Supplementary Information

Palindromic *GOLGA8* core duplicons promote chromosome 15q13.3 microdeletion and evolutionary instability

Francesca Antonacci[1*], Megan Y. Dennis[2*], John Huddleston[2,3], Peter H. Sudmant[2], Karyn Meltz Steinberg[4], Jill A. Rosenfeld[5], Mattia Miroballo[1], Tina A. Graves[4], Laura Vives[2,3], Maika Malig[2], Laura Denman[2], Archana Raja[2,3], Andrew Stuart[6], Joyce Tang[6], Brenton Munson[2], Lisa G. Shaffer[5,7], Chris T. Amemiya[6], Richard K. Wilson[4], and Evan E. Eichler[2,3†]

*These authors contributed equally to this work.
[1]Dipartimento di Biologia, Università degli Studi di Bari "Aldo Moro", Bari 70125, Italy
[2]Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA
[3]Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195, USA
[4]The Genome Institute at Washington University, Washington University School of Medicine, St. Louis, MO 63108, USA
[5]Signature Genomic Laboratories, LLC, Spokane, WA 99207, USA
[6]Benaroya Research Institute at Virginia Mason, Seattle, WA 98101, USA
[7]Genetic Veterinary Sciences, Inc., Paw Print Genetics, Spokane, WA 99202, USA

Supplementary Note

Contents

I. CNPα and CNPβ copy number analysis

We explored the spectrum of copy number variation at this locus using whole-genome sequence from 2313 human, ape and archaic hominin genomes[1] (Supplementary Table 1). Sequence read-depth analysis[2] (Supplementary Table 2) identified two large copy number polymorphic (CNP) segmental duplications (SDs) of ~300 kbp (CNPα) and ~210 kbp (CNPβ) (Figure 1a; Table 1). CNPα has six copy number (CN) states (diploid CN of 2-7), with the most common CN state (CN = 4) having a frequency of 77% in individuals from different ethnicities (Supplementary Figure 1; Supplementary Table 2; Figure 1b; Table 1).

We designed a series of three-color FISH experiments to investigate copy number and location of CNPα among different individuals. FISH mapping showed that the variable sequences corresponding to CNPα map at both breakpoints of the 15q13.3 microdeletion, with a variable CN between 0 and 1 at BP4 and between 0 and 2 at BP5 (Figure 1c) (Supplementary Table 4). CNPβ has eight diploid CN states between 5 and 12 with the most common CN state (CN = 8) having a frequency of 72% (Figure 1b; Table 1; Supplementary Table 1). Using FISH cohybridization experiments, we also attempted to investigate the structure of CNPβ, but the complex organization of this SD (patchwork of smaller duplications) and the higher CN status of this region did not allow us to visualize a clear signal by FISH. Both CNPα and CNPβ show a CN decrease among African populations compared to Europeans and Asians (Supplementary Figure 1) consistent with lower CN representing the most likely ancestral state. This is supported by the sequence read-depth analysis on 86 great ape individuals showing a diploid CN of 2 of CNPα in all nonhuman primates analyzed compared to humans (Supplementary Figure 1; Supplementary Table 2). Similarly, CNPβ results in a lower diploid CN between 4 and 5 in chimpanzee and bonobo compared to humans but has a higher CN between 6 and 11 in most gorilla and orangutan individuals analyzed. In these two species, CNPβ might have undergone species-specific duplications.

II. Discovery of the β inversion

Since complex regions of SD are frequently misassembled or missing from the human reference genome[3-7], we established a new alternate reference assembly for this 2.6 Mbp region of 15q13.3. We generated a map of 23 contiguous BAC clones (CH17) derived from a hydatidiform (haploid) mole derived human cell line (CHM1hTERT)—devoid of allelic variation—based on end-sequence mapping against the human reference genome assembly (GRCh37/hg19). We discovered eight discordant clones mapping at the 15q13.3 proximal region suggesting the presence of an inversion corresponding to CNPβ at BP4 of the 15q13.3 microdeletion (Supplementary Figure 2a)[8].

We validated the existence of the 130 kbp inversion, which we will refer to as β inversion, after performing capillary or PacBio sequencing of 21 overlapping clones generating ~4 Mbp of high-quality sequence (Supplementary Table 5 and Supplementary Figure 3). To ensure the β inversion existed in humans and was not a CHM1hTERT cell line artifact, we identified a unique inversion-tagging haplotype consisting of eight variants by comparing CNPβ at BP4 and BP5 between the human reference and CH17 assemblies (Supplementary Table 6).

We discovered this CH17-inversion haplotype in the European individual NA12891 from 1000 Genomes Project sequencing data and validated that this individual was heterozygous for the BP4 β inversion by constructing and arraying a large-insert genomic BAC library (VMRC54), performing hybridizations, and sequencing (Nextera/Illumina and PacBio) BAC clones spanning BP4 CNPβ (Supplementary Table 7; Supplementary Table 9; Figure 2a,b; Supplementary Figure 3).

III. BAC library construction and screening

We constructed individual BAC libraries from each member of the NA12878 parent–child trio, namely: NA12878 (VMRC53), NA12891 (VMRC54), and NA12892 (VMRC57) (Supplementary Table 8). High molecular weight DNA was isolated, partially *EcoRI* digested, and subcloned into the pCC1BAC vector (Epicentre) to create >150 kbp insert libraries using previously described protocols[9]. The ligation products were then transformed into DH10B (T1 resistant) electro-competent cells (Invitrogen). Each library was arrayed into 528 (384-well) microtiter dishes and subsequently gridded onto ten 22x22 cm nylon high-density filters for probe hybridization screening such that each hybridization membrane consists of over 18,000 distinct BAC clones, stamped in duplicate. Library hybridization was carried out according to the protocol available at CHORI BACPAC resources (http://bacpac.chori.org/highdensity.htm).

IV. β inversion structure and breakpoint analysis

The β inversion consists of three SDs (Supplementary Figure 2b): a pair of two highly identical (58 kbp, 99.6% identity) inversely oriented SDs flanking a 95 kbp duplication. The flanking 58 kbp palindrome corresponds to the *GOLGA8* gene family, one of the core duplicons found to be associated with most of the interspersed SD blocks across chromosome 15[10-12]. In the CH17 assembly, the 95 kbp segment aligns in direct orientation across BP4 and BP5. In the reference genome assembly we

observe the same tripartite organization with the exception that the internal 95 kbp segment is in an inverse orientation when compared to the CH17 assembly. We investigated the orientation, length and percentage of sequence identity of all the SDs mapping at 15q13.3 BP4 and BP5 and showed the presence of six paralogous regions in direct orientation between BP4 and BP5 in the reference genome assembly ($H\alpha_2$ haplotype), with the largest directly oriented SD being 58 kbp in size with 99.4% identity (Supplementary Figure 4).

Interestingly, all these SDs contain the *GOLGA8* and *ULK4P3* core duplicons. Based on the CH17 alternate assembly, the β inversion increases the directly oriented genomic region to a total of ~188 kbp of near perfect sequence (99.4% identity) making it, in principle, a more predisposed haplotype to non-allelic homologous recombination (NAHR) (Supplementary Figure 4). For simplicity, we refer to the CH17 haplotype as $H\alpha_2\beta_{inv}$ because it carries the β inversion and two haploid copies of CNPα.

The β inversion is likely mediated by the flanking 58 kbp inverted SDs within the larger CNPβ at BP4 (Supplementary Figure 2b). To assess the precise breakpoints within these SDs, we used eight sequences from BACs spanning the proximal and distal 58 kbp inverted SDs within the larger CNPβ at BP4 representing different haplotypes (inverted and direct orientations) from NA12891 (β direct and inverse), CH17 (β inverse), and the human reference (GRCh37/hg19; β direct) to generate a multiple sequence alignment (MSA) (Figure 2c; Supplementary Figure 5).

We expect that within the β inversion the proximal-β-inverse (SD2) and distal-β-direct (SD3) SD sequences and, alternatively, the proximal-β-direct (SD1) and distal-β-inverse (SD4) SD sequences should match (highlighted orange in Figure 2c and Supplementary Figure 5), while at the β inversion breakpoint and beyond SD1 and SD2 and, alternatively, SD3 and SD4 sequences should match (highlighted yellow in Figure 2c and Supplementary Figure 5). We observed this breakpoint signature across a ~12 kbp region spanning from intron 2 of *GOLGA8* to 9.6 kbp upstream of the gene, corresponding to human reference (hg19/GRCh37) coordinates chr15:30,704,161-30,716,095 (proximal SD) and chr15:30,834,548-30,846,486 (distal SD) (Supplementary Table 10). Notably, across the MSA we observed alternative sequence patterns, for instance where SD2 and SD4 within the β inversion haplotype more closely matched, suggesting gene conversion had historically occurred across these paralogous SDs within this haplotype.

We generated a 58 kbp MSA from distinct β haplotypes from NA12891 (Beta_direct_proximal, Beta_direct_distal, Beta_inverse_proximal, Beta_inverse_distal). To identify signatures of interlocus gene conversion, the alignment was analyzed using the GENECONV program (version 1.81)[13].

GENECONV identifies sequence pairs with longer than expected tracks of 100% sequence identity given the overall pattern of sequence variation within the alignment. The program was run with the same parameters as in Dumont et al.[14]. Significance was set at $p<0.05$ for global and pairwise tracks. There were 12 tracks total identified using these parameters with an average length of 2624 bp (Supplementary Table 11). There are two very long tracks from the Beta_inverse_proximal and Beta_inverse_distal pair of sequences at the end of the alignment that are much larger than expected for interlocus gene conversion. If we remove these tracks, the average length of the tracks is reduced to 1158 bp. Additionally, six tracks overlapped exons of *GOLGA8* and may not represent gene conversion but rather functional constraint due to purifying selection. However, the beginning of the alignment shows likely signatures of historical gene conversion that complicate further refinement of the breakpoint.

V. γ inversion structure and breakpoint analysis

In order to resolve the duplication architecture of the larger 1.8 Mbp BP4-BP5 inversion that is common in the human population (referred to as the γ inversion)[8,15,16] (Figure 1a) at the sequence level, we constructed and arrayed a large-insert genomic BAC library (VMRC53) from NA12878 (Supplementary Table 8), a European individual from the 1000 Genomes Project previously discovered to be heterozygous for the γ inversion[16]. Since NA12878 is a diploid genome and is heterozygous for the γ inversion, we used publicly available single nucleotide polymorphisms (SNPs) for NA12891 (father) and NA12892 (mother) to assign 17/35 clones to the inverted haplotype (maternal) and 18/35 clones to the direct haplotype (paternal) (Supplementary Figure 6; Supplementary Table 9).

Sequence analysis showed that the paternal haplotype was identical to the reference genome assembly (Hα2). Analysis of the clones mapping to the maternal inverted haplotype showed clear evidence of an inversion between BP4 and BP5 (Hα1γinv) (Figure 3a; Supplementary Figure 2c). We identified three clones mapping at the breakpoints of the γ inversion that appeared to be discordant since their collinearity was interrupted based on large-clone insert mapping to the reference. To confirm this structural configuration, we PacBio sequenced 21 VMRC53 clones (Supplementary Table 9) and generated an alternate assembly across BP4 and BP5 representing the Hα1γinv inverted haplotype with 11 nonredundant BAC clones (Figure 3b; Supplementary Figure 7).

Comparing sequences across the $H\alpha_1\gamma_{inv}$ (NA12878) and $H\alpha_2$ (hg19/GRCh37) assemblies, we find that the γ inversion spans ~1.8 Mbp from BP4 to BP5 (Figure 3b) and is flanked by inverted SDs containing two *GOLGA8* genes and a *ULK4P3* gene, (~71 kbp, 98.5% identity; Supplementary Figure 1d). The $H\alpha_1\gamma_{inv}$ assembly contains a single copy of CNPα at BP4 and CNPβ at BP5, respectively, suggesting that the γ inversion arose from the simplest human haplotype and displaced CNPα (Supplementary Figure 8).

In the absence of sequence representing $H\alpha_1$ to refine the γ inversion breakpoints, we generated an MSA using sequences from the flanking SDs represented in $H\alpha_1\gamma_{inv}$ (SD2 and SD3) and the corresponding distal SDs (SD4) represented in $H\alpha_2$ and $H\alpha_2\beta_i$ (Supplementary Figure 9). We did not include sequences representing the proximal SD (SD1) from the $H\alpha_2$ haplotypes in our MSA because historical expansions of CNPα and CNPβ to BP4 likely muddled the breakpoint signatures within these SDs. We refined the γ inversion breakpoints using a similar strategy described for the β inversion (see above) to a ~32 kbp region spanning nearly the entire *ULK4P3* gene at chr15:30,858,018-30,889,492 (GRCh37/hg19, proximal SD) and chr15:32,702,231-32,733,993 (GRCh37/hg19, distal SD) (Supplementary Figure 9; Figure 3c; Supplementary Table 10). *ULK4P3* maps within two palindromic repeats containing the *GOLGA8* core duplicon (Figure 3c; Supplementary Figure 2d). In summary, we identified a total of five alternate structural configurations of the 15q13.3 region in humans ranging in size from 2 to 3 Mbp, with $H\alpha_1$ and $H\alpha_1\gamma_{inv}$ haplotypes showing the simplest organizations (Supplementary Figure 8).

VI. Population prevalence of β and γ inversion polymorphisms

To characterize the frequency of the CH17-identified β inversion haplotype in human populations, we designed molecular inversion probes (MIPs)[17,18] to capture, sequence (Illumina), and genotype the eight haplotype-tagging variants (described above) across 904 individuals from diverse human populations from the 1000 Genomes Project (Supplementary Tables 13 and 14).

MIP design, capture, and sequencing were performed as previously described[18,19]. Briefly, target genomic coordinates of variants and dbSNP132 polymorphisms were used to generate sequences for 70 bp oligonucleotides containing a common 30 bp linker flanked by an extension arm of 16 to 20 bp and a ligation arm of 20 to 24 bp that targets a specific 112 bp genomic region using an in-house MIP design pipeline (all relevant scripts can be found here: http://krishna.gs.washington.edu/mip_pipeline/). We chose a nonredundant subset of MIPs that were

predicted to be high performing and uniformly spaced across the 15q13.3 locus, with a higher density of MIPs within BP4 and BP5 flanking SDs (Supplementary Table 13). MIPs were synthesized (Integrated DNA Technologies), pooled at equimolar concentrations, phosphorylated, and used for multiplex capture with DNA samples (100 ng) for 24 hr. To degrade uncaptured linear DNA, reactions were subsequently treated with endonuclease I and III. The captured circular DNA was amplified using a universal forward primer and a unique barcoded reverse primer for each sample. We pooled equal amounts of 96 to 384 different libraries together and purified using AMPure XP beads (Beckman Coulter).

Paired-end 150 bp reads were generated using Illumina MiSeq v2 following manufacturer instructions, split into 36-mers, mapped with mrsFAST to the human CH17-derived 15q13.3 assembly, and counts of reads with edit distance 0 were calculated at singly unique nucleotide (SUN) k-mer (SUNK) positions using previously described methods[2].

In the analysis of cases with large-scale deletions across the 15q13.3 locus, we performed a series of filters and normalizations of read-depth at SUNK positions, including (1) a filter of MIPs that was insufficiently captured by removing SUNKs with a median read-depth less than 10; (2) normalization for input DNA differences by using average read-depth of SUNKs within "unique" CN two loci outside of predicted deletions (15q13:127,218-130,659 and 15q13:2,705,257-2,723,046); and (3) standardization for different MIP capture efficiencies by using the average read-depth per SUNK across 78 control individuals from the 1000 Genomes Project.

Remarkably, despite the fact that, to our knowledge, the β inversion has never before been reported, we estimated a haplotype frequency of ~38% across European populations (n=275, CEU, TSI, and GBR). Further, we detected this β inversion haplotype at reduced frequencies of ~10% in African populations (n=299, LWK, MKK, YRI, ESN, and GWD) and ~4% in Asian populations (n=221, CHB, CDX, KHV, and JPT), with no β inversion alleles identified in CDX and KHV. These data suggest haplotype frequency differences between Europeans and Asians (average Fst=0.277), with a maximum Fst of 0.363 between Toscani (TSI) and Chinese Dai (CDX) populations (Supplementary Table 16). Fst values were calculated as previously described[20]. Notably, additional β inversion haplotypes may exist in human populations not genotyped in this analysis.

VII. FISH analysis of the γ inversion

In order to understand the prevalence of the γ inversion ($H\alpha_1\gamma_{inv}$) across humans, we coupled next-generation sequencing with cytogenetic-based assays. We selected a total of 20 individuals with CN between 2 and 5 for CNPα, obtained lymphoblastoid cell lines, and tested them for the presence of the γ inversion using a three-color interphase FISH assay (two probes inside and one outside the inversion) (Supplementary Figure 10). FISH experiments were performed using fosmid clones directly labeled by nick-translation with Cy3-dUTP (Perkin-Elmer), Cy5-dUTP (Perkin-Elmer), and fluorescein-dUTP (Enzo) as described previously[6] 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 sulphate, 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, Cy3, Cy5, and fluorescein fluorescence signals, detected with specific filters, were recorded separately as gray-scale images. Pseudocoloring and merging of images were performed using Adobe Photoshop software. A minimum of 50 interphase cells were scored for each inversion to statistically determine the orientation of the examined region.

We found an enrichment of the γ inversion among chromosomes with one copy of CNPα (10/24 chromosomes) consistent with our BAC sequencing analysis of NA12878 (see above). Without exception, all chromosomes (n=16/16) with higher CN (n=2-3) were configured in a direct orientation similar to the reference genome (Supplementary Figure 10). We designed a series of three-color FISH experiments to further investigate the location of CNPα on direct and inverted chromosomes using one probe mapping to CNPα and two probes mapping within the unique region. FISH analysis showed that when the haploid CN of CNPα is one and it carries the γ inversion ($H\alpha_1\gamma_{inv}$ configuration) (Supplementary Figure 11), there is a "missing" copy of CNPα at BP5 with respect to the reference genome, showing that the inversion transported this paralog to BP4. CNPα varies between 1 and 3 copies in the directly oriented configurations ($H\alpha_1$, $H\alpha_2$, $H\alpha_3$ configurations) (Supplementary Figure 8). Here, the CN ranges between 0 and 1 at BP4 and between 1 and 2 copies at BP5 (Supplementary Table 4). Combining this cytogenetic inference with sequence read-depth for 1311 human genomes with ethnicities matching the FISH tested individuals, we estimate a γ inversion allele frequency of 6% (Supplementary Table 12).

VIII. Evolutionary origin

In order to investigate the ancestral configuration of the 15q13.3 region, we compared the orientation of the region in human with other nonhuman primate species. We tested for the presence of the 1.8 Mbp γ inversion by FISH analysis from three orangutan (*Pongo pygmaeus*), two gorilla (*Gorilla gorilla*), and four chimpanzee (*Pan troglodytes*) cell lines (Supplementary Table 17).

All individuals from chimpanzee and orangutan species were found to be in direct orientation, while all gorilla individuals were inverted when compared to the human reference genome orientation (Supplementary Figure 12). These data suggest that either the inversion occurred in the human-chimpanzee-gorilla ancestor and reverted to the direct orientation in chimpanzee or that the γ inversion occurred independently in human and gorilla lineages.

To resolve the status of the smaller β inversion, the evolutionary history of the γ inversion, and the SD architecture in nonhuman primates, we selected chimpanzee, gorilla and orangutan BAC clones (from the CH251, CH277 and CH276 BAC libraries, respectively) (Supplementary Tables 17 and 18), based on end-sequence mapping against the GRCh37 human reference genome assembly (BAC end sequences were downloaded from the NCBI trace repository (http://www.ncbi.nlm.nih.gov/Traces)). We sequenced them using Illumina HiSeq 2000 (101 bp PE reads) and aligned 48 sequenced clones to the chromosome 15q13.3 region of human reference assembly (Supplementary Figure 13). Of these clones, 24 were selected for high-quality sequencing and *de novo* assembly using either capillary or PacBio long-read sequences and compared to the human reference genome using Miropeats[21] (Supplementary Figure 14; Supplementary Table 18).

Overall, our sequence analysis reveals a much simpler organization of the 15q13.3 region in nonhuman primates when compared to human (Figure 4). The sequenced chimpanzee, gorilla and orangutan haplotypes, for example, lack almost all of the duplications present at most human BP4 haplotypes, with the exception of the *GOLGA8* repeats (Supplementary Figure 14). This pattern is confirmed based on read-depth analysis of whole-genome sequence from chimpanzee (*Pan troglodytes*), gorilla (*Gorilla gorilla*), orangutan (*Pongo pygmaeus*), and macaque (*Macaca mulatta*) individuals[1,22] (Supplementary Figure 15). Moreover, our sequence analysis of primate clones shows that BP5 was the ancestral source for most of the duplications (Figure 4; Supplementary Figure 14).

In contrast to humans, our analysis predicts that CNPβ maps in an inverted orientation in chimpanzee

at BP5 and the *GOLGA8* repeats define the boundaries of this evolutionary inversion (Figure 4; Supplementary Figures 13 and 14). Interestingly, the proximal copy of CNPβ was found to be inverted at BP4 in $H\alpha_2\beta_i$ human haplotypes (Supplementary Figure 1a,b). The BP5 β inversion is likely a recurrent event that occurred independently on the chimpanzee lineage due to the presence of the *GOLGA8* repeats at the breakpoints of the inversion. Gorilla, orangutan, and human have BP5 β in the direct orientation suggesting this is the great ape ancestral configuration. An inversion of the distal portion of α is found in gorilla at BP5. This portion of α is partially duplicated at BP4 in gorilla, and in both instances the rearrangement (duplication at BP4 and inversion at BP5) is flanked by the *GOLGA8* repeats (Figure 4; Supplementary Figure 14). Analysis of a gorilla and an orangutan clone mapping at BP5 shows that the 58 kbp repeat found at the breakpoints of the β and γ inversions in humans seems to have a simpler organization in these two primate species since it is missing *ULK4P3* flanking *GOLGA8* in humans (Supplementary Figure 14). The fact that the γ inversion in humans maps within *ULK4P3*, and that this gene is deleted in gorilla, suggests that that the γ inversion happened as two independent events in human and gorilla lineages.

Finally, primate clone sequencing also suggests that the *CHRNA7*-adjacent SD (purple block with orange arrow in Figure 4) is completely missing at BP4 in all analyzed primates with the exception of a partial duplication in gorilla. The distal breakpoint of the *CHRNA7*-adjacent duplication at BP4 in humans maps within a *GOLGA8* repeat.

Using 30 *GOLGA8* sequences obtained from human, chimpanzee, gorilla and orangutan BAC-sequenced clones (Supplementary Table 19), we generated an MSA of 13.7 kbp and constructed a neighbor-joining phylogenetic tree (Supplementary Figure 16a). Of note, when we pull out the putative coding regions of these *GOLGA8* genes across species, 23 out of 29 homologs maintain an open reading frame (Supplementary Table 19), suggesting, despite the large-scale amplification of these genes in great apes, they may be under selective constraint due to function. Only four *GOLGA8* copies can be identified unambiguously as orthologous both by phylogenetic analysis and with respect to their position (indicated with a red arrow in Supplementary Figure 16b). These four copies are predicted not to encode a functional protein in human, chimpanzee, and orangutan. They are located at the most distal end of BP4 (red arrow in Supplementary Figure 16b) and also represent the most ancestral copies of *GOLGA8*.

Our primate analysis suggests that the 15q13.3 region has become increasingly complex during human evolution, with an increase in size from 1.8-1.9 Mbp in nonhuman primates to >3 Mbp in humans. In order to estimate the order and timing of these structural changes during the evolution, we generated a

series of MSAs corresponding to human and chimpanzee haplotypes, constructed phylogenetic trees, and estimated the coalescence time of the haplotypes using locally calibrated molecular clocks.

To study the evolutionary history of the γ inversion, we generated an MSA of a 154 kbp unique region (GRCh37, chr15:32,290,893-32,445,407) of representative haplotypes from human and chimpanzee using Clustal W[23]. We generated Illumina sequences from NA12878 BAC clones that had been resolved into either a γ inverted or directly configured haplotype as previously described. In addition, we generated Illumina sequences from chimpanzee BAC clones from library CH251. We then aligned these clone sequences to the reference using BWA[24] and called SNPs using SAMtools[25]. We required a minimum depth of coverage equal to 30 at a variant site to eliminate false calls due to misalignment of reads. The phased variant calls from the 1000 Genomes Project individuals, NA19107 (homozygous direct $H\alpha_1$) and NA20786 (homozygous inverted $H\alpha_1\gamma_{inv}$), were obtained to consider a diversity of human haplotypes. These haplotypes were then aligned to the CH17 contig generated from Sanger sequence. Using Kimura 2-parameter distances, we constructed an unrooted neighbor-joining tree using MEGA5 with 500 bootstrap replications[26] (Supplementary Figure 21; Figure 5).

We calculated the number of mutations per branch by multiplying the branch length by the number of total sites (154,471). We then calculated Kimura 2-parameter genetic distances and standard errors using MEGA5. Using the average sequence divergence of human versus chimpanzee distance (K), we calculated the average substitution rate using the equation R=K/2T assuming a chimpanzee–human divergence time (T) of 6 million years ago (mya). We then estimated the coalescence time of the haplotypes using the equation T=K/2R. We estimate that the time to most recent common ancestor (TMRCA) of the γ inverted (NA12878 $H\alpha_1\gamma_{inv}$) and direct (NA19107 $H\alpha_1$ - Hap 1) haplotypes is 578.8 ± 47 thousand years (Figure 5).

Similarly, to estimate the timing of the expansion and inversions of the CNPα and CNPβ SDs, we generated an MSA of homologous regions using sequence from the human reference ($H\alpha_2$), human CH17 contig ($H\alpha_2\beta_{inv}$), chimpanzee and gorilla. Notably, we excluded all regions containing the *GOLGA8* core duplicons on the edges of SDs that likely mediated the rearrangements and show signatures of interlocus gene conversion. We generated a 131 kbp alignment for CNPα using CH17 contig, human reference (GRCh37, chr15:30,526,580-30,650,226 (BP4) and chr15:32,464,957-32,596,497 (BP5)), a chimpanzee BAC clone (CH251-231C18), and a gorilla BAC clone (CH277-223A7) spanning the orthologous CNPα, and estimated genetic distances. We estimated the coalescence time of the haplotypes and predict the CNPα duplication from BP5 to BP4 to have occurred 995,000 ± 61,000 years ago (Supplementary Figure 20a; Figure 5).

We also generated a 33 kbp MSA of CNPβ using sequence from the human reference (GRCh37, chr15:30,787,734-30,822,007 (BP4) and chr15:32770029-32803919 (BP5)), the CH17 contig, a chimpanzee BAC clone (CH251-15E3), and a gorilla BAC clone (CH277-203H19) spanning the orthologous CNPβ. From this, we estimate CNPβ to have duplicated from BP5 to BP4 862,000 ± 99,000 years ago, similar to the time estimated for CNPα suggesting that the two segments duplicated as one unit (Supplementary Figure 20b; Figure 5).

Further, by comparing sequences between the human CH17 contig ($H\alpha_2\beta_{inv}$) and human reference ($H\alpha_2$), we predict that the haplotype carrying the β inversion to have occurred shortly thereafter 748,000 ± 92,000 years ago (Figure 5). Similarly, to estimate the timing of the human-specific *ARHGAP11* duplication from BP5 (*ARHGAP11A*) to BP4 (*ARHGAP11B*), we generated a 24.7 kbp MSA of the homologous regions using sequence from the human CH17 contig ($H\alpha_2\beta_{inv}$), a chimpanzee BAC clone (CH251-15E3), and an orangutan BAC clone (CH276-149G15) spanning the orthologous *ARHGAP11A*. We predict the duplication from the BP5 to BP4 to have occurred 5.28 ± 0.48 mya (Supplementary Figure 19). Interestingly, the proximal breakpoint of the *ARHGAP11* duplication maps within a *GOLGA8* repeat (Supplementary Figure 17).

To study the evolutionary history of the *CHRNA7*-adjacent SD, we generated a 15 kbp MSA within the duplicated region using human CH17 contig and orthologous chimpanzee (CH251-202M12), gorilla (gorGor3, chr15:10,014,983-10,029,822), and orangutan (ponabe2, chr15:27,538,748-27,554,902) sequences. We calculated the number of mutations per branch by multiplying the branch length by the number of total sites. We then calculated Kimura 2-parameter genetic distances and standard errors using MEGA5. Using the average sequence divergence of human versus orangutan distance (K), we calculated the average substitution rate using the equation R=K/2T assuming an orangutan-human divergence time (T) of 15 mya. We then estimated the coalescence time of the haplotypes using the equation T=K/2R. From this, we predict the duplication from the distal to the proximal 15q13.3 region to have occurred 12.16 ± 0.58 mya before the divergence between human, chimpanzee and gorilla (Supplementary Figure 18; Figure 5).

Our primate clone sequencing efforts confirm that the distal duplication represents the ancestral copy. The duplication to the proximal region occurred after the split of orangutan from the African great ape lineage and the proximal copy was then completely deleted in the chimpanzee lineage and partially deleted in gorilla (Figure 4).

IX. 15q13.3 microdeletions patient breakpoint analysis

We analyzed 80 total DNA samples from children with autism, intellectual disability, and/or developmental delay that were previously identified as carrying 15q13.3 microdeletions by array comparative genomic hybridization (CGH). These include 77 cases with intellectual disability and developmental delay referred to Signature Genomic Laboratories (24 unpublished and 53 described in Cooper *et al.*, 2011)[27] and three cases with idiopathic autism from the Simons Simplex Collection (SSC)[28]. We screened the 80 patients using a customized microarray targeted to the 15q13.3 region and mapped the breakpoints of the disease-critical region to a ~500 kbp region spanned by the CNPα and CNPβ SDs (see separate Supplementary Figure 22 file; Supplementary Table 20). We delineated patients into two groups: (1) a larger subset harboring 1.5 Mbp deletions appearing to break within CNPβ (n=73; 15q13 deletion-1) and (2) a smaller subset harboring larger 2 Mbp deletions appearing to break within CNPα (n=7; 15q13 deletion-2) (Supplementary Figure 23). However, this differentiation is likely an artifact of the structural differences due to α and β CN variation and not the result of two different classes of breakpoints.

In order to refine the breakpoints, we performed whole-genome sequencing of two idiopathic autism patients from the SSC carrying *de novo* 15q13.3 microdeletions along with their unaffected parents using the Illumina HiSeq 2000 (101 bp PE reads) (Supplementary Table 22). The generated sequences were aligned to the GRCh37 human reference and the alternate $H\alpha_2\beta_{inv}$ reference that we constructed using CH17 hydatidiform mole clones. We investigated paralog-specific read-depth over 1 bp windows in each trio at all sites where both parents had the expected CN of 2. Using SUN variants that allowed us to discriminate between the paralogous copies[2], we narrowed proband 13647.p1 breakpoints to a 14 kbp segment at BP4 (chr15:30,907,142-30,920,936) and 22 kbp at BP5 (chr15:32,886,788-32,909,278), and proband 13301.p1 breakpoints to a 155 kbp segment at BP4 (chr15:30,615,000-30,770,000) and 30 kbp at BP5 (chr15:32,754,000-32,784,000) (Figure 6; Supplementary Table 10). The two probands have different breakpoints but in both cases the breakpoints map at or adjacent to directly oriented copies of *GOLGA8* (Figure 7; Supplementary Table 10).

Breakpoints of structural variants in the 15q13.3 region seem to occur closer to *GOLGA8* repeats than expected by chance. We tested by simulation to determine if the apparent clustering of evolutionary and disease breakpoints within or near *GOLGA8* sequences was significant. We identified the positions of all *GOLGA8* sequences (Supplementary Table 23) within the BP4 and BP5 regions and

created a null model by randomly distributing the breakpoint intervals to the SDs mapping to this portion of 15q13.3 (chr15:30,362,914-31,196,467 and chr15:32,442,314-32,927,877) (Supplementary Table 24). For each observed breakpoint, we calculated the distance to its nearest *GOLGA8* repeat as the sum of the distances from its left edge to the *GOLGA8* gene's left edge and its right edge to the *GOLGA8* gene's right edge. We computed the number of times the mean distance of sampled breakpoints from the null distribution was less than or equal to the mean of the observed distances between 15q13.3 breakpoints and *GOLGA8* repeats (66,801 bp). The results suggest that the clustering of breakpoints with *GOLGA8* sequences is significant (p=0.002, n=100,000 permutations).

In order to assess both the β inversion frequency and microdeletion breakpoints in a larger cohort of patients, we MIP captured and sequenced the β inversion diagnostic variants (n=8; described above) and SUNs dispersed across the 15q13.3 region (n=235; Supplementary Table 13) in patients with the 15q13.3 microdeletion. We genotyped the frequency of the β inversion in a subset of patients understanding that BP4 CNPβ may be included in some 15q13.3 microdeletions (Supplementary Table 15). From this analysis, we detected the β inversion at a frequency of ~22% (n=67) in 15q13.3 microdeletion patients. Assessing only patients with known European ancestry, the frequency of the β inversion is essentially unchanged [~28% (n=40)] and reduced compared to frequencies observed in other European cohorts (~38%) suggesting that CNPβ may be deleted in a subset of patients.

The β inversion in $H\alpha_2\beta_{inv}$ configurations increases the directly oriented genomic region by ~100 kbp, for a total of 188 kbp highly identical SDs in direct orientation mapping at the breakpoints of the 15q13.3 microdeletion region. This configuration would create, in principle, a haplotype that is much more robust to 15q13.3 microdeletion by NAHR. Nevertheless, we do not observe an enrichment of the $H\alpha_2\beta_{inv}$ in our 15q13 deletion-1 patient cohort compared to comparable population controls indicating that this haplotype does not predispose to microdeletions. Instead, using whole-genome sequencing of two autism patients with *de novo* 15q13.3 microdeletions as well as targeted MIPs sequencing data of an additional 76 patients with developmental delay, we showed that all the microdeletion breakpoints map to the *GOLGA8* core duplicons within the 58 kbp inverted repeats also found at the breakpoints of the β and γ inversions. This suggests that, despite the longer extent of homology within CNPβ at BP4 and BP5 within the $H\alpha_2\beta_{inv}$ haplotype, the shorter *GOLGA8* duplicons are more susceptible to NAHR mechanisms perhaps due to the intrinsic nature of their sequence element itself.

We also attempted to use MIP sequencing data to make breakpoint predictions at the sequence level in these microdeletion patients. By capturing sequences across the 15q13.3 locus and quantifying overall

read-depth of SUNs (after normalization against 78 controls from the 1000 Genomes Project), we attempted to refine breakpoints across individual patients (see Supplementary Figures 22 and 23; Supplementary Table 21). To reduce the amount of noise observed per individual, we averaged read-depth across SUNs within distinct deletion subgroups. From this analysis, we refined breakpoints to the same BP4 and BP5 regions predicted by array CGH. Overall, this novel MIP approach represents a cheaper diagnostic tool in alternative to array CGH.

X. 15q24 and 15q25 microdeletions analysis

The *GOLGA* palindromic repeat is present in other regions of chromosome 15 associated with disease, such as the Prader-Willi/Angelman syndromes, 15q24 microdeletions, and 15q25.2 microdeletions[29-32] (Supplementary Figure 24a,b). Analysis of the SDs mapping at the 15q24 and 15q25.2 microdeletion regions shows that the paralogous SDs mapping at the microdeletion breakpoints are short and have low percentage of identity (Supplementary Figure 24c; Supplementary Figure 25), but they all contain several copies of the *GOLGA* repeat, suggesting that the same specific mechanism might underlie these genomic rearrangements.

This is supported by array CGH experiments on ten previously published 15q24 microdeletion cases showing that the *GOLGA* repeat is mapping at all the rearrangements breakpoints (Supplementary Figure 26)[29,33].

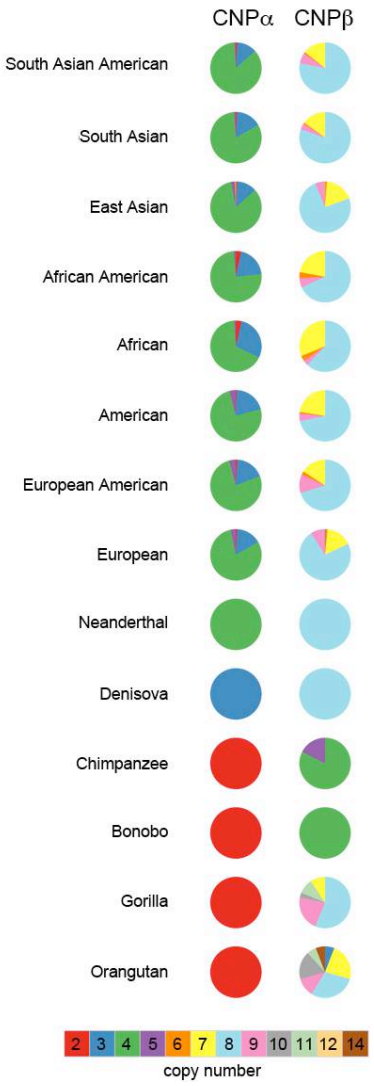
References

1. Sudmant, P.H. *et al.* Evolution and diversity of copy number variation in the great ape lineage. *Genome Res.* **23**, 1373-82 (2013).
2. Sudmant, P.H. *et al.* Diversity of human copy number variation and multicopy genes. *Science* **330**, 641-6 (2010).
3. Dennis, M.Y. *et al.* Evolution of human-specific neural *SRGAP2* genes by incomplete segmental duplication. *Cell* **149**, 912-22 (2012).
4. Itsara, A. *et al.* Resolving the breakpoints of the 17q21.31 microdeletion syndrome with next-generation sequencing. *Am. J. Hum. Genet.* **90**, 599-613 (2012).
5. She, X. *et al.* Shotgun sequence assembly and recent segmental duplications within the human genome. *Nature* **431**, 927-30 (2004).
6. Antonacci, F. *et al.* A large and complex structural polymorphism at 16p12.1 underlies microdeletion disease risk. *Nat. Genet.* **42**, 745-50 (2010).
7. Genovese, G. *et al.* Using population admixture to help complete maps of the human genome. *Nat. Genet.* **45**, 406-14, 414e1-2 (2013).
8. Sharp, A.J. *et al.* A recurrent 15q13.3 microdeletion syndrome associated with mental retardation and seizures. *Nat. Genet.* **40**, 322-8 (2008).
9. Smith, J.J., Stuart, A.B., Sauka-Spengler, T., Clifton, S.W. & Amemiya, C.T. Development and analysis of a germline BAC resource for the sea lamprey, a vertebrate that undergoes substantial chromatin diminution. *Chromosoma* **119**, 381-9 (2010).
10. Jiang, Z. *et al.* Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution. *Nat. Genet.* **39**, 1361-8 (2007).
11. Zody, M.C. *et al.* Analysis of the DNA sequence and duplication history of human chromosome 15. *Nature* **440**, 671-5 (2006).
12. Pujana, M.A. *et al.* Additional complexity on human chromosome 15q: identification of a set of newly recognized duplicons (LCR15) on 15q11-q13, 15q24, and 15q26. *Genome Res.* **11**, 98-111 (2001).
13. Sawyer, S. Statistical tests for detecting gene conversion. *Mol. Biol. Evol.* **6**, 526-38 (1989).
14. Dumont, B.L. & Eichler, E.E. Signals of historical interlocus gene conversion in human segmental duplications. *PLOS One* **8**, e75949 (2013).
15. Kidd, J.M. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56-64 (2008).
16. Antonacci, F. *et al.* Characterization of six human disease-associated inversion polymorphisms. *Hum. Mol. Gen.* **18**, 2555-66 (2009).
17. Hardenbol, P. *et al.* Multiplexed genotyping with sequence-tagged molecular inversion probes. *Nat. Biotechnol.* **21**, 673-8 (2003).
18. O'Roak, B.J. *et al.* Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. *Science* **338**, 1619-22 (2012).
19. Nuttle, X. *et al.* Rapid and accurate large-scale genotyping of duplicated genes and discovery of interlocus gene conversions. *Nature Methods* **10**, 903-9 (2013).
20. Akey, J.M., Zhang, G., Zhang, K., Jin, L. & Shriver, M.D. Interrogating a high-density SNP map for signatures of natural selection. *Genome Res.* **12**, 1805-14 (2002).
21. Parsons, J.D. Miropeats: graphical DNA sequence comparisons. *Comput Appl Biosci* **11**, 615-9 (1995).
22. Marques-Bonet, T. *et al.* A burst of segmental duplications in the genome of the African great ape ancestor. *Nature* **457**, 877-81 (2009).
23. Larkin, M.A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947-8 (2007).
24. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-60 (2009).
25. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-9 (2009).

26. Tamura, K., Dudley, J., Nei, M. & Kumar, S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol. Biol. Evol.* **24**, 1596-9 (2007).
27. Cooper, G.M. *et al.* A copy number variation morbidity map of developmental delay. *Nat. Genet.* **43**, 838-46 (2011).
28. Girirajan, S. *et al.* Refinement and discovery of new hotspots of copy-number variation associated with autism spectrum disorder. *Am. J. Hum. Genet.* **92**, 221-37 (2013).
29. Mefford, H.C. *et al.* Further clinical and molecular delineation of the 15q24 microdeletion syndrome. *J. Med. Genet.* **49**, 110-8 (2012).
30. Wat, M.J. *et al.* Recurrent microdeletions of 15q25.2 are associated with increased risk of congenital diaphragmatic hernia, cognitive deficits and possibly Diamond--Blackfan anaemia. *J. Med. Genet.* **47**, 777-81 (2010).
31. Amos-Landgraf, J.M. *et al.* Chromosome breakage in the Prader-Willi and Angelman syndromes involves recombination between large, transcribed repeats at proximal and distal breakpoints. *Am. J. Hum. Genet.* **65**, 370-86 (1999).
32. El-Hattab, A.W. *et al.* Redefined genomic architecture in 15q24 directed by patient deletion/duplication breakpoint mapping. *Hum. Genet.* **126**, 589-602 (2009).
33. Sharp, A.J. *et al.* Characterization of a recurrent 15q24 microdeletion syndrome. *Hum. Mol. Genet.* **16**, 567-72 (2007).

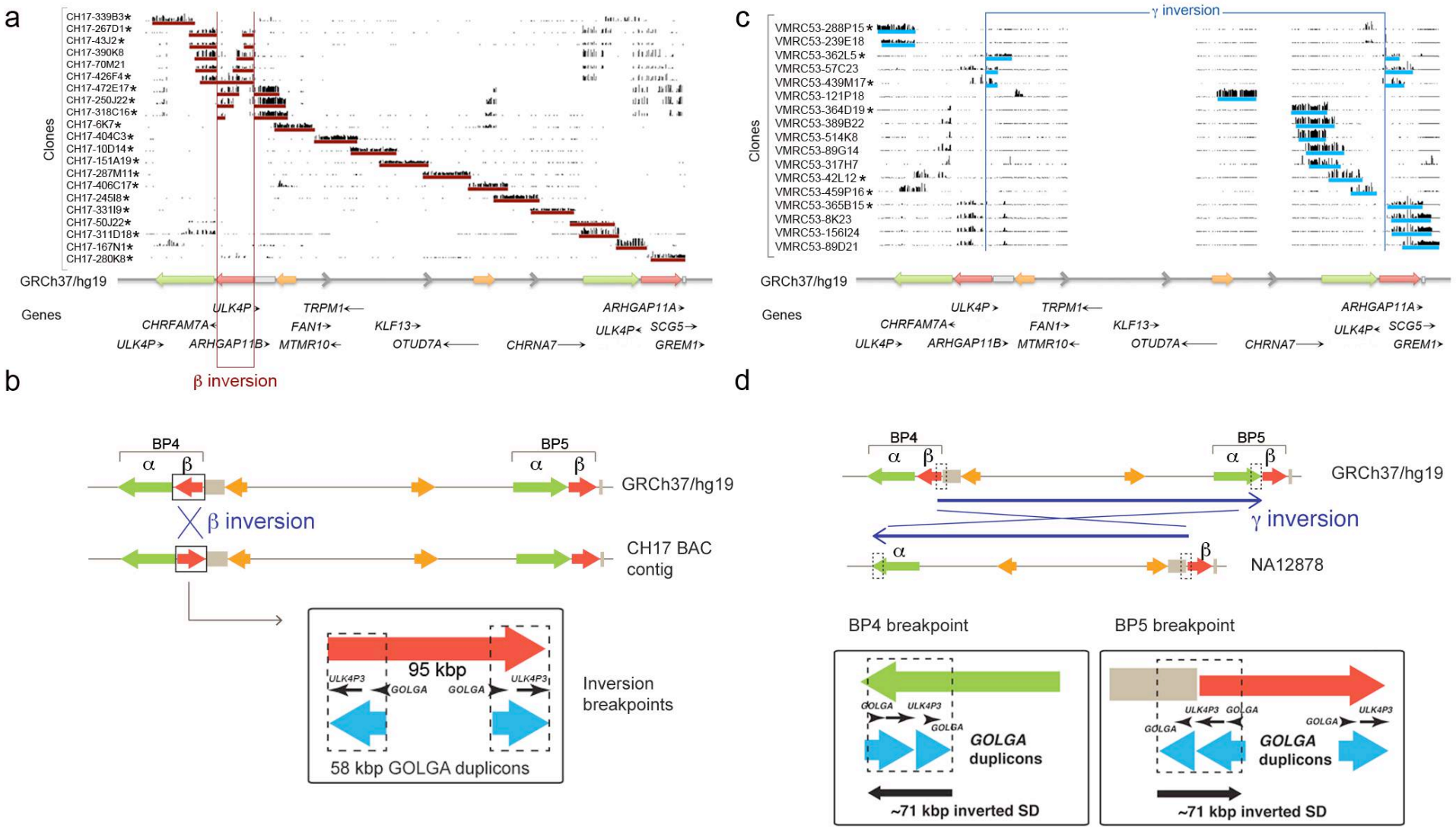
Supplementary Figures

Supplementary Figure 1



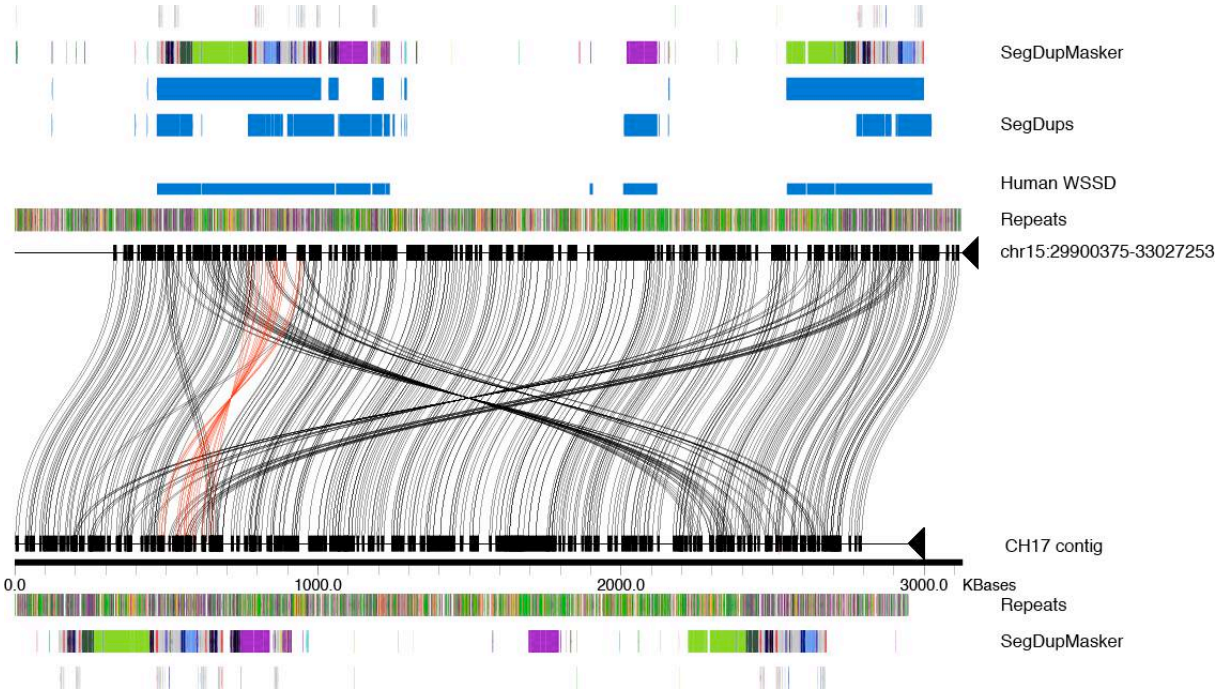
Supplementary Figure 1. Read-depth-based CN estimates of CNPα and CNPβ. The pie charts indicate the frequency of CNPα and CNPβ CN states in 2225 HapMap individuals from the 1000 Genome Project and 86 nonhuman primates, Neanderthal and Denisova genomes. The diploid CN is estimated based on sequence read-depth for each SD separately. The greatest differentiation ($Vst_{ASN:EUR}$=0.008, $Vst_{AFR:EUR}$ 0.035, $Vst_{ASN:AFR}$=0.071) is seen between African and non-African human populations (Supplementary Table 3).

Supplementary Figure 2



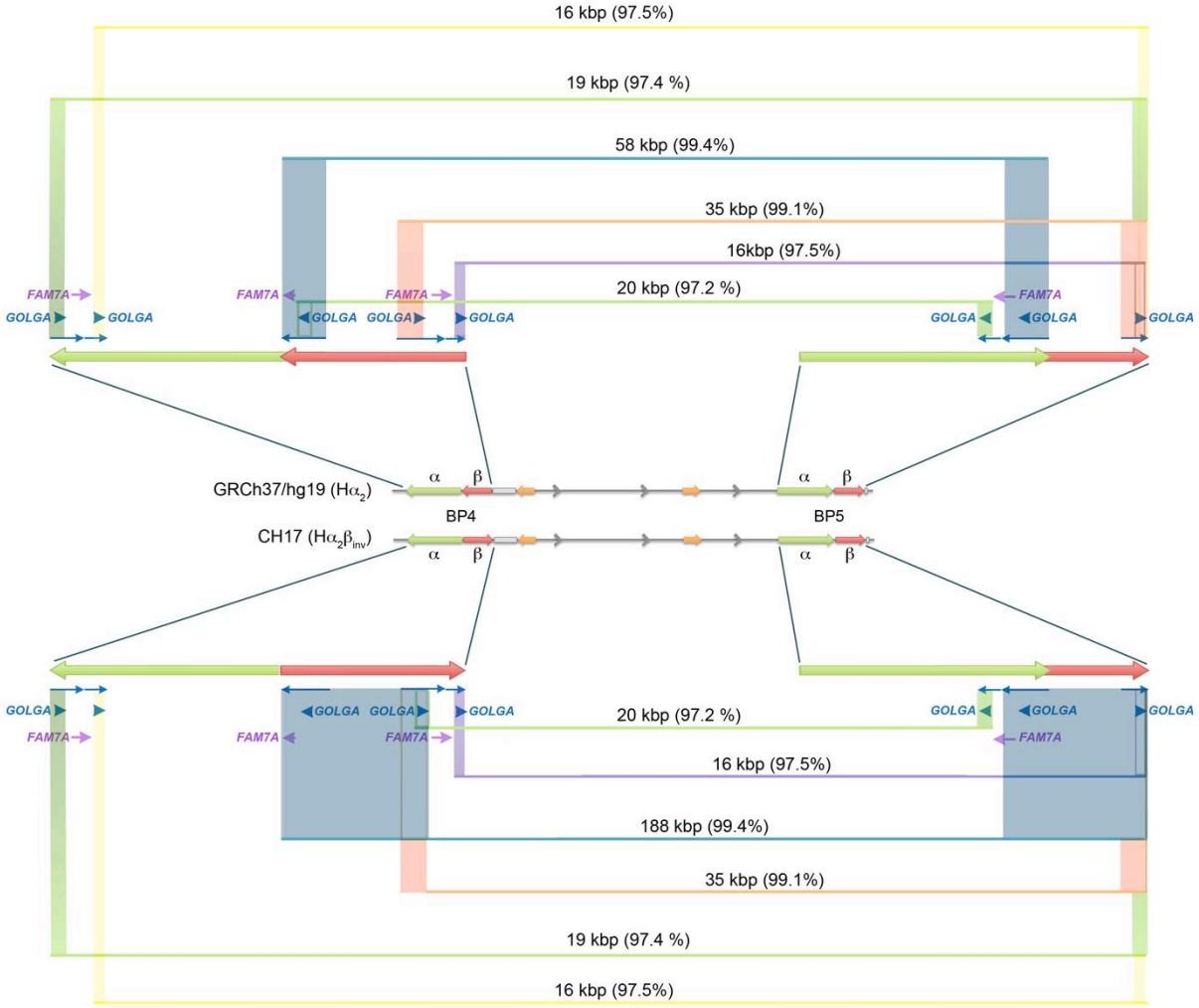
Supplementary Figure 2. Sequence reconstruction of the H$\alpha_2\beta_{inv}$ haplotype. (a) Next-generation sequencing of CHORI-17 BAC clones from a hydatidiform mole resource shows the presence of a 130 kbp inversion (b inversion) corresponding to BP4 of the 15q13.3 microdeletion. Clones sequenced using capillary or PacBio technology are indicated with asterisks. **(b)** The structure represented in the GRCh37 genome assembly is compared with the organization of the region as found in the CH17 BAC contig. The only difference between the two haplotypes is the inversion of a 130 kbp region (b inversion) consisting of three SDs: a pair of two highly identical (58 kbp, 99.6% identity) inversely oriented SDs (blue) flanking a 95 kbp duplication (red). The flanking 58 kbp palindrome corresponds to the *GOLGA8* and *ULK4P3* core duplicons. **(c)** Next-generation sequencing of VMRC53 BAC clones from NA12878 shows the presence of a 1.8 Mbp inversion (g inversion) across the 15q13.3 region. Clones sequenced using PacBio technology are indicated with asterisks. **(d)** The 15q13.3 region in the GRCh37 genome assembly is compared with the organization as found in NA12878 VMRC53 BAC clones. The breakpoints of the g inversion map to two inverted SDs containing the *GOLGA8* repeat, aligned in inverted orientation (~71 kbp, 98.5% identity) across BP4 and BP5.

Supplementary Figure 3



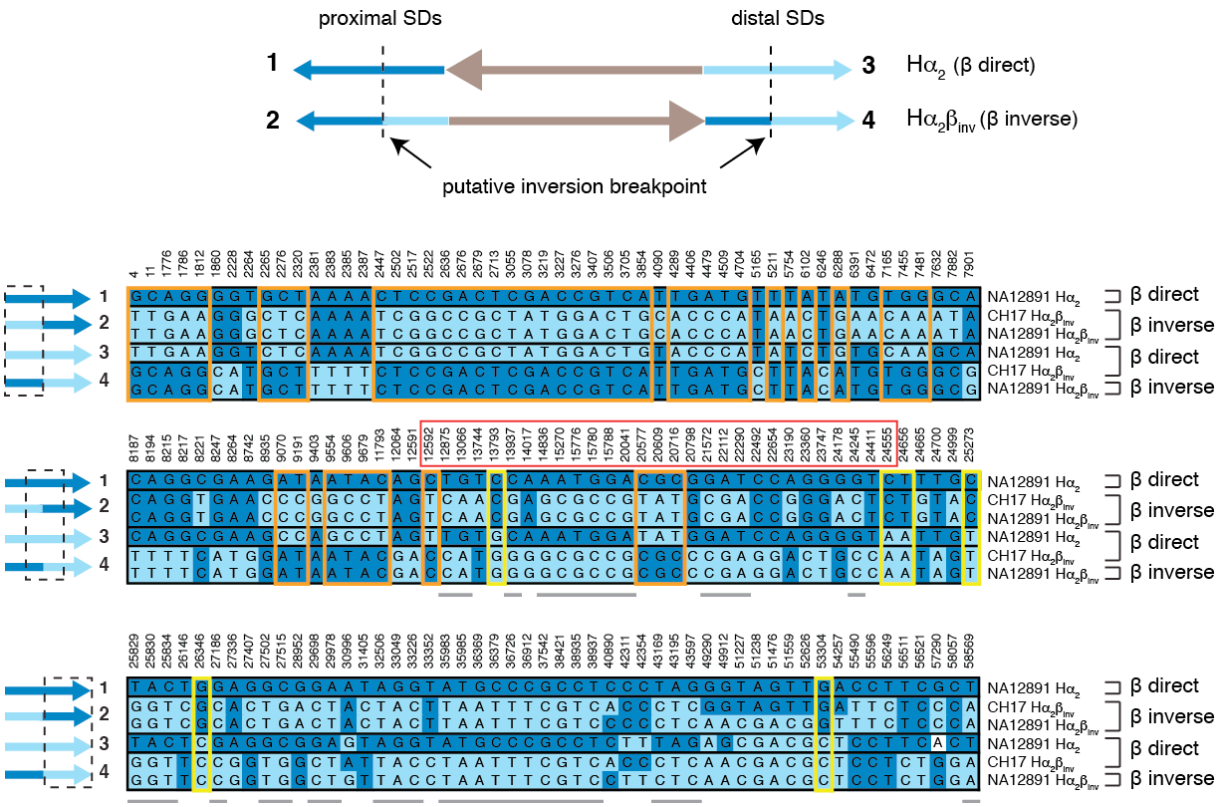
Supplementary Figure 3. Sequence reconstruction of alternative 15q13.3 haplotype. The structure represented in the GRCh37 genome assembly (chr15:29900375-33027253) is compared with a 2.6 Mbp sequence contig constructed from 21 high-quality sequenced CHORI-17 BACs using the program Miropeats. Black lines connect matching segments between the CH17 contig and chr15 reference sequence while red lines highlight the β inversion in the CH17 contig ($H\alpha_2\beta_{inv}$ haplotype). SDs were annotated using SegDupMasker.

Supplementary Figure 4



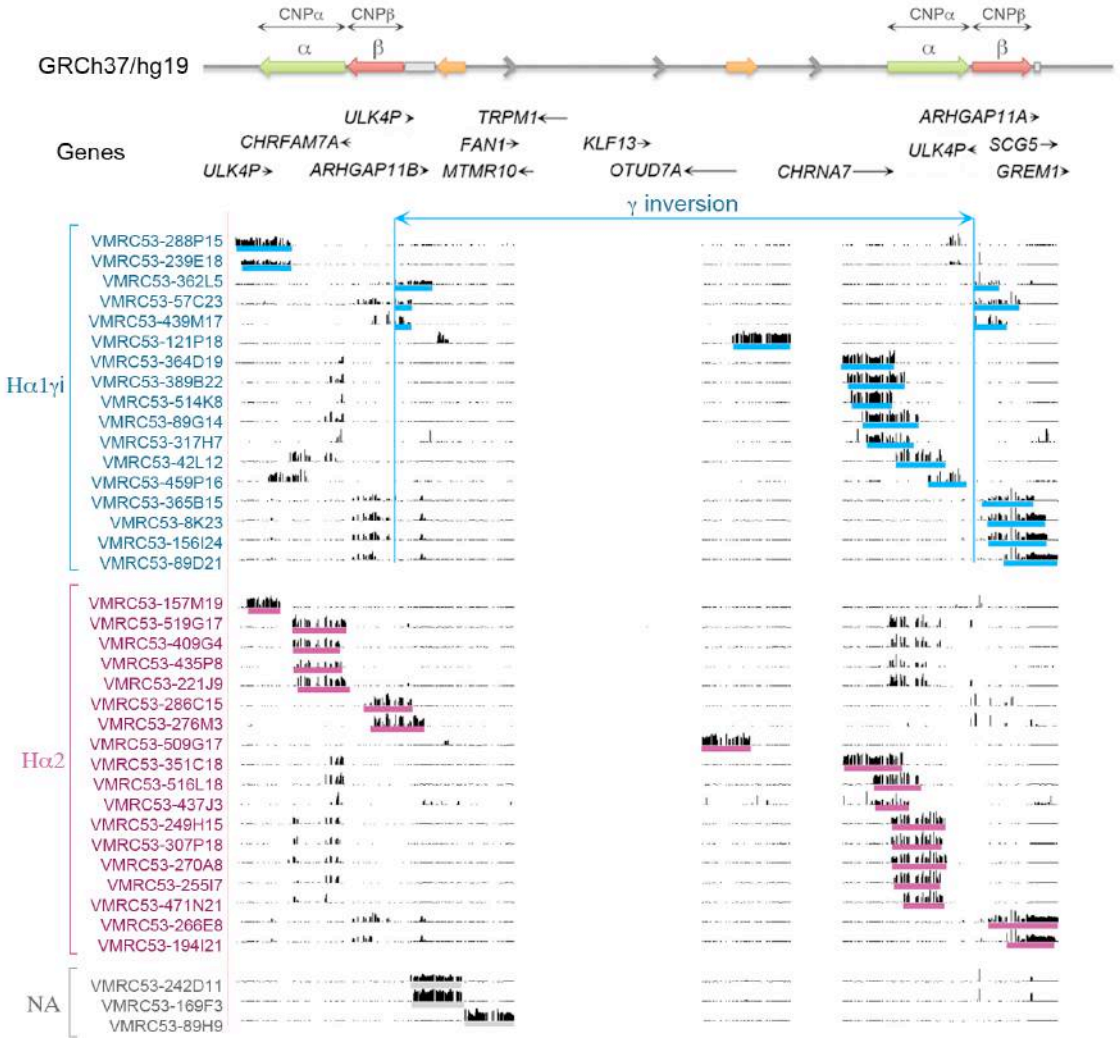
Supplementary Figure 4. Orientation, length and percentage of sequence identity of BP4 and BP5 SDs. Six paralogous regions in direct orientation between BP4 and BP5 are present in the reference genome assembly (Hα_2 haplotype), with the largest directly oriented SD being 58 kbp in size with 99.4% identity. All these SDs contain the *GOLGA8* and *ULK4P3* core duplicons at the transition boundaries. In the CH17 assembly the β inversion increases the directly oriented genomic region to a total of ~188 kbp of 99.4% sequence identity.

Supplementary Figure 5



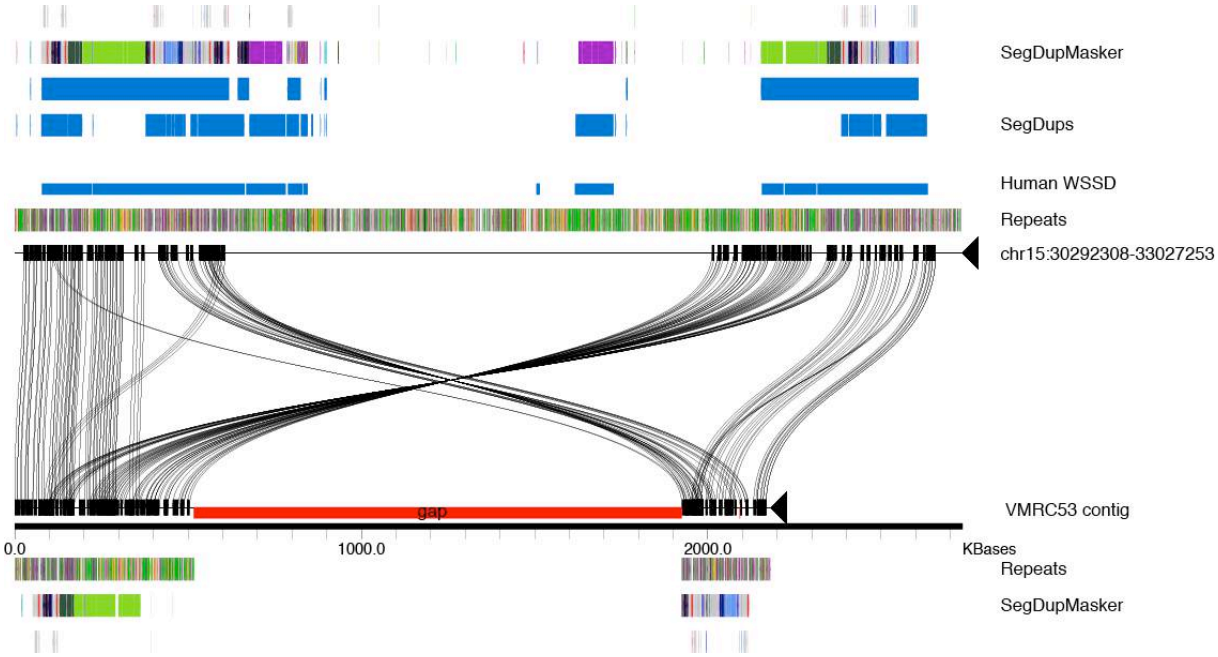
Supplementary Figure 5. Detailed alignment of CNPβ inverted SDs to refine β inversion breakpoints. Representative sequences of flanking inverted SDs (pictured as dark and light blue arrows) were aligned (58.8 kbp) for β direct and inverse haplotypes in order to refine the putative inversion breakpoints (depicted as a dashed line). Pictured are the consensus positions (numbers included above each site) from the MSA that differ across at least two homologous sequences. Orange and yellow boxes highlight positions that show expected signatures inside or outside of the β inversion breakpoints, respectively (see Figure 2 for details). Gray lines are included below sites that show signatures of gene conversion. Consensus positions 12,592-24,555 bp are boxed in red to signify the predicted β inversion breakpoint.

Supplementary Figure 6



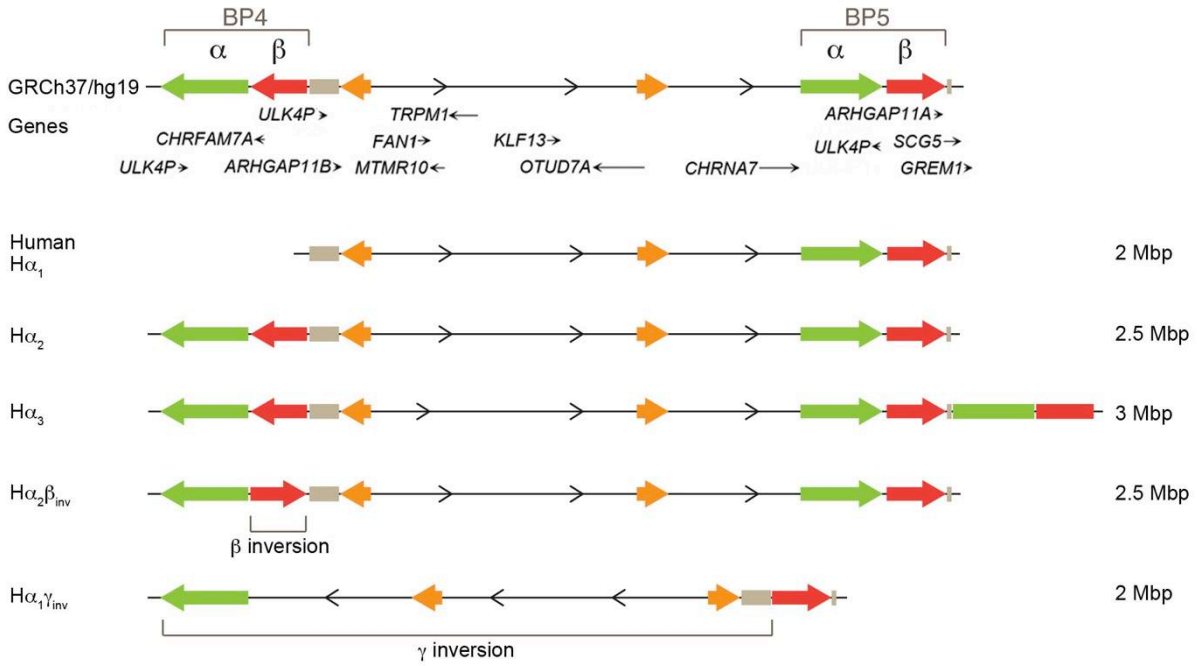
Supplementary Figure 6. Next-generation sequencing of 38 VMRC53 BAC clones from NA12878 shows the presence of a 2 Mbp inversion (γ inversion) across the BP4-BP5 region. Clones that belong to the paternal haplotype (purple clones) are concordant with the reference genome assembly ($H\alpha_2$). Analysis of the clones mapping to the maternal inverted haplotype (blue clones) shows that this configuration lacks copies of CNPα and CNPβ at BP4 but shows clear evidence of an inversion ($H\alpha_1\gamma_{inv}$).

Supplementary Figure 7



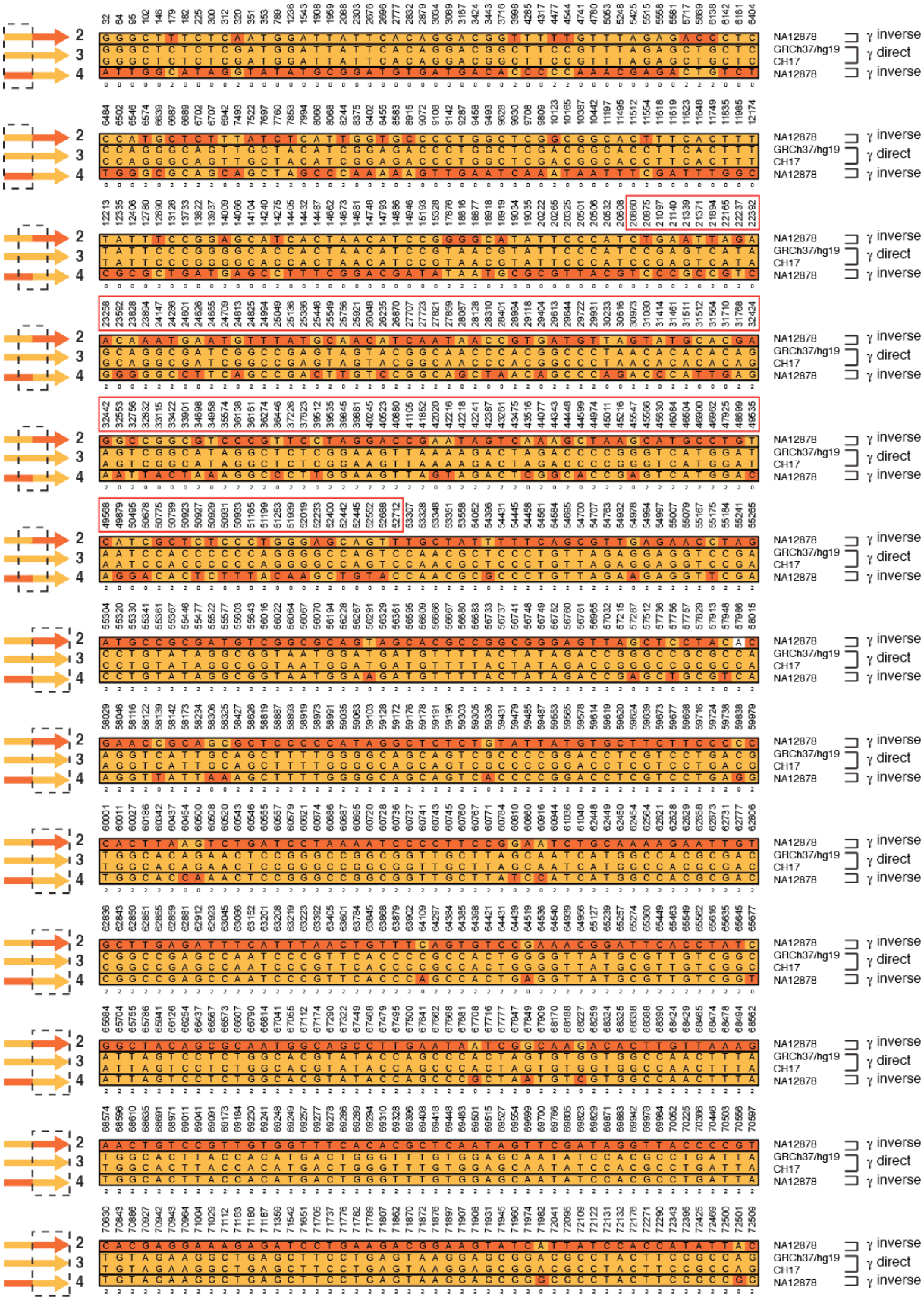
Supplementary Figure 7. Sequence reconstruction of alternative 15q13.3 haplotype. Using the program Miropeats, the GRCh37 genome assembly (chr15:30,292,308-33,027,253) is compared with a sequence contig constructed from seven high-quality sequenced VMRC53 BACs spanning the γ inversion breakpoints ($H\alpha_1\gamma_{inv}$ haplotype). Black lines connect matching segments between the VMRC53 contig and chr15 reference sequence while in red is highlighted the SDs were annotated using SegDupMasker.

Supplementary Figure 8



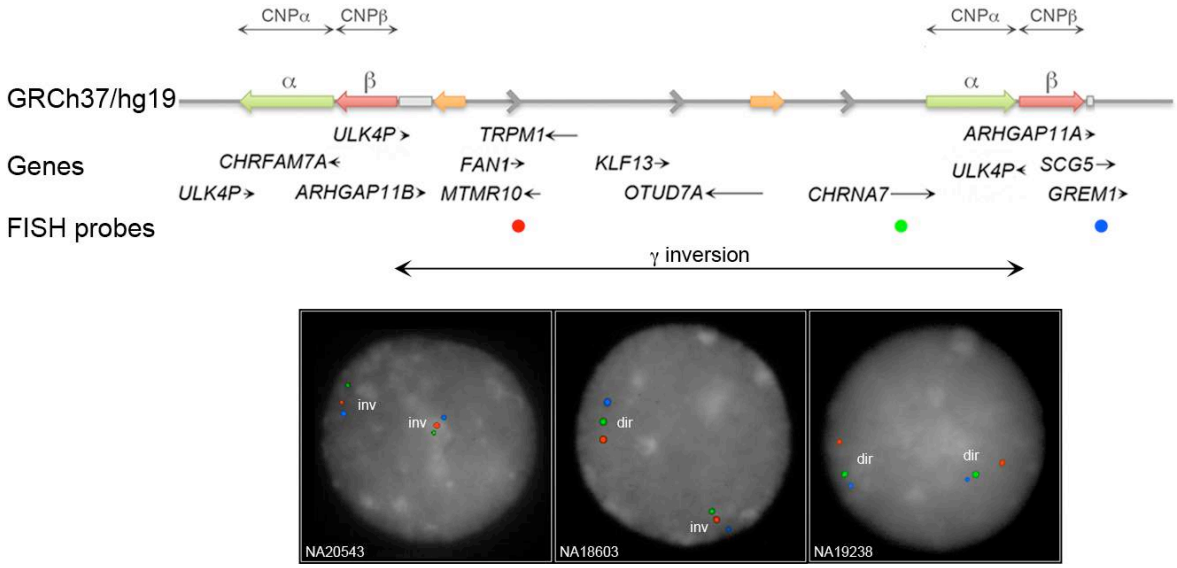
Supplementary Figure 8. Alternative structural haplotypes of 15q13.3. Shown are five distinct human structural haplotypes of the 15q13.3 region. Colored boxes indicate SDs and the arrows indicate the orientation of the region compared to the structure represented in the GRCh37 reference genome assembly ($H\alpha_2$ structural haplotype). Three direct haplotypes are defined based on the CN of two CNPs (CNPα and CNPβ) at BP4 and BP5: $H\alpha_1$ with zero copies at BP4 and one copy at BP5, $H\alpha_2$ with one copy at BP4 and one at BP5, and $H\alpha_3$ with one copy at BP4 and two copies at BP5. $H\alpha_2\beta_{inv}$ differs from the reference genome assembly for the presence of the β inversion at BP4. $H\alpha_1\gamma_{inv}$ configurations carry the γ inversion and are missing one copy of CNPα and CNPβ compared to the reference genome assembly ($H\alpha_2$ haplotype).

Supplementary Figure 9



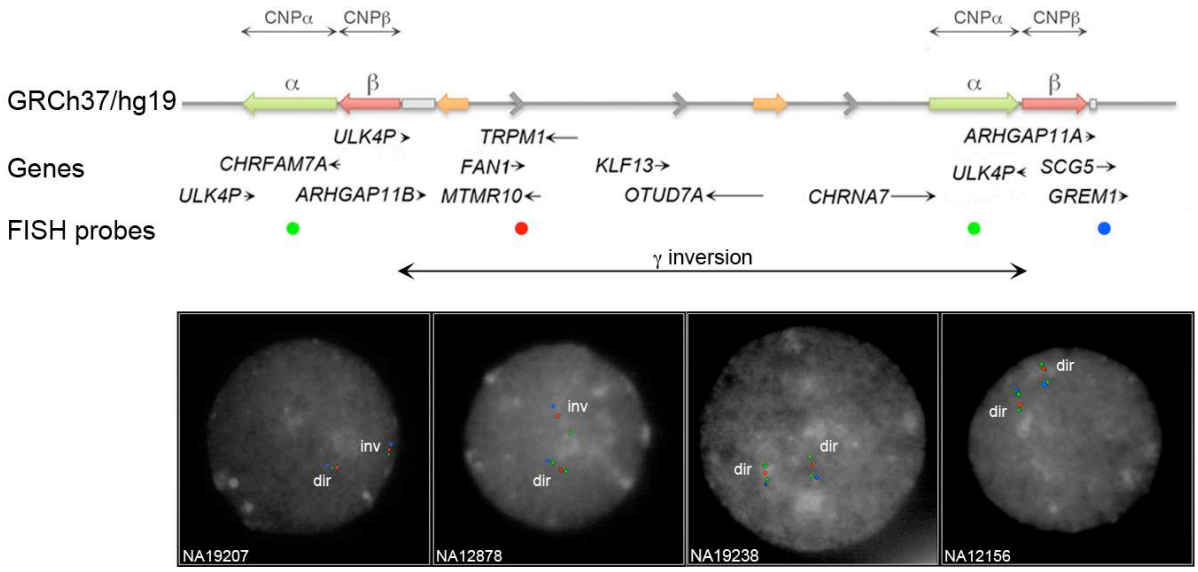
Supplementary Figure 9. Detailed alignment of flanking inverted SDs to refine γ inversion breakpoints. A 72,608 bp MSA was generated using homologous sequences from NA12878, CH17, and GRCh37/hg19 of the flanking inverted SDs that likely mediated the γ inversion. Pictured are the consensus positions (numbers included above each site) from the alignment that were shared between the distal SDs from CH17 and GRCh37/hg19 (γ direct haplotype) and variant in either SD derived from NA12878 (γ inverse haplotype). The number 0 (blue in Figure 3) or 2 (green in Figure 3) is included below the sites that show expected signatures inside or outside of the γ inversion breakpoints, respectively. Consensus positions 20,860 to 52,712 bp are boxed in red to signify the predicted γ inversion breakpoint.

Supplementary Figure 10



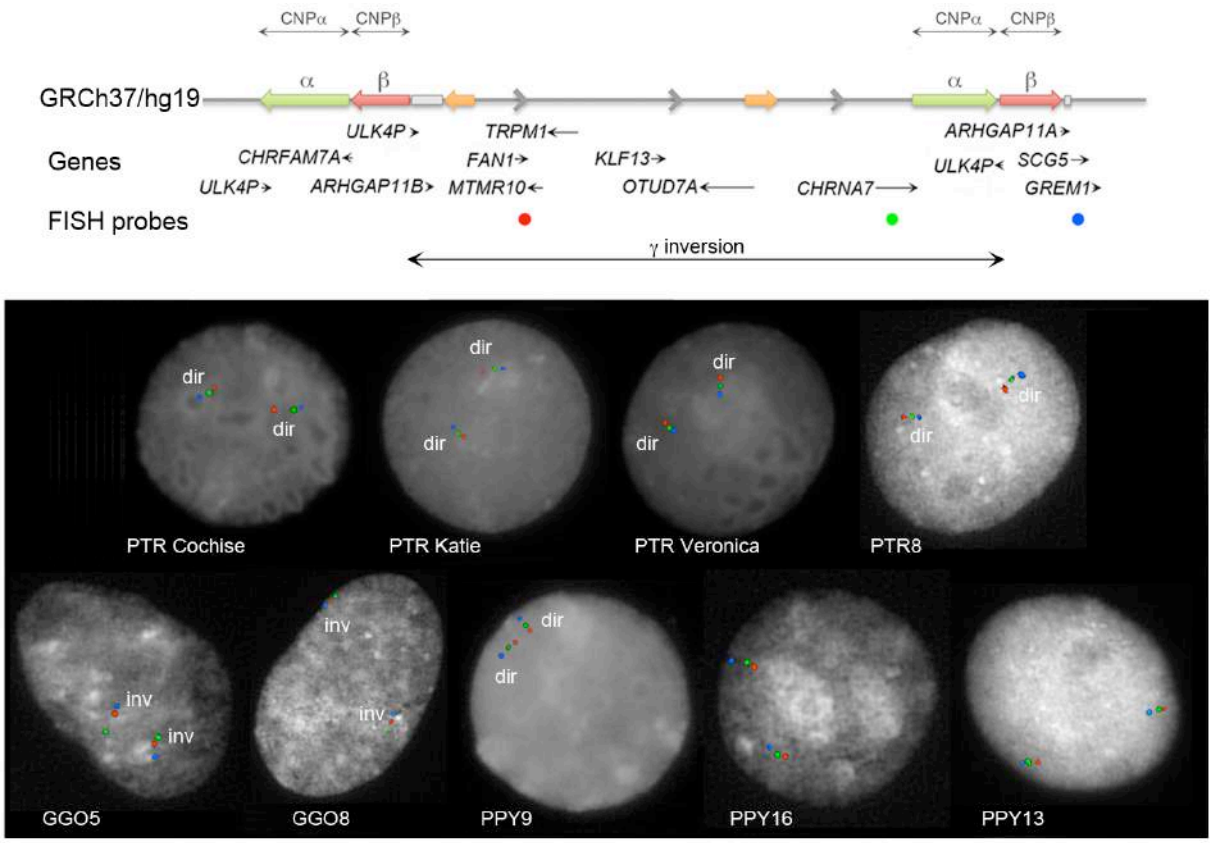
Supplementary Figure 10. Genotyping of the γ inversion. Shown is a three-color interphase FISH analysis of the γ inversion using two probes mapping inside (WIBR2-3205J20, red; WIBR2-3158E16, green) and one probe mapping outside (WIBR2-1368H18, blue) the inversion.

Supplementary Figure 11



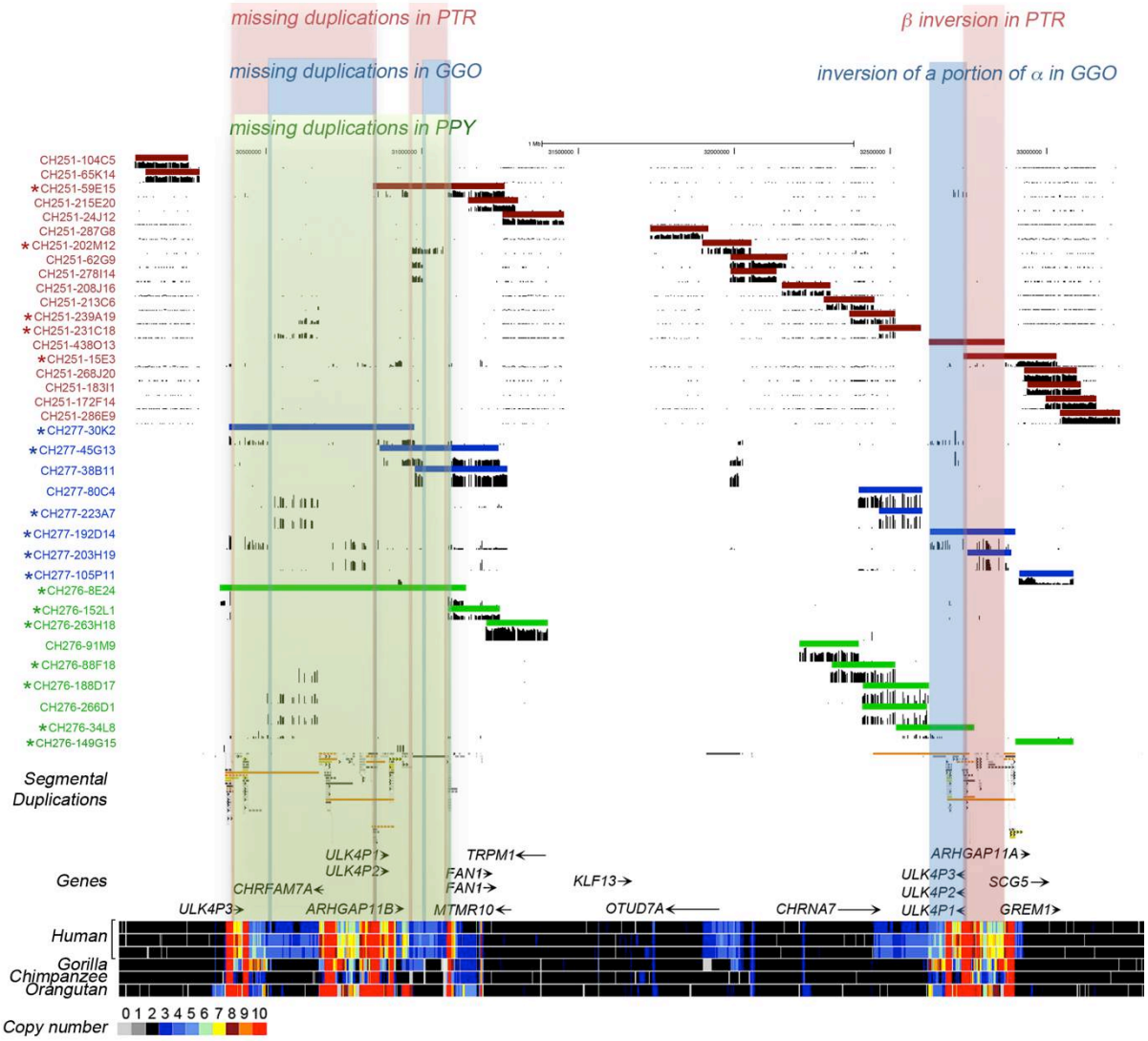
Supplementary Figure 11. Copy number analysis of CNPα in γ inverted and direct individuals. FISH analysis using a probe mapping at CNPα (WIBR2-1388I24, green) and two probes mapping in the unique sequence (WIBR2-3205J20, red; WIBR2-1368H18, blue) shows a variable CN of CNPα between 0 and 1 at BP4 and between 0 and 2 at BP5. When the haploid CN of CNPα is one and it carries the γ inversion, there is a "missing" copy of CNPα at BP5 with respect to the reference genome, showing that the inversion transported this paralog to BP4. CNPα varies between 1 and 3 copies in the directly oriented configurations. Here, the haploid CN ranges between 0 and 1 at BP4 and between 1 and 2 copies at BP5.

Supplementary Figure 12



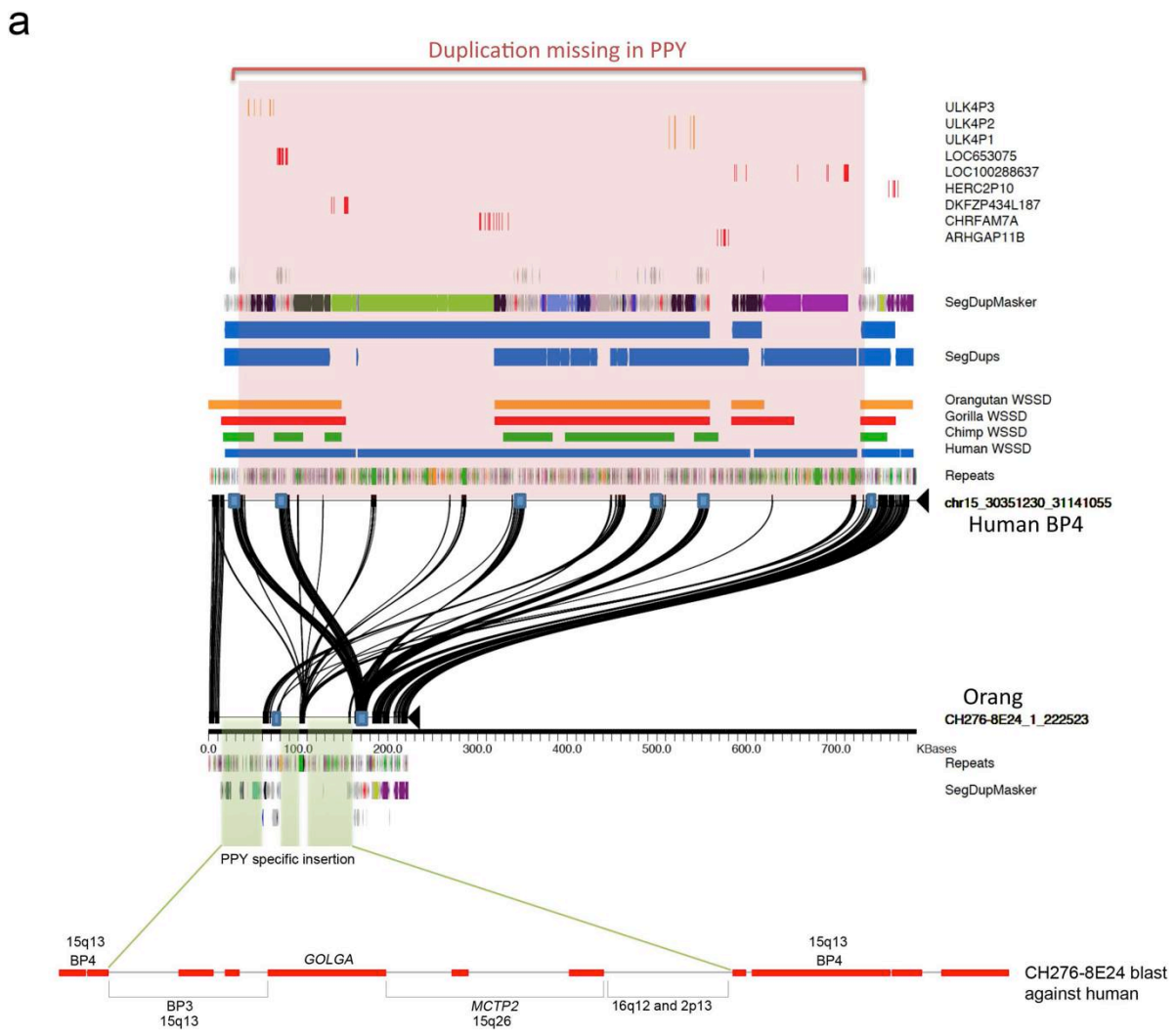
Supplementary Figure 12. γ inversion analysis in nonhuman primates. Three-color interphase FISH analysis of the γ inversion using two probes mapping inside (WIBR2-3205J20, red; WIBR2-3158E16, green) and one probe mapping outside (WIBR2-1368H18, blue) the inversion was used to test three orangutan (*Pongo pygmaeus*), two gorilla (*Gorilla gorilla*), and four chimpanzee (*Pan troglodytes*) cell lines. Orangutan and chimpanzee individuals are in direct orientation when compared to the human reference genome, while all gorilla individuals are in inverted orientation.

Supplementary Figure 13

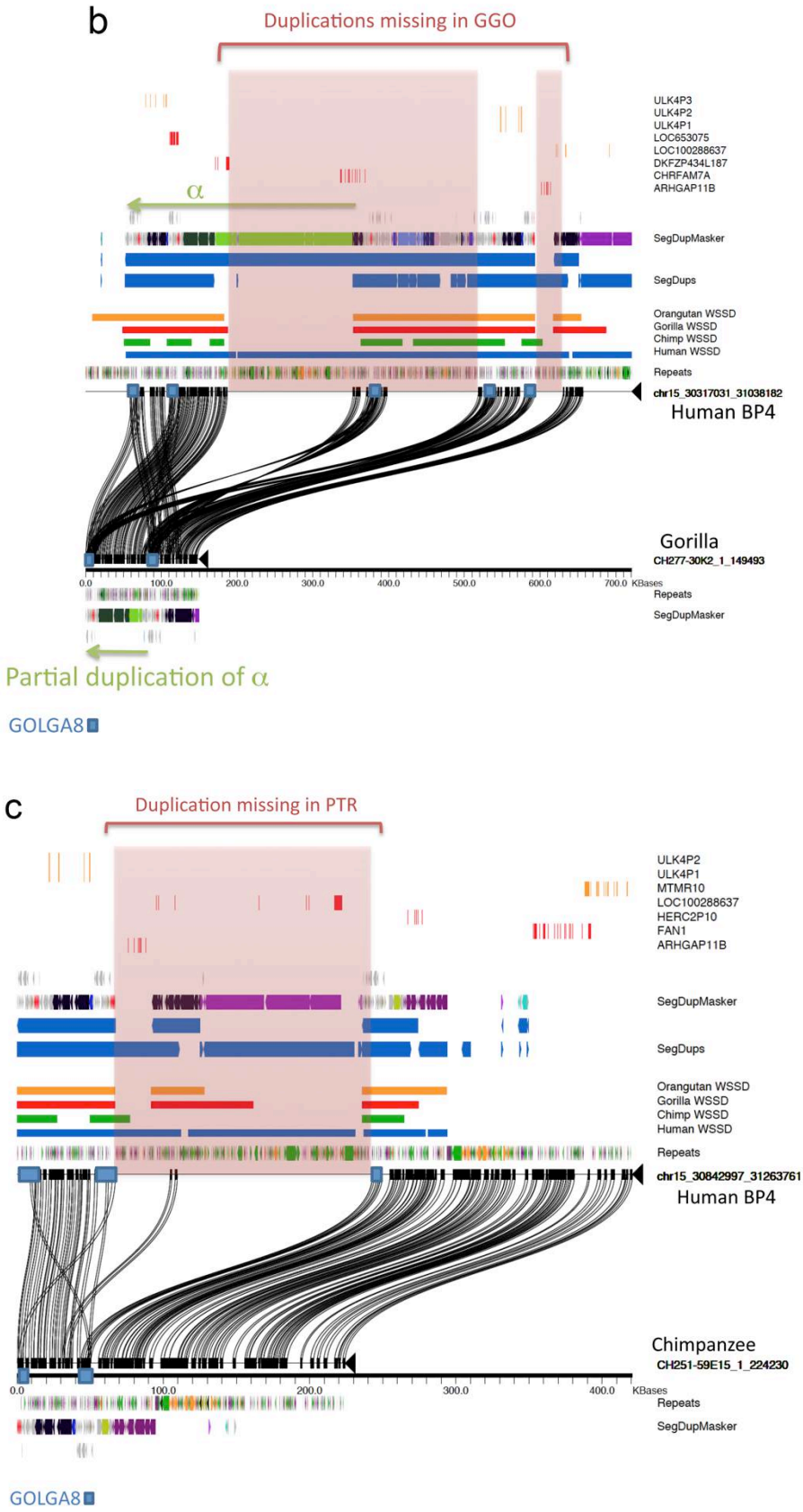


Supplementary Figure 13. Next-generation sequencing of primate BAC clones. BAC clones from chimpanzee (red), gorilla (blue), and orangutan (green) were sequenced using Illumina HiSeq 2000 and mapped to the human reference genome assembly. Clones marked with an asterisk were selected for high-quality sequencing and *de novo* assembly using either capillary or PacBio long-read sequences. Sequence analysis reveals a much simpler organization of the 15q13.3 region in nonhuman primates when compared to human, the presence of an inversion of β at BP5 in chimpanzee, and an inversion of a portion of α at BP5 in gorilla.

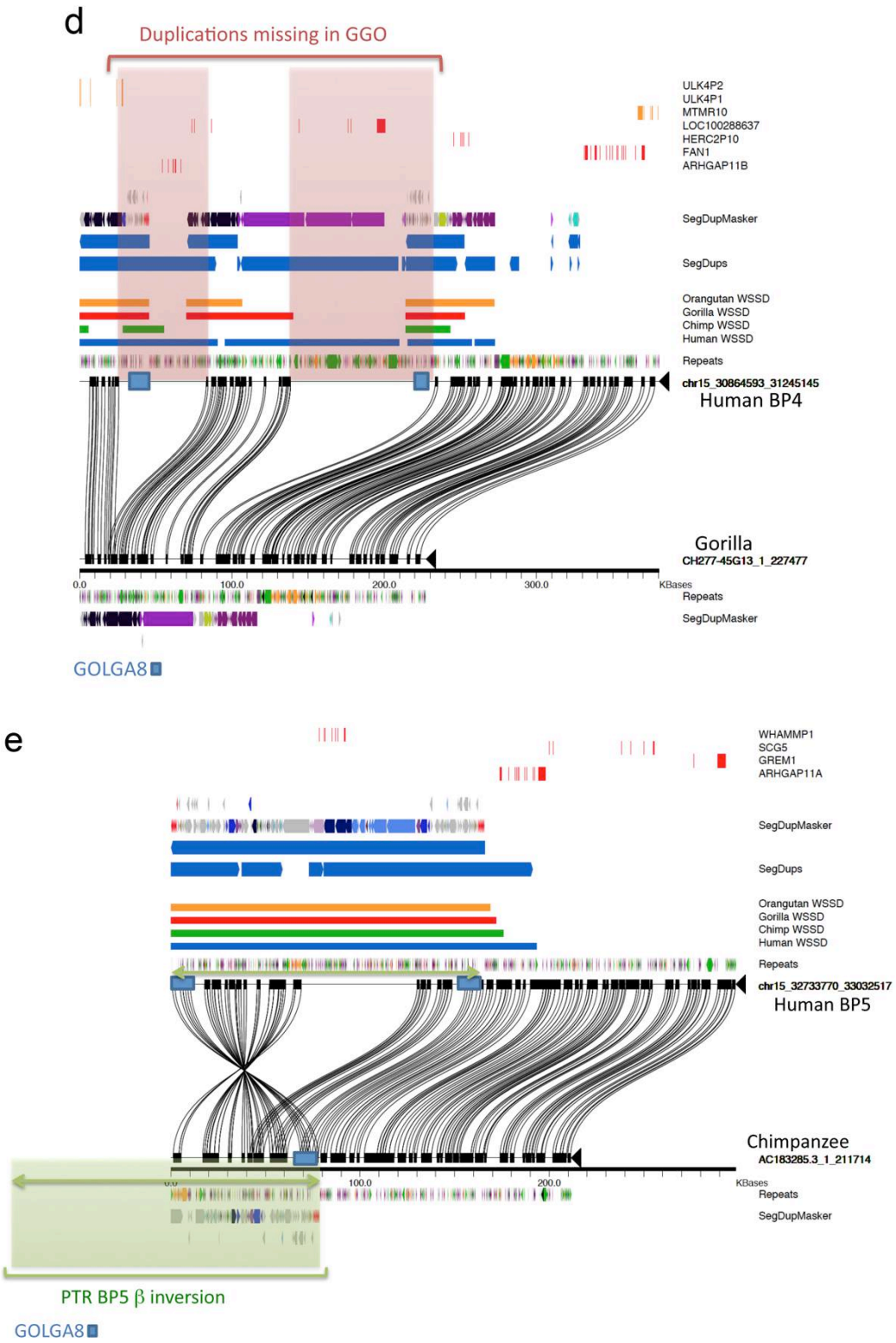
Supplementary Figure 14a



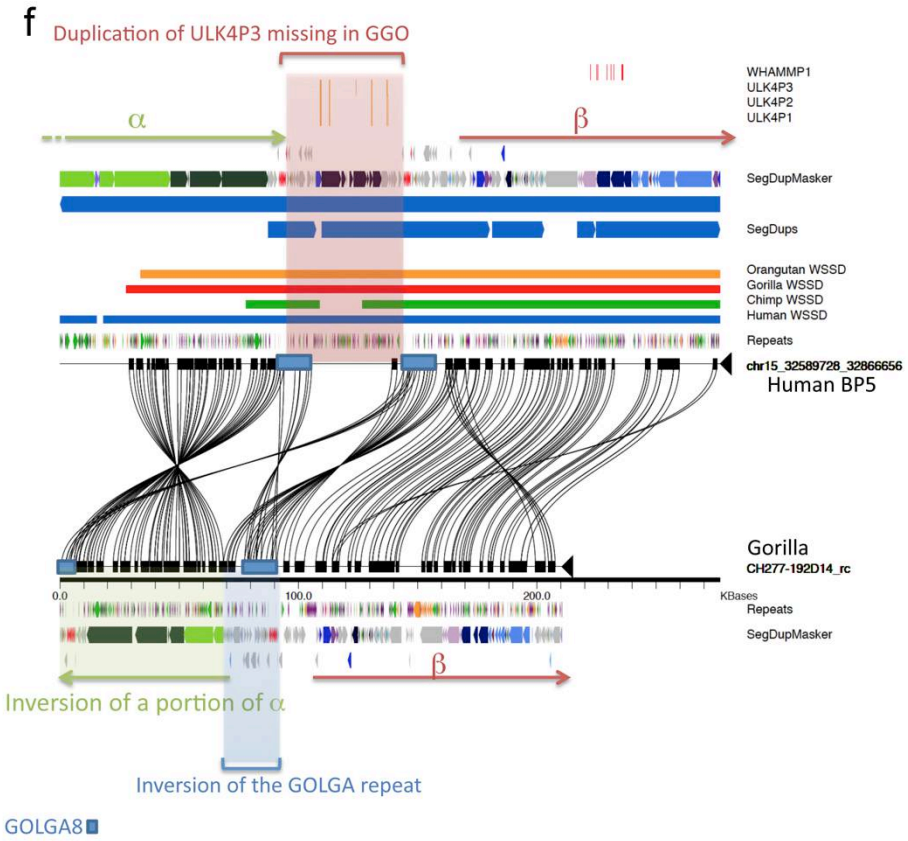
Supplementary Figures 14b and 14c

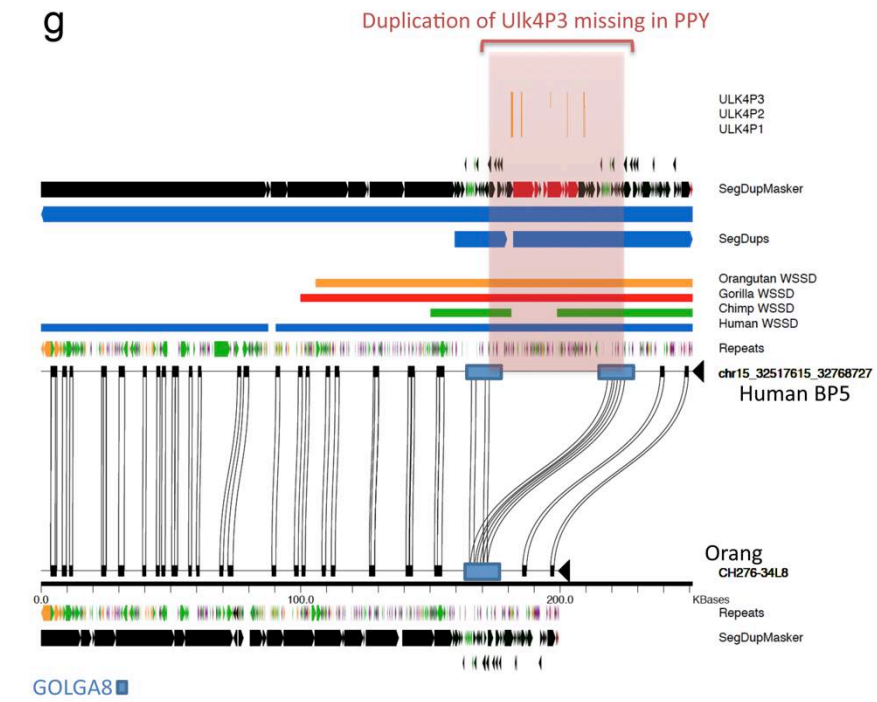


Supplementary Figures 14d and 14e



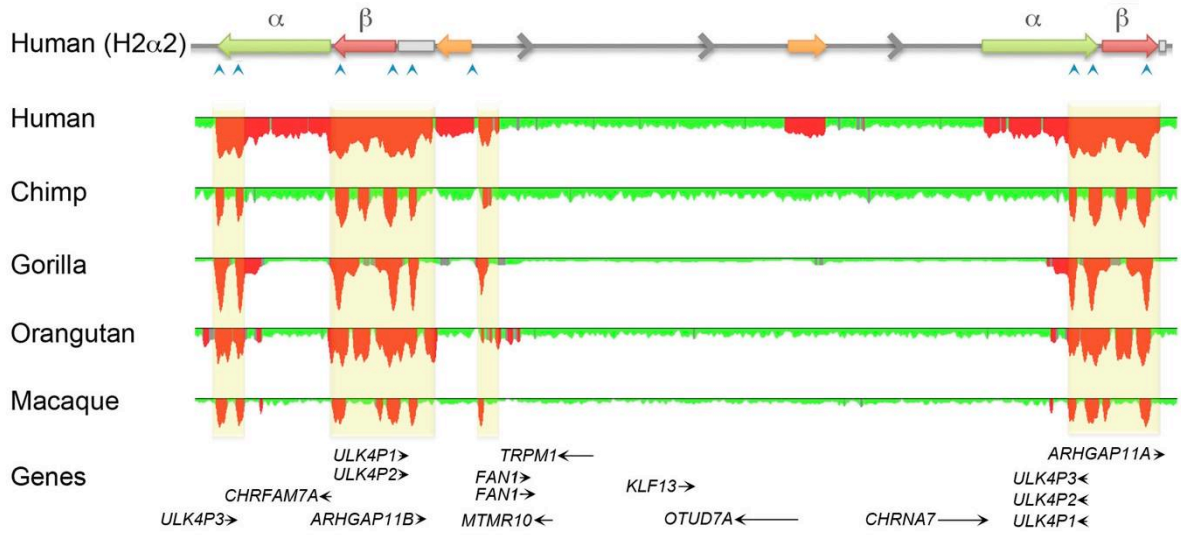
Supplementary Figures 14f and 14g





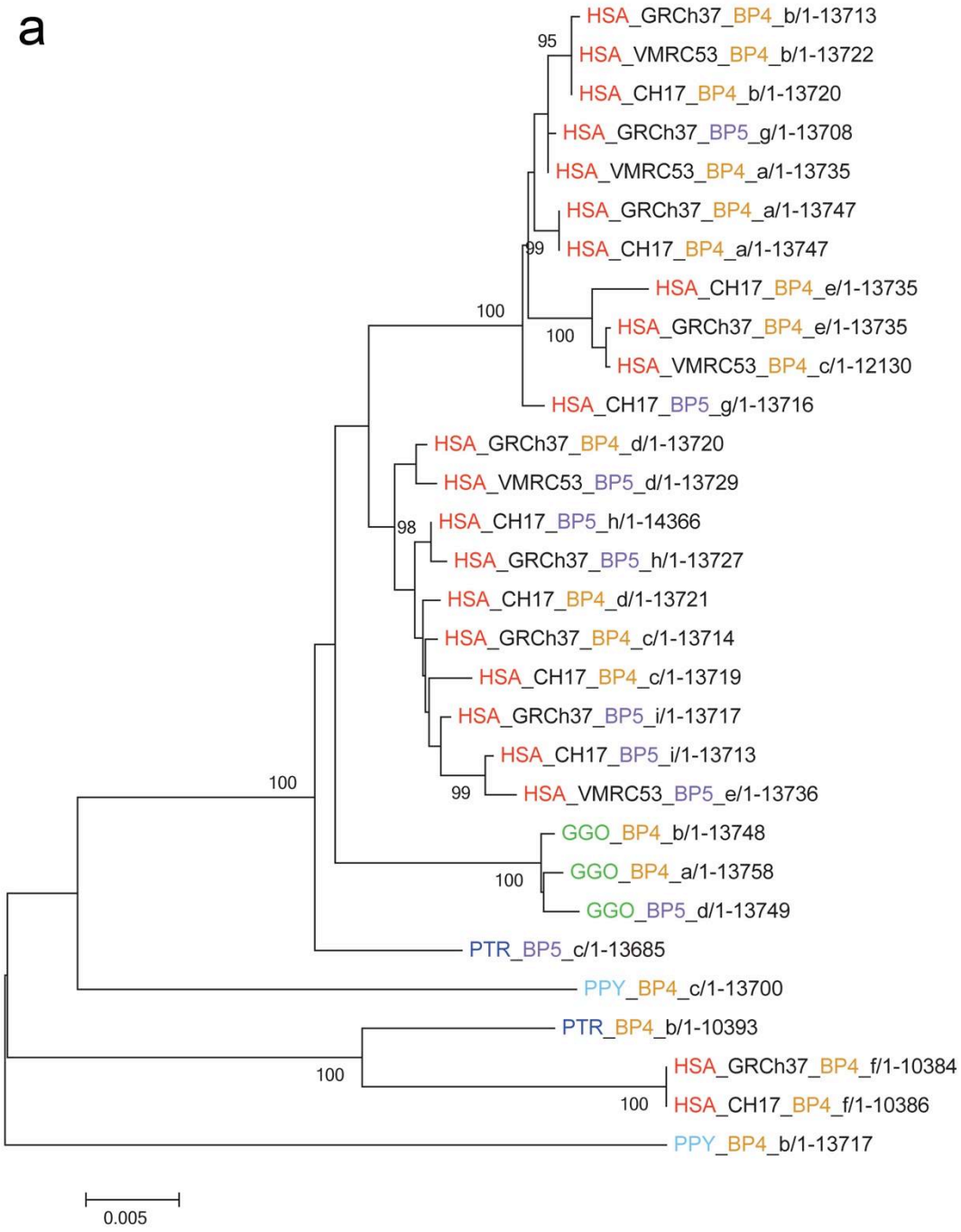
Supplementary Figure 14. Nonhuman primate clones analysis. Miropeats analysis of seven representative primate clones sequenced with either capillary or PacBio long-read sequences. **(a, b, c, d, e, f, g)** The analysis shows that nonhuman primates lack almost all of the duplications present at most human BP4 haplotypes, with the exception of the *GOLGA8* repeats. **(a)** Analysis of an orangutan clone (CH276-8E24) mapping at BP4 shows the expansion of *GOLGA8* in human, and the presence of an orangutan-specific insertion within this region. BLAST analysis against the human reference genome shows hits of this sequence with percentage identity lower the 90% in other regions of the genome (15q13 BP3, 15q26, 16q12 and 2p13). **(e)** The ancestral copy of CNPβ maps in an inverted orientation in chimpanzee at BP5 and the *GOLGA8* repeats define the boundaries of this evolutionary inversion (clone AC183285). **(f)** An inversion of the distal portion of α instead is found in gorilla at BP5 (clone CH277-192D14). **(b)** This portion of α is partially duplicated at BP4 in gorilla (clone CH277-30K2), and in both instances the rearrangement (duplication at BP4 and inversion at BP5) is flanked by the *GOLGA8* repeats. **(g)** Analysis of a gorilla (CH277-192D14) **(f)** and an orangutan (CH276-34L8) clone mapping at BP5 shows that the 58 kbp repeat found at the breakpoints of the β and γ inversion in humans seems to have a simpler organization in these two primate species since it is missing *ULK4P3* flanking *GOLGA8* in humans.

Supplementary Figure 15

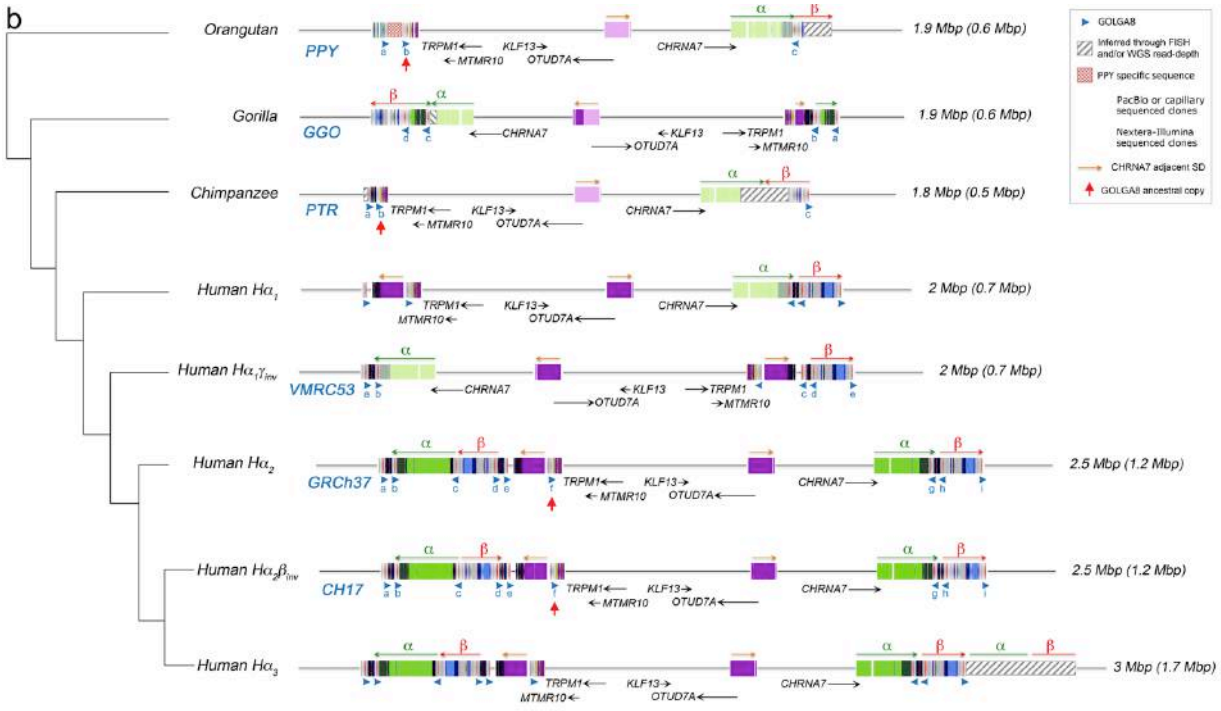


Supplementary Figure 15. Read-depth analysis in human and nonhuman primates. Read-depth analysis of whole-genome sequence from chimpanzee (*Pan troglodytes*), gorilla (*Gorilla gorilla*), orangutan (*Pongo pygmaeus*), and macaque (*Macaca mulatta*) individuals shows that nonhuman primates lack almost all of the duplications present at most human BP4 haplotypes, with the exception of the core duplicons containing the *GOLGA8* repeats (blue arrowheads).

Supplementary Figure 16a

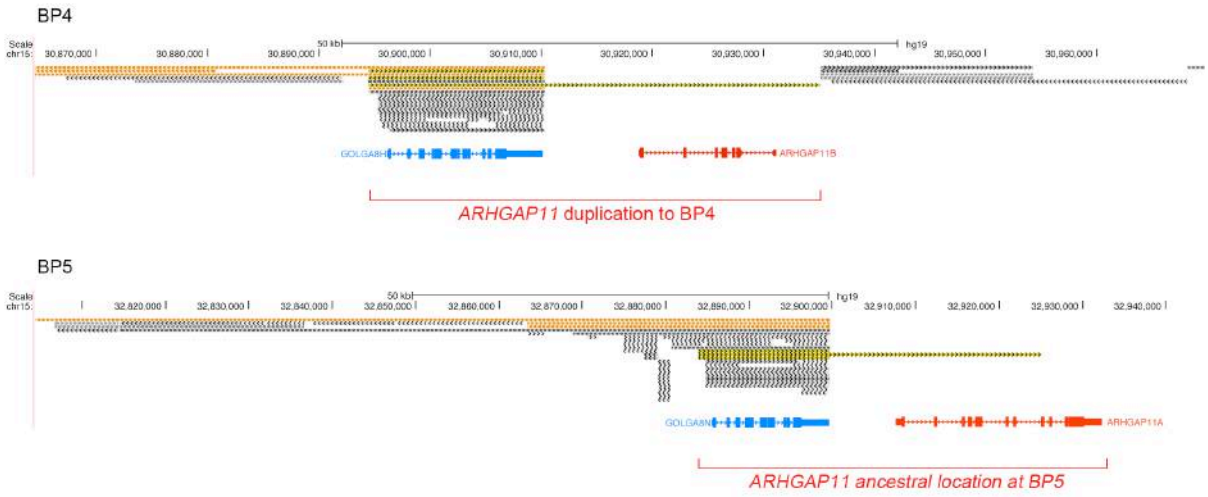


Supplementary Figure 16b



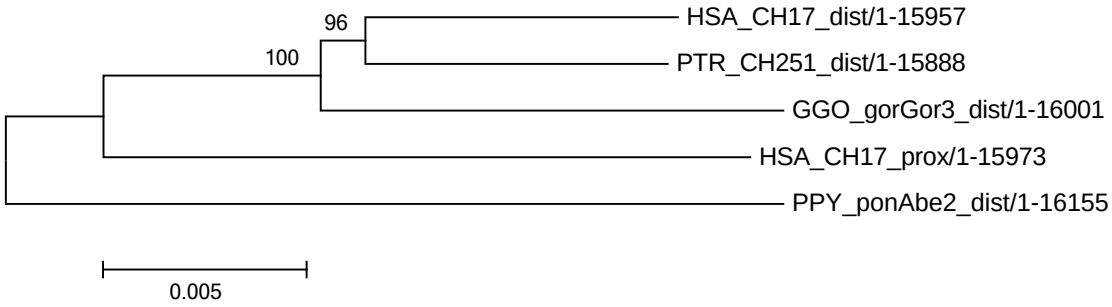
Supplementary Figure 16. Phylogenetic analysis of *GOLGA8* in primates. (a) An unrooted neighbor-joining tree was constructed using the MEGA5 complete deletion option based on 13.7 aligned kbp from 30 *GOLGA8* sequences retrieved from human, chimpanzee, gorilla and orangutan sequenced clones. **(b)** The location of *GOLGA8* across the different human and great ape haplotypes is shown with blue arrowheads. Red arrows indicate four *GOLGA8* copies that can be identified unambiguously as orthologous both by phylogenetic analysis and with respect to their position.

Supplementary Figure 17



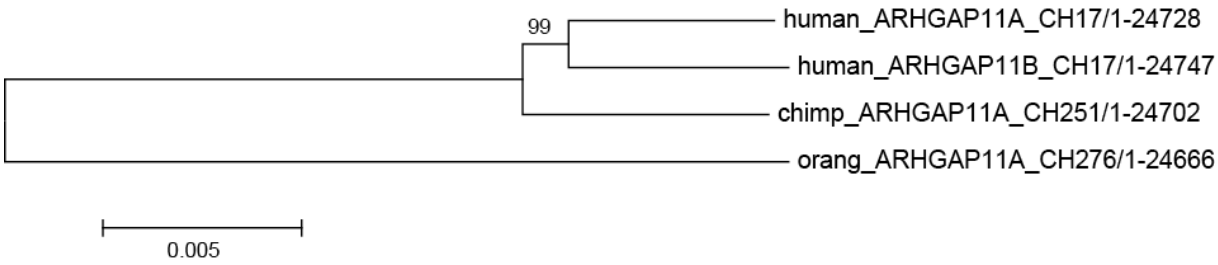
Supplementary Figure 17. Duplication of *ARHGAP11* from BP5 to BP4. A UCSC screenshot shows the location of *ARHGAP11A* (BP5) and *ARHGAP11B* (BP4) in the 15q13.3 region. The proximal breakpoint of the *ARHGAP11* duplication maps within a *GOLGA8* repeat (blue).

Supplementary Figure 18



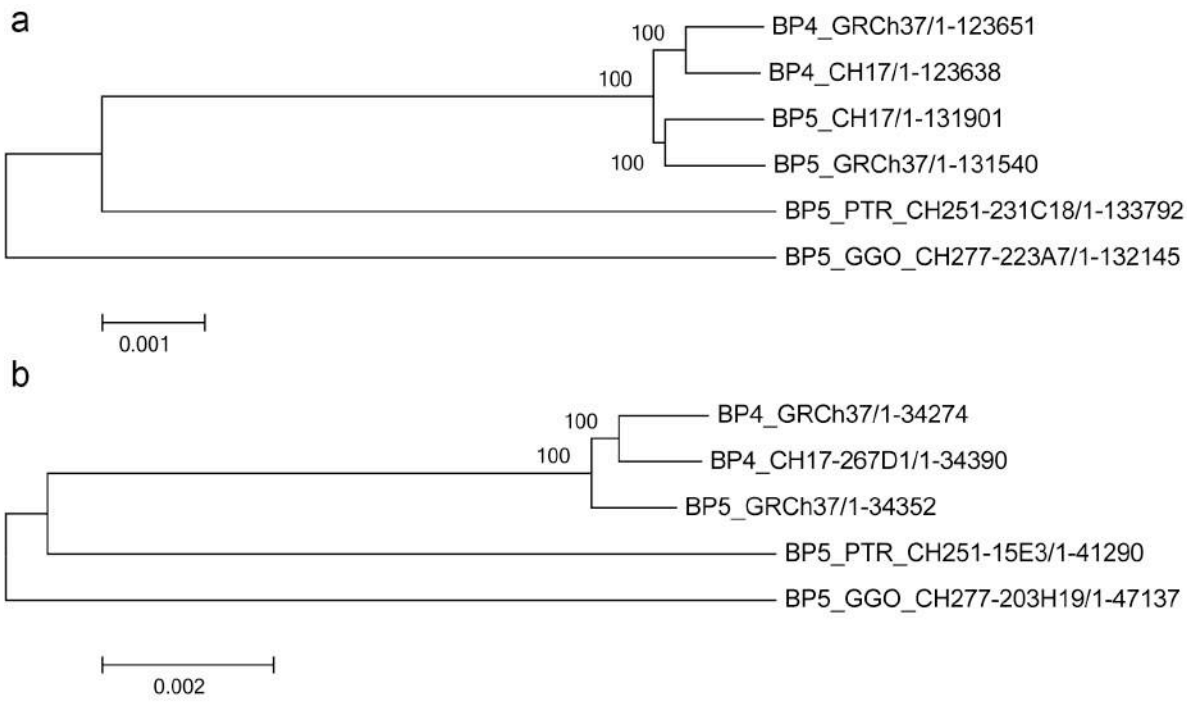
Supplementary Figure 18. Evolutionary analyses of the *CHRNA7*-adjacent SD. An unrooted neighbor-joining tree was constructed using the MEGA5 complete deletion option based on 15 kbp aligned sequence within the *CHRNA7*-adjacent SD.

Supplementary Figure 19



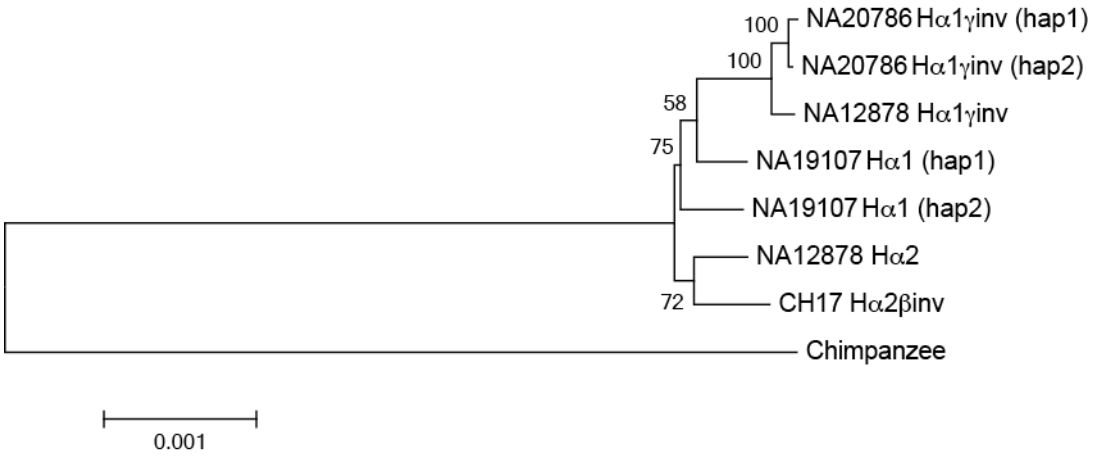
Supplementary Figure 19. Evolutionary analyses of the human-specific *ARHGAP11* SD. An unrooted neighbor-joining tree was constructed using the MEGA5 complete deletion option based on 24.7 kbp aligned sequence within the *ARHGAP11* SD.

Supplementary Figure 20



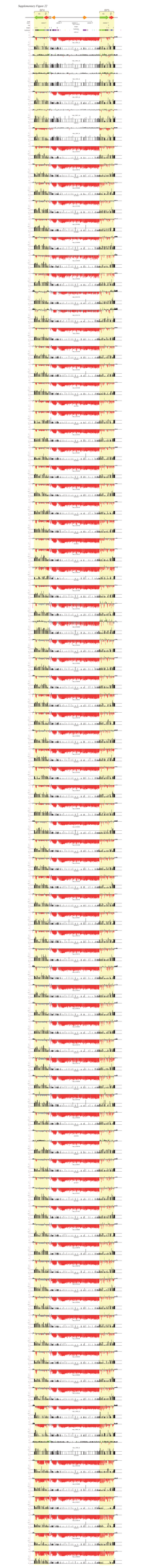
Supplementary Figure 20. Evolutionary analyses of CNPα and CNPβ. Phylogenetic trees were constructed from MSAs of homologous sequences representing **(a)** CNPα and **(b)** CNPβ SDs using the Kimura 2-parameter model to calculate genetic distances (complete deletion option) and standard errors using MEGA5 with 500 bootstrap replications.

Supplementary Figure 21



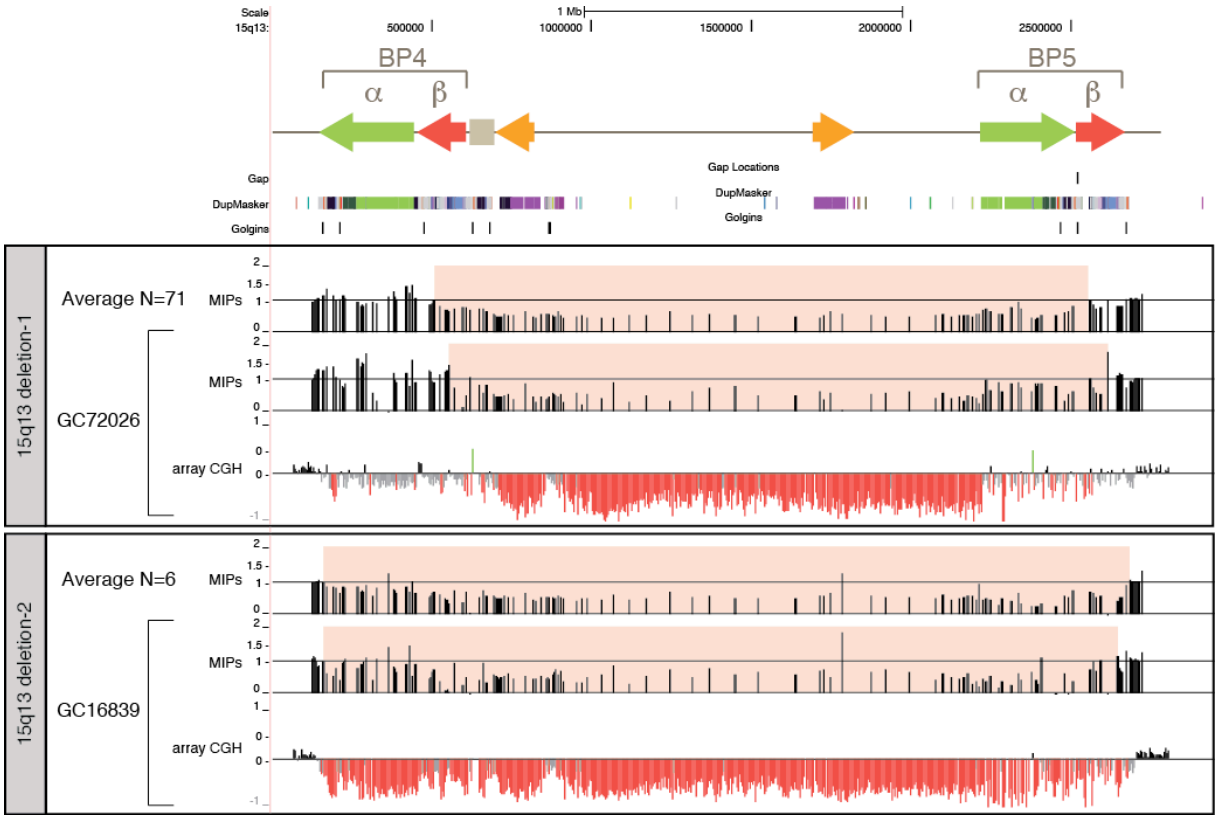
Supplementary Figure 21. Phylogenetic analysis of γ direct and inverted haplotypes. An unrooted neighbor-joining tree was constructed using the MEGA5 complete deletion option based on 154,471 aligned base pairs from unique sequence within the inversion interval. The bootstrap support for each branch is based on 500 replications.

Supplementary Figure 22



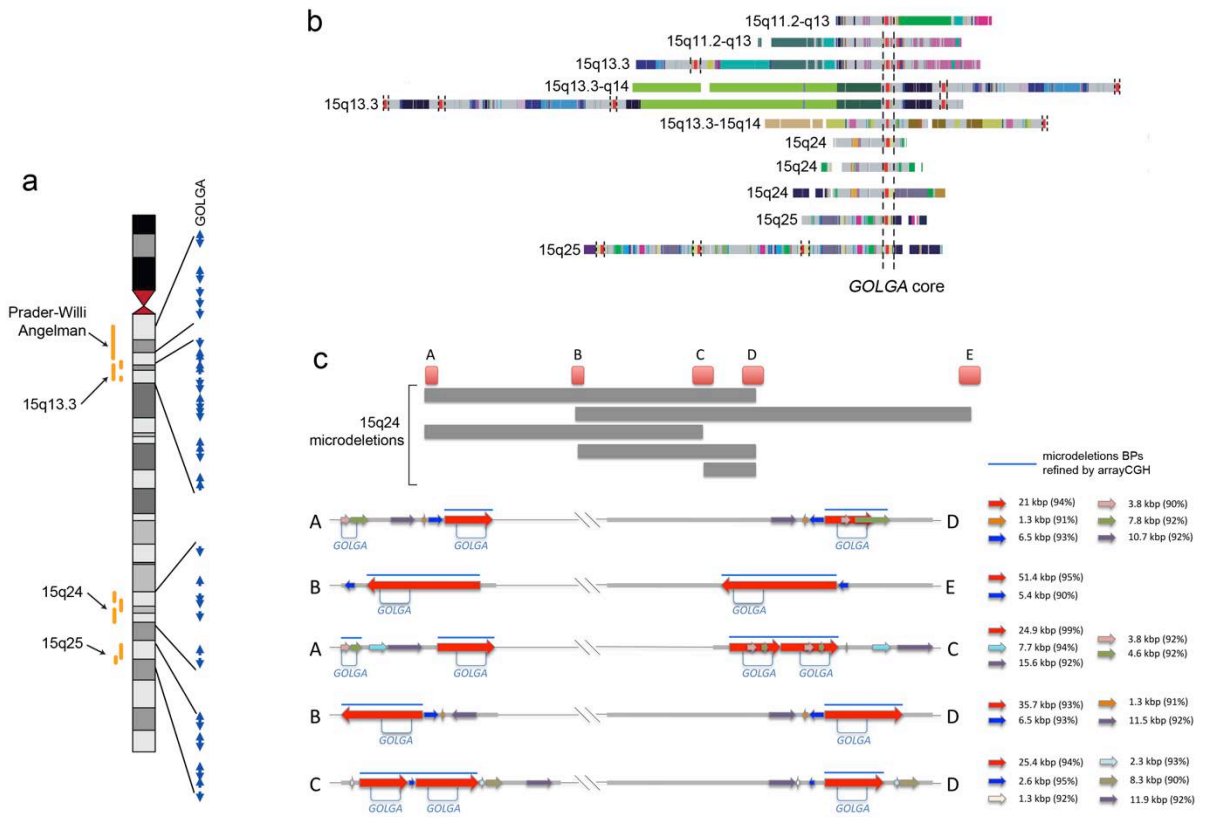
Supplementary Figure 22. 15q13.3 microdeletions in children with autism and/or developmental delay. Custom array CGH $\log_2$ relative hybridization signals, mapped against the human CH17 assembly, are depicted as histograms (probes with +1.5 standard deviation (s.d.) shown as red and -1.5 s.d. shown as green) for each child. The normalized read-depth (0-2) from the MIP capture is shown at each SUNK position.

Supplementary Figure 23



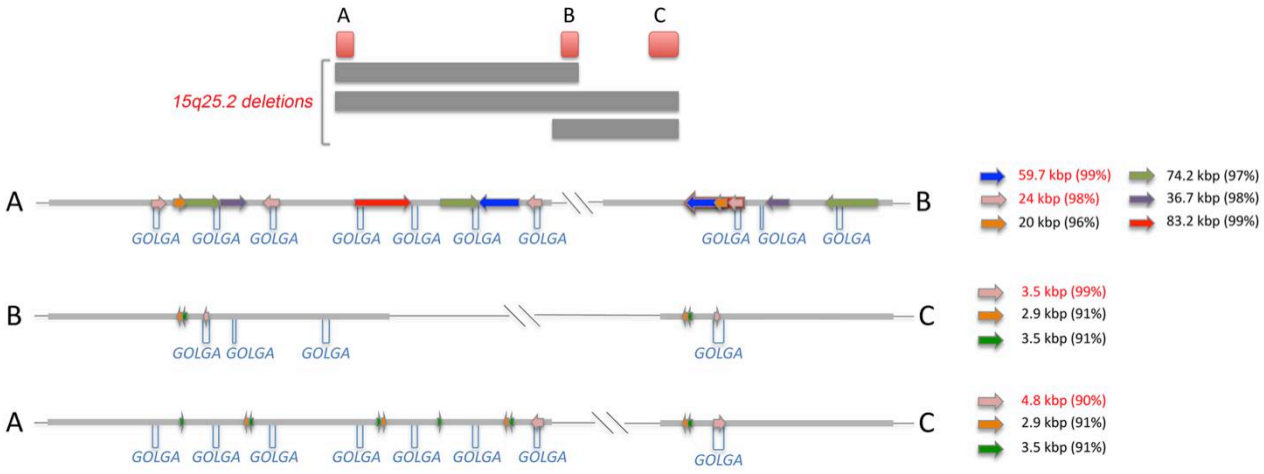
Supplementary Figure 23. Targeted array CGH and MIP sequencing analysis of 15q13.3 microdeletions. 80 15q13.3 microdeletion patients were screened using a customized microarray targeted to the 15q13.3 region, which mapped the breakpoints of the disease-critical region to a ~500 kbp region spanned by the CNPα and CNPβ SDs. We delineated patients into two groups: (1) a larger subset harboring 1.5 Mbp deletions appearing to break within CNPβ (n=73; 15q13 deletion-1) and (2) a smaller subset harboring larger 2 Mbp deletions appearing to break within CNPα (n=7; 15q13 deletion-2). Pictured are the averaged MIP-sequence read-depth results across each group.

Supplementary Figure 24



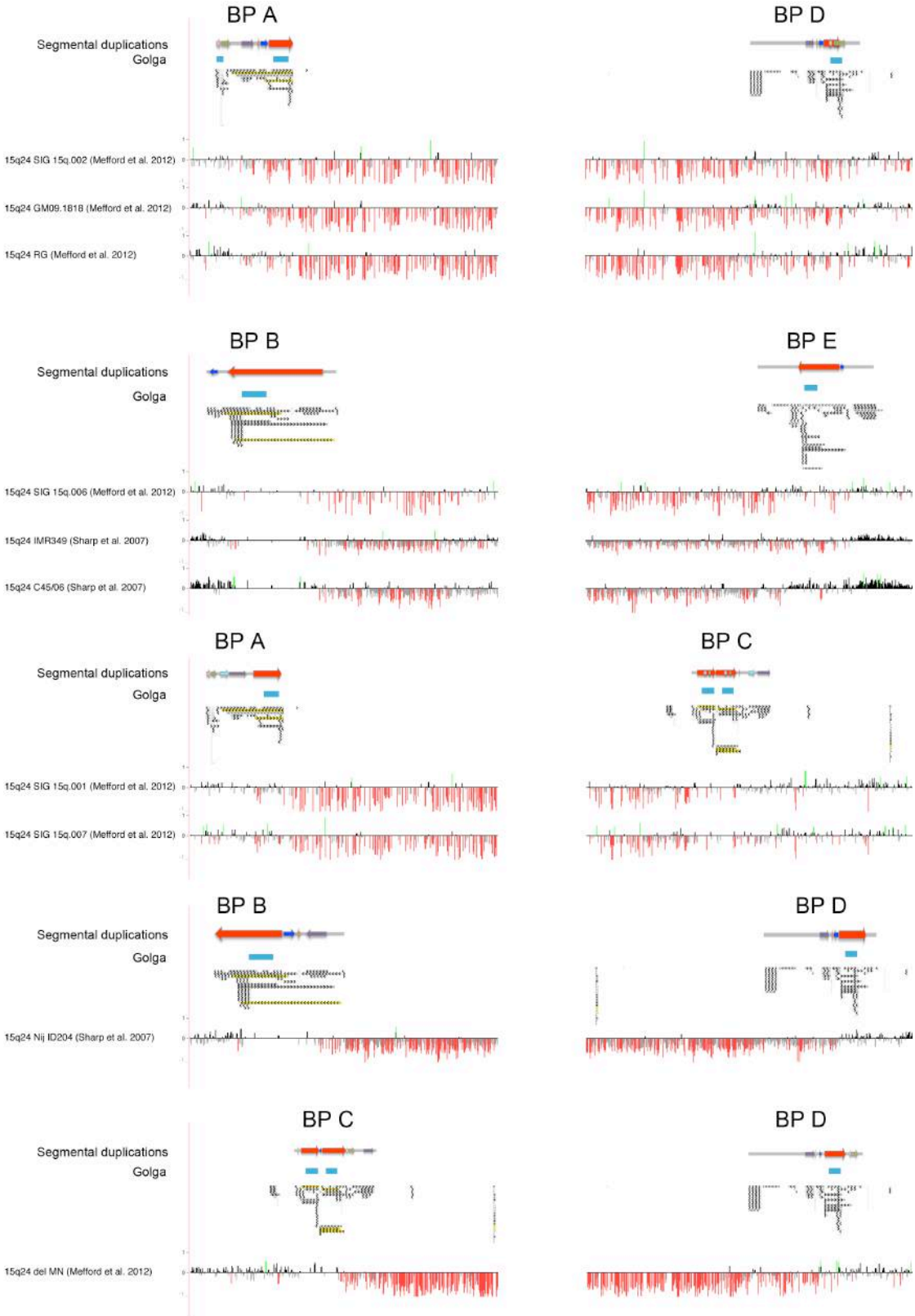
Supplementary Figure 24. Chromosome 15 rearrangement breakpoints. (a) Breakpoints of multiple chromosome 15 rearrangements, including the Prader-Willi/Angelman syndromes, 15q24 microdeletions, and 15q25.2 microdeletions coincide with the location of *GOLGA* gene family. **(b)** Diagram showing the localization of the *GOLGA* core duplicons within the duplication blocks where the rearrangement breakpoints map. **(c)** Analysis of the SDs mapping at the 15q24 microdeletion region show that the paralogous SDs mapping at the microdeletion breakpoints are short and have a low percentage of identity. The *GOLGA* repeat is mapping at all the breakpoints refined by array CGH (see Supplementary Figure 26).

Supplementary Figure 25



Supplementary Figure 25. Analysis of the SDs mapping at the 15q25 microdeletion region. Paralogous SDs mapping at the microdeletion breakpoints are short and have a low percentage of identity. The *GOLGA* repeat is mapping at all breakpoint regions analyzed.

Supplementary Figure 26



Supplementary Figure 26. High-resolution oligonucleotide-array data from ten 15q24 microdeletions cases. The location of paralogous SDs and *GOLGA* (blue bar) in each microdeletion region is indicated.

Supplementary Tables

Supplementary Table 1. Total number of individuals analyzed per each population.

Population group	Population	Number of individuals
AFR_AMR	ACB	90
AMR	CLM	79
SAN	BEB	85
AMR	PEL	78
AFR	LWK	86
AFR	MSL	83
EUR	GBR	62
EUR	IBS	104
AFR_AMR	ASW	64
EUR	TSI	87
ASN	KHV	96
EUR_AMR	CEU	94
AFR	YRI	101
ASN	CHB	83
AFR	GWD	113
SAN	STU	96
ASN	CHS	63
AFR	ESN	98
EUR	FIN	54
SAN_AMR	GIH	105
SAN	PJL	87
AMR	MXL	52
SAN	ITU	92
ASN	CDX	99
ASN	JPT	90
AMR	PUR	84
Chimpanzee		23
Bonobo		14
Gorilla		32
Orangutan		17
Denisova		1
Neanderthal		1

Supplementary Table 4. Orientation of the γ inversion locus and CN status of CNPα in 20 HapMap individuals. FISH analysis shows an enrichment of the γ inversion among individuals with a CNPα CN of 2 (p-value<0.0001, Chi-square test).

Individuals	Population	Total CN of α	CN at BP4	CN at BP5	γ inversion status
NA18603	East Asian	2	0;1	0;1	inverted/direct
NA19207	African	2	0;1	0;1	inverted/direct
NA20543	European	2	1;1	0;0	inverted/inverted
NA20786	European	2	1;1	0;0	inverted/inverted
NA19107	African	2	0;0	1;1	direct/direct
NA19373	African	2	0;1	0;1	inverted/direct
NA12878	European American	3	1;1	0;1	inverted/direct
NA18523	African	3	0;1	1;1	direct/direct
NA19138	African	3	1;1	0;1	inverted/direct
NA11831	European American	3	1;1	0;1	inverted/direct
NA19201	African	3	0;1	1;1	direct/direct
NA18498	African	3	0;1	1;1	direct/direct
NA19238	African	4	1;1	1;1	direct/direct
NA18502	African	4	1;1	1;1	direct/direct
NA06994	European American	4	1;1	1;1	direct/direct
NA07037	European American	4	1;1	1;1	direct/direct
NA18573	East Asian	4	1;1	1;1	direct/direct
NA12156	European American	5	1;1	1;2	direct/direct
NA18943	East Asian	5	1;1	1;2	direct/direct
NA12155	European American	5	1;1	1;2	direct/direct

Supplementary Table 6. CH17-derived β inversion diagnostic variants.

β inversion variant	hg19/GRCh37 coordinates	hg19 allele	CH17 allele
1	chr15:30738335-30738336	T	C
2	chr15:30743391-30743392	A	G
3	chr15:30756366-30756367	T	C
4	chr15:30799340-30799341	A	-
5	chr15:30808314-30808315	C	T
6	chr15:30808707-30808708	C	T
7	chr15:30815366-30815367	G	A
8	chr15:30816428-30816429	C	T

Supplementary Table 7. List of 26 primers used to amplify 13 sequence-tagged site and perform library hybridizations of VMRC53 and VMRC54 BAC libraries filters.

Primers	Sequence
Binv_1F	GCATCCTGGGACTCGTTCTA
Binv_1R	GCATTTCCTCCTACCAGCAG
Binv_2F	AATCTGGCAATAGGCGAGAA
Binv_2R	TGGCTGCTGTTTCATGTCTC
15q13_1 F	AGGGCAGAATTCCTCCAGAT
15q13_1 R	TGATCAGGGGTTAGCCAGAC
15q13_2 F	GGTGAGTCTGGAAGCAAAGC
15q13_2 R	TCGCTGCTCATTTCATCAAC
15q13_3 F	GAGCGTTAGTGCTGGAAAGG
15q13_3 R	GGAGCAGGATCATTCCTCAG
15q13_4 F	TCAGTTAGCTGCTGGCATTTT
15q13_4 R	AAGAAGCACAGGTGAGCAAGA
15q13_5 F	AATCTGGCAATAGGCGAGAA
15q13_5 R	TGGCTGCTGTTTCATGTCTC
15q13_6 F	ATATCAGCCCCTTGCATGTC
15q13_6 R	GGGTAGAGGCAGAAGCACAG
15q13_7 F	TCACCACCACCAAAGACAAA
15q13_7 R	ATGCCTCACGTTTTTCAACC
15q13_8 F	GGATTTGGACCCACCTTTTT
15q13_8 R	GCGACACACAGTGGAGGATA
15q13_9 F	CACAGTTTGGGCAAAGTTCA
15q13_9 R	GGGTATGCTTTGAAATGAGCA
15q13_10 F	TTGTCTGCCCTGAAACATTG
15q13_10 R	GGACCTGTGTGGCTTGAAGT
15q13_11 F	GAATGGAGCAAGTTGTGCCTA
15q13_11 R	TCAAATTCCAAAATCGAGCA

Supplementary Table 8. BAC libraries constructed for each member of the NA12878 trio.

Individual	Library	Vector	Restriction enzyme	Plate Numbers	Plate count	Recombinant Clones	Average insert size	Genomic Coverage
NA12878	VMRC53	pCC1BAC	EcoRI	1-528	528	> 95%	163 kbp	10X
NA12891	VMRC54	pCC1BAC	EcoRI	1-528	528	> 95%	177 kbp	10X
NA12892	VMRC57	pCC1BAC	EcoRI	1-528	528	> 95%	150 kbp	10X

Supplementary Table 11. Sequence pairs identified using GENECONV.

Pairwise Fragment Names	Simulated Pairwise P-value	Alignment Begin (bp)	Alignment End (bp)	Length (bp)	Number Polymorphic Sites	Number of Discordant Sites	Total Differences	Gene
Beta_inverse_proximal; Beta_direct_distal	0.045	2265	2481	217	9	0	229	
Beta_inverse_proximal; Beta_direct_distal	0	2483	4089	1607	16	0	229	
Beta_direct_proximal; Beta_inverse_distal	0.002	2419	4487	2069	22	0	139	
Beta_direct_proximal; Beta_inverse_distal	0	7482	8450	969	11	0	265	
Beta_inverse_proximal; Beta_inverse_distal	0	27408	28757	1350	35	0	118	*GOLGAR*
Beta_direct_proximal; Beta_inverse_distal	0	27516	28757	1242	33	0	139	*GOLGAR*
Beta_direct_proximal; Beta_inverse_proximal	0	27516	28951	1436	38	0	169	*GOLGAR*
Beta_direct_proximal; Beta_inverse_distal	0.002	28759	29697	939	22	0	139	*GOLGAR*
Beta_direct_proximal; Beta_inverse_proximal	0.004	28953	29697	745	17	0	169	*GOLGAR*
Beta_inverse_proximal; Beta_inverse_distal	0.026	28953	29959	1007	22	0	118	*GOLGAR*
Beta_inverse_proximal; Beta_inverse_distal	0	33353	42310	8958	53	0	118	
Beta_inverse_proximal; Beta_inverse_distal	0.002	42355	53303	10949	27	0	118	

Supplementary Table 16. β inversion pairwise Fst between populations.

	LWK	MKK	YRI	ESN	GWD	ASW	CHB	CDX	KHV	JPT	TSI	CEU	GBR	MXL
LWK	0.000													
MKK	0.000	0.000												
YRI	0.000	0.000	0.000											
ESN	0.000	0.000	0.000	0.000										
GWD	0.000	0.000	0.000	0.000	0.000									
ASW	0.105	0.079	0.075	0.070	0.051	0.000								
CHB	0.019	0.032	0.034	0.036	0.054	0.202	0.000							
CDX	0.063	0.082	0.088	0.096	0.115	0.253	0.019	0.000						
KHV	0.061	0.078	0.084	0.092	0.109	0.245	0.017	0.000	0.000					
JPT	0.000	0.000	0.000	0.000	0.006	0.123	0.010	0.055	0.052	0.000				
TSI	0.256	0.224	0.217	0.208	0.185	0.049	0.334	0.363	0.356	0.270	0.000			
CEU	0.185	0.158	0.152	0.146	0.126	0.017	0.258	0.289	0.283	0.199	0.003	0.000		
GBR	0.161	0.133	0.127	0.121	0.100	0.003	0.243	0.280	0.273	0.177	0.014	0.000	0.000	
MXL	0.045	0.028	0.026	0.023	0.011	0.003	0.124	0.174	0.168	0.059	0.105	0.058	0.036	0.000

Supplementary Table 17. Primate cell lines and BAC libraries used to study the organization of the 15q13.3 locus in nonhuman primates.

	Chimpanzee	Gorilla	Orangutan
Cell lines	PTR Cochise (University of Washington) PTR Katie (University of Washington) PTR Veronica (University of Washington) PTR8 (University of Bari)	GGO5 (University of Bari) GGO8 (University of Bari)	PPY9 (University of Washington) PPY16 (University of Bari) PPY13 (University of Bari)
BAC libraries	CHORI-251 Pan troglodytes (Clint)	CHORI-277 Gorilla gorilla gorilla (Kamilah)	CHORI-276 Pongo abelii (Susie)

Supplementary Table 22. Illumina sequencing of two 15q13.3 microdeletion probands (p1, proband) and their parents (mo, mother; fa, father).

Sample	Lanes	Reads	Read length	Coverage	Total sequence generated (bp)
13301.p1	2	714,835,326	101	16	72,198,367,926
13301.mo	3	850,024,105	101	12	85,852,434,605
13301.fa	3	892,030,336	101	14	90,095,063,936
13647.p1	2	894,508,578	101	18	90,345,366,378
13647.mo	3	1,035,744,767	101	35	104,610,221,467
13647.fa	3	910,146,515	101	17	91,924,798,015

Supplementary Table 23. Genomic coordinates of *GOLGA8* repeats at 15q13.3.

Paralog	GRCh37 coordinates			Ori
GOLGA8J	chr15	30375158	30388904	+
GOLGA8T	chr15	30427990	30439395	+
GOLGA8R	chr15	30692765	30706463	-
GOLGAb12	chr15	30844452	30857487	+
GOLGA8H	chr15	30896233	30910029	+
GOLGAb14	chr15	31083680	31094063	+
GOLGA8K	chr15	32681786	32695493	-
GOLGA8O	chr15	32734115	32747835	-
GOLGA8N	chr15	32885657	32899511	+

Supplementary Table 24. Genomic coordinates of 15q13.3 structural variation breakpoints.

Region	Rearrangement breakpoint	GRCh37 coordinates
15q13.3	15q13.3 microdeletion (13301.p1) prox BP	chr15 30615000 30770000
15q13.3	15q13.3 microdeletion (13301.p1) dist BP	chr15 32754000 32784000
15q13.3	15q13.3 microdeletion (13647.p1) prox BP	chr15 30907142 30920936
15q13.3	15q13.3 microdeletion (13647.p1) dist BP	chr15 32886788 32909278
15q13.3	γ inversion prox BP	chr15 30858018 30889492
15q13.3	γ inversion dist BP	chr15 32702231 32733993
15q13.3	β inversion prox BP	chr15 30704161 30716095
15q13.3	β inversion dist BP	chr15 30834548 30846486
15q13.3	CNPα duplication (BP4) prox BP	chr15 30370610 30447337
15q13.3	CNPα duplication (BP4) dist BP	chr15 30672134 30730016
15q13.3	CNPβ duplication (BP4) prox BP	chr15 30672134 30730016
15q13.3	CNPβ duplication (BP4) dist BP	chr15 30841068 30913084
15q13.3	CNPα duplication (BP5) prox BP	chr15 32445406 32445460
15q13.3	CNPα duplication (BP5) dist BP	chr15 32680833 32749777
15q13.3	CNPβ duplication (BP5) prox BP	chr15 32680833 32749777
15q13.3	CNPβ duplication (BP5) dist BP	chr15 32865039 32901889
15q13.3	*ARHGAP11B* duplication (BP4) prox BP	chr15 30896233 30910029
15q13.3	*ARHGAP11B* duplication (BP4) dist BP	chr15 30935158 30935159
15q13.3	*ARHGAP11A* duplication (BP5) prox BP	chr15 32885657 32899511
15q13.3	*ARHGAP11A* duplication (BP5) dist BP	chr15 32925114 32925115
15q13.3	*CHRNA7*-adj duplication (BP4) prox	chr15 30970419 30970420
15q13.3	*CHRNA7*-adj duplication (BP4) dist	chr15 31073600 31094063