

Insights into hominid evolution from the gorilla genome sequence

Aylwyn Scally¹, Julien Y. Dutheil^{2†}, LaDeana W. Hillier³, Gregory E. Jordan⁴, Ian Goodhead^{1†}, Javier Herrero⁴, Asger Hobolth², Tuuli Lappalainen⁵, Thomas Mailund², Tomas Marques-Bonet^{3,6,7}, Shane McCarthy¹, Stephen H. Montgomery⁸, Petra C. Schwalie⁴, Y. Amy Tang¹, Michelle C. Ward^{9,10}, Yali Xue¹, Bryndis Yngvadottir^{1†}, Can Alkan^{3,11}, Lars N. Andersen², Qasim Ayub¹, Edward V. Ball¹², Kathryn Beal⁴, Brenda J. Bradley^{8,13}, Yuan Chen¹, Chris M. Clee¹, Stephen Fitzgerald⁴, Tina A. Graves¹⁴, Yong Gu¹, Paul Heath¹, Andreas Heger¹⁵, Emre Karakoc³, Anja Kolb-Kokocinski¹, Gavin K. Laird¹, Gerton Lunter¹⁶, Stephen Meader¹⁵, Matthew Mort¹², James C. Mullikin¹⁷, Kasper Munch², Timothy D. O'Connor⁸, Andrew D. Phillips¹², Javier Prado-Martinez⁶, Anthony S. Rogers^{1†}, Saba Sajjadian³, Dominic Schmidt^{9,10}, Katy Shaw¹², Jared T. Simpson¹, Peter D. Stenson¹², Daniel J. Turner^{1†}, Linda Vigilant¹⁸, Albert J. Vilella⁴, Weldon Whitener¹, Baoli Zhu^{19†}, David N. Cooper¹², Pieter de Jong¹⁹, Emmanouil T. Dermitzakis⁵, Evan E. Eichler^{3,11}, Paul Flicek⁴, Nick Goldman⁴, Nicholas I. Mundy⁸, Zemin Ning¹, Duncan T. Odum^{1,9,10}, Chris P. Ponting¹⁵, Michael A. Quail¹, Oliver A. Ryder²⁰, Stephen M. Searle¹, Wesley C. Warren¹⁴, Richard K. Wilson¹⁴, Mikkel H. Schierup², Jane Rogers^{1†}, Chris Tyler-Smith¹ & Richard Durbin¹

Gorillas are humans' closest living relatives after chimpanzees, and are of comparable importance for the study of human origins and evolution. Here we present the assembly and analysis of a genome sequence for the western lowland gorilla, and compare the whole genomes of all extant great ape genera. We propose a synthesis of genetic and fossil evidence consistent with placing the human–chimpanzee and human–chimpanzee–gorilla speciation events at approximately 6 and 10 million years ago. In 30% of the genome, gorilla is closer to human or chimpanzee than the latter are to each other; this is rarer around coding genes, indicating pervasive selection throughout great ape evolution, and has functional consequences in gene expression. A comparison of protein coding genes reveals approximately 500 genes showing accelerated evolution on each of the gorilla, human and chimpanzee lineages, and evidence for parallel acceleration, particularly of genes involved in hearing. We also compare the western and eastern gorilla species, estimating an average sequence divergence time 1.75 million years ago, but with evidence for more recent genetic exchange and a population bottleneck in the eastern species. The use of the genome sequence in these and future analyses will promote a deeper understanding of great ape biology and evolution.

Humans share many elements of their anatomy and physiology with both gorillas and chimpanzees, and our similarity to these species was emphasized by Darwin and Huxley in the first evolutionary accounts of human origins¹. Molecular studies confirmed that we are closer to the African apes than to orang-utans, and on average closer to chimpanzees than gorillas² (Fig. 1a). Subsequent analyses have explored functional differences between the great apes and their relevance to human evolution, assisted recently by reference genome sequences for chimpanzee³ and orang-utan⁴. Here we provide a reference assembly and initial analysis of the gorilla genome sequence, establishing a foundation for the further study of great ape evolution and genetics.

Recent technological developments have substantially reduced the costs of sequencing, but the assembly of a whole vertebrate genome

remains a challenging computational problem. We generated a reference assembly from a single female western lowland gorilla (*Gorilla gorilla gorilla*) named Kamilah, using 5.4×10^9 base pairs (5.4 Gbp) of capillary sequence combined with 166.8 Gbp of Illumina read pairs (Methods Summary). Genes, transcripts and predictions of gene orthologues and paralogues were annotated by Ensembl⁵, and additional analysis found evidence for 498 functional long (>200-bp) intergenic RNA transcripts. Table 1 summarizes the assembly and annotation properties. An assessment of assembly quality using finished fosmid sequences found that typical (N50; see Table 1 for definition) stretches of error-free sequence are 7.2 kbp in length, with errors tending to be clustered in repetitive regions. Outside repeat masked regions and away from contig ends, the total

¹Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton CB10 1SA, UK. ²Bioinformatics Research Center, Aarhus University, C.F. Møllers Allé 8, 8000 Aarhus C, Denmark. ³Department of Genome Sciences, University of Washington School of Medicine, Seattle, Washington 98195, USA. ⁴European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton CB10 1SD, UK.

⁵Department of Genetic Medicine and Development, University of Geneva Medical School, Rue Michel-Servet 1, 1211 Geneva 4, Switzerland. ⁶Institut de Biologia Evolutiva (UPF-CSIC), 08003 Barcelona, Catalonia, Spain. ⁷Instituto Catalana de Recerca i Estudis Avançats, ICREA, 08010 Barcelona, Spain. ⁸Department of Zoology, University of Cambridge, Downing Street, Cambridge CB2 3EJ, UK. ⁹University of Cambridge, Department of Oncology, Hutchison/MRC Research Centre, Hills Road, Cambridge CB2 0XZ, UK. ¹⁰Cancer Research UK, Cambridge Research Institute, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, UK. ¹¹Howard Hughes Medical Institute, University of Washington, Seattle, Washington 20815-6789, USA. ¹²Institute of Medical Genetics, Cardiff University, Heath Park, Cardiff CF14 4XN, UK. ¹³Department of Anthropology, Yale University, 10 Sachem Street, New Haven, Connecticut 06511, USA. ¹⁴The Genome Institute at Washington University, Washington University School of Medicine, Saint Louis, Missouri 63108, USA. ¹⁵MRC Functional Genomics Unit, University of Oxford, Department of Physiology, Anatomy and Genetics, South Parks Road, Oxford OX1 3QX, UK. ¹⁶Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, UK. ¹⁷Comparative Genomics Unit, Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland, 20892-2152, USA. ¹⁸Max Planck Institute for Evolutionary Anthropology, Primatology Department, Deutscher Platz 6, Leipzig 04103, Germany. ¹⁹Children's Hospital Oakland Research Institute, Oakland, California 94609, USA. ²⁰San Diego Zoo's Institute for Conservation Research, Escondido, California 92027, USA. [†]Present addresses: Institut des Sciences de l'Évolution – Montpellier (I.S.E.-M.), Université de Montpellier II – CC 064, 34095 Montpellier Cedex 05, France (J.Y.D.); Centre for Genomic Research, Institute of Integrative Biology, University of Liverpool, Crown Street, Liverpool L69 7ZB, UK (I.G.); Division of Biological Anthropology, University of Cambridge, Fitzwilliam Street, Cambridge CB2 1QH, UK (B.Y.); EASIH, University of Cambridge, Addenbrooke's Hospital, Cambridge CB2 0QQ, UK (A.S.R.); Oxford Nanopore Technologies, Edmund Cartwright House, 4 Robert Robinson Avenue, Oxford OX4 4GA, UK (D.J.T.); Institute of Microbiology, Chinese Academy of Sciences, Datun Road, Chaoyang District, Beijing 100101, China (B.Z.); The Genome Analysis Centre, Norwich Research Park, Norwich NR4 7UH, UK (J.R.).

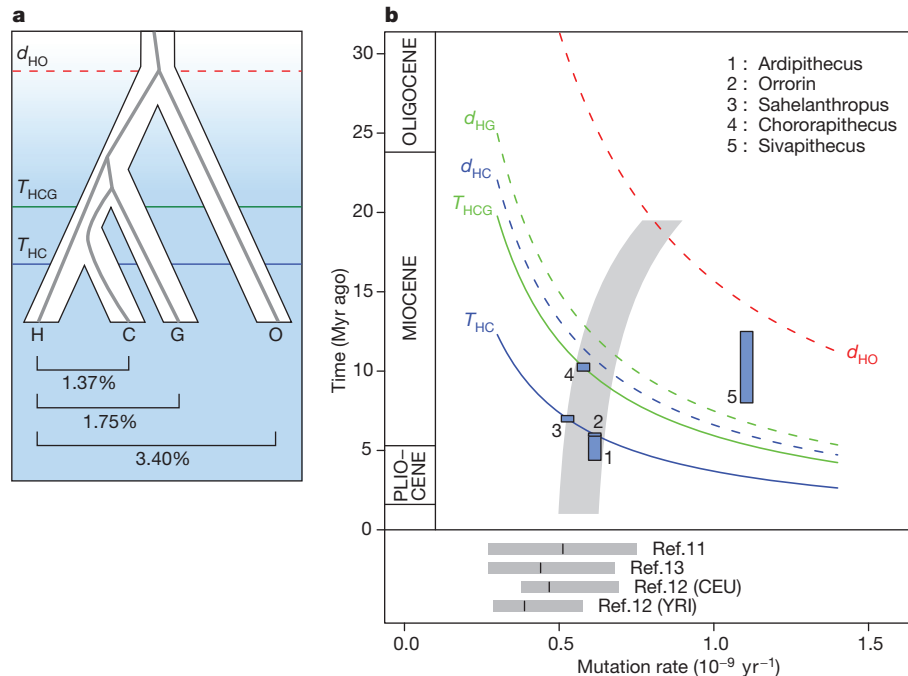


Figure 1 | Speciation of the great apes. **a**, Phylogeny of the great ape family, showing the speciation of human (H), chimpanzee (C), gorilla (G) and orang-utan (O). Horizontal lines indicate speciation times within the hominine subfamily and the sequence divergence time between human and orang-utan. Interior grey lines illustrate an example of incomplete lineage sorting at a particular genetic locus—in this case ((C, G), H), O) rather than (((H, C), G), O). Below are mean nucleotide divergences between human and the other great apes from the EPO alignment. **b**, Great ape speciation and divergence times. Upper panel, solid lines show how times for the HC and HCG speciation events estimated by CoalHMM vary with average mutation rate; dashed lines show the corresponding average sequence divergence times, as well as the HO sequence divergence. Blue blocks represent

rate of single-base and indel errors is 0.13 per kbp. See Supplementary Information for further details.

We also collected less extensive sequence data for three other gorillas, to enable a comparison of species within the *Gorilla* genus. Gorillas survive today only within several isolated and endangered populations whose evolutionary relationships are uncertain. In addition to Kamilah, our analysis included two western lowland gorillas, Kwanza (male) and EB(JC) (female), and one eastern lowland gorilla, Mukisi (male).

Speciation of the great apes

We included the Kamilah assembly with human, chimpanzee (*Pan troglodytes*), orang-utan (*Pongo abelii*) and macaque (*Macaca mulatta*) in a five-way whole-genome alignment using the Ensembl EPO pipeline⁶ (Supplementary Table 3.2). Filtering out low-quality regions of the chimpanzee assembly and regions with many alignment gaps, we obtained 2.01 Gbp of 1:1:1:1 great ape orthologous alignment blocks, to which we then applied a coalescent inference model,

Table 1 | Assembly and annotation statistics

Assembly	Annotation	
Total length	3,041,976,159 bp	Protein-coding genes 20,962
Contigs	465,847	Pseudogenes 1,553
Total contig length	2,829,670,843 bp	RNA genes 6,701
Placed contig length	2,712,844,129 bp	Gene exons 237,216
Unplaced contig length	116,826,714 bp	Gene transcripts 35,727
Max. contig length	191,556 bp	lincRNA transcripts 498
Contig N50	11.8 kbp	
Scaffolds	22,164	
Max. scaffold length	10,247,101 bp	
Scaffold N50	914 kbp	

N50: 50% of the genome is in fragments of this length or longer; lincRNA: long intergenic non-coding RNA.

hominid fossil species (key at top right): each has a vertical extent spanning the range of dates estimated for it in the literature^{9,50}, and a horizontal position at the maximum mutation rate consistent both with its proposed phylogenetic position and the CoalHMM estimates (including some allowance for ancestral polymorphism in the case of *Sivapithecus*). The grey shaded region shows that an increase in mutation rate going back in time can accommodate present-day estimates, fossil hypotheses, and a middle Miocene speciation for orang-utan. Lower panel, estimates of the average mutation rate in present-day humans^{11–13}; grey bars show 95% confidence intervals, with black lines at the means. Estimates were made by the 1000 Genomes Project for trios of European (CEU) and Yoruban African (YRI) ancestry.

CoalHMM, to estimate the timescales and population sizes involved in the speciation of the hominines (African great apes; see Supplementary Table 1.1 for terminology), with orang-utan as an outgroup (Supplementary Information).

Two issues need to be addressed in interpreting the results from CoalHMM (Supplementary Table 4.2). First, the results themselves are obtained in units of sequence divergence rather than years, and so need to be scaled by an appropriate yearly mutation rate. Second, as with any model, CoalHMM makes several simplifying assumptions whose consequences we need to understand in the context of realistic demography. We discuss these issues in turn.

Using a rate of 10^{-9} mutations per bp per year, derived from fossil calibration of the human–macaque sequence divergence and as used in previous calculations, CoalHMM's results would correspond to speciation time estimates T_{HC} (for human–chimpanzee) and T_{HCG} (for human–chimpanzee–gorilla) of 3.7 and 5.95 Myr ago, respectively (Fig. 1b). These dates are consistent with other recent molecular estimates^{7,8}, but are at variance with certain aspects of the fossil record, including several fossils which have been proposed—though not universally accepted⁹—to be hominins, and therefore to postdate the human–chimpanzee split (Fig. 1b). Indeed, the relationship between molecular and fossil evidence has remained difficult to resolve despite the accumulation of genetic data¹⁰. Direct estimates of the per-generation mutation rate in modern human populations, based on the incidence of disease-causing mutations¹¹ or sequencing of familial trios^{12,13}, indicate that a lower value of $(0.5–0.6) \times 10^{-9} \text{ bp}^{-1} \text{ yr}^{-1}$ is plausible (based on average hominine generation times of 20–25 yr). This would give substantially older estimates of approximately 6 and 10 Myr ago for T_{HC} and T_{HCG} , potentially in better agreement with the fossil record.

However, this timetable for hominine speciation must also be reconciled with older events, such as the speciation of orang-utan, which is thought to have occurred no earlier than the Middle Miocene (12–16 Myr ago), as fossil apes before that differ substantially from what we might expect of an early great ape¹⁴. This is possible if we allow for mutation rates changing over time, with a mutation rate of around $1 \times 10^{-9} \text{ bp}^{-1} \text{ yr}^{-1}$ in the common ancestor of great apes, decreasing to lower values in all extant species (Fig. 1b). Comparable changes in mutation rate have been observed previously in primate evolution on larger timescales, including an approximately 30% branch length decrease in humans compared to baboons since their common ancestor¹⁵. A decrease within the great apes is also a predicted consequence of the observed increase in body sizes over this time period and the association of small size with shorter generation times in other primates¹⁶, and is consistent with deviations from a molecular clock seen in sequence divergences of the great apes and macaque (Supplementary Table 3.3). We discuss these and other constraints on estimates of great ape speciation times in the Supplementary Information. However we note that *Sahelanthropus* and *Chororapithecus* remain difficult to incorporate in this model, and can be accommodated as hominin and gorillin genera only if most of the decrease occurred early in great ape evolution.

An alternative explanation for the apparent discrepancy in fossil and genetic dates (leaving aside the issue of whether fossil taxa have been correctly placed) is that ancestral demography may have affected the genetic inferences. Certainly CoalHMM's model does not fit the data in all respects. Perhaps most importantly, it assumes that ancestral population sizes are constant in time and that no gene flow occurred between separated populations, approximations that may not hold in reality. Simulations (details in Supplementary Information) suggest that an ancestral population bottleneck would have had limited impact on the inference of T_{HC} , its influence being captured largely by changes in the model's effective population size. Under conditions of genetic exchange between populations after the main separation of the chimpanzee and human lineages, the speciation time estimated by CoalHMM represents an average weighted by gene flow over the period of separation. This means in some cases it can be substantially older than the date of most recent exchange. However it would only be more recent than the speciation time inferred from fossils if there had been strong gene flow between populations after the development of derived fossil characteristics. To the extent that this is plausible, for example as part of a non-allopatric speciation process, it constitutes an alternative explanation for the dating discrepancy without requiring a change in mutation rate.

In summary, although whole-genome comparisons can be strongly conclusive about the ordering of speciation events, the inability to observe past mutation rates means that the timing of events from genetic data remains uncertain. In our view, possible variation in mutation rates allows hominid genomic data to be consistent with values of T_{HC} from 5.5 to 7 Myr ago and T_{HCG} from 8.5 to 12 Myr ago, with ancestral demographic structure potentially adding inherent ambiguity to both events. Better resolution may come from further integrated analysis of fossil and genetic evidence.

Incomplete lineage sorting and selection

The genealogy relating human (H), chimpanzee (C) and gorilla (G) varies between loci across the genome. CoalHMM explicitly models this and infers the genealogy at each position: either the standard ((H,C),G) relationship or the alternatives ((H,G),C) or ((C,G),H), which are the consequences of incomplete lineage sorting (ILS) in the ancestral human–chimpanzee population. We can use the pattern of ILS to explore evolutionary forces during the human–chimpanzee–gorilla speciation period. Across the genome we find 30% of bases exhibiting ILS, with no significant difference between the number sorting as ((H,G),C) and ((C,G),H). However, the fraction of ILS varies with respect to genomic position (Fig. 2a) by more than

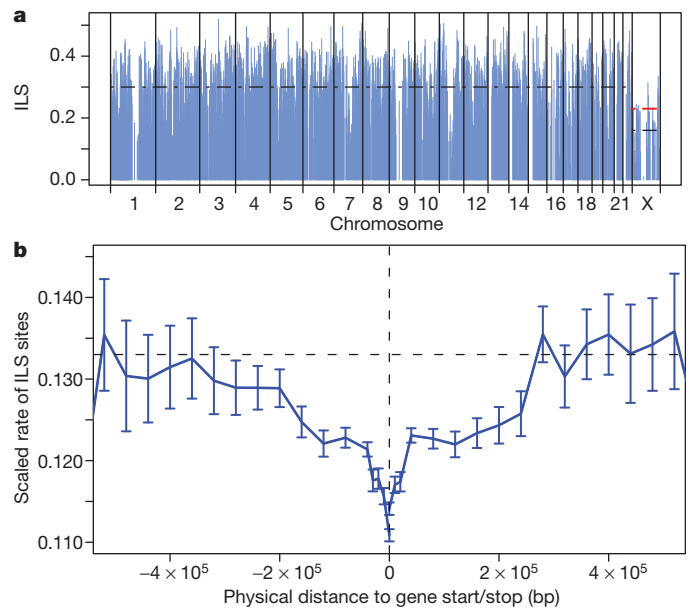


Figure 2 | Genome-wide incomplete lineage sorting (ILS) and selection. **a**, Variation in ILS. Each vertical blue line represents the fraction of ILS between human, chimpanzee and gorilla estimated in a 1-Mbp region. Dashed black lines show the average ILS across the autosomes and on X; the red line shows the expected ILS on X, given the autosomal average and assuming neutral evolution. **b**, Reduction in ILS around protein coding genes. The blue line shows the mean rate of ILS sites normalized by mutation rate as a function of distance upstream or downstream of the nearest gene (see Supplementary Information). The horizontal dashed line indicates the average value outside 300 kbp from the nearest gene; error bars are s.e.m.

expected under a model of genome-wide neutral evolution (Supplementary Fig. 5.1). This variation reflects local differences in the ancestral effective population size N_e during the period between the gorilla and chimpanzee speciation events, most probably due to natural selection reducing N_e and making ILS less likely. Within coding exons mean ILS drops to 22%, and the suppression of ILS extends out to several hundred kbp from coding genes, evident even in raw site patterns before any model inference (Fig. 2b). An analysis of ILS sites in human segmental duplications suggests that assembly errors do not contribute significantly to this signal (Supplementary Information). We therefore attribute it to the effects of linkage around selected mutations, most probably in the form of background selection¹⁷, observing that it is greater around genes with lower ratios of non-synonymous to synonymous mutation rates (dN/dS) (Supplementary Fig. 8.4). Given that more than 90% of the genome lies within 300 kbp of a coding gene, and noting the similar phenomenon reported for recent human evolution¹², this supports the suggestion that selection has affected almost all of the genome throughout hominid evolution¹⁸.

In fitting the transitions between genealogies along the alignment, CoalHMM also estimates a regional recombination rate. This is primarily sensitive to ancestral crossover events before human–chimpanzee speciation, yet despite the expectation of rapid turnover in recombination hotspots¹⁹, averaged over 1-Mbp windows there is a good correlation with estimates from present-day crossovers in humans ($R = 0.49$; $P < 10^{-13}$; Supplementary Fig. 5.5), consistent with the conservation of recombination rates between humans and chimpanzees on the 1-Mbp scale¹⁹.

As expected, we see reduced ILS (Fig. 2a) and human–chimpanzee sequence divergence d_{HC} (Supplementary Fig. 6.1) on the X chromosome, corresponding to a difference in N_e between X and the autosomes within the ancestral human–chimpanzee population. Several factors can contribute to this difference²⁰, notably the X chromosome's haploidy in males, which reduces N_e on X by 0.75, enhances purifying selection in males, and reduces the recombination rate, thereby increasing the

effect of selection via linkage. However, sequence divergence is additionally affected by the mutation rate, which is higher in males than in females, further reducing the relative divergence observed on X²¹. Incorporating the ancestral N_e estimates from CoalHMM, we estimate a ratio of 0.87 ± 0.09 between average mutation rates on X and the autosomes on the human–chimpanzee lineage, corresponding to a male/female mutation rate bias $\alpha = 2.3 \pm 0.4$ (details in Supplementary Information). Previous estimates of α in hominids have ranged from 2 to 7 (refs 22, 23). It is possible that some of the higher values, having been estimated from sequence divergence only and in smaller data sets, were inflated by underestimating the suppression of ancestral N_e on X, in particular due to purifying selection.

Our calculation of α assumes that a single speciation time applies across the genome, attributing differences between the X chromosome and autosomes to the factors mentioned above. An alternative model has been proposed²⁴, involving complex speciation, with more recent human–chimpanzee ancestry on X than elsewhere. Given potential confounding factors in demography, selection, mutation rate bias and admixture, our analyses do not discriminate between these models; however if the effective human–chimpanzee separation time on X is indeed reduced in this way it would imply a still lower value of α .

Functional sequence evolution

We looked for loss or gain of unique autosomal sequence within humans, chimpanzees and gorillas by comparing raw sequence data for each in the context of their reference assemblies (Supplementary Information). The total amount is small: 3–7 Mbp per species, distributed genome-wide in fragments no more than a few kbp in length (Supplementary Table 7.1). The vast majority (97%) of such material was also found either in orang-utan or a more distant primate, indicating loss, and consistent with the expectation that gain is driven primarily by duplication (which our analysis excludes). Some fragments found only in one species overlap coding exons in annotated genes: 6 genes in human, 5 in chimpanzee and 9 in gorilla (Supplementary Tables 7.2, 7.3, 7.4), the majority being associated with olfactory receptor proteins or other rapidly evolving functions, such as male fertility and immune response.

We did not assemble a gorilla Y chromosome, but by mapping $\sim 6\times$ reads from the male gorillas Kwanza and Mukisi to the human Y, we identified several regions in which human single-copy material is missing in gorilla, comprising almost 10% of the accessible male-specific region. Across the Y chromosome there is considerable variation in the copy number of shared material, and the pattern of coverage is quite different from that of reads from a male bonobo mapped in the same way (Supplementary Fig. 7.1). Some missing or depleted material overlaps coding genes (Supplementary Table 7.5), including for example *VCY*, a gene expressed specifically in male germ cells which has two copies in human and chimpanzee but apparently only one in gorilla (Supplementary Information). The resulting picture is consistent with rapid structural evolution of the Y chromosome in the great apes, as previously seen in the chimpanzee–human comparison²⁵.

Protein evolution

The Ensembl EPO primate alignment was filtered to produce a high-quality genome-wide set of 11,538 alignments representing orthologous primate coding sequences, which were then scored with codon-based evolutionary models for likelihoods of acceleration or deceleration of dN/dS in the terminal lineages, ancestral branch, and entire hominine subfamily (Supplementary Information). We find that genes with accelerated rates of evolution across hominines are enriched for functions associated with sensory perception, particularly in relation to hearing and brain development (Supplementary Table 8.4g, h). For example, among the most strongly accelerated genes are *OTOF* ($P = 0.0056$), *LOXHD1* ($P < 0.01$) and *GPR98* ($P = 0.0056$), which are all associated with diseases causing human deafness (Supplementary

Table 8.5). *GPR98*, which also shows significant evidence of positive selection under the branch-site test ($P = 0.0081$), is highly expressed in the developing central nervous system. The gene with the strongest evidence for acceleration along the branch leading to hominines is *RNF213* (branch-site $P < 2.9 \times 10^{-9}$), a gene associated with Moyamoya disease in which blood flow to the brain is restricted due to arterial stenosis²⁶. Given that oxygen and glucose consumption scales with total neuron number²⁷, *RNF213* may have played a role in facilitating the evolution of larger brains. Together, these observations are consistent with a major role for adaptive modifications in brain development and sensory perception in hominine evolution.

Turning to lineage-specific selection pressures, we find relatively similar numbers of accelerated genes in humans, chimpanzees and gorillas (663, 562 and 535 respectively at nominal $P < 0.05$, Supplementary Table 8.3a) and genome-wide dN/dS ratios (0.256, 0.249 and 0.239 in purifying sites, Supplementary Table 8.6). These numbers, which reflect variation in historical effective population sizes as well as environmental pressures, reveal a largely uniform landscape of recent hominine gene evolution—in accordance with previously published analyses in human and chimpanzee^{3,28} (Supplementary Table 8.7).

Genes with accelerated rates of evolution along the gorilla lineage are most enriched for a number of developmental terms, including ear, hair follicle, gonad and brain development, and sensory perception of sound. Among the most significantly accelerated genes in gorilla is *EVPL* ($P < 2.2 \times 10^{-5}$), which encodes a component of the cornified envelope of keratinocytes, and may be related to

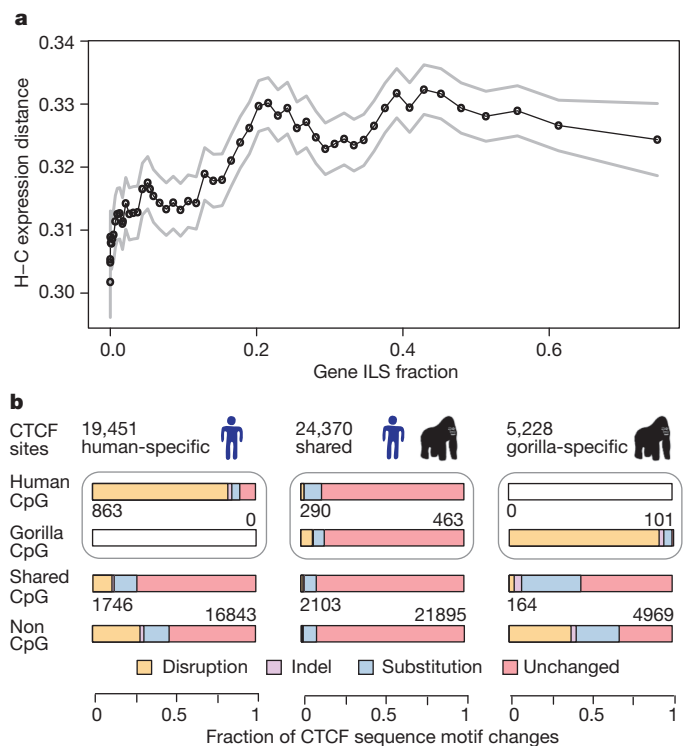


Figure 3 | Differences in expression and regulation. **a**, Mean gene expression distance between human and chimpanzee as a function of the proportion of ILS sites per gene. Each point represents a sliding window of 900 genes (over genes ordered by ILS fraction); s.d. error limits are shown in grey. **b**, Top row, classification of CTCF sites in the gorilla (EB(JC)) and human (GM12878) LCLs on the basis of species-uniqueness; numbers of alignable CTCF binding sites are shown for each category. Bottom three rows, sequence changes of CTCF motifs embedded in human-specific, shared and gorilla-specific CTCF binding sites located within shared CpG islands, species-specific CpG islands or outside CpG islands. Numbers of CTCF binding sites are shown for each CpG island category. Gorilla and human motif sequences are compared and represented as indels, disruptions (>4-bp gaps) and substitutions.

increased cornification of knuckle pads in gorilla²⁹. Interestingly, gorilla and human both yielded brain-associated terms enriched for accelerated genes, but chimpanzee did not (Supplementary Table 8.4a–c). Genes expressed in the brain or involved in its development have not typically been associated with positive selection in primates, but our results show that multiple great ape lineages show elevated dN/dS in brain-related genes when evaluated against a primate background.

We also identified cases of pairwise parallel evolution among hominines. Human and chimpanzee show the largest amount, with significantly more shared accelerations than expected by chance, whereas gorilla shares more parallel acceleration with human than with chimpanzee across a range of significance thresholds (Supplementary Fig. 8.3). Genes involving hearing are enriched in parallel accelerations for all three pairs, but most strongly in gorilla–human (Supplementary Table 8.4d–f), calling into question a previous link made between accelerated evolution of auditory genes in humans and language evolution²⁸. It is also interesting to note that ear morphology is one of the few external traits in which humans are more similar to gorillas than to chimpanzees³⁰.

Next we considered gene loss and gain. We found 84 cases of gene loss in gorilla due to the acquisition of a premature stop codon, requiring there to be no close paralogue (Supplementary Table 8.8): one such gene is *TEX14*, which codes for an intercellular bridge protein essential for spermatogenesis in mice. Genome-wide analysis of gene gain is confounded by the difficulty in assembling closely related paralogues. We therefore resequenced, by finishing overlapping fosmids, three gene clusters known to be under rapid adaptive evolution in primates: the growth hormone cluster³¹, the PRM clusters involved in sperm function and the APOBEC cluster implicated in molecular adaptation to viral defence. In the growth hormone cluster, we observed four chorionic somatomammotropin (CSH) genes in gorilla compared to three in humans and chimpanzees, with a novel highly similar pair of CSH-like genes in gorilla that share a 3' end similar to human growth hormone *GH2*, suggesting a complex evolutionary history as in other primates³¹. We saw sequence but not gene copy number changes in the PRM and APOBEC clusters (Supplementary Information).

In several cases, a protein variant thought to cause inherited disease in humans³² is the only version found in all three gorillas for which we have genome-wide sequence data (Supplementary Table 8.9). Striking examples are the dementia-associated variant Arg432Cys in the growth factor PGRN and the hypertrophic cardiomyopathy-associated variant Arg153His in the muscle Z disk protein TCAP,

both of which were corroborated by additional capillary sequencing (Supplementary Table 8.10). Why variants that appear to cause disease in humans might be associated with a normal phenotype in gorillas is unknown; possible explanations are compensatory molecular changes elsewhere, or differing environmental conditions. Such variants have also been found in both the chimpanzee and macaque genomes^{3,33}.

Gene transcription and regulation

We carried out an analysis of hominine transcriptome variation using total RNA extracted and sequenced from lymphoblastoid cell lines (LCLs) of one gorilla, two chimpanzees and two bonobos (Supplementary Information), and published RNA sequence data for eight human individuals³⁴. After quantifying reads mapping to exons and genes in each species, we calculated the degree of species-specific expression and splicing in 9,746 1:1:1 expressed orthologous genes. On average, expression levels in human and chimpanzee were more similar to each other than either was to gorilla (Supplementary Fig. 10.2). However this effect is reduced in genes with a higher proportion of ILS sites, which tend to show greater expression distance between humans and chimpanzees (Fig. 3a). More generally, patterns seen in the relative expression distances between the three species showed a significant overlap with those derived from genomic lineage sorting ($P = 0.026$; Supplementary Table 10.4), demonstrating that ILS can be reflected in functional differences between primate species.

We also explored species specific variation in splicing³⁵ by calculating the variance in differential expression of orthologous exons within each gene. In total we found 7% of genes whose between-species variance is significant at the 1% level (based on the distribution of within-human variances, Supplementary Fig. 10.5). For example, Supplementary Fig. 10.6 illustrates gorilla-specific splicing in the *SQLE* gene, involved in steroid metabolism.

We further investigated great ape regulatory evolution by comparing the binding in human and gorilla of CTCF, a protein essential to vertebrate development that is involved in transcriptional regulation, chromatin loop formation and protein scaffolding³⁶. We performed ChIP-seq (chromatin immunoprecipitation sequencing) of CTCF in a gorilla LCL (from EB(JC)), and compared this with matched human experiments³⁷, using the EPO alignments to identify species-specific and shared binding regions (Fig. 3b and Supplementary Information). Consistent with previous results reporting strong CTCF binding conservation³⁸, and in contrast to the rapid turnover of some other transcription factor binding sites³⁹, we found that approximately 70% of

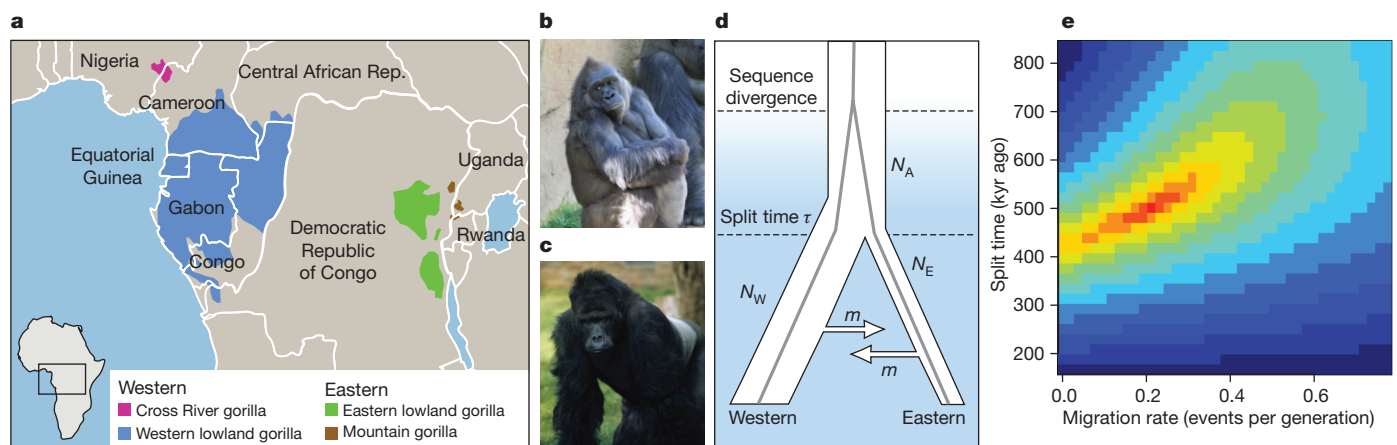


Figure 4 | Gorilla species distribution and divergence. **a**, Distribution of gorilla species in Africa. The western species (*Gorilla gorilla*) comprises two subspecies: western lowland gorillas (*G. gorilla gorilla*) and Cross River gorillas (*G. gorilla diehli*). Similarly, the eastern species (*Gorilla beringei*) is subclassified into eastern lowland gorillas (*G. beringei graueri*) and mountain gorillas (*G. beringei beringei*). (Based on data in ref. 43.) Areas of water are shown pale blue. Inset, area of main map. **b**, Western lowland gorilla Kamilah, source of the

reference assembly (photograph by J.R.). **c**, Eastern lowland gorilla Mukisi (photograph by M. Seres). **d**, Isolation–migration model of the western and eastern species. N_A , N_W and N_E are ancestral, western and eastern effective population sizes; m is the migration rate. **e**, Likelihood surface for migration and split time parameters in the isolation–migration model; colours from blue (minimum) to red (maximum) indicate the magnitude of likelihood.

Table 2 | Nucleotide polymorphism in western and eastern gorillas

Gorilla	Species	Heterozygous site rate (%)	Homozygous site rate (%)	Homozygous:heterozygous ratio
Kamilah	Western lowland	0.189	0.0015	0.008
EB(JC)	Western lowland	0.178	0.10	0.56
Mukisi	Eastern lowland	0.076	0.19	2.5

Rates (columns 3 and 4) are based on variants detected by mapping sequence data to the gorilla reference and filtering sites by depth and mapping quality (Supplementary Information). The homozygosity rate for Kamilah is low (and is effectively an error rate) because the sequence of this gorilla was used for assembly. Reduced heterozygosity in Mukisi is not due to familial inbreeding, as there are no long homozygous stretches.

gorilla CTCF binding regions are shared with human. This compares with around 80% pairwise overlaps between three human LCLs (Supplementary Fig. 11.1a). Binding regions that are shared among all three human individuals are three times more likely to be shared with gorilla than individual-specific regions (Supplementary Fig. 11.1b).

The genomic changes leading to loss of CTCF binding differ between regions within CpG islands and those in the rest of the genome. Losses of CTCF binding outside CpG islands and within species-specific CpG regions co-occur with sequence changes in the binding motif, but for shared CpG islands most binding losses have no corresponding motif sequence change (Fig. 3b). It is possible that DNA methylation differences are driving this effect, as CTCF binding can be abolished by methylation of specific target regions³⁶. Alternatively, CTCF binding within CpG islands may also depend more on other regulators' binding and less on the CTCF motif itself.

Genetic diversity within Gorilla

Recent studies of molecular and morphological diversity within the *Gorilla* genus have supported a classification into two species, eastern (*Gorilla beringei*) and western (*Gorilla gorilla*)⁴⁰, with both species further divided into subspecies (Fig. 4a). Although separated today by over 1,000 km, it has been suggested that gene flow has occurred between the eastern and western species since divergence⁴¹. To investigate this, we collected reduced representation sequence data (Supplementary Information) for another female western lowland gorilla, EB(JC), and a male eastern lowland gorilla, Mukisi.

Table 2 summarizes the sequence diversity in these individuals and in Kamilah, based on alignment of sequence data to the gorilla assembly. The ratio of homozygous to heterozygous variant rates for EB(JC) (close to 0.5) is consistent with this gorilla coming from the same population as Kamilah (Supplementary Information), and the rate of heterozygosity for this gorilla matches Kamilah's. Mukisi, on the other hand, has twice the rate of homozygous differences from the assembly, consistent with this gorilla coming from a separate population. Furthermore, heterozygosity in Mukisi is much lower, suggesting a reduced population size in the eastern species. This agrees with previous studies based on fewer loci⁴¹, and also with estimates of present-day numbers in the wild, which indicate that whereas the western lowland subspecies may number up to 200,000 individuals, the eastern population as a whole is around ten times smaller^{42,43}. Because it manifests in genetic diversity, this disparity must have existed for many millennia, and cannot have resulted solely from the current pressure of human activity in central Africa or recent outbreaks of the Ebola virus.

On the basis of an alignment of the EB(JC) and Mukisi data to the human reference sequence and comparing high confidence genotype calls for the two individuals, we estimate a mean sequence divergence time between them of 1.75 Myr ago. However the pattern of shared heterozygosity is not consistent with a clean split between western and eastern gorillas (Supplementary Information). Under a model which allows symmetric genetic exchange between the populations after an initial split (Fig. 4d; Supplementary Information), the maximum likelihood species split time is ~0.5 Myr ago with moderate subsequent exchange of ~0.2 individuals per generation each way between breeding pools, totalling ~5,000 in each direction over 0.5 Myr (Fig. 4e). Different model assumptions and parameterisations would lead to different values. More extensive sampling and sequencing of both gorilla populations will afford better resolution of this issue.

We also collected whole-genome sequence data from an additional male western lowland gorilla (Kwanza) at 12×, and further whole-genome sequence data for (eastern) Mukisi at 7× (Supplementary Information). Differences between the western gorillas and Mukisi represent a combination of inter-individual and inter-species variants. These include 1,615 non-synonymous SNPs (single nucleotide polymorphisms) in 1,326 genes, seven of which have more than four amino acid differences each (Supplementary Table 12.2), among which are two olfactory receptor genes and *EMR3*, implicated in immune and inflammatory responses⁴⁴. Nineteen of the genes annotated in Kamilah carry an apparently homozygous premature stop codon in Mukisi. These include the gene encoding the seminal fluid protein *SEMG2*, implicated in sperm competition and known to be inactivated in some gorillas, where sperm competition is rare⁴⁵. Both *EMR3* and *SEMG2* were corroborated by additional sequencing (Supplementary Tables 12.3, 12.4).

Finally, we investigated genomic duplication in gorilla using a whole-genome shotgun sequence detection method applied to data from the western lowland gorillas Kamilah and Kwanza (Supplementary Information). This revealed a level of private segmental duplication (0.9 Mbp and 1.5 Mbp in the two gorillas) well outside the range found in pairwise comparisons of humans (Supplementary Fig. 13.1), where a value of ~100 kbp is typical between any two individuals⁴⁶. These results suggest greater copy number diversity in gorillas than in humans, consistent with previous observations in the great apes⁴⁷.

Discussion

Since the middle Miocene—an epoch of abundance and diversity for apes throughout Eurasia and Africa—the prevailing pattern of ape evolution has been one of fragmentation and extinction⁴⁸. The present-day distribution of non-human great apes, existing only as endangered and subdivided populations in equatorial forest refugia⁴³, is a legacy of that process. Even humans, now spread around the world and occupying habitats previously inaccessible to any primate, bear the genetic legacy of past population crises. All other branches of the genus *Homo* have passed into extinction. It may be that in the condition of *Gorilla*, *Pan* and *Pongo* we see some echo of our own ancestors before the last 100,000 years, and perhaps a condition experienced many times over several million years of evolution. It is notable that species within at least three of these genera continued to exchange genetic material long after separation^{4,49}, a disposition that may have aided their survival in the face of diminishing numbers. As well as teaching us about human evolution, the study of the great apes connects us to a time when our existence was more tenuous, and in doing so, highlights the importance of protecting and conserving these remarkable species.

METHODS SUMMARY

Assembly. We constructed a hybrid *de novo* assembly combining 5.4 Gbp of capillary read pairs with the contigs from an initial short read assembly of 166.8 Gbp of Illumina paired reads. Improvements in long-range structure were then guided by human homology, placing contigs into scaffolds wherever read pairs confirmed collinearity between gorilla and human. Base-pair contiguity was improved by local reassembly within each scaffold, merging or extending contigs using Illumina read pairs. Finally we used additional Kamilah bacterial artificial chromosome (BAC) and fosmid end pair capillary sequences to provide longer range scaffolding. Base errors were corrected by mapping all Illumina reads back to the assembly and rectifying apparent homozygous variants, while recording the location of heterozygous sites. Further details and other methods are described in Supplementary Information.

Received 16 June 2011; accepted 10 January 2012.

- Huxley, T. H. *Evidence as to Man's Place in Nature* (Williams & Norgate, 1863).
- King, M. C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees. *Science* **188**, 107–116 (1975).
- Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69–87 (2005).
- Locke, D. P. *et al.* Comparative and demographic analysis of orang-utan genomes. *Nature* **469**, 529–533 (2011).
- Hubbard, T. J. *et al.* Ensembl 2009. *Nucleic Acids Res.* **37**, D690–D697 (2009).
- Paten, B., Herrero, J., Beal, K., Fitzgerald, S. & Birney, E. Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res.* **18**, 1814–1828 (2008).
- Bradley, B. J. Reconstructing phylogenies and phenotypes: a molecular view of human evolution. *J. Anat.* **212**, 337–353 (2008).
- Burgess, R. & Yang, Z. Estimation of hominoid ancestral population sizes under bayesian coalescent models incorporating mutation rate variation and sequencing errors. *Mol. Biol. Evol.* **25**, 1979–1994 (2008).
- Wood, B. & Harrison, T. The evolutionary context of the first hominins. *Nature* **470**, 347–352 (2011).
- Steiper, M. E. & Young, N. M. Timing primate evolution: lessons from the discordance between molecular and paleontological estimates. *Evol. Anthropol.* **17**, 179–188 (2008).
- Lynch, M. Rate, molecular spectrum, and consequences of human mutation. *Proc. Natl Acad. Sci. USA* **107**, 961–968 (2010).
- The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010); correction **473**, 544 (2011).
- Roach, J. C. *et al.* Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328**, 636–639 (2010).
- Hartwig, W. C. *et al.* *The Primate Fossil Record* (Cambridge Univ. Press, 2002).
- Kim, S. H., Elango, N., Warden, C., Vigoda, E. & Yi, S. V. Heterogeneous genomic molecular clocks in primates. *PLoS Genet.* **2**, e163 (2006).
- Fleagle, J. G. *Primate Adaptation and Evolution* 2nd edn (Academic Press, 1998).
- Charlesworth, D., Morgan, M. T. & Charlesworth, B. Mutation accumulation in finite populations. *J. Hered.* **84**, 321–325 (1993).
- McVicker, G., Gordon, D., Davis, C. & Green, P. Widespread genomic signatures of natural selection in hominid evolution. *PLoS Genet.* **5**, e1000471 (2009).
- Myers, S., Bottolo, L., Freeman, C., McVean, G. & Donnelly, P. A fine-scale map of recombination rates and hotspots across the human genome. *Science* **310**, 321–324 (2005).
- Vicoso, B. & Charlesworth, B. Evolution on the X chromosome: unusual patterns and processes. *Nature Rev. Genet.* **7**, 645–653 (2006).
- Ellegren, H. Characteristics, causes and evolutionary consequences of male-biased mutation. *Proc. R. Soc. Lond. B* **274**, 1–10 (2007).
- Goetting-Minesky, M. P. & Makova, K. D. Mammalian male mutation bias: impacts of generation time and regional variation in substitution rates. *J. Mol. Evol.* **63**, 537–544 (2006).
- Presgraves, D. C. & Yi, S. V. Doubts about complex speciation between humans and chimpanzees. *Trends Ecol. Evol.* **24**, 533–540 (2009).
- Patterson, N., Richter, D. J., Gnerre, S., Lander, E. S. & Reich, D. Genetic evidence for complex speciation of humans and chimpanzees. *Nature* **441**, 1103–1108 (2006).
- Hughes, J. F. *et al.* Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content. *Nature* **463**, 536–539 (2010).
- Kamada, F. *et al.* A genome-wide association study identifies RNF213 as the first Moyamoya disease gene. *J. Hum. Genet.* **56**, 34–40 (2011).
- Herculano-Houzel, S. Scaling of brain metabolism with a fixed energy budget per neuron: implications for neuronal activity, plasticity and evolution. *PLoS ONE* **6**, e17514 (2011).
- Clark, A. G. *et al.* Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science* **302**, 1960–1963 (2003).
- Ellis, R. A. & Montagna, W. The skin of primates. VI. The skin of the gorilla (*Gorilla gorilla*). *Am. J. Phys. Anthropol.* **20**, 79–93 (1962).
- Streeter, G. L. Some uniform characteristics of the primate auricle. *Anat. Rec. A* **23**, 335–341 (1922).
- Wallis, O. C., Zhang, Y. P. & Wallis, M. Molecular evolution of GH in primates: characterisation of the GH genes from slow loris and marmoset defines an episode of rapid evolutionary change. *J. Mol. Endocrinol.* **26**, 249–258 (2001).
- Stenson, P. D. *et al.* The Human Gene Mutation Database: 2008 update. *Genome Med.* **1**, 13 (2009).
- Gibbs, R. A. *et al.* Evolutionary and biomedical insights from the rhesus macaque genome. *Science* **316**, 222–234 (2007).
- Montgomery, S. B. *et al.* Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**, 773–777 (2010).
- Blekhman, R., Marioni, J. C., Zumbo, P., Stephens, M. & Gilad, Y. Sex-specific and lineage-specific alternative splicing in primates. *Genome Res.* **20**, 180–189 (2010).
- Phillips, J. E. & Corces, V. G. CTCF: master weaver of the genome. *Cell* **137**, 1194–1211 (2009).
- McDaniell, R. *et al.* Heritable individual-specific and allele-specific chromatin signatures in humans. *Science* **328**, 235–239 (2010).
- Kunarse, G. *et al.* Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nature Genet.* **42**, 631–634 (2010).
- Schmidt, D. *et al.* Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* **328**, 1036–1040 (2010).
- Groves, C. *Primate Taxonomy* (Smithsonian Institution Press, 2001).
- Thalmann, O., Fischer, A., Lankester, F., Paabo, S. & Vigilant, L. The complex evolutionary history of gorillas: insights from genomic data. *Mol. Biol. Evol.* **24**, 146–158 (2007).
- Stokes, E., Malonga, R., Rainey, H. & Strindberg, S. *Western Lowland Gorilla Surveys in Northern Republic of Congo 2006–2007. Summary Scientific Report* (WCS Global Conservation, 2008).
- IUCN. *The IUCN Red List of Threatened Species. Version 2010.1* (<http://www.iucnredlist.org>) (2010).
- Stacey, M., Lin, H. H., Hilyard, K. L., Gordon, S. & McKnight, A. J. Human epidermal growth factor (EGF) module-containing mucin-like hormone receptor 3 is a new member of the EGF-TM7 family that recognizes a ligand on human macrophages and activated neutrophils. *J. Biol. Chem.* **276**, 18863–18870 (2001).
- Jensen-Seaman, M. I. & Li, W. H. Evolution of the hominoid semenogelin genes, the major proteins of ejaculated semen. *J. Mol. Evol.* **57**, 261–270 (2003).
- Alkan, C. *et al.* Personalized copy number and segmental duplication maps using next-generation sequencing. *Nature Genet.* **41**, 1061–1067 (2009).
- Gazave, E. *et al.* Copy number variation analysis in the great apes reveals species-specific patterns of structural variation. *Genome Res.* **21**, 1626–1639 (2011).
- Begun, D. R. in *Handbook of Palaeoanthropology* Vol. 2, *Primate Evolution and Human Origins* (eds Henke, W. & Tattersall, I.) 921–977 (Springer, 2007).
- Green, R. E. *et al.* A draft sequence of the Neandertal genome. *Science* **328**, 710–722 (2010).
- Lebatard, A. E. *et al.* Cosmogenic nuclide dating of *Sahelanthropus tchadensis* and *Australopithecus bahrelghazali*: Mio-Pliocene hominids from Chad. *Proc. Natl Acad. Sci. USA* **105**, 3226–3231 (2008).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank H. Li and E. Birney for discussions, D. Zerbinio, J. Stalker, L. Wilming, D. Rajan and H. Clawson for technical assistance, J. Ahringer for comments on the manuscript, K. Leus of the Center for Research and Conservation of the Royal Zoological Society of Antwerp for sample material from Mukisi, and the Marmoset Genome Analysis Consortium for permission to use the unpublished assembly of the marmoset genome. This research was supported in part by Wellcome Trust grants WT062023 (to J.H., K.B., S.F., A.J.V., P.F.), WT089066 (to R.D.), WT077192 (to R.D., S.M., A.K.-K., J.T.S., W.W.), WT077009 (to Y.X., B.Y., Q.A., Y.C., C.T.-S.), WT077198 (to G.K.L.) and 075491/Z/04 (to G.L.); EMBL grants (to P.C.S., P.F.); scholarships from the Gates Cambridge Trust (to G.E.J. and T.D.O'C.); an MRC Special Fellowship in Biomedical Informatics (to A.S.); funding from the Lundbeck Foundation (to A.H.); the Academy of Finland and the Emil Aaltonen Foundation (to T.L.); a Marie Curie fellowship (to T.M.-B.); the European Community's Seventh Framework Programme (FP7/2007-2013)/ ERC Starting Grant (STG_20091118) (to T.M.-B.); an FPI grant from the Spanish Ministry of Education (BES-2010-032251) (to J.P.-M.); a BBSRC Doctoral Training Grant (to S.H.M.); grants from the UK Medical Research Council (to A.H., S.M., C.P.P.); the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health (to J.C.M.); the Danish Council for Independent Research, Natural Sciences, grant no. 09-062535 (to K.M., M.H.S.); a Commonwealth Scholarship (to M.C.W.); the Swiss National Science Foundation, Louis Jeantet Foundation (to E.T.D.); an ERC Starting Grant and an EMBO Young Investigator Award, Hutchinson Whampoa (to D.T.O.); NHGRI support (to W.C.W.); support from BIOBASE GmbH (to E.V.B., P.D.S., M.M., A.D.P., K.S., D.N.C.); US National Science Foundation grant DGE-0739133 (to W.W.); NHGRI U54 HG003079 (to R.K.W.); NIH grant HG002385 (to E.E.E.). E.E.E. is an investigator of the Howard Hughes Medical Institute.

Author Contributions Manuscript main text: A.S., R.D., C.T.-S., N.I.M., G.E.J., P.C.S., A.K.-K. Project coordination: A.S., A.S.R., A.K.-K., R.D. Project initiation: J.R., R.D., R.K.W. Library preparation and sequencing: I.G., D.J.T., M.A.Q., C.M.C., B.Z., P.d.J., O.A.R., Q.A., B.Y., Y.X., T.A.G., W.C.W. Assembly: A.S., L.W.H., Y.G., J.T.S., J.C.M., W.W., Z.N. Fosmid finishing: P.H. Assembly quality: A.S., S. Meader, G.L., C.P.P. Annotation: Y.A.T., G.K.L., A.J.V., A. Heger, S.M.S. Primate multiple alignments: J.H., K.B., S.F. Great ape speciation and ILS: J.Y.D., A.S., T.M., M.H.S., K.M., G.E.J. Sequence loss and gain: A.S., S.M., C.T.-S., Y.A.T., A.J.V. Protein evolution: G.E.J., S.H.M., N.I.M., B.J.B., T.D.O'C., Y.X., Y.C., N.G. Human disease allele analysis: Y.X., Y.C., C.T.-S., P.D.S., E.V.B., A.D.P., M.M., K.S., D.N.C. Transcriptome analysis: T.L., E.T.D. ChIP-seq experiment and analysis: P.C.S., M.C.W., D.S., P.F., D.T.O. Additional gorilla samples: B.Y., Y.X., L.V., C.T.-S. Gorilla species diversity and divergence: A.S., A.H., T.M., L.N.A., B.Y., L.V. Gorilla species functional differences: Y.X., Y.C., C.T.-S. Segmental duplication analysis: T.M.-B., C.A., S.S., E.K., J.P.-M., E.E.E.

Author Information Accession numbers for all primary sequencing data are given in Supplementary Information. The assembly has been submitted to EMBL with accession numbers FR853080 to FR853106, and annotation is available at Ensembl (http://www.ensembl.org/Gorilla_gorilla/Info/Index). Reprints and distributions information is available at www.nature.com/reprints. This paper is distributed under the terms of the Creative Commons Attribution-NonCommercial-Share Alike licence, and is freely available to all readers at www.nature.com/nature. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to R.D. (rd@sanger.ac.uk).