nature genetics

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

Francesca Antonacci[1,8], Megan Y Dennis[2,8], 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] & Evan E Eichler[2,3]

Recurrent deletions of chromosome 15q13.3 associate with intellectual disability, schizophrenia, autism and epilepsy. To gain insight into the instability of this region, we sequenced it in affected individuals, normal individuals and nonhuman primates. We discovered five structural configurations of the human chromosome 15q13.3 region ranging in size from 2 to 3 Mb. These configurations arose recently (~0.5–0.9 million years ago) as a result of human-specific expansions of segmental duplications and two independent inversion events. All inversion breakpoints map near *GOLGA8* core duplicons—a ~14-kb primate-specific chromosome 15 repeat that became organized into larger palindromic structures. *GOLGA8*-flanked palindromes also demarcate the breakpoints of recurrent 15q13.3 microdeletions, the expansion of chromosome 15 segmental duplications in the human lineage and independent structural changes in apes. The significant clustering ($P = 0.002$) of breakpoints provides mechanistic evidence for the role of this core duplicon and its palindromic architecture in promoting the evolutionary and disease-related instability of chromosome 15.

A ~2.5-Mb region on human chromosome 15q13.3, distal to the locus for the Prader-Willi and Angelman syndromes, represents one of the most genetically unstable regions of the human genome[1,2]. Rare recurrent microdeletions between blocks of segmental duplications (BP4 and BP5) are strongly associated with intellectual disability, schizophrenia, autism and other neurodevelopmental disorders[3–7]. Deletion in this region is in fact now recognized as one of the most prevalent major risk factors for idiopathic generalized epilepsy (present in ~1% of all cases)[4]. Reciprocal duplication as well as smaller internal deletions that encompass the entire *CHRNA7* gene have also been described in individuals with a range of neurodevelopmental phenotypes[8,9]. Numerous additional structural variants, including common copy number polymorphisms (CNPs) and an inversion polymorphism, have been reported within the 15q13.3 region[3,8,10,11]. The majority of the common and rare 15q13.3 structural polymorphisms are associated with complex, high-identity blocks of segmental duplications that arose recently in primate evolution[12–17]. Because of the genomic complexity of the region, neither the extent of human structural diversity nor the breakpoints of most rearrangement events are understood at the molecular genetic level.

In this study, we sought to better understand the mechanisms leading to the genomic instability of the 15q13.3 locus by characterizing the breakpoints of evolutionary and contemporary rearrangements. We used an integrated comparative genomics approach to sequence and characterize structural haplotypes from multiple human and ape genomes. This approach entailed construction of BAC libraries, high-quality finished sequencing using single-molecule real-time (SMRT) sequencing technology to resolve structural haplotypes[18] and cytogenetic-based assays to characterize the organization, orientation and segmental duplication architecture of the 15q13.3 region. We performed detailed sequence-based analysis of 80 15q13.3 microdeletions. Our results suggest a molecular convergence on specific repeat sequences as the potential source for genetic instability in these regions.

RESULTS

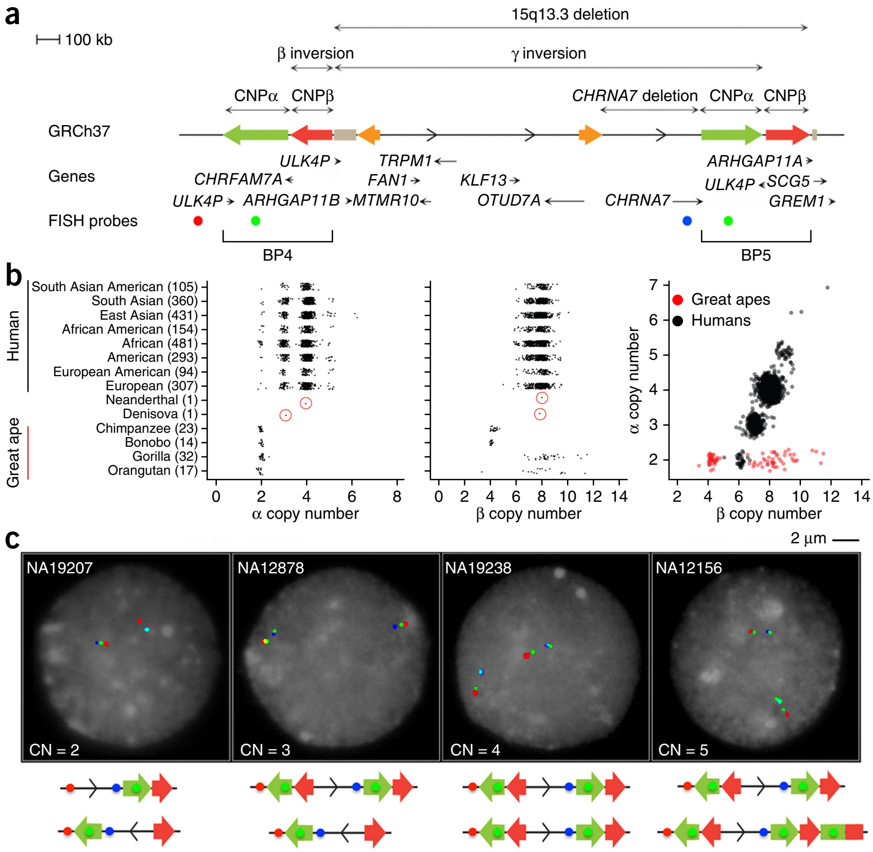
Copy number polymorphism

Because most breakpoints mapped to the large blocks of segmental duplications at BP4 and BP5 (**Fig. 1a** and **Table 1**), we first assessed the extent of CNP in these regions using sequence read-depth

[1] Dipartimento di Biologia, Università degli Studi di Bari Aldo Moro, Bari, Italy. [2] Department of Genome Sciences, University of Washington, Seattle, Washington, USA. [3] Howard Hughes Medical Institute, University of Washington, Seattle, Washington, USA. [4] Genome Institute at Washington University, Washington University School of Medicine, St. Louis, Missouri, USA. [5] Signature Genomic Laboratories, LLC, Spokane, Washington, USA. [6] Benaroya Research Institute at Virginia Mason, Seattle, Washington, USA. [7] Genetic Veterinary Sciences, Inc., Paw Print Genetics, Spokane, Washington, USA. [8] These authors contributed equally to this work. Correspondence should be addressed to E.E.E. (eee@gs.washington.edu).

Received 22 March; accepted 25 September; published online 19 October 2014; doi:10.1038/ng.3120

Figure 1 Structural variation at 15q13.3. (**a**) Different structural rearrangements at the 15q13.3 region include a 2-Mb microdeletion between BP4 and BP5 (ref. 3), a 430-kb microdeletion involving the *CHRNA7* gene[8], a 1.8-Mb polymorphic inversion of the same region (γ inversion)[3,10,11], two CNP segmental duplications (CNPα and CNPβ) mapping at BP4 and BP5 of the 15q13.3 microdeletion and a small inversion (β inversion) overlapping CNPβ at BP4. (**b**) Read depth–based copy number estimates for CNPα and CNPβ in 2,225 HapMap individuals from the 1000 Genome Project and 86 nonhuman ape, Neanderthal and Denisova genomes (circled in red). The number of individuals from each population is indicated in parentheses. A strong correlation ($r = 0.82$, Pearson correlation; $P < 2.2 \times 10^{-16}$, F test) in copy number is observed between CNPα and CNPβ in humans but not apes. (**c**) FISH analysis using a probe mapping at CNPα (WIBR2-1388I24, green) and two probes mapping in the unique sequence (WIBR2-1462O20, red; WIBR2-3158E16, blue) shows copy number varying between 0 and 1 at BP4 and between 0 and 2 at BP5.



approaches[19] applied to 2,313 human, ape and archaic hominin genomes (**Supplementary Tables 1–3**). We identified two large CNP regions of ~300 kb and ~210 kb, referred to here as CNPα and CNPβ, respectively. These two copy number–variable regions are separated by a *GOLGA8* repeat and correspond to two segmental duplications, each with >99.5% identity, in which the breakpoints of recurrent 2-Mb deletions were originally predicted to occur[3]. CNPα is a human-specific segmental duplication whose diploid copy number ranges from 2 to 7, with 77% of humans apparently fixed for the duplication (diploid copy number = 4) (**Fig. 1b** and **Supplementary Fig. 1**). In contrast, copy number states for CNPβ range from 5 to 12; 72% of humans show a diploid copy number of 8, with 4 of these copies mapping elsewhere on chromosome 15. A strong correlation ($r = 0.82$, Pearson correlation) in copy number was observed between CNPα and CNPβ, suggesting that in the human lineage (but not in the ape lineage) the two segmental duplications have expanded in concert as part of a larger 510-kb cassette.

On the basis of the extremes observed in this study, the data suggest that individuals in the human population might have genomes that differ in size by as much as 1 Mb with respect to segmental duplication

Table 1 Structural variant events at 15q13.3

Species	Structural variant	Size	GRCh37 coordinates	Disease or predisposition	Frequency
Human	15q13.3 microdeletion	2 Mb	Chr. 15: 30.7–32.7 Mb; chr. 15: 30.9–32.9 Mb	Intellectual disability, epilepsy, autism and schizophrenia	0.27% (42/15,767) cases; 0% (0/8,329) controls[1]
Human	15q13.3 microduplication	2 Mb	Chr. 15: 30.7–32.7 Mb; chr. 15: 30.9–32.9 Mb	Intellectual disability, epilepsy and autism	0.13% (20/15,767) cases; 0.3% (3/8,329) controls[1]
Human	γ inversion (BP4-BP5)	1.8 Mb	Chr. 15: 30.80–32.70 Mb	Predisposition to *CHRNA7* microdeletion	6% (haplotype freq.)
Human	β inversion (BP4)	130 kb	Chr. 15: 30.70–30.84 Mb	–	10% (haplotype freq.)
Human	CNPα duplication	300 kb	Chr. 15: 30.37–30.67 Mb; chr. 15: 32.45–32.75 Mb	–	CN = 4, 77%; CN = 3, 19%; CN = 2, 2%; CN = 5, 2%
Human	CNPβ duplication	210 kb	Chr. 15: 30.70–30.91 Mb; chr. 15: 32.68–32.89 Mb	–	CN = 8, 72%; CN = 7, 21%; CN = 9, 5%; CN = 6, 2%
Human	*CHRNA7*-adjacent duplication	124 kb	Chr. 15: 30.97–31.09 Mb	Predisposition to *CHRNA7* microdeletion	Fixed
Human	*ARHGAP11B* duplication	39 kb	Chr. 15: 30.90–30.93 Mb	–	Fixed
Chimpanzee	β inversion (BP5)	120 kb	Chr. 15: 32.7–32.8 Mb	–	ND
Gorilla	γ inversion	1.9 Mb	Chr. 15: 30.40–32.90 Mb	–	Fixed
Gorilla	Inversion of a portion of α (BP5)	80 kb	Chr. 15: 32.61–32.69 Mb	–	ND
Gorilla	Partial duplication of α (BP4)	80 kb	Chr. 15: 30.37–30.45 Mb	–	Fixed

Different structural rearrangements at the 15q13.3 region and the frequency of each rearrangement are shown. ND, not determined; CN, copy number; freq., frequency.

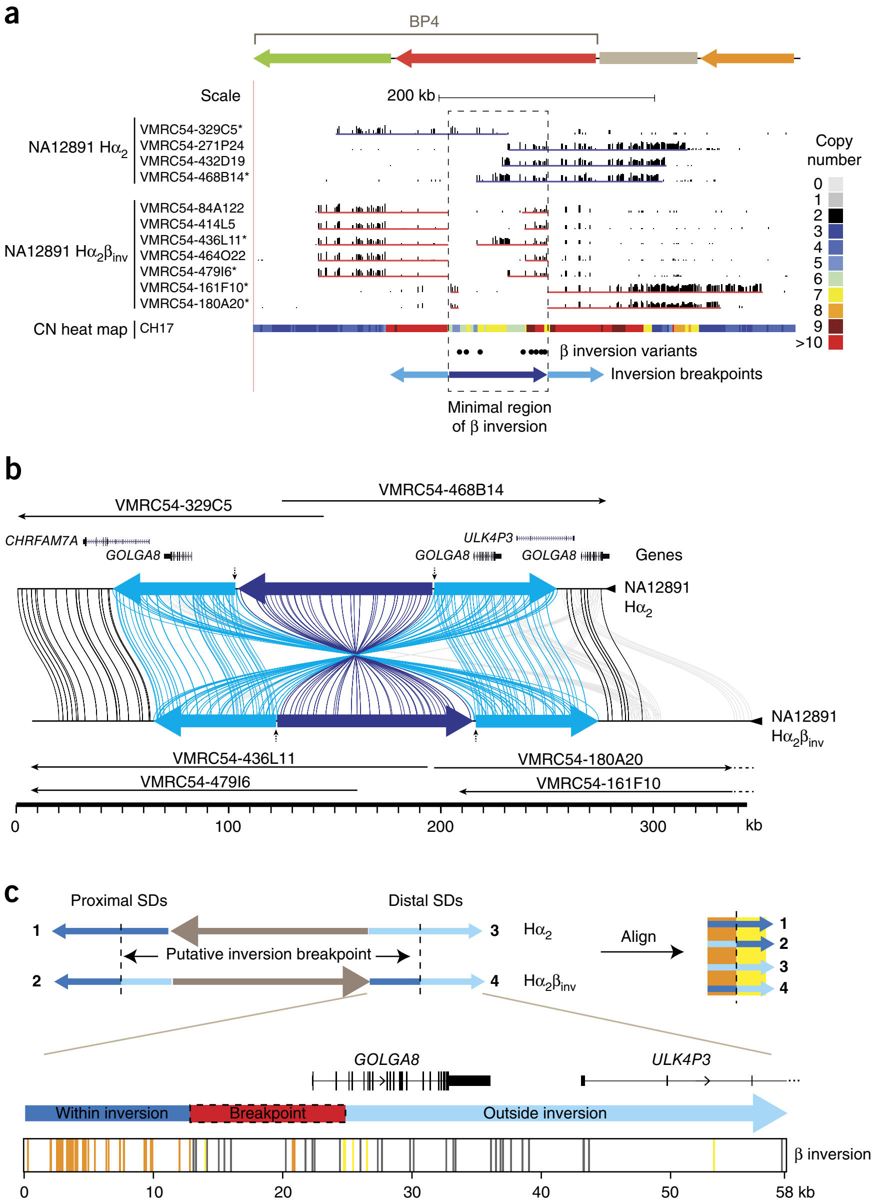


Figure 2 Sequence refinement of β inversion breakpoints. (**a**) A 210-kb β inversion was identified, validated and sequenced using the VMRC54 BAC library (NA12891 individual). Illumina-generated sequences of clones spanning the BP4 CNPβ were mapped to human reference GRCh37. Clones sequenced using SMRT sequencing are indicated with asterisks. The copy number (CN) heat map shows the total diploid copy number of a region in the CH17 hydatidiform mole cell line. The locations of the β inversion haplotype-tagging variants are shown as dots. The blue arrows represent the BP4 CNPβ (dark blue) with the flanking 58-kb inverted segmental duplications (light blue). (**b**) The homologous sequences of clones, generated using SMRT sequencing and assembled into sequence contigs, are connected with colored lines between the direct (Hα_2) and inverted (H$\alpha_2\beta_{inv}$) haplotypes from NA12891 using Miropeats[55]. Vertical arrows indicate the minimal inversion breakpoints. (**c**) Homologous sequences (58 kb) from the BP4 CNPβ flanking inverted segmental duplications (SDs) were aligned from multiple individuals (NA12891 and CH17) and haplotypes (β direct orientation, SD1 and SD3; β inverse orientation, SD2 and SD4; see **Supplementary Figure 5** for a more detailed alignment), and variant sites were compared. Variant positions showing signatures of being within or outside of the β inversion breakpoints are indicated as colored lines under the schematic of the distal β inverse segmental duplication including the following: within the inversion (orange; consensus of SD1 and SD4 and of SD2 and SD3), outside the inversion (yellow; consensus of SD1 and SD2 and of SD3 and SD4), and gene conversion (gray; consensus of SD1 and SD3 and of SD2 and SD4). The inversion breakpoint, refined to a region in which we observe a transition from orange to yellow lines, is highlighted with a dashed red box.

content between BP4 and BP5. We designed a series of three-color interphase FISH experiments to investigate the location of the copy number differences for CNPα among different individuals. The FISH analysis identified CNPs at both breakpoints of the 15q13.3 microdeletion. At a chromosomal level, we estimated a haploid variable copy number between 0 and 1 for BP4 and between 0 and 2 at BP5 (**Fig. 1c** and **Supplementary Table 4**). Because of additional copies of CNPβ mapping to chromosome 15, the BP4 and BP5 signals could not be clearly resolved by FISH for CNPβ.

Discovery and characterization of the β inversion

Because of the potential for assembly errors within segmental duplication regions[20–23], we established an alternate reference assembly for 15q13.3 from a hydatidiform (haploid) mole source (CHM1hTERT). We constructed a map of 23 contiguous BAC clones (CH17) and sequenced 21 of these using SMRT and capillary sequencing methods to establish a 4-Mb high-quality alternate reference assembly (**Supplementary Fig. 2a** and **Supplementary Table 5**). The new reference differed structurally from GRCh37 by a 130-kb inversion corresponding to CNPβ at BP4 (**Supplementary Fig. 3**). To ensure that the β inversion was not a hydatidiform mole cell line artifact, we identified a set of eight single-nucleotide variants that distinguished it from GRCh37 (**Supplementary Tables 6** and **7**) and screened additional DNA samples from the 1000 Genomes Project[24], identifying a European individual, NA12891, who was heterozygous for the β inversion. We constructed and arrayed a large-insert genomic BAC library from this DNA sample (VMRC54) as well as the two other members of the NA12878 trio (Online Methods and **Supplementary Table 8**). We recovered and sequenced VMRC54 BAC clones from the BP4 region, independently validating the sequence structure of the β inversion and the GRCh37 configuration (**Fig. 2a**,**b** and **Supplementary Table 9**). 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 in comparison to the directly oriented structural configuration (Hα_2) of the reference assembly.

The β inversion consists of three segmental duplications (**Supplementary Fig. 2b**): a pair of two highly identical (58-kb; 99.6% identity) inversely oriented segmental duplications flanking a 95-kb duplication. The flanking 58-kb palindrome corresponds to the *GOLGA8* gene family[15,16], one of the ancestral duplications, or

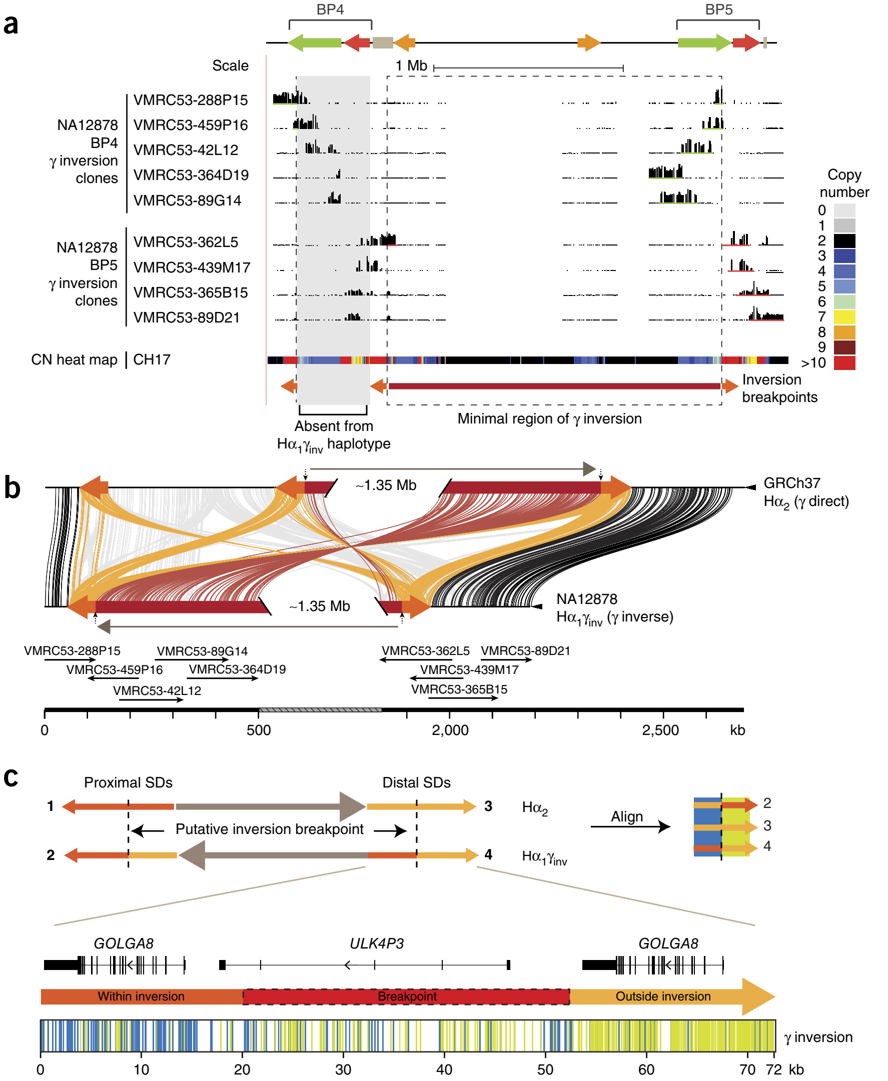


Figure 3 Sequence refinement of γ inversion breakpoints. (**a**) The γ inversion was identified, validated and sequenced using the VMRC53 BAC library (NA12878 individual). Illumina-generated sequences of clones spanning the 15q13.3 BP4 (green bars) and BP5 (red bars) loci were mapped to the human reference GRCh37. The nine clones shown were sequenced using SMRT sequencing. The copy number (CN) heat map shows the total diploid copy number of a region in the CH17 hydatidiform mole cell line. The minimal region of the inversion spans ~1.8 Mb (highlighted with a dashed box and a red bar). The orange arrows represent the flanking 72-kb inverted segmental duplications that mediate the γ inversion. The $H\alpha_1\gamma_{inv}$ haplotype likely arose from the $H\alpha_1$ haplotype, which does not harbor CNPα and CNPβ at BP4. (**b**) The homologous sequences of clones, generated using SMRT sequencing and assembled into contigs, and the human reference are connected with colored lines between the γ direct ($H\alpha_2$) and inverse ($H\alpha_1\gamma_{inv}$) haplotypes using Miropeats[55]. Vertical arrows indicate the minimal inversion breakpoints. (**c**) Homologous sequences (72 kb) from the orange flanking inverted segmental duplications (SDs) were aligned from multiple individuals (NA12878, CH17 and GRCh37) and haplotypes (γ direct orientation, SD3; γ inverse orientation, SD2 and SD4; see **Supplementary Figure 9** for a more detailed alignment), and variant sites were compared. Variant positions showing signatures of being within or outside of the γ inversion breakpoints are indicated as colored lines under the schematic of the distal γ inverse segmental duplication including the following: within the inversion (blue; consensus of SD2 and SD3) and outside the inversion (green; consensus of SD3 and SD4). The inversion breakpoint, refined to a region in which we observe a transition from blue to green lines, is highlighted with a dashed red box.

'core duplicons', found to be associated with most of the interspersed segmental duplication blocks across chromosome 15. The β inversion configuration increases the length of the largest contiguous tract of directly oriented segmental duplications between BP4 and BP5 from 58 to ~188 kb of near-perfect sequence (99.4% identity) (**Supplementary Fig. 4**), in principle creating a better substrate for unequal crossover and the instability associated with disease. To assess the breakpoints of the β inversion, we constructed a multiple-sequence alignment (MSA) from three distinct haplotypes (**Fig. 2c** and **Supplementary Fig. 5**) and used unique sequence differences in the duplicated regions to define the most likely breakpoint transition region. We narrowed the inversion breakpoint to a ~12-kb region spanning from intron 2 of the *GOLGA8* repeat to 9.6 kb upstream of the gene (**Fig. 2c** and **Supplementary Table 10**).

Sequence structure of the γ inversion

We sequence resolved the larger human inversion polymorphism spanning the entire BP4-BP5 region (referred to as the γ inversion)[3,10,11] (**Fig. 1**). This approach entailed the sequencing of 21 clones from a BAC library (VMRC53) constructed from a heterozygous individual (NA12878), SNP genotyping to assign maternal and paternal haplotypes, and high-quality sequencing of 11 non-redundant clones to generate an alternate reference assembly at the breakpoint region (**Fig. 3a**,**b**, **Supplementary Figs. 2c**, **6** and **7**, and **Supplementary Table 9**). In comparison to the reference, the γ inversion spans ~1.844 Mb from BP4 to BP5 and is flanked by palindromic segmental duplications containing two *GOLGA8* genes and a *ULK4P3* gene (~71 kb, 98.5% identity; **Fig. 3c** and **Supplementary Fig. 2d**). The $H\alpha_1\gamma_{inv}$ assembly contains a single copy of CNPα at BP4 and CNPβ at BP5, suggesting that it arose from a simpler human haplotype, $H\alpha_1$, with CNPα moved from BP5 to BP4 by the inversion (**Supplementary Fig. 8**). On the basis of sequence alignment, we refined the γ inversion breakpoints to a ~32-kb region within the palindrome containing the *ULK4P3* gene and flanked on either side by *GOLGA8* core duplicons (**Fig. 3c**, **Supplementary Fig. 9** and **Supplementary Table 10**). The high sequence identity of the duplications as well as the presence of alternative sequence signatures consistent with historical gene conversion events made it impossible to refine the breakpoint with any further precision (**Supplementary Table 11**).

Population frequency of β and γ inversion polymorphisms

To estimate the frequency of the γ inversion, we initially tested lymphoblastoid cell lines from 20 diverse HapMap individuals using a 3-color interphase FISH assay (**Supplementary Fig. 10**). Without exception, all chromosomal haplotypes (n = 16/16) with higher copy number for CNPα (n = 2–3) had the direct configuration similar to the reference genome (**Supplementary Table 4**). In contrast, all γ inversion haplotypes had a single copy of CNPα, consistent with our BAC sequencing results, and 10 of 24 of the chromosomes with a single copy of CNPα carried the γ inversion. An additional series of three-color FISH experiments confirmed that, when there is a single haploid copy number for CNPα and the haplotype carries the γ inversion (H$\alpha_1\gamma_{inv}$ configuration) (**Supplementary Fig. 11**), we always observe an absence of CNPα at BP5, consistent with a single structural haplotype for this inversion. Thus, CNPα copy number varies between 1 and 3 only in the directly oriented configurations for the BP4-BP5 region (Hα_1, Hα_2 and Hα_3) (**Supplementary Fig. 8**), with between 0 and 1 copies at BP4 and between 1 and 2 copies at BP5 (**Supplementary Table 4**). Combining this cytogenetic inference with the copy number data from 1,311 human genomes with ancestries matching those for our original FISH survey, we estimated an allele frequency of 6% for the γ inversion (**Supplementary Table 12**), with a slightly elevated frequency of

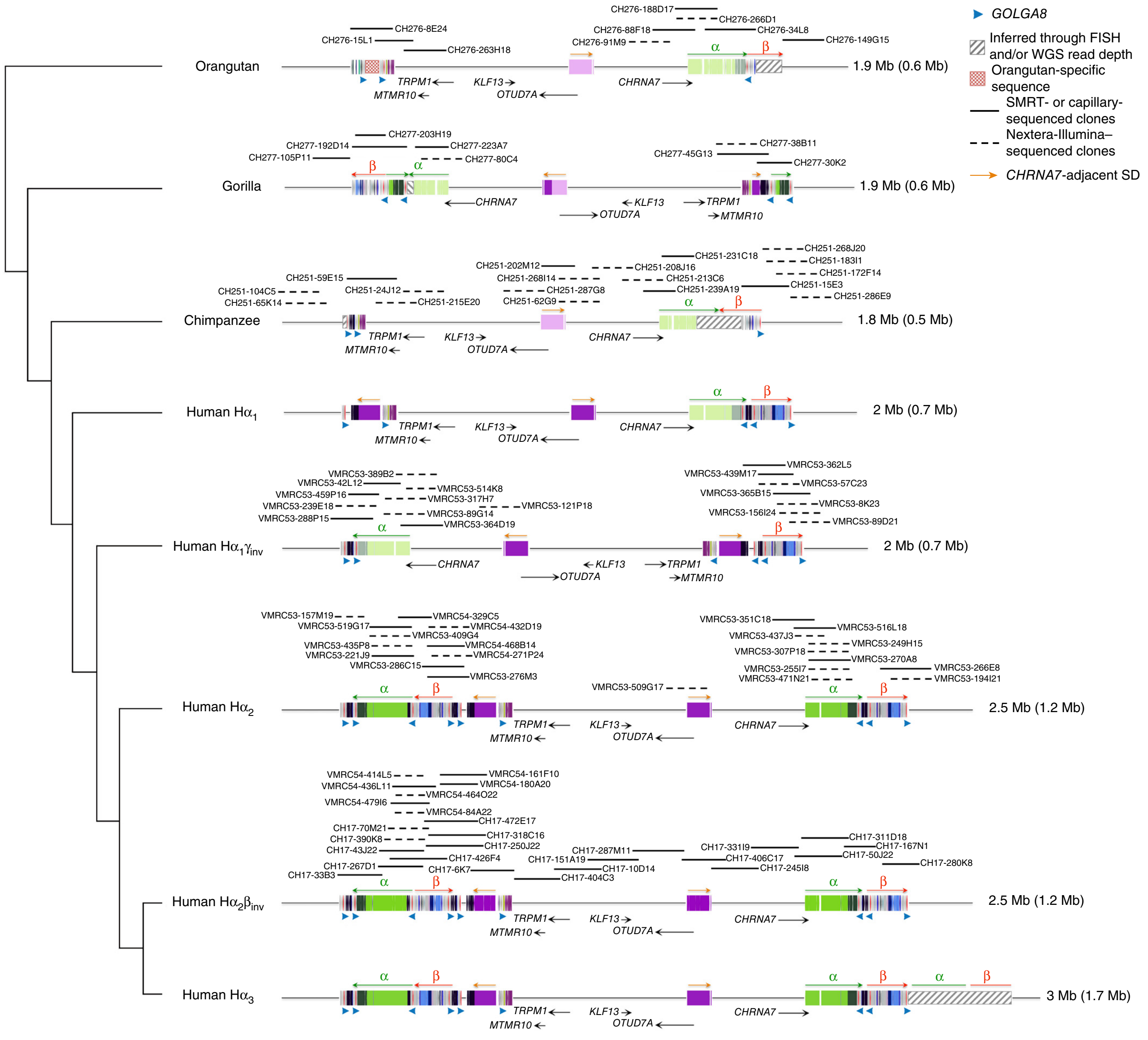


Figure 4 Comparative sequence analysis of the 15q13.3 region among apes. A schematic of the genomic structure is shown in the context of a generally accepted phylogeny of orangutan, gorilla, chimpanzee and human. A tiling path of BAC clones was sequenced for each haplotype (dashed lines, Illumina sequence; solid lines, SMRT capillary-finished sequence). A total of 66 BACs were completely sequenced and used to determine the segmental duplication organization (colored boxes). Colored boxes with lighter shades indicate segments that are single copy but are duplicated in other species. Nonhuman primates lack most of the larger duplications observed in humans (including CNPα and CNPβ) but do carry ancestral *GOLGA8* repeats. The region has expanded from 1.8–1.9 Mb in nonhuman apes to 2–3 Mb in humans as a result of the accumulation of segmental duplications. The size of each haplotype is indicated on the right, with the size of the duplicated bases shown in parentheses. The addition of a polymorphic 500-kb interval at BP4 occurred specifically in the human lineage, in association with an expansion of the *GOLGA8* repeats at BP4 (copy number = 6 in comparison to copy number = 2 in simpler human haplotypes and nonhuman primates). Using sequence and FISH data, chimpanzee and orangutan were found to have the direct orientation for the γ inversion, whereas gorilla was found to have the inverse orientation, suggesting that separate inversion events occurred at this locus across primate species. WGS, whole-genome sequencing. SD, segmental duplication.

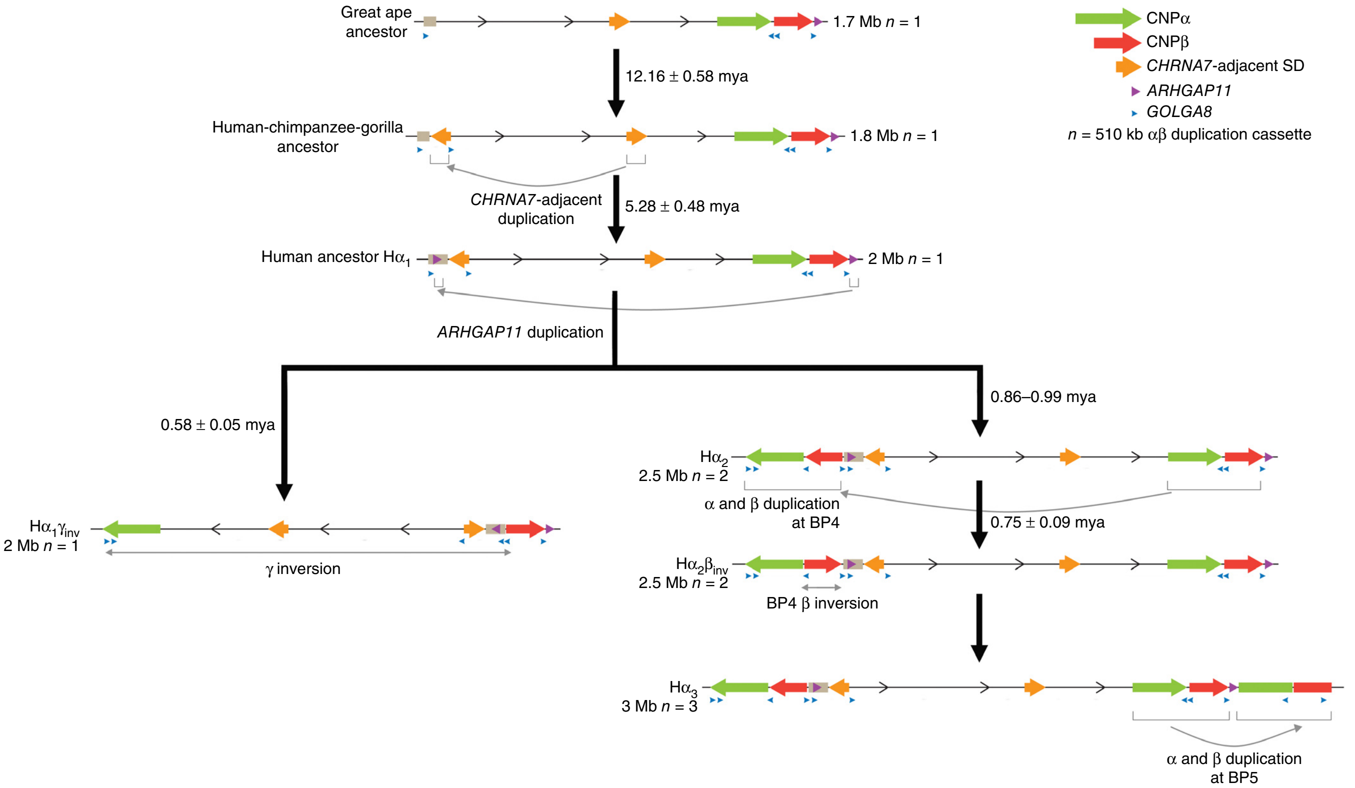


Figure 5 Model of chromosomal evolution at 15q13.3. On the basis of comparisons to outgroup primates, we propose a simpler human ancestral organization ($H\alpha_1$)—a configuration that is found to be enriched in contemporary African populations. A 510-kb duplicative transposition from BP5 to BP4 (α and β duplications) occurred potentially in a palindromic configuration ($H\alpha_2$) and was followed by an inversion of β at BP4 ($H\alpha_2\beta_{inv}$) from 700,000 to 900,000 years ago. NAHR within BP5 leads to tandemization of the 510-kb duplication ($H\alpha_3$) and larger configurations, primarily in East Asian populations. Approximately 500,000 years ago, the 1.8-Mb γ inversion independently rearranged to the $H\alpha_1\gamma_{inv}$ inverted haplotype. mya, million years ago. SD, segmental duplication.

the inversion in Tuscan and African populations. Notably, our inversion frequency estimate is lower than that previously reported[3]. In the previous study, a two-probe FISH assay was used to genotype the γ inversion, resulting in a higher error of detection in comparison to our study, which used a three-probe assay.

Because the β inversion is smaller and embedded within a complex region flanked by high-identity duplications, FISH could not be used to assess its frequency. Instead, we leveraged the unique tagging SNPs used to recover and sequence the inversion in NA12891. Using these SNPs as a surrogate, we designed molecular inversion probes (MIPs)[25,26] to capture, sequence (Illumina) and genotype the 8 haplotype-tagging variants across 904 individuals from diverse human populations in the 1000 Genomes Project (**Supplementary Tables 13–15**). We estimated a haplotype frequency of ~38% across European-ancestry populations (n = 275; CEU (Utah residents of Northern and Western European ancestry), TSI (Tuscan in Italy) and GBR (British in England and Scotland)), with reduced frequencies of ~10% in African populations (n = 299; LWK (Luhya in Webuye, Kenya), MKK (Massai in Kinyawa, Kenya), YRI (Yoruba in Ibadan, Nigeria), ESN (Esan in Nigeria) and GWD (Gambian in Western Division, The Gambia)) and ~4% in Asian populations (n = 221; CHB (Han Chinese in Beijing, China), CDX (Chinese Dai in Xishuangbanna, China), KHV (Kinh in Ho Chi Minh City, Vietnam) and JPT (Japanese in Tokyo, Japan)). No β inversion haplotypes were observed in either Chinese Dai (CDX) or Kinh (KHV) populations. These data suggest considerable stratification, especially between individuals of European and Asian ancestry (average F_{ST} = 0.28), with a maximum F_{ST} of 0.36 between the Tuscan (TSI) and Chinese Dai (CDX) populations (**Supplementary Table 16**).

Evolution of chromosome 15q13.3

We examined the organization of the region in multiple nonhuman ape samples by FISH and found that chimpanzee and orangutan showed a direct orientation between BP4 and BP5, whereas gorilla had an inverted orientation in comparison to the human reference genome (**Supplementary Fig. 12** and **Supplementary Table 17**). Next, we sequenced 48 BAC clones from chimpanzee, gorilla and orangutan to reconstruct the most likely ancestral sequence structure of the breakpoint regions (**Supplementary Fig. 13** and **Supplementary Table 18**). Sequencing data showed that both gorilla and orangutan lacked the *ULK4P3* gene where the γ inversion breakpoints mapped in humans, indicating that the 1.8-Mb γ inversion likely occurred as two independent events in the human and gorilla lineages. Recurrences of large inversion events across primate species have been reported for other regions, including the 17q21.31 and 16p12.1 microdeletion regions[23,27].

Our sequence analysis showed a much simpler organization of the 15q13.3 orthologous region in nonhuman primates in comparison to the human region (**Fig. 4**). The sequenced chimpanzee, gorilla and orangutan haplotypes, for example, lacked the large segmental duplications found in most human BP4 regions, predicting that BP5 was the ancestral source (**Supplementary Figs. 14** and **15**). Phylogenetic analysis confirmed this hypothesis and predicted that the proximal *GOLGA8* repeats at BP4 are orthologous among apes and humans

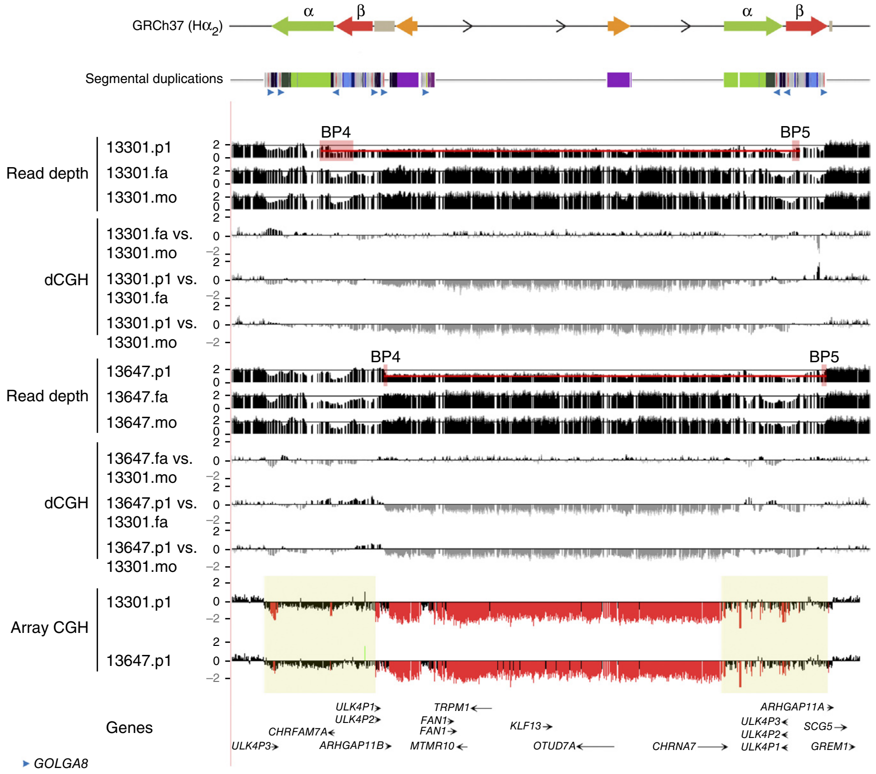


Figure 6 Analysis of 15q13.3 microdeletion breakpoints. Array CGH data for two 15q13.3 microdeletion samples from cases are mapped against the GRCh37 human reference. The microdeletion breakpoints map within a 500-kb region (yellow boxes), where both the α and β segmental duplications map. Digital comparative genomic hybridization (dCGH)[17] was used to detect regions of gain or loss in probands (.p1) in comparison to their parents (.mo, mother; .fa, father). The method measures differences in Illumina sequence read depth relative to a reference genome to define sites of copy number variation. Paralog-specific read-depth analysis in each proband and the parents was performed at all sites where both parents had the expected copy number of 2. This approach allowed us to refine the breakpoints for proband 13647.p1 to a 13-kb segment and the breakpoints for proband 13301.p1 to a 30-kb segment between BP4 and BP5 (red boxes). The two probands have different breakpoints, but in both cases the breakpoints map at or adjacent to *GOLGA8* repeats.

and, thus, preexisted the duplicative transpositions of CNPα and CNPβ to the region (**Supplementary Fig. 16** and **Supplementary Table 19**). The duplication also included *ARHGAP11*, a gene that was previously described to have undergone a human-specific expansion relative to other primate lineages[19]. In this case, we observed that the proximal breakpoint of the *ARHGAP11* duplication mapped within a *GOLGA8* repeat (13.8-kb resolution) (**Supplementary Fig. 17**).

There were numerous additional structural differences between apes and humans in this region. Our analysis showed that CNPβ maps in an inverted orientation in chimpanzee at BP5 (120-kb inversion), with a *GOLGA8* repeat defining at least one boundary of this chimpanzee-specific event (**Fig. 4** and **Supplementary Figs. 13** and **14**). A ~80-kb inversion of the distal portion of CNPα was identified in gorilla at BP5. This particular segment was also partially duplicated at BP4 in gorilla, and in both instances the rearrangement (duplication at BP4 and inversion at BP5) was flanked by *GOLGA8* repeats. Finally, the *CHRNA7*-adjacent segmental duplication (represented by the purple block with an orange arrow in **Fig. 4**) was completely absent at BP4 in all analyzed primates with the exception of the presence of a partial duplication in gorilla.

Interestingly, the distal breakpoint of the *CHRNA7*-adjacent duplication at BP4 in human mapped within a *GOLGA8* repeat.

To estimate the order and timing of the major structural changes during human evolution, we constructed a series of phylogenetic trees and estimated the coalescence/divergence time using locally calibrated molecular clocks and a predicted divergence time of 6 million years between human and chimpanzee. The earliest events in restructuring this region included the duplicative transposition of the adjacent *CHRNA7* segment to the proximal 15q13.3 region before the divergence of the African apes (12.16 ± 0.58 million years ago) (**Fig. 5** and **Supplementary Fig. 18**). This transposition was followed by the human-specific *ARHGAP11* duplication from BP5 (*ARHGAP11A*) to BP4 (*ARHGAP11B*), which occurred soon after humans and chimpanzees diverged (5.28 ± 0.48 million years ago) (**Supplementary Fig. 19**). We estimated that the largest CNPα and CNPβ duplications from BP5 to BP4 occurred in close succession or concurrently at 995 ± 61 and 862 ± 99 thousand years ago, respectively (**Supplementary Fig. 20**). These estimates are consistent with the finding that both duplications were already present before the split of Denisova and Neanderthal from the *Homo sapiens* lineage (**Fig. 1b**). Further, by comparing the sequences for the human CH17 contig (CNPβ_{inv}) and GRCh37 (CNPβ), we predicted that the β inversion occurred shortly thereafter, 748 ± 92 thousand

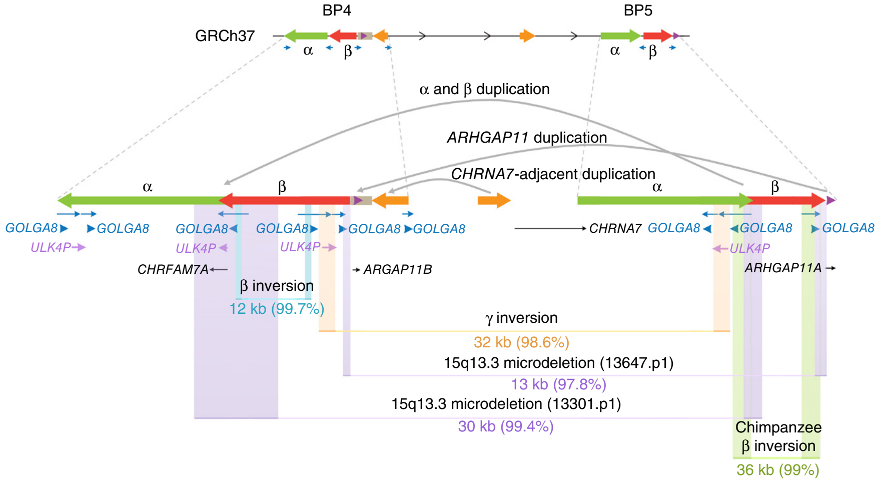


Figure 7 Summary of the 15q13.3 rearrangements mediated by *GOLGA8* repeats. Shown are eight independent rearrangements at the 15q13.3 region. Colored boxes indicate the breakpoints identified for each rearrangement (**Supplementary Table 10**). The size and percent similarity of the paralogous sequences at the rearrangement breakpoints are shown.

years ago (**Supplementary Fig. 20b**). We calculated that the time to the most recent common ancestor for the inverted NA12878 $H\alpha_1\gamma_{inv}$ and the $H\alpha_1$ haplotype (with the direct orientation for the γ inversion region) was 578 ± 47 thousand years ago (**Supplementary Fig. 21**).

Overall, these data suggest radical restructuring of this region in the *Homo* lineage over a short epoch of evolutionary time and a clear polarity of duplicative transposition events moving segments from BP4 to BP5 in association with *GOLGA8* repeats. These events have led to the emergence of at least five alternate chromosomal configurations in the human population ranging from 2 to 3 Mb in size.

15q13.3 microdeletion breakpoint analysis

We analyzed 80 total DNA samples from children with autism, intellectual disability and/or developmental delay who were previously identified as carrying 15q13.3 microdeletions by clinical array comparative genomic hybridization (CGH). These included 77 cases with intellectual disability and developmental delay referred to Signature Genomic Laboratories (24 unpublished and 53 previously reported)[1] and 3 cases with idiopathic autism from the Simons Simplex Collection (SSC)[28]. We screened the 80 subjects using complementary methods targeted to the 15q13.3 region: (i) a higher-density customized microarray and (ii) sequencing via MIP capture of singly unique nucleotide (SUN) *k*-mers (SUNKs). Both methods mapped the breakpoints of the disease-critical region to a ~500-kb region spanned by the CNPα and CNPβ segmental duplications (**Supplementary Figs. 22** and **23**, and **Supplementary Tables 20** and **21**).

Because the β inversion configuration creates a potentially more competent substrate for nonallelic homologous recombination (NAHR) owing to its longer stretch (188 kb) of directly oriented sequence, we tested whether this particular configuration was enriched in cases as has been observed for other microdeletion regions[23,29,30]. We compared the frequency of this configuration in cases and controls of European ancestry using sequence markers specific for the β inversion (**Fig. 2a**). We found that the frequency of the β inversion did not differ significantly in 15q13 microdeletion cases (~28% (n = 40) in comparison to the European-ancestry average (38%); P = 0.27, Fisher's exact two-tailed test; **Supplementary Table 15**). These data suggest that factors other than simply the length of homology promote the instability of this locus.

To refine the breakpoints with greater precision, we performed whole-genome sequencing of two idiopathic autism cases from the SSC carrying *de novo* 15q13.3 microdeletions along with their unaffected parents using the Illumina HiSeq 2000 platform (101-bp paired-end reads) (**Supplementary Table 22**). The sequences generated were aligned to the human GRCh37 reference and the alternate CH17 $H\alpha_2\beta_{inv}$ assembly. We investigated paralog-specific read depth over 1-bp windows in each trio at all sites where both parents had the expected copy number of 2. Using SUN variants that allowed us to discriminate between the paralogous copies[19], we narrowed the breakpoints for proband 13647.p1 to a 14-kb segment at BP4 and a 22-kb segment at BP5 and the breakpoints for proband 13301.p1 to a 155-kb segment at BP4 and a 30-kb segment at BP5 (**Fig. 6** and **Supplementary Table 10**). The two probands had different breakpoints, but in both cases the breakpoints mapped at or adjacent to directly oriented copies of *GOLGA8* (**Fig. 7**).

We tested by simulation whether the apparent clustering of evolutionary and disease-related 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 segmental duplications mapping to this portion of 15q13.3 (chr. 15: 30,362,914–31,196,467 and chr. 15: 32,442,314–32,927,877; **Supplementary Table 24**). We computed the number of times that the mean distance of the 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 (empirical P = 0.002, n = 100,000 permutations).

DISCUSSION

Our comparative sequence analysis of human and primate genomes shows that the 15q13.3 region has become increasingly complex over the course of human evolution, with an expansion in size from 1.8 Mb in apes to 2–3 Mb in humans. There has been a clear polarity, with most duplicative transpositions occurring from BP5 to BP4. Most of the largest structural changes, including large-scale inversion polymorphisms, arose over a narrow evolutionary period (500–900 thousand years ago)—at a time when ancestral *Homo sapiens* was diverging from archaic hominins[31,32]. We have resolved five distinct structural configurations in humans that differ radically in organization and segmental duplication content. Our results suggest that human chromosomal 15q13.3 haplotypes can vary by as much as 75% of their euchromatic length and are stratified among different populations. The simplest ancestral configurations (for example, $H\alpha_1$) show elevated frequency among African populations, whereas some of the largest and potentially disease-prone configurations are enriched in out-of-Africa populations (for example, $H\alpha_2\beta_{inv}$ in Europeans and $H\alpha_3$ in East Asians).

At least nine 15q13.3 rearrangement breakpoints (six human, one chimpanzee and two gorilla rearrangements) map at or adjacent to *GOLGA8* core duplicons (**Figs. 4**, **6** and **7**, and **Table 1**). Although our breakpoint precision ranges from 12 to 155 kb and cannot be further refined owing to the presence of virtually identical sequence within these regions (**Supplementary Table 10**), our simulations strongly suggest that this association is significant. The *GOLGA* repeat encodes a primate-specific chromosome 15 gene family of 14 kb[15] that expanded over the last 20 million years of primate evolution[12,13]. It has been dispersed to multiple locations across the long arm of chromosome 15 and is the most enriched sequence associated with segmental duplication blocks promoting disease-related instability, including those associated with the Prader-Willi and Angelman syndromes, 15q24 microdeletions and 15q25.2 microdeletions[33–36] (**Supplementary Fig. 24a**,**b**). *GOLGA* is 1 of 14 core duplicons associated with the burst of interspersed segmental duplications in the human–great ape ancestral lineage[17,37].

We propose that the *GOLGA* core duplicons are preferential sites of genomic instability that have driven both disease and the evolutionary instability of chromosome 15. In addition to showing the clustering of breakpoints on chromosome 15q13.3, other data are supportive of a more global association. We note, for example, that this same *GOLGA* repeat demarcates a pericentric inversion breakpoint between human and chimpanzee 15q11-q13 (ref. 38) and a more ancient inversion in the Catarrhini ancestor[39]. Analysis of the segmental duplications mapping at other chromosome 15 microdeletion regions (for example, 15q24 and 15q25) shows that the breakpoints often occur in directly orientated duplications that are short and have a low percentage of identity (**Supplementary Figs. 24c** and **25**) but contain multiple copies of the *GOLGA* repeats. Array CGH experiments on ten previously published 15q24 microdeletion cases confirmed that the *GOLGA* repeat maps at or near most rearrangement breakpoints

(**Supplementary Figs. 24c** and **26**)[33,40]. These observations are also consistent with our finding no evidence of an enrichment of the H$\alpha_2\beta_{\text{inv}}$ haplotype among 15q13.3 deletion cases, even though this configuration expands the directly orientated segment from 58 to 188 kb in length. Although orientation, length and degree of sequence identity for duplicated sequences are frequently deemed the most important parameters for NAHR[1,41], the presence of a *GOLGA* repeat might bias the actual position of the unequal crossover (i.e., representing an NAHR hotspot).

These results also bear striking similarities to those for the microdeletion encompassing the *NF1* gene (encoding neurofibromatosis type-1) and its flanking regions at 17q11.2. The most common *NF1* microdeletions (type 1) span 1.4 Mb and have breakpoints located within segmental duplications containing *LRRC37* core duplicons[42]. The same *LRRC37* core duplicons at 17q21.31 are known to have mediated the 970-kb polymorphic inversions of the *MAPT* locus that also underlies the syndromes associated with recurrent 17q21.31 microdeletions[27]. The presence of core duplicons at multiple evolutionary breakpoints as well as at a variety of recurring disease-associated rearrangements is indicative of the high degree of genomic instability driven by these sequences.

Our evolutionary reconstruction suggests that the *GOLGA8* core duplicon, in particular, has promoted both inversions and the formation of large palindromic segmental duplication structures. Palindromic sequences, or inverted repeats, have been known to be unstable and represent hotspots for deletion or recombination in bacteria, yeast and mammals[43–46]. This genetic instability has generally been related to DNA replication: slow replication was observed for an inverted repeat sequence in *Escherichia coli*[44], and inverted repeats lead to chromosomal rearrangements more frequently in yeast that are deficient in DNA polymerase activity[47,48]. In the events discussed here, the presence of palindromic structures might have promoted stalling of the replication fork, creating an opportunity for the chromosome to break, and recombination might have occurred in a non-allelic fashion using the homology of the *GOLGA* repeats. In humans, short palindromic AT-rich repeats (PATRRs) have been implicated in chromosomal aberrations via non-homologous end joining, leading to gross trans-chromosomal events[49] and instability in cancer cells[50]. Most experimental demonstrations of palindrome formation and instability have involved smaller structures. The putative palindromes here are massive—for instance, ~210 kb in length with 58-kb inverted arms flanking a 95-kb spacer for CNPβ—and attempts to detect their formation by *in vitro* snapback assays[50] were inconclusive.

The recurrent use of *GOLGA* core duplicons suggests that they have a fundamental role in the cycles of chromosomal rearrangement that have intertwined large-scale inversions and segmental duplication expansions in this region. We note that most of the largest interspersed segmental duplications have been transposed in an inverted orientation. Similar inverted configurations also occur for contemporary rearrangements such as at the *PLP1* locus, which is known to be associated with inverted repeats[51,52]. Microhomology-mediated break-induced replication (MMBIR) mechanisms might be responsible for the initial formation of segmental duplications[53,54], and sequences such as *GOLGA* might also represent preferred or 'fragile' sites for MMBIR. It is intriguing that the *GOLGA* repeats corresponding to sites of rearrangement and conversion maintain an ORF, whereas those repeats at the periphery have disrupted ORFs (**Supplementary Note**). We previously showed that core duplicon sequences are generally more transcriptionally active than unique or flanking duplicated sequence[16]. Thus, transcription and the maintenance of the ORF might be a critical feature of core duplicons, serving as seeds of genomic instability and punctuated segmental duplication in the human genome. The mechanism by which these elements promote evolutionary and disease-associated instability during replication will require future experimental investigation.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. BAC clone sequences generated using capillary and SMRT sequencing have been deposited in GenBank under the accessions listed in **Supplementary Tables 5**, **9** and **18**.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank M. Ventura, C. Campbell and H.C. Mefford for useful discussions and T. Brown for critical review of the manuscript. We also thank S. Diede, H. Tanaka, B. Brewer, C. Payen, L. Harshman and K. Penewit for experimental advice and support for the palindromic snapback assay. We are grateful to all of the families at the participating Simons Simplex Collection (SSC) sites, as well as the principal investigators (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). We appreciate obtaining access to phenotypic data on SFARI Base. This work was supported, in part, by US National Institutes of Health grants HG002385 and HG004120 to E.E.E. M.Y.D. is supported by the National Institute of Neurological Disorder and Stroke of the US National Institutes of Health (award K99NS083627). E.E.E. is an investigator of the Howard Hughes Medical Institute.

AUTHOR CONTRIBUTIONS

F.A., M.Y.D. and E.E.E. designed the study. F.A. performed FISH experiments, library construction for Illumina sequencing, array CGH experiments and sequence analysis. M.Y.D. performed MIP experiments, library construction for Illumina sequencing, array CGH experiments and sequence analysis. J.H. performed SMRT sequence analysis and haplotype reconstruction. P.H.S. and K.M.S. performed sequencing data analysis. T.A.G. and R.K.W. performed capillary sequencing and analysis of CH17 and nonhuman primate BAC clones. L.V. and M. Malig performed FISH experiments. M. Miroballo performed array CGH experiments. B.M. performed library construction for SMRT sequencing. L.D. performed MIP experiments and library construction for SMRT sequencing. A.R. performed SMRT sequence analysis. C.T.A., A.S. and J.T. performed library construction for the VMRC53, VMRC54 and VMRC57 BACs. J.A.R. and L.G.S. contributed to 15q13.3 microdeletion data collection. F.A., M.Y.D. and E.E.E. contributed to data interpretation. F.A., M.Y.D. and E.E.E. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Cooper, G.M. *et al.* A copy number variation morbidity map of developmental delay. *Nat. Genet.* **43**, 838–846 (2011).
2. Kaminsky, E.B. *et al.* An evidence-based approach to establish the functional and clinical significance of copy number variants in intellectual and developmental disabilities. *Genet. Med.* **13**, 777–784 (2011).
3. Sharp, A.J. *et al.* A recurrent 15q13.3 microdeletion syndrome associated with mental retardation and seizures. *Nat. Genet.* **40**, 322–328 (2008).
4. Helbig, I. *et al.* 15q13.3 microdeletions increase risk of idiopathic generalized epilepsy. *Nat. Genet.* **41**, 160–162 (2009).
5. International Schizophrenia Consortium. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* **455**, 237–241 (2008).
6. Stefansson, H. *et al.* Large recurrent microdeletions associated with schizophrenia. *Nature* **455**, 232–236 (2008).
7. Miller, D.T. *et al.* Microdeletion/duplication at 15q13.2q13.3 among individuals with features of autism and other neuropsychiatric disorders. *J. Med. Genet.* **46**, 242–248 (2009).
8. Shinawi, M. *et al.* A small recurrent deletion within 15q13.3 is associated with a range of neurodevelopmental phenotypes. *Nat. Genet.* **41**, 1269–1271 (2009).

9. Williams, N.M. *et al.* Genome-wide analysis of copy number variants in attention deficit hyperactivity disorder: the role of rare variants and duplications at 15q13.3. *Am. J. Psychiatry* **169**, 195–204 (2012).
10. Kidd, J.M. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56–64 (2008).
11. Antonacci, F. *et al.* Characterization of six human disease-associated inversion polymorphisms. *Hum. Mol. Genet.* **18**, 2555–2566 (2009).
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. Pujana, M.A. *et al.* Human chromosome 15q11-q14 regions of rearrangements contain clusters of LCR15 duplicons. *Eur. J. Hum. Genet.* **10**, 26–35 (2002).
14. Bailey, J.A. *et al.* Recent segmental duplications in the human genome. *Science* **297**, 1003–1007 (2002).
15. Zody, M.C. *et al.* Analysis of the DNA sequence and duplication history of human chromosome 15. *Nature* **440**, 671–675 (2006).
16. Jiang, Z. *et al.* Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution. *Nat. Genet.* **39**, 1361–1368 (2007).
17. Sudmant, P.H. *et al.* Evolution and diversity of copy number variation in the great ape lineage. *Genome Res.* **23**, 1373–1382 (2013).
18. Huddleston, J. *et al.* Reconstructing complex regions of genomes using long-read sequencing technology. *Genome Res.* **24**, 688–696 (2014).
19. Sudmant, P.H. *et al.* Diversity of human copy number variation and multicopy genes. *Science* **330**, 641–646 (2010).
20. Dennis, M.Y. *et al.* Evolution of human-specific neural *SRGAP2* genes by incomplete segmental duplication. *Cell* **149**, 912–922 (2012).
21. 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).
22. She, X. *et al.* Shotgun sequence assembly and recent segmental duplications within the human genome. *Nature* **431**, 927–930 (2004).
23. Antonacci, F. *et al.* A large and complex structural polymorphism at 16p12.1 underlies microdeletion disease risk. *Nat. Genet.* **42**, 745–750 (2010).
24. 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
25. Hardenbol, P. *et al.* Multiplexed genotyping with sequence-tagged molecular inversion probes. *Nat. Biotechnol.* **21**, 673–678 (2003).
26. O'Roak, B.J. *et al.* Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. *Science* **338**, 1619–1622 (2012).
27. Zody, M.C. *et al.* Evolutionary toggling of the *MAPT* 17q21.31 inversion region. *Nat. Genet.* **40**, 1076–1083 (2008).
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**, 221–237 (2013).
29. Steinberg, K.M. *et al.* Structural diversity and African origin of the 17q21.31 inversion polymorphism. *Nat. Genet.* **44**, 872–880 (2012).
30. Sharp, A.J. *et al.* Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. *Nat. Genet.* **38**, 1038–1042 (2006).
31. Meyer, M. *et al.* A mitochondrial genome sequence of a hominin from Sima de los Huesos. *Nature* **505**, 403–406 (2014).
32. Prüfer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43–49 (2014).
33. Mefford, H.C. *et al.* Further clinical and molecular delineation of the 15q24 microdeletion syndrome. *J. Med. Genet.* **49**, 110–118 (2012).
34. 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**, 777–781 (2010).
35. 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**, 370–386 (1999).
36. El-Hattab, A.W. *et al.* Redefined genomic architecture in 15q24 directed by patient deletion/duplication breakpoint mapping. *Hum. Genet.* **126**, 589–602 (2009).
37. Marques-Bonet, T. *et al.* A burst of segmental duplications in the genome of the African great ape ancestor. *Nature* **457**, 877–881 (2009).
38. Locke, D.P. *et al.* Refinement of a chimpanzee pericentric inversion breakpoint to a segmental duplication cluster. *Genome Biol.* **4**, R50 (2003).
39. Giannuzzi, G. *et al.* Hominoid fission of chromosome 14/15 and the role of segmental duplications. *Genome Res.* **23**, 1763–1773 (2013).
40. Sharp, A.J. *et al.* Characterization of a recurrent 15q24 microdeletion syndrome. *Hum. Mol. Genet.* **16**, 567–572 (2007).
41. Lupski, J.R. Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends Genet.* **14**, 417–422 (1998).
42. Bengesser, K. *et al.* A novel third type of recurrent *NF1* microdeletion mediated by nonallelic homologous recombination between *LRRC37B*-containing low-copy repeats in 17q11.2. *Hum. Mutat.* **31**, 742–751 (2010).
43. Gordenin, D.A. *et al.* Inverted DNA repeats: a source of eukaryotic genomic instability. *Mol. Cell. Biol.* **13**, 5315–5322 (1993).
44. Leach, D.R. Long DNA palindromes, cruciform structures, genetic instability and secondary structure repair. *Bioessays* **16**, 893–900 (1994).
45. Collick, A. *et al.* Instability of long inverted repeats within mouse transgenes. *EMBO J.* **15**, 1163–1171 (1996).
46. Akgün, E. *et al.* Palindrome resolution and recombination in the mammalian germ line. *Mol. Cell. Biol.* **17**, 5559–5570 (1997).
47. Ruskin, B. & Fink, G.R. Mutations in *POL1* increase the mitotic instability of tandem inverted repeats in *Saccharomyces cerevisiae*. *Genetics* **134**, 43–56 (1993).
48. Lemoine, F.J., Degtyareva, N.P., Lobachev, K. & Petes, T.D. Chromosomal translocations in yeast induced by low levels of DNA polymerase a model for chromosome fragile sites. *Cell* **120**, 587–598 (2005).
49. Inagaki, H. *et al.* Two sequential cleavage reactions on cruciform DNA structures cause palindrome-mediated chromosomal translocations. *Nat. Commun.* **4**, 1592 (2013).
50. Tanaka, H., Bergstrom, D.A., Yao, M.C. & Tapscott, S.J. Widespread and nonrandom distribution of DNA palindromes in cancer cells provides a structural platform for subsequent gene amplification. *Nat. Genet.* **37**, 320–327 (2005).
51. 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).
52. 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).
53. 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).
54. Payen, C., Koszul, R., Dujon, B. & Fischer, G. Segmental duplications arise from Pol32-dependent repair of broken forks through two alternative replication-based mechanisms. *PLoS Genet.* **4**, e1000175 (2008).
55. Parsons, J.D. Miropeats: graphical DNA sequence comparisons. *Comput. Appl. Biosci.* **11**, 615–619 (1995).

ONLINE METHODS

FISH analysis. Interphase nuclei and metaphase spreads were obtained from lymphoblast and fibroblast cell lines from 20 human HapMap individuals (Coriell Cell Repository), 4 chimpanzees (Katie, Veronica, Cochise and PTR8), 2 gorillas (GGO5 and GGO8) and 3 orangutans (PPY9, PPY16 and PPY13). All cell lines were tested for mycoplasma contamination. Primate cell lines were previously collected at the University of Washington and at the University of Bari (**Supplementary Table 17**) and have not been authenticated. FISH experiments were performed using fosmid clones directly labeled by nick translation with Cy3-dUTP (PerkinElmer), Cy5-dUTP (PerkinElmer) and fluorescein-dUTP (Enzo) as described previously[23]. A minimum of 50 interphase cells were scored for each inversion to statistically determine the orientation of the examined region.

Copy number variation analysis. Array CGH was performed on 80 samples with 15q13.3 microdeletions using custom high-density 4x180K Agilent oligonucleotide chips targeted with a density of 1 probe per 100 bp. Labeling, hybridization, scanning and data processing were performed as directed by the manufacturer. DNA sample NA19240 was used as a reference. We estimated the copy number of the 15q13.3 segmental duplications among 2,225 HapMap individuals of different ancestry[24] using a sequence read-depth method[19]. The duplication content of human, chimpanzee, gorilla, orangutan and macaque samples was determined using the whole-genome shotgun sequence detection (WSSD) method as described[37].

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). High-molecular-weight DNA was isolated, partially digested with EcoRI and subcloned into the pCC1BAC vector (Epicentre) to create libraries with an insert size of >150 kb using previously described protocols[56]. Clones were plated into 384-well microtiter plates and were transferred to high-density nylon filters for library screening.

Illumina sequencing of BAC clones. DNA from the CH17, VMRC53, VMRC54, CH251, CH276 and CH277 BAC clone libraries was isolated, prepped into barcoded genomic libraries and sequenced (PE101) on an Illumina HiSeq 2000 instrument using a Nextera protocol[29]. Sequencing data (~300-fold coverage) were mapped with mrsFAST[57] to the reference genome, and SUN identifiers were used to discriminate between highly identical segmental duplications[19].

SMRT clone sequencing and assembly. DNA was isolated from the CH17, VMRC53, VMRC54, CH251, CH277 and CH276 BAC clones, and SMRT SMRTbell libraries were prepared and sequenced using RSII C2P4 chemistry (one SMRT cell/ BAC sample with two 45-min movies). Inserts were assembled using Quiver and HGAP as described[18]. Alternate human genome assemblies, including SMRT- and capillary-sequenced clones from the CH17, RP11 and VMRC53 BAC libraries, were assembled with Sequencher and compared to the human reference genome using Miropeats[55] and BLAST[58].

Sequence analyses. MSAs of representative human haplotypes, paralogs and/or orthologs from human, chimpanzee, gorilla and orangutan were generated using Clustal W[59]. We constructed a series of phylogenetic trees using the neighbor-joining method with a complete deletion option (MEGA5)[60]. Genetic distances were calculated using the Kimura 2-parameter with standard error estimates (an interior branch test of phylogeny; n = 500 bootstrap replicates); Tajima's relative rate test was used to assess the validity of the molecular clock. We then estimated the coalescence/divergence time using the equation $T = K/2R$ (where K is divergence, R is the substitution rate and T is time) and an estimated divergence time of 6 million years between human and chimpanzee and 15 million years between human and orangutan.

Whole-genome sequencing of 15q13.3 microdeletion samples. For the SSC autism trios (proband, father and mother) 13301 and 13647, 3 µg of genomic DNA was sheared and end repaired, an A-tail was added and adaptors were ligated to the fragments as described[61]. Afterward, ligation samples were run on a 6% precast polyacrylamide gel (Invitrogen, EC6265BOX). The band at 400–550 bp was excised, diced and incubated as described above. Size-selected fragments were amplified with 0.5 µl of primers, 25 µl of 2× iProof, 0.25 µl of SYBR Green and 8.25 µl of distilled water under the following conditions: 98 °C for 30 s, 30 cycles of 98 °C for 10 s, 60 °C for 30 s, 72 °C for 30 s and 72 °C for 15 s, and a final 72 °C for 2 min. Fluorescence was assessed between the 30-s and 15-s steps at 72 °C. Amplified, size-selected libraries were quantified using an Agilent 2100 Bioanalyzer, and paired-end sequencing (101-bp reads) was performed on an Illumina HiSeq 2000 instrument. Sequence read depth corresponding to SUNs was used to refine the breakpoints as previously described[19]. dCGH was performed on the sequences from these samples using previously described methods[17].

Molecular inversion probe genotyping. We used 70-bp MIPs to capture and sequence the β inversion haplotype-tagging variants (n = 8) and SUNKs (n = 235) spanning the 15q13.3 region. The β inversion haplotype-tagging variants were identified from an MSA of CNPβ at BP4 and BP5 from our CH17-derived assembly and the human reference (**Supplementary Table 6**). We identified 3,544 SUNKs across the 15q13.3 region (chr. 15: 30,350,000–32,950,000; GRCh37) using previously described methods[19]. MIP design, capture and sequencing were performed as previously described[26,62]. MIP sequences are listed in **Supplementary Table 13**. Any individual with fewer than 5,000 reads mapping was removed from subsequent analyses. In the case of the β inversion haplotype-tagging variants, we genotyped an individual as carrying the β inversion if there was at least one read mapping to seven of the eight variants (**Supplementary Table 14**).

Human subjects. The human samples 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. Samples were collected at respective institutions after receiving informed consent and approval by the appropriate institutional review boards. There are no new health risks to participants. Samples that fall within this category include probands with autism and their parents from the SSC, probands with intellectual disability and developmental delay referred to Signature Genomic Laboratories and individuals from representative human populations from the 1000 Genomes Project.

56. 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**, 381–389 (2010).
57. Alkan, C. *et al.* Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat. Genet.* **41**, 1061–1067 (2009).
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. Larkin, M.A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947–2948 (2007).
60. Tamura, K., Dudley, J., Nei, M. & Kumar, S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol. Biol. Evol.* **24**, 1596–1599 (2007).
61. Igartua, C. *et al.* Targeted enrichment of specific regions in the human genome by array hybridization. *Curr. Protoc. Hum. Genet.* Chapter 18 Unit 18.3 (2010).
62. 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).

doi:10.1038/ng.3120