nature ecology & evolution

ARTICLES

PUBLISHED: 17 FEBRUARY 2017 | VOLUME: 1 | ARTICLE NUMBER: 0069

The evolution and population diversity of human-specific segmental duplications

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

Segmental duplications contribute to human evolution, adaptation and genomic instability but are often poorly characterized. We investigate the evolution, genetic variation and coding potential of human-specific segmental duplications (HSDs). We identify 218 HSDs based on analysis of 322 deeply sequenced archaic and contemporary hominid genomes. We sequence 550 human and nonhuman primate genomic clones to reconstruct the evolution of the largest, most complex regions with protein-coding potential ($N = 80$ genes from 33 gene families). We show that HSDs are non-randomly organized, associate preferentially with ancestral ape duplications termed 'core duplicons' and evolved primarily in an interspersed inverted orientation. In addition to *Homo sapiens*-specific gene expansions (such as *TCAF1/TCAF2*), we highlight ten gene families (for example, *ARHGAP11B* and *SRGAP2C*) where copy number never returns to the ancestral state, there is evidence of mRNA splicing and no common gene-disruptive mutations are observed in the general population. Such duplicates are candidates for the evolution of human-specific adaptive traits.

Genetic mutations have shaped the unique adaptation and evolution of the human lineage, but their characterization has been a slow and difficult endeavour. Despite a few potential success stories over the years with various degrees of support[1], the genetic basis of most of the unique aspects of human adaptation awaits discovery. As sequencing technologies have improved, more systematic efforts have been directed to discover regulatory differences among the great apes[2–6]. One potential source of genetic variation, which has been difficult to explore due to missing or erroneous sequences within reference genomes, are genes embedded in recently (<25 Ma (million years ago)) duplicated regions, also called segmental duplications (SDs)[7]. 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[8]. 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[9–12].

The genomes of apes are enriched in SDs, having experienced a burst of interspersed duplications over the past 10 Myr of evolution[13,14]. The mosaic and interspersed architecture of ape SDs offers tremendous potential for transcript innovation because duplicate paralogues may be truncated, combined with other transcripts to create fusion genes, or acquire alternative promoters directing the differential expression of new transcripts[15]. Previous investigations have been limited to microarray studies[16,17] and whole-genome sequencing read-depth comparisons[14,18,19] between humans and great apes. None of these methods provide information regarding the structure and sequence of the duplicated segments, limiting gene annotation and the understanding of the functional potential of the duplicated genes.

Here, 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 (CN) polymorphic parts of the genome and frequently embedded within larger blocks of shared ape duplications. Not surprisingly, the genome assembly builds of these regions are highly enriched for euchromatic gaps and misassembly errors even in the most recent versions of the human reference[20,21]. 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 (CH17) 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[22,23] or from postzygotic loss of a complete parental genome[24]. 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[11]. We apply the resulting high-quality sequence to more systematically investigate CN 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 both from humans and great apes, we began

[1]Genome Center, MIND Institute, and Department of Biochemistry & Molecular Medicine, University of California, Davis, California 95616, USA. [2]Department of Genome Sciences, University of Washington School of Medicine, Seattle, Washington 98195, USA. [3]Howard Hughes Medical Institute, University of Washington, Seattle, Washington 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, Missouri 63108, USA. *e-mail: eee@gs.washington.edu

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 (Fig. 1; Supplementary Fig. 1). Our approach identified 218 autosomal regions ranging in size from 5 kbp (our size threshold) to 362 kbp with HSDs dispersed non-randomly near each other (empirical median distance to nearest HSD = 440 kbp, $P < 1 \times 10^{-7}$; Supplementary Fig. 2; see Methods). Of these regions, 85 included entire or parts of RefSeq annotated genes (Supplementary Table 1). We orthogonally validated 87% (190/218) of our events as HSDs.

The set included 38 previously unreported HSDs mapping to genic regions. Among these, we included HSDs where there was evidence of independent or distinct duplications in great apes (that is, 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 CN > 8; Supplementary Fig. 3). Similarly, we identified a 6.6 kbp duplication corresponding to 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 Figs 3 and 4). We also identified a new 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 Fig. 3).

Next, we focused on the largest gene-containing HSD regions (>20 kbp; Supplementary Fig. 5). These HSDs and their ancestral counterparts reside in 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 (Fig. 1; Supplementary Table 2)[25]. Using the haploid BAC library (CH17), we generated alternate sequence assemblies (Supplementary Tables 3 and 4) of which 18.2 Mbp have now been incorporated into the most recent human reference build (GRCh38), allowing us to close 24 euchromatic gaps and correct large-scale errors in the human reference genome. The new sequence allowed us to distinguish 28 HSD events ranging in size from 11 kbp to 677 kbp corresponding to 33 HSD gene families accounting for 80 paralogous genes (Supplementary Table 5).

Most 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 (Fig. 2). Compared with a random null distribution, we found that these primary HSDs map closer to each other than by chance (empirical median distance to nearest HSD = 377 kbp, $P = 1 \times 10^{-7}$; Supplementary Fig. 2). Consistent with previous observations[26–28], we also found our primary HSDs to be significantly enriched near core duplicons—high-copy ancestral ape duplications significantly linked with the accumulation of SDs and breakpoints of rearrangement (empirical median distance to a core = 250 kbp, $P < 1 \times 10^{-7}$; Supplementary Fig. 6). We identified four secondary HSDs—additional duplications derived from a human-specific duplicate paralogue. These secondary events account for 35% (1.7 Mbp) of HSD base pairs because the events are larger when compared with primary duplications (minimum median sizes: 497 kbp versus 95 kbp, $P = 0.041$, Wilcoxon–Mann–Whitney test). The majority of HSDs are intrachromosomal and arranged in inverted orientation with respect to their ancestral paralogues (18/26, $P = 0.014$, binomial test), including all secondary duplications (4/4). HSD clustering is most pronounced on chromosome 1p12 to 1q32.1, which contains the greatest number of gene-containing HSDs (at least six independent HSD events including ~2 Mbp (0.8%) of human chromosome 1; Supplementary Note; Supplementary Fig. 7). We find that 85% of this 8 Mbp region (chr1:119,989,248–121,395,939 and chr1:143,311,826–149,876,379, GRCh38) has been duplicated in humans and great apes with only 1.15 Mbp remaining unique in humans.

Evolutionary timing of HSDs. We sequenced large-insert clones from nonhuman primate (NHP) genomic BAC libraries and applied standard phylogenetic methods under a model of gene conversion[29] to understand the evolutionary timing of each large HSD (Supplementary Tables 6–8; Supplementary Fig. 8). The results reveal differences in number and size of HSDs when we compare across three equal time periods during the evolution of the human lineage ($P = 0.017$; Fig. 2). The first was a period of relative quiescence, which occurred after the human–chimpanzee divergence. This included five smaller primary duplications corresponding to seven genes with a median minimum size of 36 kbp for a total of 285 kbp. This was followed temporally by a set of larger primary ($N = 6$) and secondary ($N = 1$) duplications containing 12 HSD genes (median minimum size of primary events 262 kbp for a total of 1.5 Mbp, $P = 0.026$). The final set of duplications involved more secondary ($N = 3$) and primary ($N = 13$) duplications and are estimated to be the most recent. Although primary duplication lengths were not significantly different in size compared with either of the other two time periods (median minimum size = 93 kbp for a total of 1.4 Mbp), they resulted in many more HSD genes (28 gene paralogues).

Human CN diversity. We undertook three different approaches to assess the potential functional significance of HSDs—namely, CN constraint, transcriptional potential and protein-coding mutations. We first assessed CN in contemporary and archaic hominin ($N = 2{,}384$)[30–33] as well as 86 NHP genomes[34] to distinguish fixed duplications from those that are highly stratified among humans (Fig. 3; Supplementary Tables 9–11). Thirteen HSD genes were among the most CN polymorphic, including genes at chromosomes 7q35 (three units: *ARHGEF5* and *OR2A*, *TCAF1*, and *TCAF2*), 5q13.1 (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 CN, showing the lowest variance among contemporary human populations (six units: *HYDIN*, *GPR89* and *PDZK1*, *CFC1* and *TISP43*, *CD8B*, *ROCK1*, and *ARHGAP11*; Fig. 3; Supplementary Figs 9 and 10; Supplementary Tables 12 and 13). As expected, higher CN HSD genes are generally more CN polymorphic (Supplementary Note). We identified 11/23 duplicated units with at least one normal individual identified who carried the ancestral state CN (diploid CN of two), suggesting the HSD paralogues are missing in these individuals (for example, *DUSP22* and *ROCK1*; Supplementary Fig. 9). Population differentiation (as measured by V_{ST})[35] generally correlated with CN variance ($R^2 = 0.32$; $\rho = 0.54$, Pearson correlation) but not CN ($R^2 = 0.01$; $\rho = -0.01$, Pearson correlation; Supplementary Fig. 11).

We also identified three genes expanded uniquely in *Homo sapiens* when compared with two sequenced archaic hominins, a Neanderthal and a Denisovan. This included the previously reported *BOLA2* on chromosome 16 (refs 32,36; Supplementary Fig. 12) and two new genes, *TRPM8*-associated *TCAF1* and *TCAF2* (formerly *FAM115A* and *FAM115C*), on chromosome 7 (Fig. 4; Supplementary Table 14; Supplementary Note)[37]. In the case of *TCAF1* and *TCAF2*, the timing estimate (0.048 ± 0.008 human–chimpanzee distance) is consistent with its absence in archaic hominins. The fact that we observe high CN in two archaic humans (Loschbour and Ust Ishim individuals with CN ≥ 6) suggests these HSDs spread rapidly in the population. The *TCAF1*/*TCAF2* HSD is differentiated (Human Genome Diversity Project (HGDP) mean $V_{ST} = 0.11$) between human populations with the highest

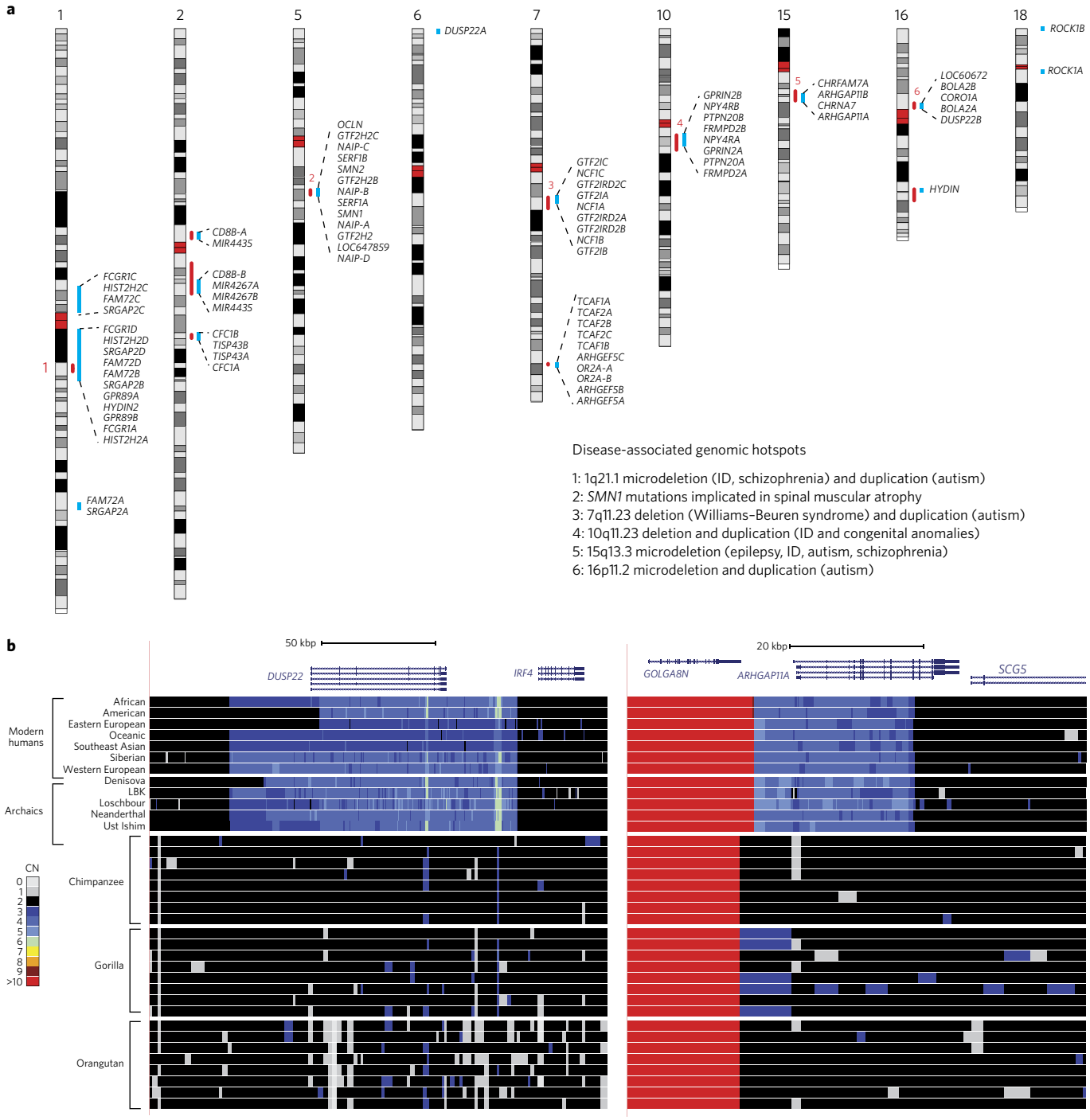

Figure 1 | Identification of HSDs. a, The locations of large, gene-containing HSDs are highlighted (blue lines) with 80 individual gene paralogues from 33 gene families listed across 9 different human autosomes. Included in this set are paralogues of *GPR89*, which duplicated in other great apes but experienced human-specific expansions. Many of these HSDs overlap known disease-implicated genomic hotspots (red lines) prone to recurrent CN variation associated with developmental delay. The genomic hotspots labelled with red numbers (1–6) have significant associations with specific disorders including epilepsy, autism, schizophrenia and intellectual disability (ID). **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 (HGDP, $N=236$)[30] and NHPs (gorillas, $N=32$; chimpanzees, $N=23$; bonobos, $N=14$; and orangutans, $N=17$)[34]. Overall CN was averaged across 500 bp sliding windows and depicted as coloured heatmaps (see pictured index). Also pictured are heatmaps for Neanderthal, Denisovan and archaic human (Linearbandkeramik (LBK), Loschbour and Ust Ishim) individuals. Any genomic region >5 kbp shown to have diploid CN ≥ 3 in 90% of humans tested compared with all NHPs was considered an HSD.

CN observed for African and European populations (in particular, Gambian and Esan from Nigeria, where multiple individuals with diploid CN = 7 are observed) and the lowest CN observed for Asian and Amerindian populations.

Patterns of HSD mRNA expression. To understand expression differences, we specifically examined RNA-sequencing (RNA-seq) data from GTEx[38] and mapped the distribution of reads to HSDs in 45 different tissues across multiple individuals (Supplementary

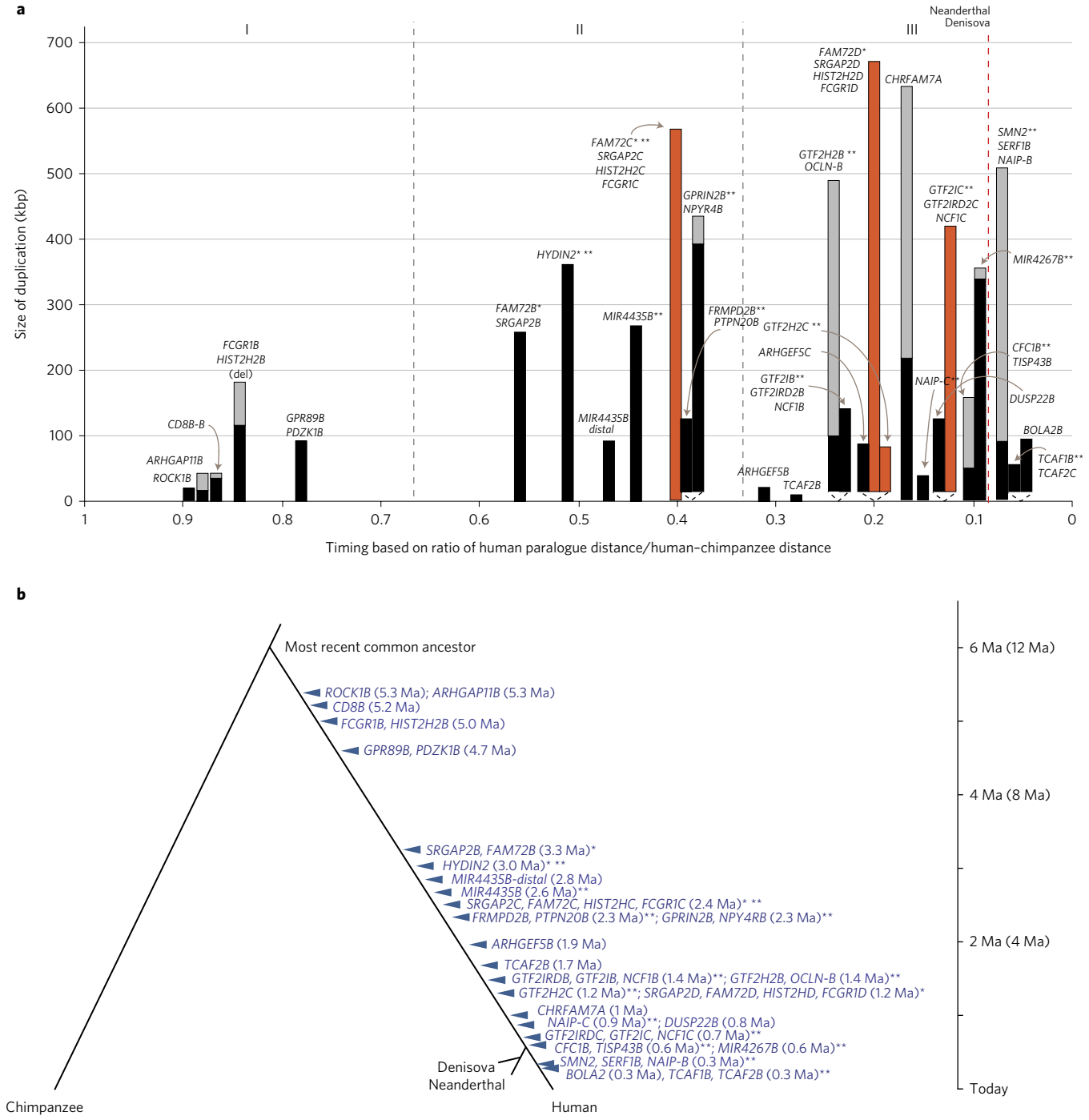

Figure 2 | Timing of HSDs. a, Duplication timing estimates are plotted as a ratio of human–chimpanzee divergence (*x* axis). The total estimated size of primary (black) and secondary (orange) duplications is shown for each event. Uncertainty in the size of each event is due to breakpoints mapping in high-identity flanking duplications (grey). **b**, Generally accepted phylogeny indicating timing of each event assuming chimpanzee and human lineage divergence of 6 Ma (refs [76–78]). Recent estimates based on recalibrated substitution rates suggest an earlier divergence time of 12 Mya, with this maximum scale indicated in parentheses. The analysis is based on high-quality sequencing, assembly and alignment of large-insert clones (human $N = 224$; NHP $N = 269$). Asterisks indicate adjusted timing estimates because of failed Tajima's *D* relative rate (*) and genes with evidence of IGC (**).

Tables 15 and 16; Supplementary Figs 13 and 14). Of the 26 comparisons that could be made between known ancestral and duplicate paralogues (Supplementary Fig. 13), 65% (17/26) of duplicate paralogues showed significantly lower expression levels compared with their ancestral paralogue (versus 19% (5/26) showing significantly greater expression and 15% (4/26) showing no difference in expression). In contrast, human-specific *FRMPD2B* and *CHRFAM7A* each show increased expression in specific tissues compared with their ancestral paralogues (*FRMPD2A* and *CHRNA7*). Both of the derived duplicates are incomplete, lacking the 5′ portion when compared with the ancestral gene. *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.

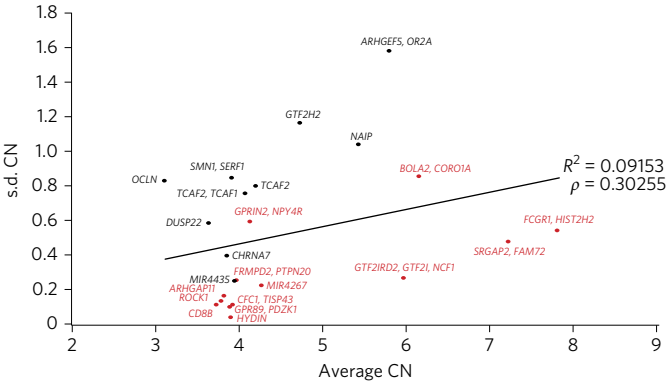

Figure 3 | Human CN diversity. Overall average CN was calculated per individual from read depth produced from Illumina mappings across a set region defining each duplication (Supplementary Table 9) in human populations, including the HGDP ($N=236$; GRCh38) and 1KG ($N=2,143$; GRCh37) cohorts, NHPs, archaic humans, a Denisovan and a Neanderthal. From these results, the mean, standard deviation (s.d.), V_{ST} and number of individuals with CN = 2 indicating no duplicate paralogues exist were calculated for average CN of each duplicated gene family (Supplementary Table 11). For each gene family, plots are shown for the CN average versus s.d. across all HGDP individuals with duplicate gene family names indicated next to each data point. Red data points indicate genes with no homozygous HSD deletions in any human tested. Genes with higher s.d. are considered CN polymorphic and tend to have higher CN ($R^2=0.09$; $\rho=0.30$, Pearson correlation) and average V_{ST} ($R^2=0.32$; $\rho=0.54$, Pearson correlation; Supplementary Fig. 11).

Discovery of likely gene-disruptive events. As pseudogenization is the most likely fate of duplicated genes, we tested whether paralogues had accumulated likely gene-disrupting (LGD) mutations by targeted sequencing of canonical protein-coding exons of 30 gene families using molecular inversion probes (MIPs)[39]. In 658 individuals from the 1000 Genomes Project (1KG), we identified 96 LGD variants (25 gene families), of which 33 could be definitively assigned to a copy (Supplementary Tables 10, 17–19). Ten duplicate paralogues and no ancestral paralogues harboured common LGD mutations (population frequency >5%) suggesting potential functional constraint. The remaining LGD variants ($N=63$) could not be unambiguously assigned but typically fell into gene families with more than one duplicate paralogue. Overall, we identified no LGD variants in five gene families and discovered only rare LGD variants (<5% population frequency) in an additional 15 gene families (Supplementary Table 10).

Next, we sequenced and compared the LGD frequency in 3,444 children with autism and 2,617 unaffected siblings. 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 20 and 21). Most 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 ($P_{\text{uncorrected}}=0.02$; Supplementary Fig. 15), although this result did not pass Bonferroni multiple testing correction. 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 paralogues represent pseudogenes. Combining these results with our assessment of 1KG individuals, 16/30 gene families showed an absence of common protein-disrupting variants in the 6,719 humans tested (Supplementary Table 10).

The complexity of HSD evolutionary history. To highlight the complex evolutionary history associated with such regions, we selected three loci for further investigation (see Supplementary Note for details of genomic hotspot chromosome 10q11.23). 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[40]. The directly oriented flanking HSDs (labelled B; Fig. 5a) contain three genes: *GTF2I*, *GTF2IRD2* and *NCF1*. Our analysis predicts that the most common human haplotype arose through a three-step evolutionary process. The first two events occurred within the distal duplication cluster of the region (Supplementary Fig. 16). They involved an inverted duplication of a ~116 kbp SD (termed A, containing paralogues of the high-copy duplicon (*SPDYE*)) and a possible 90 kbp inversion (0.313 ± 0.028 human–chimpanzee distance) followed by a separate ~106 kbp inverted duplication of B (0.229 ± 0.019 human–chimpanzee distance). These events created truncated paralogues *GTF2IB* and *GTF2IRD2B*, and a full-length version of *NCF1B*. A third large-scale inverted duplication transposed an ~395 kbp region comprising SDs A, B and C (containing *POM121L*) from the distal to proximal breakpoints of the disease-associated region (0.122 ± 0.014 human–chimpanzee distance). This tertiary duplication established 'granddaughter' truncated copies of *GTF2IRDC* and *GTF2IC*, as well as the full-length paralogue *NCF1C*, probably overwriting the 3′ end of the ancestral *POM121* with *POM121L*. This final event created directly oriented SDs A and B, providing a substrate for non-allelic homologous recombination leading to disease-associated CN variants. The great ape sequence (Supplementary Fig. 17) matched nearly perfectly the deduced genomic configuration hypothesized previously[41], with the exception of a large-scale inversion of the region proximal to BP1 in orangutan.

We also characterized one of the youngest HSD regions unique to modern humans on chromosome 7q35 containing *TCAF1*, *TCAF2* and primate-duplicated *CTAGE6*[42] (Fig. 4; Supplementary Note; Supplementary Fig. 18). We note that expansion of a *CTAGE* paralogue 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; Fig. 5b). We estimate an initial 10 kbp inverted duplication of HSD B containing the 3′ end of *TCAF2A* (0.275 ± 0.041 human–chimpanzee distance) creating a truncated *TCAF2B*. The subsequent events occurred very recently during human evolution, potentially during or after the split from a common ancestor of Denisova and Neanderthal. These subsequent rearrangements created a new full-length paralogue of *CTAGE6* (contained in A; 0.091 ± 0.008 human–chimpanzee distance) and truncated paralogues *TCAF1A* (the putative ancestral paralogue contained in C1; 0.048 ± 0.008 human–chimpanzee distance) and *TCAF2C* (contained in C2). Notably, we estimate that the full-length and functional *TCAF1B* and *TCAF2A* now reside on distinct SD paralogues 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.

Discussion

In this study, we generated new reference sequence for some of the most complex and gap-ridden sequence of the human genome. Several important features emerge from our targeted sequencing (48.4 Mbp) and evolutionary reconstruction of HSD regions (Supplementary Discussion). The largest HSDs are significantly clustered near core duplicons, including at chromosomes 1q21,

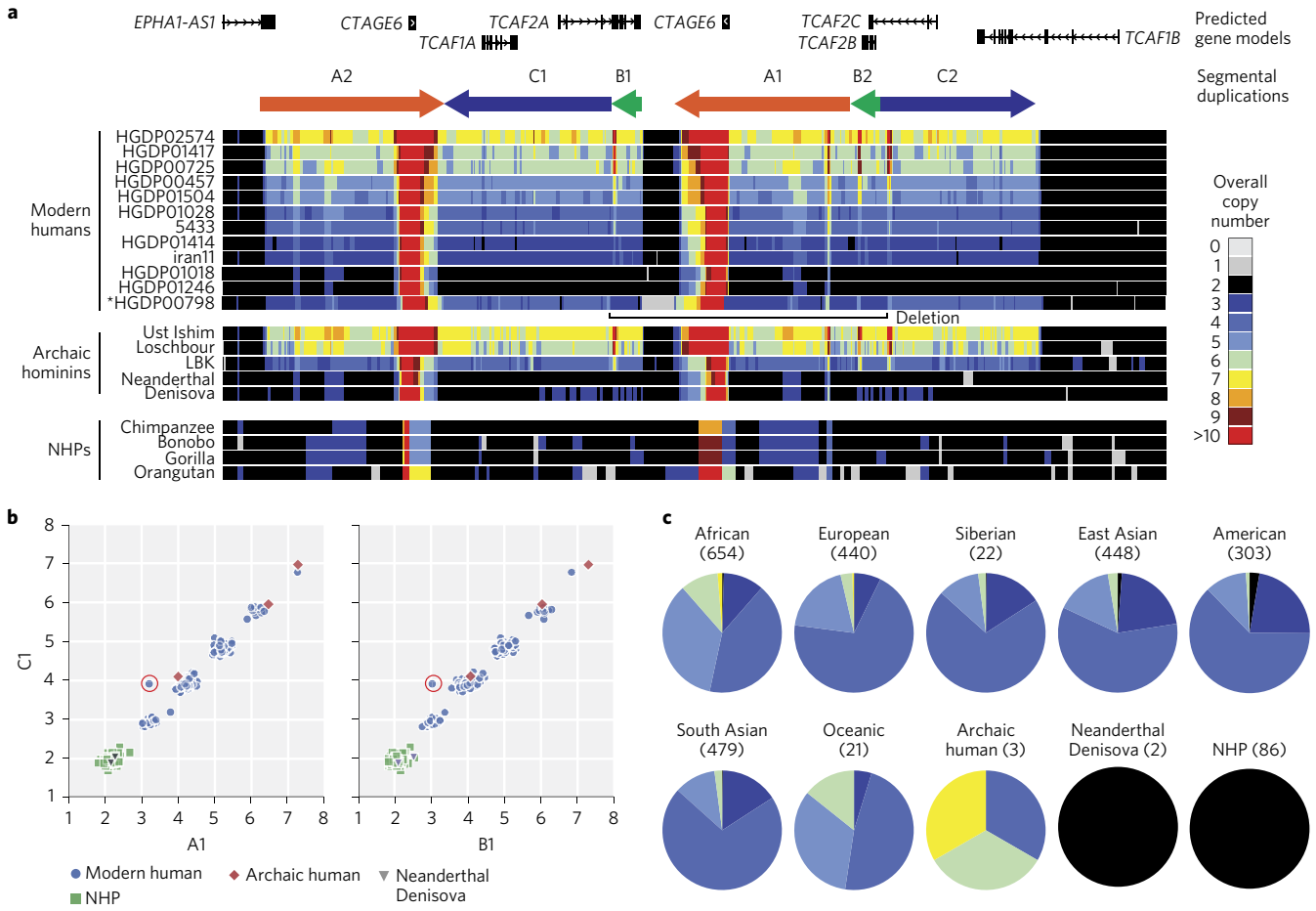

Figure 4 | CN polymorphism across diverse populations of *TCAF1* and *TCAF2* HSDs. a, Heatmap of overall CN of *TCAF1* and *TCAF2* HSD region on human chromosome 7 with predicted gene models and SDs (depicted as coloured arrows) pictured above. Representative modern humans are shown for each genotyped CN across the locus with a single person (*HGDP00798) showing deletion of the region, likely to be due to non-allelic homologous recombination between directly oriented SDs B1 and B2 (Supplementary Note). **b**, A scatter plot of *TCAF1* and *TCAF2* SDs (A1, B1 and C1) with 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' CNs is the deletion carrier pictured in **a**. **c**, CN predictions across modern humans from the 1KG and HGDP ($N=2,379$), archaic hominins, and NHPs were made across a representative region (chr7:143,533,137–143,571,789; GRCh37). Overall CNs in the pie charts per population are represented as colours depicted in the legend shown in **a**.

5q13 and 7q11.3. 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, $P=0.006$; Supplementary Table 5). Inverted SDs have been noted before in complex structural rearrangements associated with genomic disorders, such as Pelizaeus–Merzbacher disease[43,44] and Smith–Magenis syndrome[45], and may be a product of replication-based mechanisms, such as fork-stalling and template switching[43], and/or microhomology-mediated break-induced repair[46].

We enriched for potential functional HSD genes by applying three criteria: (1) all humans must carry the duplicate paralogue; (2) no common truncating mutations are observed in the human population; and (3) duplicates show evidence of spliced mRNA expression. Ten HSD gene families met all criteria, including two genes previously implicated in cortical development and neuronal spine density, *ARHGAP11B*[12] and *SRGAP2C*[10,11], as well as the gene families *BOLA2*, *CD8B*, *CFC1*, *FAM72*, *GPR89*, *GPRIN2*, *NPY4R* and *TISP43*. *GPRIN2* (G protein regulated inducer of neurite outgrowth 2) has been shown to interact directly with G-coupled proteins (GNAO1 and GNAZ)[47] and has been implicated in the control of neurite outgrowth[48]. Our RNA-seq analysis points to localized expression in various regions of the brain, including the cerebellum and hypothalamus (Fig. 5). 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[49], and *NPY4R* (neuropeptide Y receptor Y4), a gene involved in energy homeostasis. Large CN variants of the region are associated with obesity[50] (Supplementary Discussion; Supplementary Fig. 20).

Although our analysis provides a framework for the evolution of new human-specific genes, there are a number of limitations. First, additional mutation events, such as interlocus gene conversion (IGC), frequently occur between high-identity paralogues[51–53] (Supplementary Fig. 19). We identified 2.9% of the sequence showing signatures of IGC, consistent with previous estimates[51] (Supplementary Table 7). Though such duplications will make HSDs appear evolutionarily 'younger', excluding these regions increases our timing estimates by only a small degree (on average an increase of 0.008 human–chimpanzee distance across 17 HSDs; Supplementary Table 8). 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 past 15 Myr. Breakpoints and boundaries

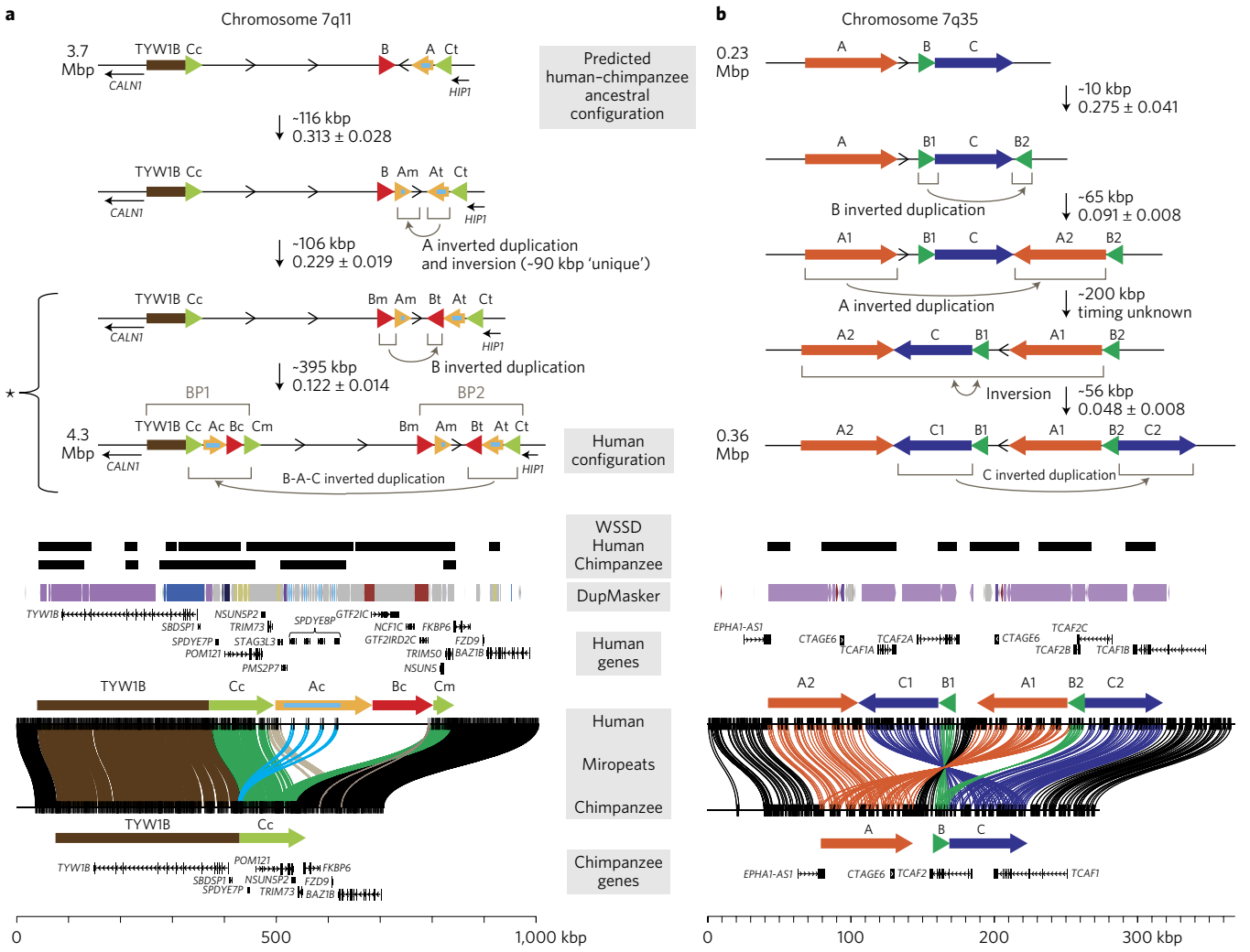

Figure 5 | Complex models of HSD evolutionary history. a,**b**, BACs tiling across human chromosome 7q11 (**a**) and 7q35 (**b**) regions were sequenced and assembled (representing human and additional great apes) and supercontigs were created. Estimates of sizes and evolutionary timing (human–chimpanzee distance; Supplementary Table 8) of events are denoted between each predicted intermediate genomic structure. SD organization is depicted as coloured arrows across the 7q11 (SDs annotated with subscripts representing relative positions, including centromeric (c), middle (m) and telomeric (t) as previously defined[41]) and 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 based on a chosen threshold (s), defined as the number of matching bases minus the number of mismatching bases (s = 500 for **a**, s = 1,000 for **b**) and match the arrow colours when they connect SD blocks. Additional annotations include whole-genome shotgun sequence detection (WSSD) in human and chimpanzee (indicating duplicated regions identified by sequence read depth[56]), DupMasker[79] and genes.

become challenging to delineate due to a series of overlapping complex rearrangements (for example, *SMN1* region on chromosome 5q13.3; Supplementary Discussion; Supplementary Fig. 21). Third, we assume that any individual with two copies of a gene family represents the ancestral non-duplicated state. It is possible that the alternative scenario of duplication followed by subsequent deletion of the ancestral paralogue may have occurred (Supplementary Discussion; Supplementary Figs 22 and 23; Supplementary Tables 22 and 23). Fourth, this study focuses on protein-encoding gene models 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 (Ensembl Variant Effect Predictor). Moreover, the canonical gene model being investigated in our analysis is heavily weighted by the ancestral intron–exon structure (Supplementary Discussion; Supplementary Fig. 24). Thus, new fusion genes and transcripts not previously annotated that have gained alternative promoters would not have been considered[36]. It is likely that long-read genome and transcriptome data will be required to explore such gene innovations[20].

Finally, although we focused on HSDs that had become fixed in the human population, it may be that some of the most CN polymorphic loci (or, additionally, loci that exist at <90% frequency in the population) are candidates for more recent adaptations between populations[30]. 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[54,55] 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[37]. Our CN 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 CNs, while Asians show the lowest with some humans showing no duplication of the region (Fig. 4). The model suggests that a single full-length paralogue of *TCAF1B* (predicted HSD duplicate paralogue) and *TCAF2A* (predicted ancestral paralogue) 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 paralogue after truncation of the ancestral paralogue, a mechanism we also suggest occurred for duplicate family *PTPN20* (Supplementary Figs 25 and 26). Although the function of the truncated duplicates awaits 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, with the potential effect of altering regulation of these genes important in cold sensation.

Methods

Characterization of HSD regions. HSD regions >5 kbp in length were initially identified by read-depth analysis of 236 human[30] and 86 NHP[34] whole-genome Illumina sequence data sets mapped against the human reference genome (GRCh37/hg19). We defined HSDs as regions with evidence of CN gain in >90% of all humans (>2.5 copies), but where >90% of all great apes did not harbour duplications of the locus (<2.8 copies). Previously uncharacterized HSDs were validated by whole-genome shotgun sequence detection (WSSD)[56] and whole-genome analysis comparison (WGAC)[57] methods (Supplementary Table 1), and comparison with a genome assembly of the CHM1 haploid hydatidiform mole (NCBI Assembly PacBioCHM1_r2_GenBank_08312015) using BLASR. A combination of BLAST[58], BLAT[59] and WGAC[57] methods were used to annotate HSD paralogous regions and identify duplication breakpoints. HSD clustering simulations were performed 10 million times using BEDTools shuffle (v2.23.0)[60], median midpoint distances to specific genomic features for each iteration of the simulation were calculated and comparisons of these distribution were made to the empirical values.

BAC clone sequencing. Large-insert clones from primate BAC libraries (CH17, CH251, CH276 and CH277) were sequenced using either capillary-based methods or single-molecule, real-time (SMRT) sequencing using Pacific Biosciences (PacBio) RSII P4C2 or P6C4 chemistry. Inserts were assembled using Quiver and hierarchical genome assembly process (HGAP) as described previously[61] (Supplementary Tables 3 and 6). Tiling paths of human BAC clones (CH17) were subjected to capillary ($N=205$) or PacBio ($N=76$) sequence and assembly, resulting in 47.2 Mbp of high-quality sequence from 224 BACs. Of these, 85 BACs were included in the most recent human reference assembly (GRCh38). Contig sequences not included in the human reference may be found in Supplementary Dataset 1. All NHP clones were subjected to SMRT sequence and assembly ($N=269$ clones).

Evolutionary analyses. Multiple sequence alignments were generated using MAFFT[62] (Supplementary Dataset 1), visualized for manual editing using Jalview[63], and phylogenetic analyses were performed using MEGA6[64] (Supplementary Dataset 2). Evolutionary timing of HSDs was estimated as a fraction of the human–chimpanzee branch length. IGC regions were identified by GENECONV[29], masked using BEDTools and timing estimates repeated with masked alignments. Duplication mechanisms were predicted using a combined approach of defining ancestral paralogues/configurations using genomic synteny taken from chimpanzee and/or orangutan and evolutionary timing estimates to predict the order of rearrangements.

CN genotyping. CN genotyping was performed from genome sequence data from 2,379 humans from the HGDP[30] and Phase 3 of the 1KG[31], 86 NHP individuals from the Great Ape Genome Project (including bonobo ($N=14$), chimpanzee ($N=23$), gorilla ($N=32$) and orangutan ($N=17$))[34], a Denisovan individual[33], a Neanderthal individual[32] and three archaic hominids[65,66]. CN variant genotypes were determined based on mrsFAST sequence alignment[67] and paralogue-specific read-depth (SUNK) mapping[19]. We used the V_{ST} statistic[35] (custom python script available at https://github.com/EichlerLab/vst_calc) to measure CN stratification between populations. In some cases, gene CNs were validated via fluorescence *in situ* hybridization (FISH) using fosmid clones performed on lymphoblast cell lines (Coriell Cell Repository, Camden, New Jersey) as described previously[68] (Supplementary Tables 12 and 13).

RNA-seq. GTEx RNA-seq data from different subtissues (dbGaP version phs000424.v3.p1) were used to analyse the expression of a set of representative transcripts from hg38 RefSeq annotation. We quantified relative levels of expression using an adjusted version of reads per kilobase of transcript per million mapped reads (RPKM) with reads intersecting unique genomic 30-mers of a canonical isoform (RefSeq) corresponding to each gene paralogue. Alternatively, we also applied the Sailfish[69] method version 0.63 with the default parameters and $k=20$.

MIP sequencing. Single-molecule MIPs ($N=1,105$, capturing 415 exonic regions of 30 gene families) designed using MIPgen[70] were phosphorylated, captured, barcoded and sequenced as previously described[71]. Variants were identified using FreeBayes (https://github.com/ekg/freebayes) with relaxed constraints allowing for reduced allele ratios (0.07) and annotated with the Ensembl Variant Effect Predictor[72] based on the canonical transcript for each gene. We sequenced a total of 1,096 MIPs from 6,719 individuals, including population controls from the 1KG, and cases and controls from the Simons Simplex Collection (SSC)[73], Autism Genetic Resource Exchange (AGRE)[74] and The Autism Simplex Collection (TASC)[75] cohorts.

Statistical analysis. We applied the Wilcoxon–Mann–Whitney test when comparing primary versus secondary HSD sizes and the Kruskal–Wallis rank sum test to assess size differences across three different evolutionary periods. We applied a Wilcoxon–Mann–Whitney test *post hoc* to identify the duplication period(s) with significant differences and adjusted for multiple comparisons using the Holm method. For paralogous gene expression comparisons, median RPKM values of annotated RefSeq transcripts were compared across all tissue types using a Wilcoxon signed rank test and a Bonferroni correction applied for multiple test comparisons. A one-tailed Fisher's exact test was used to compare frequency of HSD-exonic mutations in autism cases versus unaffected sibling controls and Bonferroni-corrected for multiple testing comparisons.

Human subjects. The 1KG and SSC cohorts included in this study did not meet the US federal definitions for human subjects research. All samples were publicly available or encoded, with no individual identifiers available to the study authors. The University of Washington institutional review board (IRB) approved the AGRE and TASC cohorts for human subjects research. All samples were collected at respective institutions after receiving informed consent and approval by the appropriate IRBs. There are no new health risks to participants.

Data availability. BAC sequencing data generated during the current study are available in GenBank with the primary accession numbers provided in Supplementary Tables 3 and 6. Targeted MIP sequencing data generated during this study are available from NCBI BioProject (1KG cohort, ID# PRJNA356308) and the National Database for Autism Research (autism cohorts, NDAR project number #431; doi:10.15154/1338620).

Received 8 August 2016; accepted 22 December 2016; published 17 February 2017

References

1. O'Bleness, M., Searles, V. B., Varki, A., Gagneux, P. & Sikela, J. M. Evolution of genetic and genomic features unique to the human lineage. *Nat. Rev. Genet.* **13,** 853–866 (2012).
2. Gallego Romero, I. *et al.* A panel of induced pluripotent stem cells from chimpanzees: a resource for comparative functional genomics. *eLife* **4,** e07103 (2015).
3. Khan, Z. *et al.* Primate transcript and protein expression levels evolve under compensatory selection pressures. *Science* **342,** 1100–1104 (2013).
4. McLean, C. Y. *et al.* Human-specific loss of regulatory DNA and the evolution of human-specific traits. *Nature* **471,** 216–219 (2011).
5. Prescott, S. L. *et al.* Enhancer divergence and *cis*-regulatory evolution in the human and chimp neural crest. *Cell* **163,** 68–83 (2015).
6. Vermunt, M. W. *et al.* Epigenomic annotation of gene regulatory alterations during evolution of the primate brain. *Nat. Neurosci.* **19,** 494–503 (2016).
7. 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 (2004).
8. Ohno, S. *Evolution by Gene Duplication* (Springer-Verlag, 1970).
9. 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 (2015).
10. Charrier, C. *et al.* Inhibition of SRGAP2 function by its human-specific paralogs induces neoteny during spine maturation. *Cell* **149,** 923–935 (2012).
11. Dennis, M. Y. *et al.* Evolution of human-specific neural *SRGAP2* genes by incomplete segmental duplication. *Cell* **149,** 912–922 (2012).
12. Florio, M. *et al.* Human-specific gene *ARHGAP11B* promotes basal progenitor amplification and neocortex expansion. *Science* **347,** 1465–1470 (2015).
13. Marques-Bonet, T. *et al.* A burst of segmental duplications in the genome of the African great ape ancestor. *Nature* **457,** 877–881 (2009).
14. Sudmant, P. H. *et al.* Evolution and diversity of copy number variation in the great ape lineage. *Genome Res.* **23,** 1373–1382 (2013).

15. Bailey, J. A. & Eichler, E. E. Primate segmental duplications: crucibles of evolution, diversity and disease. *Nat. Rev. Genet.* **7,** 552–564 (2006).
16. Fortna, A. *et al.* Lineage-specific gene duplication and loss in human and great ape evolution. *PLoS Biol.* **2,** E207 (2004).
17. Locke, D. P. *et al.* Large-scale variation among human and great ape genomes determined by array comparative genomic hybridization. *Genome Res.* **13,** 347–357 (2003).
18. Cheng, Z. *et al.* A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature* **437,** 88–93 (2005).
19. Sudmant, P. H. *et al.* Diversity of human copy number variation and multicopy genes. *Science* **330,** 641–646 (2010).
20. Chaisson, M. J. *et al.* Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517,** 608–611 (2015).
21. Eichler, E. E. Segmental duplications: what's missing, misassigned, and misassembled—and should we care? *Genome Res.* **11,** 653–656 (2001).
22. Fan, J. B. *et al.* Paternal origins of complete hydatidiform moles proven by whole genome single-nucleotide polymorphism haplotyping. *Genomics* **79,** 58–62 (2002).
23. Kajii, T. & Ohama, K. Androgenetic origin of hydatidiform mole. *Nature* **268,** 633–634 (1977).
24. Destouni, A. *et al.* Zygotes segregate entire parental genomes in distinct blastomere lineages causing cleavage-stage chimerism and mixoploidy. *Genome Res.* **26,** 567–578 (2016).
25. Itsara, A. *et al.* Population analysis of large copy number variants and hotspots of human genetic disease. *Am. J. Hum. Genet.* **84,** 148–161 (2009).
26. Antonacci, F. *et al.* Palindromic *GOLGA8* core duplicons promote chromosome 15q13.3 microdeletion and evolutionary instability. *Nat. Genet.* **46,** 1293–1302 (2014).
27. Jiang, Z. *et al.* Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution. *Nat. Genet.* **39,** 1361–1368 (2007).
28. Steinberg, K. M. *et al.* Structural diversity and African origin of the 17q21.31 inversion polymorphism. *Nat. Genet.* **44,** 872–880 (2012).
29. Sawyer, S. Statistical tests for detecting gene conversion. *Mol. Biol. Evol.* **6,** 526–538 (1989).
30. Sudmant, P. H. *et al.* Global diversity, population stratification, and selection of human copy-number variation. *Science* **349,** aab3761 (2015).
31. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526,** 75–81 (2015).
32. Prufer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505,** 43–49 (2014).
33. Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338,** 222–226 (2012).
34. Prado-Martinez, J. *et al.* Great ape genetic diversity and population history. *Nature* **499,** 471–475 (2013).
35. Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444,** 444–454 (2006).
36. Nuttle, X. *et al.* Emergence of a *Homo sapiens*-specific gene family and chromosome 16p11.2 CNV susceptibility. *Nature* **536,** 205–209 (2016).
37. 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 (2015).
38. The GTEx Consortium The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348,** 648–660 (2015).
39. Hiatt, J. B., Pritchard, C. C., Salipante, S. J., O'Roak, B. J. & Shendure, J. Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation. *Gen. Res.* **23,** 843–854 (2013).
40. 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 (2011).
41. 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. *Gen. Res.* **15,** 1179–1188 (2005).
42. Zhang, Q. & Su, B. Evolutionary origin and human-specific expansion of a cancer/testis antigen gene family. *Mol. Biol. Evol.* **31,** 2365–2375 (2014).
43. 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 (2007).
44. Carvalho, C. M. *et al.* Inverted genomic segments and complex triplication rearrangements are mediated by inverted repeats in the human genome. *Nat. Genet.* **43,** 1074–1081 (2011).
45. Park, S. S. *et al.* Structure and evolution of the Smith–Magenis syndrome repeat gene clusters, SMS-REPs. *Gen. Res.* **12,** 729–738 (2002).
46. 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 (2009).
47. Iida, N. & Kozasa, T. Identification and biochemical analysis of GRIN1 and GRIN2. *Methods Enzymol.* **390,** 475–483 (2004).
48. Chen, L. T., Gilman, A. G. & Kozasa, T. A candidate target for G protein action in brain. *J. Biol. Chem.* **274,** 26931–26938 (1999).
49. Bamford, R. N. *et al.* Loss-of-function mutations in the EGF-CFC gene *CFC1* are associated with human left-right laterality defects. *Nat. Genet.* **26,** 365–369 (2000).
50. 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 (2009).
51. Dumont, B. L. Interlocus gene conversion explains at least 2.7% of single nucleotide variants in human segmental duplications. *BMC Genomics* **16,** 456 (2015).
52. Fawcett, J. A. & Innan, H. The role of gene conversion in preserving rearrangement hotspots in the human genome. *Trends Genet.* **29,** 561–568 (2013).
53. Nuttle, X. *et al.* Rapid and accurate large-scale genotyping of duplicated genes and discovery of interlocus gene conversions. *Nat. Methods* **10,** 903–909 (2013).
54. Colburn, R. W. *et al.* Attenuated cold sensitivity in TRPM8 null mice. *Neuron* **54,** 379–386 (2007).
55. Bautista, D. M. *et al.* The menthol receptor TRPM8 is the principal detector of environmental cold. *Nature* **448,** 204–208 (2007).
56. Bailey, J. A. *et al.* Recent segmental duplications in the human genome. *Science* **297,** 1003–1007 (2002).
57. 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 Res.* **11,** 1005–1017 (2001).
58. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).
59. Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12,** 656–664 (2002).
60. Quinlan, A. R. BEDTools: the Swiss-Army tool for genome feature analysis. *Curr. Protoc. Bioinformatics* **47,** 11–34 (2014).
61. Huddleston, J. *et al.* Reconstructing complex regions of genomes using long-read sequencing technology. *Genome Res.* **24,** 688–696 (2014).
62. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30,** 772–780 (2013).
63. 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 (2009).
64. Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* **30,** 2725–2729 (2013).
65. Fu, Q. *et al.* Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* **514,** 445–449 (2014).
66. Lazaridis, I. *et al.* Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513,** 409–413 (2014).
67. Hach, F. *et al.* mrsFAST: a cache-oblivious algorithm for short-read mapping. *Nat. Methods* **7,** 576–577 (2010).
68. Antonacci, F. *et al.* A large and complex structural polymorphism at 16p12.1 underlies microdeletion disease risk. *Nat. Genet.* **42,** 745–750 (2010).
69. Patro, R., Mount, S. M. & Kingsford, C. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat. Biotechnol.* **32,** 462–464 (2014).
70. 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 (2014).
71. O'Roak, B. J. *et al.* Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. *Science* **338,** 1619–1622 (2012).
72. Cunningham, F. *et al.* Ensembl 2015. *Nucleic Acids Res.* **43,** D662–D669 (2015).
73. Fischbach, G. D. & Lord, C. The Simons Simplex Collection: a resource for identification of autism genetic risk factors. *Neuron* **68,** 192–195 (2010).
74. Geschwind, D. H. *et al.* The autism genetic resource exchange: a resource for the study of autism and related neuropsychiatric conditions. *Am. J. Hum. Genet.* **69,** 463–466 (2001).
75. Buxbaum, J. D. *et al.* The Autism Simplex Collection: an international, expertly phenotyped autism sample for genetic and phenotypic analyses. *Mol. Autism* **5,** 34 (2014).
76. Brunet, M. *et al.* New material of the earliest hominid from the Upper Miocene of Chad. *Nature* **434,** 752–755 (2005).
77. Brunet, M. *et al.* A new hominid from the Upper Miocene of Chad, Central Africa. *Nature* **418,** 145–151 (2002).
78. Vignaud, P. *et al.* Geology and palaeontology of the Upper Miocene Toros-Menalla hominid locality, Chad. *Nature* **418,** 152–155 (2002).
79. Jiang, Z., Hubley, R., Smit, A. & Eichler, E. E. DupMasker: a tool for annotating primate segmental duplications. *Genome Res.* **18,** 1362–1368 (2008).

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 with 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. We thank M. Kremitzki for technical support in generating sequence maps and contigs of BACs not currently included in the human reference sequence. For DNA samples used in MIP sequencing, we would like to thank the investigators and families participating in the 1KG Project, Autism Speaks, TASC, and 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 and E. Wijsman). Approved researchers can obtain the SSC population data set 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 US National Institutes of Health (NIH) grants from NINDS (R00NS083627 to M.Y.D.), NIMH (R01MH101221 to E.E.E.) and NHGRI (R01HG002385 to 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.

Author contributions

M.Y.D. and E.E.E. conceived and designed the experiments. M.Y.D., L.H., B.J.N., O.P., S.C., J.H., F.A., L.D., K.M. and C.E. performed the experiments. M.Y.D., L.H., B.J.N., O.P., S.C. and J.H. analysed data. M.Y.D., K.P., A.R., C.B., M.M., N.J. and K.H. provided technical support. H.A.F.S., X.N., T.A.L.G., R.K.W. and E.E.E. provided materials or analyses tools. M.Y.D., L.H., B.J.N., O.P., S.C. and E.E.E. wrote the paper.

Additional information

Supplementary information is available for this paper.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to E.E.E.

How to cite this article: Dennis, M. Y. *et al.* The evolution and population diversity of human-specific segmental duplications. *Nat. Ecol. Evol.* **1,** 0069 (2017).

Competing interests

E.E.E. is on the scientific advisory board (SAB) of DNAnexus, Inc., is a consultant for Kunming University of Science and Technology (KUST) as part of the 1000 China Talent Program (2014–2016), and was an SAB member of Pacific Biosciences, Inc. (2009–2013).