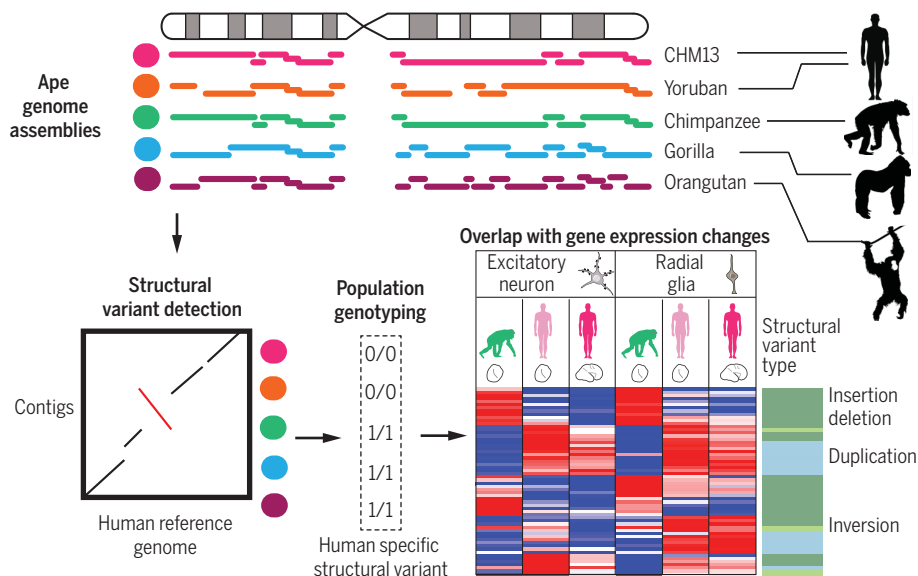## RESEARCH ARTICLE SUMMARY

GREAT APE GENOMICS

High-resolution comparative analysis of great ape genomes

Zev N. Kronenberg, Ian T. Fiddes*, David Gordon*, Shwetha Murali*, Stuart Cantsilieris*, Olivia S. Meyerson*, Jason G. Underwood*, Bradley J. Nelson*, Mark J. P. Chaisson, Max L. Dougherty, Katherine M. Munson, Alex R. Hastie, Mark Diekhans, Fereydoun Hormozdiari, Nicola Lorusso, Kendra Hoekzema, Ruolan Qiu, Karen Clark, Archana Raja, AnneMarie E. Welch, Melanie Sorensen, Carl Baker, Robert S. Fulton, Joel Armstrong, Tina A. Graves-Lindsay, Ahmet M. Denli, Emma R. Hoppe, PingHsun Hsieh, Christopher M. Hill, Andy Wing Chun Pang, Joyce Lee, Ernest T. Lam, Susan K. Dutcher, Fred H. Gage, Wesley C. Warren, Jay Shendure, David Haussler, Valerie A. Schneider, Han Cao, Mario Ventura, Richard K. Wilson, Benedict Paten, Alex Pollen, Evan E. Eichler†

INTRODUCTION: Understanding the genetic differences that make us human is a long-standing endeavor that requires the comprehensive discovery and comparison of all forms of genetic variation within great ape lineages.

RATIONALE: The varied quality and completeness of ape genomes have limited comparative genetic analyses. To eliminate this contiguity and quality disparity, we generated human and nonhuman ape genome assemblies without the guidance of the human reference genome. These new genome assemblies enable both coarse and fine-scale comparative genomic studies.

RESULTS: We sequenced and assembled two human, one chimpanzee, and one orangutan genome using high-coverage (>65x) single-molecule, real-time (SMRT) long-read sequencing technology. We also sequenced more than 500,000 full-length complementary DNA samples from induced pluripotent stem cells to construct de novo gene models, increasing our knowledge of transcript diversity in each ape lineage. The new nonhuman ape genome assemblies improve gene annotation and genomic contiguity (by 30- to 500-fold), resulting in the identification of larger synteny blocks (by 22- to 74-fold) when compared to earlier assemblies.

Including the latest gorilla genome, we now estimate that 83% of the ape genomes can be compared in a multiple sequence alignment.

We observe a modest increase in single-nucleotide variant divergence compared to previous genome analyses and estimate that 36% of human autosomal DNA is subject to incomplete lineage sorting. We fully resolve most common repeat differences, including full-length retrotransposons such as the African ape-specific endogenous retroviral element PtERV1. We show that the spread of this element independently in the gorilla and chimpanzee lineage likely resulted from a founder element that failed to segregate to the human lineage because of incomplete lineage sorting.

The improved sequence contiguity allowed a more systematic discovery of structural variation (>50 base pairs in length) (see the figure). We detected 614,186 ape deletions, insertions, and inversions, assigning each to specific ape lineages. Unbiased genome scaffolding (optical maps, bacterial artificial chromosome sequencing, and fluorescence in situ hybridization) led to the discovery of large, unknown complex inversions in gene-rich regions. Of the 17,789 fixed human-specific insertions and deletions, we focus on those of potential functional effect. We identify 90 that are predicted to disrupt genes and an additional 643 that likely affect regulatory regions, more than doubling the number of human-specific deletions that remove regulatory sequence in the human lineage. We investigate the association of structural variation with changes in human-chimpanzee brain gene expression using cerebral organoids as a proxy for expression differences. Genes associated with fixed structural variants (SVs) show a pattern of down-regulation in human radial glial neural progenitors, whereas human-specific duplications are associated with up-regulated genes in human radial glial and excitatory neurons (see the figure).

CONCLUSION: The improved ape genome assemblies provide the most comprehensive view to date of intermediate-size structural variation and highlight several dozen genes associated with structural variation and brain-expression differences between humans and chimpanzees. These new references will provide a stepping stone for the completion of great ape genomes at a quality commensurate with the human reference genome and, ultimately, an understanding of the genetic differences that make us human. ∎

ON OUR WEBSITE

Read the full article at http://dx.doi.org/10.1126/science.aar6343

SMRT assemblies and SV analyses. (Top) Contiguity of the de novo assemblies. (Bottom, left to right) For each ape, SV detection was done against the human reference genome as represented by a dot plot of an inversion). Human-specific SVs, identified by comparing ape SVs and population genotyping (0/0, homozygous reference), were compared to single-cell gene expression differences [range: low (dark blue) to high (dark red)] in primary and organoid tissues. Each heatmap row is a gene that intersects an insertion or deletion (green), duplication (cyan), or inversion (light green).

GREAT APE GENOMICS

High-resolution comparative analysis of great ape genomes

Zev N. Kronenberg[1]*, Ian T. Fiddes[2]*, David Gordon[1,3], Shwetha Murali[1,3]*,
Stuart Cantsilieris[1]*, Olivia S. Meyerson[4]*, Jason G. Underwood[1,5]*,
Bradley J. Nelson[1]*, Mark J. P. Chaisson[1,6], Max L. Dougherty[1], Katherine M. Munson[1],
Alex R. Hastie[7], Mark Diekhans[2], Fereydoun Hormozdiari[8], Nicola Lorusso[9],
Kendra Hoekzema[1], Ruolan Qiu[1], Karen Clark[10], Archana Raja[1,3],
AnneMarie E. Welch[1], Melanie Sorensen[1], Carl Baker[1], Robert S. Fulton[11],
Joel Armstrong[2], Tina A. Graves-Lindsay[11], Ahmet M. Denli[12], Emma R. Hoppe[1],
PingHsun Hsieh[1], Christopher M. Hill[1], Andy Wing Chun Pang[7], Joyce Lee[7],
Ernest T. Lam[7], Susan K. Dutcher[11], Fred H. Gage[12], Wesley C. Warren[11],
Jay Shendure[1,3], David Haussler[2,13], Valerie A. Schneider[10], Han Cao[7], Mario Ventura[9],
Richard K. Wilson[11], Benedict Paten[2], Alex Pollen[4,14], Evan E. Eichler[1,3]†

Genetic studies of human evolution require high-quality contiguous ape genome
assemblies that are not guided by the human reference. We coupled long-read
sequence assembly and full-length complementary DNA sequencing with a multiplatform
scaffolding approach to produce ab initio chimpanzee and orangutan genome
assemblies. By comparing these with two long-read de novo human genome assemblies
and a gorilla genome assembly, we characterized lineage-specific and shared great
ape genetic variation ranging from single– to mega–base pair–sized variants. We
identified ~17,000 fixed human-specific structural variants identifying genic and
putative regulatory changes that have emerged in humans since divergence from
nonhuman apes. Interestingly, these variants are enriched near genes that are
down-regulated in human compared to chimpanzee cerebral organoids, particularly in
cells analogous to radial glial neural progenitors.

Scientists have long been interested in the functional genetic differences that distinguish humans from other ape species (*1*). Human and chimpanzee protein-encoding changes and structural differences in regulatory DNA or in the copy number of gene families have all been implicated in adaptation (*2*, *3*). Indeed, several potentially high-impact reg-ulatory changes (*4*, *5*) and human-specific genes (*6*–*9*) that are important in synapse density, neuronal count, and other morphological differences have been identified. Most of these genetic differences, however, were not initially recognized upon comparison of human and ape genomes because the genetic changes mapped to regions of rapid genomic structural change that were not resolved in draft genome assemblies.

Despite recent efforts to sequence and assemble ape genomes (*10*–*12*), our understanding of structural differences, and particularly those specific to the human lineage, remains far from complete. There are two fundamental problems. First, there is considerable heterogeneity in the contiguity of ape genome assemblies. The presence of tens to hundreds of thousands of gaps in ape genomes limits the proportion of the genome that can be compared in a multispecies sequence alignment. Therefore, a large fraction of human-specific insertions and deletions, including those that alter regulatory sequences, are not resolved. Second, the higher-quality human genome assemblies have often been used to guide the final stages of nonhuman genome projects, including the order and orientation of sequence contigs and, perhaps more importantly, the annotation of genes. This bias has effectively "humanized" other ape genome assemblies, minimizing po-tential structural and transcript differences observed between the species. Using long-read, long-range sequence and mapping technologies (*13*–*15*), we generated new great ape genome assemblies, along with full-length cDNA annotation, without guidance from the human genome. We also generated and analyzed an African genome and an effectively haploid human genome complement to distinguish fixed differences in the human ancestral lineage and to further mitigate human genome reference biases.

Results

Genome assembly

We sequenced two human, one chimpanzee, and one orangutan genome to high depth (>65-fold coverage) using single-molecule, real-time (SMRT; PacBio) sequence data and assembled each ab initio using the same underlying assembly algo-rithm (Table 1) (*16*). For each species, we gener-ated assemblies ranging from 2.9 to 3.1 giga–base pairs (Gbp) in size, where most of the euchromatic DNA mapped to <1000 large contigs (Table 1). We error-corrected sequence contigs with Quiver (*17*) and Pilon (*18*), followed by a procedure that re-duced the remaining 1- to 2-bp indels (insertions or deletions) specifically in regions with clustered single-nucleotide variants (SNVs) (*16*). We next scaffolded the chimpanzee and orangutan ge-nomes without guidance from the human reference genome. In total, 93% (2.79 Gbp, excluding chro-mosome Y) of the chimpanzee-assembled bases and 92.7% (2.82 Gbp) of the orangutan-assembled bases were incorporated into chromosomal-level scaffolds (Table 1). We confirmed most large-scale chromosomal inversions among the great apes (*19*), some of which were absent from previous assemblies.

Sequence accuracy and quality assessment

More than 96% of our assembled sequence was concordant by length and orientation by differ-ent metrics (Table 1) (*16*). We conservatively estimate that these assemblies have improved contiguity for the chimpanzee and orangutan genomes by 32- and 533-fold, respectively (Fig. 1, A and B). Consistent with the gorilla genome (*20*), the application of long-read sequence data closed most of the genome gaps in earlier assem-blies. The extent of the change varied, however, depending on the prior level of finishing. In the case of the chimpanzee, 52% of the remaining 27,797 gaps were closed. We added 6.9 Mbp of new sequence and removed at least 27.2 Mbp of duplicated or extraneous sequence, possibly artifacts of scaffolding and gap filling (*21*). In the case of the orangutan, we added 54.5 Mbp of sequence while removing 4.2 Mbp, closing an estimated 96.8% (305,069/315,124) of the remaining euchromatic gaps. We determined the sequence contigs to be highly accurate at the base-pair level (>99.9%) on the basis of comparisons of each genome to Sanger end-sequence data, completely sequenced clone in-serts, and Illumina whole-genome sequencing data generated from the same source individ-uals (Table 1) (*16*).

[1]Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA 98195, USA. [2]UC Santa Cruz Genomics Institute, University of California, Santa Cruz, Santa Cruz, CA 95064, USA. [3]Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195, USA. [4]Department of Neurology, University of California, San Francisco, CA 94158, USA. [5]Pacific Biosciences (PacBio) of California, Inc., Menlo Park, CA 94025, USA. [6]Computational Biology and Bioinformatics, University of Southern California, Los Angeles, CA 90089, USA. [7]Bionano Genomics, San Diego, CA 92121, USA. [8]Department of Biochemistry and Molecular Medicine, University of California, Davis, Davis, CA 95817, USA. [9]Department of Biology, University of Bari, Aldo Moro, Bari 70121, Italy. [10]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA. [11]Departments of Medicine and Genetics, McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO 63108, USA. [12]The Salk Institute for Biological Studies, La Jolla, CA 92037, USA. [13]Howard Hughes Medical Institute, University of California, Santa Cruz, Santa Cruz, CA 95064, USA. [14]Eli and Edythe Broad Center of Regeneration Medicine and Stem Cell Research, University of California, San Francisco, San Francisco, CA 94143, USA.
*These authors contributed equally to this work.
†Corresponding author. Email: eee@gs.washington.edu

Gene annotation

Nonhuman primate (NHP) genome assemblies have typically relied almost exclusively on the human reference to define gene models (table S1). To provide a less biased source of gene annotation, we generated long-read transcriptome sequencing data to produce an average of 658,546 full-length nonchimeric (FLNC) transcripts from induced pluripotent stem cells (iPSCs) derived from each of the three nonhuman ape lineages (16). We selected iPSC material to maximize transcript diversity and enrich for early developmental genes. We next annotated the genomes of chimpanzee, gorilla, and orangutan, using FLNC transcripts along with short-read RNA-sequencing (RNA-seq) to guide gene and previously undescribed isoform predictions (22).

The number of genes and most gene models (coding and noncoding, including long noncoding RNA) are consistent among the different ape genomes (Table 2). However, we saw differential mapping of FLNC transcripts that favored the SMRT assemblies, especially in repeat-rich transcripts (Fig. 1C). Concordantly, human transcript models (GENCODE V27) aligned better to SMRT assemblies (Fig. 1D). For chimpanzee, 17,744 human protein-coding transcript models showed an increase of mapping coverage, which averaged 5.6%. This pattern was more pronounced in orangutan, where 28,033 of the 91,578 protein-coding transcript annotations showed an average improvement of 5.7% in mapping coverage. Overall, human protein-coding transcript models mapped to chimpanzee and orangutan SMRT assemblies with 99.1 and 98.8% average coverage, respectively—1.5 and 2.5% improvements. These improvements stemmed largely from gap closures, which rescue missing exons and re-

cover more full-length transcripts, including untranslated regions (UTRs).

We identified a small fraction (~1.5%) of putative protein-encoding genes present among NHPs that were absent in human annotations (GENCODE V27). In addition, a larger fraction (3.1 to 3.8%) of transcripts exhibited RNA-seq– or isoform sequencing (Iso-Seq)–supported splice junctions present in NHPs but not in human transcripts. Finally, we evaluated the NHP annotations, identifying full exons that affect coding sequences, which have been gained or lost between humans and other great apes (table S1).

Comparative sequence analyses

We constructed a five-way genome-wide multiple sequence alignment (MSA) of the ab initio assembled genomes (Table 1) by identifying syntenic (20 kbp) blocks against the human reference genome. In total, 83% of the ape genome was represented in MSAs. This allowed us to identify a comprehensive set of SNVs, indels, and structural variants (SVs); calculate divergence; and perform genome-wide phylogenetic analyses (Fig. 2). We observed a modest elevation in SNV divergence compared to previous genome comparisons (Fig. 2A and table S2) and estimated that 35.6% of the human genome is subject to incomplete lineage sorting among the African apes (Fig. 2B). Human and chimpanzee branch lengths are remarkably similar within coding regions (0.026% difference in branch length); however, we observed a 3.5% slowdown of the human mutation rate in noncoding regions (23, 24) (Fig. 2C). Human and chimpanzee branch lengths were considerably shorter compared to the other apes, consistent with the hominid slowdown hypothesis (25).

Repeat comparisons

Although the general repeat content of primate genomes has been well established (16), the longest and most complex repetitive regions have been more difficult to assay. Because long-read sequence data resolve most microsatellites and high-copy interspersed repeats (20, 26), we focused on comparative analysis of short tandem repeats (STRs) and endogenous retrovirus elements. Previous studies have suggested differential expansion of STR sequences between humans and other NHPs (27, 28). However, these studies suffer from ascertainment bias owing to methodological differences in genome sequencing or STR enrichment, differential access to GC-rich regions, and discovery bias in the human reference genome.

We analyzed each genome independently and, after clustering STRs that mapped within 25 bp, identified a consistent number of STRs per ape genome (344,354 to 358,622 STR regions; table S3). Because STRs often map within or adjacent to other classes of repetitive DNA, we restricted our analysis to the subset where orthology and STR lengths were clearly defined (12,694 to 16,138 STRs; fig. S28 and table S4). The average length difference between human and chimpanzee STR loci is 0.02 bp, with only a slight difference in distributions [$P = 0.015$, Kolmogorov-Smirnov (KS) test; table S5 and Fig. 2D]. Other ape comparisons show a modest increase in overall STR length (for example, a 1.2-bp average increase in gorilla versus chimpanzee; $P = 8.76 \times 10^{12}$, KS test). We found no significant difference between human and chimpanzee STR length in coding sequences ($n = 2199$, $P = 0.28$, KS test) or UTRs of genes ($n = 2794$, $P = 0.16$, KS test), although we identified 4920 loci preferentially

Table 1. Assembly statistics for the great ape genomes. QV, quality value score; AGP, a golden path assembly; ND, no data.

Statistic	Ape assembly				
	CHM13_HSAv1* (human)	YRI_HSAv1 (human)	Clint_PTRv1 (chimpanzee)	GSMRT3.2 (gorilla)	Susie_PABv1 (orangutan)
Estimated depth†	72	116	117	86.3	94.9
Subread length N50 (kbp)‡	16.2	13.4	17.4	18.6	16.6
Contigs, initial and final counts§	1,923 and 1,916	3,645 and 3,642	4,912 and 5,037	15,997	5,771 and 5,813
Assembly size (Gbp)	2.88	2.88	2.99	3.08	3.04
Contig length >3 Mbp (Gbp)	2.65	2.27	2.45	2.42	2.48
Initial and final contig N50 lengths (Mbp)§	29.26	6.60	12.76 and 12.42	10.02	11.27 and 11.07
Scaffold N50 (Mbp)	83.02	ND	53.1	ND	98.47
Longest contig (Mbp)‖	81	27	80	36	53
BAC concordance (%)	97.11	97.73	99.13	96.85	96.75
Sequence accuracy (QV)	36	31	33 to 38	30 to 38	28 to 33
Iso-Seq transcripts	710,974	ND	565,691	881,801	528,145
Contigs in AGP	ND	ND	685	794	544
Contigs aligned to GRCh38¶ (length in Gbp)	407 (2.8)	1,167 (2.8)	656 (2.8)	907 (2.8)	524 (2.8)

*Haploid genome assembly derived from a complete hydatidiform mole. †Estimated coverage in raw SMRT subreads based on 3.5-Gbp (gorilla) or 3.2-Gbp (all others) estimated genome size. ‡N50 subread lengths of raw input data. §Initial and final contigs before and after resolving chimeras by optical map comparison; note that optical maps are not available for each assembly. These stats do not consider the National Center for Biotechnology Information (NCBI) minimum–contig length filter. ‖Longest contig without gross assembly error. ¶Contigs with less than 95% of sequence aligning to GRCh38, depth-of-coverage greater than two SDs above the mean, or no coverage were excluded.

expanded in the human lineage (table S6), including loci associated with genomic instability and disease.

Endogenous retroelements are among the longest retrotransposons within mammalian genomes (up to 10 kbp) and are frequently mis-assembled because of their copy number and sequence identity. The chimpanzee and gorilla lineages carry an endogenous retrovirus, PtERV1, that is absent in orangutan and human genomes (29, 30). None of the PtERV1 integrations between chimpanzees and gorillas appear orthologous,

suggesting either that independent retroviral integrations occurred in these two lineages (29, 30) or that humans and orangutans contain extrinsic factors that differentially restricted propagation (31). A high-quality map of 540 PtERV1 elements [both full-length and solo long terminal repeat (LTR)] in chimpanzee and gorilla (table S7) (16) shows that their integration events are non-orthologous (99.8%), biased against genes, and integrated in the antisense orientation (figs. S30 and S31), consistent with the action of purifying selection.

Using the more complete ape genomes, we identified only one chimpanzee-gorilla ortho-logous PtERV1 element, not present in modern humans, that was lost through incomplete lineage sorting and integrated roughly 4.7 million years ago [95% highest posterior density: 1.9, 7.2 million years ago; Fig. 2E]. We named this element the "source PtERV1," as it was present in the common ancestor of all African apes and was likely the progenitor for independent expansions to nonorthologous loci in the chimpanzee and gorilla genomes. The source PtERV1
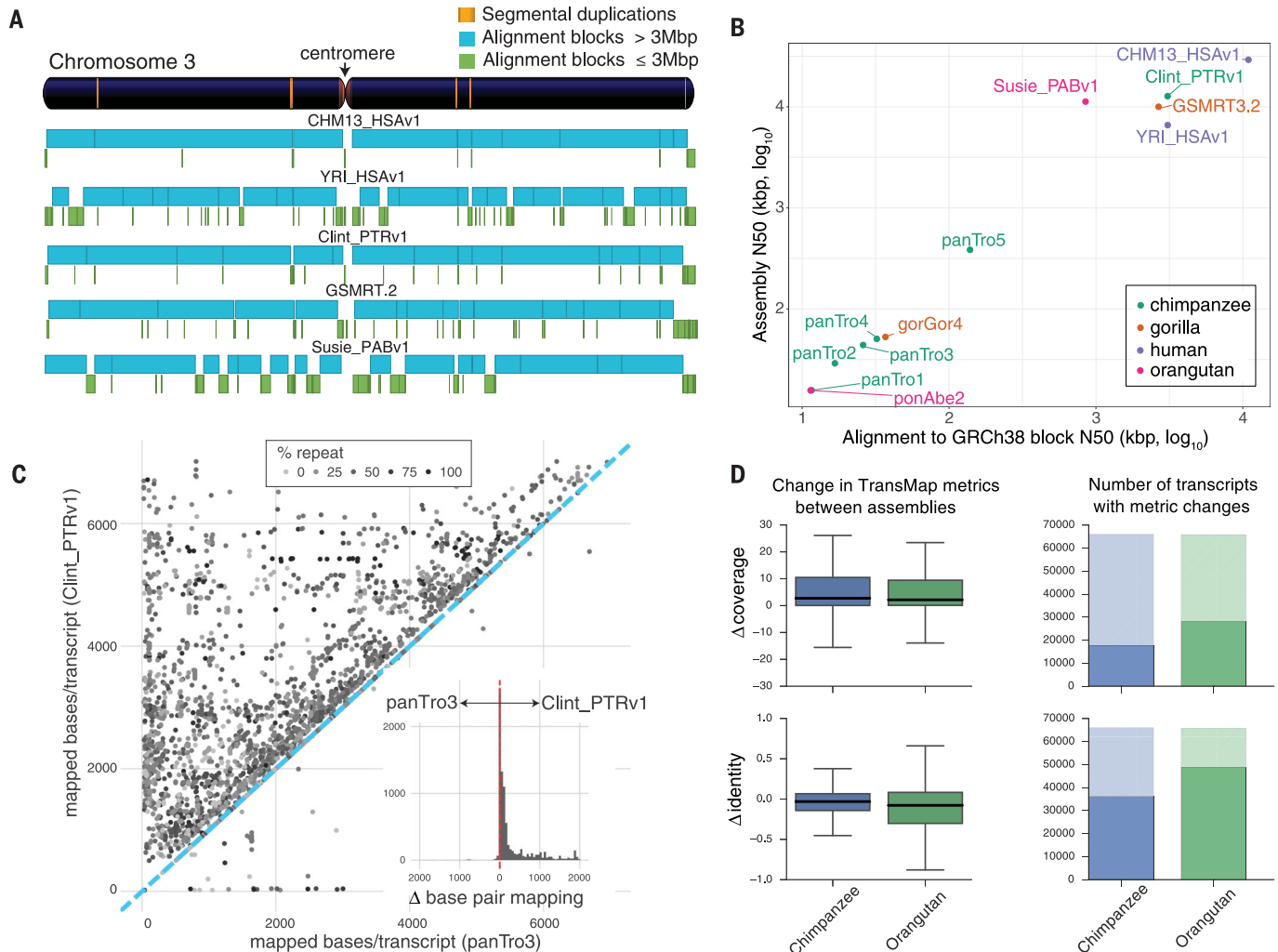
Fig. 1. Assembly and annotation of great ape genomes. (**A**) Comparison of genome sequence contiguity. Chromosome 3 contiguity is compared among the great ape genome assemblies by alignment to human reference genome sequence GRCh38. Contigs larger than (blue) and smaller than (green) 3 Mbp are compared with the position of SDs (SDs >50 kbp in size, orange) shown in the reference ideogram. (**B**) Scatterplot of syntenic-alignment block lengths (x axis) against GRCh38 versus FALCON assembly contig N50 length (y axis) of the great ape assemblies. The SMRT assemblies are Clint_PTRv1, Susie_PABv1, GSMRT3.2, CHM13_HSAv1, and YRI_HSAv1. The previous reference genomes are ponAbe2 (GCF_000001545.3), gorGor4 (GCA_000151905.3), panTro2 (GCF_000001515.2), panTro3 (GCA_000001515.3), panTro4 (GCA_000001515.4), and panTro5 (GCA_000001515.5). (**C**) Full-length assembled transcripts mapped to Clint_PTRv1 and panTro3. Each point denotes the number of bases per

transcript matching the two assemblies. Repeat content is indicated by gray shading of the points. Although most of the transcripts map well to both assemblies (Pearson's correlation = 0.95), the subset of differentially mapped transcripts (12,724; 60% of 21,118) aligns better to Clint_PTRv1 (dots above the blue dashed line). The histogram inset shows the effect, per transcript, with a total of 4.8 million more bases aligned to Clint_PTRv1. Δ, difference in mapped bases per transcript. (**D**) Comparative Annotation Toolkit was used to project transcripts from GRCh38 to Clint_PTRv1, panTro3, Susie_PABv1, and ponAbe2. Alignment coverage and identity were compared for orthologous transcripts found in each assembly pair. The boxplots (left) summarize TransMap differences between the short-read and SMRT assemblies, in terms of coverage and identity. The solid-shaded portions of the bar plots (right) represent alignments, which had identical coverage or identity in both assemblies.

was likely missed in earlier genomic studies of draft genomes because the locus (sharing orthology with human chromosome 19) (*16*) is repeat rich and the integration site is an ancient LTR element.

Structural variation analyses

We focused on identifying all SVs >50 bp in size within ape genomes because these are the least well-characterized differences and are more likely to affect gene function than SNVs (*32*). SVs were identified by mapping each assembly back to the human reference genome, by using the two newly assembled human genomes as a control for reference effects and fixed human differences

(CHM13_HSAv1 and YRI_HSAv1). We detected 614,186 ape deletions, insertions, and inversions, with the number of SVs increasing as a function of evolutionary distance from human (Fig. 3 and Table 3). We confirmed 92% of 61 events (from 2.7 to 95 kbp in size) by bacterial artificial chromosome (BAC) sequencing (table S8) (three of the remaining events were polymorphic among the great apes, suggesting a validation rate of >95%). We assigned SVs as shared or lineage-specific and genotyped each at the population level, with a panel of 86 great apes (*33*) (Fig. 3A). We identified 17,789 fixed human-specific structural variants (fhSVs), including 11,897 fixed human-specific insertions (fhINSs) and 5892 fixed

human-specific deletions (fhDELs) (Fig. 3A and table S9). Projecting these onto the human genome identifies potential hotspots of structural variation (Fig. 3B).

We annotated fhSVs against chimpanzee and human gene models (table S10). The Variant Effect Predictor annotated the loss of 13 start codons, 16 stop codons, and 61 exonic deletions in the human lineage. By contrast, we estimate that fhSVs disrupt 643 regulatory regions near 479 genes (for example, Fig. 3, C to E). Interestingly, 139 of the fhSVs intersect with regions recently classified as super-enhancers (*34*). A comparison with a previous analysis of human-conserved deletions (hCONDELs) from earlier

Fig. 2. Ape genetic diversity and lineage sorting. (**A**) SNV divergence between each primate assembly and GRCh38 was calculated in 1-Mbp non-overlapping windows across all autosomes and chromosome X (excluding X and Y homologous regions). Mean autosomal divergence is 1.27 ± 0.20% (human-chimpanzee), 1.61 ± 0.21% (human-gorilla), and 3.12 ± 0.33% (human-orangutan). The African genome (YRI_HSAv1) shows a 17% increase in SNV diversity. (**B**) Proportion of phylogenetic trees supporting standard species topology and incomplete lineage sorting (ILS). The mean and 95% confidence intervals (in brackets) are based on 100 genome-wide permutations. (**C**) A phylogenetic tree (maximum clade credibility consensus tree) comparing genic regions [~9000 consensus coding sequences (CCDSs) and 1000-bp flanking sequence (orange)] to a randomly genome-shuffled set matched to coding-sequence lengths (green). The analysis excludes regions of SDs, SVs, and large tandem repeats. Branch lengths (noted above the lines) and proportion of trees supporting each bifurcation (internal nodes) are shown. Violin plots summarize the distribution and mean divergence (substi-

tutions per base pair) for a subset of trees consistent with the species tree. YRI_HSAv1 is the representative human in the violin plots. (**D**) A comparison of the expanded STR sequences (*n* = 16,138 loci) between human (African) and chimpanzee ab initio genome assemblies shows little to no species bias (0.02 bp). (**E**) Top, a MSA of ape genomes (gorilla BAC CH277-16N20 and chimpanzee CH251-550G17) identifies an orthologous 379-bp PtERV1 element nested within another LTR and shared
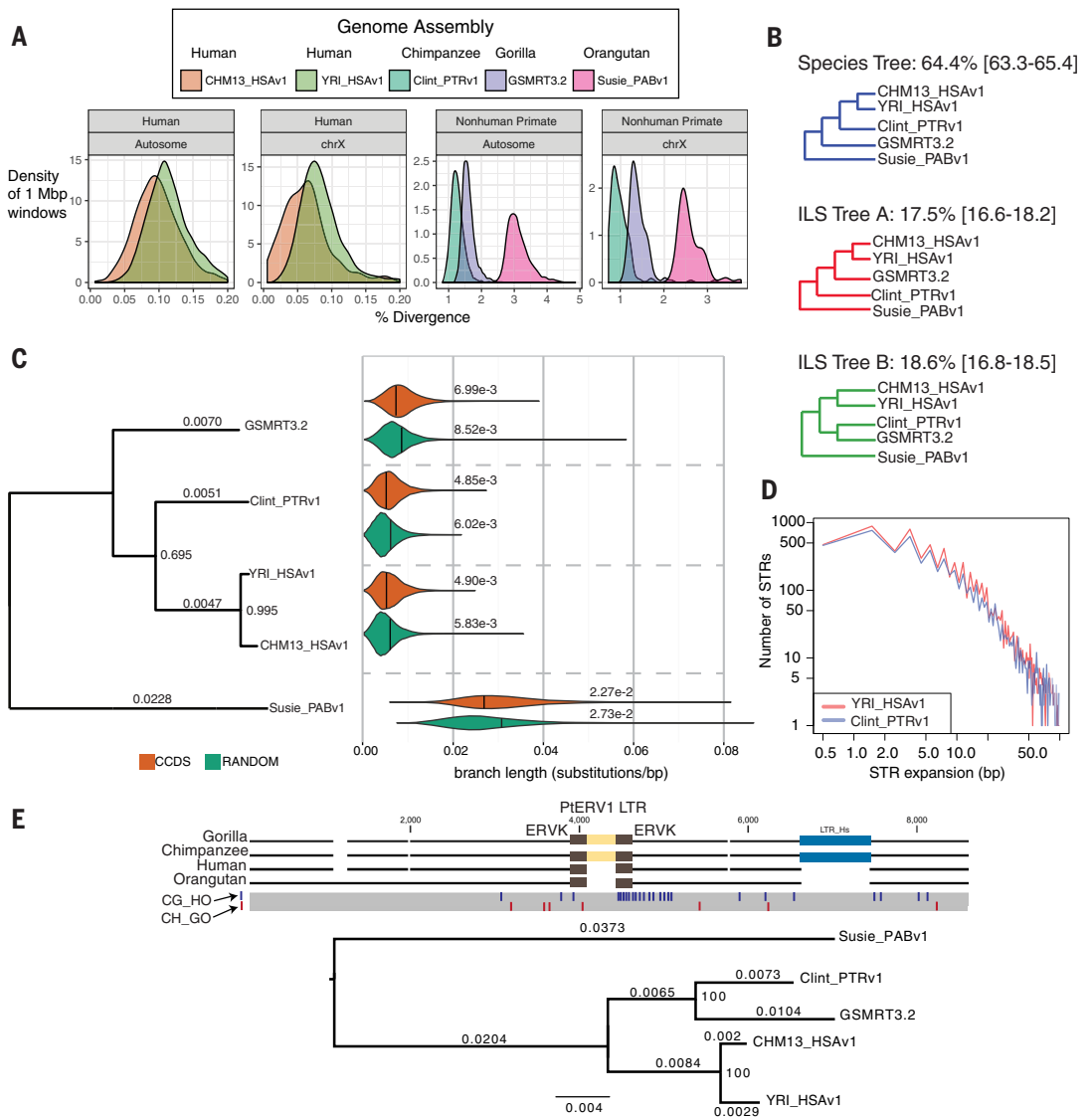
between gorilla and chimpanzee. Bottom, a maximum likelihood phylogenetic tree (GTR+Gamma) built from 12,108 bp that supports ILS. Single-nucleotide polymorphisms that support chimpanzee-gorilla sorting (CG_HO) and the species tree (CH_GO) are shown as blue and red lines, respectively. Branch lengths (substitutions per site) are shown above the lineages, and internal nodes are labeled with bootstrap support (proportion of replicates supporting split; 1000 replicates).

versions of the human, chimpanzee, and macaque genomes (5) confirms that 77% (451/583) of the hCONDELs intersect the fhDELs, with the remainder corresponding primarily to polymorphic events in the human population (Fig. 3F). We also predicted an additional 694 hCONDELs (table S11). A comparison of the SMRT gorilla assembly to the human reference genome identified an hCONDEL sequence previously reported as affecting an androgen receptor (AR) enhancer and associated with the loss of penile spines in humans. In gorilla, this fhDEL involves a complex SV, including an inversion, that may independently influence AR gene expression in the gorilla lineage (Fig. 3G) (35).

The spectrum of structural variation ranges from simple insertion and deletion events to larger events of increasing complexity (Fig. 4). We identified 46 fhSV deletions that putatively disrupt the orthologous chimpanzee gene, of which only six were previously reported (5). Seven of the 46 fhSV deletions can also be seen in the transcript data (Iso-Seq). The largest previously unidentified fhSV deletion is 61,265 bp in size. It contains almost all of the caspase recruitment domain family member 8 (CARD8) gene and removes 13 exons that are transcribed into full-length cDNA in the chimpanzee (Fig. 4A). We also resolve a 65-kbp human-specific deletion in FADS1 and FADS2, genes involved in fatty acid biosynthesis that have been the target of positive selection (36) and potential dietary changes in human evolution (37, 38). The deletion brings the promoters of FADS1 and FADS2 (major isoform) in closer proximity and shortens the first intron of the other two FADS2 isoforms (Fig. 4B). The fhDEL might alter the relative abundance of the FADS2 isoforms, as supported by quantifying the number of splice junction–containing reads specific to each isoform (16).

The relative abundance of the minor FADS2 isoforms is significantly increased in humans (χ^2 = 165.65, df = 1, $P < 2.2 \times 10^{-16}$). These minor isoforms differ only in their N terminus, and, of the two, one (NM_001281502.1) shows evidence of encoding a signal peptide (39), potentially altering the protein's subcellular location. Because great ape diets range from herbivorous to omnivorous, genic and structural changes related to diet metabolism may be of particular relevance for the evolution of ape species.

We further discovered two fhDELs in WEE1 (Fig. 4C) and CDC25C (Fig. 4D), two highly conserved cell-cycle genes that act as ultrasensitive antagonists during the interphase to mitotic transition, G_2/M (40). WEE1 encodes a serine-threonine protein kinase that delays mitosis by phosphorylating cyclin-dependent kinase 1 (CDK1), whereas CDC25C is a member of the phosphatase gene family and encodes a protein that dephosphorylates CDK1, triggering entry into mitosis. Expression of these genes in radial glia is particularly interesting because additional cell divisions are thought to have played a role in increasing the number of cortical neurons in human evolution (41). These cell-cycle regulators that display different protein sequence or differential expression between chimpanzee and human are, thus, candidates for future investigation to explain neocortical expansion in the human lineage.

We also identified several larger, subcytogenetic structural differences using optical (Bionano) (42, 43) and BAC end-sequence mapping data that were not detected or sequence-resolved in previous genome assemblies. We validated large inversions and more-complex SV events by integrating fluorescence in situ hybridization (FISH) and large-insert clone sequencing at the breakpoints (table S12). We identified 29 human-chimpanzee-orangutan inversions (16 in chim-panzee, 10 in orangutan, and 3 shared between chimpanzee and orangutan) ranging from 100 kbp to 5 Mbp in size, of which 55% (16/29) have not been previously described (table S12 and Fig. 5) (44–48). More than 93% of inversions are flanked by large complex segmental-duplication (SD) blocks, 38% of which show evidence of other structural and copy-number variation at the boundaries of the inversion (Fig. 5).

Interestingly, ~28% (8/29) of these ape-human inversions are also polymorphic among humans (49, 50), some in regions previously shown to be hotspots of recurrent rearrangement and disease (48, 51). Notably, these regions of genomic instability also associate with expression differences in radial glial and excitatory neurons between the species. For example, among the 18 chimpanzee-human inversions (table S12), we identified 18 differentially expressed brain genes between chimpanzee and human (10 radial glia, 11 excitatory neurons, and 3 common to both sets), of which 78% resided in SD regions. Three of these genes (GLG1, ST3GAL2, and EXOSC6) were significantly up-regulated in human and associated with a 5-Mbp human-specific inversion on chromosome 16q22 (Fig. 5D). ST3GAL2 encodes the main mammalian sialyltransferase for GD1a and GT1b ganglioside biosynthesis in the brain (52).

Radial glial neural progenitor expression differences and human-specific SVs

Over the course of human evolution, human brain volume has nearly tripled compared to that of chimpanzees (53), likely owing to differential expression of genes during brain development (6, 8, 54). We investigated the association of structural variation with changes in human-chimpanzee brain gene expression using cerebral organoids as a proxy for brain expression differences (55). Importantly, because great ape brain tissue is largely inaccessible, these organoid models provide a realistic window into developmental cell behavior and gene expression differences between human and ape radial glia and other early developmental cell types (56). We processed several single-cell RNA-seq brain data sets from primary human cortex and from human and chimpanzee cortical organoids, focusing on cortical excitatory neurons and radial glia (55–57). Using the new chimpanzee SMRT assembly and genome annotations increases the sensitivity of gene expression analyses; our data set reveals 2625 additional chimpanzee genes with expression in the brain relative to previous studies (58). After performing unsupervised clustering, we analyzed 52,875 orthologous genes in 320 primary neurons, 176 human organoid cells, and 210 chimpanzee organoid cells expressing cortical radial glia and excitatory neuron genes.

Our analysis identified 383 and 219 genes up-regulated in human radial glial and excitatory neurons, respectively, when compared to chimpanzee (table S13) (16). Conversely, we defined a set of 285 and 165 genes down-regulated in human radial glia and excitatory neurons (Fig. 6), respectively; most of these changes have not been identified previously (56, 59). Because SVs

Table 2. Great ape gene and transcript annotation summary. TPM, transcripts per million; NA, not applicable to this genome; ND, no data.

Statistic	Ape assembly		
	Clint_PTRv1	GSMRT3.2	Susie_PABv1
Genes	55,894	55,985	55,522
Orthologs in human	55,594 (95.4*)	55,570 (95.4*)	54,900 (94.2*)
Isoforms	192,725	192,734	190,716
Coding genes	19,153	19,311	19,043
Previously undescribed predicted genes†	300	415	322
Coding isoforms	92,610	92,713	91,578
Transcript predictions with previously undescribed splice junctions‡	2,809	2,902	2,333
Transcripts with TPM > 0.1 (%)	66.3	67.3	50.6
Transcripts supported by Iso-Seq reads (%)	66.5	46.5	63.4
Previously unannotated exons identified	29	16	16
Putative exons gained in human	57	NA	ND
Putative exons lost in human	13	NA	NA

*Percent of GENCODE V27 represented.　†Previously undescribed predicted genes based on GENCODE V27 annotation.　‡Previously undescribed splice junctions compared to liftover annotation set from the human reference genome, where splice junction is supported by NHP RNA-seq.
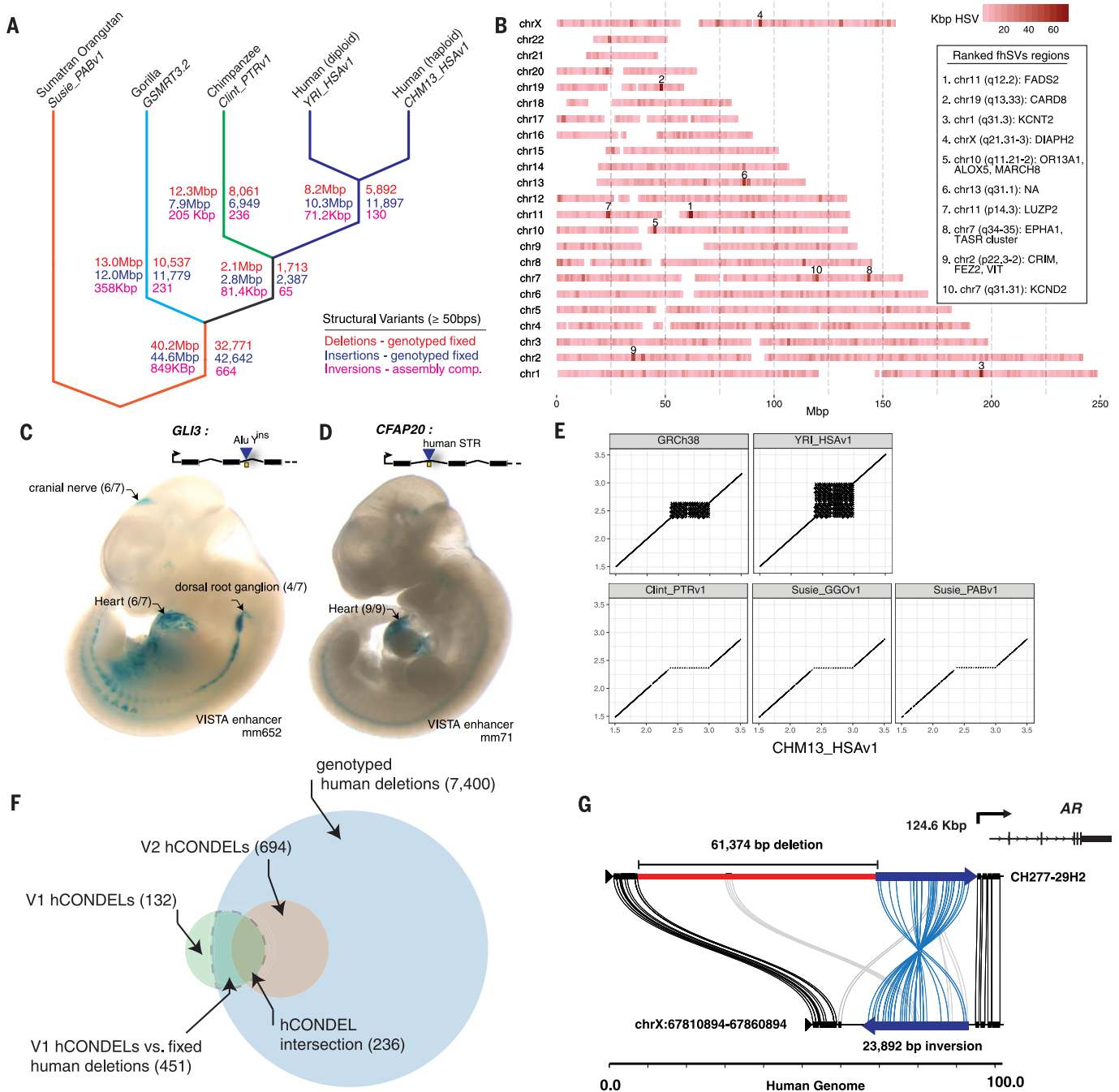
Fig. 3. Fixed structural variation and regulatory mutation. (**A**) The great ape cladogram with fixed structural variation assigned to lineages on the basis of assembly comparison, genotyping, and stratification (except for inversions). The total amount of sequence is shown on the left side of the branches, and the number of SVs is shown on the right for deletions (blue), insertions (red), and inversions (magenta). Inversions were assigned to branches on the basis of the comparison of our five assemblies because genotyping was less reliable. The cladogram was rooted against Susie_PABv1, meaning that the assignment of SVs to the orangutan or the common ancestor of human, chimpanzee, and gorilla is arbitrary. (**B**) A map of fhSVs. The color denotes the number of fhSVs bases (kbp), within a 1-Mbp sliding window (0.5-Mbp step). Each chromosome is labeled on the y axis. Key regions are annotated with genes. (**C**) The cell specificity for a mouse enhancer element (mm652, represented as a yellow box) that shares orthology in chimpanzee. In human, an AluY element has been inserted directly into the mm652

enhancer. GLI3, a GLI family zinc finger 3 gene. (**D**) A human-specific STR interrupts a mouse heart-specific enhancer shared with chimpanzee (yellow box). The STR is contained within a CFAP20 intron (CFAP20 encodes cilia and flagella associated protein 20). (**E**) Dot-plots of the human-specific STR expansion. The x axis corresponds to the CHM13_HSAv1 sequence (0.1-kbp units); the other sequences are on the y axis. The two human assemblies, CHM13_HSAv1 and YRI_HSAv1, show additional STR expansion relative to GRCh38, suggesting that the reference is collapsed. (**F**) A comparison of the hCONDEL set reported by McLean et al. (5) (V1) versus the hCONDELs reported here (V2). The current hCONDELs are from conservation (25-bp MSA windows) between chimpanzee, macaque, and mouse. The area enclosed by the dashed gray line shows the overlap between all fixed human deletions and all V1 hCONDELs. (**G**) A Miropeats diagram of the gorilla complex SV (inversion and deletion) upstream of the AR locus; the human reference genome is shown on the bottom (100-kbp window).
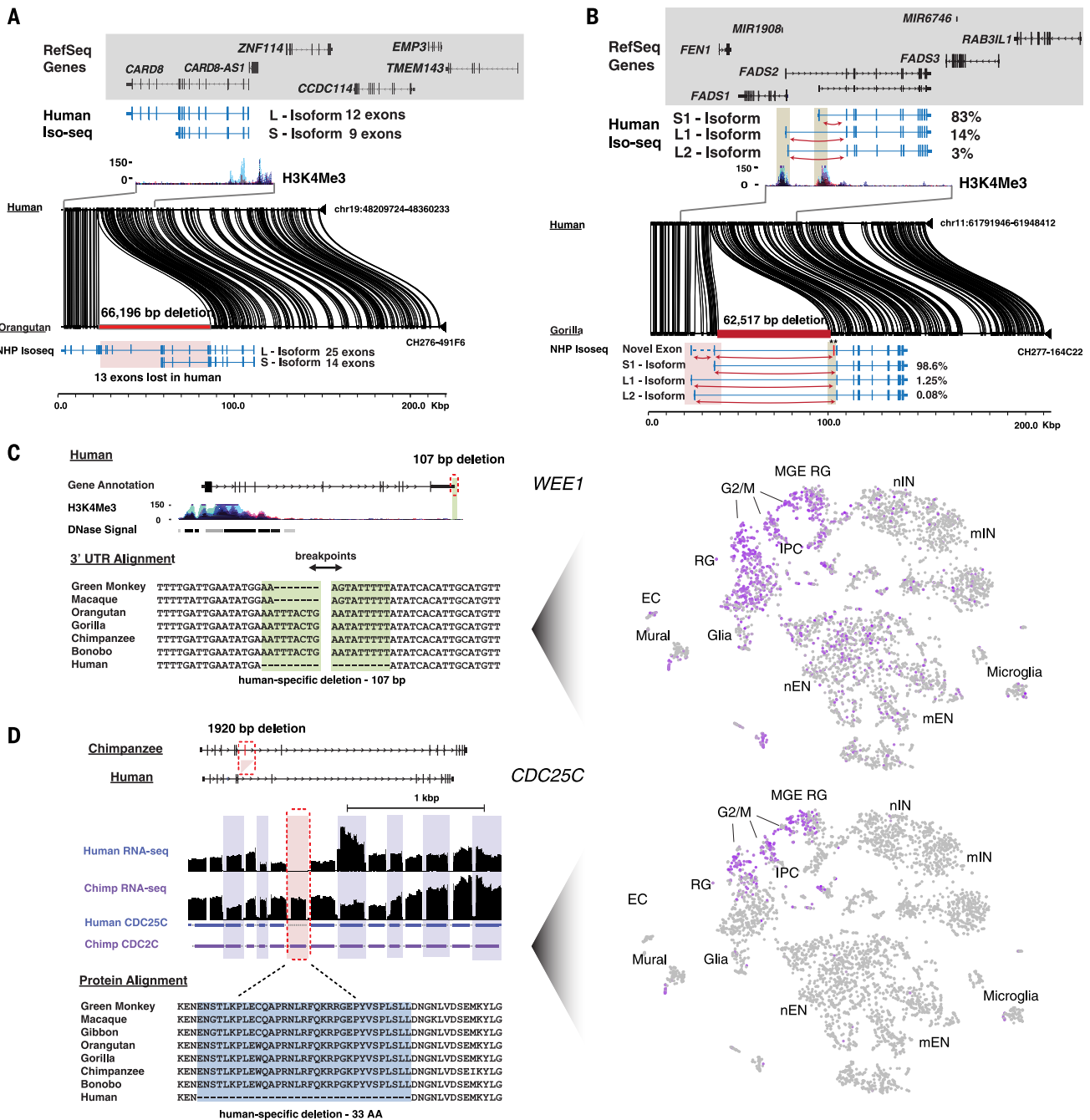
Fig. 4. Examples of intragenic human-specific structural variation. Shown are annotated MSAs between the human reference (GRCh38) and NHPs generated with Multiple Alignment using Fast Fourier Transform or visualized with Miropeats against sequenced large-insert primate clones. Single-cell gene expression for select genes is highlighted across 4261 cells from developing human telencephalon plotted using t-distributed stochastic neighbor embedding (tSNE) (66). (**A**) A 66.2-kbp intragenic deletion of *CARD8* removes 13 putative coding exons in human. Iso-Seq data from chimpanzee and human iPSCs identifies isoforms with and without the deleted exons, respectively. L, long; S, short; H3K4Me3, trimethylated histone H3 lysine 4. (**B**) A 62.5-kbp intergenic deletion of *FADS2* is found in humans, along with an altered isoform ratio: The relative abundance of the long isoforms is increased in humans relative to chimpanzee, as seen in the counts of junction-spanning short reads specific to each isoform. Additionally, a previously undescribed, rare (<5%) 75-bp exon is observed in chimpanzee and gorilla but absent in human, likely resulting from a human-specific splice-site mutation. (**C**) A 107-bp deletion in the 3′UTR of *WEE1* (red dashed box) reduces AU-rich sequence content in the mRNA. The tSNE plot illustrates that *WEE1* is highly expressed in cortical radial glia (RG), intermediate progenitor cells (IPCs), and medial ganglionic eminence progenitors (MGE RG) but shows limited expression in newborn and maturing inhibitory and excitatory neurons (nIN, mIN, nEN, and mEN, respectively), microglia, endothelial cells (ECs), and glia. DNase, deoxyribonuclease. (**D**) A 1920-bp deletion of cell-cycle regulator *CDC25C* (red dashed box) removes a 99-bp constitutive exon conserved in mouse, resulting in a 33–amino acid deletion and shorter N-terminal regulatory domain in humans. The tSNE plot illustrates that *CDC25C* shows restricted expression to telencephalon progenitors in the G2/M cell-cycle phase. Human and chimpanzee RNA-seq data were aligned directly to the exonic regions of *CDC25C*.
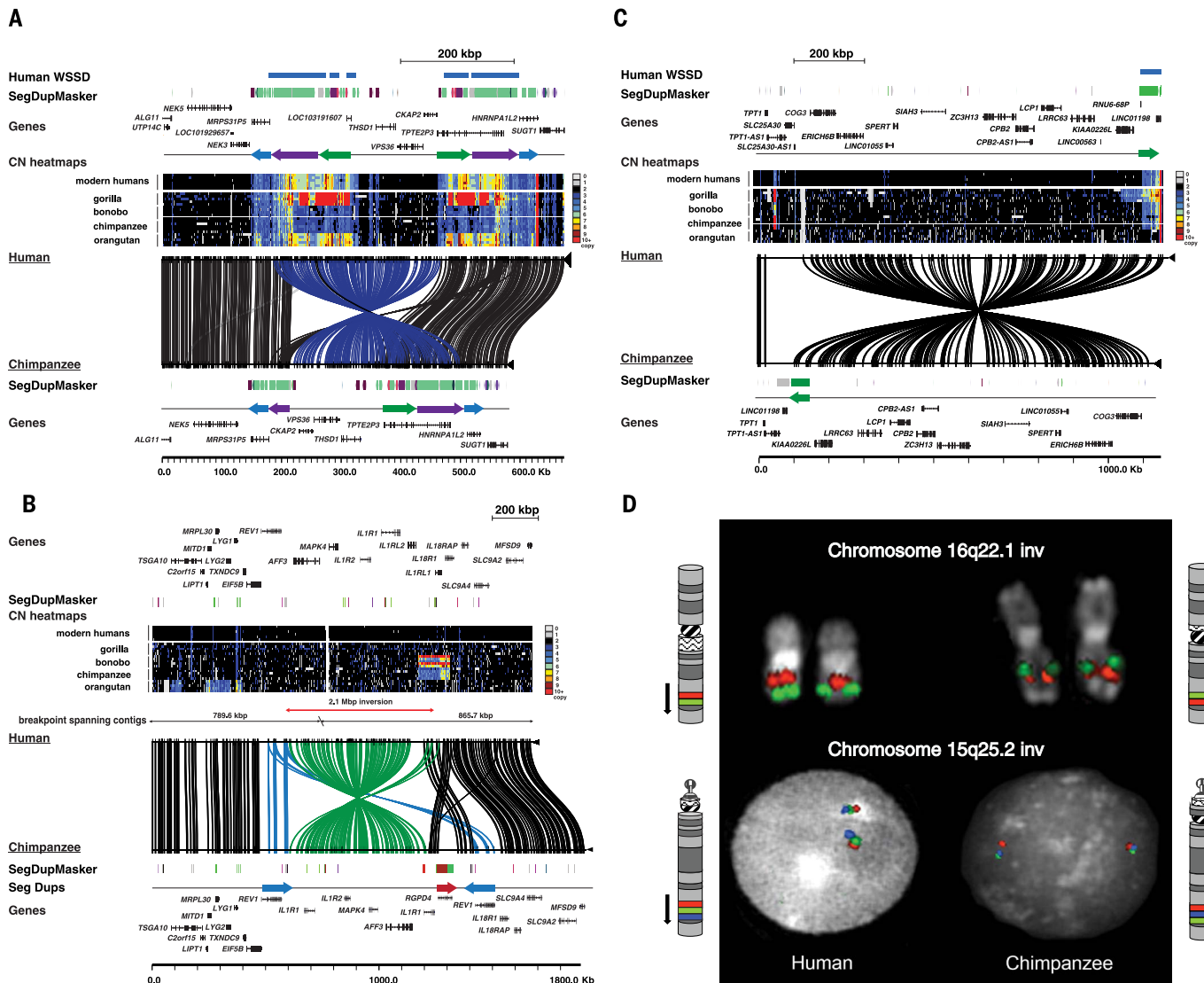
Fig. 5. Complex structural variation. Large-scale inversions between human and chimpanzee are depicted. The human reference genome sequence (GRCh38) with gene annotation is compared to large-insert clone-based assemblies from the chimpanzee BAC library CH251 using Miropeats. Connecting lines identify homologous regions of high sequence identity. SD organization is depicted with colored arrows, as defined by whole-genome shotgun sequence detection (WSSD) and DupMasker. Heatmap indicates copy number (CN) estimated by read depth from ape genome sequence. (**A**) A ~265-kbp inversion on chromosome 13q14.3 detected by optical mapping in chimpanzee (annotated blue lines). The inverted region is flanked by large ~180-kbp inverted SD blocks that vary with respect to copy number among great apes. (**B**) A 2.7-Mbp inversion on chromosome 2q12 to 13 detected by BAC end sequencing in chimpanzee (annotated green lines). The inverted region is flanked by duplication blocks containing lineage-specific expansions of the interleukins, an inverted duplication of *REV1*, and an additional copy of the *RGPD4* core duplicon. (**C**) A ~1.1-Mbp inversion at chromosome 13q14.13 identified by optical mapping in chimpanzee encompassing 15 genes. On the telomeric side of the inversion lies a ~60-kbp duplication block that demonstrates lineage-specific duplications in great apes. (**D**) Chromosome inversions, originally detected by optical mapping and BAC end sequencing, confirmed by metaphase analysis and interphase FISH experiments. A human-specific inversion of the chromosome 16q22.1 region was confirmed with orangutan clones CH276-89P20 (red) and CH276-192M7 (green) (top), and the 15q25.2 inversion was confirmed using chimpanzee clones CH251-321P13 (red), CH251-511D5 (green), and CH251-66E11 (blue) (bottom).

are more likely (*32*) to affect gene expression, we considered fhSV overlap on the basis of variant effect predictor annotations (including GRCh38 and Clint_PTRv1 annotation sets), which correlate both coding and noncoding variation to genes (Fig. 6A). Of the differentially expressed genes, 252 radial glia genes ($P = 9.78 \times 10^{-8}$, χ^2 test; 252/668) and 123 excitatory neuron genes ($P = 0.27$, χ^2 test; 123/360) had annotated fhSVs

associated with them. To test if this observation was an artifact of gene size, we shuffled fhSVs and counted the number of fhSVs that mapped within 50 kbp of a differentially expressed gene.

Overall, genes down-regulated in humans remain enriched for fhSVs, compared to the null distribution, whereas up-regulated genes did not show a significant overlap. In particular, genes down-regulated in human radial glial neural pro-

genitors showed significant enrichment for structural variation ($P = 0.02$; 10^4 permutations) (Fig. 6B). Although we observe the same trend in excitatory neurons, the effect did not reach significance. As a control, we repeated the same analysis for genes mapping to human-specific SDs (*54*), a form of structural variation not accessed in this study. Genes mapping to human SDs were up-regulated in radial glial and excitatory neurons
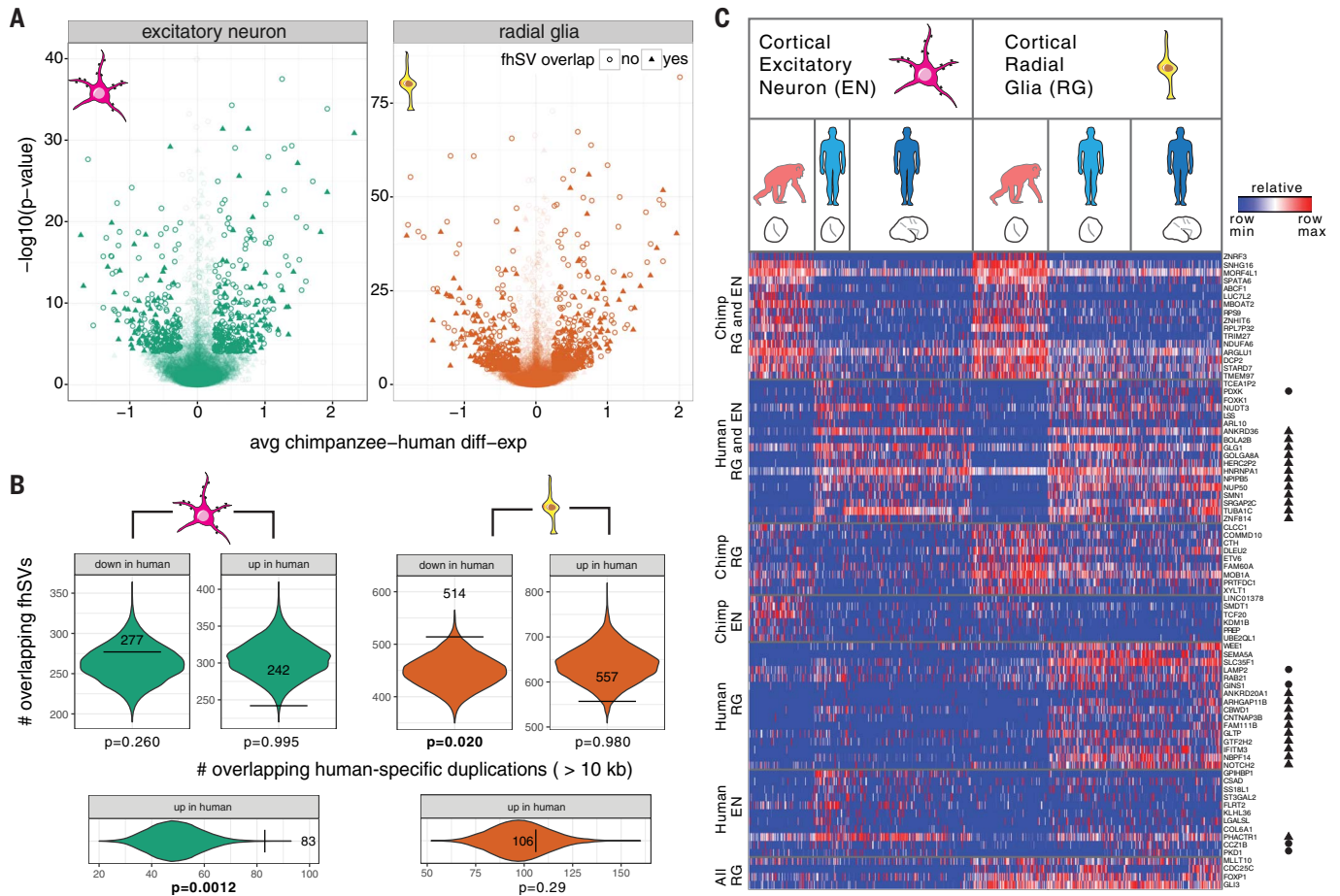
Fig. 6. Structural variation and neural progenitor expression differences between human and chimpanzee. (**A**) Volcano plots for chimpanzee-human gene expression in excitatory neuron (left) and radial glia (right) organoid single-cell data. Each point represents a gene, with sufficient data to assess significance between human and chimpanzee organoid cells. Genes with fhSVs within 50 kbp of their start or end are indicated with a triangle. The data points are shaded by significance (lighter shade indicates less significance). (**B**) Spatial permutation test for overlap between fhSVs and differentially expressed genes. Each violin plot shows the null distribution of human-specific SV overlap (±50 kbp of transcript start or end) with genes that are significantly differentially down- or up-regulated, relative to chimpanzee. The

horizontal bars and observed counts are overlaid on the null distribution. (**C**) Heatmap illustrating the percentile gene expression of differentially expressed genes near fhSVs (rows) across single cells (columns), including genes near the start or end of inversions (circles) and duplicated regions (WSSD) (triangles). Cells include 333 excitatory neurons (97 chimpanzee organoid, 53 human organoid, and 183 human primary cells) and 373 radial glia (113 chimpanzee organoid, 123 human organoid, and 137 human primary cells) (*56*, *57*). Expression patterns include concerted changes between chimpanzee and human cells across radial glia and excitatory neurons (chimp RG and EN and human RG and EN), cell-type-specific changes (human EN and human RG), and conserved radial glia expression (all RG).

Table 3. Summary of great ape genome structural variation. SV events (>50 bp in size) called against the human reference genome (GRCh38) using smartie-sv.

Statistic	Ape assembly				
	CHM13_HSAv1	**YRI_HSAv1**	**Clint_PTRv1**	**GSMRT3.2**	**Susie_PABv1**
Deletion count	9,126	11,747	63,634	73,681	136,980
Insertion count	14,962	14,528	68,589	76,230	142,631
Inversion count	74	55	446	533	969
Deletion size (Mbp)	4.76	4.85	41.88	45.48	84.76
Insertion size (Mbp)	6.85	7.17	40.34	47.53	120.35
Average deletion size (bp)	552	413	658	617	618
Average insertion size (bp)	458	493	588	623	843
Size of largest variation (kbp) [type]	84 [deletion]	124 [insertion]	133 [insertion]	90 [insertion]	123 [insertion]

when compared to chimpanzee (Fig. 6). This association identifies dozens of putative candidates for functional investigation, including some of the most differentially expressed genes between humans and chimpanzees in neural progenitor cells (Fig. 6 and table S14).

Discussion

Our great ape genome assemblies improve sequence contiguity by orders of magnitude (*20*, *60*), leading to a more comprehensive understanding of the evolution of structural variation. Coupling this effort with full-length cDNA sequencing improved gene annotation, especially for the discovery of transcripts and isoforms that have recently diverged between closely related species. Because genomes of species may be sequenced and assembled using the same platforms and experimental designs, we minimized biases introduced by ascertainment or an uneven sequencing quality between genomes.

These improved genomes yield a comprehensive view of intermediate-size structural variation among apes. As we focused on SVs that potentially disrupt genes or regulatory sequences, we began to address potential functional effect. Differential gene expression, especially in cortical radial glia, has been hypothesized to be a critical effector of brain size and a likely selective target of human brain evolution (*41*). Nearly 41% of the genes down-regulated in human radial glia, when compared to chimpanzee radial glial analogs from cerebral organoids, associate with an fhSV and most often as a deletion or a retroposon insertion. These findings are consistent with the "less-is-more" hypothesis (*61*), which argues that the loss of functional elements underlies critical aspects of human evolution. By contrast, human-specific gene duplications associate with up-regulated expression in both neural progenitors and excitatory neurons, although the effect is stronger for the latter. This finding is consistent with recent studies evidencing that human-specific SDs contribute to cortical differences between humans and chimpanzees (*6*–*8*). It is intriguing that the repeat-rich nature of ape genomes and, in particular, the expansion of SDs in the common ancestral lineage of the African ape lineage (*62*) may have made great ape genomes particularly prone to both deletion and duplication events, accelerating the rate of structural changes and large-effect mutations during the evolution of these species.

Despite this more comprehensive assessment of structural variation, not all SV types have been fully resolved among the great apes. In particular, we are still missing many larger, more complex events, including inversions and SDs that have differentially evolved between the lineages. For example, we recovered only one of five ape inversions identified by comparative BAC-based sequencing of a 2-Mbp region of chromosome 16p11.2 (*63*), although optical mapping techniques did identify four of the events. In this case, all inversions are flanked by large blocks of SDs (>200 kbp) that cannot be currently assembled by long-read whole-genome sequencing. We predict that such large, multi–mega–base pair inversions represent a common uncharacterized source of human-ape genetic variation that has been underestimated. Long-range sequencing and mapping technologies, such as Strand-seq (*49*), BAC-based sequencing (*63*), optical mapping (table S12), and longer-read sequencing (*64*) will be necessary to sequence resolve such large, more complex SVs.

Materials and Methods

We sequenced and assembled four genomes [chimpanzee (Clint), Sumatran orangutan (Susie), CHM13 (human), and YRI19240 (human)] using long-read PacBio RS II sequencing chemistry and the Falcon genome assembler. Sequence contigs were error-corrected using Quiver (*17*), Pilon (*18*), and a FreeBayes-based (*65*) indel correction pipeline. A chromosomal-level AGP was generated using optical maps (Bionano Genomics Saphyr platform) for scaffold building and bicolor FISH of ~700 large-insert clones. The Comparative Annotation Toolkit (CAT) (*22*) was used to annotate all of the great ape genomes using the human GENCODE V27 as reference with a combination of RNA-seq obtained from SRA as well as Iso-Seq data specifically from NHP iPSCs. STRs were defined using RepeatMasker v4.0.1 and Tandem Repeats multiple sequence Finder v4.07b. Syntenic regions and MSAs were constructed with MUSCLE (v3.8.31), phylogenetic analyses were performed using a general time-reversible model ("GTR+GAMMA") under a maximum likelihood RAxML (8.2.3) framework, and phylogenetic trees were generated using DendroP. A BLASR-based computational pipeline, smartie-sv, was developed to align, compare, and call insertions, deletions, and inversions (https://github.com/zeeev/smartie-sv). Insertions and deletions were genotyped against a panel of 45 ape genomes using SVTyper (paired-end) and WSSD (read depth). FISH and BAC clone sequencing was used to estimate sequence accuracy and validate the breakpoints of complex rearrangements. We compared SV locations with genes showing differential expression during human and chimpanzee cortical development using single-cell gene expression data from cerebral organoid models and from primary cortex.

REFERENCES AND NOTES

1. A. Varki, D. H. Geschwind, E. E. Eichler, Human uniqueness: Genome interactions with environment, behaviour and culture. *Nat. Rev. Genet.* **9**, 749–763 (2008). doi: 10.1038/nrg2428; pmid: 18802414
2. M. C. King, A. C. Wilson, Evolution at two levels in humans and chimpanzees. *Science* **188**, 107–116 (1975). doi: 10.1126/science.1090005; pmid: 1090005
3. A. Fortna *et al.*, Lineage-specific gene duplication and loss in human and great ape evolution. *PLOS Biol.* **2**, e207 (2004). doi: 10.1371/journal.pbio.0020207; pmid: 15252450
4. J. L. Boyd *et al.*, Human-chimpanzee differences in a FZD8 enhancer alter cell-cycle dynamics in the developing neocortex. *Curr. Biol.* **25**, 772–779 (2015). doi: 10.1016/j.cub.2015.01.041; pmid: 25702574
5. C. Y. McLean *et al.*, Human-specific loss of regulatory DNA and the evolution of human-specific traits. *Nature* **471**, 216–219 (2011). doi: 10.1038/nature09774; pmid: 21390129
6. M. Y. Dennis *et al.*, Evolution of human-specific neural SRGAP2 genes by incomplete segmental duplication.

7. C. Charrier *et al.*, Inhibition of SRGAP2 function by its human-specific paralogs induces neoteny during spine maturation. *Cell* **149**, 923–935 (2012). doi: 10.1016/j.cell.2012.03.034; pmid: 22559944
8. M. Florio *et al.*, Human-specific gene ARHGAP11B promotes basal progenitor amplification and neocortex expansion. *Science* **347**, 1465–1470 (2015). doi: 10.1126/science.aaa1975; pmid: 25721503
9. X.-C. Ju *et al.*, The hominoid-specific gene TBC1D3 promotes generation of basal neural progenitors and induces cortical folding in mice. *eLife* **5**, 206 (2016). doi: 10.7554/eLife.18197; pmid: 27504805
10. A. Scally *et al.*, Insights into hominid evolution from the gorilla genome sequence. *Nature* **483**, 169–175 (2012). doi: 10.1038/nature10842; pmid: 22398555
11. D. P. Locke *et al.*, Comparative and demographic analysis of orangutan genomes. *Nature* **469**, 529–533 (2011). doi: 10.1038/nature09687; pmid: 21270892
12. Chimpanzee Sequencing and Analysis Consortium, Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69–87 (2005). doi: 10.1038/nature04072; pmid: 16136131
13. E. T. Lam *et al.*, Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nat. Biotechnol.* **30**, 771–776 (2012). doi: 10.1038/nbt.2303; pmid: 22797562
14. J. N. Burton *et al.*, Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat. Biotechnol.* **31**, 1119–1125 (2013). doi: 10.1038/nbt.2727; pmid: 24185095
15. J. Eid *et al.*, Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133–138 (2009). doi: 10.1126/science.1162986; pmid: 19023044
16. Materials and methods are available as supplementary materials.
17. C.-S. Chin *et al.*, Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10**, 563–569 (2013). doi: 10.1038/nmeth.2474; pmid: 23644548
18. B. J. Walker *et al.*, Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLOS ONE* **9**, e112963 (2014). doi: 10.1371/journal.pone.0112963; pmid: 25409509
19. J. J. Yunis, O. Prakash, The origin of man: A chromosomal pictorial legacy. *Science* **215**, 1525–1530 (1982). doi: 10.1126/science.7063861; pmid: 7063861
20. D. Gordon *et al.*, Long-read sequence assembly of the gorilla genome. *Science* **352**, aae0344 (2016). doi: 10.1126/science.aae0344; pmid: 27034376
21. L. F. K. Kuderna *et al.*, A 3-way hybrid approach to generate a new high-quality chimpanzee reference genome (Pan_tro_3.0). *Gigascience* **6**, 1–6 (2017). doi: 10.1093/gigascience/gix098; pmid: 29092041
22. I. T. Fiddes, J. Armstrong, M. Diekhans, S. Nachtweide, Z. N. Kronenberg, J. G. Underwood, D. Gordon, D. Earl, T. Keane, E. E. Eichler, D. Haussler, M. Stanke, B. Paten, Comparative Annotation Toolkit (CAT)—simultaneous clade and personal genome annotation. bioRxiv 231118 [Preprint]. 8 December 2017. doi: 10.1101/231118
23. N. Elango, J. W. Thomas, S. V. Yi; NISC Comparative Sequencing Program, Variable molecular clocks in hominoids. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 1370–1375 (2006). doi: 10.1073/pnas.0510716103; pmid: 16432233
24. P. Moorjani, C. E. G. Amorim, P. F. Arndt, M. Przeworski, Variation in the molecular clock of primates. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 10607–10612 (2016). doi: 10.1073/pnas.1600374113; pmid: 27601674
25. W. H. Li, M. Tanimura, P. M. Sharp, An evaluation of the molecular clock hypothesis using mammalian DNA sequences. *J. Mol. Evol.* **25**, 330–342 (1987). doi: 10.1007/BF02603118; pmid: 3118047
26. D. M. Bickhart *et al.*, Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nat. Genet.* **49**, 643–650 (2017). doi: 10.1038/ng.3802; pmid: 28263316
27. D. C. Rubinsztein *et al.*, Microsatellite evolution—evidence for directionality and variation in rate between species. *Nat. Genet.* **10**, 337–343 (1995). doi: 10.1038/ng0795-337; pmid: 7670473
28. M. T. Webster, N. G. C. Smith, H. Ellegren, Microsatellite evolution inferred from human-chimpanzee genomic

sequence alignments. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 8748–8753 (2002). doi: 10.1073/pnas.122067599; pmid: 12070344

29. C. T. Yohn *et al.*, Lineage-specific expansions of retroviral insertions within the genomes of African great apes but not humans and orangutans. *PLOS Biol.* **3**, e110 (2005). doi: 10.1371/journal.pbio.0030110; pmid: 15737067

30. N. Polavarapu, N. J. Bowen, J. F. McDonald, Identification, characterization and comparative genomics of chimpanzee endogenous retroviruses. *Genome Biol.* **7**, R51 (2006). doi: 10.1186/gb-2006-7-6-r51; pmid: 16805923

31. S. M. Kaiser, H. S. Malik, M. Emerman, Restriction of an extinct retrovirus by the human TRIM5alpha antiviral protein. *Science* **316**, 1756–1758 (2007). doi: 10.1126/science.1140579; pmid: 17588933

32. P. H. Sudmant *et al.*, An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015). doi: 10.1038/nature15394; pmid: 26432246

33. J. Prado-Martinez *et al.*, Great ape genetic diversity and population history. *Nature* **499**, 471–475 (2013). doi: 10.1038/nature12228; pmid: 23823723

34. Y. A. Pérez-Rico *et al.*, Comparative analyses of super-enhancers reveal conserved elements in vertebrate genomes. *Genome Res.* **27**, 259–268 (2017). doi: 10.1101/gr.203679.115; pmid: 27965291

35. P. L. Reno *et al.*, A penile spine/vibrissa enhancer sequence is missing in modern and extinct humans but is retained in multiple primates with penile spines and sensory vibrissae. *PLOS ONE* **8**, e84258 (2013). doi: 10.1371/journal.pone.0084258; pmid: 24367647

36. A. Ameur *et al.*, Genetic adaptation of fatty-acid metabolism: A human-specific haplotype increasing the biosynthesis of long-chain omega-3 and omega-6 fatty acids. *Am. J. Hum. Genet.* **90**, 809–820 (2012). doi: 10.1016/j.ajhg.2012.03.014; pmid: 22503634

37. K. Ye, F. Gao, D. Wang, O. Bar-Yosef, A. Keinan, Dietary adaptation of FADS genes in Europe varied across time and geography. *Nat. Ecol. Evol.* **1**, 0167 (2017). doi: 10.1038/s41559-017-0167; pmid: 29094686

38. M. T. Buckley *et al.*, Selection in Europeans on fatty acid desaturases associated with dietary changes. *Mol. Biol. Evol.* **34**, 1307–1318 (2017). doi: 10.1093/molbev/msx103; pmid: 28333262

39. T. N. Petersen, S. Brunak, G. von Heijne, H. Nielsen, SignalP 4.0: Discriminating signal peptides from transmembrane regions. *Nat. Methods* **8**, 785–786 (2011). doi: 10.1038/nmeth.1701; pmid: 21959131

40. N. B. Trunnell, A. C. Poon, S. Y. Kim, J. E. Ferrell Jr., Ultrasensitivity in the Regulation of Cdc25C by Cdk1. *Mol. Cell* **41**, 263–274 (2011). doi: 10.1016/j.molcel.2011.01.012; pmid: 21292159

41. P. Rakic, A small step for the cell, a giant leap for mankind: A hypothesis of neocortical expansion during evolution. *Trends Neurosci.* **18**, 383–388 (1995). doi: 10.1016/0166-2236(95)93934-P; pmid: 7482803

42. M. Pendleton *et al.*, Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat. Methods* **12**, 780–786 (2015). doi: 10.1038/nmeth.3454; pmid: 26121404

43. A. C. Y. Mak *et al.*, Genome-wide structural variation detection by genome mapping on nanochannel arrays. *Genetics* **202**, 351–362 (2016). doi: 10.1534/genetics.115.183483; pmid: 26510793

44. L. Feuk *et al.*, Discovery of human inversion polymorphisms by comparative analysis of human and chimpanzee DNA sequence assemblies. *PLOS Genet.* **1**, e56 (2005). doi: 10.1371/journal.pgen.0010056; pmid: 16254605

45. T. L. Newman *et al.*, A genome-wide survey of structural variation between human and chimpanzee. *Genome Res.* **15**, 1344–1356 (2005). doi: 10.1101/gr.4338005; pmid: 16169929

46. J. M. Szamalek *et al.*, Polymorphic micro-inversions contribute to the genomic variability of humans and chimpanzees. *Hum. Genet.* **119**, 103–112 (2006). doi: 10.1007/s00439-005-0117-6; pmid: 16362346

47. M. F. Cardone *et al.*, Hominoid chromosomal rearrangements on 17q map to complex regions of segmental duplication.

Genome Biol. **9**, R28 (2008). doi: 10.1186/gb-2008-9-2-r28; pmid: 18257913

48. M. C. Zody *et al.*, Evolutionary toggling of the MAPT 17q21.31 inversion region. *Nat. Genet.* **40**, 1076–1083 (2008). doi: 10.1038/ng.193; pmid: 19165922

49. A. D. Sanders *et al.*, Characterizing polymorphic inversions in human genomes by single-cell sequencing. *Genome Res.* **26**, 1575–1587 (2016). doi: 10.1101/gr.201160.115; pmid: 27472961

50. M. J. P. Chaisson, A. D. Sanders, X. Zhao, A. Malhotra, D. Porubsky, T. Rausch, E. J. Gardner, O. Rodriguez, L. Guo, R. L. Collins, X. Fan, J. Wen, R. E. Handsaker, S. Fairley, Z. N. Kronenberg, X. Kong, F. Hormozdiari, D. Lee, A. M. Wenger, A. Hastie, D. Antaki, P. Audano, H. Brand, S. Cantsilieris, H. Cao, E. Cerveira, C. Chen, X. Chen, C.-S. Chin, Z. Chong, N. T. Chuang, D. M. Church, L. Clarke, A. Farrell, J. Flores, T. Galeev, G. David, M. Gujral, V. Guryev, W. Haynes-Heaton, J. Korlach, S. Kumar, J. Y. Kwon, J. E. Lee, J. Lee, W.-P. Lee, S. P. Lee, P. Marks, K. Valud-Martinez, S. Meiers, K. M. Munson, F. Navarro, B. J. Nelson, C. Nodzak, A. Noor, S. Kyriazopoulou-Panagiotopoulou, A. Pang, Y. Qiu, G. Rosanio, M. Ryan, A. Stutz, D. C. J. Spierings, A. Ward, A. E. Welsch, M. Xiao, W. Xu, C. Zhang, Q. Zhu, X. Zheng-Bradley, G. Jun, L. Ding, C. L. Koh, B. Ren, P. Flicek, K. Chen, M. B. Gerstein, P.-Y. Kwok, P. M. Lansdorp, G. Marth, J. Sebat, X. Shi, A. Bashir, K. Ye, S. E. Devine, M. Talkowski, R. E. Mills, T. Marschall, J. Korbel, E. E. Eichler, C. Lee, Multi-platform discovery of haplotype-resolved structural variation in human genomes. bioRxiv 193144 [Preprint]. 23 September 2017. doi: 10.1101/193144

51. B. P. Coe *et al.*, Refining analyses of copy number variation identifies specific genes associated with developmental delay. *Nat. Genet.* **46**, 1063–1071 (2014). doi: 10.1038/ng.3092; pmid: 25217958

52. E. R. Sturgill *et al.*, Biosynthesis of the major brain gangliosides GD1a and GT1b. *Glycobiology* **22**, 1289–1301 (2012). doi: 10.1093/glycob/cws103; pmid: 22735313

53. S. Herculano-Houzel, The human brain in numbers: A linearly scaled-up primate brain. *Front. Hum. Neurosci.* **3**, 31 (2009). doi: 10.3389/neuro.09.031.2009; pmid: 19915731

54. M. Y. Dennis *et al.*, The evolution and population diversity of human-specific segmental duplications. *Nat. Ecol. Evol.* **1**, 0069 (2017). doi: 10.1038/s41559-016-0069; pmid: 28580430

55. J. G. Camp *et al.*, Human cerebral organoids recapitulate gene expression programs of fetal neocortex development. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 15672–15677 (2015). pmid: 26644564

56. F. Mora-Bermúdez *et al.*, Differences and similarities between human and chimpanzee neural progenitors during cerebral cortex development. *eLife* **5**, 166 (2016). doi: 10.7554/eLife.18683; pmid: 27669147

57. A. A. Pollen *et al.*, Molecular identity of human outer radial glia during cortical development. *Cell* **163**, 55–67 (2015). doi: 10.1016/j.cell.2015.09.004; pmid: 26406371

58. Z. He *et al.*, Comprehensive transcriptome analysis of neocortical layers in humans, chimpanzees and macaques. *Nat. Neurosci.* **20**, 886–895 (2017). doi: 10.1038/nn.4548; pmid: 28414332

59. M. C. N. Marchetto *et al.*, Differential L1 regulation in pluripotent stem cells of humans and apes. *Nature* **503**, 525–529 (2013). doi: 10.1038/nature12686; pmid: 24153179

60. J. Korlach *et al.*, De novo PacBio long-read and phased avian genome assemblies correct and add to reference genes generated with intermediate and short reads. *Gigascience* **6**, 1–16 (2017). doi: 10.1093/gigascience/gix085; pmid: 29020750

61. M. V. Olson, When less is more: Gene loss as an engine of evolutionary change. *Am. J. Hum. Genet.* **64**, 18–23 (1999). doi: 10.1086/302219; pmid: 9915938

62. T. Marques-Bonet *et al.*, A burst of segmental duplications in the genome of the African great ape ancestor. *Nature* **457**, 877–881 (2009). doi: 10.1038/nature07744; pmid: 19212409

63. X. Nuttle *et al.*, Emergence of a *Homo sapiens*-specific gene family and chromosome 16p11.2 CNV susceptibility. *Nature* **536**, 205–209 (2016). doi: 10.1038/nature19075; pmid: 27487209

64. M. Jain, H. E. Olsen, B. Paten, M. Akeson, The Oxford Nanopore MinION: Delivery of nanopore sequencing to the genomics

community. *Genome Biol.* **17**, 239 (2016). doi: 10.1186/s13059-016-1103-0; pmid: 27887629

65. E. Garrison, G. Marth, Haplotype-based variant detection from short-read sequencing. arXiv:1207.3907 [q-bio.GN] (17 July 2012).

66. T. J. Nowakowski *et al.*, Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. *Science* **358**, 1318–1323 (2017). doi: 10.1126/science.aap8809; pmid: 29217575

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/360/6393/eaar6343/suppl/DC1
Materials and Methods
Figs. S1 to S53
Tables S1 to S58
References (67–104)

Science

High-resolution comparative analysis of great ape genomes

Zev N. Kronenberg, Ian T. Fiddes, David Gordon, Shwetha Murali, Stuart Cantsilieris, Olivia S. Meyerson, Jason G. Underwood, Bradley J. Nelson, Mark J. P. Chaisson, Max L. Dougherty, Katherine M. Munson, Alex R. Hastie, Mark Diekhans, Fereydoun Hormozdiari, Nicola Lorusso, Kendra Hoekzema, Ruolan Qiu, Karen Clark, Archana Raja, AnneMarie E. Welch, Melanie Sorensen, Carl Baker, Robert S. Fulton, Joel Armstrong, Tina A. Graves-Lindsay, Ahmet M. Denli, Emma R. Hoppe, PingHsun Hsieh, Christopher M. Hill, Andy Wing Chun Pang, Joyce Lee, Ernest T. Lam, Susan K. Dutcher, Fred H. Gage, Wesley C. Warren, Jay Shendure, David Haussler, Valerie A. Schneider, Han Cao, Mario Ventura, Richard K. Wilson, Benedict Paten, Alex Pollen and Evan E. Eichler

A spotlight on great ape genomes

Most nonhuman primate genomes generated to date have been "humanized" owing to their many gaps and the reliance on guidance by the reference human genome. To remove this humanizing effect, Kronenberg *et al.* generated and assembled long-read genomes of a chimpanzee, an orangutan, and two humans and compared them with a previously generated gorilla genome. This analysis recognized genomic structural variation specific to humans and particular ape lineages. Comparisons between human and chimpanzee cerebral organoids showed down-regulation of the expression of specific genes in humans, relative to chimpanzees, related to noncoding variation identified in this analysis.

Science, this issue p. eaar6343

ARTICLE TOOLS	http://science.sciencemag.org/content/360/6393/eaar6343
SUPPLEMENTARY MATERIALS	http://science.sciencemag.org/content/suppl/2018/06/06/360.6393.eaar6343.DC1
REFERENCES	This article cites 102 articles, 30 of which you can access for free http://science.sciencemag.org/content/360/6393/eaar6343#BIBL
PERMISSIONS	http://www.sciencemag.org/help/reprints-and-permissions

Use of this article is subject to the Terms of Service