nature ecology & evolution

SUPPLEMENTARY INFORMATION

VOLUME: 1 | ARTICLE NUMBER: 0069

In the format provided by the authors and unedited.

The evolution and population diversity of human-specific segmental duplications

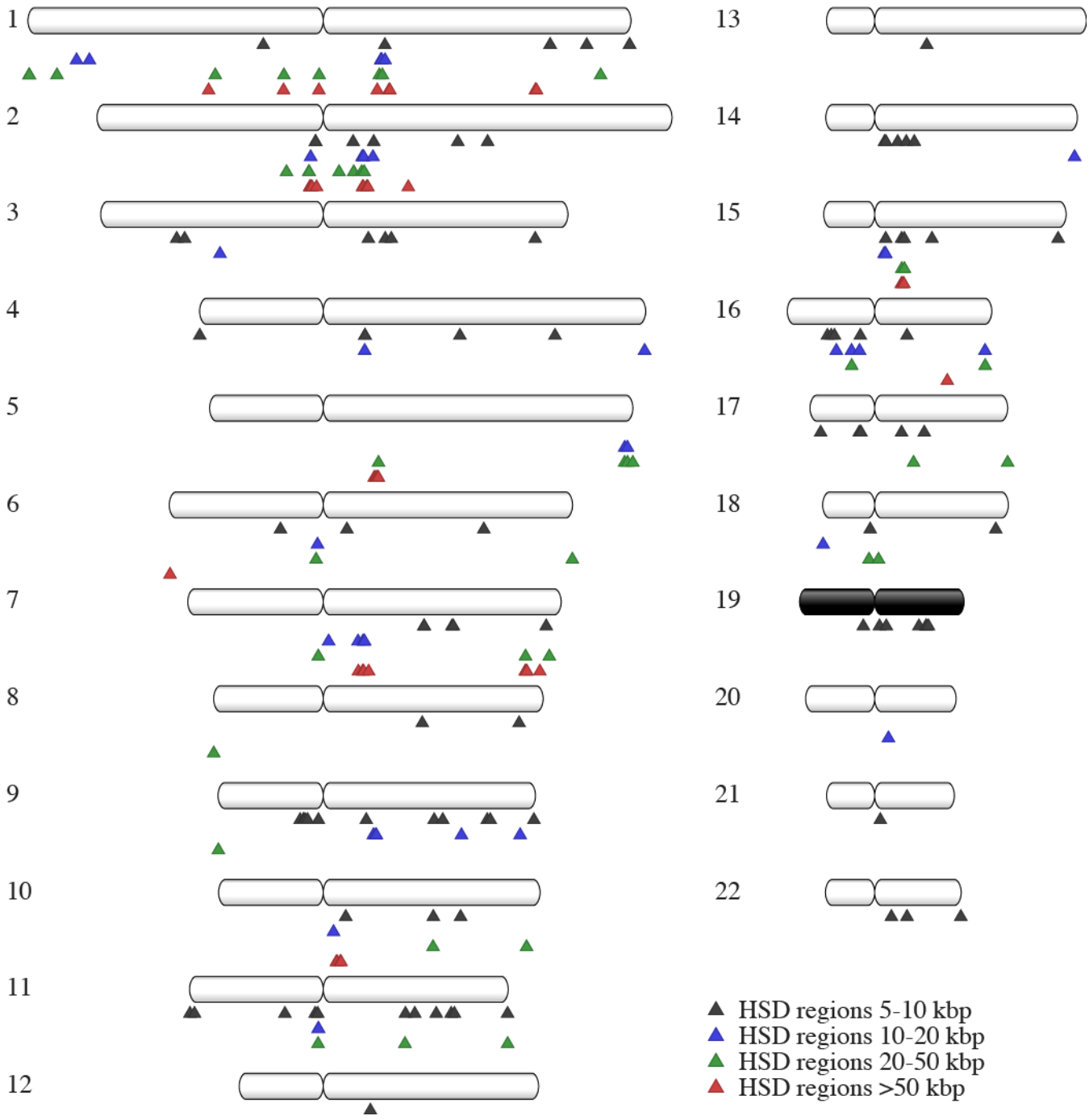
Megan Y. Dennis[1,2], Lana Harshman[2], Bradley J. Nelson[2], Osnat Penn[2], Stuart Cantsilieris[2], John Huddleston[2,3], Francesca Antonacci[4], Kelsi Penewit[2], Laura Denman[2], Archana Raja[2,3], Carl Baker[2], Kenneth Mark[2], Maika Malig[2], Nicolette Janke[2], Claudia Espinoza[2], Holly A.F. Stessman[2], Xander Nuttle[2], Kendra Hoekzema[2], Tina A. Lindsay-Graves[5], Richard K. Wilson[5], Evan E. Eichler[2,3*]

[1]Genome Center, MIND Institute, and Department of Biochemistry & Molecular Medicine, University of California, Davis, CA 95616, USA
[2]Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA 98195, USA
[3]Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195, USA
[4]Dipartimento di Biologia, Università degli Studi di Bari "Aldo Moro", Bari 70125, Italy
[5]McDonnell Genome Institute at Washington University, Washington University School of Medicine, St. Louis, MO 63108, USA

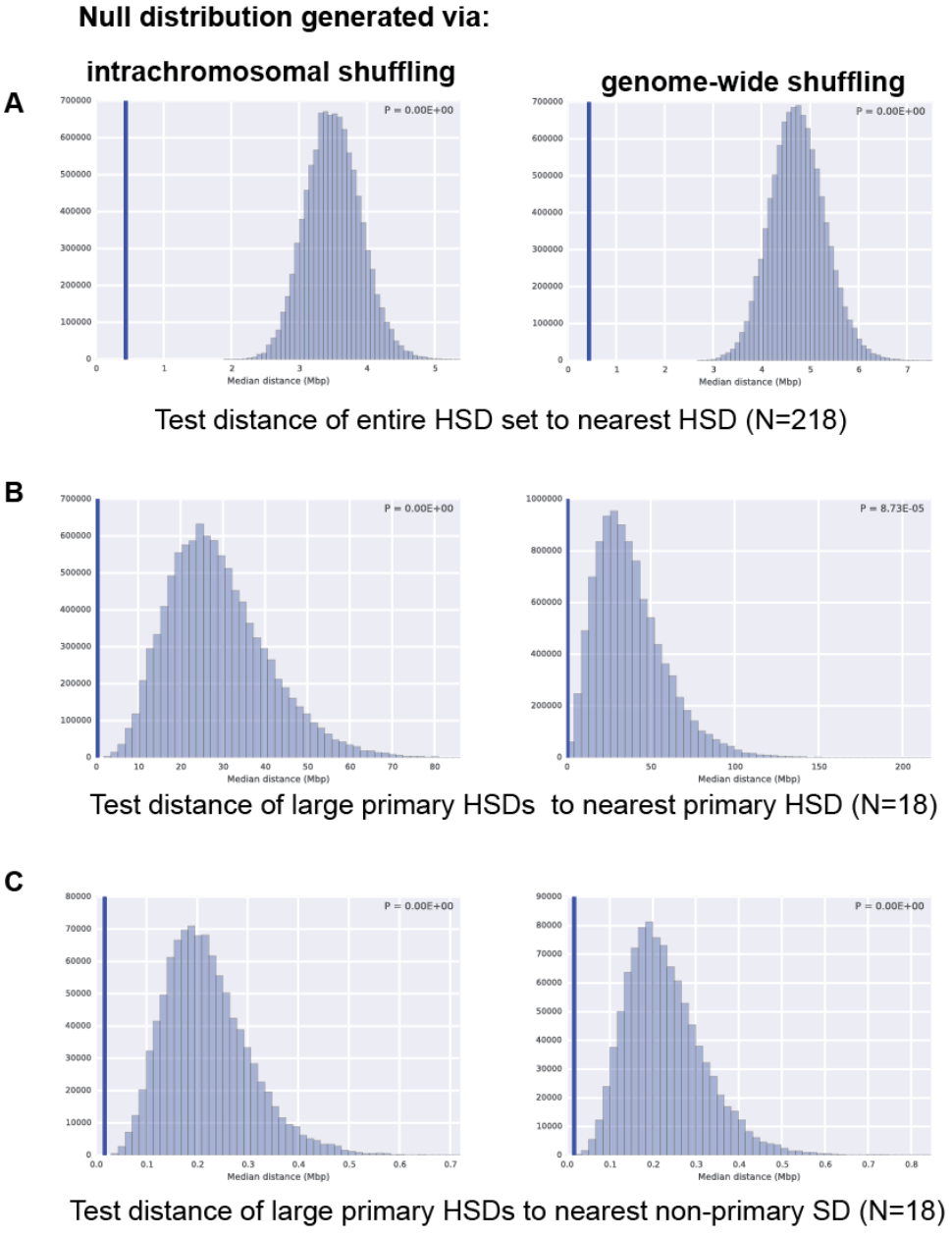
[*]Corresponding author: Evan E. Eichler, Ph.D.
University of Washington School of Medicine
Howard Hughes Medical Institute
Box 355065
Foege S413C, 3720 15th Ave NE
Seattle, WA 98195
E-mail: eee@gs.washington.edu

TABLE OF CONTENTS

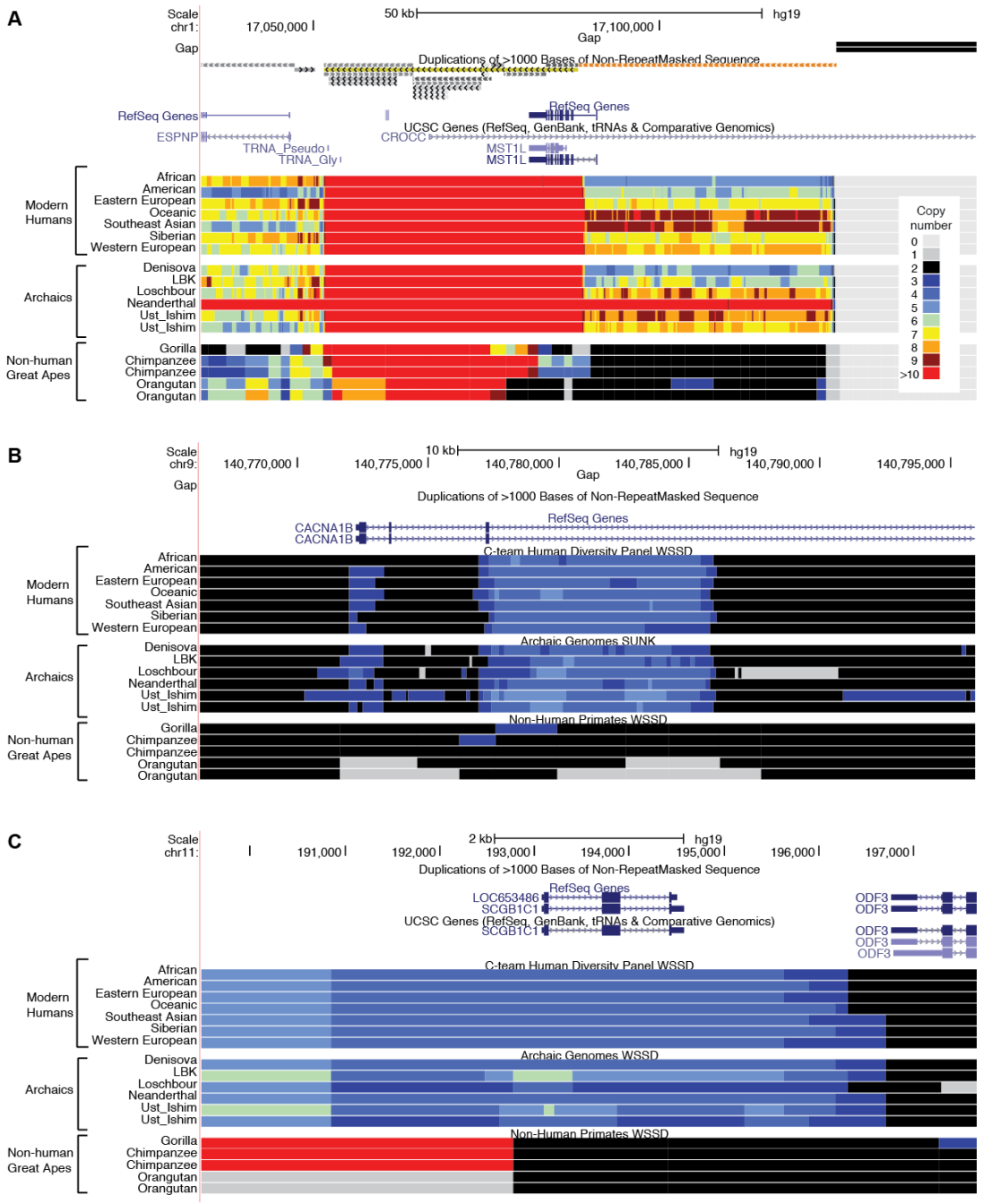
SUPPLEMENTARY FIGURES



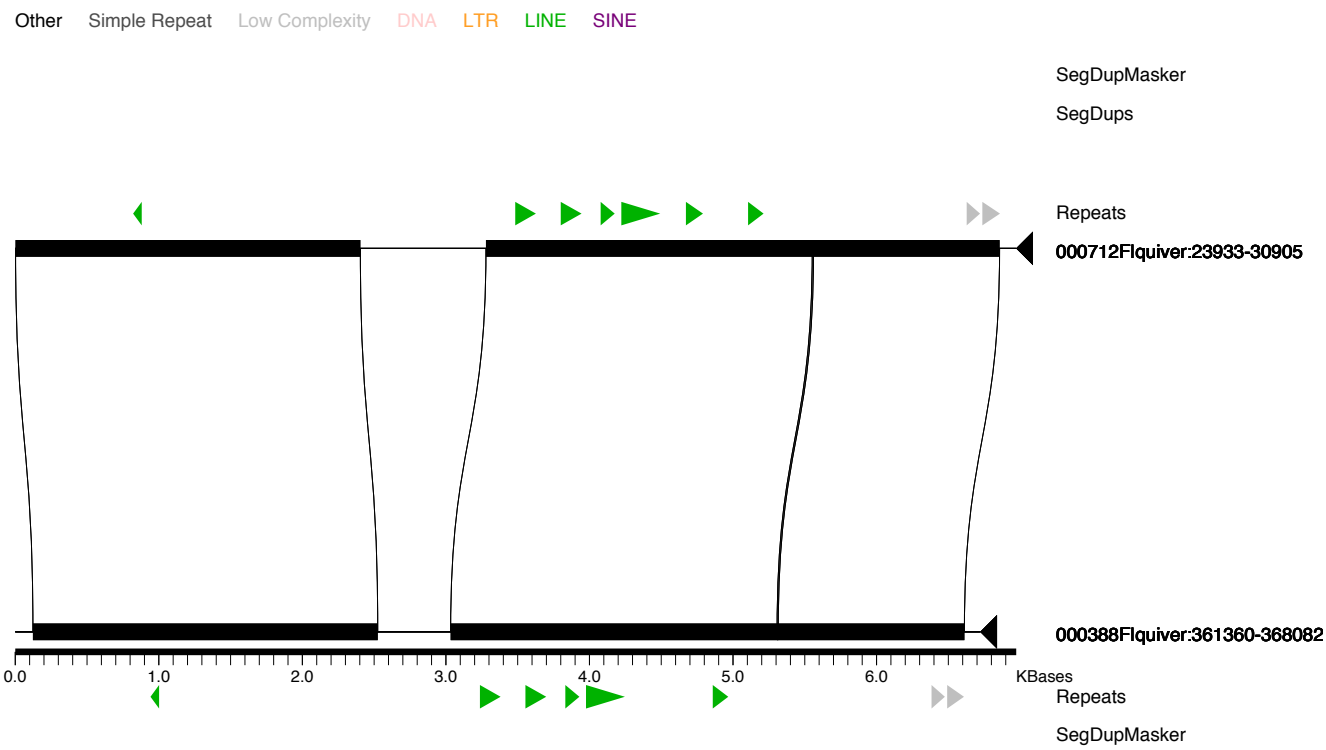
Supplementary Figure 1. Location of human-specific duplicated regions mapped onto chromosomal ideograms. HSDs identified by comparisons of overall Illumina sequence read depth of a diverse panel of 236 human and 86 chimpanzee, gorilla and orangutan genomes to the human reference genome (GRCh37) identified regions of the genome uniquely duplicated in humans (colored triangles mapped to autosomal chromosomes with numbers depicted to their left; Supplementary Table 1). Analysis was performed only on human autosomes.



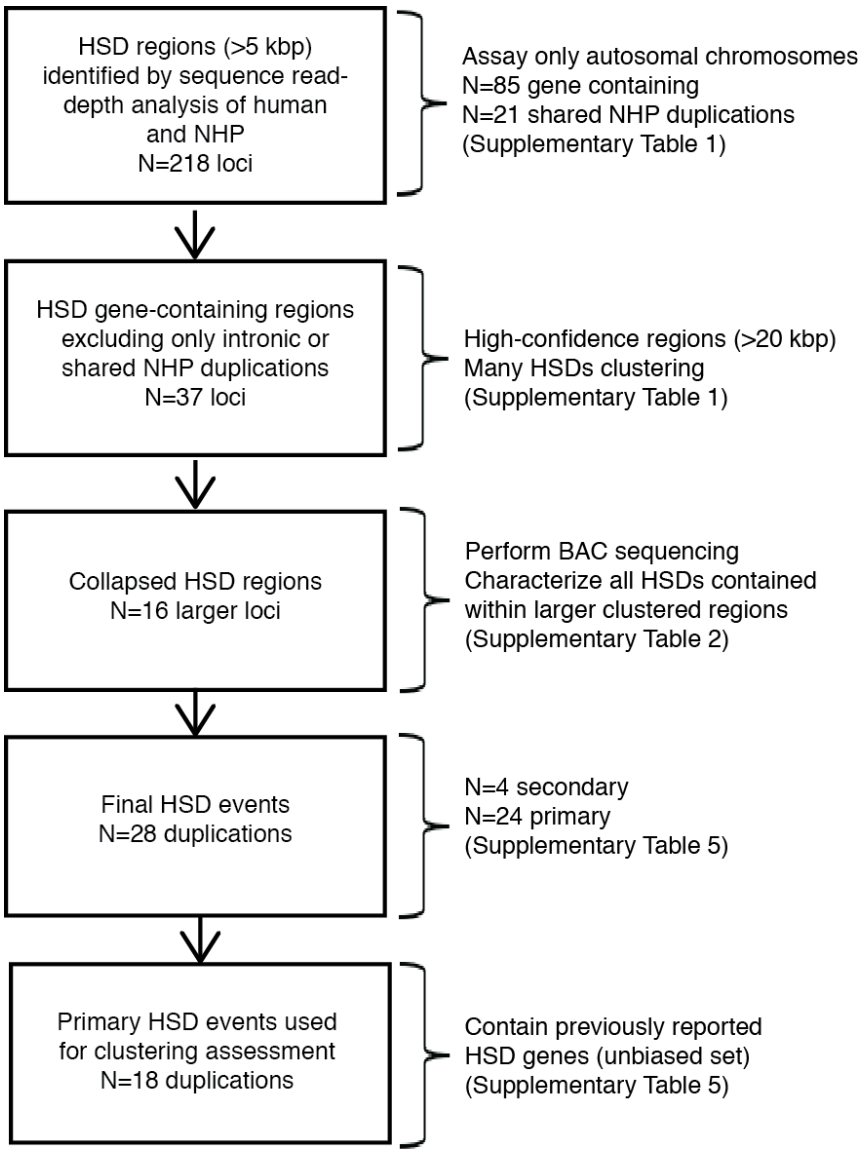
Supplementary Figure 2. HSDs significantly enriched near each other and also other non-primary SDs than expected by chance. We simulated a null distribution by shuffling all **(A)** 218 identified HSDs within the same chromosome 10 million times and, when multiple duplications occurred on a single chromosome, calculated the distance to the nearest duplication using midpoint coordinates. We calculated the median distance for each iteration of the simulation and compared this distribution (pictured as histograms) to the empirical value (indicated with the blue line). The empirical median distance between the nearest HSDs of 440,707 bp was significantly less than a null distribution generated by intrachromosomal and genome-wide shuffling ($P < 1x10^{-7}$). **(B)** The same simulation was performed for 18 primary duplications resulting in an empirical median distance between duplications of 376,575 bp, which was significantly less than a null distribution generated by intrachromosomal ($P < 1x10^{-7}$) and genome-wide ($P = 8.73x10^{-5}$) shuffling. **(C)** Likewise, the empirical median distance of 17,523 bp to the nearest SD (as defined by whole-genome analysis comparison (WGAC) [1]) was significantly less than the null distribution (1 million simulations) performed with intrachromosomal and genome-wide shuffling ($P < 1x10^{-6}$).



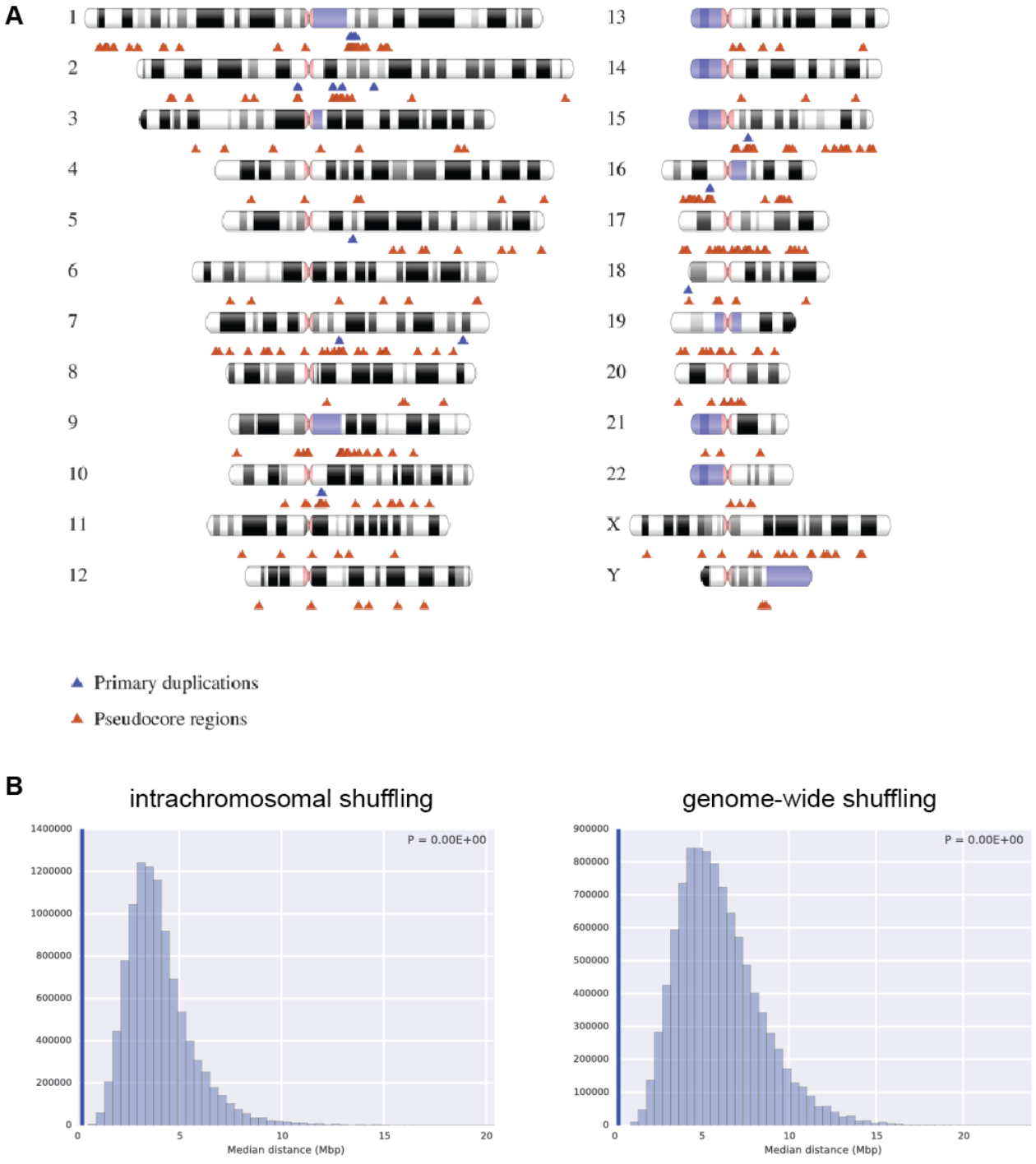
Supplementary Figure 3. Novel human-specific duplicated genes. Pictured are UCSC Genome Browser snapshots (human build GRCh37) of three genes, **(A)** *MST1L*, **(B)** *CACNA1B*, and **(C)** *SCGB1C1*, showing expansions in human compared to other great apes. Starting from top to bottom, tracks include annotated reference gaps, WGAC SDs (>1 kbp, >90% sequence identity; color coded: gray = lowest sequence identity; orange = highest sequence identity), gene annotations from RefSeq and UCSC gene libraries, and copy number (CN) heatmaps, with CN index shown, produced from Illumina read-depth predictions (see Supplementary Methods) for representative modern humans from the Human Genome Diversity Project (HGDP) cohort, archaic humans, and nonhuman primates (NHPs). Note, for *CACNA1B* and *SCGB1C1*, the WGAC analysis shown by the WGAC track at the top of the browser image does not identify an SD at these loci suggesting the paralogous sequence is missing from the reference genome.



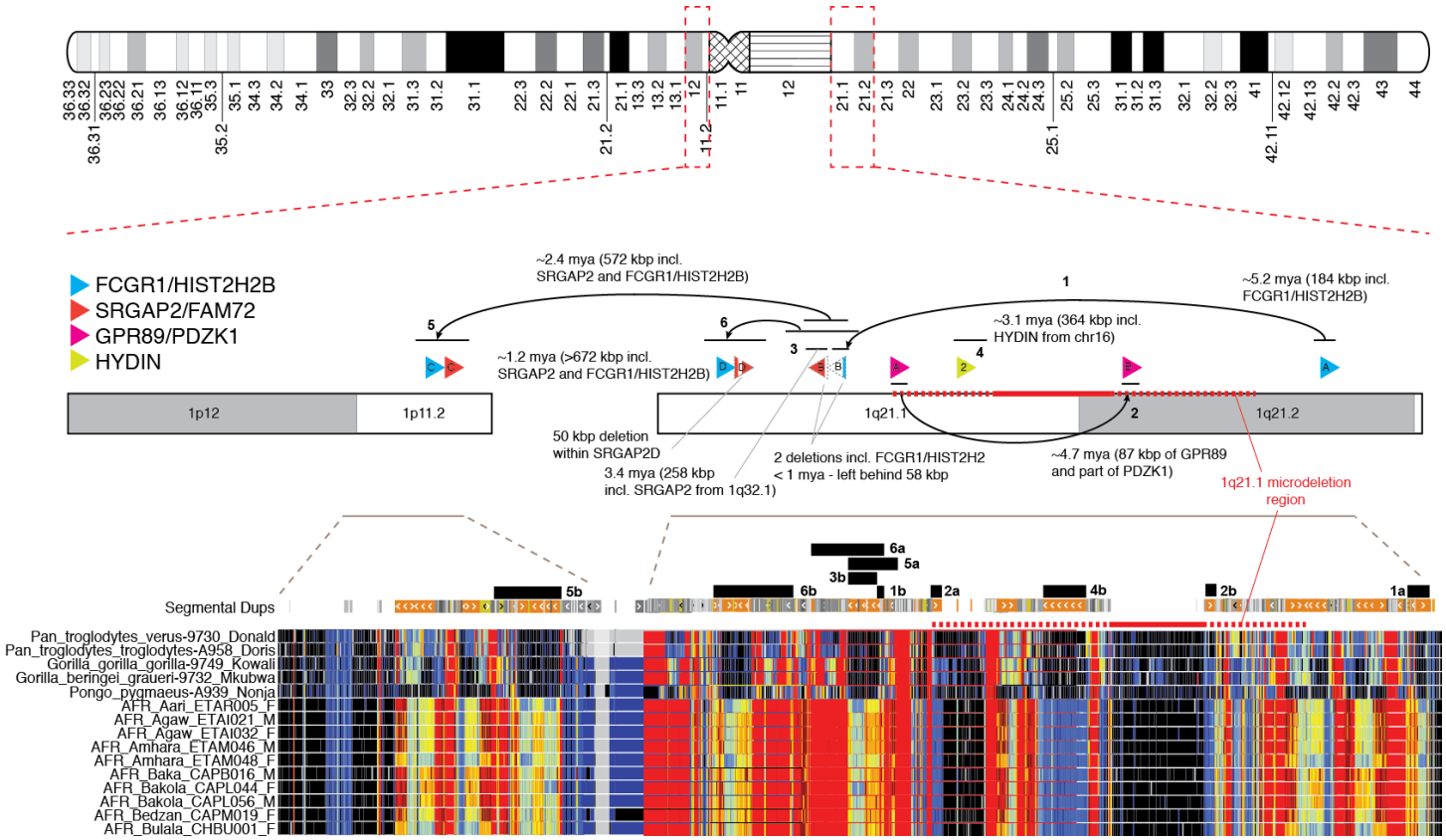
Supplementary Figure 4. Pairwise sequence comparison of the HSD paralogs of *CACNA1B* intron 3 taken from the CHM1 human haploid single-molecule, real-time (SMRT) sequenced assembly, visualized using Miropeats. Homologous segments are shown as black bars and connected by lines. Simple repeats identified by RepeatMasker are depicted as colored arrows.



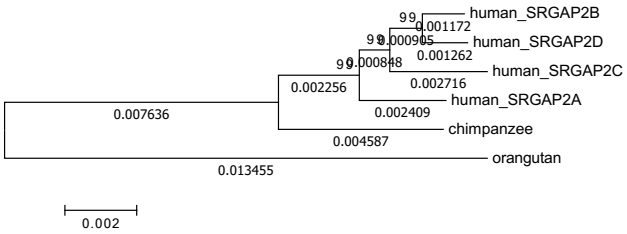
Supplementary Figure 5. Flowchart of HSD filtering. We started with a set of 218 HSD loci identified using our read-depth great ape comparative approach. Notably, these represent duplicated *regions* because the complete extent of duplications was still uncertain from this analysis. We filtered these regions to a higher confidence set (i.e., large size of >20 kbp, exon containing, and no overlapping duplication in a NHP), honing in on a set of 37 HSD loci, many of which clustered together. When we accounted for this clustering, we collapsed our loci to 16 larger regions and selected a tiling path of bacterial artificial chromosomes (BACs) to characterize sequences at these loci. We subsequently characterized all 28 HSDs residing in these regions (including some duplications that may have been originally filtered based on a size threshold). To avoid any inherent genomic positional bias resulting from our targeting these 16 regions, we selected only the 18 primary duplications containing previously identified HSD genes as an unbiased set of duplications to perform clustering simulations. The HSDs included in the clustering simulations are indicated in Supplementary Table 5.

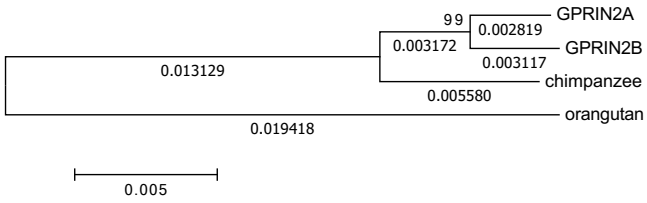


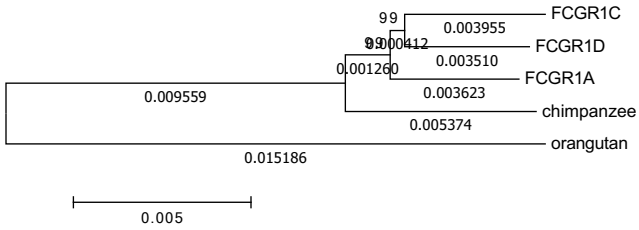
Supplementary Figure 6. HSDs significantly enriched near core duplicons than expected by chance. (A) Pictured is an ideogram of 18 tested primary duplications (blue triangles) and core duplicons (red triangles) mapped to the human reference (GRCh38). **(B)** A null distribution of median midpoint distance between 18 primary duplications and core duplicons is shown as a histogram using intrachromosomal and genome-wide shuffling, respectively (10 million times). The empirical median distance of 249,679 bp was significantly shorter than both null distributions ($P < 1x10^{-7}$).

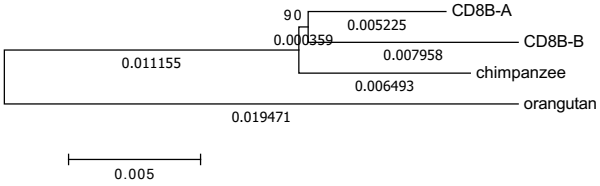


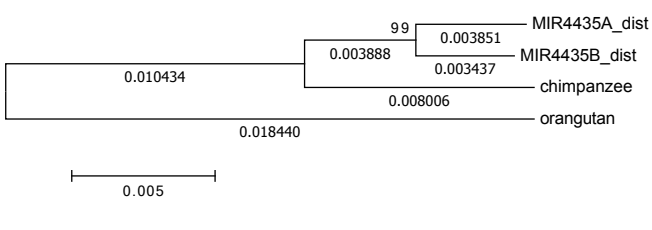
Supplementary Figure 7. Proposed mechanism and evolution of SDs on human chromosome 1. The schematic depicts the duplication landscape across chromosomal regions 1p12-11.2 (pictured chr1:118991014-125780799, GRCh38) and 1q21.1-21.2 (pictured chr1:143145441-149935226, GRCh38), highlighted with red dashed boxes on the chromosome 1 ideogram. This hotspot of duplication has been the target of at least six independent HSD events, including four primary duplications and two secondary duplications (shown as colored arrowheads indicating the respective direction). Black lines show the extent of duplications with ancestral and derived regions connected by curved arrows and numbers above conveying the evolutionary order of events. The red line spanning from 1q21.1 to 1q21.2 shows the extent of the 1q21.1 disease-associated microdeletion region, with dashed lines showing uncertainty in the breakpoints. At the bottom is a UCSC Genome Browser snapshot (GRCh38) with black boxes showing the extent of duplicated regions with numbers corresponding to evolutionary ordering (Supplementary Table 5; a = ancestral and b = derived), annotated SDs, and CN heatmaps from Illumina sequence mapping of chimpanzee, gorilla, and orangutan, and 10 African individuals (see Figure 1 for color coding index). We note timing estimates shown here are based on an evolutionary human–chimpanzee divergence time of 6 mya, though recent estimates suggest as much as 12 mya.

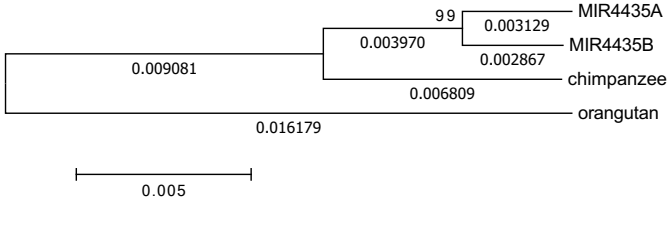


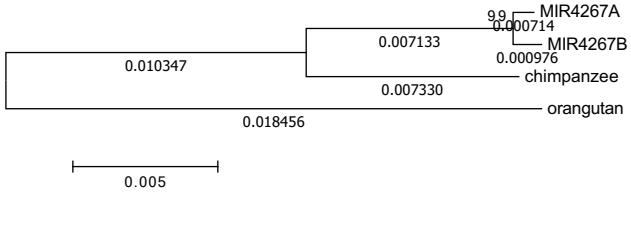


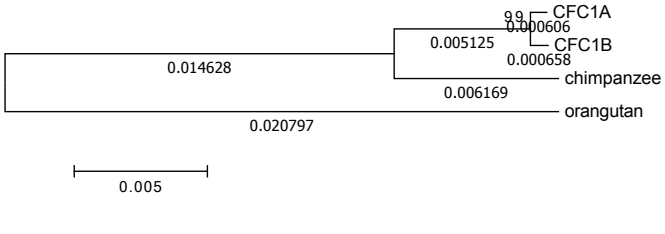


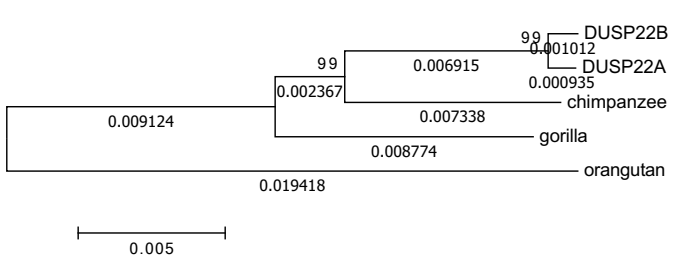


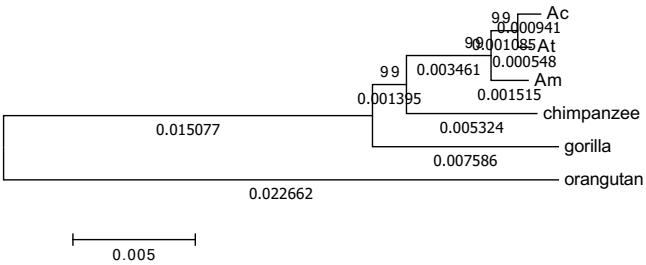


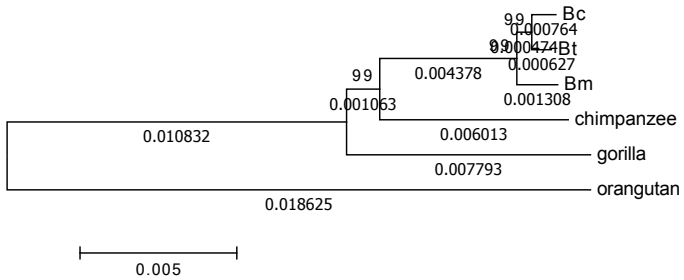


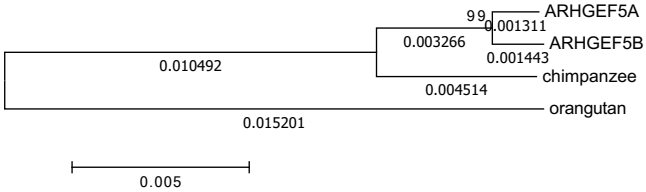


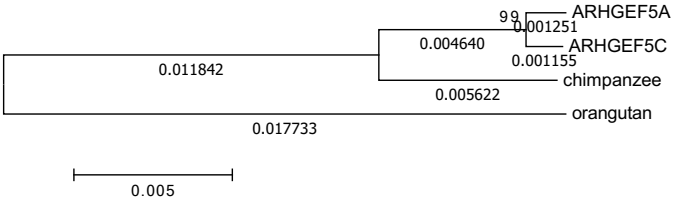


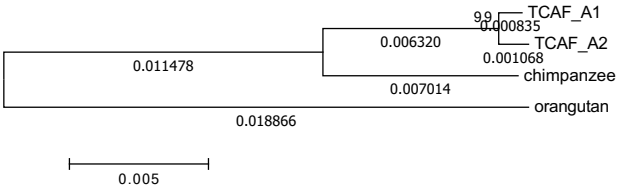


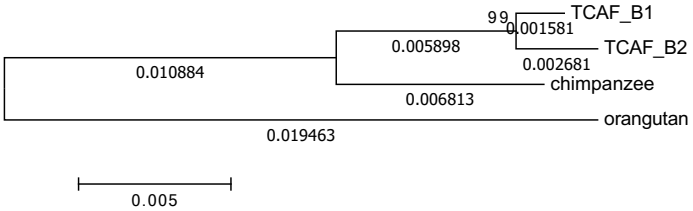


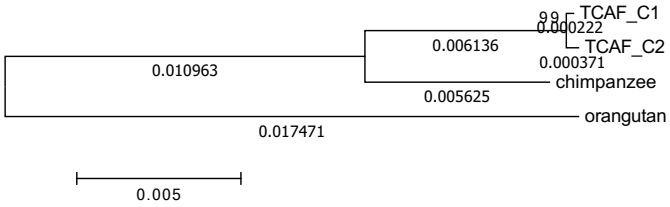


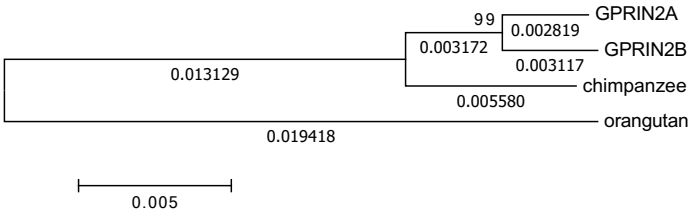


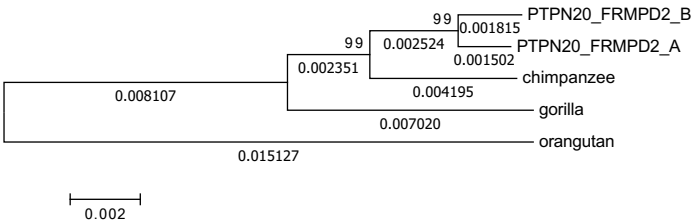


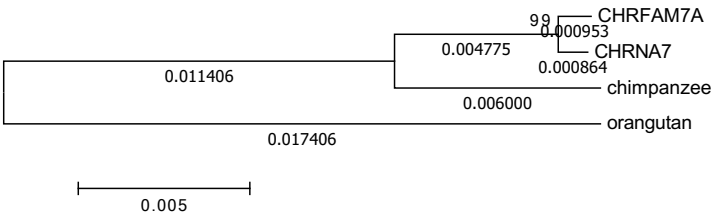


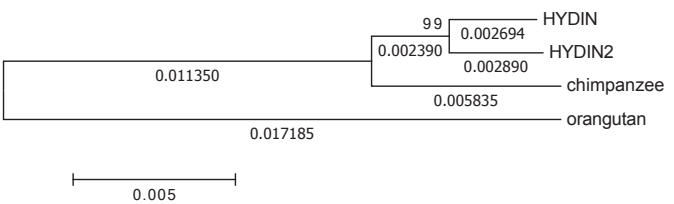


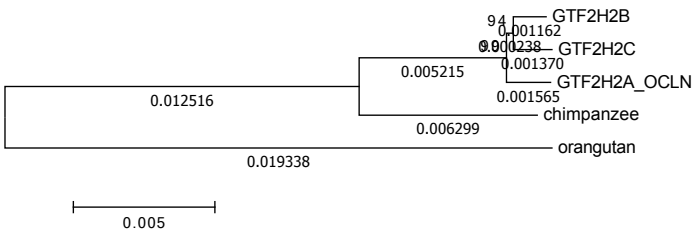


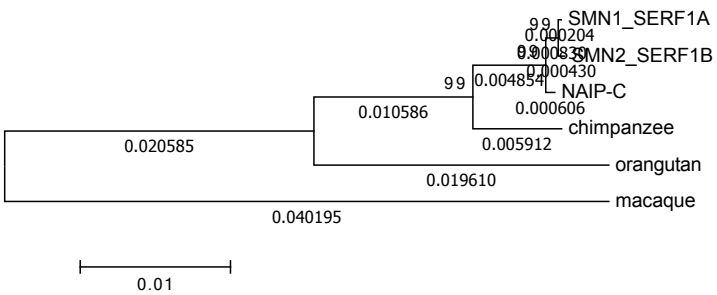




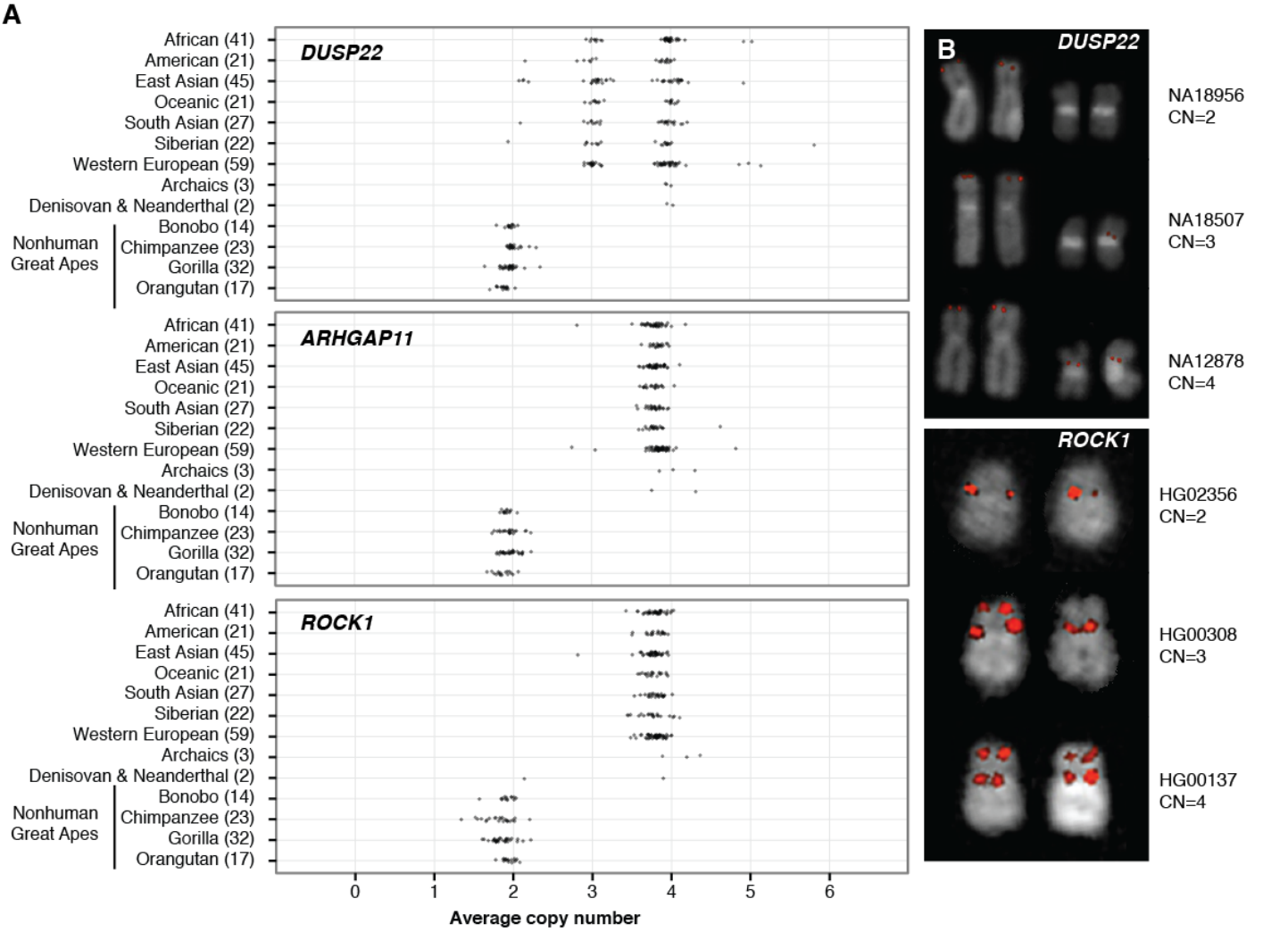




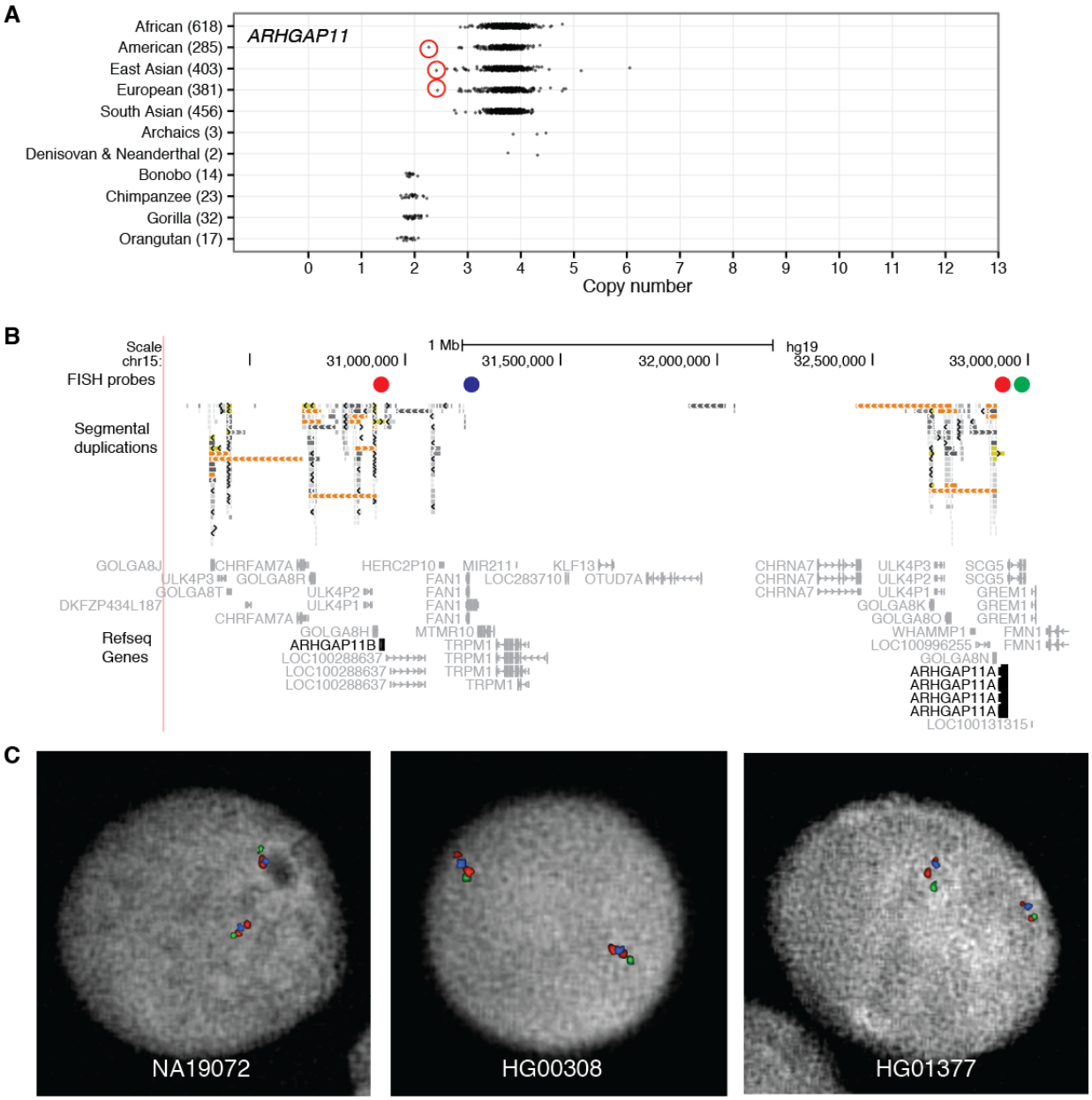




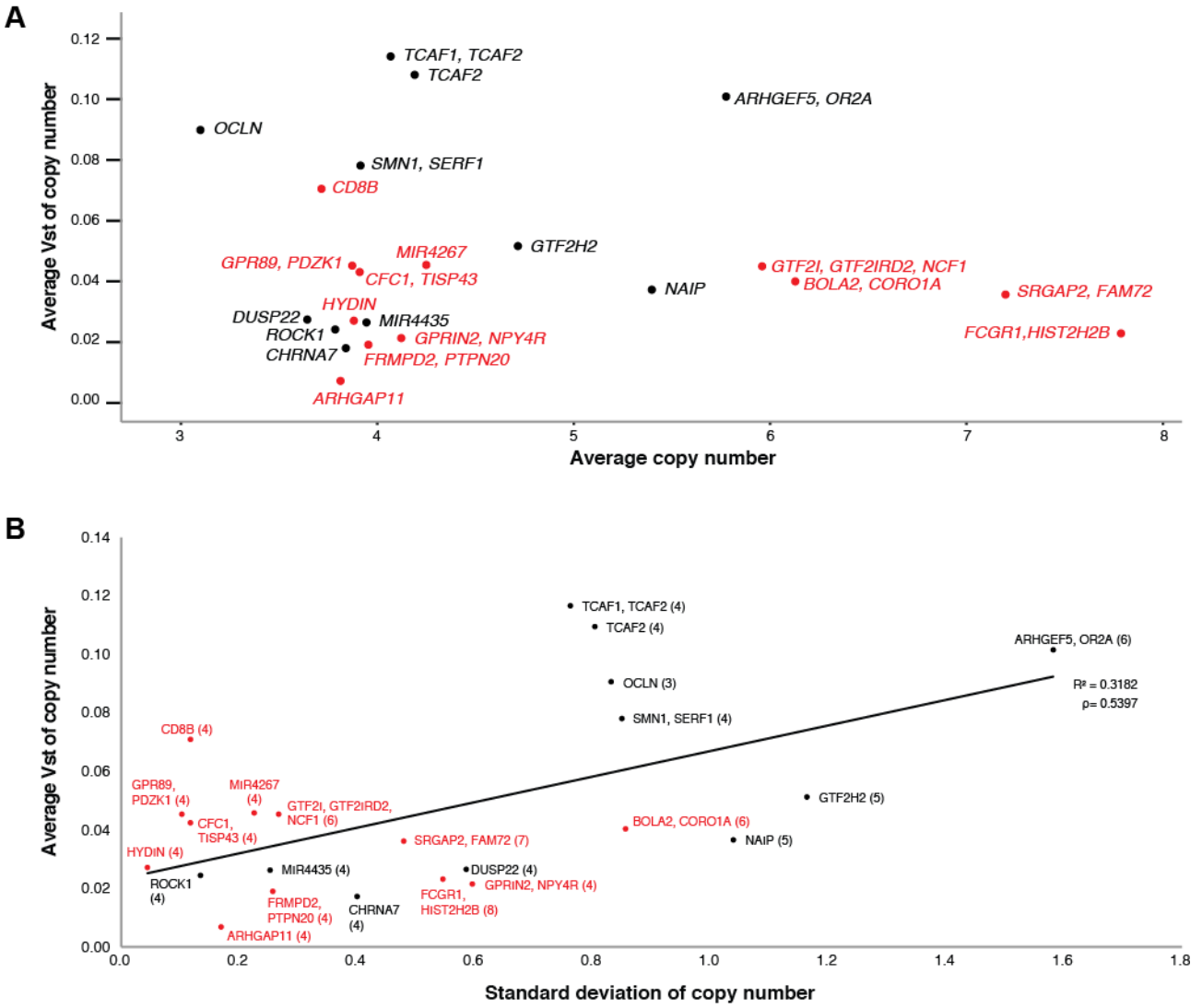
Supplementary Figure 8. Phylogenetic trees of HSD homologs. Muliple-species sequence alignments were generated using the human reference genome (GRCh38) or CH17 BAC contigs for human paralogs and orthologs from NHPs, including chimpanzee (panTro2 or CH251 BACs), gorilla (CH277 BACs), and orangutan (ponAbe2 or CH276 BACs). Trees were generated using MEGA6 [2] using the neighbor-joining method [3]. See the confidence probability (multiplied by 100) that the interior branch length is greater than 0, as estimated using the bootstrap test (500 replicates shown next to the branches [4,5]). The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Kimura 2-parameter method [6] and are in the units of the number of base substitutions per site. All positions containing gaps and missing data were eliminated. The alignments used to generate pictured trees were inputs into Tajima's relative rate test [7]. Subsequent timing estimates were performed using pairwise comparisons of human and chimpanzee orthologs using the complete deletion option (with all outgroups removed from the alignment). Alignment lengths and timing estimates are shown in Supplementary Table 8.



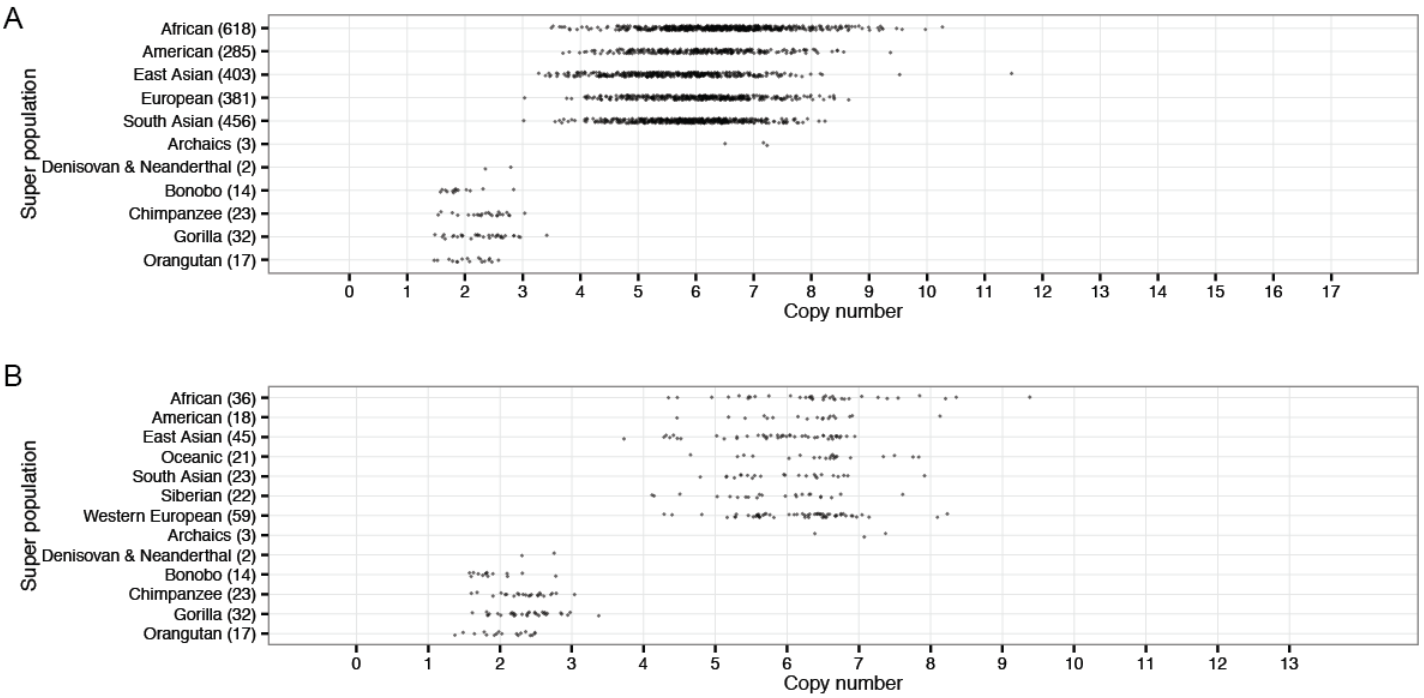
Supplementary Figure 9. Human copy number diversity. Overall average CN was calculated per individual from read depth produced from Illumina mappings across a set region defining each duplication (Supplementary Table 9) in human populations, including the HGDP (N = 236; GRCh38) and 1000 Genomes Project (1KG; N = 2,143; GRCh37) cohorts, NHPs, archaic humans, a Denisovan and a Neanderthal. From these results, the mean, standard deviation, V_{st}, and number of individuals with CN = 2 indicating no duplicate paralogs exist were calculated for average CN of each duplicated gene family (Supplementary Table 11). **(A)** Overall CN of individuals from human populations (HGDP), archaic hominins, and NHPs are shown for three examples, with total number of individuals depicted next to each population: *DUSP22*, a highly polymorphic gene with 7 individuals from HGDP (pictured) and 28 individuals from 1KG cohorts (not pictured) with homozygous deletions of *DUSP22B*; *ARHGAP11*, a "fixed" gene with some individuals showing reduced overall copy number (CN = 3) but no homozygous deletions of *ARHGAP11B* across any human individual tested; and *ROCK1*, a gene that appeared "fixed" within the HGDP cohort but a single individual was identified as homozygously deleted for *ROCK1B* in the 1KG cohort (not pictured in plot but validated by fluorescent *in situ* hybridization (FISH) as shown in B). **(B)** FISH validations were performed in representative individuals from 1KG for different diploid CN states for *DUSP22* (shown), *ARHGAP11* (not shown), and *ROCK1* (shown). The ancestral *DUSP22A* resides on chromosome 6 and is fixed across all individuals while the human-specific duplicated *DUSP22B* resides on chromosome 16 and varies in CN. Three individuals from the 1KG cohort were genotyped as homozygously deleted for *ARHGAP11B*, but FISH analyses showed CN = 4 for all three individuals (Supplementary Figure 10). The ancestral *ROCK1A* resides on human chromosome 18q11 while the human-specific *ROCK1B* shows polymorphism at 18p11.32, with one individual homozygously deleted for the paralog.



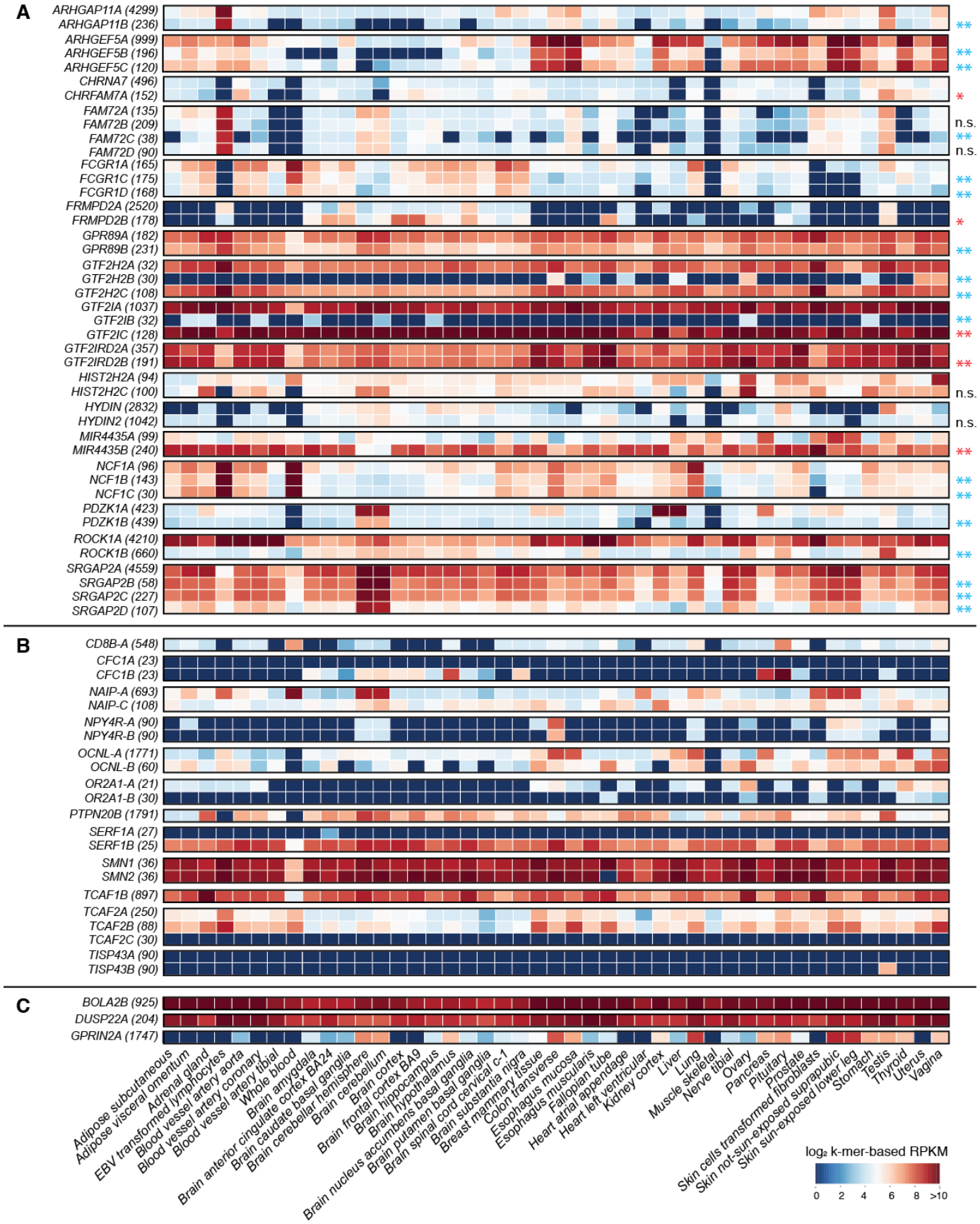
Supplementary Figure 10. FISH validation of *ARHGAP11* overall copy number. (A) CN estimates from the 1KG populations, archaic humans, Denisovans and Neanderthals, and NHPs (with total number per population indicated in parentheses). Three human individuals who were predicted to be homozygously deleted for the human-specific *ARHGAP11B* (overall CN = 2) were subjected to FISH validation. **(B)** UCSC Genome Browser snapshot of the 15q13.3 region, where *ARHGAP11A* (distal) and *ARHGAP11B* (proximal) are located, with RefSeq genes and SDs annotated. Location of the FISH probes and their corresponding fluorescence coloring are shown as colored circles. **(C)** Corresponding interphase FISH results in lymphoblastoid cell lines from the three individuals highlighted in **A**. The red signals indicate overall *ARHGAP11* CN, which is four in all three individuals, despite their genotyped predictions. The red signal adjacent to the blue signal indicates human-specific *ARHGAP11B* and the red signal adjacent to the green signal indicates ancestral *ARHGAP11A*. This same procedure was performed for individuals with suggestive rare homozygous HSD deletions for *GPRIN2/NPYR4* (not pictured), *MIR4267* (not pictured) and *ROCK1* (Figure 3B). See the Supplementary Table 11 "Note" column for results of those experiments.



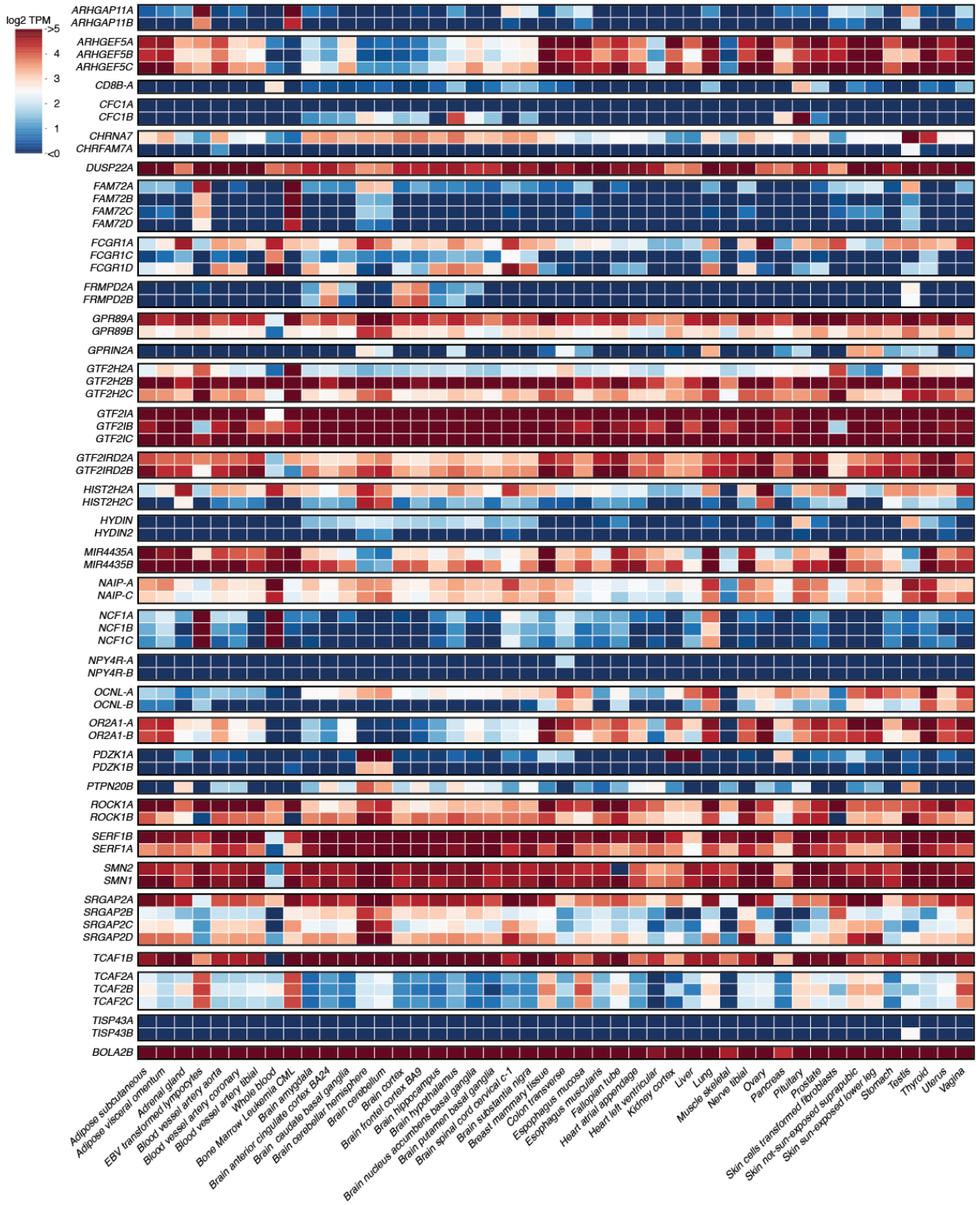
Supplementary Figure 11. Correlation of average gene family CN versus standard deviation (s.d.) and versus average V_{st}. For each gene family, plots are shown for CN statistics of HGDP individuals versus **(A)** average CN versus average V_{st} and **(B)** s.d. versus average CN polymorphic diversity statistic V_{st}. Duplicate gene family names and average CN in parentheses indicated next to each data point (for B). Red data points indicate genes with no homozygous deletions in any human tested. Genes with higher s.d. are considered CN polymorphic and tend to have higher CN and average V_{st} ($R^2 = 0.32$; $\rho = 0.54$, Pearson correlation, for B). Notably, no correlation exists between V_{st} and average CN, and thus is not pictured ($R^2 = 0.01$; $\rho = -0.01$, Pearson correlation, for A). Raw data for these plots can be found in Supplementary Table 11.



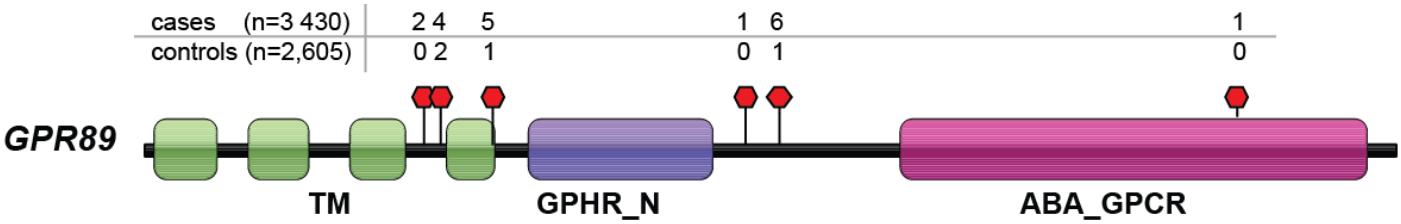
Supplementary Figure 12. **Overall copy number of *BOLA2*.** Pictured are overall CN predictions of *BOLA2* in human populations from the **(A)** 1KG and **(B)** HGDP, archaic humans, Denisovan and Neanderthal, and NHPs. Notably, this CN of the Denisovan and Neanderthal are diploid CN 2, the same as NHPs, while all archaic and modern day humans show diploid CN of 3 or greater.



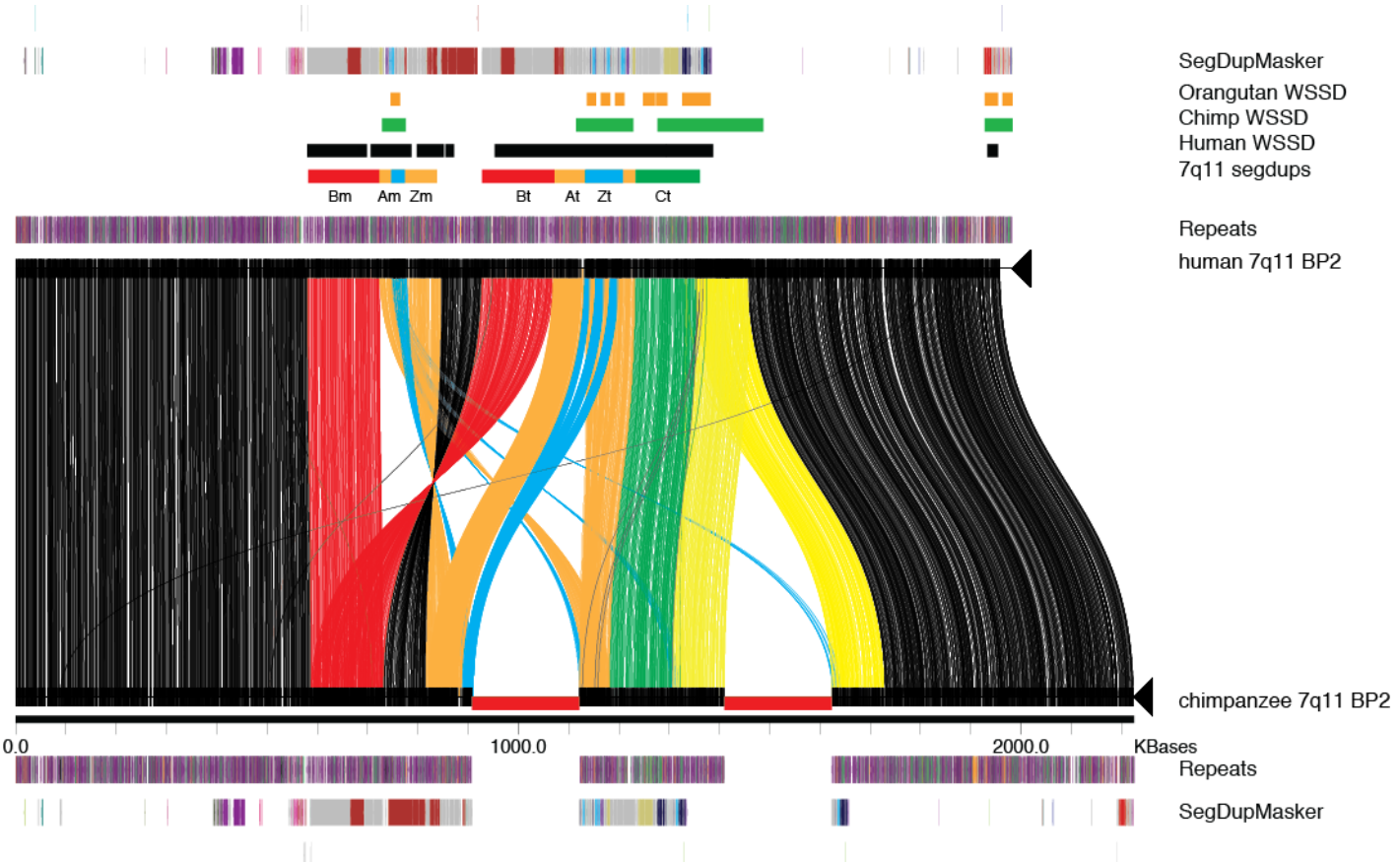
Supplementary Figure 13. Gene expression analysis of duplicated genes using a k-mer approach. Expression analysis using the GTEx RNA-seq dataset [8]. The heatmap shows a $\log_2$ scale of the median normalized read counts per subtissue, based on unique 30-mers that differentiate between the paralogs (Methods; Supplementary Table 16). Color-scale minimum and maximum values were set to 0.0 and 10.0, respectively. **(A)** Expression of paralogs within gene families with known ancestral and duplicate paralogs and annotated RefSeq transcripts were compared across all tissue types using a Wilcoxon signed-rank test of the median RPKM (reads per kilobase of transcript per million mapped reads) values of subtissues (n.s.: not significant; *: $p < 0.05$; **: $p < 0.001$ Bonferroni-corrected for 26 tests with color indicating lower (blue) or higher (red) expression of duplicate paralog compared to the ancestral by taking the mean ratio of median expression of each tissue). **(B)** For a subset of gene families (N = 12)—where the ancestral paralog was uncertain or an annotated RefSeq transcript existed for only one paralog—expression of paralogs was determined but no comparisons were made. **(C)** For a smaller subset of gene families (N = 3)—where accurate paralog-specific expression could not be determined due to missing paralogs in the reference genome (*DUSP22* and *GPRIN2*) or insufficient number of 30-mers to distinguish between paralogs (*BOLA2*)—total expression was calculated.



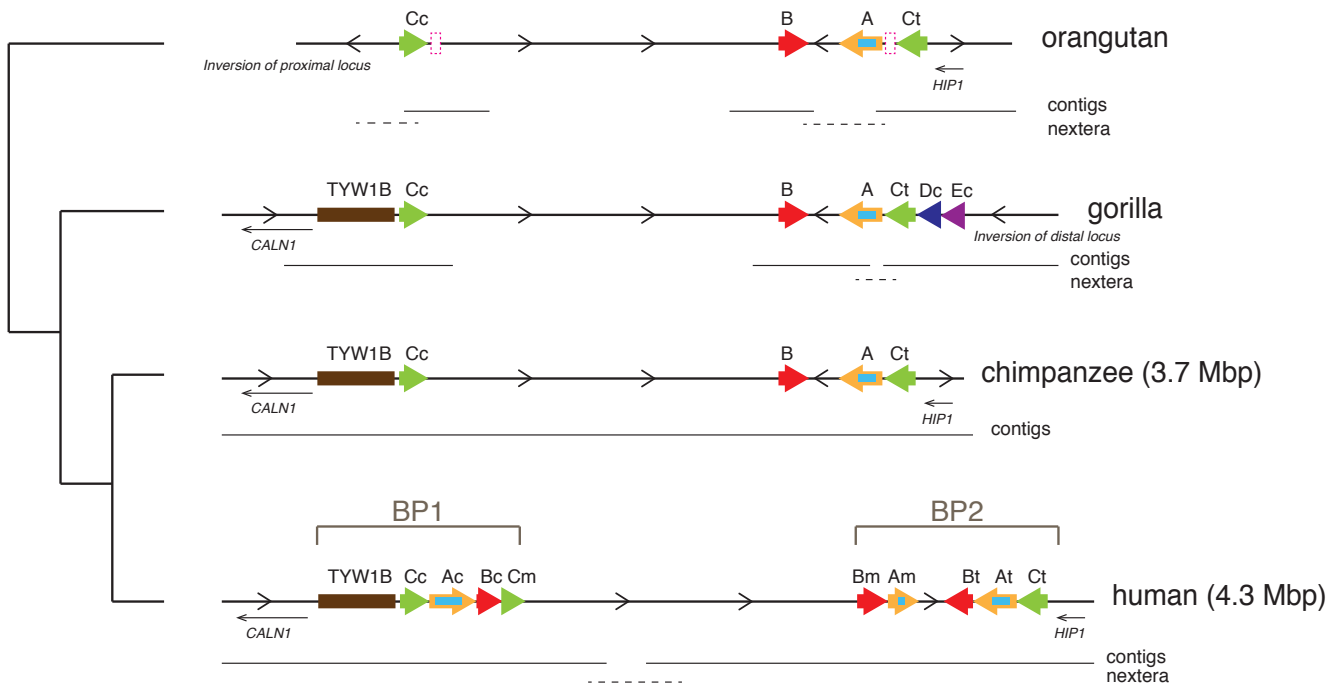
Supplementary Figure 14. Gene expression analysis of duplicated genes using Sailfish method. Expression analysis of the GTEx RNA-seq dataset [8] used the Sailfish method [9]. For each gene, the estimated transcripts per million (TPM) values were summed over all its transcripts. The heatmap shows a $\log_2$ scale of the median TPM per subtissue. Color-scale minimum and maximum values were set to 0.0 and 5.0, respectively.



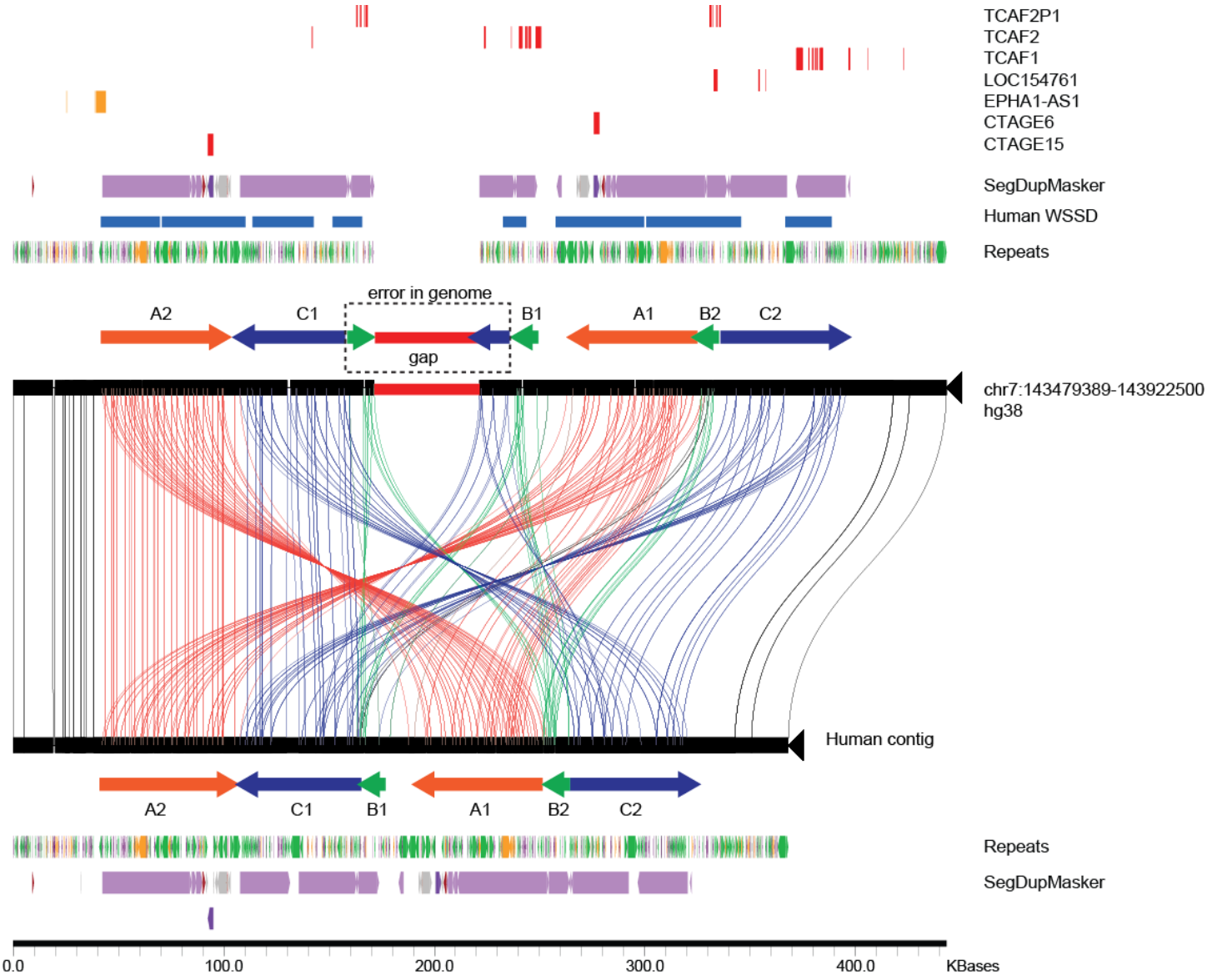
Supplementary Figure 15. Likely gene-disrupting (LGD) mutations identified in *GPR89* from targeted molecular inversion probe (MIP) sequencing of children with autism spectrum disorder (cases) versus unaffected siblings (controls). Red hexagons represent locations of identified nonsense and frameshift mutations with the total number of cases and controls with variants shown above each site. Colored boxes represent predicted functional domains (using SMART (http://smart.embl-heidelberg.de/)) with labels beneath. Schematic created using IBS [10]. Details of mutations are described in Supplementary Table 20.



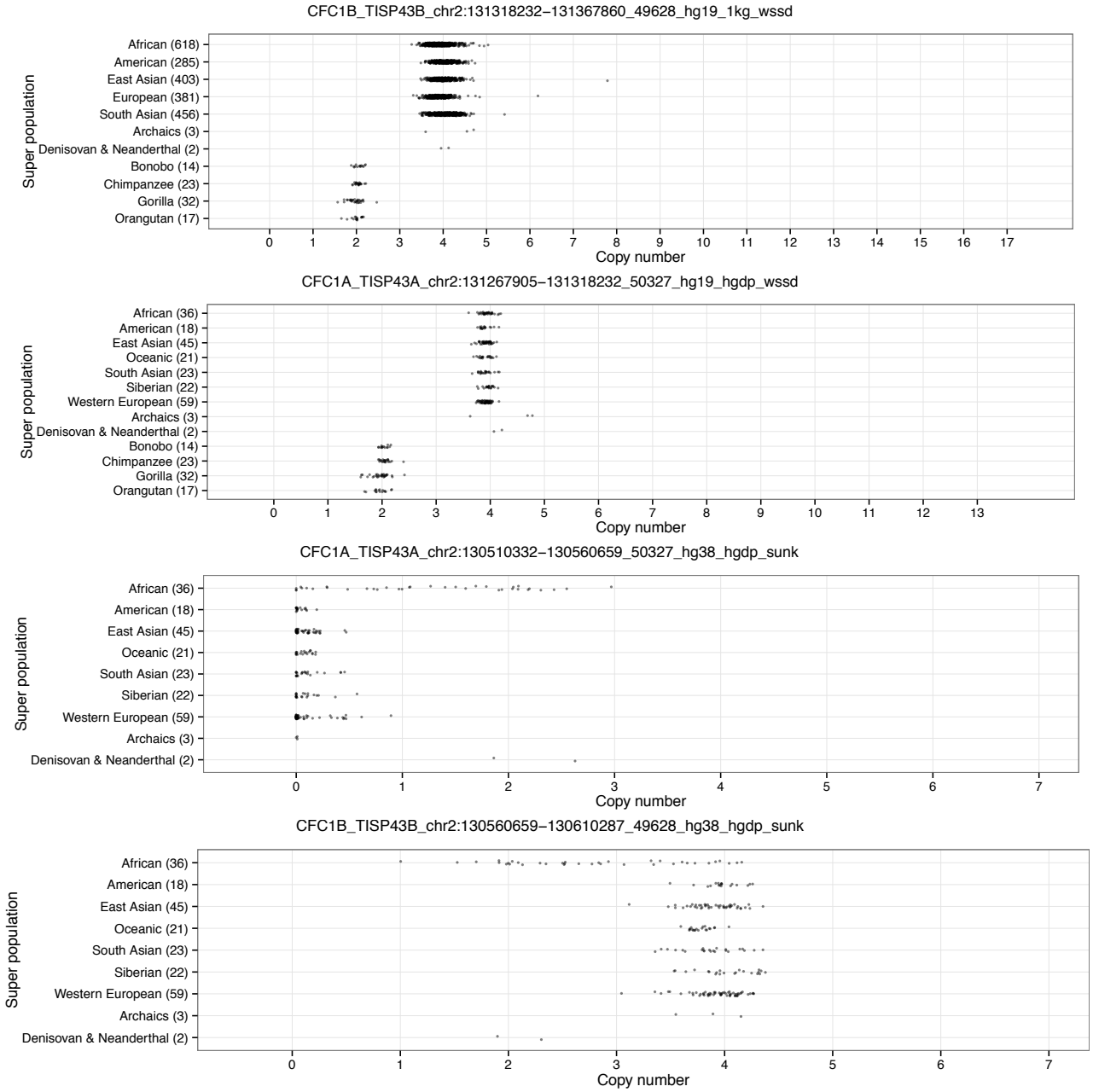
Supplementary Figure 16. Pairwise comparison of genomic organization of human and chimpanzee at the 7q11.23 distal breakpoint using Miropeats. A Miropeats comparison of the human and chimpanzee contigs shows the pairwise differences between the orthologous regions. Lines connect stretches of homologous regions (threshold s = 1000) and match the arrow colors when they connect SD blocks (defined in Figure 6A). Additional annotations include: whole-genome shotgun sequence detection (WSSD) in human and chimpanzee, indicating duplicated regions identified by sequence read depth [11]; DupMasker [12]; and exons of genes. Gaps in the chimpanzee contig are shown as red boxes.



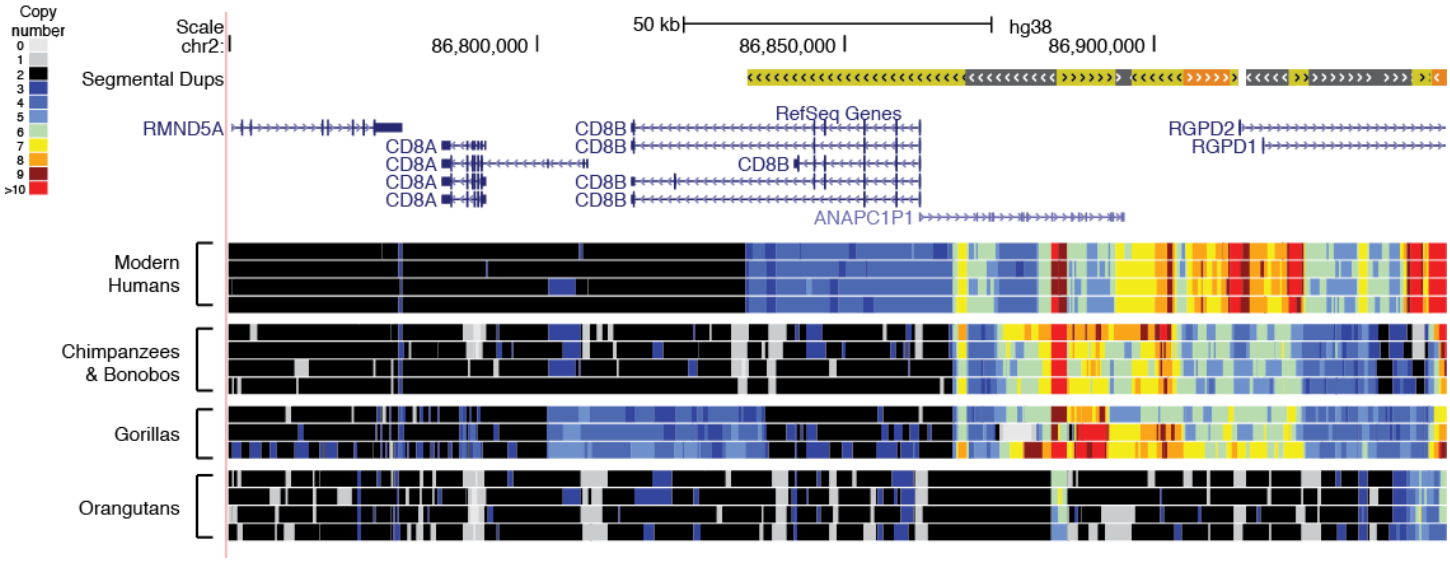
Supplementary Figure 17. SD configuration of chromosome 7q11.23 in great apes. Tiling paths of BACs were selected and sequenced using Illumina Nextera, capillary, and SMRT methods. Finished BAC sequences were compared and assembled into contigs (shown as gray lines beneath each haplotype). In some cases, BACs could not be assembled due to complex structural configurations contained within the clone, and Illumina Nextera sequencing was used to deduce the structural configuration and connections of larger contigs (shown as dashed gray line). SDs were identified within contigs from each species, as previously defined [13], and shown as colored arrows: A is orange, B is red, and C is green. Additional SDs from a distal region are also depicted as blue (D) and purple (E) arrows in gorilla, as they were transplanted to BP2 via large-scale inversion.



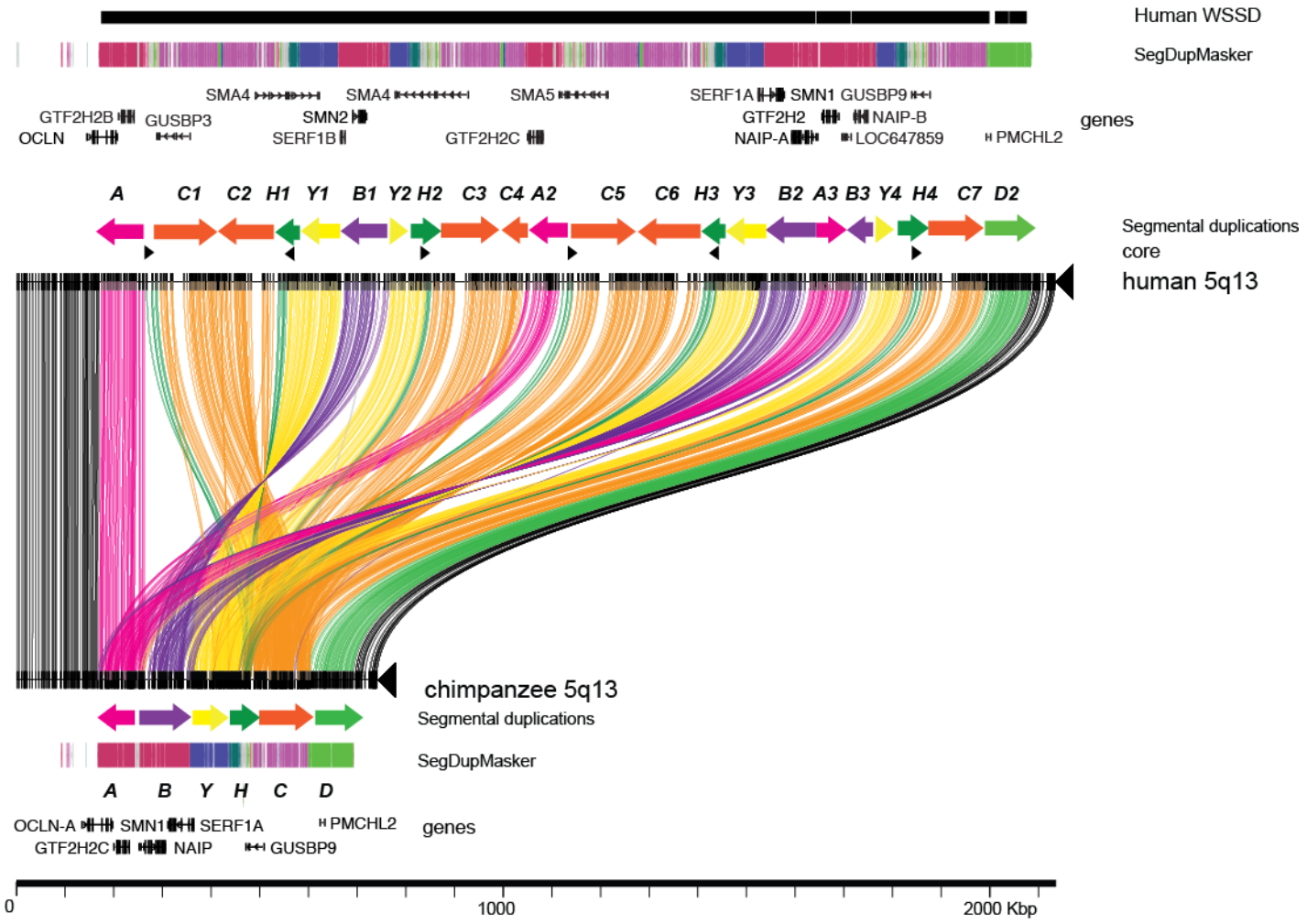
Supplementary Figure 18. Pairwise comparison of human reference (GRCh38) and newly constructed contig using CH17 BAC clones. A Miropeats comparison of the human and chimpanzee contigs shows the pairwise differences between the orthologous regions. Lines connect stretches of homologous regions (threshold s = 1000) and match the arrow colors when they connect SD blocks (defined in Figure 3C). Additional annotations include: WSSD in human and chimpanzee, indicating duplicated regions identified by sequence read depth [11]; DupMasker [12]; and exons of genes. In the newly created contig, errors in the human reference genome were corrected, including removal of a gap (pictured as a red box) and ~29 kbp of extra sequence (chr7:143336323-143347897 and chr7:143397898-143415538; hg38).



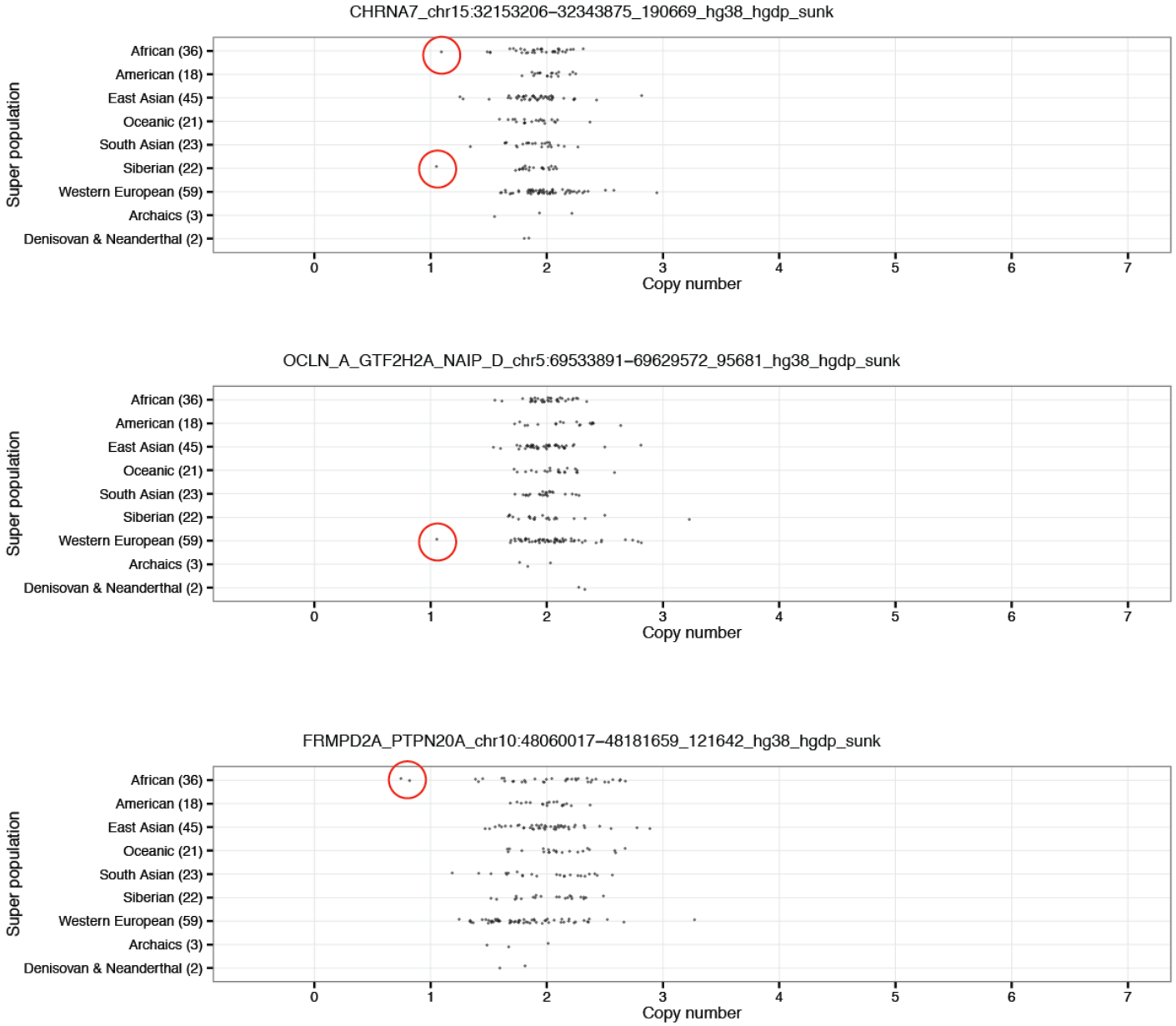
Supplementary Figure 19. An example of HSD gene family (*CFC1*) showing evidence of interlocus gene conversion from paralog-specific duplication analysis. We leveraged singly unique nucleotide k-mers (SUNKs) calculated per individual from read depth produced from Illumina mappings across a set region defining each duplication in human populations, including the HGDP (N = 236; GRCh38), NHPs, archaic humans, a Denisovan and a Neanderthal. SUNK analysis of *CFC1*/*TISP43*, which we originally determined was "fixed" by our aggregate diploid copy number analysis, revealed that nearly all non-African individuals carry only the B duplicate paralog and almost no copies of the A paralog. Africans show a more diverse distribution with the B locus ranging from 1 to 4 copies while the A locus ranges from 0 to 3. In contrast, Neanderthal and Denisova both carry two copies of each paralog indicating that gene conversion has likely occurred within modern humans making the two paralogs nearly identical among the out-of-Africa populations (Supplementary Tables 22 and 23).



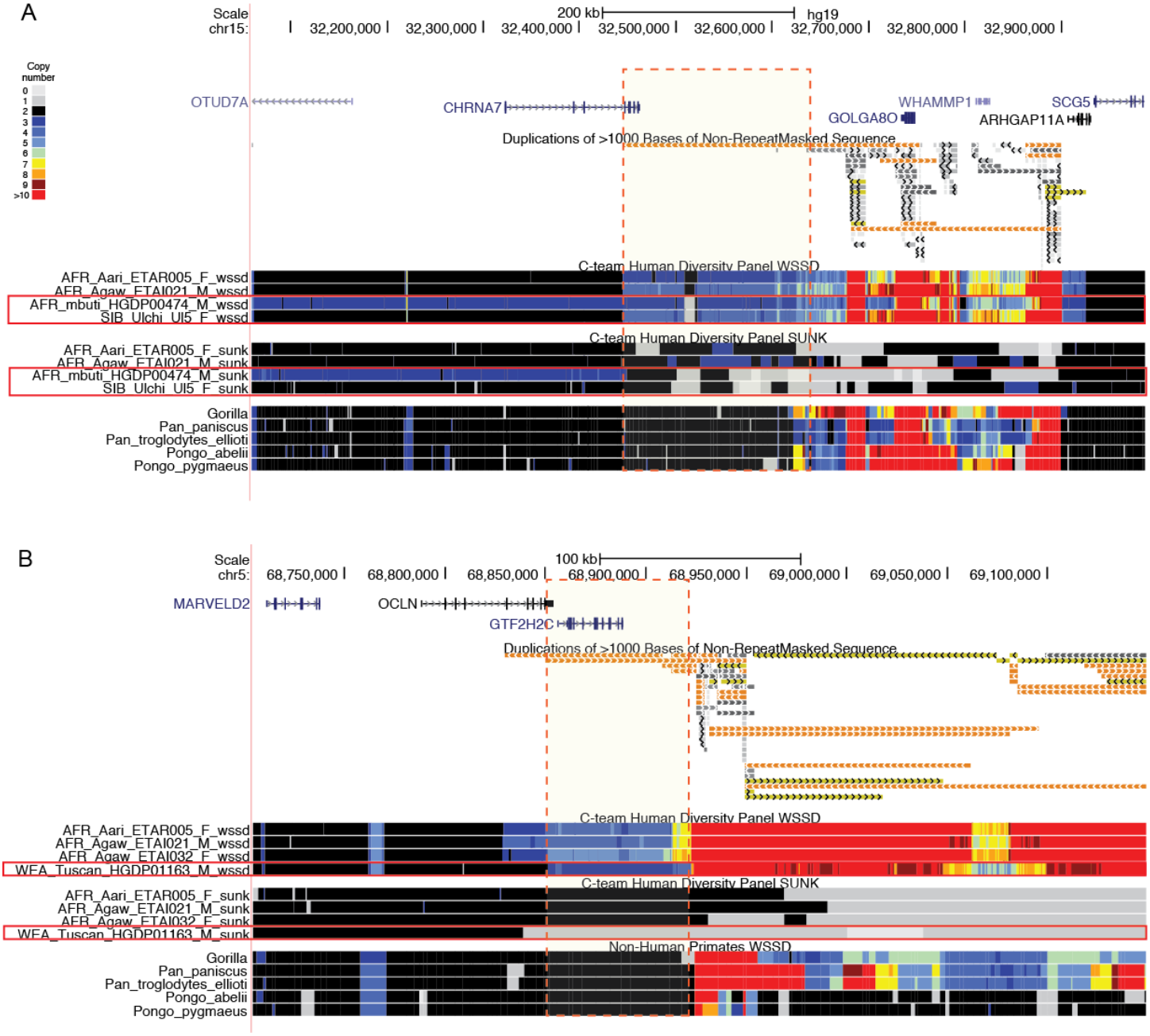
Supplementary Figure 20. Heatmap of HSD *CD8B* showing copy number across different primate species. A UCSC Genome Browser snapshot with tracks shown from top to bottom: SDs, gene models from RefSeq, and a heatmap of overall diploid CN in representative individuals from modern humans, chimpanzees, bonobos, gorillas, and orangutans. A ~40 kbp SD exists in gorillas of the 5' end of *CD8B* that is not shared with human, chimpanzee or bonobo, as suggested by a previous study [14]. The HSD region encompasses the 3' end of *CD8B*, with only a very small amount of overlap with the gorilla-specific duplication indicating this was a separate event unique to humans. The timing estimate for the *CD8B-B* paralog supports this finding (Figure 2).



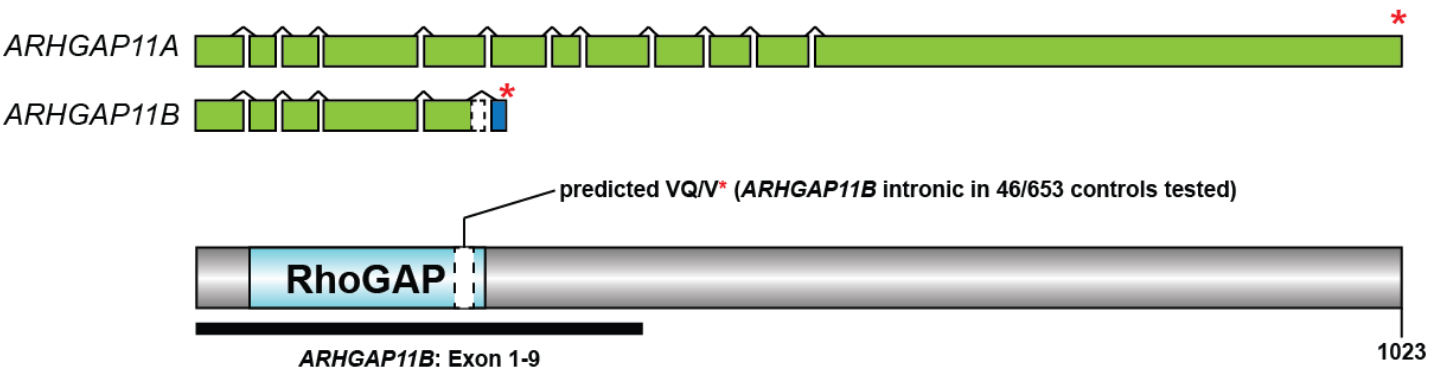
Supplementary Figure 21. Pairwise alignment of human and chimpanzee 5q13.3 orthologous loci. A Miropeats comparison of the human (reference build GRCh38) and chimpanzee contigs shows the pairwise differences between the orthologous regions. Lines connect stretches of homologous regions (threshold s = 1000) and match the arrow colors when they connect SD blocks.



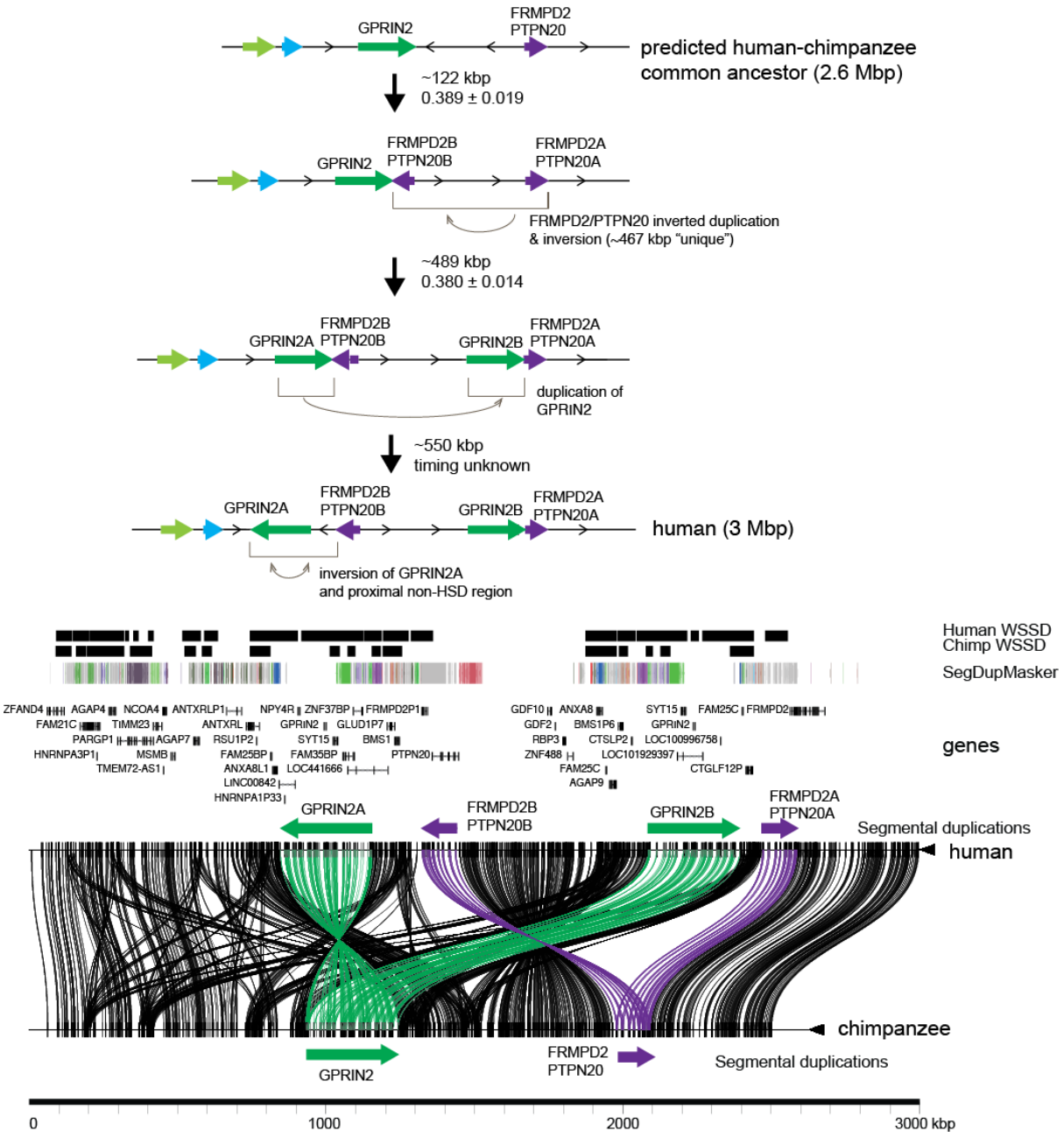
Supplementary Figure 22. Ancestral paralogs showing evidence of deletion from SUNK copy number estimates. SUNK-based paralog-specific CN was calculated per individual from read depth produced from Illumina mappings across a set region defining each duplication in human populations, including the HGDP (N = 236; GRCh38), NHPs, archaic humans, a Denisovan and a Neanderthal. The individuals predicted to have deletions are circled in red. Heatmaps for individuals with predicted heterozygous deletions of *CHRNA7* and *OCLN* are pictured in Supplementary Figure 23.



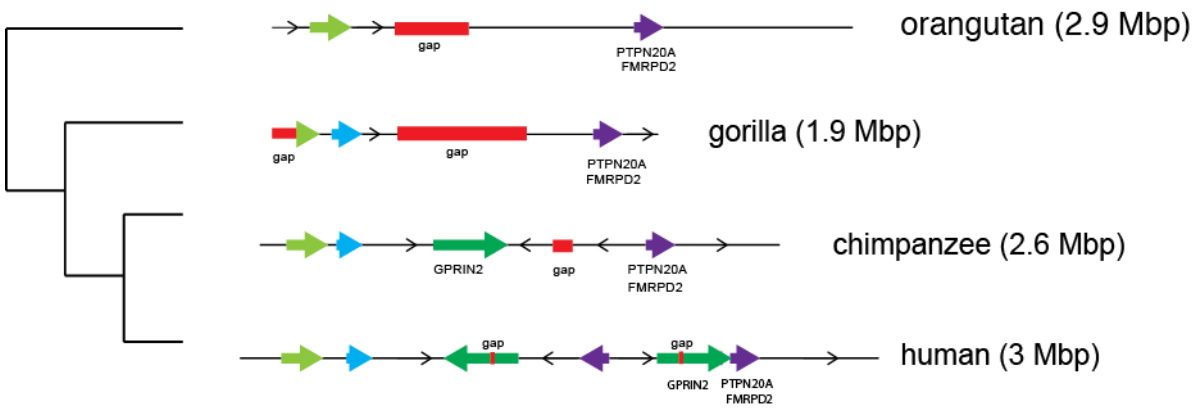
Supplementary Figure 23. Copy number heatmaps of individuals predicted to harbor heterozygous deletions of *CHRNA7* and *OCLN-A*. Paralog-specific genotypes identified individuals predicted to harbor heterozygous deletions of ancestral paralogs (shown in Supplementary Figure 22; individuals highlighted in red). From top to bottom, pictured are RefSeq genes, SD tracks (WGAC), and CN heatmaps for overall CN (WSSD) and paralog-specific CN (SUNK) from the HGDP cohort, and overall CN of NHPs.



Supplementary Figure 24. Schematic of *ARHGAP11* common variant identified from 1KG MIP analysis. Splicing differences between ancestral *ARHGAP11A* (encoding 12 coding exons) and HSD *ARHGAP11B* (encoding 6 coding exons including a shortened exon 5 and novel amino acids due to a frame shift in exon 6) are depicted on the top. Molecular phasing places the common variant within the *ARHGAP11B*-specific intronic region (Supplementary Table 19), which would normally impact the RhoGAP domain of the full-length ARHGAP11A protein.



Supplementary Figure 25. Complex models of chromosome 10q11.23 HSD evolutionary history. BACs tiling across human chromosome 10q11.23 region were sequenced and assembled (representing human and additional great apes) and supercontigs were created. Estimates of sizes and evolutionary timing (human–chimpanzee distance; Supplementary Table 8) of events are denoted between each predicted intermediate genomic structure. SD organization is depicted as colored arrows and the orientations of intervening regions are shown with arrows. Models of the predicted evolutionary histories of the HSDs at all loci are depicted starting with the predicted human–chimpanzee common ancestor to the most common haplotype present in modern humans. A Miropeats comparison of the human and chimpanzee contigs shows the pairwise differences between the orthologous regions. Lines connect stretches of homologous regions based on a chosen threshold (s), defined as the number of matching bases minus the number of mismatching bases (s = 500) and match the arrow colors when they connect SD blocks. Additional annotations include WSSD in human and chimpanzee, indicating duplicated regions identified by sequence read depth [11], DupMasker [12], and genes.



Supplementary Figure 26. SD configuration of chromosome 10q11 in great apes. Tiling paths of BACs were selected and sequenced using Illumina Nextera, capillary, and SMRT methods. Finished BAC sequences were compared and assembled into contigs. SDs were identified within contigs from each species, as previously defined [13], and shown as colored arrows: *PTPN20/FRMPD2* as purple and *GPRIN2/NPY4R* as dark green. Additional SDs from a distal region are also depicted as light blue and light green. Gaps in the contigs are shown as red boxes.

SUPPLEMENTARY TABLES

Supplementary Table 1. Complete list of HSD regions identified by comparisons of human and NHP genome sequencing. See accompanying Supplementary Tables Excel file.

Supplementary Table 2. High-confidence HSD targeted regions and accompanying genes. Information summarized in Figure 1. See accompanying Supplementary Tables Excel file.

Supplementary Table 3. Sequenced human CH17 BAC clones. See accompanying Supplementary Tables Excel file.

Supplementary Table 4. CH17 BAC contigs not included in human reference (GRCh38). See accompanying Supplementary Tables Excel file.

Supplementary Table 5. High-confidence HSDs and genomic locations of breakpoints. See accompanying Supplementary Tables Excel file.

Supplementary Table 6. Sequenced NHP BAC clones. See accompanying Supplementary Tables Excel file.

Supplementary Table 7. Genomic regions of interlocus gene conversion identified by GENCONV.

Region	Alignment pairwise	coordStart	coordEnd	GI global sim p value	gene conv tract length	comment
TCAF_B						No conversion detected
TCAF_A						No conversion detected
TCAF_C	C1; C2	26007	38360	0.0155	12354	
TCAF_C	C1; C2	7761	17211	0.0259	9451	
ROCK1						No conversion detected
HYDIN	HYDIN;HYDIN2	121063	122118	0.0129	1056	
CHRNA7						No conversion detected
ARHGAP11						No conversion detected
FRMPD2_PTPN20	A;B	102242	109700	0.0000	7459	
GPRIN	A;B	169913	172028	0.0005	2116	
ARHGEF5AC						No conversion detected
ARHGEF5AB						No conversion detected
CFC1	A;B	41170	46046	0.0004	4877	
CD8A_B						No conversion detected
FCGR1						No conversion detected
MIR4267	A;B	324915	329234	0.0211	4320	
MIR4435_dist						No conversion detected
DUSP22						No conversion detected
GPR89						No conversion detected
7q11_A						No conversion detected
7q11_B	Bc;Bt	6351	9052	0.0138	2702	
7q11_B	Bm;Bt	14660	17621	0.0138	2962	
7q11_B	Bc;Bt	50312	52948	0.0172	2637	
7q11_B	Bc;Bm	12455	14396	0.0417	1942	
SRGAP2	SRGAP2A;SRGAP2C	102627	104263	0.022	1637	added RheMac to the MSA
MIR4435	A;B	191289	194049	0.0000	2761	
MIR4435	A;B	239430	241631	0.0264	2202	
SMN_NAIP	NAIP-C:SMN2_SERF1B	13323	15689	0.0013	2367	added RheMac to the MSA
SMN_NAIP	SMN1_SERF1A:SMN2_SERF1B	10838	17331	0.0072	6494	added RheMac to the MSA
SMN_NAIP	NAIP-C:SMN1_SERF1A	13323	15689	0.0171	2367	added RheMac to the MSA
GTF2H2	B;C	25117	30180	0.0000	5064	

Supplementary Table 8. Evolutionary characterization of HSDs. Information summarized in Figure 2. See accompanying Supplementary Tables Excel file.

Supplementary Table 9. Human genome reference coordinates used for HSD CN estimates. See accompanying Supplementary Tables Excel file.

Supplementary Table 10. Summary of overall CN and exon mutational profiles of HSD genes. See accompanying Supplementary Tables Excel file.

Supplementary Table 11. Overall CN estimates of HSDs in modern-day humans. Information summarized in Figure 3 and Supplementary Figure 11. See accompanying Supplementary Tables Excel file.

Supplementary Table 12. Fosmids used for FISH experiments.

Region	Fosmid	chr (GRCh37)	start	end	Location	Fluorescence
DUSP22	WIBR2-2930l21_G248P89200E11	chr6	243883	284894	overlapping_DUSP22	cy3
ARHGAP11A	167098_ABC8_000040990300_G1	chr15	32901632	32938073	overlapping_ARHGAP11A	cy3
ARHGAP11A	167098_ABC8_000040990300_G1	chr15	30912172	30935159	overlapping_ARHGAP11B	cy3
ARHGAP11A	G248P85460C11	chr15	32938920	32978653	flanking_ARHGAP11A	fx
ARHGAP11A	G248P89213G9	chr15	31193865	31231670	flanking_ARHGAP11B_dist	cy5
ROCK1	1200277_ABC11_2_1_000048323300_G11	chr18	18521220	18561062		cy3
GPRIN_PPYR	G248P83429G12_WIBR2-2629N23	chr10	46985704	47022808	alternative_overlapping_GPRIN2A	cy3
GPRIN_PPYR	G248P800478D10	probe2 in 10q11 contig				fx
GPRIN_PPYR	G248P86891H2	probe3 in 10q11 contig				cy5
MIR4267	G248P88615G4	chr2	1.11E+08	1.11E+08	overlapping_MIR4267A	cy3
MIR4267	G248P81805G1	chr2	1.11E+08	1.11E+08	between_MIR4267A_and_MIR4267B	fx
MIR4267	G248P801670H4	chr2	1.11E+08	1.12E+08	dist_MIR4267B	cy5

Supplementary Table 13. Details and results of of FISH validation experiments.

Region	Individuals	Need interphase?	Genotyped overall copy number	FISH results
ARHGAP11A	HG01377	Yes	CN=2 homozygous deletion	CN=4; all paralogs present
ARHGAP11A	NA19072	Yes	CN=2 homozygous deletion	CN=4; all paralogs present
ARHGAP11A	HG00308	Yes	CN=2 homozygous deletion	CN=4; all paralogs present
ARHGAP11A	NA19434	Yes	CN=3	CN=4; all paralogs present
ARHGAP11A	HG00137	Yes	CN=4	CN=4; all paralogs present
DUSP22	NA18956	No	CN=2	CN=4; all paralogs present
DUSP22	NA18507	No	CN=3	CN=3; CN=2 at DUSP22A on chr6 and CN=1 at DUSP22B on chr16
DUSP22	NA12878	No	CN=4	CN=4; CN=2 at DUSP22A on chr6 and CN=2 AT DUSP22B on chr16
ROCK1	HG02356	No	CN=2 homozygous deletion	CN=2; CN=2 at ROCK1A on 18q-cen and CN=0 at ROCK1B on 18p-tel
ROCK1	HG00308	No	CN=3	CN=3; CN=2 at ROCK1A on 18q-cen and CN=1 at ROCK1B on 18p-tel
ROCK1	HG00137	No	CN=4	CN=4; all paralogs present
GPRIN2_PPYR1	HG00566	Yes	CN=2 homozygous deletion	confusing results, the cell line might be rearranged
GPRIN2_PPYR1	HG02485	Yes	CN=2 homozygous deletion	CN=2; CN=1 at GPRIN2A and CN=1 at GPRIN2B
MIR4267	NA18558	Yes	CN=2 homozygous deletion	CN=2; CN=1 at MIR4267A and CN=1 at MIR4267B - one deletion encompasses both plus intervening unique region

Supplementary Table 14. Overall CN of TCAF1/2 duplication in humans (1KG cohort), archaic humans, and NHPs. Data used in pie charts included in Figure 4. See accompanying Supplementary Tables Excel file.

Supplementary Table 15. GTEX tissues included in RNA-seq analysis. Data shown in Supplementary Figures 13 and 14).

#samples	Tissue	subtissue
107	Adipose	Adipose-Subcutaneous
19	Adipose	Adipose-Visceral (Omentum)
12	Adrenal Gland	Adrenal Gland
39	Blood	Cells-EBV-transformed lymphocytes
24	Blood Vessel	Artery-Aorta
10	Blood Vessel	Artery-Coronary
118	Blood Vessel	Artery-Tibial
177	Blood	Whole Blood
33	Bone Marrow	Cells-Leukemia cell line (CML)
26	Brain	Brain-Amygdala
22	Brain	Brain-Anterior-cingulate-cortex-(BA24)
38	Brain	Brain-Caudate-(basal-ganglia)
26	Brain	Brain-Cerebellar-Hemisphere
38	Brain	Brain-Cerebellum
31	Brain	Brain-Cortex
27	Brain	Brain-Frontal-Cortex-(BA9)
29	Brain	Brain-Hippocampus
25	Brain	Brain-Hypothalamus
31	Brain	Brain-Nucleus-accumbens-(basal-ganglia)
22	Brain	Brain-Putamen-(basal-ganglia)
17	Brain	Brain-Spinal-cord-(cervical-c-1)
26	Brain	Brain-Substantia-nigra
27	Breast	Breast-Mammary-Tissue
12	Colon	Colon-Transverse
21	Esophagus	Esophagus-Mucosa
22	Esophagus	Esophagus-Muscularis
1	Fallopian	Tube-Fallopian-Tube
26	Heart	Heart-Atrial-Appendage
97	Heart	Heart-Left-Ventricle
3	Kidney	Kidney-Cortex
6	Liver	Liver
133	Lung	Lung
146	Muscle	Muscle-Skeletal
98	Nerve	Nerve-Tibial
6	Ovary	Ovary
20	Pancreas	Pancreas
14	Pituitary	Pituitary
10	Prostate	Prostate-
14	Skin	Cells-Transformed-fibroblasts
23	Skin	Skin-Not-Sun-Exposed-(Suprapubic)
109	Skin	Skin-Sun-Exposed-(Lower-leg)
13	Stomach	Stomach
15	Testis	Testis
113	Thyroid	Thyroid
7	Uterus	Uterus
6	Vagina	Vagina

Supplementary Table 16. GTEX RNA-seq kmer analysis. Data shown in Supplementary Figure 13. See accompanying Supplementary Tables Excel file.

Supplementary Table 17. HSD MIP summary.

Gene	Region	Exons	amino acids	cds	Haploid CN	Number MIPs
ARHGAP11A	15q13.3	12	1023	3072	2	41
ARHGEF5	7q35	15	1597	4794	3	57
BOLA2	16p11.2	3	152	459	3	7
CD8B	2p11.2/2q13	6	210	633	2	13
CFC1	2q21.1	6	223	672	2	11
CHRNA7	15q13.3	10	502	1509	2	27
CORO1A	16p11.2	11	461	1386	3	27
DUSP22	16p12.1/6p	8	205	618	2	17
FAM72A	chr1	4	149	450	4	6
FCGR1	chr1	6	374	1125	3	19
FRMPD2	10q11.23	29	1309	3930	2	64
GPR89	1q21	14	455	1368	3	26
GPRIN2	10q11.23	1	458	1377	2	16
GTF2H2C	5q13.2	16	395	1188	2	26
GTF2I	7q11.23	35	998	2997	3	66
GTF2IRD2	7q11.23	16	949	2850	3	47
HIST2H2BF	chr1	2	134	405	3	7
HYDIN	16q22/1q21.1	86	5121	15366	2	240
NAIP	5q13.2	17	1403	4212	3	55
NCF1	7q11.23	11	390	1173	3	24
NPY4R	10q11.23	3	375	1128	2	12
OCLN	5q13.2	8	522	1569	2	21
PTPN20B	10q11.23	10	339	1020	2	25
ROCK1	chr18	33	1354	4065	2	65
SERF1	5q13.2	3	110	333	2	7
SMN1	5q13.2	9	294	885	2	18
SRGAP2	chr1	22	1071	3216	4	82
TCAF1	7q35	9	921	2766	2	33
TCAF2	7q35	7	815	2448	2	40
TISP43	2q21.1	3	160	483	2	6

Supplementary Table 18. HSD MIP design and sequence coverage. See accompanying Supplementary Tables Excel file.

Supplementary Table 19. HSD MIP LGD variants identified from the 1KG cohort. See accompanying Supplementary Tables Excel file.

Supplementary Table 20. HSD MIP exonic variants identified from autism cohorts. See accompanying Supplementary Tables Excel file.

Supplementary Table 21. LGD-variant frequencies of HSD genes in autism cases and controls. See accompanying Supplementary Tables Excel file.

Supplementary Table 22. SUNK-based CN estimates of HSDs using GRCh37 human reference. See accompanying Supplementary Tables Excel file.

Supplementary Table 23. SUNK-based CN estimates of HSDs using GRCh38 human reference.

					HGDP (N=236)			
chr	start	end	Paralog name	Number of SUNKs	Mean CN	sd CN	Mean Vst	Notes
chr15	32607507	32632914	ARHGAP11A	147	1.99	0.13	0.02	
chr15	30618104	30642956	ARHGAP11B	141	2.15	0.20	0.02	
chr7	144296183	144377283	ARHGEF5A/OR2A1	150	2.38	0.74	0.06	
chr7	144186950	144254793	ARHGEF5B/OR2A2	122	2.68	0.69	0.06	
chr7	144272352	144296188	ARHGEF5C	44	2.50	0.81	0.12	
chr16	30188533	30198679	BOLA2A/CORO1A/SLX1A	8	3.38	0.88	0.10	Not enough SUNKs to genotype
chr16	29449195	29459337	BOLA2B/CORO1B/SLX1B	7	2.16	0.90	0.19	Not enough SUNKs to genotype
chr2	86834206	86869480	CD8B-A	226	1.95	0.11	0.03	
chr2	106479411	106515037	CD8B-B	224	1.88	0.12	0.09	
chr2	130510332	130560659	CFC1A/TISP43A	60	0.29	0.59	0.14	
chr2	130560659	130610287	CFC1B/TISP43B	59	3.72	0.58	0.17	
chr15	30194794	30377808	CHRFAM7A	327	1.95	0.36	0.03	
chr15	32153206	32343875	CHRNA7	331	1.94	0.23	0.04	
chr6	256519	382461	DUSP22	83	3.67	0.58	0.03	Incorrect in GRCh38
chr7	143836044	143874696	TCAF1B/TCAF2C	54	3.28	0.65	0.07	Incorrect in GRCh38
chr7	143587138	143621615	TCAF1A/TCAF2A	54	1.00	0.44	0.10	Incorrect in GRCh38
chr7	143700805	143728107	TCAF2B	36	1.62	0.43	0.08	Incorrect in GRCh38
chr1	149681544	149817485	FCGR1A/HIST2H2A	432	2.03	0.18	0.07	
chr1	145129136	145188250	FCGR1B/HIST2H2B	115	2.10	0.25	0.07	
chr1	120989760	121133389	FCGR1C/HIST2H2C	550	1.93	0.13	0.04	
chr1	143772885	143909760	FCGR1D/HIST2H2D	398	1.68	0.54	0.04	
chr10	48060017	48181659	FRMPD2A/PTPN20A	S	1.97	0.36	0.02	
chr10	46870207	46991846	FRMPD2B/PTPN20B	353	2.18	0.26	0.04	
chr1	145600141	145686997	GPR89A/PDZK1A	389	2.07	0.11	0.02	
chr1	147920587	148009863	GPR89B/PDZK1B	401	2.09	0.10	0.02	
chr10	46396229	46585038	GPRIN2A/NPYR1A	423	3.55	0.52	0.03	
chr10	47756785	47988907	GPRIN2B/NPYR1B	374	1.96	0.36	0.04	
chr5	69554620	69612650	GTF2H2A	98	1.97	0.30	0.04	
chr5	70409766	70467792	GTF2H2B	61	1.20	0.56	0.10	
chr5	71015225	71073279	GTF2H2C	85	1.02	0.55	0.06	
chr7	74728174	74869950	GTF2IRD2A/GTF2IA/NCF1A	291	1.45	0.33	0.13	
chr7	75074151	75217878	GTF2IRD2B/GTF2IB/NCF1B	246	2.40	0.25	0.08	
chr7	73174796	73280573	GTF2IRD2C/GTF2IC/NCF1C	226	2.67	0.46	0.21	
chr16	70811377	71168670	HYDIN-A	1443	2.10	0.07	0.02	
chr1	146541436	146905930	HYDIN-B	1454	2.01	0.07	0.07	
chr2	109930243	110095177	MIR4267A	198	2.07	0.25	0.01	
chr2	110276211	110441265	MIR4267B	194	2.02	0.27	0.05	
chr2	111252926	111615296	MIR4435A	1474	2.00	0.07	0.10	
chr2	87431440	87697988	MIR4435B	1093	1.98	0.24	0.03	
chr5	69534817	69553693	OCLN-A	32	2.48	0.43	0.10	
chr5	69533891	69629572	OCLN-A/GTF2H2A/NAIP-D	148	2.03	0.25	0.03	
chr5	71015225	71093996	OCLN-B	119	1.02	0.57	0.07	
chr18	20940379	20959847	ROCK1A	97	2.33	0.20	0.06	
chr18	112547	131692	ROCK1B	91	2.59	0.29	0.11	
chr5	70025150	70129604	SMN2/SERF1B	87	2.62	0.67	0.03	
chr5	70900569	71005160	SMN1/SERF1A	72	1.51	0.41	0.12	
chr1	206151620	206409977	SRGAP2A/FAM72A	749	2.00	0.09	0.08	
chr1	144888777	145129140	SRGAP2B/FAM72B	404	2.02	0.25	0.02	
chr1	121151202	121391237	SRGAP2C/FAM72C	599	1.68	0.14	0.08	
chr1	143938830	144071736	SRGAP2D/FAM72D	164	1.50	0.52	0.02	

SUPPLEMENTARY NOTE

Proposed mechanism and evolution of SDs on human chromosome 1. Previous directed sequencing efforts using the haploid BAC library have completely corrected this locus in the most recent build (GRCh38) [15,16]. This hotspot of duplication has been the target of at least six independent HSD events, including four primary duplications and two secondary duplications (Supplementary Figure 7). The majority of events occurred within bands 1p11.2 to 1q21.2 and contain seven HSD gene families, many of which have been previously implicated with neurological diseases and phenotypes. *HYDIN*, a gene in which homozygous mutations lead to congenital hydrocephalus in mouse [17] and primary ciliary dyskinesia in humans [18], duplicated from chromosome 16q22.2 (357 kbp) to the 1q21.1 locus ~3.1 million years ago (mya) [19] creating partial paralog *HYDIN2* (Supplementary Table 8, assuming a human–chimpanzee split of 6 mya). An 87 kbp duplication occurred ~4.7 mya, including *GPR89* and a partial *PDZK1*, flank the 1q21.1 microdeletion locus in direct orientation, creating a potential substrate for non-allelic homologous recombination (NAHR). Our analysis also provided additional insight into the evolutionary history of *SRGAP2*, a gene previously shown to have duplicated in the human lineage [15] potentially contributing to innovative neurological phenotypes, including alterations in synaptogenesis [20]. We predicted two separate loci, a ~184 kbp region containing complete copies of *FCGR1* and *HIST2H2* (5.2 mya, originating from 1q21.2) and a 258 kbp region containing *FAM72* and *SRGAP2* (~3.4 mya, originating from chromosome 1q32 [15]), duplicated side-by-side on 1q21.1. Subsequently, larger inverted duplications containing all four of these genes established additional paralogs at 1p11.2 (~2.4 mya, *FCGR1C/HIST2H2C/FAM72C/SRGAP2C*, 572 kbp) and directly proximal at 1q21.1 creating a palindromic configuration (<1 mya, *FCGR1D/HIST2H2D/FAM72D/SRGAP2D*, >672 kbp). Notably, deletions upstream of *SRGAP2B* may have altered expression patterns of this gene paralog [15]. We note timing estimates are based on an evolutionary human–chimpanzee divergence time of 6 mya, though recent estimates suggest as much as 12 mya.

Human copy number diversity
A notable exception to the general finding that higher copy number gene families have greater variance are duplication blocks containing *GTF2I*, *GTF2IRD2*, and *NCF1* (diploid CN = 6; HGDP standard deviation (s.d.) = 0.27), located within the Williams-Beuren syndrome region on chromosome 7q11.23. This may be due to functional constraints of the genes.

Homo sapiens*-specific duplications of *BOLA2* and *TCAF1*/*TCAF2
Comparisons with the genomes of Neanderthal and Denisova allowed us to identify the most recent duplications to emerge specifically on the lineage of *Homo sapiens* since divergence from archaic hominins. This includes duplications of the TRP channel-associated factors, *TCAF1* and *TCAF2*, as well as the recently characterized *BOLA2* gene family [21,22], which shows some of the most significant expression differences between human and chimpanzee induced pluripotent stem cells [23] and whose SD mediates large-scale copy number variation associated with 1% of autism cases [24,25]. It is interesting that both gene families are involved with pathways related to "environmental sensing". For example, *BOLA2* regulates intracellular iron levels while *TCAF1/TCAF2* are associated with posttranslational regulation of the primary cold sensor, *TRPM8* [26]. Our sequence analysis suggests that both loci have undergone extensive expansion and restructuring creating not only additional copies but the potential for novel fusion genes and truncated copies to have emerged specifically *Homo sapiens* [21].

For the novel *TCAF*-duplications finding, we characterized copy number of the *TCAF* locus and show that among modern humans copy number is uniform over a ~131 kbp cassette (involving A, B, and C SDs) and ranges from 2 to 7 (Figure 4). The only exception to this pattern of copy number polymorphism was a Western European individual (HGDP00798) (A and B segments were discordant with C copy number), which we predict arose as a result of a NAHR-mediated deletion between directly oriented B1 and B2. Further, we identified 17 individuals (n = 2,367) where the copy number was consistent with the non-duplicated ancestral state observed in Denisova and Neanderthal. We remapped Illumina data of HGDP individuals to our newly constructed human contig of this locus in order to determine the paralogs experiencing polymorphism but were unable to perform SUNK analysis due to the extremely high similarity of the duplicate paralogs.

The complexity of HSD evolutionary history
To highlight the complex evolutionary history associated with HDS regions, we selected three loci (*TCAF1/2* locus at

chromosome 7q35 (main text); Williams-Beuren disease region at chromosome 7q11.23 (main text); and genomic hotspot-associated chromosome 10q11.23) for further investigation taking advantage of high-quality NHP BAC sequence generated as part of this study (N = 196; Supplementary Table 6) or taken from GenBank (N = 35). At the chromosome 10q11.21 locus, large-scale deletions and duplications have been identified in children with developmental delay with variable expressivity and penetrance [27,28]. HSD genes *FRMPD2*, *PTPN20*, *GPRIN2*, and *NPY4R* reside within two separate SDs proximal to the disease-associated region (Supplementary Figures 25 and 26). Our data predict an initial inversion of 589 kbp, which resulted in a duplication of a 122 kbp segment containing a partial paralog of *FRMPD2B* and a full-length paralog of *PTPN20B* (0.389 ± 0.019 human–chimpanzee distance). The inversion breakpoint maps within *PTPN20A* and truncates the likely ancestral version of this gene by removing the last two exons, leaving the human duplicate *PTPN20B* as the only functional paralog. A 489 kbp duplication containing full-length *GPRIN2* and *NPY4R* along with additional great ape-duplicated genes occurred potentially concurrently with the previous event (0.380 ± 0.014 human–chimpanzee distance). Comparing this region in human and chimpanzee identified an additional 550 kbp inversion that included the *GPRIN2*/*NPY4R* SD and adjoining proximal region.

Our BAC-based targeted sequencing of the 7q35 locus (containing *TCAF1* and *TCAF2*) not only eliminated the gap annotated in the reference sequence (GRCh38) but also dramatically reorganized the structure of the region removing incorrectly assigned paralogous sequence, including ~29 kbp of extra sequence (Supplementary Figure 18). Errors also existed within the sequence itself as evidenced by dramatically different sequence identities of paralogs between the reference and the corrected contig, which would have incorrectly estimated the timing of the final *TCAF1* and *TCAF2* duplications before the split with Denisova and Neanderthal.

SUPPLEMENTARY DISCUSSION

Additional important HSD features emerging from targeted sequencing of loci

We distinguish primary duplications (as those ancestrally derived from the unique ortholog in NHP species) from secondary duplications (those that have duplicated from an HSD). Secondary duplications (average 437 kbp) are twice as large when compared to primary duplications (average 136 to 196 kbp). The size discrepancy between older and younger duplications may be explained, in part, by subsequent internal deletions accruing over time within older duplications, as was the case for the *SRGAP2B* paralog [15]. Nevertheless, unlike primary duplications, secondary duplications are always distributed intrachromosomally in inverted orientation with the majority mapping within 5 Mbp of their progenitor.

We also found an enrichment of HSDs near primate ancestral "core duplicons." It is possible that the complex genomic architecture associated with core duplicons residing in proximity to HSDs acted to perturb the replisome leading to fork-stalling and template switching (FoSTeS) or represented sites of fragility predisposing to microhomology-mediated break-induced repair (MMBIR). The enrichment of inverted SDs also emphasizes the intimate association between inversions and the dispersal of SDs [29-31]. The association of inversion breakpoints with SDs has contributed to this type of variation being underestimated in studies of human genomes [32-34].

Additional information on putatively functional HSD genes

While the functional relevance of most transcripts mapping to HSDs remains to be determined, several recent studies have suggested that these regions may encode genes relevant to human neurocognitive and neuroanatomical adaptation. The human-specific duplicate *SRGAP2C*, for example, has been shown *in vivo* to alter dendrite formation and potentially spine density in developing neurons [15,20,35], while the HSD gene *ARHGAP11B* appears to promote apical basal radial progenitor amplification in the subventricular zone [36]. Microinjection of the *ARHGAP11B* RNA into the developing mouse brain increases the number of basal radial glial divisions leading to cortical expansion as well as gyrification.

There are also a handful of HSD-associated genes implicated in immune response, including *CD8B* (CD8 antigen, beta polypeptide), which encodes the beta chain of the heterodimeric CD8 glycoprotein responsible for recognizing antigenic peptides on the surface of immune T cells. This gene was previously thought to have duplicated in a common ancestor of human, gorilla and chimpanzee [14]. Our analysis shows a gorilla-specific 40 kbp duplication of *CD8B* exon 1 not shared with human, chimpanzee, bonobo or orangutan (Supplementary Figure 20). The gorilla duplication has no overlap with our HSD, which encompasses *CD8B* exons 2 to 5; our timing estimate of *CD8B-B*, which places the duplication at just after the human–chimpanzee divergence from a common ancestor, supports this finding (Figure 2).

Genomic characterization of *SMN1* locus

When we compared our new chimpanzee reference for chromosome 5q13.3 containing *SMN1* [37,38], we found that human and chimpanzee differ by 1.3 Mbp (Supplementary Figure 21). Despite having complete sequences of the two species, we were unable to delineate the precise mechanism of duplications leading from the assumed ancestral structure in the chimpanzee genome to the configuration in the human reference today. Our efforts were hindered by the presence of high copy (CN > 8), polymorphic, nearly identical, palindromic duplications peppered throughout this locus that likely arose in the last two million years. Sequencing additional human haplotypes will be necessary to delineate the precise mechanisms leading to the rapid expansion of this complex region.

Evidence of ancestral paralog deletions

We leveraged SUNKs to investigate copy number variation diversity at the level of individual paralogs [39] (Supplementary Tables 10, 23 and 24). Among the 72 gene paralogs assayed here, we identified apparent homozygous deletions for 24 paralogs in at least one human individual. Genotyping implicated three ancestral paralogs as showing evidence of heterozygous deletions (Supplementary Figure 22); closer inspection of copy number heatmaps found that *FRMPD2A* and *OCLN-A* appeared to be truly deleted, though in the latter case gene conversion may be at play based on the known disease relevance of this gene [40] (Supplementary Figure 23). Meanwhile, though deletions appear to exist within portions of the HSD-containing *CHRNA7*, the gene itself appeared undeleted. These results highlight certain limitations of this SUNK approach, especially when paralogs share high-sequence similarity [15,41].

Alternative derived gene model of *ARHGAP11B* places common LGD variant in intron of gene

A common LGD variant was identified in *ARHGAP11B*, an HSD previously implicated in neuronal migration [36], suggesting the gene may not be functionally relevant (Supplementary Tables 10 and 19). Closer inspection shows the LGD variant falling within an *ARHGAP11B*-specific intron due to an alternative gene model of *ARHGAP11B* compared to its ancestral counterpart (Supplementary Figure 24). We propose that systematic characterization of gene models for other HSD genes is necessary to fully understand the impact of LGD variants on function.

SUPPLEMENTARY METHODS

Identification of HSD regions from Illumina data

Whole-genome Illumina short-read mappings against the human reference genome (GRCh37/hg19) of a diverse, high-coverage set of 236 human genomes from the HGDP [42] and a set of 86 NHP genomes [43] were used to estimate aggregate copy number in 500 bp windows using previously described methods [39]. Windows with >90% of human genomes at copy number >2.5 and >90% of NHP genomes at copy number <2.8 were identified as HSDs. The copy number cutoff for NHP genomes was empirically determined using previously identified HSD genes to reduce false negatives [39,44]. These regions were merged if within 1 kbp of each other with the final HSD regions, including merged windows 5 kbp in size or greater. To connect adjacent HSD regions punctuated by higher ancient "core" duplicons [45,46] (found duplicated across all great apes), HSDs within 20 kbp of each other were additionally merged if, in all the windows between the regions, >90% of human genomes had copy number >2.8.

Validation of HSD regions

We intersected HSD regions with SDs identified using whole-genome shotgun sequence detection (WSSD) using Sanger sequence read depth [11] and whole-genome analysis comparison (WGAC) [1] methods (Supplementary Table 1). We compared the 37 regions that did not intersect previously identified SDs to a genome assembly of the CHM1 haploid hydatidiform mole (NCBI Assembly PacBioCHM1_r2_GenBank_08312015) using BLASR to look for orthogonal evidence that these regions might be resolved or collapsed duplications [47,48]. We counted a query region as resolved if it had multiple mappings with match length minus edit distance greater than 90% of the query length. To find collapsed duplications, we calculated coverage across the CHM1 assembly in 100 bp windows and identified regions >5 kbp of elevated coverage, where the threshold for elevated coverage was set at the third quartile plus two times the interquartile range (or 72.7X).

Sequencing of BAC clones

DNA from CH17, CH251, CH276 and CH277 BAC clone libraries was isolated, prepped into barcoded genomic libraries, and sequenced (150 bp paired end) on an Illumina MiSeq using a Nextera protocol [31]. Sequencing data (~300-fold coverage) were mapped with mrFAST [49] to the human reference genome (GRCh37) and SUNKs were used to discriminate between highly identical SDs [39]. Some clones were subject to a hierarchical clone-based strategy with high-quality capillary fluorescent-based sequencing. This entailed the construction of genomic libraries, sequencing of paired-end shotgun libraries, and assembly of inserts into a finished sequencing contig. The majority of clones were subject to PacBio-based sequencing entailing the preparation of SMRTbell libraries sequenced using PacBio RSII C2P4 or C4P6 chemistry (one SMRT cell/BAC sample with two 45-minute movies) for a subset of clones spanning HSD regions based on Illumina sequence analyses. Inserts were assembled using Quiver and HGAP (Hierarchical Genome Assembly Process) as described [50]. BACs were assembled into contigs with PacBio- and capillary-sequenced clones using Sequencher and compared to the human reference genome using Miropeats [51] and BLAST [52]. Sequences for contigs not included in the human reference can be found in Supplementary Dataset 1. Additional BAC clones mapping to HSD regions in human and NHPs were identified within GenBank, previously sequenced by the Wellcome Trust Sanger Center and McDonnell Genome Institute at Washington University as part of separate projects, which we included in breakpoint and evolutionary analyses (Supplementary Tables 3 and 6). These clones were not included in counts of BACs sequenced for this study.

Breakpoint identification

A combination of BLAST [52], BLAT [53], and WGAC [1] methods were used to identify HSD paralogous regions. Sequences (±500 bp) were subsequently extracted from the human reference genome (GRCh38) or from BAC-assembled contigs and pairwise comparisons performed and visualized using BLAST and Miropeats [51] to identify the maximal duplicon breakpoints. When precise breakpoints could not be defined due to flanking core duplicons, minimal and maximal breakpoints were reported for both ancestral and duplicate paralogs. When comparing sizes of HSDs, we used the greater of the two sizes between ancestral and duplicate paralogs. In size estimates and comparisons of HSDs, we excluded the *NAIP-C* HSD due to uncertainty of its status as a primary or secondary duplication. We found that some sets of duplication sizes were drawn from a non-normal distribution using the Shapiro-Wilk test (e.g., maximum size of HSDs in

period III); hence, we compared all duplications sizes using nonparametric tests. Specifically, we applied the Wilcoxon-Mann-Whitney test—using the wilcox.test() function in R (version 2.15.0)—when comparing primary versus secondary duplication sizes and the Kruskal-Wallis rank sum test to assess size differences across all three evolutionary periods. After identifying significant differences between time periods of minimum primary duplication sizes, we applied a Wilcoxon-Mann-Whitney test post hoc to identify the duplication waves that were significantly different and adjusted for multiple comparisons using the Holm method. All cumulative sizes were reported as medians.

HSD clustering simulations

We simulated a null distribution by shuffling 218 HSD regions (defined in Supplementary Table 1; GRCh37) and an unbiased set of 18 primary HSDs (of our total 24) containing previously identified HSD genes [39] (derived genomic coordinates only, Supplementary Table 5; GRCh38) within the same chromosome 10 million times using BEDTools shuffle (v2.23.0) and, when multiple duplications occurred on a single chromosome, calculated the distance to the nearest duplication using midpoint coordinates. In this and all subsequent simulations, shuffled intervals were not allowed to intersect non-scaffold gap regions. We calculated the median distance for each iteration of the simulation and compared this distribution to the empirical value. We also determined midpoint distances of the 18 HSD primary duplication to the nearest non-primary HSD defined by WGAC [1] except shuffling was performed one million times. All analyses were repeated allowing shuffling genome-wide.

We recalculated the sequence and location of core duplicons using coordinates of all duplicon clades previously defined [45]. As the original definition of a core duplicon required all duplicons to occur on the same chromosome, we limited our analysis to clades with intrachromosomal duplicons. We extracted the GRCh35/hg17 sequence corresponding to each duplicon and aligned all duplicon sequences against themselves with BLASR using parameters tuned for high-quality queries (-affineAlign -affineOpen 8 -affineExtend 0 -bestn 30 -maxMatch 30 -sdpTupleSize 13) [54]. We retained all alignments >100 bp between duplicons from the same clade except for alignments of each duplicon to itself. For each clade, we clustered all alignments that reciprocally overlapped by 50% or more and selected the cluster with the most components (i.e., alignments from other duplicons in the same clade) as the representative core for the clade. We aligned the core sequences to GRCh38/hg38 with BLASR using the same parameters described above and filtering matches <75% of the query length. A null distribution of median distance to the nearest core duplicon was created using simulations performed as described above with shuffling of the 18 HSD primary duplications 10 million times.

Evolutionary analysis

Sequences from HSD orthologs were identified and extracted from genome reference or BAC assemblies for chimpanzee (panTro4 and CH251), gorilla (gorGor4.1 and CH277), and orangutan (ponAbe2 and CH276). Multiple sequence alignments (MSAs) were generated using MAFFT and included the maximal shared genomic regions of human paralogs and nonhuman orthologs excluding any flanking core duplicons [55]. Alignments were visualized for manual editing using Jalview [56]. MSAs are available in Supplementary Dataset 1. Phylogenetic analyses were performed using MEGA6 [2] (Supplementary Dataset 2). If a full-length alignment did not pass the Tajima's D relative rate test [7], using orangutan as the outgroup, pairwise sequence identities were calculated across 500 bp sliding windows with 100 bp increments using the PopGenome statistical package [57] and visualized using ggplot in R. Portions of the alignment that exhibited aberrant spikes in sequence identity were excluded and phylogenetic analyses were repeated on a refined region (Supplementary Table 8). In the case of HSD regions containing *SRGAP2*, corrections were made to distance estimates to account for differences in substitution rates of paralogous regions as previously described [15]. Duplication mechanisms were predicted using a combined approach of defining ancestral paralogs/configurations using genomic synteny taken from chimpanzee and/or orangutan and evolutionary timing estimates to predict the order of rearrangements.

Detection of interlocus gene conversion

We implemented two approaches to detect signatures consistent with recent interlocus gene conversion among HSD paralogs. First, we created pairwise sequence alignments between all annotated HSDs containing the ancestral and duplicate paralogs. We next calculated the identity of two aligned sequences over 2 kbp sliding windows across the alignment with a stepwise increment of 100 bp (Supplementary Dataset 1). Pairwise sequence identity was plotted against the length of the alignment and potential interlocus gene conversion events were identified by the presence of sharp

sequence identity transitions from low <99% to high ≥1.0. All sequence identity plots are included in Supplementary Dataset 1. We next used the program GENECONV to scan each MSA used in the phylogenetic analysis of HSD regions. GENECONV identifies pairs of sequences with longer than expected tracks of 100% sequence identity conditioning on the overall pattern of variable sites in the alignment [58,59]. The program was run using default parameters and tracks with a global $P < 0.05$ were considered significant for follow-up analysis [58]. Tracks of perfect sequence identity were then hardmasked from the original MSAs using BEDTools and the resulting branch lengths were then recalculated from the phylogenies (Supplementary Table 8).

Copy number genotyping

Raw sequences from 236 human individuals from HGDP [42], 2,143 human individuals through Phase 3 of the 1000 Genomes Project [33], 86 NHP individuals from the Great Ape Genome Project [including bonobo (N = 14), chimpanzee (N = 23), gorilla (N = 32), and orangutan (N = 17)] [43], a Denisovan individual [60], a Neanderthal individual [22], and three archaic hominids [61,62] were mapped to the human reference genome (GRCh37) using mrsFAST [63]. Overall read-depth (WSSD) and paralog-specific read-depth (SUNK) approaches were performed genome-wide across 500 bp sliding windows in 100 bp increments using previously described methods [39] and visualized as heatmaps using bigBed tracks within the UCSC Genome Browser. Using these same data, we genotyped the average copy number across 23 HSD units (i.e., regions where the same HSD gene families were always found together on duplicate paralogs) containing duplicated gene families (Supplementary Table 9). The genotypes were used in subsequent downstream analyses. Human population diversity of individual gene paralogs (SUNK) and families (WSSD) was determined by calculating the mean, median and standard deviation of genotyped copy numbers. We identified the most copy number polymorphic gene families, defined here as those showing the greatest copy number variance as measured by standard deviation in the human species. We used the V_{st} statistic [64] (calculated using a custom python script) to measure copy number stratification between populations. Since many of these duplicated genes reside clustered together in groups (Figure 1), we defined 16 genomic regions containing the 33 HSD gene families to assess variation in humans across these loci (Supplementary Table 2).

FISH analysis

Metaphase spreads and interphase nuclei were obtained from lymphoblast cell lines from HapMap and 1000 Genomes Projects (Supplementary Table 13; Coriell Cell Repository, Camden, NJ). FISH experiments were performed using fosmid clones (Supplementary Table 12) directly labeled by nick-translation with Cy3-dUTP (Perkin-Elmer), as described previously [65], with minor modifications. Briefly, 300 ng of labeled probe were used for the FISH experiments; hybridization was performed at 37°C in 2xSSC, 50% (v/v) formamide, 10% (w/v) dextran sulfate, and 3 µg sonicated salmon sperm DNA, in a volume of 10 µL. Posthybridization washing was at 60°C in 0.1xSSC (three times, high stringency). Nuclei were simultaneously DAPI stained. Digital images were obtained using a Leica DMRXA2 epifluorescence microscope equipped with a cooled CCD camera (Princeton Instruments). DAPI and Cy3 fluorescence signals, detected with specific filters, were recorded separately as gray-scale images. Pseudocoloring and merging of images were performed using Adobe Photoshop software.

RNA-seq analysis

GTEx RNA-seq data from different subtissues (dbGaP version phs000424.v3.p1) were used to analyze the expression of a set of representative transcripts from hg38 RefSeq annotation. First, 30-mers within these transcript's exons that do not appear anywhere else in hg38 genome were detected. Then, for each such unique 30-mer, the number of reads that include this 30-mer was normalized by dividing by the total number of reads in the sample, and multiplying by 10^9. Next, for each subtissue and each unique 30-mer, the median of the normalized counts was calculated over all the samples from this subtissue. Finally, for each subtissue and each transcript, a median value was calculated over all the unique 30-mers found in this transcript, if the normalized value was higher than zero. Due to an insufficient number of k-mers to distinguish paralogs (i.e., *BOLA2*) or remaining gaps within our HSD genomic builds causing only single paralogs to be represented (i.e., *DUSP22* and *GPRIN2*), total overall expression was instead calculated for three gene families. Also, we did not quantify expression if the genomic paralogous sequence was present in the human reference but no representative RefSeq transcript could be assigned to the paralog (e.g., *CD8B-B*, *PTPN20A*, and *TCAF1B*). Finally, in a few gene families with short transcript lengths we detected very low expression for nearly all human tissues (i.e., *NYP4R*, *TISP43*, and *OR21A*), which we note could be the result of a technical artefact caused by too few SUNKs. In addition, we also quantified

transcript expression for each sample in the GTEx dataset by applying the Sailfish method version 0.63 with the default parameters and k = 20, using the hg38 transcriptome (downloaded on April 16, 2015) [9] (Supplementary Figure 14).

Molecular inversion probe (MIP) targeted sequencing

Human reference sequence (GRCh37) of the CDS exons from each ancestral paralog (±5 bp) was used as input to design single-molecule MIPs using MIPgen [66]. Each MIP was designed to capture 112 bp of genomic sequence and included 40 bp unique to the target of a region (split between a ligation and an extension arm of the MIP), a universal 30 bp backbone, and a degenerate 8 bp unique tag included on the extension arm (Supplementary Table 18). In cases where the ancestral paralog was unclear, a paralog was arbitrarily chosen. We ran a separate pipeline for *SRGAP2* using the most recent human reference genome (GRCh38) for the MIP design (since this gene had been resolved in the most recent build). MIP phosphorylation, capture, and barcoding were performed as previously described [67]. Briefly, oligos were pooled together at equal concentrations (100 uM), phosphorylated, and an 800:1 excess of oligos were used for the genomic DNA capture (100 ng). Capture reactions were incubated at 60°C for 18 hours. Finished libraries were pooled together and sequenced using either MiSeq (2 x 150 bp) or HiSeq2000 (2 x 101 bp). Sequence capture was first performed on control DNA from the 1000 Genome Project [N = 658 including European (N = 395) and African (N = 263) populations]. The high coverage per MIP (on average 86-fold sequence coverage per individual) allowed us to sensitively detect single-nucleotide and small indel events in the exonic regions and to estimate their frequency in the human population (1,030 MIPs with >10-fold coverage). Subsequently, capture was performed on DNA from the Simons Simplex Collection (SSC) [68], Autism Genetic Resource Exchange (AGRE) [69], and The Autism Simplex Collection (TASC) [70] cohorts. In total, we targeted an additional 6,061 individuals (N = 3,444 children with autism and N = 2,617 unaffected siblings) for sequencing [~261X average coverage (N = 1,096 MIPs) with 1,058 MIPs with >10-fold coverage].

We used the MIPgen data analysis pipeline to map and filter reads in FASTQ format to a minimal human reference containing only the regions included in our MIP design with the remaining regions masked out. This masking ensured reads mapped to only one paralog per gene family. Discovery variant calling was performed across the entire 1000 Genomes Project cohort or for the autism spectrum disorder cohort per pooled sequence set containing up to 384 samples using FreeBayes (https://github.com/ekg/freebayes) with the following command: freebayes -b <sorted_bams> -f <masked_reference> -t <targeted_regions> -F 0.07 -C 2 -n 4. We removed any variants with the following feature: trinucleotide or homopolymer repeat, read depth ≤10, quality score ≤20, or with no alleles using previously described methods [28]. The resulting variant set was annotated utilizing the Ensembl Variant Effect Predictor (VEP) [71] using the canonical transcript for each gene. Subsequently, for the autism spectrum disorder study, the complete list of coding variants was used to separately genotype cases and controls to assess overall frequency of events in each cohort: freebayes -b <sorted_bams> -f <masked_reference> -s <sample_list> -@ <variant_vcf> --only-use-input-alleles -F 0.07 -C 2 -n 4 --min-coverage 10. LGD variants, which included frameshift, stop gain and loss, and splice donor and acceptor mutations, were highlighted. MAXENT, which can monitor the dependencies between different positions by using a maximum-entropy distribution consistent with lower order marginal constraints, was used to predict the effect of splice-site mutations [3,72]. The severity of missense mutations was predicted using the Combined Annotation Dependent Depletion (CADD) score for all genic variants precomputed for human reference GRCh37 except for *SRGAP2*, in which variants were annotated in GRCh38 [73]. In some cases, variants were definitively assigned to a paralogous copy by manual inspection in the Integrative Genomic Viewer [74] and manually annotated by identifying molecularly phased reads containing the identified variant and a known paralogous sequence variant (identified from our MSAs).

SUPPLEMENTARY INFORMATION REFERENCES

1 Bailey, J. A., Yavor, A. M., Massa, H. F., Trask, B. J. & Eichler, E. E. Segmental duplications: organization and impact within the current human genome project assembly. *Genome research* **11**, 1005-1017, doi:10.1101/gr.187101 (2001).

2 Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Molecular biology and evolution* **30**, 2725-2729, doi:10.1093/molbev/mst197 (2013).

3 Jian, X., Boerwinkle, E. & Liu, X. In silico prediction of splice-altering single nucleotide variants in the human genome. *Nucleic Acids Res* **42**, 13534-13544, doi:10.1093/nar/gku1206 (2014).

4 Rzhetsky, A. & Nei, M. A Simple Method for Estimating and Testing Minimum-Evolution Trees. *Molecular biology and evolution* **9**, 945-967 (1992).

5 Dopazo, J. Estimating errors and confidence intervals for branch lengths in phylogenetic trees by a bootstrap approach. *J Mol Evol* **38**, 300-304 (1994).

6 Kimura, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* **16**, 111-120 (1980).

7 Tajima, F. Simple methods for testing the molecular evolutionary clock hypothesis. *Genetics* **135**, 599-607 (1993).

8 GTEx_Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648-660, doi:10.1126/science.1262110 (2015).

9 Patro, R., Mount, S. M. & Kingsford, C. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nature biotechnology* **32**, 462-464, doi:10.1038/nbt.2862 (2014).

10 Liu, W. *et al.* IBS: an illustrator for the presentation and visualization of biological sequences. *Bioinformatics* **31**, 3359-3361, doi:10.1093/bioinformatics/btv362 (2015).

11 Bailey, J. A. *et al.* Recent segmental duplications in the human genome. *Science* **297**, 1003-1007, doi:10.1126/science.1072047 (2002).

12 Jiang, Z., Hubley, R., Smit, A. & Eichler, E. E. DupMasker: a tool for annotating primate segmental duplications. *Genome research* **18**, 1362-1368, doi:10.1101/gr.078477.108 (2008).

13 Antonell, A., de Luis, O., Domingo-Roura, X. & Perez-Jurado, L. A. Evolutionary mechanisms shaping the genomic structure of the Williams-Beuren syndrome chromosomal region at human 7q11.23. *Genome research* **15**, 1179-1188, doi:10.1101/gr.3944605 (2005).

14 Delarbre, C., Nakauchi, H., Bontrop, R., Kourilsky, P. & Gachelin, G. Duplication of the CD8 beta-chain gene as a marker of the man-gorilla-chimpanzee clade. *Proc Natl Acad Sci U S A* **90**, 7049-7053 (1993).

15 Dennis, M. Y. *et al.* Evolution of human-specific neural SRGAP2 genes by incomplete segmental duplication. *Cell* **149**, 912-922, doi:10.1016/j.cell.2012.03.033 (2012).

16 O'Bleness, M. *et al.* Finished sequence and assembly of the DUF1220-rich 1q21 region using a haploid human genome. *BMC Genomics* **15**, 387, doi:10.1186/1471-2164-15-387 (2014).

17 Davy, B. E. & Robinson, M. L. Congenital hydrocephalus in hy3 mice is caused by a frameshift mutation in Hydin, a large novel gene. *Human molecular genetics* **12**, 1163-1170 (2003).

18 Olbrich, H. *et al.* Recessive HYDIN mutations cause primary ciliary dyskinesia without randomization of left-right body asymmetry. *American journal of human genetics* **91**, 672-684, doi:10.1016/j.ajhg.2012.08.016 (2012).

19 Doggett, N. A. *et al.* A 360-kb interchromosomal duplication of the human HYDIN locus. *Genomics* **88**, 762-771, doi:10.1016/j.ygeno.2006.07.012 (2006).

20 Charrier, C. *et al.* Inhibition of SRGAP2 function by its human-specific paralogs induces neoteny during spine maturation. *Cell* **149**, 923-935, doi:10.1016/j.cell.2012.03.034 (2012).

21 Nuttle, X. *et al.* Emergence of a Homo sapiens-specific gene family and chromosome 16p11.2 CNV susceptibility. *Nature*, doi:10.1038/nature19075 (2016).

22 Prufer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43-49, doi:10.1038/nature12886 (2014).

23 Marchetto, M. C. *et al.* Differential L1 regulation in pluripotent stem cells of humans and apes. *Nature* **503**, 525-529, doi:10.1038/nature12686 (2013).

24 Weiss, L. A. *et al.* Association between microdeletion and microduplication at 16p11.2 and autism. *The New England journal of medicine* **358**, 667-675, doi:10.1056/NEJMoa075974 (2008).

25 Kumar, R. A. *et al.* Recurrent 16p11.2 microdeletions in autism. *Human molecular genetics* **17**, 628-638, doi:10.1093/hmg/ddm376 (2008).

26 Gkika, D. *et al.* TRP channel-associated factors are a novel protein family that regulates TRPM8 trafficking and activity. *J Cell Biol* **208**, 89-107, doi:10.1083/jcb.201402076 (2015).

27 Stankiewicz, P. *et al.* Recurrent deletions and reciprocal duplications of 10q11.21q11.23 including CHAT and SLC18A3 are likely mediated by complex low-copy repeats. *Hum Mutat* **33**, 165-179, doi:10.1002/humu.21614 (2012).

28 Coe, B. P. *et al.* Refining analyses of copy number variation identifies specific genes associated with developmental delay. *Nature genetics* **46**, 1063-1071, doi:10.1038/ng.3092 (2014).

29 Antonacci, F. *et al.* Palindromic GOLGA8 core duplicons promote chromosome 15q13.3 microdeletion and evolutionary instability. *Nature genetics* **46**, 1293-1302, doi:10.1038/ng.3120 (2014).

30 Boettger, L. M., Handsaker, R. E., Zody, M. C. & McCarroll, S. A. Structural haplotypes and recent evolution of the human 17q21.31 region. *Nature genetics* **44**, 881-885, doi:10.1038/ng.2334 (2012).

31 Steinberg, K. M. *et al.* Structural diversity and African origin of the 17q21.31 inversion polymorphism. *Nature genetics* **44**, 872-880, doi:10.1038/ng.2335 (2012).

32 Hermetz, K. E. *et al.* Large inverted duplications in the human genome form via a fold-back mechanism. *PLoS Genet* **10**, e1004139, doi:10.1371/journal.pgen.1004139 (2014).

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

34 Weckselblatt, B. & Rudd, M. K. Human Structural Variation: Mechanisms of Chromosome Rearrangements. *Trends Genet* **31**, 587-599, doi:10.1016/j.tig.2015.05.010 (2015).

35 Guerrier, S. *et al.* The F-BAR domain of srGAP2 induces membrane protrusions required for neuronal migration and morphogenesis. *Cell* **138**, 990-1004, doi:10.1016/j.cell.2009.06.047 (2009).

36 Florio, M. *et al.* Human-specific gene ARHGAP11B promotes basal progenitor amplification and neocortex expansion. *Science* **347**, 1465-1470, doi:10.1126/science.aaa1975 (2015).

37 Lefebvre, S. *et al.* Identification and characterization of a spinal muscular atrophy-determining gene. *Cell* **80**, 155-165 (1995).

38 Roy, N. *et al.* The gene for neuronal apoptosis inhibitory protein is partially deleted in individuals with spinal muscular atrophy. *Cell* **80**, 167-178 (1995).

39 Sudmant, P. H. *et al.* Diversity of human copy number variation and multicopy genes. *Science* **330**, 641-646, doi:10.1126/science.1197005 (2010).

40 O'Driscoll, M. C. *et al.* Recessive mutations in the gene encoding the tight junction protein occludin cause band-like calcification with simplified gyration and polymicrogyria. *American journal of human genetics* **87**, 354-364, doi:10.1016/j.ajhg.2010.07.012 (2010).

41 Nuttle, X. *et al.* Rapid and accurate large-scale genotyping of duplicated genes and discovery of interlocus gene conversions. *Nature methods* **10**, 903-909, doi:10.1038/nmeth.2572 (2013).

42 Sudmant, P. H. *et al.* Global diversity, population stratification, and selection of human copy-number variation. *Science* **349**, aab3761, doi:10.1126/science.aab3761 (2015).

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

44 Fortna, A. *et al.* Lineage-specific gene duplication and loss in human and great ape evolution. *PLoS Biol* **2**, E207, doi:10.1371/journal.pbio.0020207 (2004).

45 Jiang, Z. *et al.* Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution. *Nature genetics* **39**, 1361-1368, doi:10.1038/ng.2007.9 (2007).

46 Ji, X. & Zhao, S. DA and Xiao-two giant and composite LTR-retrotransposon-like elements identified in the human genome. *Genomics* **91**, 249-258, doi:10.1016/j.ygeno.2007.10.014 (2008).

47 Chaisson, M. J. *et al.* Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517**, 608-611, doi:10.1038/nature13907 (2015).

48 Steinberg, K. M. *et al.* Single haplotype assembly of the human genome from a hydatidiform mole. *Genome research* **24**, 2066-2076, doi:10.1101/gr.180893.114 (2014).

49 Alkan, C. *et al.* Personalized copy number and segmental duplication maps using next-generation sequencing. *Nature genetics* **41**, 1061-1067, doi:10.1038/ng.437 (2009).

50 Huddleston, J. *et al.* Reconstructing complex regions of genomes using long-read sequencing technology. *Genome research* **24**, 688-696, doi:10.1101/gr.168450.113 (2014).

51 Parsons, J. D. Miropeats: graphical DNA sequence comparisons. *Computer applications in the biosciences : CABIOS* **11**, 615-619 (1995).

52 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403-410, doi:10.1016/S0022-2836(05)80360-2 (1990).

53 Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome research* **12**, 656-664, doi:10.1101/gr.229202. Article published online before March 2002 (2002).

54 Chaisson, M. J. & Tesler, G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* **13**, 238, doi:10.1186/1471-2105-13-238 (2012).

55 Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution* **30**, 772-780, doi:10.1093/molbev/mst010 (2013).

56 Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M. & Barton, G. J. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189-1191, doi:10.1093/bioinformatics/btp033 (2009).

57 Pfeifer, B., Wittelsburger, U., Ramos-Onsins, S. E. & Lercher, M. J. PopGenome: an efficient Swiss army knife for population genomic analyses in R. *Molecular biology and evolution* **31**, 1929-1936, doi:10.1093/molbev/msu136 (2014).

58 Dumont, B. L. & Eichler, E. E. Signals of historical interlocus gene conversion in human segmental duplications. *PloS one* **8**, e75949, doi:10.1371/journal.pone.0075949 (2013).

59 Sawyer, S. Statistical tests for detecting gene conversion. *Molecular biology and evolution* **6**, 526-538 (1989).

60 Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222-226, doi:10.1126/science.1224344 (2012).

61 Fu, Q. *et al.* Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* **514**, 445-449, doi:10.1038/nature13810 (2014).

62 Lazaridis, I. *et al.* Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513**, 409-413, doi:10.1038/nature13673 (2014).

63 Hach, F. *et al.* mrsFAST: a cache-oblivious algorithm for short-read mapping. *Nature methods* **7**, 576-577, doi:10.1038/nmeth0810-576 (2010).

64 Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444**, 444-454, doi:10.1038/nature05329 (2006).

65 Antonacci, F. *et al.* A large and complex structural polymorphism at 16p12.1 underlies microdeletion disease risk. *Nature genetics* **42**, 745-750, doi:10.1038/ng.643 (2010).

66 Boyle, E. A., O'Roak, B. J., Martin, B. K., Kumar, A. & Shendure, J. MIPgen: optimized modeling and design of molecular inversion probes for targeted resequencing. *Bioinformatics* **30**, 2670-2672, doi:10.1093/bioinformatics/btu353 (2014).

67 O'Roak, B. J. *et al.* Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. *Science* **338**, 1619-1622, doi:10.1126/science.1227764 (2012).

68 Fischbach, G. D. & Lord, C. The Simons Simplex Collection: a resource for identification of autism genetic risk factors. *Neuron* **68**, 192-195, doi:10.1016/j.neuron.2010.10.006 (2010).

69 Geschwind, D. H. *et al.* The autism genetic resource exchange: a resource for the study of autism and related neuropsychiatric conditions. *American journal of human genetics* **69**, 463-466, doi:10.1086/321292 (2001).

70 Buxbaum, J. D. *et al.* The Autism Simplex Collection: an international, expertly phenotyped autism sample for genetic and phenotypic analyses. *Mol Autism* **5**, 34, doi:10.1186/2040-2392-5-34 (2014).

71 Cunningham, F. *et al.* Ensembl 2015. *Nucleic Acids Res* **43**, D662-669, doi:10.1093/nar/gku1010 (2015).

72 Yeo, G. & Burge, C. B. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol* **11**, 377-394, doi:10.1089/1066527041410418 (2004).
73 Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics* **46**, 310-315, doi:10.1038/ng.2892 (2014).
74 Robinson, J. T. *et al.* Integrative genomics viewer. *Nature biotechnology* **29**, 24-26, doi:10.1038/nbt.1754 (2011).