesch, C. Crockford, K. Fawcett, E. Inoue, M. Inoue-Muruyama, J. C. Mitani, M. N. Muller, M. M. Robbins, G. Schubert, T. S. Stoinski, B. Viola, D. Watts, R. M. Wittig, R. W. Wrangham, K. Zuberbühler, S. Pääbo, L. Vigilant, Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 15716–15721 (2012). doi:10.1073/pnas.1211740109

54. A. H. Freedman, I. Gronau, R. M. Schweizer, D. Ortega-Del Vecchyo, E. Han, P. M. Silva, M. Galaverni, Z. Fan, P. Marx, B. Lorente-Galdos, H. Beale, O. Ramirez, F. Hormozdiari, C. Alkan, C. Vilà, K. Squire, E. Geffen, J. Kusak, A. R. Boyko, H. G. Parker, C. Lee, V. Tadigotla, A. Siepel, C. D. Bustamante, T. T. Harkins, S. F. Nelson, E. A. Ostrander, T. Marques-Bonet, R. K. Wayne, J. Novembre, Genome sequencing highlights the dynamic early history of dogs. *PLOS Genet.* **10**, e1004016 (2014). Medline doi:10.1371/journal.pgen.1004016

55. J. Runge, in *Dynamics of Forest Ecosystems in Central Africa During the Holocene*, J. Runge, Ed. (CRC Press, Boca Raton, FL, 2008), pp. 15–27.

56. R. E. Green, J. Krause, A. W. Briggs, T. Maricic, U. Stenzel, M. Kircher, N. Patterson, H. Li, W. Zhai, M. H. Fritz, N. F. Hansen, E. Y. Durand, A. S. Malaspinas, J. D. Jensen, T. Marques-Bonet, C. Alkan, K. Prüfer, M. Meyer, H. A. Burbano, J. M. Good, R. Schultz, A. Aximu-Petri, A. Butthof, B. Höber, B. Höffner, M. Siegemund, A. Weihmann, C. Nusbaum, E. S. Lander, C. Russ, N. Novod, J. Affourtit, M. Egholm, C. Verna, P. Rudan, D. Brajkovic, Z. Kucan, I. Gusic, V. B. Doronichev, L. V. Golovanova, C. Lalueza-Fox,

M. de la Rasilla, J. Fortea, A. Rosas, R. W. Schmitz, P. L. Johnson, E. E. Eichler, D. Falush, E. Birney, J. C. Mullikin, M. Slatkin, R. Nielsen, J. Kelso, M. Lachmann, D. Reich, S. Pääbo, A draft sequence of the Neandertal genome. *Science* **328**, 710–722 (2010). Medline doi:10.1126/science.1188021

57. Y. Fitzpatrick, Cover stories: Abstract ideas. *Science* **338**, 1–10 (2012). doi:10.1126/science.338.6103.1

58. N. Patterson, P. Moorjani, Y. Luo, S. Mallick, N. Rohland, Y. Zhan, T. Genschoreck, T. Webster, D. Reich, Ancient admixture in human history. *Genetics* **192**, 1065–1093 (2012). Medline doi:10.1534/genetics.112.145037

59. W. McLaren, B. Pritchard, D. Rios, Y. Chen, P. Flicek, F. Cunningham, Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**, 2069–2070 (2010). Medline doi:10.1093/bioinformatics/btq330

60. V. Colonna, Q. Ayub, Y. Chen, L. Pagani, P. Luisi, M. Pybus, E. Garrison, Y. Xue, C. Tyler-Smith, G. R. Abecasis, A. Auton, L. D. Brooks, M. A. DePristo, R. M. Durbin, R. E. Handsaker, H. M. Kang, G. T. Marth, G. A. McVean, Human genomic regions with exceptionally high levels of population differentiation identified from 911 whole-genome sequences. *Genome Biol.* **15**, R88 (2014). Medline doi:10.1186/gb-2014-15-6-r88

61. E. Huerta-Sánchez, M. Degiorgio, L. Pagani, A. Tarekegn, R. Ekong, T. Antao, A. Cardona, H. E. Montgomery, G. L. Cavalleri, P. A. Robbins, M. E. Weale, N. Bradman, E. Bekele, T. Kivisild, C. Tyler-Smith, R. Nielsen, Genetic signatures reveal high-altitude adaptation in a set of Ethiopian populations. *Mol. Biol. Evol.* **30**, 1877–1888 (2013). Medline doi:10.1093/molbev/mst089

62. W. Zhang, Z. Fan, E. Han, R. Hou, L. Zhang, M. Galaverni, J. Huang, H. Liu, P. Silva, P. Li, J. P. Pollinger, L. Du, X. Zhang, B. Yue, R. K. Wayne, Z. Zhang, Hypoxia adaptations in the grey wolf (*Canis lupus chanco*) from Qinghai-Tibet Plateau. *PLOS Genet.* **10**, e1004466 (2014). Medline doi:10.1371/journal.pgen.1004466