nature genetics

A large and complex structural polymorphism at 16p12.1 underlies microdeletion disease risk

Francesca Antonacci[1], Jeffrey M Kidd[1], Tomas Marques-Bonet[1,2], Brian Teague[3], Mario Ventura[4], Santhosh Girirajan[1], Can Alkan[1,5], Catarina D Campbell[1], Laura Vives[1], Maika Malig[1], Jill A Rosenfeld[6], Blake C Ballif[6], Lisa G Shaffer[6], Tina A Graves[7], Richard K Wilson[7], David C Schwartz[4] & Evan E Eichler[1,5]

There is a complex relationship between the evolution of segmental duplications and rearrangements associated with human disease. We performed a detailed analysis of one region on chromosome 16p12.1 associated with neurocognitive disease and identified one of the largest structural inconsistencies in the human reference assembly. Various genomic analyses show that all examined humans are homozygously inverted relative to the reference genome for a 1.1-Mb region on 16p12.1. We determined that this assembly discrepancy stems from two common structural configurations with worldwide frequencies of 17.6% (S1) and 82.4% (S2). This polymorphism arose from the rapid integration of segmental duplications, precipitating two local inversions within the human lineage over the last 10 million years. The two human haplotypes differ by 333 kb of additional duplicated sequence present in S2 but not in S1. Notably, we show that the S2 configuration harbors directly oriented duplications, specifically predisposing this chromosome to disease-associated rearrangement.

Numerous studies have shown that segmental duplications and the flanking unique regions are sites of both rare and common copy-number polymorphism (CNP)[1–3]. Segmental duplications are blocks of DNA >1 kb in size that occur at more than one site within the genome and typically share a high level (>90%) of sequence identity[4–6]. Duplicated blocks may be substrates for nonallelic homologous recombination (NAHR), resulting in large structural polymorphisms and chromosomal rearrangements that directly lead to genomic disorders[5,7–13]. NAHR between directly oriented segmental duplications results in deletions or reciprocal duplications of the genomic segment between them, whereas NAHR between inverted segmental duplications leads to an inversion of the intervening sequence.

Recently, we reported a recurrent microdeletion on chromosome 16p12.1 that acts as a risk factor for childhood intellectual disability and developmental delay[14]. The microdeletion was found to be inherited in 95.6% of the cases, and 24% of the probands carried an additional large duplication or deletion elsewhere in the genome. The data suggested a two-hit copy-number variation (CNV) model in which the 16p12.1 microdeletion results in severe neurodevelopmental phenotypes when coupled to an additional genetic, epigenetic or environmental abnormality.

Using high-density and targeted array-based comparative genomic hybridization (CGH) experiments, we mapped the 16p12.1 microdeletion breakpoints to large blocks of segmental duplications, which we posited might mediate the recurrent rearrangement associated with disease[15]. The extensive CNV and inconsistencies between the reference genome and various genomic analyses, however, complicated breakpoint assessment, suggesting that large alternative structural configurations might exist within the human population[16,17]. We therefore investigated this region by conducting a detailed analysis using fluorescence *in situ* hybridization (FISH), array CGH, optical mapping and sequencing of large-insert bacterial artificial chromosome (BAC) clones to understand the extent of human genetic variation in this region, its origin and its impact on disease.

RESULTS

Resolution of a reference genome assembly error

We initially began our investigation of the region by testing whether the gene order within this ~1-Mb region was consistent with published reference genome assemblies (GRCb37 and build 36). We performed a series of cohybridization FISH experiments on ten HapMap cell lines using probes corresponding to unique sequences flanking the duplication blocks (**Supplementary Note**). FISH results showed that 20 of 20 chromosomes tested were inverted relative to build 36 and GRCb37, suggesting a potential error in the orientation of the reference genome assembly involving 18 genes (**Supplementary Note**). To confirm this notably large-scale difference, we used optical mapping[18,19] to generate single-molecule restriction maps from the genomes of GM18994 and GM10860 cell lines. We compared the consensus maps to a restriction map generated *in silico* from the build 36 human genome

[1]Department of Genome Sciences, University of Washington, Seattle, Washington, USA. [2]Institut de Biologia Evolutiva (UPF-CSIC), Barcelona, Catalonia, Spain. [3]The Laboratory for Molecular and Computational Genomics, Department of Chemistry, Laboratory of Genetics and Biotechnology Center, University of Wisconsin, Madison, Wisconsin, USA. [4]Department of Genetics and Microbiology, University of Bari, Bari, Italy. [5]Howard Hughes Medical Institute, University of Washington, Seattle, Washington, USA. [6]Signature Genomic Laboratories, Spokane, Washington, USA. [7]Genome Sequencing Center, Washington University School of Medicine, St. Louis, Missouri, USA. Correspondence should be addressed to E.E.E. (eee@gs.washington.edu).

Received 8 April; accepted 15 July; published online 22 August 2010; doi:10.1038/ng.643

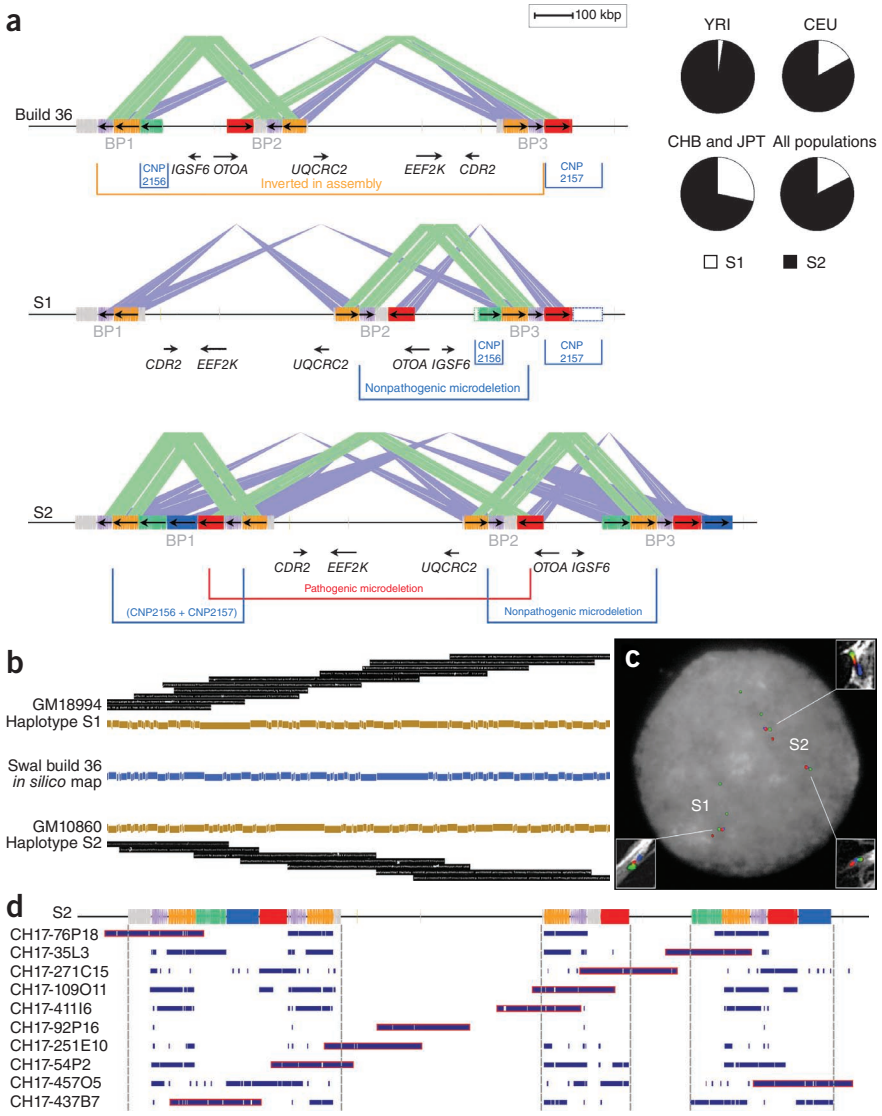

Figure 1 Alternative structural configurations of the 16p12.1 region. (**a**) The organization in the reference genome (build 36, top schematic) is compared against two experimentally validated structural configurations (S1 and S2). Locations are indicated for the inversion, CNPs[2] (CNP2156 and CNP2157), a rare (20 of 6,712) nonpathogenic deletion variant[1] and segmental duplications (colored rectangles). Dashed empty boxes at the S1 structure correspond to regions duplicated in S2 but present in single copy in the S1 haplotype. The S1 and S2 structures differ because of the presence of the distal duplication segment (CNP2156 and CNP2157 at BP1) on the S2 haplotype. On the basis of this structure, the S1 configuration is predicted to be protective against occurrence of the pathogenic 16p12.1 microdeletion. The red block corresponds to the 68-kb segmental duplication that probably mediates, through NAHR, the recurrent 16p12.1 microdeletion[14]. Segments duplicated in a direct orientation are connected by green lines; sequences duplicated in an inverted orientation are connected by blue lines. (**b**) The organization of the region was experimentally validated by optical mapping. SwaI single-molecule restriction maps are depicted and summarized for both configurations (**Supplementary Note**). (**c**) The large-scale orientation of each block was confirmed by FISH experiments on interphase nuclei and stretched chromosomes (white rectangles) using probes mapping at segmental duplications shown in red, blue and green in **a**. (**d**) A contig of ten BAC clones along the 16p12.1 region from the genome of the complete hydatidiform mole (CHM1hTERT) was sequenced. All clones mapped against the S2 structure were concordant.

reference sequence. Maps from both genomes confirmed a large inversion spanning from the duplication blocks defined as breakpoint (BP) regions BP1 and BP3 (build 36, chr16:21421324-22464053; **Supplementary Note** and **Supplementary Fig. 1**).

As a final test, we generated a map of contiguous clones of the region from the CHORI-17 BAC library from a hydatidiform mole-derived (haploid) human cell line (CHM1hTERT)[20]. Complete hydatidiform moles arise from the fertilization of an enucleated egg from a single sperm and therefore carry a haploid complement of the human genome, eliminating allelic variation that may confound mapping and assembly. We constructed a contiguous set of ten BAC clones corresponding to this 1.6-Mb region on 16p12.1 and then sequenced the inserts using Illumina technology. We generated 406 Mb of sequence (270-fold coverage) from these clones and aligned it to both the human reference genome assembly and our reconstructed inverted version of the region (see below). The mapped sequence data from these clones were consistent with the entire region being inverted within the hydatidiform mole (**Supplementary Note**). Thus, all three analyses indicate that orientation of the sequence between BP1 and BP3 should be flipped with respect to published versions of the human genome (**Fig. 1**).

Copy number and structural polymorphism

One of the predicted consequences of this inverted orientation of the human genome is that the locations of previously described segmental duplications and CNPs change with respect to disease-associated breakpoints. The deletion breakpoints associated with intellectual disability now map to BP1 and BP2 using the correct orientation (build 36, chr16:21716331-22464053; **Fig. 1a**). These variable regions correspond, in part, to two sites of common CNP (CNP2156 and CNP2157) identified in the HapMap sample collection[2]. Both loci have three reported copy-number states (diploid copy numbers of 2, 3 and 4), with the highest-copy-number state (copy number of 4) having a frequency of 73% in Europeans (CEU), 95% in Yorubans (YRI) and 52% in Asians (CHB and JPT) (**Supplementary Note**). We performed a series of FISH and array CGH experiments to determine the absolute copy number and the location and extent of CNP within this region (**Supplementary Note**).

We analyzed 11 DNA control samples (**Supplementary Note**) using a customized oligonucleotide microarray and found good correspondence between predicted CNP2157 genotypes and expected signal-intensity differences among samples (**Fig. 2**). Array CGH data for CNP2156 was less clear, and the data suggested more extensive CNV than was originally defined, although the location of this variation could not be determined solely on the basis of hybridization data. We therefore designed a series of three-color FISH experiments to investigate copy number and location. FISH analysis showed that

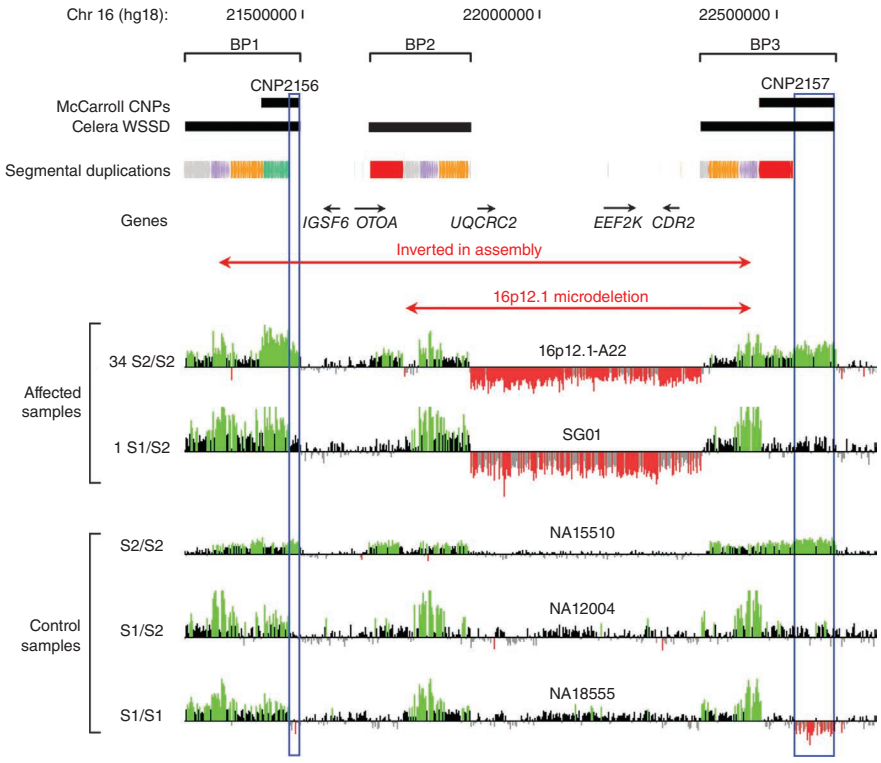

Figure 2 Array CGH data for 16p12.1 microdeletion samples and control HapMap samples (NA15510, NA12004 and NA18555). Probes with $\log_2$ ratios above or below a threshold of 1.5 s.d. from the normalized mean $\log_2$ ratio are colored green (duplication) or red (deletion), respectively. The positions of CNPs (CNP2156 and CNP2157) and segmental duplications are indicated. Blue empty boxes highlight the S2-specific duplications that have a diploid copy number of 2 in S1/S1 individuals, 3 in S1/S2 heterozygotes and 4 in S2/S2 homozygotes. HapMap sample NA18956 with S1/S2 genotype was used as reference.

the absolute copy number of the 68-kb segment corresponding to the distal region of CNP2157 differed by two copies with respect to previous reports (copy numbers of 4, 5 and 6). Similarly, FISH analysis for the CNP2156 region showed an absolute count that is four copies greater than previously reported genotype estimates (**Supplementary Note**)[2]. FISH mapping showed that the variable sequences corresponding to CNP2156 and CNP2157 map adjacent to one another within the BP1 region (**Fig. 1** and **Supplementary Note**). Thus, the two reported CNP regions actually correspond to a single segment of variable sequence that has been duplicatively transposed from BP3 to BP1. Together, these experiments revealed the presence of two distinct structural configurations for the 16p12.1 region, which we refer to as S1 and S2, with the S2 haplotype showing the greater duplication complexity (**Fig. 1**).

As our analyses predicted a large alternative structural polymorphism, we searched GenBank for additional sequenced BACs from this region. We identified clones anchored within the unique region distal to BP1 and constructed an alternative assembly from four BAC clones not included in the human reference genome assembly (**Supplementary Note**). We assembled a 433-kb alternative sequence haplotype corresponding to most of the additional duplicated sequence in BP1. Detailed comparisons with FISH, optical mapping and fosmid end-sequence pair data all provide strong support for the orientation and location of the additional duplicated copies on the S2 chromosomal configuration (**Supplementary Note**).

The combined analysis identifies one of the largest common CNPs in human euchromatin. We identify a total of 333 kb of duplicated sequence that is specific to S2 when compared to the BP1 region of S1. As this additional sequence is homologous to BP1 and BP2, this polymorphism creates additional direct and inverted blocks of high sequence identity, making S2 prone to rearrangement events mediated by NAHR[15]. Only the S2 configuration has segmental duplications in the direct orientation necessary to drive the formation of microdeletions associated with disease. We note that the S2-specific segmental duplications at BP1 show the highest sequence identity (99.85%) with BP3 when compared to BP2 (99.47%), consistent with a recent duplicative transposition event from BP3 placing a large inverted duplication within BP1.

Disease risk

The large-scale structural polymorphism between S1 and S2 allowed us to make some testable predictions regarding differences in susceptibility to microdeletion and disease. As only the S2 configuration has directly oriented duplications, we hypothesized that the breakpoints would map to this 68-kb segment and that only carriers of the S2 configuration would be predisposed to the 16p12.1 microdeletion. Notably, we found that the S2 structure is the most common haplotype worldwide, with frequencies of 97.5% in Africans (YRI), 83.1% in Europeans (CEU) and 71.6% in Asian populations (CHB and JPT)[2] (**Table 1**). This general observation is confirmed by an examination of a larger group of African samples, which show an almost complete absence of the protective S1 haplotype (**Supplementary Note**). Thus, we hypothesize that African and European populations should be more at risk for the 16p12.1 microdeletion event than Asians.

One way to test whether the S2 haplotype predisposes to microdeletion is to determine on which structure the microdeletion occurs. However, most of the identified cases are inherited, and parental DNA for additional genotyping is not available[14]. We therefore determined the structural genotype present in each of the cases using array CGH. The presence of any S1/S1 homozygotes who also have the 16p12.1 microdeletion would be inconsistent with the proposed rearrangement structures and mechanism. As the S2 haplotype has a more extended segmental duplication architecture than S1, differences in the chromosomal configuration can easily be deduced (**Fig. 2**). In particular, the S2-specific duplication block corresponding to the distal segment of CNP2157 (blue empty box

Table 1 S1 and S2 haplotype frequencies

Population	S1 frequency	S2 frequency
Asians (CHB and JPT)	0.28	0.72
Yorubans (YRI)	0.03	0.98
Europeans (CEU)	0.17	0.83
Microdeletion samples	0.01	0.99

Shown are the frequencies of S1 and S2 haplotypes in three HapMap populations. Analysis of 35 individuals with the 16p12.1 microdeletion confirmed a non–Hardy-Weinberg equilibrium enrichment of the S2 haplotype ($P = 0.0088$), suggesting that this structural polymorphism predisposes to deletion and disease.

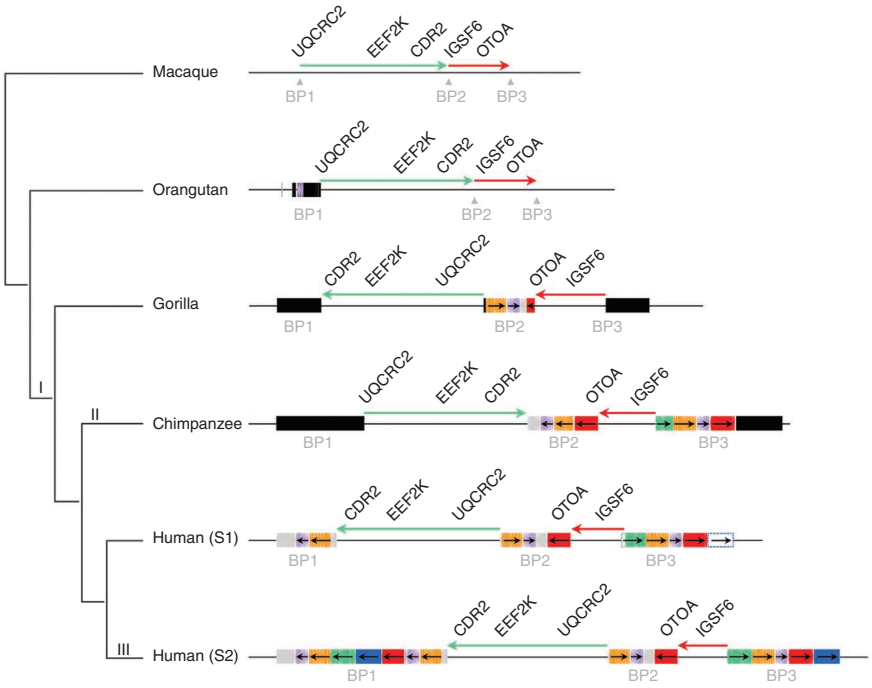

Figure 3 Expansion and multiple inversions of the 16p12.1 region in humans and the syntenic regions in nonhuman primates during primate evolution. The genomic organization is compared within a generally accepted phylogeny of macaque, orangutan, gorilla, chimpanzee and human. The region has expanded from 726 kb (macaque) to 1.6 Mb (human S2) as a result of segmental duplication accumulation (black and colored rectangles). Sequence and FISH data indicate that the inverted configuration as found in orangutan and macaque is probably the ancestral state in all mammals (I). The expansion of segmental duplications in the African great ape ancestor occurred in conjunction with two inversions, between BP1 and BP2 (green arrow) and between BP2 and BP3 (red arrow), which may have reverted back to the direct orientation in the chimpanzee lineage (II). The region has become increasingly complex in humans, leading to the addition of another polymorphic 333 kb at BP1 specifically in the human lineage (III). Colored boxes indicate segmental duplications as determined by complete sequencing of large-insert BAC clones from primate genomic libraries (**Supplementary Note**).

in **Fig. 2**) has a diploid copy number of 2 in S1/S1 individuals, 3 in S1/S2 heterozygotes and 4 in S2/S2 homozygotes.

We examined 35 microdeletion samples by array CGH using two reference samples with known genotypes (NA15724, S2/S2; and NA18956, S1/S2). Self-identified ethnicity was provided for 27 of these individuals (21 of European and 6 of African descent). On the basis of the observed mean $\log_2$ values for the S2-specific duplication block, the genotype of each sample was determined (**Fig. 2**, **Supplementary Figs. 2** and **3**, and **Supplementary Note**). We found that 97% (34 of 35) of the affected individuals were homozygous for the S2/S2 haplotype, with only a single heterozygous carrier (S1/S2) being identified in the affected population (**Table 1**). This represents a significant enrichment of the S2 haplotype when matching for ethnicity of the sample collection ($P = 0.0088$, Hardy-Weinberg equilibrium test). Furthermore, array CGH data from 15 of 16 cases were consistent with breakpoints mapping within the 68-kb S2-specific duplication (**Supplementary Note**). These combined data strongly suggest that the S2, and not S1, haplotype predisposes to the 16p12.1 microdeletion associated with intellectual disability and neurocognitive disease (**Table 1**).

Evolutionary origin

To investigate the ancestral configuration of the 16p12.1 region, we compared the orientation of the region in humans with that in several nonhuman primate species. Notably, sequence comparison of the orangutan (WUGSC 2.0.2/ponAbe2) and human sequence at 16p12.1 revealed an expansion of the region in humans owing to the integration of segmental duplications accompanied by two local inversions of 481 kb and 142 kb (**Supplementary Note**). We tested for the presence of the larger inversion between BP1 and BP2 (481 kb) by FISH analysis of cell lines from three chimpanzees (*Pan troglodytes*), three orangutans (*Pongo pygmaeus*), two gorillas (*Gorilla gorilla*) and one macaque (*Macaca mulatta*) (**Supplementary Note**). Macaque, orangutan and chimpanzee were found to be inverted when compared to the true human genome orientation, suggesting that this represents the likely ancestral state. To resolve the status of the smaller inversion (BP2-BP3) as well as duplications at the boundaries, we identified and sequenced nine large-insert chimpanzee, orangutan and gorilla BAC clones, generating 1.8 Mb of high-quality ape sequence from the region (**Supplementary Fig. 4**). Our results indicated that all African great apes are inverted for the smaller BP2-BP3 interval (142 kb) when compared to orangutan (ponAbe2) and macaque (rheMac2) genome assemblies. We conclude that the two inversions occurred in the human–African great ape ancestor and that the region spanning BP1 to BP2 probably flipped back to the ancestral orientation in the chimpanzee lineage (**Fig. 3**). Alternatively, the chimpanzee configuration may represent incomplete lineage sorting of an ancestral state.

Next, we compared the extent of segmental duplications in the 16p12.1 region among human, chimpanzee, gorilla, orangutan, gibbon and macaque using a whole-genome shotgun sequence detection method and interspecies array CGH[5,21]. These analyses showed an expansion of segmental duplications among African great apes (human, chimpanzee and gorilla) with respect to orangutan, gibbon and macaque (**Fig. 4** and **Supplementary Note**). Sequencing of orangutan BAC clones suggests that this region is largely devoid of segmental duplications in orangutan, with the exception of BP1, where the composition of the duplication block differs radically from that of human (**Fig. 3**). Sequence analysis of the BAC clones reveals the presence of duplicated sequences that are not present at this location in human or chimpanzee, with the exception of a 20-kb segment corresponding to the *NPIP* gene. Overall, we determined that this particular region of 16p12.1 has increased in size from 726 kb to 1,259 kb (S1) or 1,671 kb (S2) during the last 10 million years, primarily as a result of a duplicative transposition of segmental duplications in the region. Our primate analysis suggests that the region has become increasingly complex in the human–African great ape lineage. The euchromatin has expanded 2.3-fold in size. These changes were accompanied by two local inversions, 481 kb and 142 kb in length, creating the genomic architecture that now predisposes this region to microdeletion associated with neuropsychiatric disease.

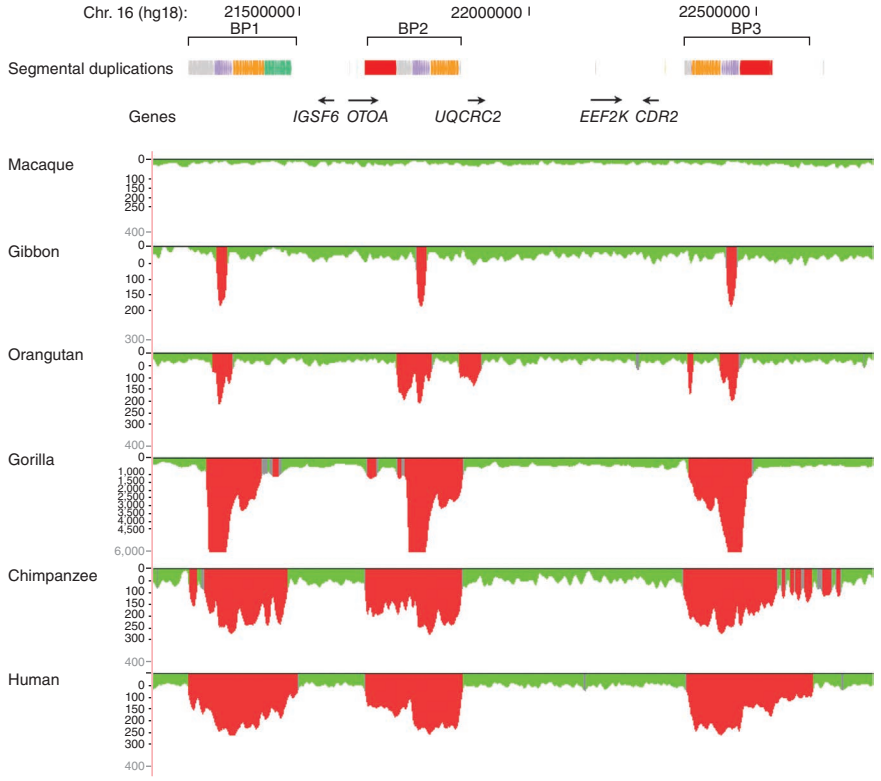

Figure 4 Regions of segmental duplication based on read-depth mapping of whole-genome shotgun sequences against the human genome. The figure shows an expansion of segmental duplications in the African great apes (human, chimpanzee and gorilla) with respect to orangutan, gibbon and macaque. Also shown are the segmental duplications in human annotated using SegDupMasker[43].

DISCUSSION

Our analyses highlight three important findings regarding the organization and evolution of the human genome. First, the data show that the structure and copy number of even very large-scale euchromatic regions may yet be unresolved in the human reference assembly. We describe a 333-kb polymorphism that has changed in copy, orientation and location over a 1-Mb portion of chromosome 16p12.1. With estimated frequencies of 17.6% and 82.4% for the S1 and S2 configurations, respectively, this represents one of the largest CNPs mapping within human euchromatin.

We show that previous analyses of genome structural variation[2,3,16] have not adequately deciphered the true structure and copy number of this polymorphism. In particular, CNP analysis using Affymetrix 6.0 microarrays[2] did not accurately determine the extent of the CNP (76 kb at CNP2156 and 146 kb at CNP2157) owing to the insensitivity of probes mapping within the duplicated regions. Moreover, FISH analyses revealed that the absolute copy number was incorrect, as a baseline copy number of 2 (diploid) was assumed to represent the population average in previous analyses. This was compounded by the fact that the reference genome assemblies (GRC37 and build 36) are missing duplicated copies and present an organization that can not be validated over 1.1 Mb. We postulate that the presence of the inverted 333-kb duplication polymorphism led to large-scale misassembly and misorientation of sequence involving 18 genes (**Fig. 1**). It may be somewhat surprising that such a large 'error' has been uncovered nearly 10 years after the sequence and assembly of the human genome[22,23]; however, it should be pointed out that at least five different types of molecular, optical mapping and cytogenetic analyses were required to resolve the architecture of this region. We anticipate that other regions of comparable complexity and variation will be uncovered and that similarly detailed analyses of large-insert clones will be required to resolve the true architecture of these regions.

Second, our comparative analyses of human and African great ape genomes reveal the evolutionary rapidity of these complex changes and their intimate association with larger chromosomal rearrangements. The 16p12.1 region has experienced a remarkable 'bloating' of euchromatin, doubling the size of this region from 726 kb to 1.6 Mb as a result of duplicative transposition of sequences from other portions of chromosome 16. Most of these changes occurred in a ~6-million-year window of evolution before the emergence of humans and great apes as distinct lineages (**Fig. 3**), consistent with the burst of duplications in their common ancestor[21]. In concert with these changes, there have been multiple local inversions specific to humans and African great apes. These findings reinforce the strong association between evolutionary inversions and segmental duplications[24–27]. It is interesting that all of the 16p12.1 changes are associated with the spread of the human–great ape gene family morpheus (*NPIP*)[28]. The core duplicon carrying this gene (LCR16a)[29] maps to each of the breakpoint regions, including the boundaries of the complex CNP. Sequencing of large-insert ape clones suggests that these sequences also demarcate the breakpoints of the evolutionary inversions. Notably, the segmental duplication associated with the *NPIP* gene family appears to be at the breakpoints of other recurrent microdeletions[30–35] on chromosome 16.

Third, our findings emphasize the impact of this genetic variation with respect to human health and genomic susceptibility to neurocognitive disease. The marked changes in the S2 chromosome architecture mean that it is the only configuration with homologous segmental duplications in direct orientation flanking the disease-critical region. Accordingly, we find that S1 chromosomes are depleted from individuals with microdeletions ($P = 0.0088$ rejecting Hardy-Weinberg equilibrium) and that the breakpoints map specifically to the directly oriented duplication on S2. Combined, these results suggest that S2 chromosomes are likely to predispose to 16p12.1 microdeletion, whereas the S1 chromosomes are immune to such rearrangement. Notably, Asian HapMap samples are enriched for S1 chromosomes, predicting that this particular cause of intellectual disability may be less common among these populations. These results bear striking similarity to another region of the human genome, on 17q21.31, where a largely Mediterranean European–specific duplication arose in direct orientation, predisposing H2 chromosomes to microdeletion associated with the 17q21.31 syndrome[25,36–39]. In both of these cases, changes in disease-causing architecture are also associated with inversions. We posit that this will be the underlying molecular basis for other associations that have been seen with inverted chromosomal haplotypes[40–42]. These observations emphasize the importance of correctly defining alternative human genomic configurations to assess

variable risk of subsequent pathogenic rearrangements. Molecular cytogenetics, genomic approaches and sequencing of long molecules from single haplotypes remain the only way to correctly resolve these complex architectures of the human genome.

URLs. CHORI-17 BAC library, http://bacpac.chori.org/library.php?id=231; mrsFAST, http://mrsfast.sourceforge.net; Integrated Genomics Viewer, http://www.broadinstitute.org/ig.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

Accession codes. Sequence data are available from GenBank under accession numbers AC196535, AC142201, AC009124, AC142205 and AC142206 (human), AC183674, AC183685, AC120780, AC183619 and AC183100 (chimpanzee), AC145243 (gorilla), and AC206441, AC206011 and AC207090 (orangutan). The sequence obtained from the ten CHORI-17 BAC clones is available in the Sequence Read Archive under accession number SRP002828. The sequence of the reassembled BAC contig is available in the Third Party Annotation section of the DDBJ/EMBL/GenBank databases under accession number TPA: BK007104.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank P. Sudmant for useful discussions, G.M. Cooper and T. Brown for critical review of the manuscript and L. Zhou, Y. Fu, R. Shi, J. Wu, S. Shaull and B.A. Roe for sequencing of clone AC120780. This work was supported by a US National Science Foundation Graduate Research Fellowship (to J.M.K.) and a Marie Curie fellowship (FP7 to T.M.-B.), and by the US National Institutes of Health (grants T32 GM07215 and 5T15 LM007359 to B.T., HG000225 to D.C.S. and HG002385 to E.E.E.). E.E.E. is an investigator of the Howard Hughes Medical Institute.

AUTHOR CONTRIBUTIONS

F.A. and E.E.E. designed the study. F.A. performed FISH experiments and constructed shotgun sequencing libraries. J.M.K. performed sequence analysis and haplotype reconstructions. B.T. and D.C.S. performed optical mapping analysis. T.M.-B., T.A.G. and R.K.W. performed nonhuman primate BAC clone sequencing and analysis. M.V. performed FISH experiments on stretched chromosomes. C.A. performed Illumina sequencing data analysis. S.G., C.D.C. and L.V. performed high-density array CGH experiments. M.M. performed PCR experiments. J.A.R., B.C.B. and L.G.S. contributed to 16p12.1 microdeletion data collection. F.A., J.M.K. and E.E.E. contributed to data interpretation. F.A. and E.E.E. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturegenetics/.

Published online at http://www.nature.com/naturegenetics/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Itsara, A. *et al.* Population analysis of large copy number variants and hotspots of human genetic disease. *Am. J. Hum. Genet.* **84**, 148–161 (2009).
2. McCarroll, S.A. *et al.* Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat. Genet.* **40**, 1166–1174 (2008).
3. Conrad, D.F. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704–712 (2010).
4. Cheung, V.G. *et al.* Integration of cytogenetic landmarks into the draft sequence of the human genome. *Nature* **409**, 953–958 (2001).
5. Bailey, J.A. *et al.* Recent segmental duplications in the human genome. *Science* **297**, 1003–1007 (2002).
6. Cheung, J. *et al.* Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence. *Genome Biol.* **4**, R25 (2003).
7. Ji, Y., Eichler, E.E., Schwartz, S. & Nicholls, R.D. Structure of chromosomal duplicons and their role in mediating human genomic disorders. *Genome Res.* **10**, 597–610 (2000).
8. Inoue, K. & Lupski, J.R. Molecular mechanisms for genomic disorders. *Annu. Rev. Genomics Hum. Genet.* **3**, 199–242 (2002).
9. Stankiewicz, P. & Lupski, J.R. Genome architecture, rearrangements and genomic disorders. *Trends Genet.* **18**, 74–82 (2002).
10. Scherer, S.W. *et al.* Human chromosome 7: DNA sequence and biology. *Science* **300**, 767–772 (2003).
11. Eichler, E.E., Clark, R.A. & She, X. An assessment of the sequence gaps: unfinished business in a finished human genome. *Nat. Rev. Genet.* **5**, 345–354 (2004).
12. Shaw, C.J. & Lupski, J.R. Implications of human genome architecture for rearrangement-based disorders: the genomic basis of disease. *Hum. Mol. Genet.* **13** Spec No 1, R57–R64 (2004).
13. Lupski, J.R. & Stankiewicz, P. Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes. *PLoS Genet.* **1**, e49 (2005).
14. Girirajan, S. *et al.* A recurrent 16p12.1 microdeletion supports a two-hit model for severe developmental delay. *Nat. Genet.* **42**, 203–209 (2010).
15. 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).
16. Kidd, J.M. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56–64 (2008).
17. Tuzun, E. *et al.* Fine-scale structural variation of the human genome. *Nat. Genet.* **37**, 727–732 (2005).
18. Zhou, S. *et al.* A single molecule scaffold for the maize genome. *PLoS Genet.* **5**, e1000711 (2009).
19. Teague, B. *et al.* High-resolution human genome structure by single-molecule analysis. *Proc. Natl. Acad. Sci. USA* **107**, 10848–10853 (2010).
20. Fan, J.B. *et al.* Paternal origins of complete hydatidiform moles proven by whole genome single-nucleotide polymorphism haplotyping. *Genomics* **79**, 58–62 (2002).
21. Marques-Bonet, T. *et al.* A burst of segmental duplications in the genome of the African great ape ancestor. *Nature* **457**, 877–881 (2009).
22. Lander, E.S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
23. Martin, J. *et al.* The sequence and analysis of duplication-rich human chromosome 16. *Nature* **432**, 988–994 (2004).
24. Cáceres, M., Sullivan, R.T. & Thomas, J.W. A recurrent inversion on the eutherian X chromosome. *Proc. Natl. Acad. Sci. USA* **104**, 18571–18576 (2007).
25. Zody, M.C. *et al.* Evolutionary toggling of the *MAPT* 17q21.31 inversion region. *Nat. Genet.* **40**, 1076–1083 (2008).
26. Kehrer-Sawatzki, H. & Cooper, D.N. Molecular mechanisms of chromosomal rearrangement during primate evolution. *Chromosome Res.* **16**, 41–56 (2008).
27. Murphy, W.J. *et al.* A rhesus macaque radiation hybrid map and comparative analysis with the human genome. *Genomics* **86**, 383–395 (2005).
28. Johnson, M.E. *et al.* Positive selection of a gene family during the emergence of humans and African apes. *Nature* **413**, 514–519 (2001).
29. Jiang, Z. *et al.* Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution. *Nat. Genet.* **39**, 1361–1368 (2007).
30. Weiss, L.A. *et al.* Association between microdeletion and microduplication at 16p11.2 and autism. *N. Engl. J. Med.* **358**, 667–675 (2008).
31. Kumar, R.A. *et al.* Recurrent 16p11.2 microdeletions in autism. *Hum. Mol. Genet.* **17**, 628–638 (2008).
32. Ullmann, R. *et al.* Array CGH identifies reciprocal 16p13.1 duplications and deletions that predispose to autism and/or mental retardation. *Hum. Mutat.* **28**, 674–682 (2007).
33. Hannes, F.D. *et al.* Recurrent reciprocal deletions and duplications of 16p13.11: the deletion is a risk factor for MR/MCA while the duplication may be a rare benign variant. *J. Med. Genet.* **46**, 223–232 (2009).
34. Ballif, B.C. *et al.* Discovery of a previously unrecognized microdeletion syndrome of 16p11.2-p12.2. *Nat. Genet.* **39**, 1071–1073 (2007).
35. Bochukova, E.G. *et al.* Large, rare chromosomal deletions associated with severe early-onset obesity. *Nature* **463**, 666–670 (2010).
36. Koolen, D.A. *et al.* A new chromosome 17q21.31 microdeletion syndrome associated with a common inversion polymorphism. *Nat. Genet.* **38**, 999–1001 (2006).
37. 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).
38. Stefansson, H. *et al.* A common inversion under selection in Europeans. *Nat. Genet.* **37**, 129–137 (2005).
39. Shaw-Smith, C. *et al.* Microdeletion encompassing *MAPT* at chromosome 17q21.3 is associated with developmental delay and learning disability. *Nat. Genet.* **38**, 1032–1037 (2006).
40. Osborne, L.R. *et al.* A 1.5 million-base pair inversion polymorphism in families with Williams-Beuren syndrome. *Nat. Genet.* **29**, 321–325 (2001).
41. Giglio, S. *et al.* Olfactory receptor-gene clusters, genomic-inversion polymorphisms, and common chromosome rearrangements. *Am. J. Hum. Genet.* **68**, 874–883 (2001).
42. Antonacci, F. *et al.* Characterization of six human disease-associated inversion polymorphisms. *Hum. Mol. Genet.* **18**, 2555–2566 (2009).
43. Jiang, Z., Hubley, R., Smit, A. & Eichler, E.E. DupMasker: a tool for annotating primate segmental duplications. *Genome Res.* **18**, 1362–1368 (2008).

ONLINE METHODS

Fluorescence *in situ* hybridization analysis. Metaphase spreads were obtained from lymphoblast and fibroblast cell lines from ten human HapMap individuals (Coriell Cell Repository), three chimpanzees (Douglas, Veronica and Cochise), three orangutans (Susie, ISIS no. 71; PPY9; PPY6), two gorillas (AG20600 and AG05251) and one macaque (MMU2). Stretched chromosomes were prepared as described[44]. Briefly, cells were resuspended in hypotonic solution (HCM: 100 mM HEPES, 1 M glycerol, 100 mM $CaCl_2$, 0.5 M $MgCl_2$) for 15 min. The suspension was then centrifuged using a Shandon CytoSpin III Cytocentrifuge (WS-CYTOSPIN3; 800–1,200 r.p.m. for 5–15 min). 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[45] with minor modifications. Briefly, 300 ng of labeled probe was used for the FISH experiments; hybridization was performed at 37 °C in 2× SSC, 50% (v/v) formamide, 10% (w/v) dextran sulfate and 3 μg sonicated salmon sperm DNA, in a volume of 10 μl. Post-hybridization washing was at 60 °C in 0.1× SSC (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 grayscale 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.

Copy-number variation analysis. Microarray-based CGH was performed on 35 individuals who had 16p12.1 microdeletions, intellectual disability or developmental delay, and congenital malformation[14]. Array CGH experiments on 16p12.1 microdeletion samples and HapMap samples were performed with custom, high-density oligonucleotide arrays (12-plex NimbleGen chip with a density of 1 probe per 40 bp within the 16p12.1 region; 4× 180K Agilent chip targeted to copy number–polymorphic regions of the human genome (C.D.C. and E.E.E., unpublished data), containing 50 probes in the CNP2157 at chr16:22533636-22618896).

The duplication content of human, chimpanzee, gorilla, orangutan, gibbon and macaque was determined using the whole-genome shotgun sequence detection method[21,46]. We also assessed copy-number differences in shared duplications by interspecific array CGH as previously reported[21] (GEO GSE13885). We performed cross-species array CGH with human (Coriell GM15510) as a reference (GEO GSE13884) using chimpanzee (Clint, Coriell S006006), gorilla (Bahati), orangutan (Susie, ISIS no. 71) and macaque (ID17573) samples.

Optical mapping. We examined the 16p12.1 locus in optical mapping data sets for two genomes, those of HapMap panel members GM10860 and GM18994. Briefly, optical mapping[18,19,47,48] is a whole-genome, single-molecule system for the discovery and characterization of structural variation. Individual genomic DNA molecules are restriction-mapped using light microscopy, producing large data sets that are assembled into multimegabase map contigs covering up to 98% of the euchromatic genome. These map contigs provide a global, detailed assessment of genome structure. We recovered consensus restriction maps matching the S1 haplotype from the GM18994 assembly and the S2 haplotype from GM10860; the consensus maps, their alignments back to the build 36 reference sequence (build 36) and a montage of representative single-molecule micrographs are depicted in **Supplementary Figure 1**.

Illumina sequencing. DNA was extracted from ten BAC clones (CHORI-17) (**Supplementary Note**) from the genome of a complete hydatidiform mole (CHM1hTERT) using Roche high pure plasmid isolation kit. We used 3 μg of DNA from each BAC for construction of a shotgun sequencing library as described previously[49,50], using adaptors for paired-end sequencing on an Illumina Genome Analyzer IIX (GAIIX). To allow the simultaneous sequencing of multiple BAC clones, we differentially ligated modified adaptors (**Supplementary Note**) to each sample during library preparation, enabling the *in silico* separation of samples post-sequencing[51]. We obtained a total of 34,206,404 76-bp reads (17,103,202 pairs) and separated them into ten pools using 12-bp barcodes, resulting in 20,316,752 reads of 64 bp. To control for contamination, we first aligned the reads to the *Escherichia coli* reference genome (K12 strain) using mrsFAST, allowing at most 4-bp mismatches. This experiment resulted in removal of 2,363,518 reads (1,181,759 pairs) from consideration owing to contamination. The remaining reads (a total of 406 Mb of generated sequence) were then mapped to the 16p12 region in build 36 and the S1 and S2 haplotype sequences that we constructed. We tracked all possible map locations for the concordant pairs and discarded the discordant mappings. This resulted in reliable mapping of 6,345,136 reads (3,172,568 pairs; 406,088,704 bp of sequence) to the 16p12, S1 and S2 reference sequences, corresponding to 270.7-fold coverage per BAC sequence on average (minimum coverage, 132.5×; maximum coverage, 520.8×). Next, we merged the map locations of the overlapping pairs into contiguous segments and removed any segment <2 kb from analysis. We reasoned that the smaller segments are mapping artifacts resulting from short repeats in the sequenced BAC clones and the reference sequences (16p12, S1 and S2). Finally, we visualized the resulting segments using the Integrated Genomics Viewer software.

Nonhuman primate BAC clone sequencing. We selected nine BAC clones from the libraries of chimpanzee (CH251), orangutan (CH276) and gorilla (CH255) genomes mapping to the 16p12.1 segmental duplications in humans (**Supplementary Note**). We generated a clone shotgun sequence library and completely sequenced the insert of each clone. We aligned the sequence to the human genome and to the S1 haplotype that we reconstructed with miropeats[52]. Final annotation with common repeats and DupMasker output[43] describing the composition of segmental duplications was also included with customized Perl scripts.

44. Laan, M. *et al.* Mechanically stretched chromosomes as targets for high-resolution FISH mapping. *Genome Res.* **5**, 13–20 (1995).
45. Lichter, P. *et al.* High-resolution mapping of human chromosome 11 by *in situ* hybridization with cosmid clones. *Science* **247**, 64–69 (1990).
46. Bailey, J.A., Yavor, A.M., Massa, H.F., Trask, B.J. & Eichler, E.E. Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res.* **11**, 1005–1017 (2001).
47. Church, D.M. *et al.* Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLoS Biol.* **7**, e1000112 (2009).
48. Zhou, S. *et al.* Validation of rice genome sequence by optical mapping. *BMC Genomics* **8**, 278 (2007).
49. Ng, S.B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272–276 (2009).
50. Quail, M.A. *et al.* A large genome center's improvements to the Illumina sequencing system. *Nat. Methods* **5**, 1005–1010 (2008).
51. Craig, D.W. *et al.* Identification of genetic variants using bar-coded multiplexed sequencing. *Nat. Methods* **5**, 887–893 (2008).
52. Parsons, J.D. Miropeats: graphical DNA sequence comparisons. *Comput. Appl. Biosci.* **11**, 615–619 (1995).