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. Stessman[2], Xander Nuttle[2], Kendra Hoekzema[2], Tina A. 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

ABSTRACT

Segmental duplications contribute significantly to the evolution, adaptation and disease-associated instability of the human genome. The largest and most identical duplications suffer from the poorest characterization, often corresponding to genome gaps and misassembly. Here we focus on creating a framework to understand the evolution, copy number variation and coding potential of human-specific segmental duplications (HSDs). We identify 218 HSDs (>5 kbp in length) based on analysis of 322 deeply sequenced ape and human genomes. We target 268 large-insert human bacterial artificial chromosomes, 85 of which have been incorporated into the most recent human reference build (GRCh38) correcting 24 large euchromatic gaps, and 269 nonhuman primate clones for finished sequencing in order to resolve the structure and evolution of the largest, most complex regions with protein-coding potential (n=80 genes/33 gene families). Our analyses indicate that these HSDs (28 duplications ranging in length from 11–677 kbp) are non-randomly organized ($P<1x10^{-6}$), cluster in association with core duplicons ($P<1x10^{-7}$) and the majority represent intrachromosomal events arranged predominantly in an interspersed inverted orientation (18/26; P=0.014). Phylogenetic reconstruction suggests different waves of HSD with the latest burst occurring <1.3 million years ago. These 16 duplications and 28 genes would be specific to the genus *Homo*, including three gene families absent in ancient Neanderthal and Denisova genomes. Of particular interest are the *TCAF1*/*TCAF2* family, which is the most stratified of the *Homo sapiens*-specific duplications and has been implicated in the somatosensation of cold. Overall, copy number variation analysis (n=2,379 genomes), RNA sequence mapping (GTEx) and targeted resequencing of the protein-coding regions (n=3,275 controls) identify ten gene families where copy number never returns to the ancestral state, there is evidence of mRNA splicing and expression, and no common gene-disruptive mutation events are observed in the general population. We propose that this subset of genes, including functional paralogs *ARHGAP11B* and *SRGAP2C*, represents excellent candidates for the evolution of human-specific adaptive traits.

INTRODUCTION

Genetic mutations have shaped the unique adaptation and evolution of the human lineage, but their characterization has been a slow and difficult endeavor. Despite a few potential success stories over the years with various degrees of support (e.g., *FOXP2* [1,2], *HAR1F* [3], *AQP7* [4], *HACNS1* [5], *MYH16* [6], and *GADD45G* [7]), the genetic basis of most of the unique aspects of human adaptation await discovery. As sequencing technologies have improved, more systematic efforts have been directed to discover regulatory differences among the great apes [7-11]. One potential source of genetic variation, which has been difficult to explore due to missing or erroneous sequences within reference genomes, are genes embedded within recently duplicated regions also called segmental duplications (SDs) [12]. Unlike the focus on regulatory mutations or gene loss, which typically modify the expression of ancestral genes mapping to unique regions, duplicated regions have long been recognized as a potential source for the rapid evolution of new genes with novel functions [13]. Recent functional studies have emphasized the potential importance of SDs with respect to unique features of synaptogenesis, neuronal migration and neocortical expansion in the human lineage [14-17].

The genomes of apes are enriched in SDs having experienced a burst of interspersed duplications over the last 10 million years of evolution [18,19]. The mosaic and interspersed architecture of ape SDs offers tremendous potential for transcript innovation because duplicate paralogs may be truncated, combined with other transcripts to create fusion genes, or acquire alternate promoters directing the differential expression of novel transcripts [20]. Previous investigations have been limited to microarray studies [21,22] and whole-genome sequencing read-depth comparisons [19,23,24] between humans and great apes. None of these methods provided information regarding the structure and sequence of the duplicated segments limiting gene annotation and an understanding of the functional potential of the duplicated genes.

In this study, we focus on understanding the sequence structure, genetic variation and transcriptional potential of the largest human-specific segmental duplications (HSDs). HSDs are particularly problematic because they are highly identical (~99%), among the

most copy number polymorphic parts of the genome, and are frequently embedded within larger blocks of shared ape duplications. Not surprisingly, the regions are highly enriched for euchromatic gaps and misassembly errors even within the most recent versions of the human genome [25,26]. We specifically target 33 human-specific gene families contained within these HSDs for high-quality sequence assembly by selecting large-insert bacterial artificial chromosome (BAC) clones from a library (CHORI-17) generated from a well-characterized complete hydatidiform mole cell line (CHM1tert). The mole derives from the fertilization of an enucleated human oocyte with a single spermatozoon [27,28] or from postzygotic loss of a complete parental genome [29]. The end result is a haploid as opposed to a diploid equivalent of the human genome where the absence of allelic variation allows high-identity paralogous regions of the genome to be rapidly resolved [16,30]. We apply the resulting high-quality sequence to more systematically investigate copy number variation, transcriptional potential and human genetic variation in an effort to understand their evolutionary history as well as discover regions that have become fixed and potentially functional in the human species.

RESULTS

Refining regions of HSDs

With the wealth of deep-coverage Illumina sequence data from both humans and great apes, we began by first redefining the map location of HSDs. We mapped a genetically diverse panel of 236 human and 86 chimpanzee, gorilla and orangutan genomes to the human reference (GRCh37) to identify regions uniquely duplicated in humans (Figure 1, Supplementary Figure 1). Operationally, we defined HSDs as regions with evidence of copy number gain in >90% of all humans but where >90% of all great apes were diploid for the locus (<3 copies). The approach identified 217 autosomal regions ranging in size from 5 kbp (our size threshold) to 362 kbp with HSDs dispersed non-randomly near each other (median distance to nearest HSD 440 kbp, empirical $P < 1x10^{7}$; Supplementary Figure 2A). Of these regions, 85 corresponded to entire or parts of RefSeq annotated genes (Supplementary Table 1). We orthogonally validated 88% (190/217) of our events as HSDs by whole-genome analysis comparison (WGAC) of the human reference [31], whole-genome shotgun sequence detection (WSSD) using Sanger sequence read depth [32],

or by analysis of a recent *de novo* assembly of a human haploid genome (CHM1) generated using single-molecule, real-time (SMRT) sequencing data [25,33].

Overall, we identified 38 previously unreported HSDs corresponding to genic regions in the human genome. Among these, we included HSDs where there was evidence of independent or distinct duplications in great apes (i.e., homoplasy; N = 21) and duplications corresponding to introns (N = 12). For example, the 3' portion of *MST1L* (macrophage stimulating 1 like) on chromosome 1p36.13 is partially duplicated in chimpanzee and gorilla, but a complete duplication of the gene (>36 kbp) has risen to high copy uniquely in humans (diploid copy number (CN) > 8; Supplementary Figure 3A). Similarly, we identified a 6.6 kbp duplication corresponding the third intron of *CACNA1B* (calcium voltage-gated channel subunit alpha1 B)—a pore-forming subunit of an N-type voltage-dependent calcium channel that controls neurotransmitter release from neurons (Supplementary Figure 3B). This HSD was missed by previous SD analyses (i.e., WGAC and WSSD) but was represented as two distinct paralogs within the *de novo* assembly of haploid CHM1 genome (Supplementary Figure 4). We also identified a novel duplication of *SCGB1C1* (secretoglobin family 1C member 1), a gene family whose products are secreted at large concentrations in the lung, lacrimal and salivary glands (Supplementary Figure 3C).

Next, we focused on the largest gene-containing HSD regions (>20 kbp). These HSDs and their ancestral counterparts reside on 16 autosomal regions with many appearing to cluster with other smaller HSDs and at "genomic hotspots"—regions prone to recurrent large-scale microdeletions and microduplications associated with neurodevelopmental disorders (Figure 1, Supplementary Table 2) [34]. Spanning these HSD regions, we selected a tiling path of clones and subjected them to capillary (N = 203) or SMRT (N = 65) sequence and assembly (Methods), of which 211 resulted in complete high-quality BAC sequences (Supplementary Table 3). The alternate sequence assemblies, many of which have now been incorporated into the most recent human reference build (GRCh38; N = 85) allowed us to close 24 euchromatic gaps and correct large-scale errors in the human reference genome. From this, we identified smaller HSDs residing within these loci that

did not originally meet our size threshold. The new sequence allowed us to distinguish 28 duplication events ranging in size from 11 kbp to 677 kbp (mean size: minimum 176 kbp and maximum 234 kbp, median size: minimum 95 kbp and maximum 152 kbp) corresponding to 33 HSD gene families accounting for 80 paralogous "genes" (Supplementary Table 4).

The majority of events (N = 24 events or 3.2 Mbp) were primary duplications—defined here as the initiating SD from the ancestral locus shared between human and chimpanzee (Figure 2). Many of these duplications occur in the vicinity of flanking core duplicons (high-copy shared human–great ape duplications) [35,36], making it difficult to delineate the precise breakpoints, thus minimum and maximum sizes for these regions were calculated (Supplementary Table 4). We identified four secondary HSDs—additional duplications derived from a human-specific duplicate paralog. Although few in number, these secondary events account for 35% (1.7 Mbp) of HSD base pairs because the events are, on average, larger (437 kbp) when compared to primary duplications (minimum: 136 kbp; maximum: 196 kbp). The majority of HSDs are intrachromosomal and arranged in inverted orientation with respect to their ancestral paralogs (18/26, P = 0.014, binomial test), including all secondary duplications (4/4). Further, when we considered the minimum extent of breakpoints, we observed a significant difference in sizes between primary and secondary duplications (136 kbp vs. 437 kbp, P = 0.05, Wilcoxon-Mann-Whitney test).

HSDs appear to be distributed in a highly nonrandom fashion in the human genome. To further test this apparent clustering of HSDs near each other, we performed simulations using an unbiased set of duplications (18 of our 24 primary duplications) containing previously identified HSD genes [24]. Compared to a random null distribution (see Methods), we found that these primary HSDs map closer to each other than by chance (empirical median distance to nearest HSD 377 kbp, $P = 1x10^{-7}$; Supplementary Figure 2B). Likewise, HSDs were also significantly enriched near any non-primary SD (empirical median distance to nearest HSD 18 kbp, $P < 1x10^{-6}$; Supplementary Figure 2C). Noting that most HSDs appear to reside in the vicinity of a core duplicon [35], we also

found our primary HSDs to be significantly enriched near core duplicons (empirical median distance to a core 250 kbp, $P < 1x10^{-7}$; Supplementary Figure 5) consistent with recent studies of specific genomic hotspots regions [35,37,38].

This clustering is most pronounced on chromosome 1p12 to 1q32.1, which contains the greatest number of gene-containing HSDs (~2 Mbp or ~0.8% of the euchromatin on human chromosome 1). As such, this region has been a source of considerable error with respect to the human reference genome (GRCh37) and directed efforts have been undertaken using the haploid BAC library to correct and improve the reference genome [16,39]. By mapping the same genetically diverse panel of human and great ape genomes (described above) against this corrected build, our analysis found that 85% of the 8 Mbp has been duplicated in humans and great apes with only 1.15 Mbp remaining unique in humans (across regions chr1:119,989,248-121,395,939 and chr1:143,311,826-149,876,379, GRCh38; Supplementary Figure 6). This hotspot of duplication has been the target of at least six independent HSD events, including four primary duplications and two secondary duplications. Besides the primary duplications of *SRGAP2* [16] and *HYDIN* [40], the majority of events occurred intrachromosomally between chromosomal bands 1p11.2 to 1q21.2 and contain seven HSD gene families, including complete genic paralogs of *FAM72*, *FCGR1*, *HIST2H2*, and *GPR89* (which also show separate recurrent duplications in other great apes) and a partial duplication of *PDZK1*.

Evolutionary timing of HSDs

We estimated the evolutionary timing of each HSD event by constructing multiple sequence alignments using nonhuman primate (NHP) sequences as outgroups and a previously described molecular clock approach adjusting for alignments that failed a relative rate test [16]. Since many of the HSD donor regions were complex and not properly assembled in ape reference genomes, it was necessary to target NHP BAC libraries and sequence corresponding ancestral loci (N = 269 clones; Supplementary Table 5). We estimated an evolutionary time of each event based on human–chimpanzee branch length as well as in terms of millions of years ago assuming a divergence of 6 million years ago

(mya) [41-43] (Figure 2, Supplementary Table 6, Supplementary Figure 7, Supplementary Data).

Our results reveal significant differences in number and size of HSDs when we compare across three equal time periods during the evolution of the human lineage (P = 0.017, Kruskal-Wallis rank sum test of minimum sizes) (Figure 2). The first was a period of relative quiescence, which occurred after the human–chimpanzee divergence (~4.7-6 mya). This included five smaller primary duplications corresponding to seven genes with an average minimum size of 57 kbp for a total of 285 kbp. This was followed by an apparent burst of larger primary (N = 6) and secondary (N = 1) duplications between ~2.3 to 3.1 mya. This set of duplications was substantially larger containing 12 HSD genes (average minimum size of primary events 249 kbp for a total of 1.5 Mbp, P = 0.026, Wilcoxon-Mann-Whitney test). The final set of duplications involved more secondary (N = 3) and primary (N = 13) duplications and are estimated to have occurred in the last 1.9 million years. Although primary duplication lengths were not significantly different in size compared to either of the other two time periods (average minimum size 113 kbp for a total of 1.4 Mbp), they resulted in many more HSD genes (28 gene paralogs including primary (20 genes) and secondary (8 genes) events).

Human copy number diversity

We undertook three different approaches to assess the potential functional significance of HSDs; namely, copy number constraint, transcriptional potential and protein-coding mutations. We first assessed copy number in the human species in order to distinguish fixed duplications from those that are highly stratified among human populations. We determined the average copy number of 23 HSD units (i.e., regions where the same HSD gene families were always found together on duplicate paralogs) containing duplicated gene families across a diversity panel of humans (n = 2,379) [44,45], two archaic hominins (one Neanderthal [46] and a Denisovan [47]), three archaic humans (two representatives of Neolithic and Mesolithic populations [48] as well as an Ust'-Ishm individual, estimated to have lived 45 thousands years ago [49]) and 86 NHPs (Great Ape Genome Project [50]) (Figure 3, Table 1, Supplementary Tables 7 and 8). We identified the most copy number

polymorphic gene families, defined here as those showing the greatest copy number variance as measured by standard deviation (s.d.) in the human species. These included genes at the 7q35 locus (three units: *ARHGEF5* and *OR2A*, *TCAF1*, and *TCAF2*), the 5q13.1 locus (four units: *SMN1* and *SERF1*, *GTF2H2*, *OCLN*, and *NAIP*), 16p11.2 (two units: *BOLA2* and *DUSP22*), and 10q11.23 (one unit: *GPRIN2* and *NPY4R*). Conversely, eight HSD genes were largely fixed for copy number, showing the lowest variance among contemporary human populations (six units: *HYDIN*, *GPR89* and *PDZK1*, *CFC1* and *TISP43*, *CD8B*, *ROCK1*, and *ARHGAP11*; Figure 3, Supplementary Figure 8, Supplementary Tables 9 and 10).

As expected, HSD gene families with an overall higher copy number are generally more copy number polymorphic (Supplementary Figure 9A). For example, less copy number variable genes had lower overall copies (average diploid copy number 4.0, N = 13; Human Genome Diversity Project (HGDP) s.d. = 0.04-0.27) compared to genes with the greatest variance (average diploid CN of 5.1, N = 13; s.d. = 0.48-1.58). Notable exceptions are *GTF2I*, *GTF2IRD2*, and *NCF1* (diploid CN = 6; HGDP s.d. = 0.27), located within the Williams-Beuren syndrome region on chromosome 7q11.23, which shows little polymorphism across humans. Importantly, our analysis identified 11/23 duplicated units with at least one normal individual identified who carried the ancestral state copy number (diploid CN of two) suggesting the HSD paralogs are missing in these individuals (e.g., *DUSP22* and *ROCK1*; Figure 3B). Population differentiation (as measured by Vst [51]) generally correlated with copy number variance ($R^2 = 0.32$; $\rho = 0.54$, Pearson correlation; Supplementary Figure 9B) but not copy number ($R^2 = 0.01$; $\rho = -0.01$, Pearson correlation; Figure 3C). Some potential Vst outliers include *TCAF1* and *TCAF2* (HGDP mean Vst = 0.11), *OCLN* (HGDP mean Vst = 0.09), and *SMN1* and *SERF1* (HGDP mean Vst = 0.08). Alternatively, some polymorphic gene families, including *CHRNA7* (HGDP mean Vst = 0.018), exhibit very little population differentiation.

We leveraged singly unique nucleotide k-mers (SUNKs) to investigate copy number variation diversity at the level of individual paralogs [24] (Table 1, Supplementary Tables

11 and 12). Among the 72 gene paralogs assayed here, we identified homozygous deletions for 24 paralogs in at least one human individual. One interesting example was *CFC1*/*TISP43*, which we originally determined was "fixed" by our aggregate diploid copy number analysis (Table 1). SUNK analysis, however, revealed that nearly all non-African individuals carry only the B duplicate paralog and almost no copies of the A paralog (Supplementary Figure 10). 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-African populations. Further, initial genotyping implicated three ancestral paralogs as showing evidence of heterozygous deletions (Supplementary Figure 11); 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 [52] (Supplementary Figure 12). Meanwhile, though deletions appear to exist within portions of the HSD-containing *CHRNA7*, the gene itself appeared undeleted. These results further highlight certain limitations of this SUNK approach, especially when paralogs share high-sequence similarity [16,53].

We also identified three genes expanded uniquely in *Homo sapiens* when compared to two sequenced archaic hominins, a Neanderthal and a Denisovan, including the previously reported *BOLA2* on chromosome 16 [46,54] (Supplementary Figure 13), and two novel genes, TRPM8-associated *TCAF1* and *TCAF2* (formerly *FAM115A* and *FAM115C*) on chromosome 7 (Figure 4, Supplementary Table 13) [55]. In the case of *TCAF1*and *TCAF2*, the timing estimate (~0.19-0.75 mya) is consistent with this distribution. The fact that we observe high copy number in two archaic humans (Loschbour and Ust Ishim individuals with CN $\geq$ 6), suggests these HSDs spread rapidly in the population. Among modern humans copy number ranges from 2 to 7 and occurs over a ~131 kbp cassette (involving A, B, and C SDs). 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 non-allelic

homologous recombination (NAHR)-mediated deletion between directly oriented B1 and B2. The *TCAF1/TCAF2* HSD is differentiated between human populations with the highest copy number observed for African and European populations (in particular, Gambian and Esan from Nigeria where multiple individuals with CN 7 are observed) and the lowest copy number observed for Asian and Amerindian populations. Further, we identified 17 individuals (n = 2,367) where the copy number was consistent with the diploid 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.

Patterns of HSD mRNA expression

In order to assess HSD transcriptional potential, we leveraged the specificity of sequence differences between exonic paralogs and searched for evidence of spliced mRNA products. We specifically examined RNA-seq data from GTEx [56] and mapped the distribution of reads to HSDs in 45 different tissues across multiple individuals (Supplementary Table 14). We quantified relative levels of expression using an adjusted version of RPKM (reads per kilobase of transcript per million mapped reads intersecting unique genomic k-mers of a canonical isoform (RefSeq); Supplementary Table 15, Supplementary Figure 14) corresponding to each gene paralog (Figure 5A and B). Due to an insufficient number of k-mers to distinguish paralogs (i.e., *BOLA2*) or mistakes in the human reference genome causing only single paralogs to be represented (i.e., *DUSP22* and *GPRIN2*), total overall expression was instead calculated for three gene families (Figure 5C). 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*). Of the 26 comparisons that could be made between known ancestral and duplicate paralogs (Figure 5A), 73% (19/26) of duplicate paralogs showed significantly lower expression levels compared to their ancestral paralog (versus 12% (3/26) showing significantly greater and 15% (4/26) no difference). Although the chromosome 5q13 region has been the target of active gene conversion [57], all genes that we predict to represent derived paralogs, including *OCLN-B*,

NAIP-C, *SMN2*, and *SERF1B*, show greater expression. In a few gene families we detected very low expression for nearly all human tissues (i.e., *NYP4R*, *TISP43*, and *OR21A*), though we note this could be the result of a technical artefact due to small gene sizes resulting in few SUNKs. Human-specific *TCAF2C*, for example, shows little expression compared to its closely related paralogs (*TCAF2A* and *TCAF2B*). This is consistent with its gene structure that suggests it is likely incomplete and thus a non-processed pseudogene. In contrast, human-specific *FRMPD2B* and *CHRFAM7A* both show increased expression in specific tissues compared to their ancestral paralogs (*FRMPD2A* and *CHRNA7*). Both of the derived duplicates are incomplete, lacking the 5' portion when compared to the ancestral gene. It is possible that the altered expression results from co-opted new promoters or regulatory elements. *CHRFAM7A*, for example, is the product of a gene fusion of *FAM7A* and *CHRNA7* duplications and shows increased expression in the aorta, liver, lung, testis, and thyroid. *FRMPD2B* shows increased expression compared with *FRMPD2A* in several regions of the brain cortex as well as reproductive organs, including fallopian tubes and uterus.

Discovery of likely gene-disruptive events

Since pseudogenization is the most likely fate of duplicated genes, we tested whether paralogs had accumulated likely gene-disrupting (LGD) mutations by targeted sequencing of canonical protein-coding exons. We designed 1,105 molecular inversion probes (MIPs) to capture the 415 exonic regions of 30 gene families with coding exons and performed massively parallel barcoded Illumina sequencing of 658 individuals from the 1000 Genomes Project including European (n = 395) and African populations (n = 263). (Table 1, Supplementary Tables 16-18). 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). From these data we identified 96 LGD variants—which included frameshift, stop gain and loss, and splice donor and acceptor mutations—in 25 out of 30 gene families assayed.

A subset of LGD variants (n = 33) could be definitively associated with a specific paralog because of proximity to a paralog-specific variant (PSV) on the same sequence read (Table 1, Supplementary Table 18). We identified 10 duplicate paralogs and no ancestral paralogs harboring common loss-of-function mutations (population frequency >5%) suggesting these genes may not be under strong functional constraint, including *ARHGAP11B*, an HSD previously implicated in neuronal migration [17]; closer inspection shows the variant falling within an *ARHGAP11B*-specific intron due to alternative splicing of the paralog compared to its ancestral counterpart (Supplementary Figure 15). The remaining LGD variants (n = 63) could not be unambiguously assigned but typically fell into gene families with more than one duplicate paralog. Overall, we identified no LGD variants in four gene families and discovered only rare LGD variants (<5% population frequency) in an additional 15 gene families (Table 1). Notably *SRGAP2C* [15,16], another duplicate paralog previously shown to contribute to human-specific neurological traits, was included in these 19 gene families that exhibit a paucity of LGD variants.

Using these same assays, we compared the burden of mutation in a case versus control design for neurodevelopmental disease. We sequenced and compared the LGD frequency in 3,444 children with autism and 2,617 unaffected siblings from the Simons Simplex Collection (SSC) [58], Autism Genetic Resource Exchange (AGRE) [59], and The Autism Simplex Collection (TASC) [60] cohorts. In total, we targeted an additional 6,061 individuals for sequencing (~261X average coverage (N = 1,096 MIPs) with 1,058 MIPs with >10-fold coverage). From this set, we identified 4,069 total coding and splice variants, of which 247 were LGD, in both cases and controls for 30 genes (Supplementary Tables 18-20). The majority of LGD variants (N = 231) were considered rare and collectively found in equal proportions of cases and controls (24% of individuals). Examining burden of rare LGD variants of individual genes, we identified a nominal enrichment in cases versus controls of *GPR89* with seven variants exhibiting an overall frequency of 19/3,430 cases versus 5/2,605 controls ($p_{uncorr} = 0.02$, Fisher's exact one-sided test), though this result did not pass Bonferroni multiple-testing correction (Supplementary Figure 16).

Common LGD variants existed in 11/30 genes, with six genes (*FCGR1*, *GTF2I*, *GTF2IRD2*, *HIST2H2BF*, *HYDIN*, and *ROCK1*) carrying fixed LGD variants in nearly all individuals tested. Notably, five of these genes represent partial duplications where we might expect a greater likelihood of disruptive mutations if the paralogs represented pseudogenes. Combining these results with our assessment of 1000 Genomes Project individuals, 16/30 gene families showed an absence of common protein-disrupting variants in the 6,719 humans tested (Table 1).

The complexity of HSD evolutionary history

To highlight the complex evolutionary history associated with such regions, we selected three loci for further investigation. Two are known to be important in human disease (chromosomes 7q11.23 and 10q11.23) because the SDs promote NAHR resulting in deletions associated with neurodevelopmental delay. The other region shows duplications unique to modern humans corresponding to potential novel *Homo sapiens*-specific paralogs of *TCAF1* and *TCAF2*. In order to reconstruct the breakpoints and the likely order of events, we constructed tiling paths using our BACs sequenced in NHPs, including chimpanzee, gorilla, and orangutan (N = 196; Supplementary Table 5) and previously sequenced clones (N = 35) across the complete extent of each region.

Large deletions ~1.8 Mbp in size of the chromosome 7q11.23 region lead to Williams-Beuren syndrome (OMIM #194050) and reciprocal duplications are associated with autism and intellectual disability [61]. The directly oriented flanking human-specific SDs (termed B; Figure 6A) contain three genes—*GTF2I*, *GTF2IRD2*, and *NCF1*. Comparative analysis with great ape genomes (Supplementary Figure 17) predicts that the most common human haplotype present today arose through a three-step evolutionary process. The first two events occurred within the distal breakpoint of the region (Supplementary Figure 18). They involved an inverted duplication of ~116 kbp SD (termed A, containing paralogs of the high-copy duplicon (*SPDYE*)) and a possible 90 kbp inversion, ~1.7-2.1 mya (0.318 ± 0.029 human–chimpanzee distance) followed by a separate ~106 kbp inverted duplication of B around 1.2-1.4 mya (0.212 ± 0.018 human–

chimpanzee distance). These events created truncated paralogs *GTF2IB* and *GTF2IRD2B* and a full-length version of *NCF1B*. A third large-scale inverted duplication transposed an ~395 kbp region comprised of SDs A, B, and C (containing *POM121L*) from the distal to proximal breakpoints of the disease-associated region less than 1 mya (0.118 ± 0.014 human–chimpanzee distance). This tertiary duplication established "granddaughter" truncated copies of *GTF2IRDC*, *GTF2IC* as well as a full-length paralog *NCF1C*. The event also appears to have overwrote the 3' end of the ancestral *POM121* with *POM121L*. This final event created the susceptible genomic hotspot configuration, with directly oriented SDs A and B, providing a substrate for NAHR leading to disease-associated copy number variants. Notably, our sequence analysis of other great apes (Supplementary Figure 17) matched nearly perfectly the deduced genomic configuration hypothesized previously by [62], with the exception of a large-scale inversion of the region proximal to BP1 in orangutan.

Likewise, at the chromosome 10q11.21 locus, large-scale deletions and duplications have been identified in children with developmental delay with variable expressivity and penetrance [63,64]. HSD genes *FRMPD2*, *PTPN20*, *GPRIN2*, and *NPY4R* reside within two separate SDs proximal to the disease-associated region (Figure 6B). 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* ~2.4 mya (0.396 ± 0.033 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 ~2.2 mya (0.372 ± 0.006 human–chimpanzee distance), potentially concurrently with the previous event. Comparing this region in human and chimpanzee identified an additional 550 kbp inversion that included the *GPRIN2*/*NPY4R* SD and adjoining proximal region. The precise evolutionary history of this inversion could not be deduced because we were unable to identify BAC clones corresponding to this segment in either gorilla or orangutan (Supplementary Figure 19).

Finally, we characterized one of the youngest HSD regions unique to modern humans on chromosome 7q35 containing *TCAF1* and *TCAF2* and primate-duplicated *CTAGE6* [65]. We note that expansion of a *CTAGE*-paralog also occurred in the duplication of HSD gene *ARHGEF5*, located less than 500 kbp distal to this locus. Pairwise comparisons between human and chimpanzee suggest the possibility of three distinct duplication events (A: 65 kbp, B: 10 kbp, and C: 56 kbp) as well as a large-scale inversion (~200 kbp). We estimate an initial 10 kbp inverted duplication of SD B containing the 3' end of *TCAF2A* ~1.4 to 1.9 mya (0.091 ± 0.008 human–chimpanzee distance) creating a truncated *TCAF2B*. The subsequent events occurred very recently during human evolution, 0.24 to 0.6 mya, potentially during or after the split from a common ancestor of Denisova and Neanderthal (Figure 6C). These subsequent rearrangements created a new full-length paralog of *CTAGE6* (contained in A) and truncated paralogs *TCAF1A* (the putative ancestral paralog contained in C1) and *TCAF2C* (contained in C2). Notably, we estimate the full-length and functional *TCAF1B* and *TCAF2A* now reside on distinct SD paralogs that are separated by 130 kbp transcribed on opposite strands—as opposed to the ancestral configuration where the genes are tandem, adjacent, and transcribed on the same strand. Our BAC-based targeted sequencing of the locus not only eliminated the gap in the sequence but also dramatically reorganized the structure of the region removing incorrectly assigned paralogous sequence, including ~29 kbp of extra sequence (Supplementary Figure 20). 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.

DISCUSSION

In this study we generated new reference sequence for some of the most complex and gap-ridden sequence of the human genome. Over the course of this work we generated 44.5 Mbp of high-quality sequence from 211 BACs derived from a CHM1 hydatidiform mole BAC clone resource (CH17). The haploid origin of the human genome facilitated the resolution of high-identity paralogs and allowed us to rapidly assemble these regions without being confounded by the allelic structural variation that has complicated previous

genome assemblies. 18.2 Mbp of these data have already been incorporated into new human genome reference (GRCh38) closing 24 large euchromatic gaps and in many cases completely revising the structure of the region. In addition to providing a better substrate for gene annotation, the improved references are key to defining rearrangement breakpoints [37,53] since most of the regions are associated with recurrent microdeletions leading to neurodevelopmental disorders such as Prader-Willi (OMIM #176270) and Williams-Beuren syndromes (OMIM #194050).

Several important features emerge from our targeted sequencing (48.4 Mbp) of HSD ancestral and target sites of integration in NHPs over the course of human evolution. The largest HSDs are significantly clustered in the human genome and near core duplicons with chromosomes 1q21, 5q13 and 7q11.3 showing the greatest density of independent HSD events (N = 18). Detailed phylogenetic reconstruction of several regions indicate that most regions have been subjected to multiple large structural variation events during human evolution with inverted duplications being the predominant mode of structural change (71.4% of the total predicted 28 intrachromosomal duplication events, P = 0.006, binomial test; Supplementary Table 4). Inverted SDs have been noted before in complex structural rearrangements associated with genomic disorders such as Pelizaeus-Merzbacher disease [66,67] and Smith-Magenis Syndrome [68] and may be a product of replication-based mechanisms such as fork-stalling and template switching (FoSTeS) [66] and/or microhomology-mediated break-induced repair (MMBIR) [69]. It is possible that the complex genomic architecture associated with core duplicons residing in proximity to HSDs acted to perturb the replisome leading to FoSTeS or represented sites of fragility predisposing to MMBIR. The enrichment of inverted SDs also emphasizes the intimate association between inversions and the dispersal of SDs [37,38,70]. The association of inversion breakpoints with SDs has contributed to this type of variation being underestimated in studies of human genomes [45,71,72].

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 [16]. Nevertheless, unlike primary duplications, secondary duplications are always distributed intrachromosomally in inverted orientation with the majority mapping within 5 Mbp of their progenitor. We infer from our timing estimates that HSD activity was relatively quiescent during the first three million years after divergence from chimpanzee. Events during this time are almost exclusively restricted to primary duplication events. Over the last three million years of human evolution, we predict that the tempo of primary and secondary duplications increased with two noticeable waves between 2-3 mya and another set in the last two million years. It is intriguing that the first wave of HSDs and the associated structural changes in the human genome occur at a time when the genus *Homo* is thought to have diverged from that of Australopithecine-like precursors [73,74].

Comparisons with the genomes of Neanderthal and Denisova also 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/TCAF2,* as well as the recently characterized *BOLA2*[46,54] gene family [54], which shows some of the most significant expression differences between human and chimpanzee induced pluripotent stem cells [75] and whose segmental duplication mediates large-scale copy-number variation associated with 1% of autism cases [76,77]. 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 [55]. 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 in our lineage [54].

Although our estimates provide an evolutionary framework for the timing of HSDs, there are two important caveats. First, additional mutation events, such as interlocus gene

conversion, frequently occur between high-identity paralogs [53,78,79]. Such duplications will make HSDs appear evolutionarily "younger" than they are because the original molecular signatures, including divergence, have been erased by gene conversion. Nevertheless, we note that the most recent timing estimates of *BOLA2* (~0.3 mya) [54] and *TCAF1* and *TCAF2* (~0.3-0.5 mya) are consistent with their absence in ancient hominin genomes such as Neanderthal and Denisova, which diverged from humans ~600-650 thousand years ago [46]. The second caveat is that the full extent of HSDs is often difficult to assess because they frequently occur in duplication blocks where there have been multiple rounds of structural variation over the last 15 million years. Breakpoints and boundaries become challenging to delineate due to a series of overlapping complex rearrangements. For example, when we compared our new chimpanzee reference for chromosome 5q13.3 containing *SMN1* [80,81], 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.

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 [15,16,82], while the HSD gene *ARHGAP11B* appears to promote apical basal radial progenitor amplification in the subventricular zone [17]. 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. As a group, ancestral HSD genes are enriched for neurological functions [24]

and are disproportionately expressed in the developing neocortex and prefrontal cortex when compared to older duplicated genes [83].

In this study, we used human diversity and transcriptional potential to enrich for potential functional paralogs among the 35 HSD gene families—composed of 33 ancestral and 47 human-specific paralogous genes. We applied three criteria: (1) all humans must carry the duplicate paralog based on copy number polymorphism data from 2,367 sequenced human genomes (i.e., no controls have returned to the chimpanzee ancestral state); (2) no common truncating mutations, including frameshift, splice site or nonsense events, exist for a gene family based on targeted sequencing of 3,262 controls; and (3) duplicates show evidence of spliced mRNA expression in at least one of 45 human tissues surveyed using the GTEx RNA-seq resource. 10 HSD gene families met all criteria, making them top candidates for further functional investigation. This list included two genes previously implicated in cortical development, *ARHGAP11B* [17] and *SRGAP2C* [15,16], as well as the gene families *BOLA2*, *CD8B*, *CFC1*, *FAM72*, *GPR89*, *GPRIN2*, *NPY4R*, and *TISP43*. Though, the latter two genes show low expression across nearly all tissues from GTEx dataset (Figure 5), their small sizes could lead to potential false negatives due to the paucity of SUNKs representative in their exons. *GPRIN2* (G protein-regulated inducer of neurite outgrowth 2) has been shown to interact directly with G-coupled proteins (GNAO1 and GNAZ) [84] and has been implicated in the control of neurite outgrowth [85]. Our RNA-seq analysis points to localized expression in various regions of the brain, including the cerebellum and hypothalamus (Figure 5), confirming Northern analyses (OMIM #611240). Other genes of interest include *CFC1* (cripto, FRL-1, cryptic family 1), which encodes a member of the epidermal growth factor important in patterning the left-right embryonic axis [86], and *NPYR4* (neuropeptide Y receptor Y4)—previously known as *PPYR1* (pancreatic polypeptide receptor 1), which has been implicated in energy homeostasis. Large copy number variants of the corresponding region on chromosome 10q11.23 are associated with obesity [87]. *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, was previously thought to have duplicated in a common ancestor of human, gorilla and

chimpanzee [88]. Our analysis shows a gorilla-specific 40 kbp duplication of *CD8B* exon 1 not shared with human, chimpanzee, bonobo or orangutan (Supplementary Figure 22). 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 ~5.2 mya, supports this finding (Figure 2).

While this work provides a useful starting point for further investigation, there are some important limitations. First, our assessment enriches in paralogs that are more likely to encode proteins and does not consider the possibility of functional noncoding RNA. Notably, three of the annotated genes (*MIR4435*, *MIR4267*, and *OR2A)* mapping to HSDs are identified as noncoding RNA (Ensemble Variant Effect Finder), thus, disruptions in the open reading frame have no meaning in this context. Second, the canonical gene model being investigated is heavily weighted by the ancestral intron-exon structure. Thus, novel fusion genes and transcripts not previously annotated that have gained alternate promoters would not have been considered in this analysis [54]. Third, higher copy gene families and paralogs with higher sequence identity become more difficult to discern using paralogous sequence variants. It is likely that long-read genome and transcriptome data will be required to explore these particular paralogs [25]. Nevertheless, in cases where these data were informative, we were able to rule out 28 paralogs by identification of homozygous deletions or a common LGD variant in population controls. In some cases, such as *OCLN*, we predicted the loss of ancestral paralogs in a few individuals (Supplementary Figures 11 and 12). Since such events are thought to result in disease [52], it is possible that interlocus gene conversion has occurred between the ancestral and duplicate copies complicating our analysis.

Finally, although we focused on HSDs that had become fixed in the human population, it may be that some of the most copy number polymorphic loci are candidates for more recent adaptations between populations [44]. In this regard, duplications of *TCAF1*/*TCAF2* are particularly intriguing. The genes encode TRP channel-associated factors that bind to *TRPM8*—the primary detector of environmental cold [89,90] expressed in 10-15% of somatosensory neurons. The two TCAF proteins are thought to exert opposing effects in

TRPM8 gating and insertion into the plasma membrane [55]. Our copy-number analysis agrees with our evolutionary finding that duplications of this locus are *Homo sapiens*-specific – not existing in Neanderthal and Denisova but at high copy in archaic humans. In modern humans, African and European populations show the greatest copy numbers while Asians show the lowest with some humans showing no duplication of the region (Figure 4). Our evolutionary model suggests that a single full-length paralog of *TCAF1B* (predicted HSD duplicate paralog) and *TCAF2A* (predicted ancestral paralog) exist at the locus, respectively, while additional *TCAF1/TCAF2* copies appear to be truncated or incomplete. It is interesting to note that the conserved function of full-length TCAF2 may have been co-opted by a duplicate paralog after truncation of the ancestral paralog, a mechanism we also suggest occurred for duplicate family *PTPN20* (Figure 6B). Although the function of the truncated duplicates await further characterization, it is clear that this locus has been radically restructured in most humans resulting in the ancestral functional loci being separated by hundreds of kilobase pairs and being transcribed in opposite orientations. It is likely that regulatory changes occurred as a result of duplication and restructuring of this locus.

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 [44] and a set of 86 NHP genomes [50] were used to estimate aggregate copy number in 500 bp windows using previously described methods [24]. Windows with >90% of human genomes at copy number >2.5 and >90% of NHP genomes at copy number <3 were identified as HSD. 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 [35,36] (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 >2.5.

Validation of HSD regions

We intersected HSD regions with SDs identified using WSSD [32] and WGAC [31] 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. 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 [38]. Sequencing data (~300-fold coverage) were mapped with mrFAST [91] to the human reference genome (GRCh37) and SUNKs were used to discriminate between highly identical SDs [24]. PacBio SMRTbell libraries were prepared and sequenced using 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 as described [30]. BACs were assembled into contigs with PacBio- and capillary-sequenced clones using Sequencher and compared to the human reference genome using Miropeats [92] and BLAST [93]. 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, that we included in breakpoint and evolutionary analyses (Supplementary Tables 3 and 5). These clones were not included in counts of BACs sequenced for this study.

Breakpoint identification

A combination of BLAST [93], BLAT [94], and WGAC [31] 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 [92] 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 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.

HSD clustering simulations

We simulated a null distribution by shuffling 218 HSD regions (defined in Supplementary Table 1; GRCh37) and 18 primary HSDs (derived genomic coordinates only, Supplementary Table 4; 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 [31] 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 [35]. 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) [95]. 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 orthologs of HSDs 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 were generated using MAFFT and included the maximal shared genomic regions of human paralogs and nonhuman orthologs excluding any flanking core duplicons [96]. Alignments were visualized for manual editing using Jalview [97]. Phylogenetic analyses were performed using MEGA6 [98]. If a full-length alignment did not pass the Tajima's D relative rate test [99], using orangutan as the outgroup, pairwise sequence identities were calculated across 500 bp sliding windows with 100 bp increments using the PopGenome statistical package [100] 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 6). 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 [16]. 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.

Copy number genotyping

Raw sequences from 236 human individuals from HGDP [44], 2,143 human individuals through Phase 3 of the 1000 Genomes Project [45], 86 NHP individuals from the Great Ape Genome Project [including bonobo (N = 14), chimpanzee (N = 23), gorilla (N = 32), and orangutan (N = 17)] [50], a Denisovan individual [47], a Neanderthal individual [46], and three archaic hominids [48,49] were mapped to the human reference genome (GRCh37) using mrsFAST [101]. 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 [24] and visualized as heatmaps using bigBed tracks within the UCSC Genome Browser. Using these same data, we genotyped average copy number across HSD-defined regions (Supplementary Table 7). 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 used the Vst statistic [51] 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 three human HapMap individuals: GM18956, GM18507 and GM12878 (Coriell Cell Repository, Camden, NJ). FISH experiments were performed using the fosmid clone

WIBR2-2930I21 directly labeled by nick-translation with Cy3-dUTP (Perkin-Elmer), as described previously [102], 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, if the normalized value was higher than zero. Finally, for each subtissue and each transcript, a median value was calculated over all the unique 30-mers found in this transcript. 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 16th, 2015) [103] (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 [104]. Each MIP was designed to capture 112 bp of genomic sequence and included 40 bp unique to the target 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 17). 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) and only sequenced the 1000 Genomes Project control cohort. MIP phosphorylation, capture, and barcoding were performed as previously described [105]. 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 37°C for 18 hours. Finished libraries were pooled together and sequenced using either MiSeq (2 x 150 bp) or HiSeq2000 (2 x 101 bp).

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 [64]. The resulting variant set was annotated utilizing the Ensembl Variant Effect Predictor (VEP) [106] 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. 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 [107,108]. The severity of missense mutations was predicted using the Combined Annotation Dependent Depletion

(CADD) score for all genic variants pre-computed for human reference GRCh37 except from *SRGAP2*, in which variants were annotated in GRCh38 [109].

FIGURE LEGENDS

Figure 1. Identification of human-specific segmental duplications or HSDs. (A) The locations of large, gene-containing HSDs are highlighted (blue lines) with 80 individual gene paralogs from 33 gene families listed across 9 different human autosomal chromosomes. Included in this set are paralogs of *GPR89*, which duplicated in other great apes but experienced human-specific expansions. Many of these HSDs overlap known disease-associated genomic hotspots (red lines) prone to recurrent copy number variation associated with developmental delay. ID: intellectual disability. **(B)** Duplicated regions were detected based on read-depth analysis of Illumina reads mapped to the human reference genome (GRC37). The set included a diversity panel of humans (Human Genome Diversity Project or HGDP (N = 236) [44]) and nonhuman primates or NHPs (gorillas (N = 32), chimpanzees (N = 23), bonobos (N = 14), and orangutans (N = 17) [50]). Overall copy number (CN) was averaged across 500 bp sliding windows and depicted as colored heat maps (see pictured index). Any genomic region >5 kbp shown to have diploid CN ≥ 3 in 90% of humans tested compared to all NHPs was considered an HSD.

Figure 2. Timing of HSDs. (A) Duplication timing estimates are plotted as a ratio of human–chimpanzee divergence (x-axis). The total estimated size of primary (black) and secondary (orange) duplications is shown for each event. Uncertainty in the size of each event is due to breakpoints mapping in high-identity flanking duplications (gray). Primary duplications are colored as black bars and secondary duplications are colored as orange bars. **(B)** Generally accepted phylogeny indicating timing of each event assuming chimpanzee and human lineage divergence of 6 million years ago (mya). The analysis is based on high-quality sequencing, assembly and alignment of large-insert clones (human N = 211; NHP N = 269). Asterisks indicate adjusted timing estimates because of failed Tajima's D Relative Rate (*) and genes with evidence of gene conversion (**).

Figure 3. 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 (see Supplementary Table 5) in human populations, including the HGDP (N = 236; GRCh38) and 1000 Genomes Project (1KG, N = 2,143; GRCh37) cohorts,

NHPs, archaic humans, and a Denisovan and Neanderthal. From these results, the mean, standard deviation, Vst, and number of individuals with CN 2 indicating no duplicate paralogs exist were calculated for average CN of each duplicated gene family (Supplementary Table 8). **(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 FISH 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 1000 Genomes Project cohort were genotyped as homozygously deleted for *ARHGAP11B*, but FISH analyses showed CN = 4 for all three individuals (Supplementary Figure 8). The ancestral *ROCK1A* resides on at human chromosome 18q11 while the human-specific *ROCK1B* shows polymorphism at the 18p11.32, with one individual homozygously deleted for the paralog. **(C)** For each gene family, plots are shown for the average CN vs. average Vst across all HGDP individuals with duplicate gene family names indicated next to each data point. Red data points indicate genes with no homozygous deletions in any human tested.

Figure 4. Copy number polymorphism across diverse populations of *TCAF1* and *TCAF2* HSDs. (A) Heatmap of overall CN of *TCAF1* and *TCAF2* HSD region on human chromosome 7 with predicted gene models and segmental duplications (SDs; depicted as colored arrows) pictured above. Representative modern humans are represented for each genotyped CN across the locus with a single person (*HGDP00798) showing deletion of

the region, likely due to non-allelic homologous recombination between directly oriented SDs B1 and B2. **(B)** A scatter plot of *TCAF1* and *TCAF2* SDs (A1, B1, and C1) overall CN of individuals from modern human (HGDP cohort), archaic humans, a Denisova and a Neanderthal, and NHPs (chimpanzee, bonobo, gorilla, and orangutan) plotted on each axis. The one Western European individual circled in red that deviates from the rest of the individuals' copy numbers is the deletion carrier pictured in **A**. **(C)** CN predictions across modern humans from the 1000 Genomes Project and HGDP (N = 2,379), archaic hominins, and NHPs were made across a representative region (chr7:143533137-143571789; GRCh37).

Figure 5. Gene expression analysis of duplicated genes using a k-mer approach. Expression analysis using the GTEx RNA-seq dataset [56]. 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 (see Methods; Supplementary Table 20). 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 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.

Figure 6. Complex models of HSD evolutionary history. BACs tiling across human chromosome 7q11, 10q11, and 7q35 regions were sequenced and assembled for human and additional great apes and supercontigs were created. SD organization is depicted as

colored arrows across the **(A)** 7q11 (as previously defined [62]), **(B)** 10q11, and **(C)** 7q35 regions. 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 (threshold s = 500 for A and B, s = 1000 for C) 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 [32]; DupMasker [110]; and genes.

ACCESSION NUMBERS

Accession numbers for all sequenced BACs included in this study can be found in Supplementary Tables 3 and 5.

ACKNOWLEDGEMENTS

We would like to thank many individuals that contributed to the results described here. We thank B. Coe for assistance in statistical analyses, T. Brown for manuscript editing, and L. Vives, T. Wang, and B. Xiong for technical assistance MIP sequencing. We would also like to thank B. Dumont, C. Campbell, K. Meltz Steinberg, S. Girirajan, C. Payan, C. Alkan, and E. Karakoc for helpful discussion and advice. For DNA samples used in MIP sequencing, we would like to thank the investigators and families participating in the 1000 Genomes Project, Autism Speaks, the Autism Simplex Collection, and Simons Simplex Collection (SSC). Additionally, we would like to thank the principal investigators involved in the SSC (A. Beaudet, R. Bernier, J. Constantino, E. Cook, E. Fombonne, D. Geschwind, R. Goin-Kochel, E. Hanson, D. Grice, A. Klin, D. Ledbetter, C. Lord, C. Martin, D. Martin, R. Maxim, J. Miles, O. Ousley, K. Pelphrey, B. Peterson, J. Piggot, C. Saulnier, M. State, W. Stone, J. Sutcliffe, C. Walsh, Z. Warren, E. Wijsman). Approved researchers can obtain the SSC population dataset described in this study (https://sfari.org/resources/autism-cohorts/simons-simplex-collection) by applying at https://base.sfari.org. The BAC clones from the complete hydatidiform mole were

derived from a cell line created by U. Surti. This work was supported, in part, by U.S. National Institutes of Health (NIH) grants from NINDS (R00NS083627, M.Y.D.) and NHGRI (R01HG002385 and P01HG004120, E.E.E. and U41HG007635 to R.K.W. and E.E.E.) as well as The Paul G. Allen Family Foundation (11631 to E.E.E.). S.C. is supported by a National Health and Medical Research Council (NHMRC) CJ Martin Biomedical Fellowship (#1073726). E.E.E. is an investigator of the Howard Hughes Medical Institute.

REFERENCES

1 Enard, W. *et al.* A humanized version of Foxp2 affects cortico-basal ganglia circuits in mice. *Cell* **137**, 961-971, doi:10.1016/j.cell.2009.03.041 (2009).

2 Enard, W. *et al.* Molecular evolution of FOXP2, a gene involved in speech and language. *Nature* **418**, 869-872, doi:10.1038/nature01025 (2002).

3 Pollard, K. S. *et al.* An RNA gene expressed during cortical development evolved rapidly in humans. *Nature* **443**, 167-172, doi:10.1038/nature05113 (2006).

4 Dumas, L. *et al.* Gene copy number variation spanning 60 million years of human and primate evolution. *Genome research* **17**, 1266-1277, doi:10.1101/gr.6557307 (2007).

5 Prabhakar, S. *et al.* Human-specific gain of function in a developmental enhancer. *Science* **321**, 1346-1350, doi:10.1126/science.1159974 (2008).

6 Stedman, H. H. *et al.* Myosin gene mutation correlates with anatomical changes in the human lineage. *Nature* **428**, 415-418, doi:10.1038/nature02358 (2004).

7 McLean, C. Y. *et al.* Human-specific loss of regulatory DNA and the evolution of human-specific traits. *Nature* **471**, 216-219, doi:10.1038/nature09774 (2011).

8 Gallego Romero, I. *et al.* A panel of induced pluripotent stem cells from chimpanzees: a resource for comparative functional genomics. *Elife* **4**, e07103, doi:10.7554/eLife.07103 (2015).

9 Khan, Z. *et al.* Primate transcript and protein expression levels evolve under compensatory selection pressures. *Science* **342**, 1100-1104, doi:10.1126/science.1242379 (2013).

10 Prescott, S. L. *et al.* Enhancer divergence and cis-regulatory evolution in the human and chimp neural crest. *Cell* **163**, 68-83, doi:10.1016/j.cell.2015.08.036 (2015).

11 Vermunt, M. W. *et al.* Epigenomic annotation of gene regulatory alterations during evolution of the primate brain. *Nat Neurosci* **19**, 494-503, doi:10.1038/nn.4229 (2016).

12 Eichler, E. E., Clark, R. A. & She, X. An assessment of the sequence gaps: unfinished business in a finished human genome. *Nat Rev Genet* **5**, 345-354, doi:10.1038/nrg1322 (2004).

13 Ohno, S. *Evolution by gene duplication*. (Allen & Unwin; Springer-Verlag, 1970).

14 Boyd, J. L. *et al.* Human-chimpanzee differences in a FZD8 enhancer alter cell-cycle dynamics in the developing neocortex. *Curr Biol* **25**, 772-779, doi:10.1016/j.cub.2015.01.041 (2015).

15 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).

16 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).

17 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).

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

19 Sudmant, P. H. *et al.* Evolution and diversity of copy number variation in the great ape lineage. *Genome research* **23**, 1373-1382, doi:10.1101/gr.158543.113 (2013).

20 Bailey, J. A. & Eichler, E. E. Primate segmental duplications: crucibles of evolution, diversity and disease. *Nat Rev Genet* **7**, 552-564, doi:10.1038/nrg1895 (2006).

21 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).

22 Locke, D. P. *et al.* Large-scale variation among human and great ape genomes determined by array comparative genomic hybridization. *Genome research* **13**, 347-357, doi:10.1101/gr.1003303 (2003).

23 Cheng, Z. *et al.* A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature* **437**, 88-93, doi:10.1038/nature04000 (2005).

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

25 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).

26 Eichler, E. E. Segmental duplications: what's missing, misassigned, and misassembled--and should we care? *Genome research* **11**, 653-656, doi:10.1101/gr.188901 (2001).

27 Fan, J. B. *et al.* Paternal origins of complete hydatidiform moles proven by whole genome single-nucleotide polymorphism haplotyping. *Genomics* **79**, 58-62, doi:10.1006/geno.2001.6676 (2002).

28 Kajii, T. & Ohama, K. Androgenetic origin of hydatidiform mole. *Nature* **268**, 633-634 (1977).

29 Destouni, A. *et al.* Zygotes segregate entire parental genomes in distinct blastomere lineages causing cleavage-stage chimerism and mixoploidy. *Genome research* **26**, 567-578, doi:10.1101/gr.200527.115 (2016).

30 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).

31 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).

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

33 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).

34 Itsara, A. *et al.* Population analysis of large copy number variants and hotspots of human genetic disease. *American journal of human genetics* **84**, 148-161, doi:10.1016/j.ajhg.2008.12.014 (2009).

35 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).

36 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).

37 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).

38 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).

39 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).

40 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).

41 Brunet, M. *et al.* New material of the earliest hominid from the Upper Miocene of Chad. *Nature* **434**, 752-755, doi:10.1038/nature03392 (2005).

42 Brunet, M. *et al.* A new hominid from the Upper Miocene of Chad, Central Africa. *Nature* **418**, 145-151, doi:10.1038/nature00879 (2002).

43 Vignaud, P. *et al.* Geology and palaeontology of the Upper Miocene Toros-Menalla hominid locality, Chad. *Nature* **418**, 152-155, doi:10.1038/nature00880 (2002).

44 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).

45 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).

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

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

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

49 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).

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

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

52 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).

53 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).

54 Nuttle, X. *et al.* Emergence of a Homo sapiens-specific gene family and chromosome 16p11.2 CNV susceptibility. *Nature* (2016).

55 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).

56 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).

57 Burghes, A. H. When is a deletion not a deletion? When it is converted. *American journal of human genetics* **61**, 9-15, doi:10.1086/513913 (1997).

58 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).

59 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).

60 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).

61 Sanders, S. J. *et al.* Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron* **70**, 863-885, doi:10.1016/j.neuron.2011.05.002 (2011).

62 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).

63 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).

64 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).

65 Zhang, Q. & Su, B. Evolutionary origin and human-specific expansion of a cancer/testis antigen gene family. *Molecular biology and evolution* **31**, 2365-2375, doi:10.1093/molbev/msu188 (2014).

66 Lee, J. A., Carvalho, C. M. & Lupski, J. R. A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell* **131**, 1235-1247, doi:10.1016/j.cell.2007.11.037 (2007).

67 Carvalho, C. M. *et al.* Inverted genomic segments and complex triplication rearrangements are mediated by inverted repeats in the human genome. *Nature genetics* **43**, 1074-1081, doi:10.1038/ng.944 (2011).

68 Park, S. S. *et al.* Structure and evolution of the Smith-Magenis syndrome repeat gene clusters, SMS-REPs. *Genome research* **12**, 729-738, doi:10.1101/gr.82802 (2002).

69 Hastings, P. J., Ira, G. & Lupski, J. R. A microhomology-mediated break-induced replication model for the origin of human copy number variation. *PLoS Genet* **5**, e1000327, doi:10.1371/journal.pgen.1000327 (2009).

70 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).

71 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).

72 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).

73 Kimbel, W. H. *et al.* Late PlioceneHomoand Oldowan Tools from the Hadar Formation (Kada Hadar Member), Ethiopia. *J Hum Evol* **31**, 549-561, doi:10.1006/jhev.1996.0079 (1996).

74 Villmoare, B. *et al.* Paleoanthropology. Early Homo at 2.8 Ma from Ledi-Geraru, Afar, Ethiopia. *Science* **347**, 1352-1355, doi:10.1126/science.aaa1343 (2015).

75 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).

76 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).

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

78 Dumont, B. L. Interlocus gene conversion explains at least 2.7% of single nucleotide variants in human segmental duplications. *BMC Genomics* **16**, 456, doi:10.1186/s12864-015-1681-3 (2015).

79 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).

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

81 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).

82 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).

83 Zhang, Y. E., Landback, P., Vibranovski, M. D. & Long, M. Accelerated recruitment of new brain development genes into the human genome. *PLoS Biol* **9**, e1001179, doi:10.1371/journal.pbio.1001179 (2011).

84 Iida, N. & Kozasa, T. Identification and biochemical analysis of GRIN1 and GRIN2. *Methods Enzymol* **390**, 475-483, doi:10.1016/S0076-6879(04)90029-8 (2004).

85 Chen, L. T., Gilman, A. G. & Kozasa, T. A candidate target for G protein action in brain. *J Biol Chem* **274**, 26931-26938 (1999).

86 Bamford, R. N. *et al.* Loss-of-function mutations in the EGF-CFC gene CFC1 are associated with human left-right laterality defects. *Nature genetics* **26**, 365-369, doi:10.1038/81695 (2000).

87 Sha, B. Y. *et al.* Genome-wide association study suggested copy number variation may be associated with body mass index in the Chinese population. *J Hum Genet* **54**, 199-202, doi:10.1038/jhg.2009.10 (2009).

88 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).

89 Colburn, R. W. *et al.* Attenuated cold sensitivity in TRPM8 null mice. *Neuron* **54**, 379-386, doi:10.1016/j.neuron.2007.04.017 (2007).

90 Bautista, D. M. *et al.* The menthol receptor TRPM8 is the principal detector of environmental cold. *Nature* **448**, 204-208, doi:10.1038/nature05910 (2007).

91 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).

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

93 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).

94 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).

95 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).

96 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).

97 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).

98 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).

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

100 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).

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

102 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).

103 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).

104 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).

105 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).

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

107 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).

108 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).

109 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).

110 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).

Table 1. Copy number and single-nucleotide variant diversity of HSD genes

		HGDP (n=236)		1000 Genomes (n=2,143)				1000 Genomes (N=658)					ASD unaffected controls (N=2,617)					
Gene Family	Duplication Type	Average CN	Individual with no HSD	Average CN	Individual with no HSD	Number exons	Number CDS	Number MIPS*	Rare LGD variants	Common LGD variants	Total variants	Total individuals	Number MIPS*	Rare LGD variants	Common LGD variants	Total variants	Total individuals	high prior.
ARHGAP11	partial	3.8	0	4.0	0	12	3072	40/41	1	1	2	60	40/41	8	0	8	43	x
ARHGEF5	partial	5.8	2	6.7	4	15	4794	52/57	5	3	8	469	57/57	16	2	18	2003	
*BOLA2***	complete	6.1	0	6.1	0	3	459	7/7	0	0	0	0	7/7	3	0	3	3	x
CD8B	partial	3.7	0	3.7	0	6	633	13/13	2	0	2	17	13/13	1	0	1	11	x
CFC1	complete	3.9	0	4.0	0	6	672	8/11	4	0	4	12	9/11	2	0	2	18	x
CHRNA7	partial-fusion	3.9	1	3.8	34	10	1509	26/27	1	1	2	313	25/27	1	1	2	1709	
*CORO1A***	partial	6.1	0	6.1	0	11	1386	21/27	1	0	1	6	25/27	1	0	1	2	
*DUSP22***	complete	3.6	7	3.8	28	8	618	14/17	0	0	0	0	16/17	3	0	3	3	
FAM72	complete	7.2	0	6.7	0	4	450	5/6	0	0	0	0	5/6	1	0	1	1	x
*FCGR1^*	complete	7.8	0	5.5	0	6	1125	19/19	5	4	9	658	19/19	5	4	9	2561	
FRMPD2	partial	4.0	0	4.3	0	29	3930	63/64	5	0	5	19	64/64	9	1	10	158	
GPR89	complete	3.9	0	3.9	0	14	1368	25/26	2	0	2	10	25/26	4	0	4	5	x
*GPRIN2***	complete	4.1	0	4.3	0	1	1377	12/16	1	0	1	1	14/16	2	0	2	2	x
GTF2H2	complete	4.7	2	4.4	17	16	1188	22/26	6	0	6	20	23/26	6	1	7	220	
*GTF2I^*	partial	6.0	0	6.0	0	35	2997	60/66	7	1	8	655	62/66	10	2	12	2604	
*GTF2IRD2^*	partial	6.0	0	6.0	0	16	2850	44/47	3	1	4	618	46/47	7	1	8	2601	
*HIST2H2BF***^*	complete	7.8	0	5.5	0	2	405	7/7	2	1	3	635	7/7	2	1	3	2561	
*HYDIN^*	partial	3.9	0	3.9	0	86	15366	233/240	8	0	8	36	239/240	31	1	32	2593	
MIR4267	complete	4.3	0	4.2	1	1	0	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	
MIR4435	complete	4.0	1	3.9	0	1	0	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	
*NAIP****	partial	5.4	2	6.2	5	17	4212	53/55	8	1	9	143	54/55	17	0	17	199	
*NCF1^*	complete	6.0	0	6.0	0	11	1173	21/24	1	1	2	622	21/24	3	1	4	2597	
NPY4R	complete	4.1	0	4.2	0	3	1128	12/12	2	0	2	5	12/12	3	0	3	5	x
OCLN	partial	3.1	47	3.0	582	8	1569	19/21	3	0	3	12	21/21	2	0	2	48	
OR3A	complete	4.7	2	4.4	17	1	0	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	
PTPN20	partial	4.0	0	4.3	0	10	1020	24/25	2	0	2	16	25/25	6	0	6	15	
*ROCK1^*	partial	3.8	0	3.8	1	33	4065	57/65	0	1	1	656	59/65	0	1	1	2595	
SERF1	complete	3.9	19	3.5	109	3	333	6/7	1	0	1	8	6/7	3	0	3	4	
SMN1	complete	3.9	19	3.5	109	9	885	17/18	1	0	1	1	18/18	1	0	1	2	
SRGAP2	partial	7.2	0	6.7	0	22	3216	79/82	3	0	3	8	79/82	7	0	7	11	x
TCAF1	partial	4.1	3	4.0	24	9	2766	33/33	0	0	0	0	33/33	5	0	5	6	
TCAF2#	partial	4.0	3	4.2	16	7	2448	33/40	6	1	7	103	28/31	7	0	7	8	
TISP43	complete	3.9	0	4.0	0	3	483	5/6	0	0	0	0	6/6	1	0	1	3	x

** Average MIP read-depth > 10*

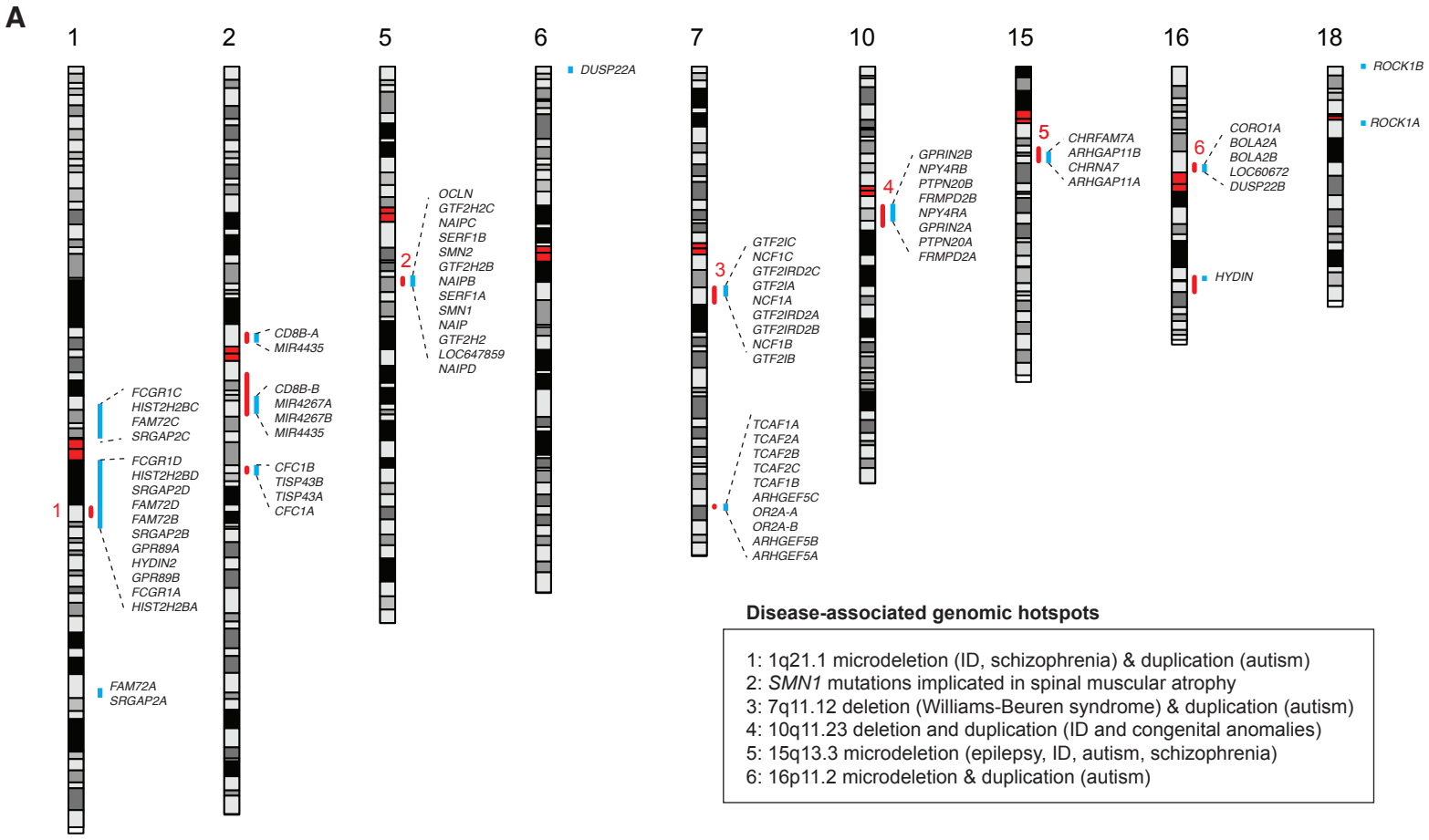
** Cannot differentiate paralogs

*** Gene paralog deleted from duplication

^ Gene paralog has a fixed LGD

MIPs were excluded in the ASD sequence analysis

FIGURE 1


A
1
2
5
6
7
10
15
16
18
DUSP22A
ROCK1B
ROCK1A
FCGR1C
HIST2H2BC
FAM72C
SRGAP2C
FCGR1D
HIST2H2BD
SRGAP2D
FAM72D
FAM72B
SRGAP2B
GPR89A
HYDIN2
GPR89B
FCGR1A
HIST2H2BA
FAM72A
SRGAP2A
CD8B-A
MIR4435
CD8B-B
MIR4267A
MIR4267B
MIR4435
CFC1B
TISP43B
TISP43A
CFC1A
OCLN
GTF2H2C
NAIPC
SERF1B
SMN2
GTF2H2B
NAIPB
SERF1A
SMN1
NAIP
GTF2H2
LOC647859
NAIPD
GTF2IC
NCF1C
GTF2IRD2C
GTF2IA
NCF1A
GTF2IRD2A
GTF2IRD2B
NCF1B
GTF2IB
TCAF1A
TCAF2A
TCAF2B
TCAF2C
TCAF1B
ARHGEF5C
OR2A-A
OR2A-B
ARHGEF5B
ARHGEF5A
GPRIN2B
NPY4RB
PTPN20B
FRMPD2B
NPY4RA
GPRIN2A
PTPN20A
FRMPD2A
CHRFAM7A
ARHGAP11B
CHRNA7
ARHGAP11A
CORO1A
BOLA2A
BOLA2B
LOC60672
DUSP22B
HYDIN
Disease-associated genomic hotspots
1: 1q21.1 microdeletion (ID, schizophrenia) & duplication (autism)
2: SMN1 mutations implicated in spinal muscular atrophy
3: 7q11.12 deletion (Williams-Beuren syndrome) & duplication (autism)
4: 10q11.23 deletion and duplication (ID and congenital anomalies)
5: 15q13.3 microdeletion (epilepsy, ID, autism, schizophrenia)
6: 16p11.2 microdeletion & duplication (autism)

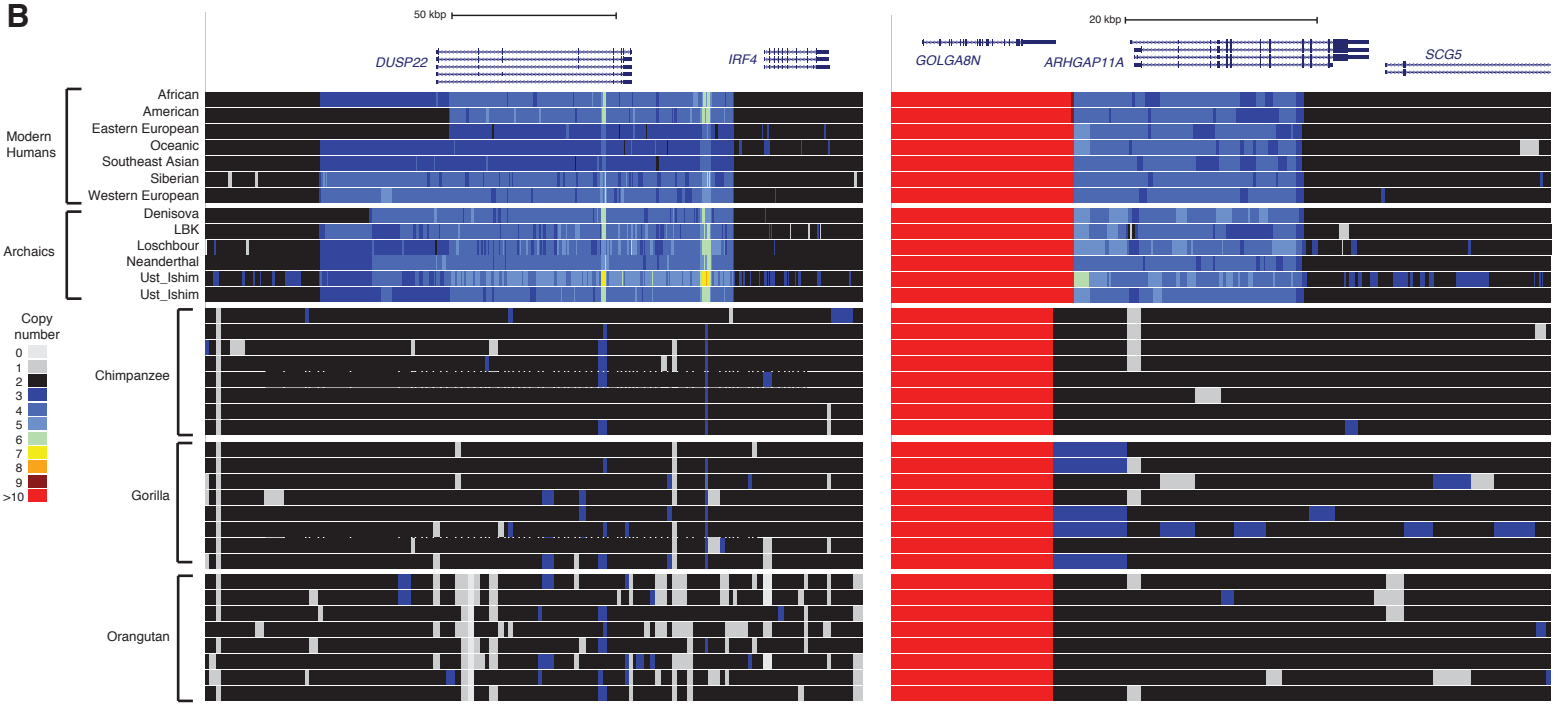

B
50 kbp
DUSP22
IRF4
20 kbp
GOLGA8N
ARHGAP11A
SCG5
Modern Humans
African
American
Eastern European
Oceanic
Southeast Asian
Siberian
Western European
Archaics
Denisova
LBK
Loschbour
Neanderthal
Ust_Ishim
Ust_Ishim
Copy number
0
1
2
3
4
5
6
7
8
9
>10
Chimpanzee
Gorilla
Orangutan

FIGURE 2

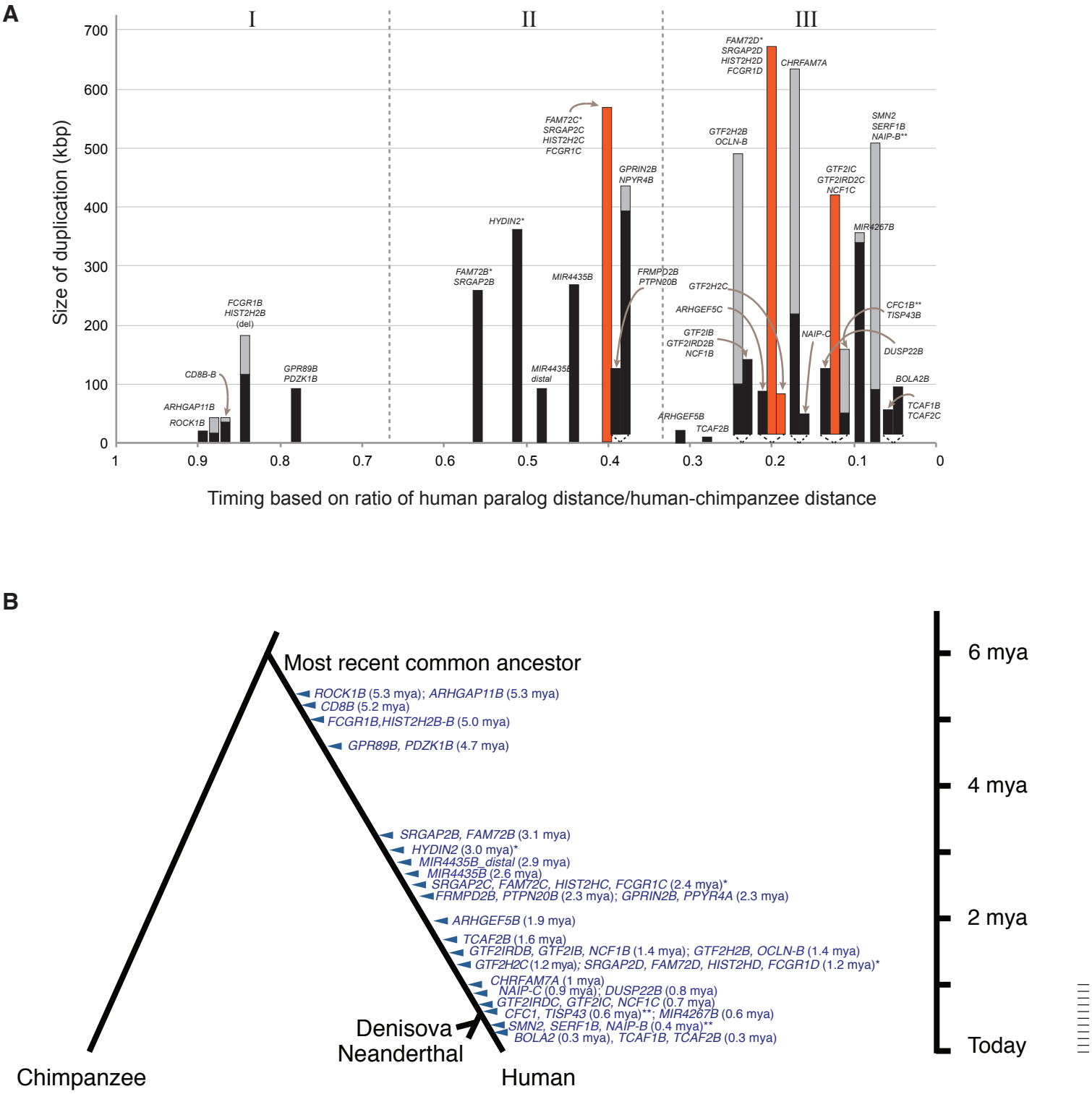

A
I
II
III
Size of duplication (kbp)
700
600
500
400
300
200
100
0
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Timing based on ratio of human paralog distance/human-chimpanzee distance
FCGR1B HIST2H2B (del)
CD8B-B
ARHGAP11B
ROCK1B
GPR89B PDZK1B
FAM72B* SRGAP2B
HYDIN2*
MIR4435B
MIR4435B distal
FAM72C* SRGAP2C HIST2H2C FCGR1C
GPRIN2B NPYR4B
FRMPD2B PTPN20B
ARHGEF5B
TCAF2B
GTF2H2B OCLN-B
FAM72D* SRGAP2D HIST2H2D FCGR1D
CHRFAM7A
GTF2H2C
ARHGEF5C
GTF2IB GTF2IRD2B NCF1B
NAIP-C
GTF2IC GTF2IRD2C NCF1C
SMN2 SERF1B NAIP-B**
MIR4267B
CFC1B** TISP43B
DUSP22B
BOLA2B
TCAF1B TCAF2C
B
Most recent common ancestor
ROCK1B (5.3 mya); ARHGAP11B (5.3 mya)
CD8B (5.2 mya)
FCGR1B,HIST2H2B-B (5.0 mya)
GPR89B, PDZK1B (4.7 mya)
SRGAP2B, FAM72B (3.1 mya)
HYDIN2 (3.0 mya)*
MIR4435B_distal (2.9 mya)
MIR4435B (2.6 mya)
SRGAP2C, FAM72C, HIST2HC, FCGR1C (2.4 mya)*
FRMPD2B, PTPN20B (2.3 mya); GPRIN2B, PPYR4A (2.3 mya)
ARHGEF5B (1.9 mya)
TCAF2B (1.6 mya)
GTF2IRDB, GTF2IB, NCF1B (1.4 mya); GTF2H2B, OCLN-B (1.4 mya)
GTF2H2C (1.2 mya); SRGAP2D, FAM72D, HIST2HD, FCGR1D (1.2 mya)*
CHRFAM7A (1 mya)
NAIP-C (0.9 mya); DUSP22B (0.8 mya)
GTF2IRDC, GTF2IC, NCF1C (0.7 mya)
CFC1, TISP43 (0.6 mya)**; MIR4267B (0.6 mya)
SMN2, SERF1B, NAIP-B (0.4 mya)**
BOLA2 (0.3 mya), TCAF1B, TCAF2B (0.3 mya)
Denisova
Neanderthal
Chimpanzee
Human
6 mya
4 mya
2 mya
Today

FIGURE 3

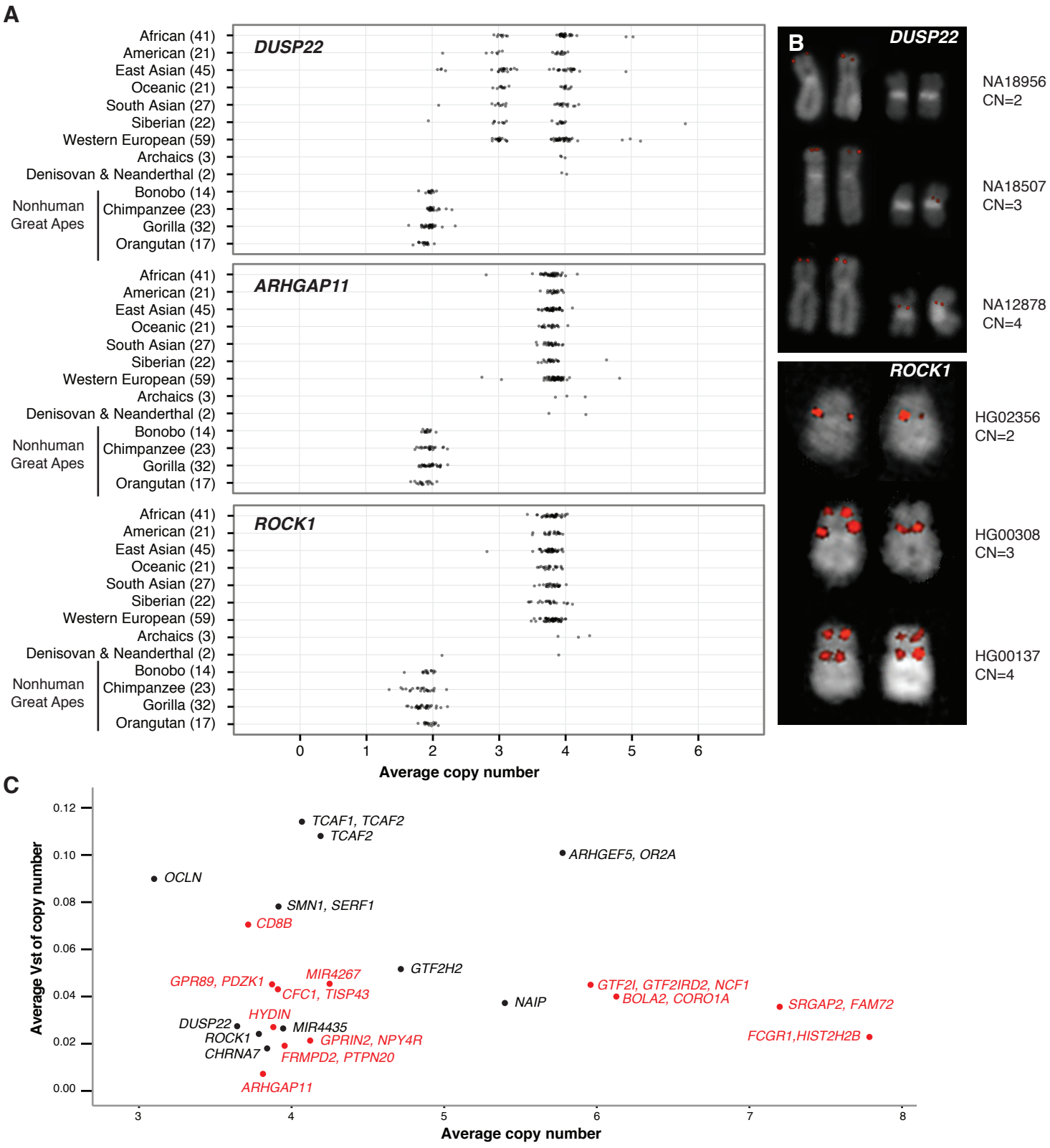

A
DUSP22
ARHGAP11
ROCK1
African (41)
American (21)
East Asian (45)
Oceanic (21)
South Asian (27)
Siberian (22)
Western European (59)
Archaics (3)
Denisovan & Neanderthal (2)
Bonobo (14)
Chimpanzee (23)
Gorilla (32)
Orangutan (17)
Nonhuman Great Apes
0
1
2
3
4
5
6
Average copy number
B
DUSP22
NA18956
CN=2
NA18507
CN=3
NA12878
CN=4
ROCK1
HG02356
CN=2
HG00308
CN=3
HG00137
CN=4
C
Average Vst of copy number
0.12
0.10
0.08
0.06
0.04
0.02
0.00
TCAF1, TCAF2
TCAF2
ARHGEF5, OR2A
OCLN
SMN1, SERF1
CD8B
GTF2H2
GPR89, PDZK1
MIR4267
CFC1, TISP43
GTF2I, GTF2IRD2, NCF1
BOLA2, CORO1A
NAIP
SRGAP2, FAM72
HYDIN
DUSP22
MIR4435
ROCK1
GPRIN2, NPY4R
FCGR1,HIST2H2B
CHRNA7
FRMPD2, PTPN20
ARHGAP11
3
4
5
6
7
8
Average copy number

FIGURE 4

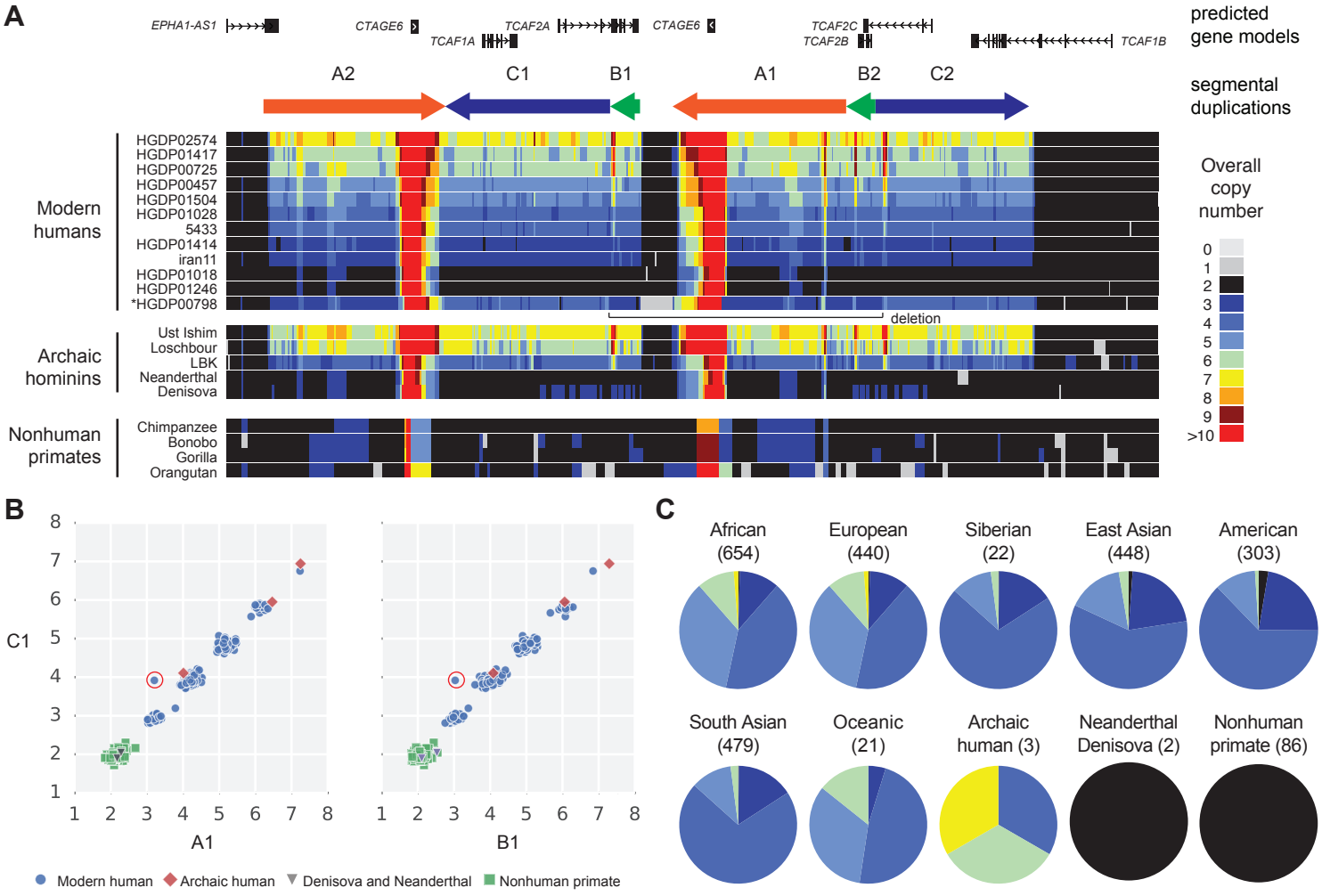

A
EPHA1-AS1
CTAGE6
TCAF1A
TCAF2A
CTAGE6
TCAF2C
TCAF2B
TCAF1B
predicted gene models
A2
C1
B1
A1
B2
C2
segmental duplications
Modern humans
HGDP02574
HGDP01417
HGDP00725
HGDP00457
HGDP01504
HGDP01028
5433
HGDP01414
iran11
HGDP01018
HGDP01246
*HGDP00798
deletion
Archaic hominins
Ust Ishim
Loschbour
LBK
Neanderthal
Denisova
Nonhuman primates
Chimpanzee
Bonobo
Gorilla
Orangutan
Overall copy number
0
1
2
3
4
5
6
7
8
9
>10
B
C1
A1
B1
Modern human
Archaic human
Denisova and Neanderthal
Nonhuman primate
C
African (654)
European (440)
Siberian (22)
East Asian (448)
American (303)
South Asian (479)
Oceanic (21)
Archaic human (3)
Neanderthal Denisova (2)
Nonhuman primate (86)

FIGURE 5

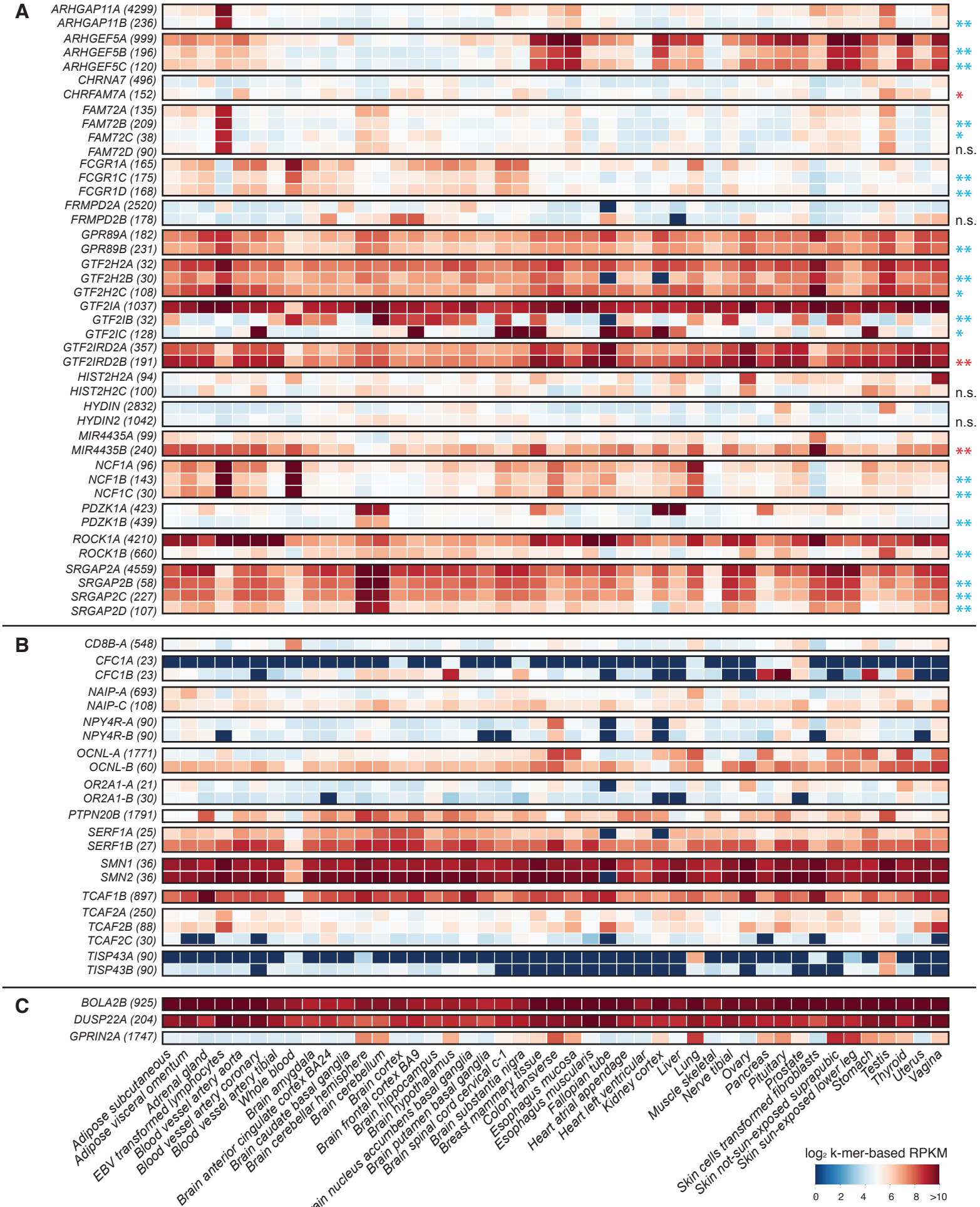

A
ARHGAP11A (4299)
ARHGAP11B (236)
ARHGEF5A (999)
ARHGEF5B (196)
ARHGEF5C (120)
CHRNA7 (496)
CHRFAM7A (152)
FAM72A (135)
FAM72B (209)
FAM72C (38)
FAM72D (90)
FCGR1A (165)
FCGR1C (175)
FCGR1D (168)
FRMPD2A (2520)
FRMPD2B (178)
GPR89A (182)
GPR89B (231)
GTF2H2A (32)
GTF2H2B (30)
GTF2H2C (108)
GTF2IA (1037)
GTF2IB (32)
GTF2IC (128)
GTF2IRD2A (357)
GTF2IRD2B (191)
HIST2H2A (94)
HIST2H2C (100)
HYDIN (2832)
HYDIN2 (1042)
MIR4435A (99)
MIR4435B (240)
NCF1A (96)
NCF1B (143)
NCF1C (30)
PDZK1A (423)
PDZK1B (439)
ROCK1A (4210)
ROCK1B (660)
SRGAP2A (4559)
SRGAP2B (58)
SRGAP2C (227)
SRGAP2D (107)
**
*
n.s.
B
CD8B-A (548)
CFC1A (23)
CFC1B (23)
NAIP-A (693)
NAIP-C (108)
NPY4R-A (90)
NPY4R-B (90)
OCNL-A (1771)
OCNL-B (60)
OR2A1-A (21)
OR2A1-B (30)
PTPN20B (1791)
SERF1A (25)
SERF1B (27)
SMN1 (36)
SMN2 (36)
TCAF1B (897)
TCAF2A (250)
TCAF2B (88)
TCAF2C (30)
TISP43A (90)
TISP43B (90)
C
BOLA2B (925)
DUSP22A (204)
GPRIN2A (1747)
Adipose subcutaneous
Adipose visceral omentum
Adrenal gland
EBV transformed lymphocytes
Blood vessel artery aorta
Blood vessel artery coronary
Blood vessel artery tibial
Whole blood
Brain amygdala
Brain anterior cingulate cortex BA24
Brain caudate basal ganglia
Brain cerebellar hemisphere
Brain cerebellum
Brain cortex
Brain frontal cortex BA9
Brain hippocampus
Brain hypothalamus
Brain nucleus accumbens basal ganglia
Brain putamen basal ganglia
Brain spinal cord cervical c-1
Brain substantia nigra
Breast mammary tissue
Colon transverse
Esophagus mucosa
Esophagus muscularis
Fallopian tube
Heart atrial appendage
Heart left ventricular
Kidney cortex
Liver
Lung
Muscle skeletal
Nerve tibial
Ovary
Pancreas
Pituitary
Prostate
Skin cells transformed fibroblasts
Skin not-sun-exposed suprapubic
Skin sun-exposed lower leg
Stomach
Testis
Thyroid
Uterus
Vagina
log2 k-mer-based RPKM
0
2
4
6
8
>10

FIGURE 6A

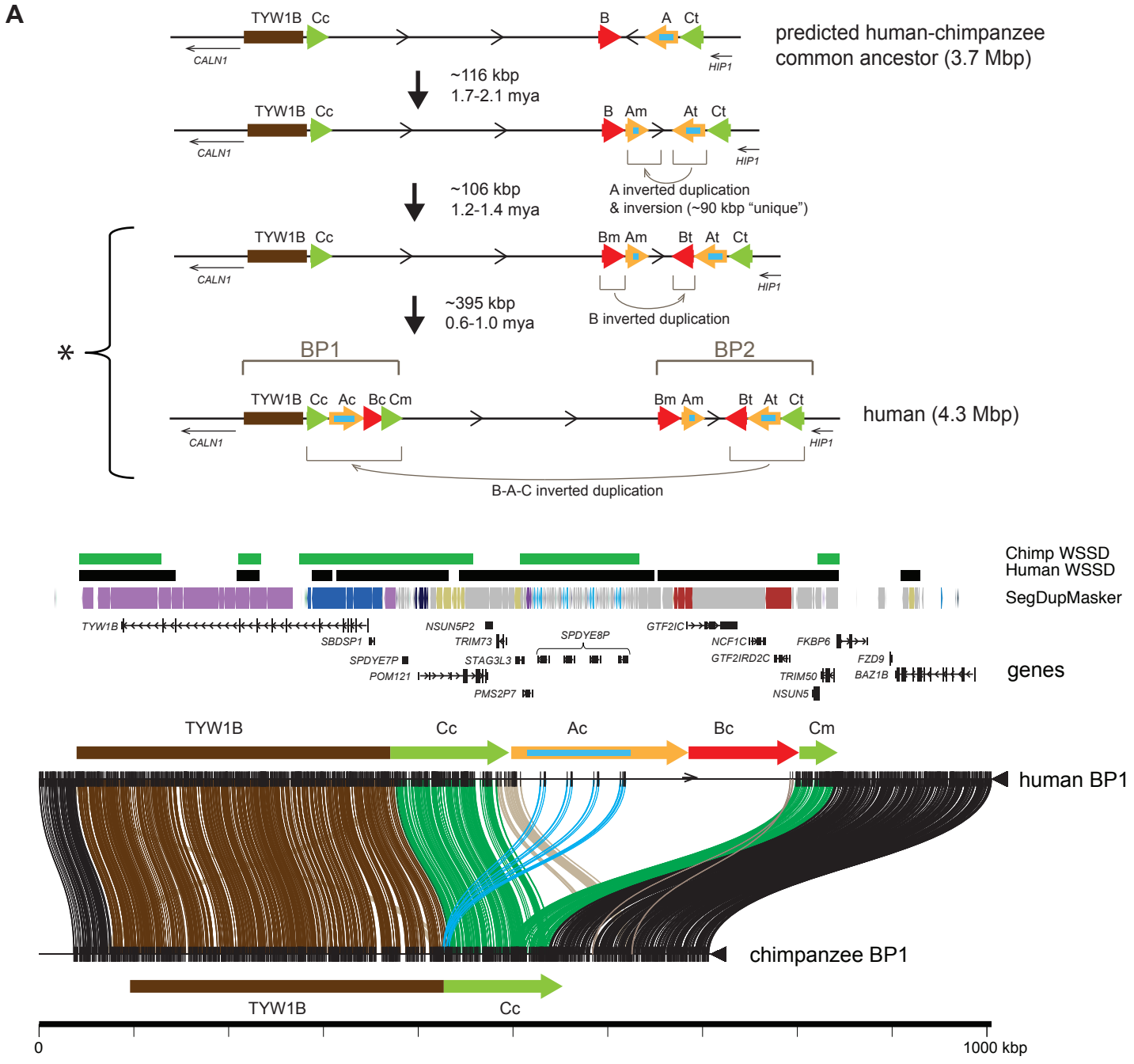

A
TYW1B Cc
B A Ct
CALN1
HIP1
predicted human-chimpanzee common ancestor (3.7 Mbp)
~116 kbp
1.7-2.1 mya
TYW1B Cc
B Am At Ct
CALN1
HIP1
~106 kbp
1.2-1.4 mya
A inverted duplication
& inversion (~90 kbp "unique")
TYW1B Cc
Bm Am Bt At Ct
CALN1
HIP1
~395 kbp
0.6-1.0 mya
B inverted duplication
*
BP1
BP2
TYW1B Cc Ac Bc Cm
Bm Am Bt At Ct
CALN1
HIP1
human (4.3 Mbp)
B-A-C inverted duplication
Chimp WSSD
Human WSSD
SegDupMasker
TYW1B
SBDSP1
NSUN5P2
TRIM73
SPDYE8P
SPDYE7P
STAG3L3
POM121
PMS2P7
GTF2IC
NCF1C
GTF2IRD2C
FKBP6
FZD9
TRIM50
BAZ1B
NSUN5
genes
TYW1B
Cc
Ac
Bc
Cm
human BP1
chimpanzee BP1
TYW1B
Cc
0
1000 kbp

FIGURE 6B

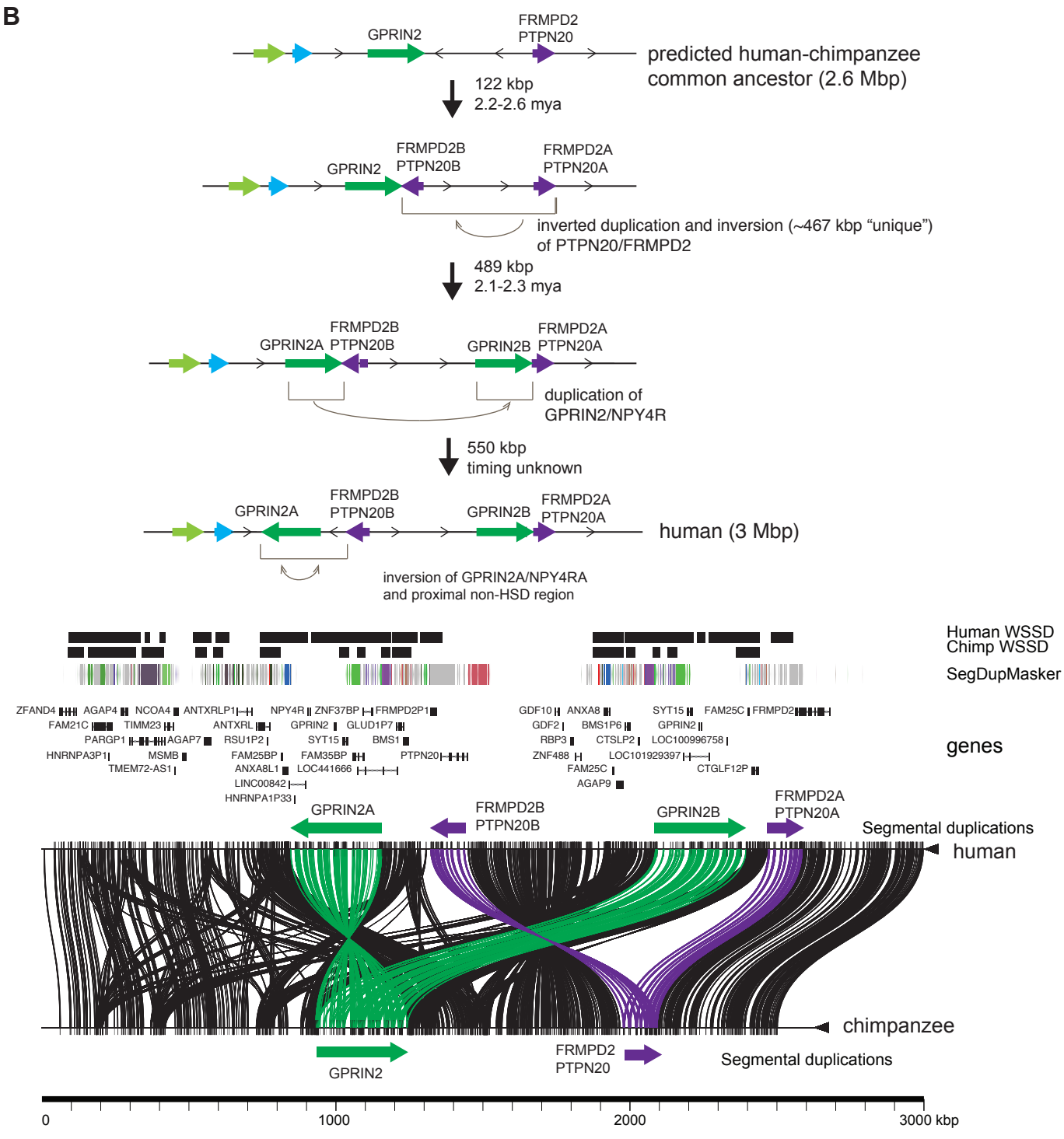
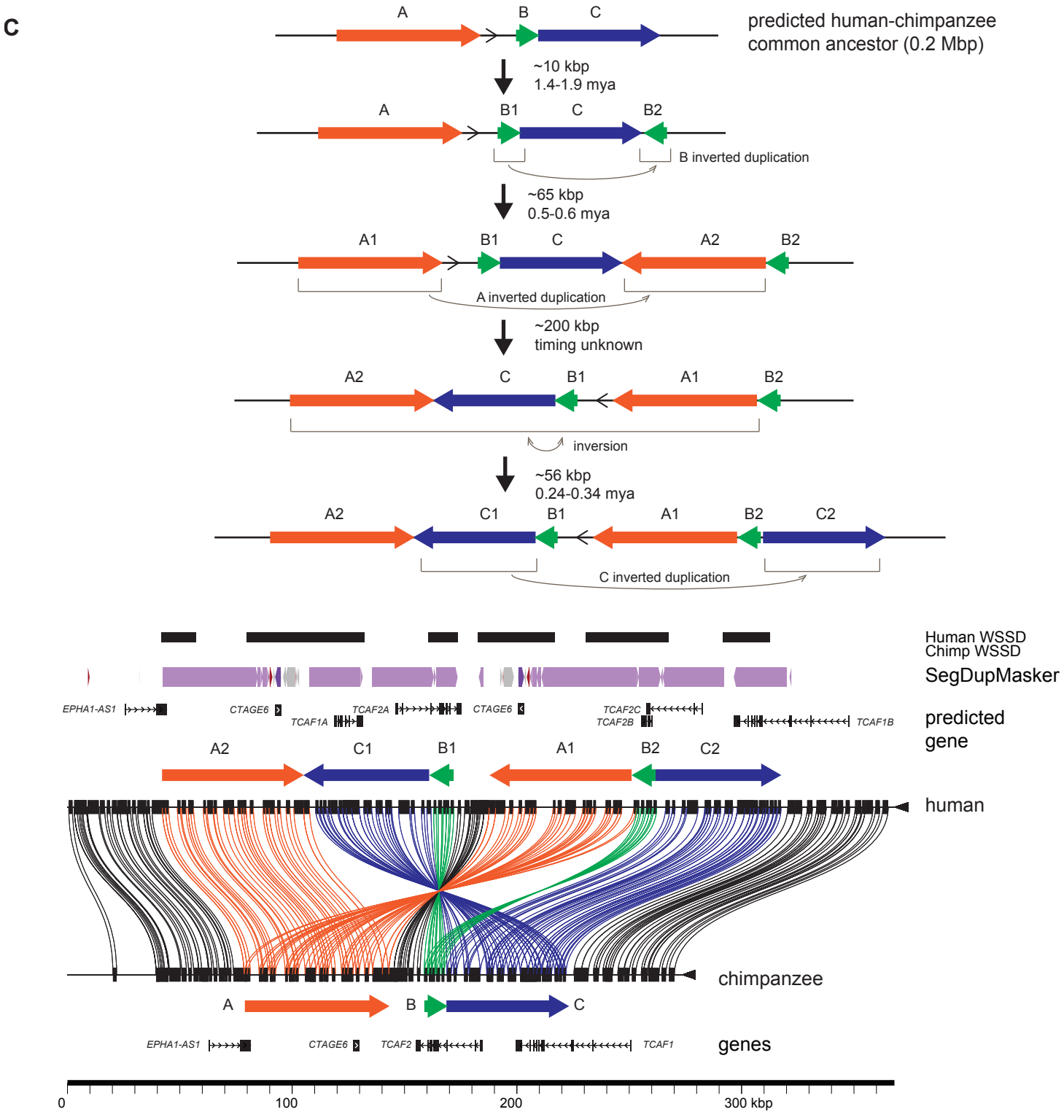

B
GPRIN2
FRMPD2
PTPN20
predicted human-chimpanzee
common ancestor (2.6 Mbp)
122 kbp
2.2-2.6 mya
GPRIN2
FRMPD2B
PTPN20B
FRMPD2A
PTPN20A
inverted duplication and inversion (~467 kbp "unique")
of PTPN20/FRMPD2
489 kbp
2.1-2.3 mya
GPRIN2A
FRMPD2B
PTPN20B
GPRIN2B
FRMPD2A
PTPN20A
duplication of
GPRIN2/NPY4R
550 kbp
timing unknown
GPRIN2A
FRMPD2B
PTPN20B
GPRIN2B
FRMPD2A
PTPN20A
human (3 Mbp)
inversion of GPRIN2A/NPY4RA
and proximal non-HSD region
Human WSSD
Chimp WSSD
SegDupMasker
genes
ZFAND4
AGAP4
NCOA4
ANTXRLP1
NPY4R
ZNF37BP
FRMPD2P1
FAM21C
TIMM23
ANTXRL
GPRIN2
GLUD1P7
PARGP1
AGAP7
RSU1P2
SYT15
BMS1
HNRNPA3P1
MSMB
FAM25BP
FAM35BP
PTPN20
TMEM72-AS1
ANXA8L1
LOC441666
LINC00842
HNRNPA1P33
GDF10
ANXA8
SYT15
FAM25C
FRMPD2
GDF2
BMS1P6
GPRIN2
RBP3
CTSLP2
LOC100996758
ZNF488
LOC101929397
FAM25C
CTGLF12P
AGAP9
GPRIN2A
FRMPD2B
PTPN20B
GPRIN2B
FRMPD2A
PTPN20A
Segmental duplications
human
chimpanzee
GPRIN2
FRMPD2
PTPN20
Segmental duplications
0
1000
2000
3000 kbp

FIGURE 6C


C
A
B
C
predicted human-chimpanzee common ancestor (0.2 Mbp)
~10 kbp
1.4-1.9 mya
A
B1
C
B2
B inverted duplication
~65 kbp
0.5-0.6 mya
A1
B1
C
A2
B2
A inverted duplication
~200 kbp
timing unknown
A2
C
B1
A1
B2
inversion
~56 kbp
0.24-0.34 mya
A2
C1
B1
A1
B2
C2
C inverted duplication
Human WSSD
Chimp WSSD
SegDupMasker
EPHA1-AS1
CTAGE6
TCAF1A
TCAF2A
CTAGE6
TCAF2C
TCAF2B
TCAF1B
predicted gene
A2
C1
B1
A1
B2
C2
human
chimpanzee
A
B
C
EPHA1-AS1
CTAGE6
TCAF2
TCAF1
genes
0
100
200
300 kbp