Evolution and diversity of copy number variation in the great ape lineage

Peter H Sudmant, John Huddleston, Claudia R Catacchio, et al.

Genome Res. published online July 3, 2013
Access the most recent version at doi:10.1101/gr.158543.113

Evolution and diversity of copy number variation in the great ape lineage

Peter H. Sudmant[1], John Huddleston[1,7], Claudia R. Catacchio[2], Maika Malig[1], LaDeana W. Hillier[3], Carl Baker[1], Kiana Mohajeri[1], Ivanela Kondova[4], Ronald E. Bontrop[4], Stephan Persengiev[4], Francesca Antonacci[2], Mario Ventura[2], Javier Prado-Martinez[5], Tomas Marques-Bonet[5,6], and Evan E. Eichler[1,7]

1. Department of Genome Sciences, University of Washington, Seattle, WA, USA
2. University of Bari, Bari, Italy
3. The Genome Institute, Washington University School of Medicine, St. Louis, MO, USA
4. Department of Comparative Genetics, Biomedical Primate Research Centre, Rijswijk, The Netherlands
5. Institut de Biologia Evolutiva, (UPF-CSIC) Barcelona, Spain
6. Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain
7. Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA

Correspondence to:

Evan Eichler
Department of Genome Sciences
University of Washington School of Medicine
Foege S-413A, Box 355065
3720 15th Ave NE
Seattle, WA 98195
E-mail: eee@gs.washington.edu

ABSTRACT

Copy number variation (CNV) contributes to the genetic basis of disease and has significantly restructured the genomes of humans and great apes. The diversity and rate of this process, however, has not been extensively explored among the great ape lineages. We analyzed 97 deeply sequenced great ape and human genomes and estimate that 16% (469 Mbp) of the hominid genome has been affected by recent copy number changes. We identify a comprehensive set of fixed gene deletion (n=340) and duplication (n=405) events as well as more than 13.5 Mbp of genomic sequence that has been specifically lost on the human lineage over the last 16 million years of evolution. We compared the diversity and rates of copy number and single nucleotide variation across different time points of the hominid phylogeny. We find that CNV diversity partially correlates with single nucleotide polymorphism diversity (r^2=0.5) and recapitulates the phylogeny of apes with few exceptions. Duplications significantly outpace deletions (2.8-fold), especially along ancestral African great ape branches. The load of segregating duplications remains significantly higher in bonobos, Western chimpanzees, and Sumatran orangutans—populations that have experienced recent genetic bottlenecks (P=0.0014, 0.02 and 0.0088, respectively). We find that the rate of fixed deletion has been more clocklike with the exception of the chimpanzee lineage where we observe a twofold increase in the chimpanzee-bonobo ancestor (P=4.79 X 10^{-9}) and evidence of increased deletion load among Western chimpanzees (P=0.002). The latter includes the first evidence of a genomic disorder in a chimpanzee with features resembling Smith-Magenis syndrome mediated by a chimpanzee-specific increase in segmental duplication complexity. We hypothesize that demographic effects, such as bottlenecks, have contributed to larger and more gene-rich segments being deleted in the chimpanzee lineage and that this effect, more generally, may account for episodic bursts in CNV during hominid evolution.

INTRODUCTION

Sequence and assembly of great ape reference genomes has consistently revealed that copy number variation (CNV) affects more base pairs than single nucleotide variation (SNV) (Chimpanzee Sequencing and Analysis Consortium 2005; Locke et al. 2011; Cheng et al. 2005). Segmental duplications, in particular, have disproportionately affected the African great ape (human, chimpanzee and gorilla) lineages where they appear to have accumulated at an accelerated rate (Marques-Bonet et al. 2009; Cheng et al. 2005). This has led to speculation that differences in fixation and copy number polymorphism may have contributed to the phenotypic "plasticity" and species-specific differences between humans and great apes (Varki et al. 2008; Olson 1999). While there is some evidence that fixed deletions and duplications contribute to morphological differences between humans and great apes (McLean et al. 2011; Dennis et al. 2012; Charrier et al. 2012), a comprehensive assessment of these differences at the level of the genome has not yet been performed. Previous studies of CNV have been predominated by array comparative genomic hybridization (CGH) experiments (Locke et al. 2011; Perry et al. 2006; Gazave et al. 2011; Dumas et al. 2007; Fortna et al. 2004), which provide limited size resolution, are imprecise in absolute copy number differences, and are biased by probes derived from the human reference genome. Comparisons of reference genomes have been complicated by assessments of a single individual and distinguishing CNVs from assembly errors (Chimpanzee Sequencing and Analysis Consortium 2005; Locke et al. 2011; Ventura et al.

2011; Prüfer et al. 2012). Here, we compare the evolution and diversity of deletions, duplications, and SNVs in 97 great ape individuals sequenced to high coverage (median ~25X) (Sudmant and Prado-Martinez et al., in press). The set includes multiple individuals from the four great ape genera, including Bornean and Sumatran orangutans, each of the four recognized chimpanzee subspecies, bonobos, and both Eastern and Western gorillas in addition to ten diverse humans and a high-coverage archaic Denisovan individual. This dataset provides unprecedented genome-wide resolution to interrogate multiple forms of genetic variation and a unique opportunity to directly compare mutational processes and patterns of diversity in great apes.

RESULTS

Patterns and Diversity: We constructed maps of deletions and segmental duplications by measuring sequence read-depth in 500 bp unmasked windows across the genome (Sudmant et al. 2010). We employed a scale-space filtering algorithm to identify deletion and duplication breakpoints (Supplementary section 3, **Figure 1a,b**). In addition to the breakpoints of deletions and duplications, read-depth genotyping allows us to determine the absolute copy number of loci at an individual genome level. We partitioned CNVs into three categories: fixed (i.e., the deletion or duplication was seen as a homozygous event in most individuals), copy number polymorphic, and private (observed only once) (see Supplementary Material for definitions). Fixed lineage-specific (events occurring on edges between nodes in the species tree) segmental duplications are non-randomly distributed ($P<0.0002$, permutation test) with more than 20% mapping within 5 kbp of shared ancestral duplications (Supplementary section 7)—a phenomenon we previously described as duplication shadowing (Marques-Bonet et al. 2009; Cheng et al. 2005). Deletions, in contrast, are randomly distributed across great ape genomes with respect to one another ($P>0.2$, permutation test).

We parsimoniously assigned fixed events to ancestral branches based on comparisons between populations. In total, we identify 469 Mbp of CNVs (**Table 1**). This set includes 11,836 fixed duplicated loci (325 Mbp; median length of 3,778 bp), 5,528 fixed deletions (47 Mbp; median size=4,227 bp), and 6,406 private and segregating copy number variants (96.2 Mbp) (**Table 1** and Supplementary section 3). In order to assess the accuracy of these calls, we performed 104 fluorescent *in situ* hybridization (FISH) experiments confirming 102 of the loci tested (98.1%). We also designed three custom duplication and deletion array comparative genomic hybridization (CGH) microarrays confirming 85% of CNPs (1,294/1,520 of events >2 kbp), 96.9% (3,660/3,776) of fixed duplications, and 98.6% (3,966/4,021) of fixed deletions (Supplementary section 4). As part of our assessment of deletions, we also screened sequence absent from the human reference genome yet present in one or more of the great ape reference genomes (Supplementary section 6). Since these "missing sequences" may represent artifacts or polymorphisms, we additionally estimated the frequency of each segment in 624 diverse humans from 13 different populations (1000 Genomes Project Consortium et al. 2012). We assigned 13.54 Mbp of human deletions unambiguously to specific time intervals during the evolution of our species. Notably, ~5% of these deleted sequences are still segregating in the human population, consistent with known population relationships among extant humans (**Figure 1c**).

Since fixed deletions are less likely than duplications to be subjected to recurrent mutation events, we assessed whether they might serve as reliable genetic markers for phylogenetic reconstruction of ape populations. The resulting neighbor-joining tree of deletion genotypes (**Figure 2a**) accurately recapitulates the ape phylogeny, including separation of Bornean and Sumatran orangutans, Eastern and Western gorillas, and bonobos and chimpanzees with high confidence. In contrast, however, to trees built from mitochondrial haplotypes or autosomal single nucleotide polymorphism (SNP) data from the same population (Sudmant and Prado-Martinez et al., in press), Central chimpanzees emerge as an out-group to the other chimpanzee subspecies (96% support). Interestingly, we observed a slight distortion towards increased branch length for the chimpanzee-bonobo ancestral lineage, which becomes more pronounced for larger deletions (see below section *Rates and CNV Load*; Supplementary section 9). Principal component analysis (PCA) of segregating structural variants also captures the subspecies relationships in addition to inter-population diversity (**Figure 2b**). Our analysis shows that estimates of SNP diversity and segregating copy number variants (as measured by Watterson's θ) are correlated (r^2=0.5 Pearson, P=0.02).

Genes: The availability of multiple sequenced genomes allows us to generate a comprehensive list of fixed deletions and duplications that disrupt genes along each branch of the ape lineage (see Supplementary section 5). We identified 407 lineage-specific gene duplications and 340 deletions with complete or partial exon loss (**Figure 3a,b,c**) with an excess of gene duplication events in the African great ape and chimpanzee-human ancestor. Lineage-specific duplications include a chimpanzee expansion of *PRDM7*, a high-identity paralog of *PRDM9*, in common chimpanzees (10-20 copies) and bonobos (35-40 copies) that is stratified among chimpanzee populations; a 75 kbp gorilla-specific expansion of *C1QTNF* and *AMACR*—genes important in brain and skeletal development; and 33 genes duplicated specifically in human since divergence from chimpanzee. This includes two genes that appear to have been duplicated, or to have increased in frequency, in the human lineage after the divergence from Denisova, ~700 kya (Meyer et al. 2012), with the caveat that only a single Denisovan individual was assessed. These potential Homo sapiens-specific genes include *BOLA2*, which resides just inside the critical region of the 16p11.2 locus, the deletion of which results in developmental delay, intellectual disability, and features of autism (Kumar et al. 2008; Weiss et al. 2008).

Among the 340 exonic gene-loss events, orangutans show the highest number (90), commensurate with their divergence from African great apes ~16 mya. Strikingly, the second highest number of gene-loss events occurs in the chimpanzee-bonobo ancestral lineage where 57 genes exhibit exonic deletions. As expected, we find a massive enrichment for olfaction genes (96/340) in addition to fixed deletions of immunity (*IL36*, *IL37* in chimp, *CCL26* in gorilla), drug detoxification (*CYP3A43* in Denisova, *CYP2C18* in humans and chimps), and sperm surface membrane genes (*ADAM2* in gorilla; *ADAM3A* in gorilla and *Pan* genus). Some genes appear to have undergone both lineage-specific duplication and loss. Of note is the carboxyl-esterase gene family (**Figure 3c**; *CES1*, *2*, *3*), which appears to have expanded independently in all great ape lineages with the exception of human where it remains diploid or alternatively has been subjected to deletion.

We were also interested in genes that were lost in the human lineage, and therefore absent from the human genome, since these have been hypothesized to contribute disproportionately to the evolution of human adaptive traits (Olson 1999). We, thus, analyzed the 13.54 Mbp of human fixed deletions (see above) for the presence of open reading frames (ORFs) where there was also support for a multi-exon spliced transcript from RNAseq data from multiple nonhuman primate tissues (Brawand et al. 2011) (Supplementary section 6). By this definition, we identified 86 putative gene losses along the branches leading to the human lineage—40 since divergence from chimpanzee. A search of these ORFs against the RefSeq protein database yielded not only previously annotated gene-loss events, such as the human-specific *SIGLEC13* (Wang et al. 2012) and *CLECM4* (Ortiz et al. 2008) deletions, but 42 previously unannotated or only predicted protein-coding genes with homology to other genes, 28 of which intersect highly conserved elements (HCEs, (Siepel et al. 2005)). In total we identified 180 kbp of highly conserved sequence within these fixed deletions, a marked depletion compared to the 3-8% of the human reference genome encompassed by HCEs. However, 18% and 12% of regions were located within introns or within 10 kbp upstream or downstream of annotated genes, respectively, suggesting some of these loci may have a potential regulatory impact as has been previously suggested (McLean et al. 2011).

Rates and CNV Load: Comparing deletions and duplications among different great ape lineages (>2 kbp), we find that the number of base pairs added by duplication significantly exceeds that of deletions by a factor of 2.8, although this ratio varies considerably depending on the specific lineage (**Table 1**). In this analysis, we considered only those base pairs added by new duplication excluding the ancestral locus. Overall, we find that the contribution of fixed base pairs by deletion and duplication is ~1.4-fold greater than that of single base-pair substitutions. We estimated rates of duplication and deletion throughout great ape evolution by normalizing the number of fixed base pairs that were lost or gained as a function of genetic branch length as well as divergence time (**Figure 4a,b**). All analyses were additionally computed in units of the number of events per millions of years (Supplementary Table 9.1) and exhibited the same observed trends. Although there was an acceleration of duplicated base pairs along the ancestral African great ape lineage ($P=9.786 \times 10^{-12}$), we predict that the rate of fixation subsequently declined in the ancestral lineage of human and chimpanzee and at a slower rate in the gorilla lineage. Our analysis shows that the rate of duplication in base pairs exceeds by threefold the rate of substitution in the African great ape lineage and is ~7-fold higher than the rate of duplication in the human lineage. This results in a significant excess of fixed gene duplication events occurring at this time point (**Figure 4c**; $P=1.66 \times 10^{-20}$).

The corresponding analysis for deletions shows a markedly different pattern with the rate occurring in a more clocklike manner throughout most of the tree with the notable exception of the ancestral lineage of chimpanzees and bonobos. We observe an approximate twofold increase in the rate of deleted base pairs leading to a distortion specifically along this branch ($P=4.79 \times 10^{-9}$). This increase results from an excess of large (>5 kbp) chimpanzee-bonobo ancestral deletions, which affect significantly more genes when compared to all other great ape lineages (**Figure 4c**; $P=4.397 \times 10^{-8}$). Notably, this excess of deletions corresponds to a predicted collapse in the ancestral chimpanzee-bonobo effective

population size (Ne) ~3 mya (Sudmant and Prado-Martinez et al., in press; Supplementary section 5).

As demography may have played a significant role in the excess rate of deletion in the chimpanzee-bonobo ancestor, we sought to estimate the relative burden of segregating duplications (**Figure 4d**) and deletions (**Figure 4e**) in each of the great ape populations by comparing CNV and SNP diversity (Methods). Specific populations showed an increased burden of CNV load, both in the total number of base pairs affected and in the number of events (Supplementary section 10), although humans were not remarkable in this regard as has been hypothesized (Varki et al. 2008). Western chimpanzees, bonobos, and Sumatran orangutans all showed an excess of segregating duplications >30 kbp consistent with an increased duplication burden in these populations (P=0.02, 0.0014 and 0.0088, respectively; Supplementary section 9). Western chimpanzees were the only population to show an additional excess of segregating deletions >30 kbp (P=0.002). All of these populations are predicted to have experienced striking collapses in their effective population sizes during recent evolution (Sudmant and Prado-Martinez et al., in press; Supplementary section 5). Western chimpanzees in particular exhibit the lowest overall nucleotide diversity and effective population size (8×10^{-4} Het/bp, Ne=9800) among all populations assessed. This subspecies also harbors the largest number of fixed deletions (34 events encompassing 276 kbp), consistent with a population that experienced a severe bottleneck.

A Putative Chimpanzee Genomic Disorder: Among the Western chimpanzees assessed, we identified one particularly striking private structural variant—a ~1.7 Mbp microdeletion on 17p11.2 in the individual Susie-A (BPRC, Holland) (**Figure 5a**). This deletion encompasses 29 genes, including *RAI1* (retinoic acid-induced 1). In humans, deletions of this locus cause Smith-Magenis syndrome (SMS). SMS is a rare syndrome with an incidence of 1 in 15,000-25,000 (Elsea and Girirajan 2008) resulting in severe behavioral abnormalities, mental retardation, and developmental delay. The clinical features of this chimpanzee bear striking similarity to many of the phenotypes observed in SMS patients (**Table 2**), including common SMS maladaptive behaviors such as aggression and disobedience, obesity, a humped back indicative of kyphoscoliosis, renal abnormalities, and velopharyngeal insufficiency (Supplementary section 10). The chimpanzee deletion is flanked by multiple loci that have undergone expansion in the *Pan* genus (**Figure 5b**). The typical human SMS deletion spans an additional 2 Mbp and has breakpoints mapping to different locations and different segmental duplication blocks (**Figure 5c**). In order to resolve the chimpanzee duplication organization, we sequenced to high quality a total of 20 large-insert BAC clones (2.9 Mbp, ~1.73 Mbp nonredundant sequence) identifying ~765 kbp of sequence absent from panTro3. We find these blocks have increased in size and complexity in the chimpanzee lineage with at least an additional 600 kbp of duplicated sequence compared to humans (**Figure 5d**). These results predict that the chimpanzee genome harbors a novel 17p11.2 architecture whose more complex organization predisposes to a deletion resulting in an SMS-like phenotype. This identifies the first chimpanzee-specific genomic disorder mediated by lineage-specific expansion and restructuring of segmental duplications creating a putative chimpanzee-specific hotspot for deletion.

DISCUSSION

We present the first genome-wide assessment of duplication and deletion diversity where single nucleotide substitutions have been used to calibrate CNV accumulation over the course of great ape evolution. There are three novel findings in this study. First, chimpanzees show an excess of large deletions early in their history. This is in stark contrast to almost every other population of great ape where deletions have accumulated in a more clocklike fashion. The ancestral human lineage does not show an excess in the number of duplicated or deleted base pairs despite previous predictions (Varki et al. 2008; Olson 1999). Second, specific populations of great apes show an excess of copy number polymorphic duplications, notably Western chimpanzees, bonobos, and Sumatran orangutans. Only the Western chimpanzee shows evidence of increased deletion polymorphism. These three populations stand out in that they are predicted to have experienced sudden rises and crashes in effective population size. The Western chimpanzees are the most extreme in this regard, showing the strongest signal of genetic drift and the largest excess of ancestry-informative markers—consistent with the strongest bottleneck.

One possibility may be that CNPs (both duplications and deletions), in general, increase with small effective population sizes but that a severe bottleneck is necessary in order to result in an increase in deletion burden as a result of strong selection against deletions. The neutral nature of the vast majority of SNPs suggests that reductions in diversity may, in some cases, have little effect on overall fitness, in contrast to large structural variants. Human investigations as well as Drosophila studies have additionally shown that deletions affecting genes are significantly more deleterious than duplications (Cooper et al. 2011; Emerson et al. 2008). Indeed, analyses of the theoretical relationship between N_e and rates of deletion and duplication have suggested fluctuations in effective population size may play a significant role in overall variations in genome size among organisms (Lynch 2007). These findings would explain the excess of deletions specifically in the ancestral chimpanzee branch as this species shows the most drastic decline in effective population size when compared to orangutan, human, and gorilla. Humans once again are similar to other great apes with respect to CNP burden and do not particularly stand out, although the number of genomes compared are few.

Finally, we report the first evidence of a genomic disorder in the chimpanzee lineage. The phenotype is remarkably similar to SMS but the breakpoints are not shared with the common recurrent deletion seen in humans. Our sequencing analysis shows that the chimpanzee 17p11.2 breakpoints have radically changed in structure and content facilitating non-allelic homologous recombination. Owing to the evolution of this chimpanzee-specific architecture, we predict that this locus represents a chimpanzee genomic hotspot of mutation and that additional recurrent microdeletions may be encountered among the chimpanzee population. It is somewhat surprising that Susie-A was captured from the wild, albeit as a young chimp. In light of her behavioral anomalies, it is unlikely that she would have survived to adulthood outside of captivity. This raises the intriguing possibility that additional cases, and perhaps novel recurrent genomic disorders, may be encountered as apes continue to be bred in captivity. Most comparative sequencing studies of human genomic disorder breakpoint regions have reported increasing complexity in the human lineage as a predisposing factor to rearrangement associated with disease

(Antonacci et al. 2010; Boettger et al. 2012; Rochette et al. 2001). Our results show that loci of increasing complexity are present in other great ape lineages creating species-specific hotspots prone to deletion and disease.

METHODS

Read-depth profiles were initially constructed from whole-genome sequence from 120 great ape individuals. We assessed the quality of each of these genomes by assessing the sequence read-depth in regions of the genome (1.1 Gbp) regarded as copy number invariant (Supplementary section 1). We excluded 23 individual genomes that showed considerable heterogeneity in their read-depth presumably due to non-uniformity (Supplementary Figure 1.1). We report analysis on the remaining 97 genomes: 75 were sequenced as part of the Great Ape Genome Diversity Project (Sudmant and Prado-Martinez et al., in press) to a mean coverage of ~25X on an Illumina HiSeq 2000 while an additional nine orangutans, ten humans, and the Denisova individual were sequenced as part of the Orangutan Genome Project and the Denisova Genome Project (Locke et al. 2011; Meyer et al. 2012). Individuals sequenced as part of the Great Ape Genome Project were originally selected to best represent wild natural diversity by focusing on captive individuals of known wild-born origin in addition to individuals from protected areas in Africa (Table S1). Individual genome subspecies designations were assigned as reported by sample sources and confirmed by SNP genotyping and PCA analysis. All reads were first divided into their 36 bp constituents and mapped to the human reference genome (NCBI36) using the mrsFASTc read aligner (Hach et al. 2010). Read-depth estimates across the genome were corrected for the underlying GC content, and a calibration curve from regions of known copy number was used to assign copy number estimates to windows of the genome. These regions were then segmented using a scale-space filtering algorithm (Supplementary section 3).

Briefly, the scale-space filtering algorithm transforms the windowed copy number waveform, $f(\mathbf{x})$, into a set of waveforms, $f(\mathbf{x}, \sigma)$, where values of σ represents the standard deviation of a Gaussian smoothing kernel applied to the original waveform. Contours of this transform are then traversed from large values of σ as $\sigma \rightarrow 0$ and the resulting segments are hierarchically clustered. We also masked regions of high GC content (>57%, corresponding to 2.23% of the genome). Array CGH validation experiments were performed in duplicate for every sample tested with Cy3 and Cy5 labeling dyes swapped. Probes giving opposite signals in the dye swap experiment were discarded. Only loci with at least three probes were considered for validation. CNV load comparisons were performed using Kaplan-Meier survival curves and statistical tests were corrected for sample size. BAC clones were selected from the chimpanzee BAC library CHORI-251 corresponding to the male chimpanzee Clint. Clones were sequenced using a PacBio RS system using standard protocols. The library was prepared with a 10 kbp insert size and sequence generated with C2 chemistry in 90-minute movies.

DATA ACCESS

Copy number maps for the 97 individuals assessed in this study are available online (http://eichlerlab.gs.washington.edu/greatape-cnv). All lineage-specific and segregating copy number variants are additionally reported in Supplementary Tables S2-S11. All structural variants have been deposited into the database of genomic structural variation

(dbVAR, http://www.ncbi.nlm.nih.gov/dbvar/) under accession nstd82. Underlying raw sequence reads have been deposited in the short read archive (SRA, http://www.ncbi.nlm.nih.gov/sra/) under accession SRP018689.

ACKNOWLEDGEMENTS
PHS is supported by a Howard Hughes International Student Fellowship. TMB is supported by an ERC Starting Grant (260372). TMB is an ICREA Research Investigator (Institut Catala d'Estudis i Recerca Avancats de la Generalitat de Catalunya). This work was supported, in part, by U.S. National Institutes of Health (NIH) grant HG002385 to EEE, BFU2009-13409-C02-02 to JPM, and MICINN (Spain) BFU2011-28549 to TMB. EEE is an investigator of the Howard Hughes Medical Institute.

FIGURE LEGENDS

Figure 1. Duplication and Deletion Landscape a. Ideograms of human autosomes 5 and 6 overlaid with copy number heat maps of the deletion landscape of great apes across 7 species and 11 distinct populations. Each row represents one of 97 individuals sorted by species; each column shows the estimated copy number in each of these individuals for deleted loci in 500 bp unmasked windows. Arrows above the chromosome ideogram indicate deletions identified along the lineages leading to the human species, the African great ape, chimpanzee-human, and human lineages, respectively. **b.** Ideograms of human autosomes 5 and 6 overlaid with copy number heat maps of the duplication landscape of great apes. **c.** Breakdown of the number of base pairs lost along the lineage leading to humans identified by screening sequence absent from the human reference genome yet present in the orangutan, gorilla, or chimpanzee reference genomes against the 97 great apes sequenced in this study. A total of 13.54 Mbp have been lost in these lineages since the divergence of African great apes and orangutans. We find an additional 680 kbp (316 loci) of sequence absent from the human reference genome (4.8% of the total) is fixed in all nonhuman great apes and segregating humans. For these loci a hierarchically clustered heat map is shown. Colors indicate the frequency of sequences absent in the human reference genome assessed in 624 diverse individuals from 13 different populations sequenced to low coverage by the 1000 Genomes Project and found to be segregating ≥5% frequency in at least one population. The hierarchical clustering recapitulates all the relationships between the individual human populations and the different great ape species assessed in this study. We identify 53.8 kbp of sequence segregating exclusively in African populations compared to only 1.4 kbp of sequence segregating specifically in Europeans.

Figure 2. Hominid Deletion Phylogeny a. Neighbor-joining tree constructed from pairwise edit distance of genotypes for fixed and segregating deletions >5 kbp. Branch length confidence estimates were generated by repeatedly subsampling 50% of the variants and regenerating the topology. All species and subspecies relationships are reconstructed with high confidence and are concordant with the topology identified from SNPs with the exception of Central chimpanzees, which form an out-group to the other chimpanzee subspecies as a result of their increased diversity. SNP-based trees cluster Central and Eastern chimpanzees on a single clade. Among chimpanzees, the three individuals Yolanda, Andromeda and Vincent, the Eastern-most individuals assessed in this study from Gombe National Reserve in Tanzania, cluster together with strong support. Additionally,

the individuals Tobi and Julie, a distinct subpopulation of Nigerian chimpanzees by SNP analysis, cluster together. Eastern lowland gorillas form an out-group to the gorilla clade and the Cross River gorilla clusters as an out-group to Western lowland gorillas. The archaic Denisova individual clusters as an out-group to all humans with 97% support. **b.** PCA of segregating deletion genotypes recapitulates intra-population relationships and additionally the relative diversity within the populations assessed.

Figure 3. Genic Variation a. Summary of the number of genes with exonic deletions and genes duplicated in each of the lineages assessed in this study. We identify 340 genes lost throughout the great ape lineage. While orangutans show the highest number of gene-exon-loss events (89), strikingly, the second highest number of gene-exon-loss events was in the chimpanzee-bonobo ancestral lineage where 55 were lost. **b.** A line plot of the copy number over the *DUF1220* domain of *NBPF10*. This domain has expanded specifically in the African great ape lineage with human exhibiting ~300 copies compared to 50-100 in chimpanzee, bonobo, and gorilla. **c.** Lineage-specific great ape duplication events encompassing genes. Gene models are drawn with the duplication breakpoints shown below colored by lineage and dot-plots of the copy number in all individuals assessed in this study. *PRDM7* is a close paralog of *PRDM9*, binding of which associates with recombination hotspots in humans. We find *PRDM7* is specifically duplicated in the *Pan* genus and highly stratified; Nigerian-Cameroon chimpanzees have 10-15 copies while Eastern and Central chimpanzees have 16-20. Bonobos exhibit 30-40 copies of the gene. FISH assays demonstrate the extra copies to be the result of subtelomeric duplicative transposition (**d**). *AMACR* and *C1QTNF3* are also specifically duplicated in gorillas. Mutations in *AMACR* have been shown to result in adult onset neurological disorders (Ferdinandusse et al. 2000) and *C1QTNF* plays a key role in skeletal development, inducing increased growth of murine mesenchymal cells with overexpression (Maeda et al. 2001).

Figure 4. CNV Rates and Polymorphism. Rates of **a)** duplication and **b)** deletion accumulation as a function of the number of substitutions along each branch of the great ape phylogeny. Tree branch lengths are scaled proportionally to the number of substitutions while tree widths are scaled proportionally to the number of duplicated base pairs per substituted base pair. Duplicated base pairs were added ~2.6-fold the rate of substitution along the African great ape ancestral branch, which rapidly declined in the chimpanzee-human ancestral lineage and more slowly in the gorilla lineage. In contrast, the rate of deletion in the great ape lineage is fairly consistent along all branches (mean of 0.32 deleted bp per substitution) with the exception of the chimpanzee-bonobo ancestral lineage where an approximate twofold increase in the rate of deletion is observed (0.71 deleted bp per substitution). **c.** The rate of genic deletion events and gene duplication events per million years plotted for each of the lineages assessed in this study. The rate of gene deletion events is significantly higher in the chimpanzee-bonobo ancestor ($p=5.262e-9$). An acceleration in the number of gene duplications is observed in the African great ape ancestor, the human-chimpanzee ancestor, and the ancestral gorilla lineage ($p=1.663e-20$). **d.** A survival curve of the total load of segregating duplications >30 kbp in Western chimpanzees, Sumatran orangutans, and bonobos compared to all other great apes shows that these three populations harbor an increased total number of duplicated base pairs (results significant for each individual population and combined). **e.** A survival curve for the

total load of deletions >30 kbp in Western chimpanzees compared to all other great apes shows a significant excess of deletions in this population. Western chimpanzee populations show the lowest diversity of any of the populations assessed in this study and the most fixed deletions of all chimpanzee species assessed.

Figure 5. A Chimpanzee Genomic Disorder a. A genome browser snapshot of 17p11.2 Smith-Magenis syndrome (SMS) critical region with a copy number heat map of the Western chimpanzee Susie-A and the Nigeria-Cameroon chimpanzee Koto. Susie-A has a 1.7 Mbp deletion of this locus, which encompasses *RAI1,* the critical gene associated the SMS phenotype. We confirm this deletion by array CGH. **b.** Copy number of great apes assessed in this study over the Susie-A deletion breakpoint 2 H-duplicon. **c.** Organization of the 17p11.2 SMS locus and 17p12 in humans with four blocks of segmental duplication. The typical human SMS deletion spans ~3.7 Mbp with different breakpoints than the Susie-A deletion (Elsea and Girirajan 2008). **d.** Segmental duplication architecture of the 17p11.2 locus as represented in the human reference genome and constructed in chimpanzees from high-quality sequencing of 22 BAC clones. We were able to assemble and anchor 11 of these clones into 7 contigs. The remaining 11 contigs were placed at their most likely locations and orientations based on their underlying duplication architecture and read-depth analysis of Susie-A compared to normal chimpanzees. We hypothesize a non-allelic homologous recombination event between the directly oriented chimpanzee G duplicons resulted in Susie-A's deletion.

REFERENCES

1000 Genomes Project Consortium. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**: 56–65.

Antonacci F, Kidd JM, Marques-Bonet T, Teague B, Ventura M, Girirajan S, Alkan C, Campbell CD, Vives L, Malig M, et al. 2010. A large and complex structural polymorphism at 16p12.1 underlies microdeletion disease risk. *Nat Genet* **42**: 745–750.

Boettger LM, Handsaker RE, Zody MC, McCarroll SA. 2012. Structural haplotypes and recent evolution of the human 17q21.31 region. *Nat Genet* **44**: 881–885.

Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, Harrigan P, Weier M, Liechti A, Aximu-Petri A, Kircher M, et al. 2011. The evolution of gene expression levels in mammalian organs. *Nature* **478**: 343–348.

Charrier C, Joshi K, Coutinho-Budd J, Kim J-E, Lambert N, de Marchena J, Jin W-L, Vanderhaeghen P, Ghosh A, Sassa T, et al. 2012. Inhibition of SRGAP2 function by its human-specific paralogs induces neoteny during spine maturation. *Cell* **149**: 923–935.

Cheng Z, Ventura M, She X, Khaitovich P, Graves T, Osoegawa K, Church D, DeJong P, Wilson RK, Pääbo S, et al. 2005. A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature* **437**: 88–93.

Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**: 69–87.

Cooper GM, Coe BP, Girirajan S, Rosenfeld JA, Vu TH, Baker C, Williams C, Stalker H, Hamid R, Hannig V, et al. 2011. A copy number variation morbidity map of developmental delay. *Nat Genet* **43**: 838–846.

Dennis MY, Nuttle X, Sudmant PH, Antonacci F, Graves TA, Nefedov M, Rosenfeld JA, Sajjadian S, Malig M, Kotkiewicz H, et al. 2012. Evolution of human-specific neural SRGAP2 genes by incomplete segmental duplication. *Cell* **149**: 912–922.

Dumas L, Kim YH, Karimpour-Fard A, Cox M, Hopkins J, Pollack JR, Sikela JM. 2007. Gene copy number variation spanning 60 million years of human and primate evolution. *Genome Res* **17**: 1266–1277.

Elsea SH, Girirajan S. 2008. Smith–Magenis syndrome. *Eur J Hum Genet* **16**: 412–421.

Emerson JJ, Cardoso-Moreira M, Borevitz JO, Long M. 2008. Natural selection shapes genome-wide patterns of copy-number polymorphism in Drosophila melanogaster. *Science* **320**: 1629–1631.

Ferdinandusse S, Denis S, Clayton PT, Graham A, Rees JE, Allen JT, McLean BN, Brown AY, Vreken P, Waterham HR, et al. 2000. Mutations in the gene encoding peroxisomal alpha-methylacyl-CoA racemase cause adult-onset sensory motor neuropathy. *Nat Genet* **24**: 188–191.

Fortna A, Kim Y, MacLaren E, Marshall K, Hahn G, Meltesen L, Brenton M, Hink R, Burgers S, Hernandez-Boussard T, et al. 2004. Lineage-specific gene duplication and loss in human and great ape evolution. *PLoS Biol* **2**: E207.

Gazave E, Darré F, Morcillo-Suarez C, Petit-Marty N, Carreño A, Marigorta UM, Ryder OA, Blancher A, Rocchi M, Bosch E, et al. 2011. Copy number variation analysis in the great apes reveals species-specific patterns of structural variation. *Genome Res* **21**: 1626–1639.

Hach F, Hormozdiari F, Alkan C, Hormozdiari F, Birol I, Eichler EE, Sahinalp SC. 2010. mrsFAST: a cache-oblivious algorithm for short-read mapping. *Nat Methods* **7**: 576–577.

Kumar RA, KaraMohamed S, Sudi J, Conrad DF, Brune C, Badner JA, Gilliam TC, Nowak NJ, Cook EH, Dobyns WB, et al. 2008. Recurrent 16p11.2 microdeletions in autism. *Hum Mol Genet* **17**: 628–638.

Locke DP, Hillier LW, Warren WC, Worley KC, Nazareth LV, Muzny DM, Yang S-P, Wang Z, Chinwalla AT, Minx P, et al. 2011. Comparative and demographic analysis of orang-utan genomes. *Nature* **469**: 529–533.

Lynch M. 2007. *The origins of genome architecture*. Sinauer Associates Inc.

Maeda T, Abe M, Kurisu K, Jikko A, Furukawa S. 2001. Molecular cloning and characterization of a novel gene, CORS26, encoding a putative secretory protein and its

possible involvement in skeletal development. *J Biol Chem* **276**: 3628–3634.

Marques-Bonet T, Kidd JM, Ventura M, Graves TA, Cheng Z, Hillier LW, Jiang Z, Baker C, Malfavon-Borja R, Fulton LA, et al. 2009. A burst of segmental duplications in the genome of the African great ape ancestor. *Nature* **457**: 877–881.

McLean CY, Reno PL, Pollen AA, Bassan AI, Capellini TD, Guenther C, Indjeian VB, Lim X, Menke DB, Schaar BT, et al. 2011. Human-specific loss of regulatory DNA and the evolution of human-specific traits. *Nature* **471**: 216–219.

Meyer M, Kircher M, Gansauge M-T, Li H, Racimo F, Mallick S, Schraiber JG, Jay F, Prüfer K, de Filippo C, et al. 2012. A High-Coverage Genome Sequence from an Archaic Denisovan Individual. *Science* **338**: 222–226. http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=22936568&retmode=ref&cmd=prlinks.

Olson MV. 1999. When less is more: gene loss as an engine of evolutionary change. *Am J Hum Genet* **64**: 18–23.

Ortiz M, Kaessmann H, Zhang K, Bashirova A, Carrington M, Quintana-Murci L, Telenti A. 2008. The evolutionary history of the CD209 (DC-SIGN) family in humans and non-human primates. *Genes Immun* **9**: 483–492.

Perry GH, Tchinda J, McGrath SD, Zhang J, Picker SR, Cáceres AM, Iafrate AJ, Tyler-Smith C, Scherer SW, Eichler EE, et al. 2006. Hotspots for copy number variation in chimpanzees and humans. *Proc Natl Acad Sci USA* **103**: 8006–8011.

Prüfer K, Munch K, Hellmann I, Akagi K, Miller JR, Walenz B, Koren S, Sutton G, Kodira C, Winer R, et al. 2012. The bonobo genome compared with the chimpanzee and human genomes. *Nature* 1–5.

Rochette CF, Gilbert N, Simard LR. 2001. SMN gene duplication and the emergence of the SMN2 gene occurred in distinct hominids: SMN2 is unique to Homo sapiens. *Hum Genet* **108**: 255–266.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**: 1034–1050.

Sudmant PH, Kitzman JO, Antonacci F, Alkan C, Malig M, Tsalenko A, Sampas N, Bruhn L, Shendure J, 1000 Genomes Project, et al. 2010. Diversity of human copy number variation and multicopy genes. *Science* **330**: 641–646.

Varki A, Geschwind DH, Eichler EE. 2008. Explaining human uniqueness: genome interactions with environment, behaviour and culture. *Nat Rev Genet* **9**: 749–763.

Ventura M, Catacchio CR, Alkan C, Marques-Bonet T, Sajjadian S, Graves TA, Hormozdiari

F, Navarro A, Malig M, Baker C, et al. 2011. Gorilla genome structural variation reveals evolutionary parallelisms with chimpanzee. *Genome Res* **21**: 1640–1649.

Wang X, Mitra N, Secundino I, Banda K, Cruz P, Padler-Karavani V, Verhagen A, Reid C, Lari M, Rizzi E, et al. 2012. Specific inactivation of two immunomodulatory SIGLEC genes during human evolution. *Proc Natl Acad Sci USA* **109**: 9935–9940.

Weiss LA, Shen Y, Korn JM, Arking DE, Miller DT, Fossdal R, Saemundsen E, Stefansson H, Ferreira MAR, Green T, et al. 2008. Association between microdeletion and microduplication at 16p11.2 and autism. *N Engl J Med* **358**: 667–675.

Figure 1

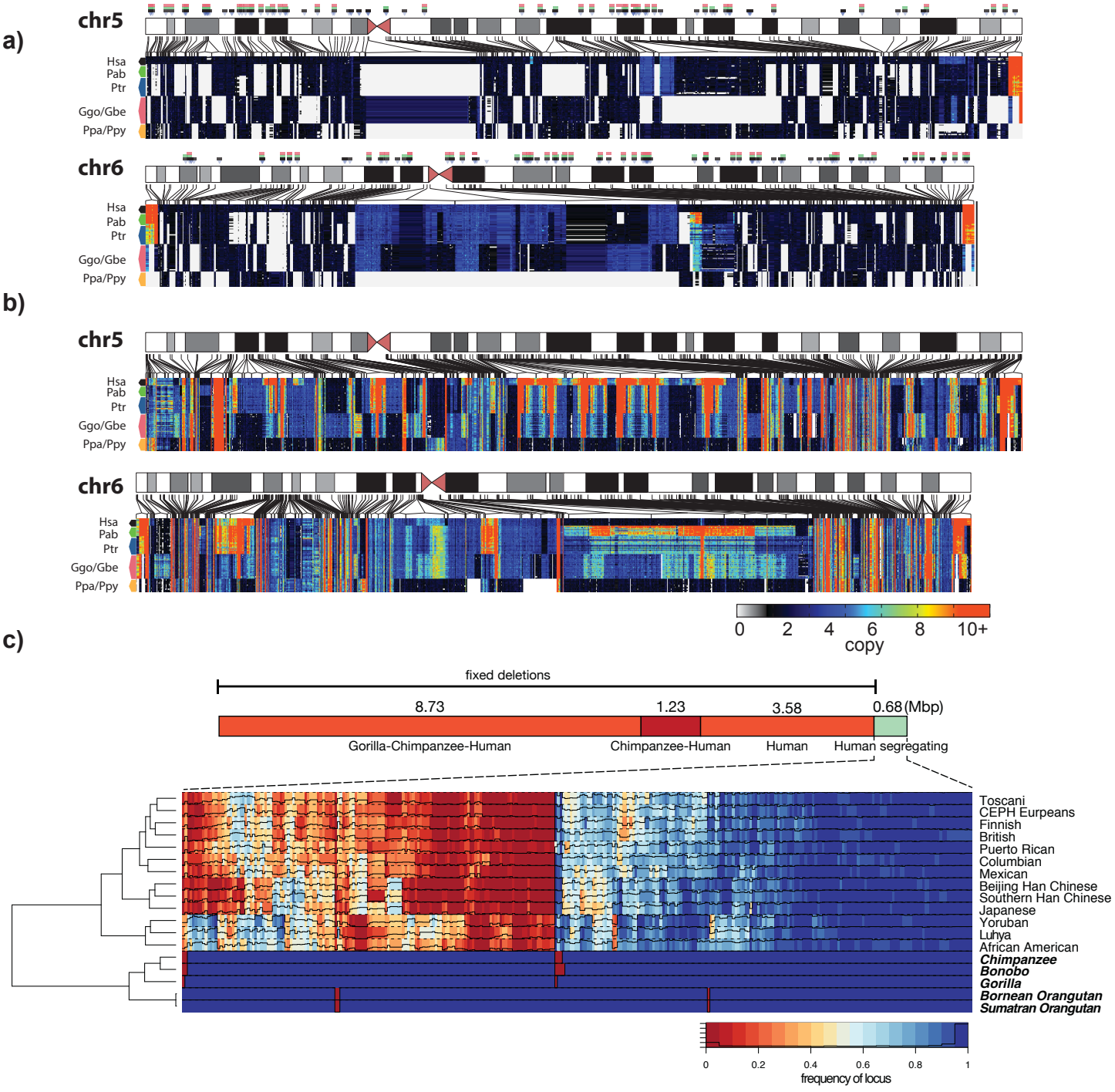

Figure 2

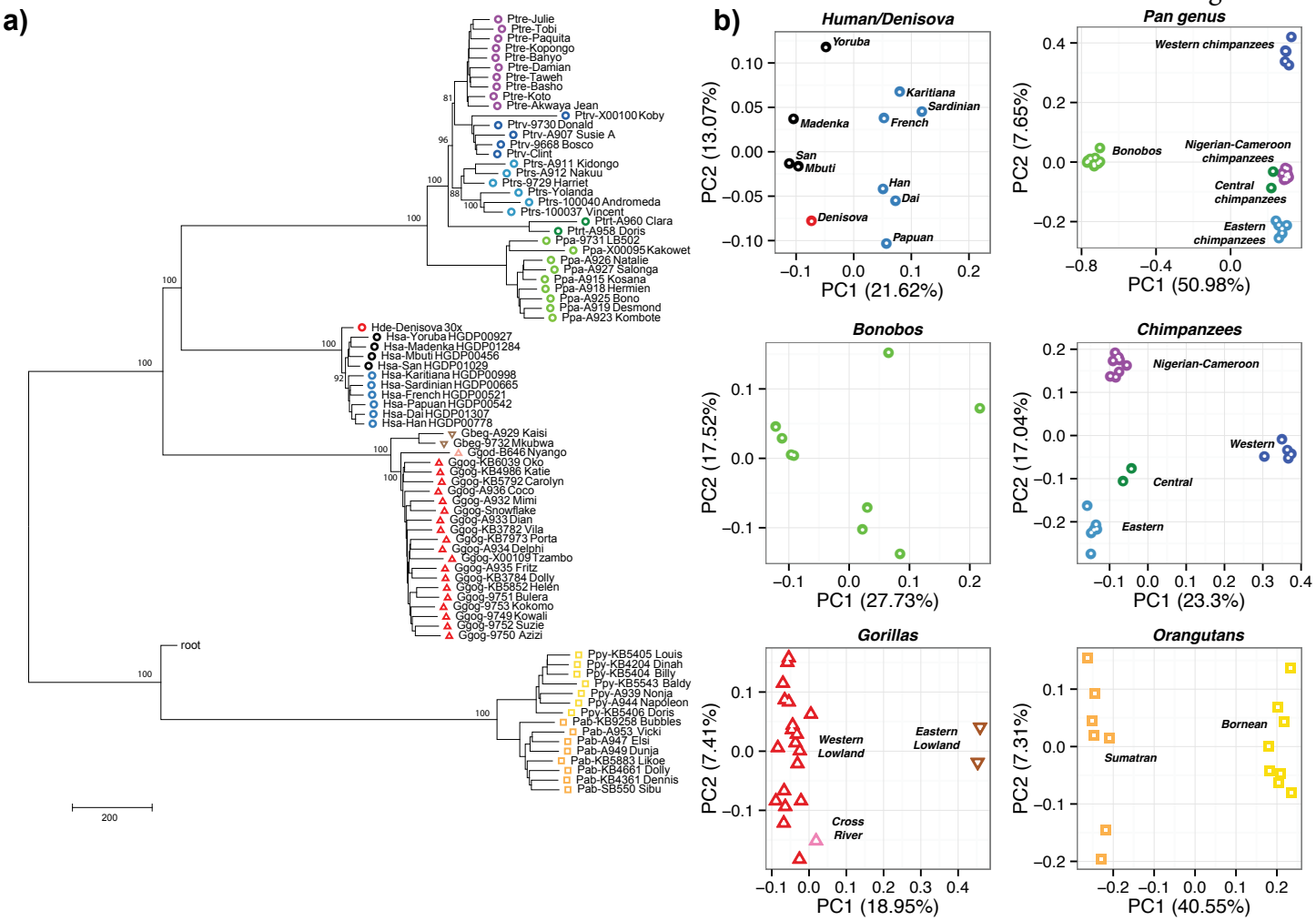

Figure 3

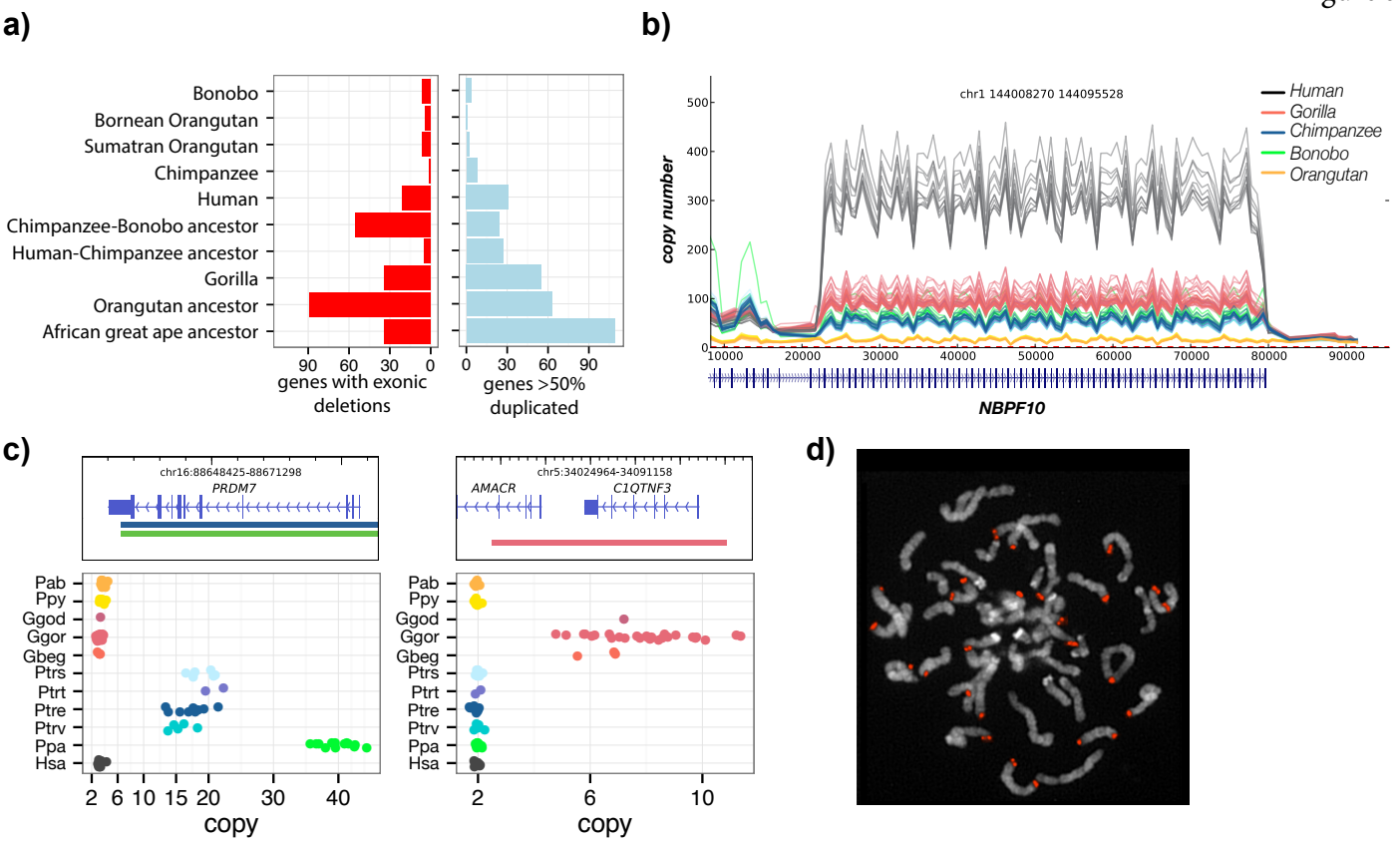

Figure 4

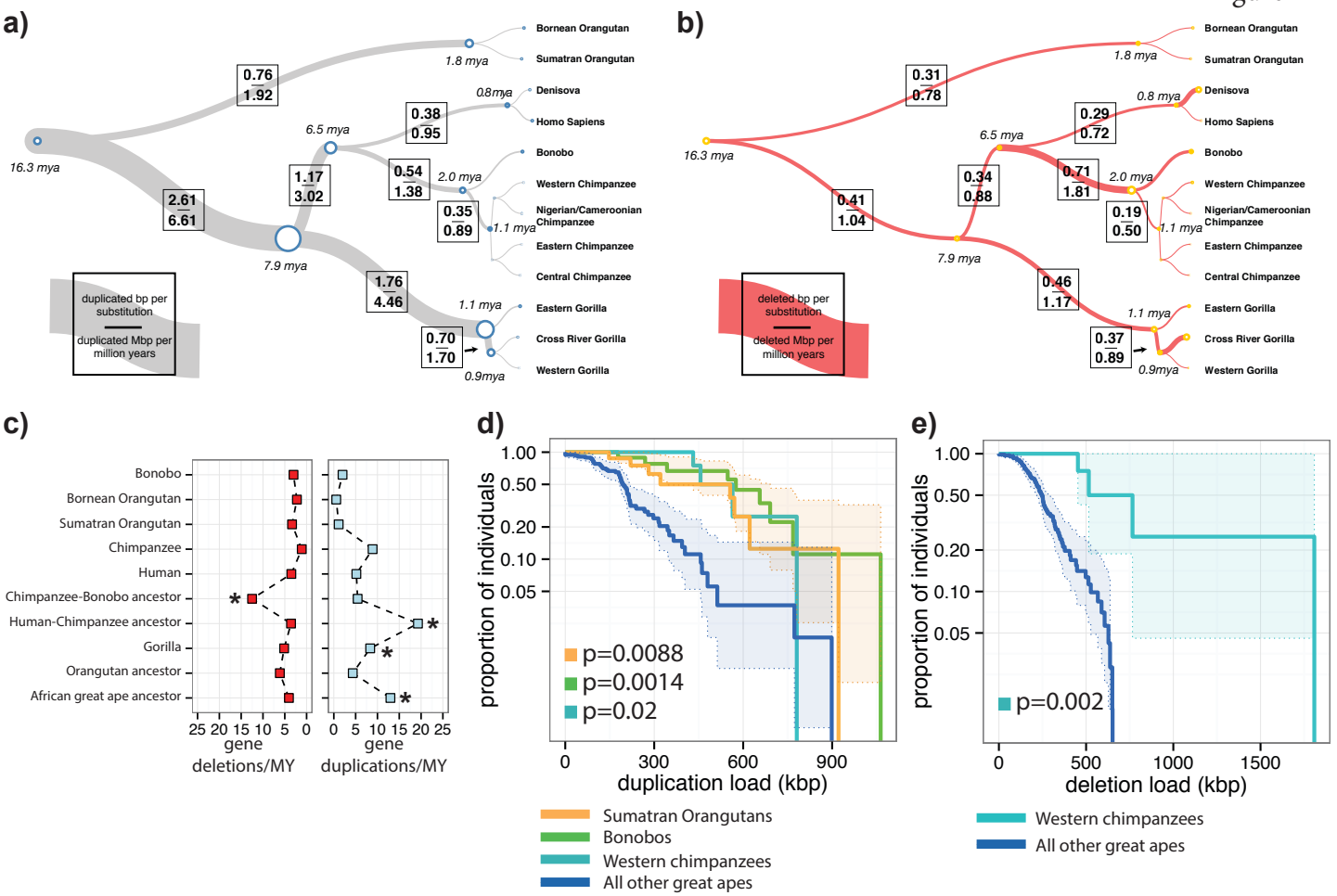

Figure 5