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

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

SUPPLEMENTARY NOTE

Table of contents

Nature Genetics: doi:10.1038/ng.643

1. Structural variation at 16p12.1

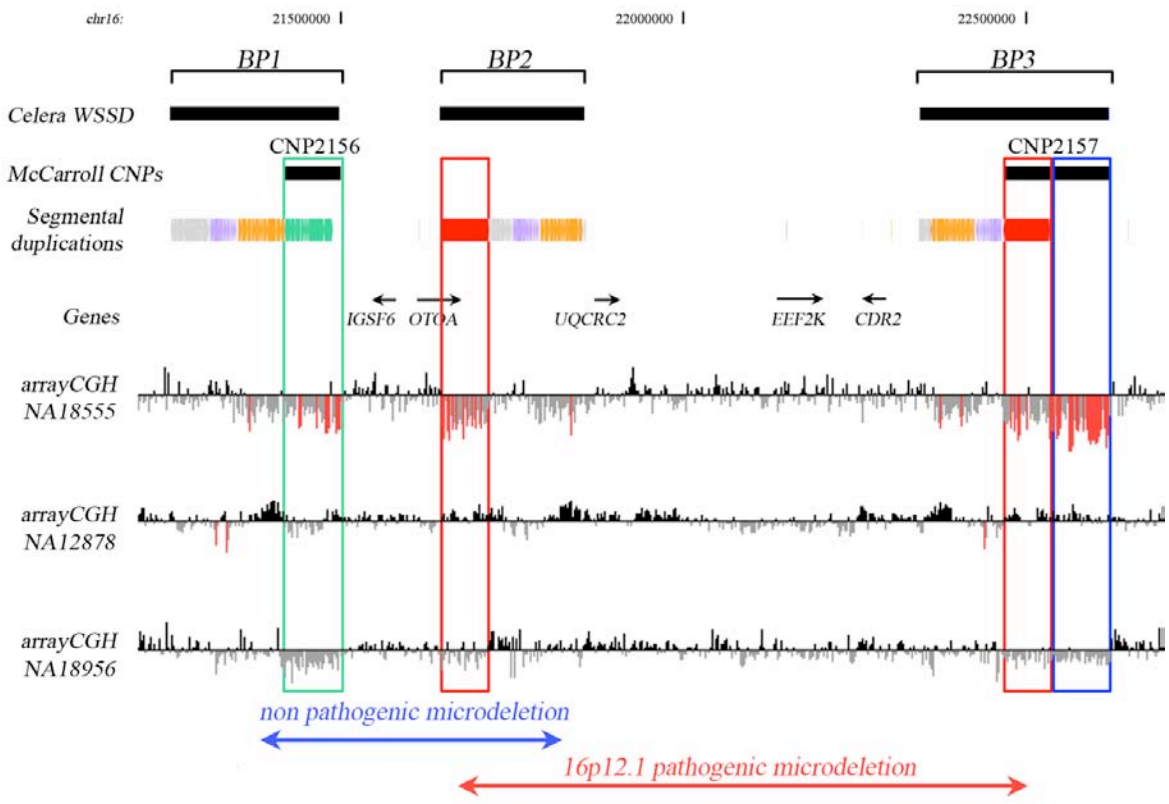
1.1 Copy-number variation analysis by arrayCGH

We initially investigated copy-number variation of the segmental duplications flanking the 520-kbp microdeleted region by array comparative genomic hybridization (arrayCGH). We constructed a high density oligonucleotide microarray (NimbleGen, 50,000 probes with a density of 1 probe per 40 bp along 2 Mb at 16p12.1) and performed arrayCGH on 56 representative HapMap samples (NA07056, NA11831, NA11839, NA11993, NA11994, NA12003, NA12004, NA12043, NA12056, NA12057, NA12707, NA12750, NA12751, NA12753, NA12873, NA12891, NA18524, NA18526, NA18529, NA18547, NA18558, NA18563, NA18576, NA18582, NA18940, NA18942, NA18943, NA18951, NA18956, NA18974, NA18981, NA18501, NA18502, NA18505, NA18507, NA18516, NA19099, NA19102, NA19127, NA19128, NA19138, NA19145, NA19160, NA19171, NA19201, NA19238, NA19239 using reference NA15724, and NA11832, NA12004, NA12878, NA18861, NA18555, NA18947, NA11840, NA18502, NA19240, NA18564, NA15510 using reference NA18956) (Supplementary Note Figure 1). Several regions of copy-number polymorphism were noted based on this survey. First, variation in copy number was observed for an 80-kbp region (green box at BP1) corresponding to a previously described copy-number polymorphism (CNP2156). Note the underlying segmental duplication also delimits the boundary of a non-pathogenic microdeletion, rarely (20/6712) found in controls [1]. Second, we also observed copy-number polymorphism for a 68-kbp segmental duplication mapping to both BP2 and BP3 (red

Nature Genetics: doi:10.1038/ng.643

boxes). This segment was hypothesized by Girirajan *et al.* [2] to promote non-allelic homologous recombination (NAHR) resulting in the 16p12.1 microdeletions associated with intellectual disability and neuropsychiatric disease. Due to the high degree of sequence identity of the underlying segmental duplication, arrayCGH can not be used to unambiguously assign it to one of the two locations. However, we note that copy-number variation extends to the adjacent 85-kbp region at BP3 (blue box) annotated as duplicated based on a previous analysis of the Venter genome [3] but represented as unique sequence in the hg18 (build 36) genome assembly (Supplementary Note Figure 1). In addition to these copy-number polymorphic regions, we observe reproducible $\log_2$ signal intensity differences for other segmental duplications in the region. The higher copy number of the underlying segmental duplications corresponding to these regions, however, reduces the sensitivity and our power to call these regions.

Nature Genetics: doi:10.1038/ng.643



Supplementary Note Figure 1 ArrayCGH data from three representative HapMap samples (NA18555, NA12878 and NA18956) are shown for the 16p12.1 microdeletion region. Also shown are the locations of the 16p12.1 pathogenic microdeletion [2] and a previously described non-pathogenic microdeletion found in 20/6712 controls [1]. The positions of copy-number polymorphisms (CNP2156 and CNP2157) from McCarroll *et al.* [4] are depicted. Segmental duplications were annotated using SegDupMasker [5]. The location of duplications identified by excess whole-genome shotgun sequence read-depth analysis (Celera WSSD) is also shown. In the empty red boxes are shown the 68-kbp segmental duplication paralogous copies at BP2 and BP3. The blue and green boxes indicate respectively the 85-kbp region at BP3 and the CNP2156 region at BP1.

Using SNP microarrays, McCarroll and colleagues previously reported a subset of the underlying copy-number polymorphism. Two sites of common copy-number

polymorphism, CNP2156 at BP1 and CNP2157 at BP3, were identified in the HapMap sample collection [4]. Both loci have three reported copy-number (CN) states (diploid copy numbers of 2, 3, and 4 reported by McCarroll *et al.* [4]), with the highest copy-number state (CN = 4) having a frequency of 73% in Europeans (CEU), 95% in Yorubans (YRI), and 52% in Asians (CHB/JPT) (Supplementary Note Table 1).

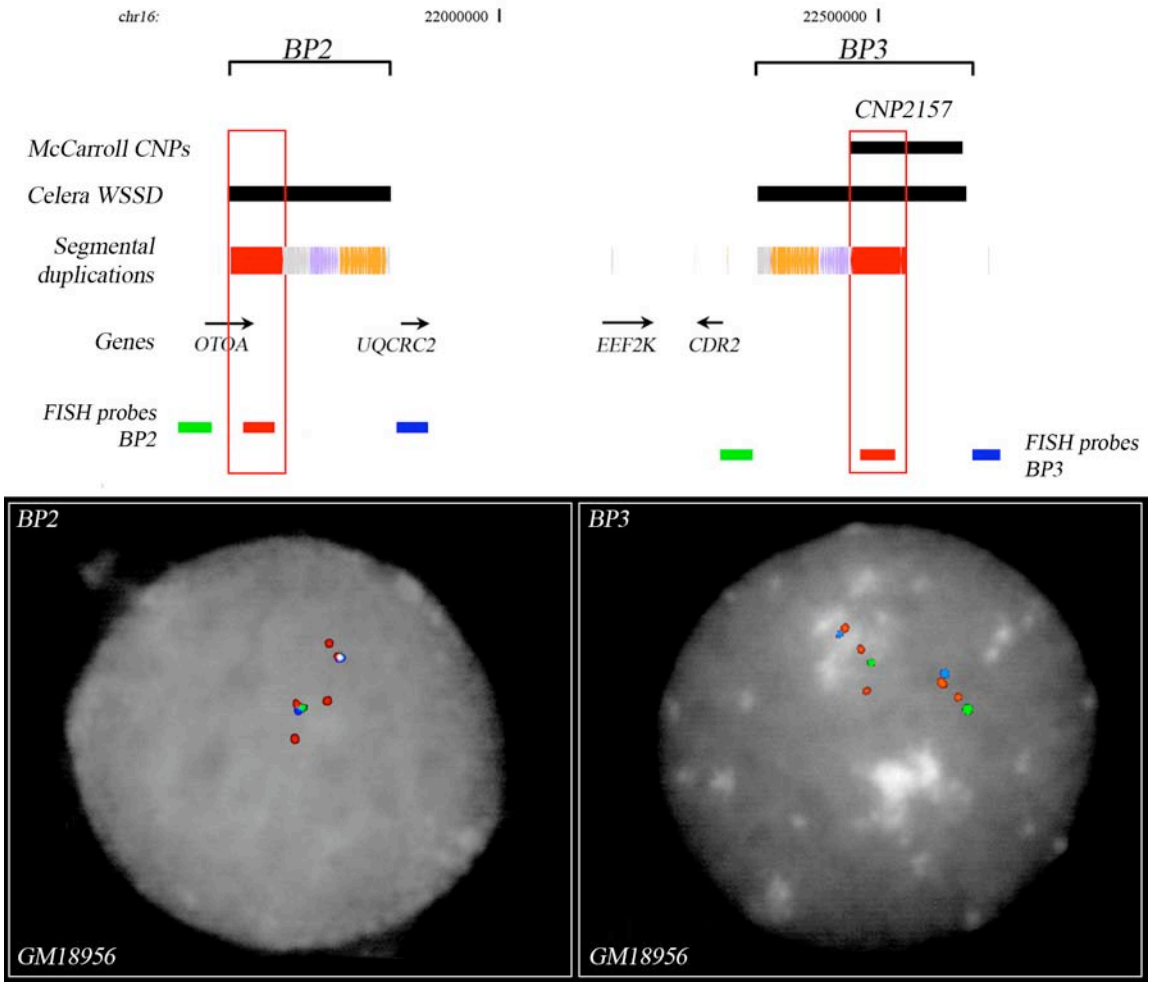
Population	frequency CN=2 (%)	frequency CN=3 (%)	frequency CN=4 (%)
CHB+JPT	9.1	38.6	52.3
YRI	0	5	95
CEU	6.8	20.3	72.9
all populations	5.8	23.7	70.5

Supplementary Note Table 1 Frequencies of CNP2156 and CNP2157 variable regions in 207 HapMap individuals identified by McCarroll and colleagues [4].

1.2 Structural polymorphism analysis and haplotype determination by FISH

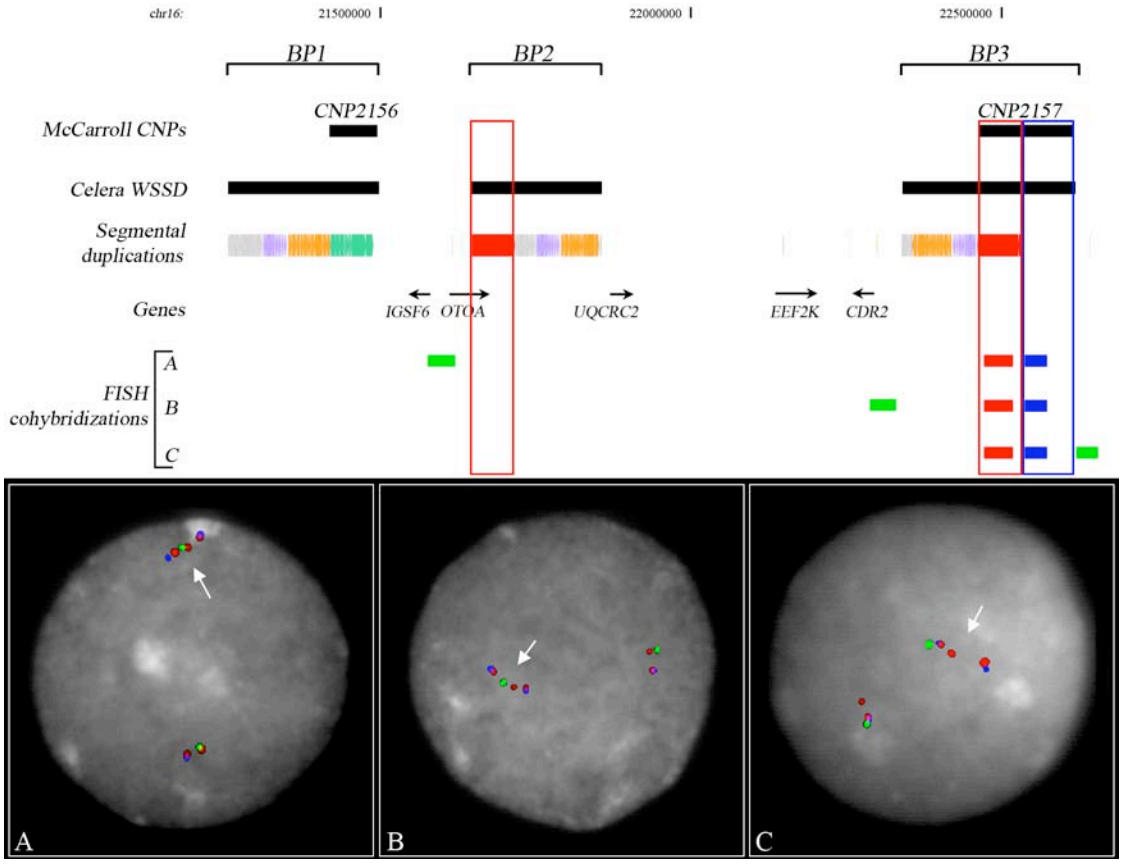
Since copy-number variation of the flanking duplications could potentially be related to increased likelihood for occurrence of the 16p12.1 microdeletion, we performed a series of fluorescence *in situ* hybridization (FISH) experiments to determine the absolute copy number and orientation of these duplications. First, using FISH we analyzed the 68-kbp segmental duplication (red box in Supplementary Note Figure 2) that corresponds to part of CNP2157 in 10 HapMap individuals. The FISH results showed three different copy-number states (diploid CN = 4, 5, and 6). Compared to the absolute copy numbers of McCarroll *et al.* [4], the FISH-determined copy numbers for this segment differed by a count of two in all of the HapMap samples analyzed (Supplementary Note Figure 2; Supplementary Note Table 2).

Nature Genetics: doi:10.1038/ng.643



Supplementary Note Figure 2 The figure shows two cohybridization experiments at BP2 and BP3 using fosmid clones mapping at the 68-kbp directly oriented segmental duplication (WIBR2-0658O17 at BP2 and WIBR2-2031K01 at BP3 (red)) and two flanking single copy probes (WIBR2-2529G18 (green) and WIBR2-1064L24 (blue) at BP2 and WIBR2-3632J22 (green) and WIBR2-1829F15 (blue) at BP3). Five copies of the red probes mapping at the 68-kbp segmental duplication were detected in GM18956. The green and blue probes mapping at the single copy region are not duplicated. Also shown is the position of copy-number polymorphism CNP2157 from McCarroll *et al.* [4]. Segmental duplications were annotated using SegDupMasker [5]. The location of duplications identified by read-depth is also depicted (Celera WSSD).

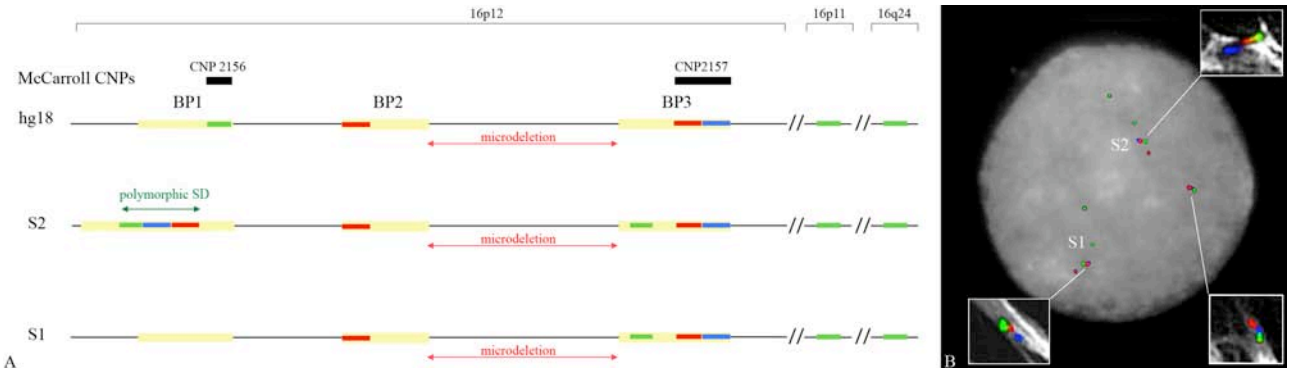
We performed three cohybridization experiments to determine the copy number and position of the 85-kbp WSSD-positive region (blue box) relative to the 68-kbp polymorphic segmental duplication (red box) using three different single-copy probes (Supplementary Note Figure 3). These experiments revealed the presence of two distinct structural configurations for the 16p12.1 region, which we refer to as S1 and S2, with two and three copies, respectively, of the 68-kbp segmental duplication (red box) and one and two copies, respectively, of the 85-kbp WSSD-positive segmental duplication (blue box) (Supplementary Note Figure 3; Supplementary Note Table 2). All three FISH experiments showed that the additional duplication copies found on the S2 structure map to BP1.



Supplementary Note Figure 3 The figure shows three FISH experiments on GM18956 using probe WIBR2-2031K01 (red) mapping at the 68-kbp segmental duplication (highlighted in the empty red box) and probe WIBR2-3608M06 (blue) mapping at the 85-kbp WSSD-positive region (highlighted in the empty blue box) in cohybridization with three single copy probes (green) (WIBR2-2529G18 (A), WIBR2-3632J22 (B), WIBR2-1829F15 (C)). All three FISH experiments confirmed the location of these additional copies in the S2 haplotype at the duplication block in BP1. For example, in panel A the order of the probes as visualized in interphase nuclei is blue, red, green, red, red, blue, indicating that the additional copies of the red and blue probes are located proximally to single copy probe labeled in green. White arrows indicate the S2 haplotype. Also shown are the positions of copy-number polymorphisms (CNP2156 and CNP2157) from McCarroll *et al.* [4]. Segmental duplications were annotated using SegDupMasker [5]. The location of duplications identified by read-depth is also depicted (Celera WSSD).

Additionally, we determined the copy number and relative locations of the BP1 variable regions (green box, corresponding to CNP2156). FISH experiments indicate that the CNP2156 region is present in three copies in the S1 haplotype and in four copies in the S2 (Supplementary Note Figure 4). This corresponds to an absolute count that is four copies greater than the genotypes reported by McCarroll (Supplementary Note Table 2) [4]. The FISH probe for the CNP2156 region co-localizes with the 68-kbp polymorphic segmental duplication and the 85-kbp polymorphic WSSD-positive region at BP3 and BP1. FISH experiments on stretched chromosomes show a different ordering for the green, red and blue probes at BP1 and BP3 (Supplementary Note Figure 4).

Nature Genetics: doi:10.1038/ng.643



Supplementary Note Figure 4 (A) Schematic showing the S1 and S2 structural configurations. Shown are the positions of copy-number polymorphisms (CNP2156 and CNP2157) from McCarroll *et al.* [4]. Segmental duplication blocks are indicated by yellow, green, red and blue boxes. Red, blue and green boxes correspond respectively the 68-kbp polymorphic segmental duplication, the 85-kbp polymorphic WSSD-positive region and the CNP2156. (B) Three-color interphase FISH experiment on GM18956 using probes WIBR2-2031K01 (red) mapping at the 68-kbp segmental duplication, WIBR2-3608M06 (blue) mapping at the 85-kbp WSSD-positive region and WIBR2-0590C03 (green) mapping at the CNP2156 region. The same probes were used in FISH experiments on stretched chromosomes to define the relative order of the red, blue and green duplication blocks at BP1 and BP3. Extracted fibers from FISH on the stretched chromosomes are shown in the white rectangles.

Cell line	Population	CNP 2157 McCarroll calls	CNP2157 FISH calls		CNP 2156 McCarroll calls	CNP2156 FISH calls	Haplotype
			WIBR2-2031K01	WIBR2-3608M06		WIBR2-0590C03	
GM11832	CEU	2	4	2	2	6	S1/S1
GM12004	CEU	3	5	3	3	7	S1/S2
GM12878	CEU	4	6	4	4	8	S2/S2
GM10860	CEU	4	6	4	4	8	S2/S2
GM18861	YRI	4	6	4	4	8	S2/S2
GM18555	CHB	2	4	2	2	6	S1/S1
GM18947	JPT	3	5	3	3	7	S1/S2
GM18956	JPT	3	5	3	3	7	S1/S2
GM18994	JPT	3	5	3	3	7	S1/S2
GM15510	n.a.	n.a.	6	4	n.a.	8	S2/S2

Supplementary Note Table 2 Copy number of CNP2157 and CNP2156 polymorphic regions determined by FISH in 10 HapMap individuals. Probe WIBR2-2031K01 maps to the 68-kbp

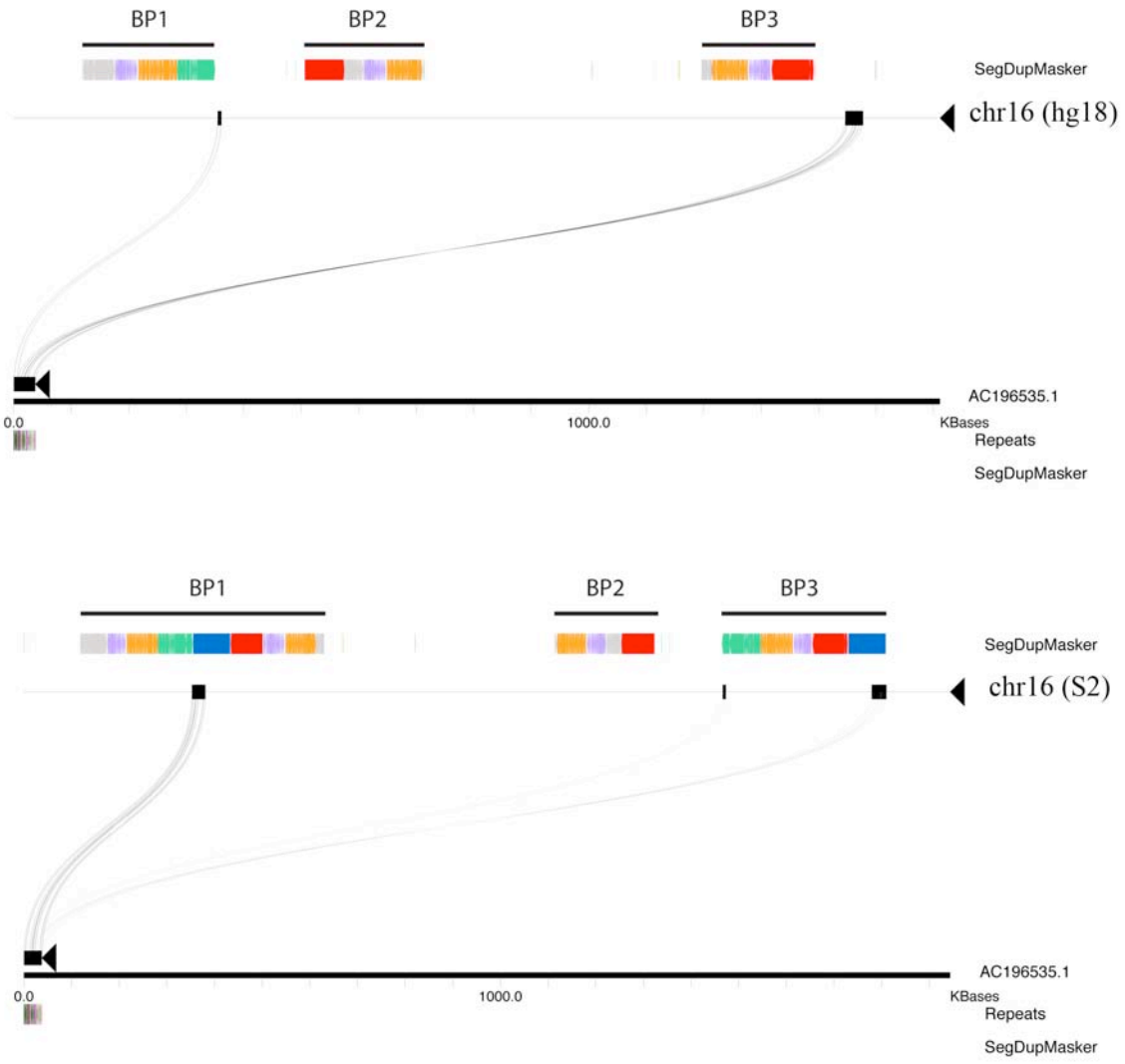
segmental duplication predisposing to 16p12.1 microdeletion [2].

Our FISH analyses confirm the reported polymorphism at the CNP2156 and CNP2157 regions. However, these regions have a mosaic duplication structure that complicates the assignment of a single absolute copy number (Supplementary Note Table 2). Additionally, both FISH on interphase nuclei and stretched chromosomes indicate that the variable sequences corresponding to CNP2156 and CNP2157 are located adjacent to each other at the BP1 region (Supplementary Note Figure 4). Thus, the two reported CNV regions actually correspond to a single segment of variable sequence.

1.3 Sequence analysis and construction of the S2 haplotype

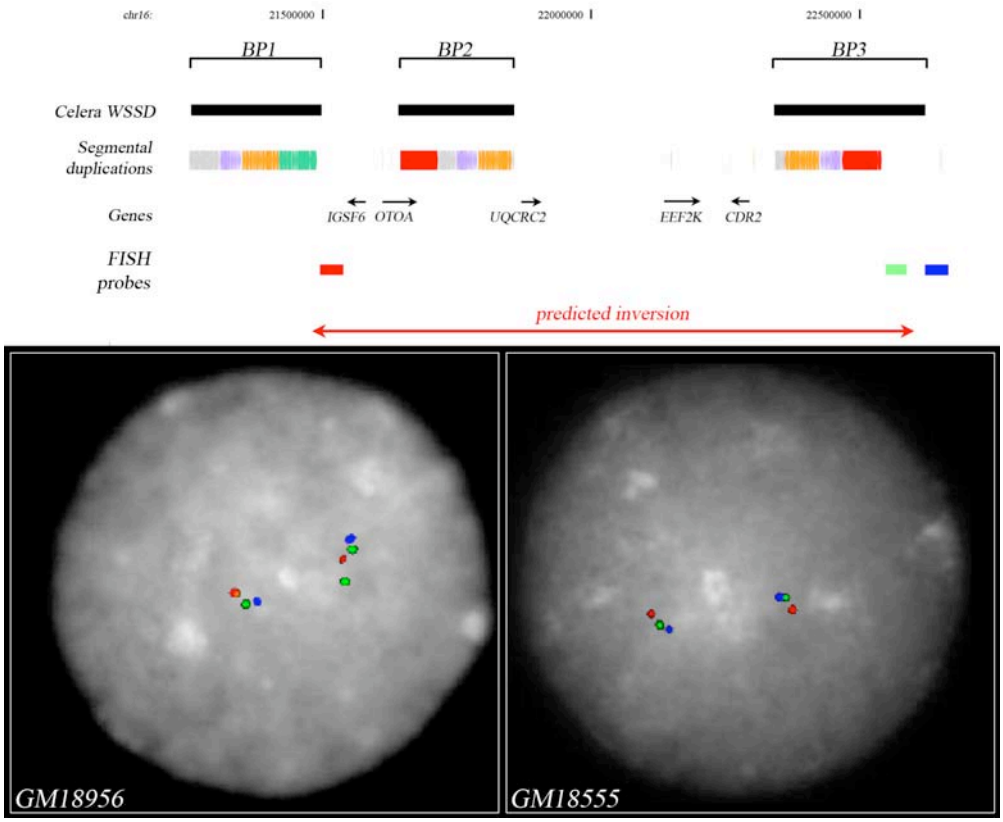
Given the complexity of this locus, we sought to confirm at the sequence level the relative orientation of these polymorphic duplications in the S1 and S2 configurations. Previously, using a fosmid clone-based analysis, Kidd *et al.* [6] identified the presence of a putative inversion spanning the region between BP1 and BP3 in eight out of the nine individuals analyzed [6]. Analysis of the sequenced fosmid clone (clone WI2-2793I8, accession: AC196535) supports the alternate structure of segmental duplications in the S2 haplotype (Supplementary Note Figure 5). This alternate structural polymorphism led to the inference of an inversion in some individuals relative to the genome reference assembly because of the presence of additional segmental duplications, which are located in a different position (BP1) and orientation.

Nature Genetics: doi:10.1038/ng.643



Supplementary Note Figure 5 Sequence alignment of a fosmid clone (AC196535) (black rectangle on the left) with the hg18 (build 36) reference genome and the S2 reconstructed haplotype. Black lines connect matching segments between the clone and the hg18 (top) and S2 (bottom) haplotype sequences [7]. Segmental duplications were annotated using SegDupMasker [5]. The location of duplications identified by read-depth is also depicted (WSSD). Alignment of the clone against hg18 identified the presence of an inversion spanning the region between BP1 and BP3 [6]. Subsequent alignment of the clone against S2 is consistent with the alternate S2 structure.

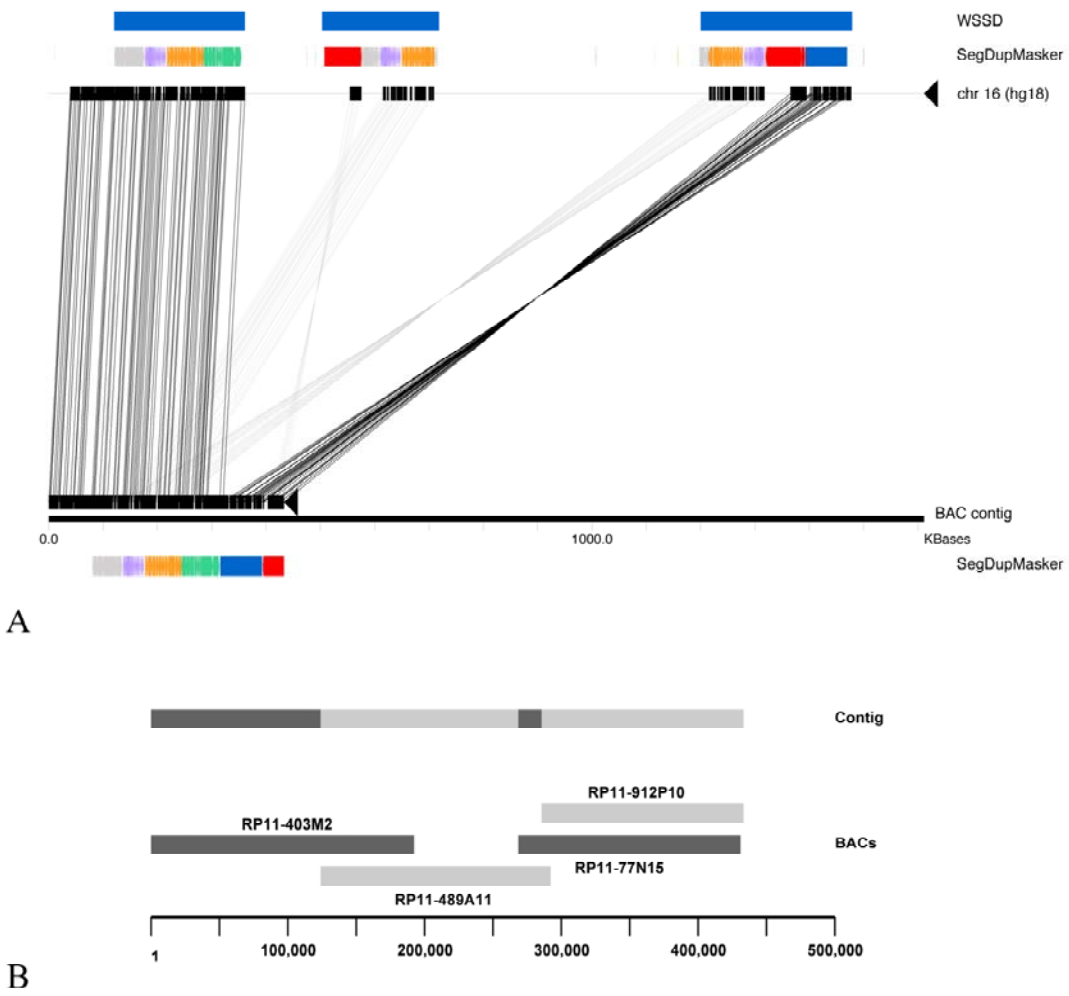
FISH analysis of the proximal breakpoint of the inversion showed evidence of duplication. FISH probe (WIBR2-3608M06) was not duplicated in GM18555 (not predicted to be inverted) and duplicated in GM18956 (predicted to be inverted) due to the polymorphism associated with the S2 haplotype (Supplementary Note Figure 6).



Supplementary Note Figure 6 Interphase triple-color FISH using two probes inside the predicted inversion [6] (WIBR2-0927E20 in red; WIBR2-3608M06 in green) and one outside the inversion (WIBR2-1724F12 in blue) showed duplication of the FISH probe mapping at the inversion proximal breakpoint (WIBR2-3608M06). Note that this probe was duplicated in GM18956 (S1/S2, left) (predicted to be inverted) due to the polymorphism associated with the S2 haplotype and not duplicated in GM18555 (S1/S1, right) (not predicted to be inverted). Segmental duplications were annotated using SegDupMasker [5]. The location of duplications identified by read-depth is also depicted (Celera WSSD).

The comparatively small size of the fosmid clone (~40 kbp) and the large size of the structural polymorphism (>100 kbp) prohibited anchoring of the clone into the adjacent unique sequence at BP1. Construction of a full-tiling path of clones is difficult. Therefore, we searched GenBank for additional sequenced BACs from this region. This was aided by the fact that analysis of the human genome suggested that the S2 configuration would represent the major allele. During the construction of this sequence contig, we required that overlaps among the BACs were consistent with allelic overlap (reported in Supplementary Note Table 3) and a contig of 433,163 bp was constructed from four BACs all derived from the RP11 library (Supplementary Note Figure 7).

Nature Genetics: doi:10.1038/ng.643



Supplementary Note Figure 7 Sequence reconstruction of alternative 16p12.1 structures. (A) The structure represented in the hg18 genome assembly (chr16:21140710-22752037) is compared with a 433,163-bp sequence contig constructed from four RP11 BACs using the program miropeats [7]. Black lines connect matching segments between the RP11 contig and chr16 sequence while light-gray lines indicate lower-identity matches to distinct paralogues. Segmental duplications were annotated using SegDupMasker [5]. The location of duplications identified by read-depth is also depicted (WSSD, blue box). (B) Construction of 433,163-bp contig from four RP11 BACs.

Sequence 1		Sequence 2		Overlapping bp	Identity
Clone	Insert size	Clone	Insert size		
RP11-403M2 (AC142201)	194,943 bp	RP11-489A11 (AC009124)	168,268 bp	69,389	99.99%
RP11-489A11 (AC009124)	168,268 bp	RP11-77N15 (AC142205)	162,377 bp	23,629	99.94%
RP11-77N15 (AC142205)	162,377 bp	RP11-912P10 (AC142206)	147,555 bp	145,396	99.99%

Supplementary Note Table 3 The length and sequence identity of overlapping BACs as determined by BLAST2SEQ is shown.

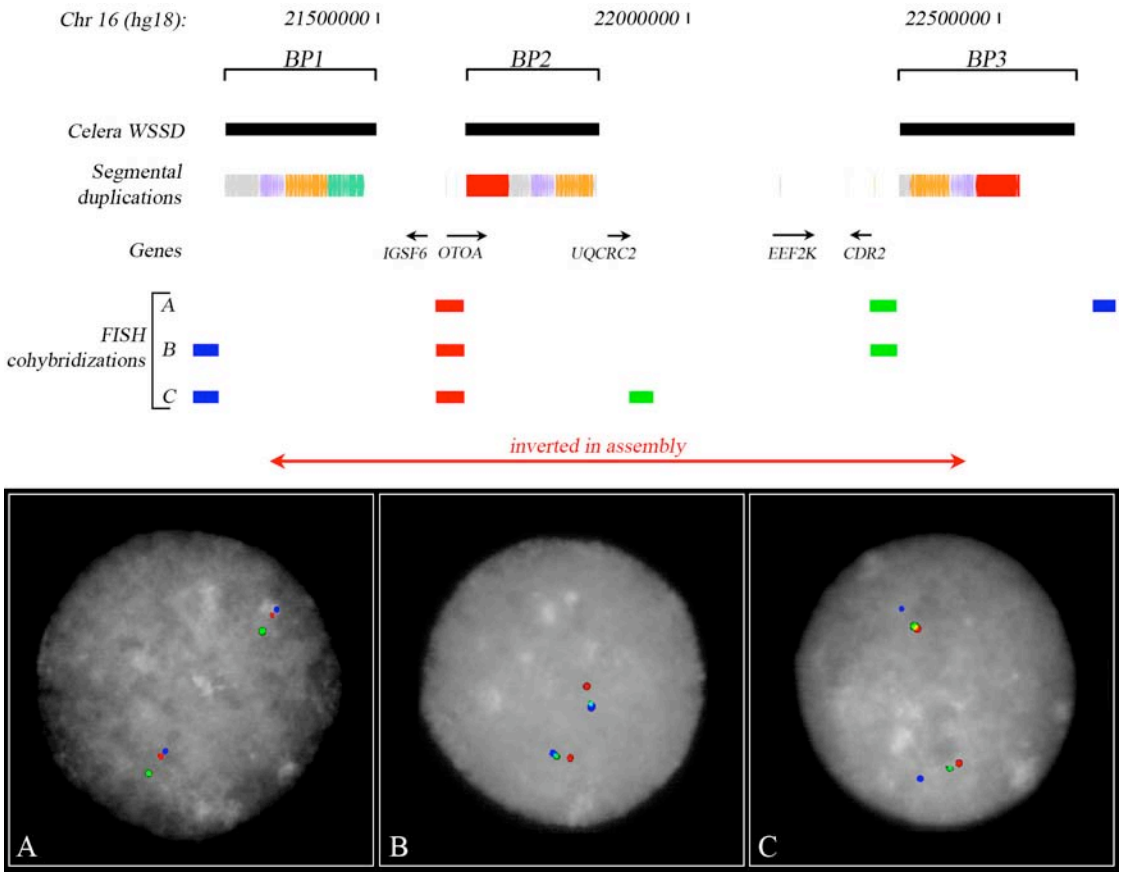
Since clone RP11-403M2 (AC142201.1) is represented as a working draft sequence, we limited the analysis to the largest single contig (192364 bp) assembled for this BAC. This resulting BAC-derived contig was compared with the genome reference sequence (build 36) using the program miropeats [7]. Analysis of these sequences supports the presence of additional inverted duplications including CNP2157 on the S2 haplotype located at BP1. This finding is also consistent with the clone WI2-2793I8 sequence analysis shown in Supplementary Note Figure 5.

2. Orientation error in the reference genome assembly

2.1 FISH analysis

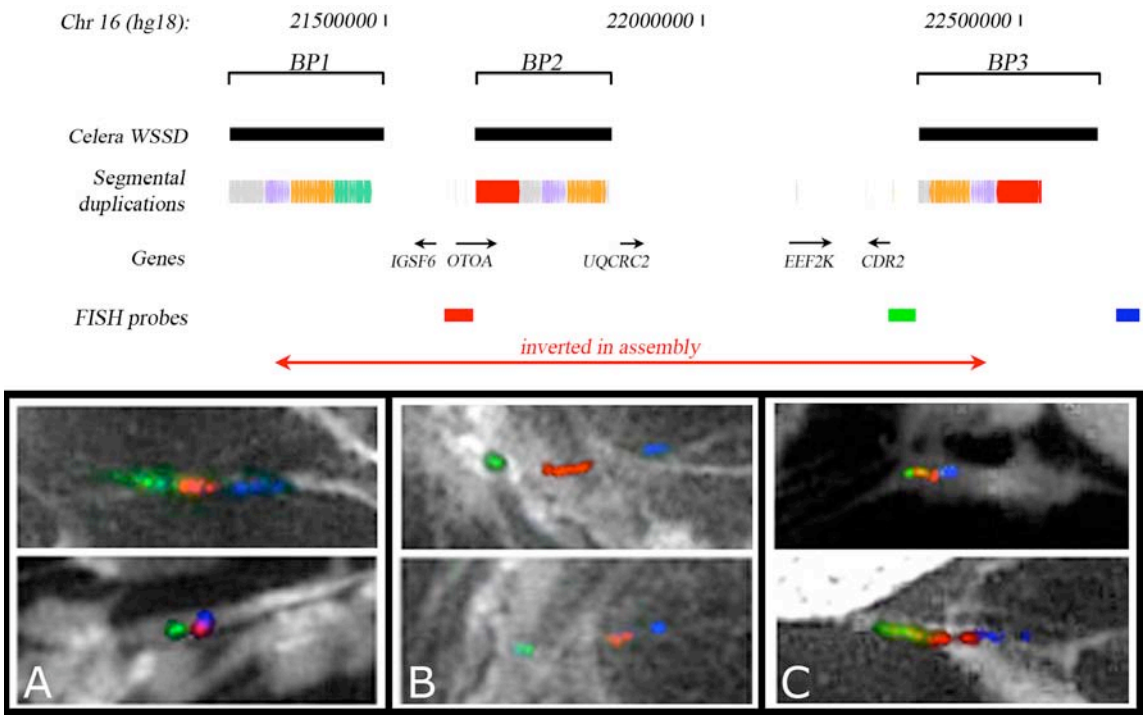
We investigated the orientation of the region by multi-color FISH to determine if the region was inverted. We tested the orientation by performing a series of cohybridization FISH experiments on 10 HapMap cell lines (Supplementary Note Table 4). We used probes anchored within unique regions (no segmental duplication and no copy-number polymorphism) to unambiguously resolve the order of the genes within this 1.1-Mbp region. Surprisingly, FISH results showed that 20/20 chromosomes tested were actually

inverted relative to build 36 (hg18) and GRCb37, suggesting an error in the orientation of the reference genome assembly (Supplementary Note Figure 8).



Supplementary Note Figure 8 The figure shows three cohybridization experiments on GM18956 in order to determine the orientation of the region between BP1 and BP3. (A) Shows a cohybridization of probes WIBR2-2529G18 (red), WIBR2-3632J22 (green) and WIBR2-1829F15 (blue); (B) WIBR2-2019O07 (blue), WIBR2-2529G18 (red) and WIBR2-3632J22 (green); (C) WIBR2-2019O07 (blue), WIBR2-2529G18 (red) and WIBR2-1520O04 (green). All three FISH experiments show the presence of an inversion with respect to the reference genome assembly. The inversion changes the order of the red and green probes and their relative position with respect to the blue probes. The location of duplications identified by read-depth is also depicted (Celera WSSD).

FISH analysis on stretched chromosomes from GM15510 (S2/S2), GM18956 (S1/S2) and GM18555 (S1/S1) cell lines confirmed that all six chromosomes tested were inverted relative to hg18 human reference genome (Supplementary Note Figure 9). This represents one of the largest inconsistencies within the human reference genome where the order of the 18 genes (as defined by RefSeq) should be flipped with respect to hg18/NCBI36. While the fosmid end-sequence pair data had also predicted an inversion between BP1 and BP3, we found that the breakpoints detected by the fosmid corresponded to CNPs within the segmental duplication region (see above and data not shown) and do not correspond to the true breakpoints of the inverted segment.



Supplementary Note Figure 9 The figure shows a cohybridization experiment on stretched chromosomes on GM18555 (A), GM18956 (B) and GM15510 (C) using probes WIBR2-2529G18 (red), WIBR2-3632J22 (green) and WIBR2-1829F15 (blue) (same probes used in

Supplementary Note Figure 8) in order to confirm the inversion of the region between BP1 and BP3 relative to the hg18 human genome reference assembly. The inversion changes the order of the red and green probes and their relative position with respect to the blue probe. Segmental duplications were annotated using SegDupMasker [5]. The location of duplications identified by read-depth is also depicted (Celera WSSD).

Cell line	Population	Inversion status
GM11832	CEU	inv/inv
GM12004	CEU	inv/inv
GM12878	CEU	inv/inv
GM10860	CEU	inv/inv
GM18861	YRI	inv/inv
GM18555	CHB	inv/inv
GM18947	JPT	inv/inv
GM18956	JPT	inv/inv
GM18994	JPT	inv/inv
GM15510	n.a.	inv/inv

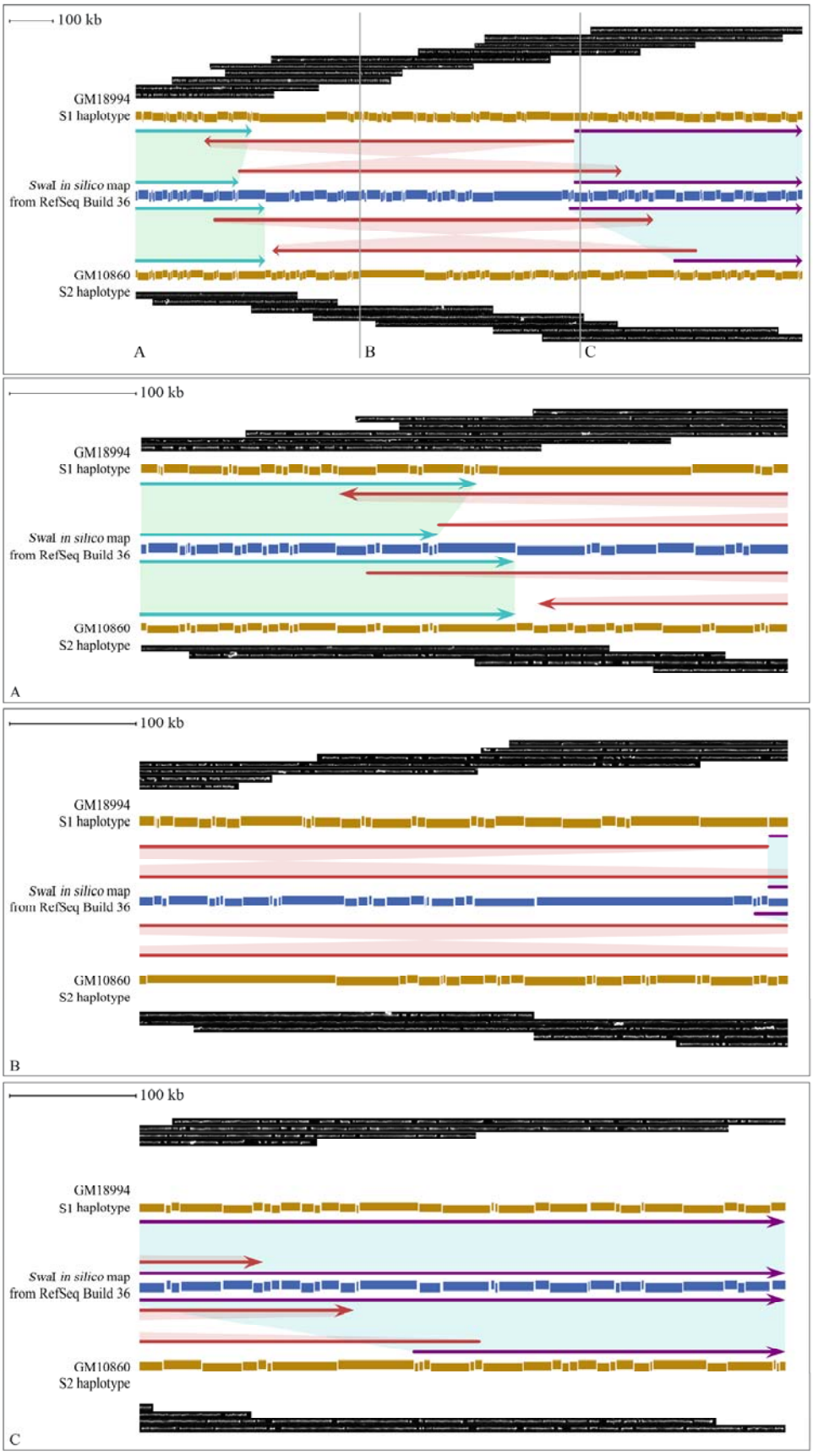
Supplementary Note Table 4 Inversion status assessed by FISH in 10 HapMap individuals.

2.2 Optical mapping

As another orthogonal-based approach, we confirmed the structure of this region using optical mapping single-molecule restriction maps from the genomes of GM18994 and GM10860 cell lines [8,9]. In order to identify the inversion, we compared the consensus maps to a restriction map generated *in silico* from the hg18 human genome reference sequence. Both maps demonstrate a large inversion spanning from BP1 through BP3 (Supplementary Note Figure 10; Figure S1 is a higher resolution version of Supplementary Note Figure 10) and are not consistent with current reference sequence assembly. We were able match the consensus restriction map from GM18994 to the S1

haplotype and the consensus restriction map from GM10860 to the S2 haplotype. Based on the large-scale structure of this region, the map for the S2 haplotype would suggest a different ordering for the green, red and blue probes at BP1 and BP3, as shown by previous FISH experiments.

Nature Genetics: doi:10.1038/ng.643


100 kb
GM18994
S1 haplotype
SwaI in silico map
from RefSeq Build 36
GM10860
S2 haplotype
A
B
C
100 kb
GM18994
S1 haplotype
SwaI in silico map
from RefSeq Build 36
GM10860
S2 haplotype
A
100 kb
GM18994
S1 haplotype
SwaI in silico map
from RefSeq Build 36
GM10860
S2 haplotype
B
100 kb
GM18994
S1 haplotype
SwaI in silico map
from RefSeq Build 36
GM10860
S2 haplotype
C

Supplementary Note Figure 10 Optical mapping data for the 16p12.1 locus. The top panel shows the 16p12.1 locus examined in two whole-genome optical mapping analyses, those of the HapMap panel members GM10860 and GM18994. The figure also shows three enlarged sections (A, B, and C) of the 16p12.1 optical map. The arrows demonstrate the alignment between the optical mapping consensus restriction maps (yellow) and a restriction map created *in silico* from the hg18 reference sequence (build 36) (blue). Places where multiple arrows overlap indicate apparent duplications in the optical map as compared to the reference. A montage of representative single-DNA molecule micrographs is provided for each consensus restriction map.

2.3 Sequencing of the S2 haplotype from large-insert BAC clones

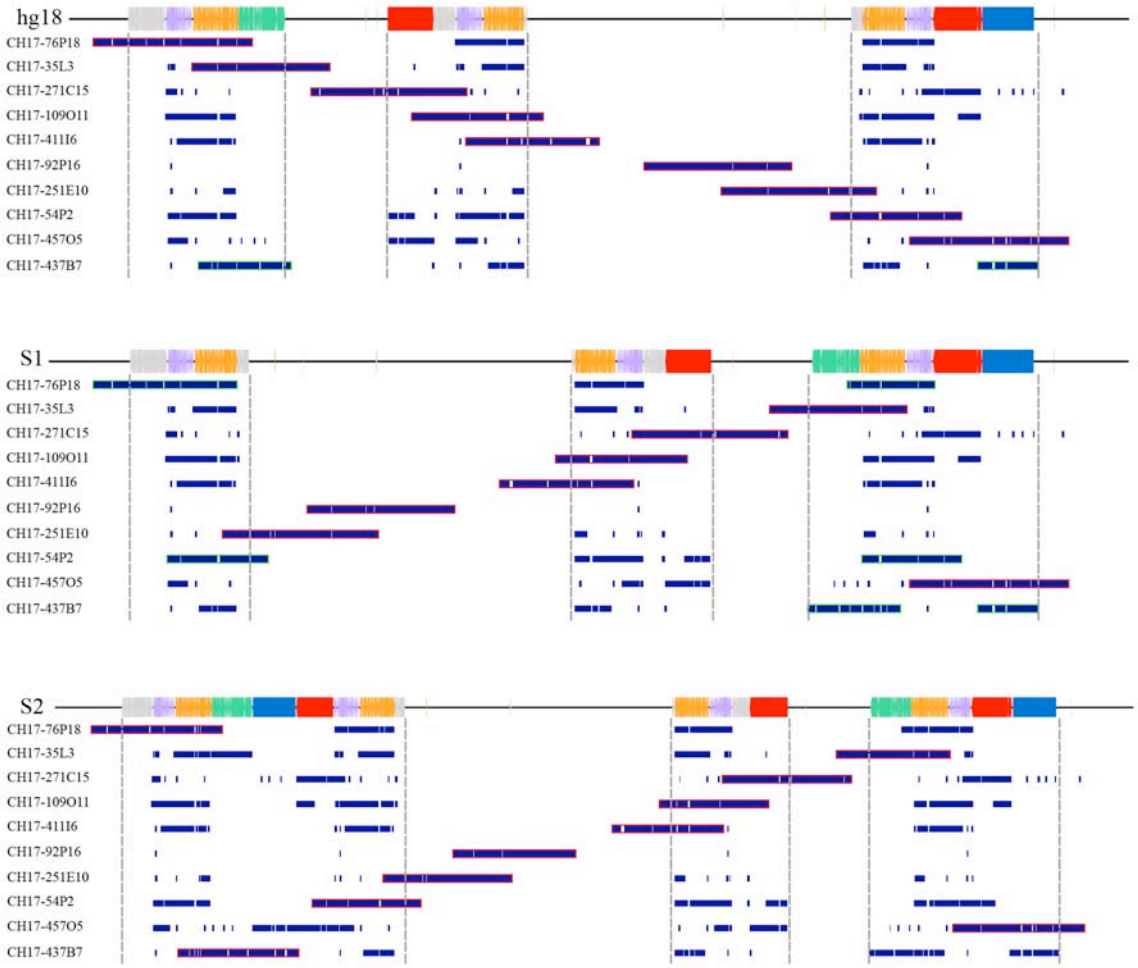
As a final step to resolve the genomic architecture of the region, we selected BAC clones from the genome of a complete hydatidiform mole (CHM1hTERT) corresponding to this region of 16p12.1. Since complete hydatidiform moles result from fertilization of an enucleated egg by a single sperm, they represent a haploid genome where all sequence variation must be non-allelic in origin [10]. We selected 10 clones based on BAC end-sequence mapping against the human reference genome, constructed individual bar-coded libraries for each clone, and subjected the pooled library to Illumina GAIIX sequencing (Supplementary Note Table 5) [11-13].

Nature Genetics: doi:10.1038/ng.643

Clones	Barcodes	
CH17-76P18	BCLO-1	[Phos]GATGACTTCGTAAGATCGGAAGAGCGTCGTGTA
	BCHI-1	TACACGACGCTCTTCCGATCTTACGAAGTCATC*T
CH17-35L3	BCLO-2	[Phos]GATAATCTCGTCAGATCGGAAGAGCGTCGTGTA
	BCHI-2	TACACGACGCTCTTCCGATCTGACGAGATTATC*T
CH17-271C15	BCLO-3	[Phos]GATTCTTACGGTAGATCGGAAGAGCGTCGTGTA
	BCHI-3	TACACGACGCTCTTCCGATCTACCGTAAGAATC*T
CH17-109O11	BCLO-4	[Phos]GATTTGCCACTAAGATCGGAAGAGCGTCGTGTA
	BCHI-4	TACACGACGCTCTTCCGATCTTAGTGGCAAATC*T
CH17-411I6	BCLO-5	[Phos]GATGCGTTAATGAGATCGGAAGAGCGTCGTGTA
	BCHI-5	TACACGACGCTCTTCCGATCTCATTAACGCATC*T
CH17-92P16	BCLO-6	[Phos]GATCTTCAACGAAGATCGGAAGAGCGTCGTGTA
	BCHI-6	TACACGACGCTCTTCCGATCTTCGTTGAAGATC*T
CH17-251E10	BCLO-7	[Phos]GATAGCGTACTAAGATCGGAAGAGCGTCGTGTA
	BCHI-7	TACACGACGCTCTTCCGATCTTAGTACGCTATC*T
CH17-54P2	BCLO-8	[Phos]GATTGATCTGAGAGATCGGAAGAGCGTCGTGTA
	BCHI-8	TACACGACGCTCTTCCGATCTCTCAGATCAATC*T
CH17-457O5	BCLO-9	[Phos]GATTACGGTGAAAGATCGGAAGAGCGTCGTGTA
	BCHI-9	TACACGACGCTCTTCCGATCTTTCACCGTAATC*T
CH17-437B7	BCLO-10	[Phos]GATATGCATGACAGATCGGAAGAGCGTCGTGTA
	BCHI-10	TACACGACGCTCTTCCGATCTGTCATGCATATC*T

Supplementary Note Table 5 The table shows 10 BAC clones from the genome of a complete hydatidiform mole (CHM1hTERT) sequenced using Illumina technology. The barcodes ligated to each sample during genomic library preparation are shown.

A total of 406 Mbp (6,345,136 X 64-bp paired-end reads) were mapped using mrsFAST [14] to the 16p12.1 region in hg18 corresponding to ~270-fold sequence coverage. We searched for the best fit for sequences to three reconstructed versions of the region (S1, S2 and hg18) (Supplementary Note Figure 11). Inconsistent alignments (highlighted in green) were found for both hg18 and S1 structures, while all clone-binned sequence showed colinearity (highlighted in red) for the S2 structure, concluding that the complete hydatidiform mole carries the S2 haplotype.



Supplementary Note Figure 11 A contig of 10 BAC clones from the 16p12.1 region were identified from a complete hydatidiform mole (CHM1hTERT) BAC library (CHORI-17). Inserts were sequenced using Illumina technology. Reads corresponding to each clone were mapped to the human genome reference assembly (hg18) and the alternate reconstructed S1 and S2 assemblies. Shown are the segmental duplications (colored boxes) annotated using SegDupMasker [5] for hg18 S1 and S2 haplotype structures. The gray dashed lines delimit regions where reads mapped to multiple locations due to the presence of segmental duplications. All concordant clones are highlighted in red and the discordant ones in green. All clones mapped against the S2 structure are concordant and colinear, while clone CH17-437B7 is discordant when mapped against the hg18 reference genome assembly and the S1 structure. Clones CH17-76P18

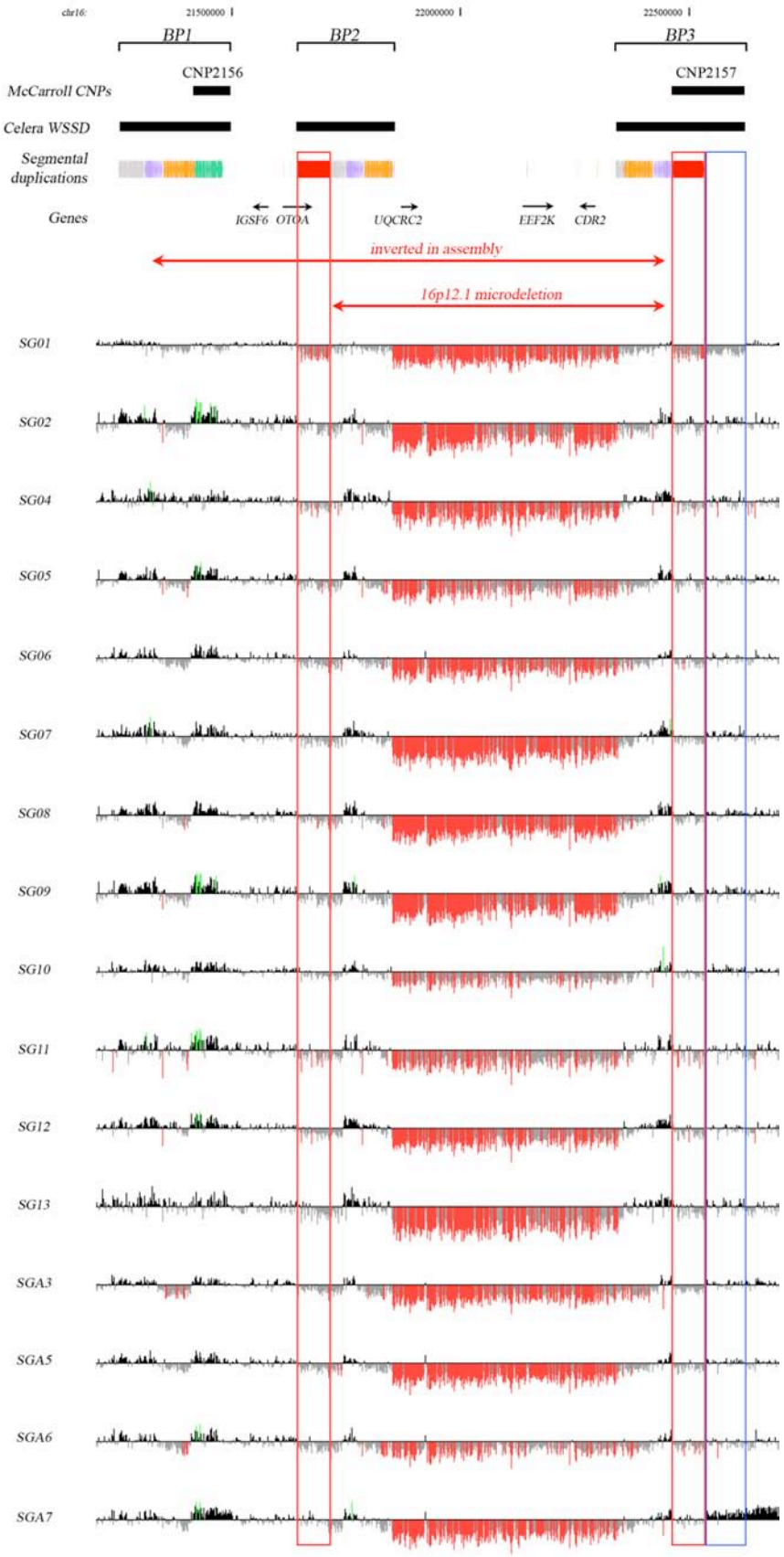
and CH17-54P2 are discordant only when mapped against the S1 haplotype. These data confirm genomically the existence of the S2 structure and define a minimal tiling path of clones for future complete high quality sequencing.

3. 16p12.1 microdeletion samples analysis

3.1 ArrayCGH breakpoints analysis and genotyping

Using high density and targeted array-based comparative genomic hybridization (CGH) experiments (NimbleGen, 50,000 probes with a density of one probe per 40 bp along 2 Mb at 16p12.1), we mapped the 16p12.1 microdeletion breakpoints in 15/16 patients (Supplementary Note Table 6) to the 68 kbp of polymorphic duplicated sequence present in direct orientation only on S2 (red box). HapMap sample NA15724 with S2/S2 genotype was used as reference (Supplementary Note Figure 12; Figure S2 is a higher resolution version of Supplementary Note Figure 12).

Nature Genetics: doi:10.1038/ng.643

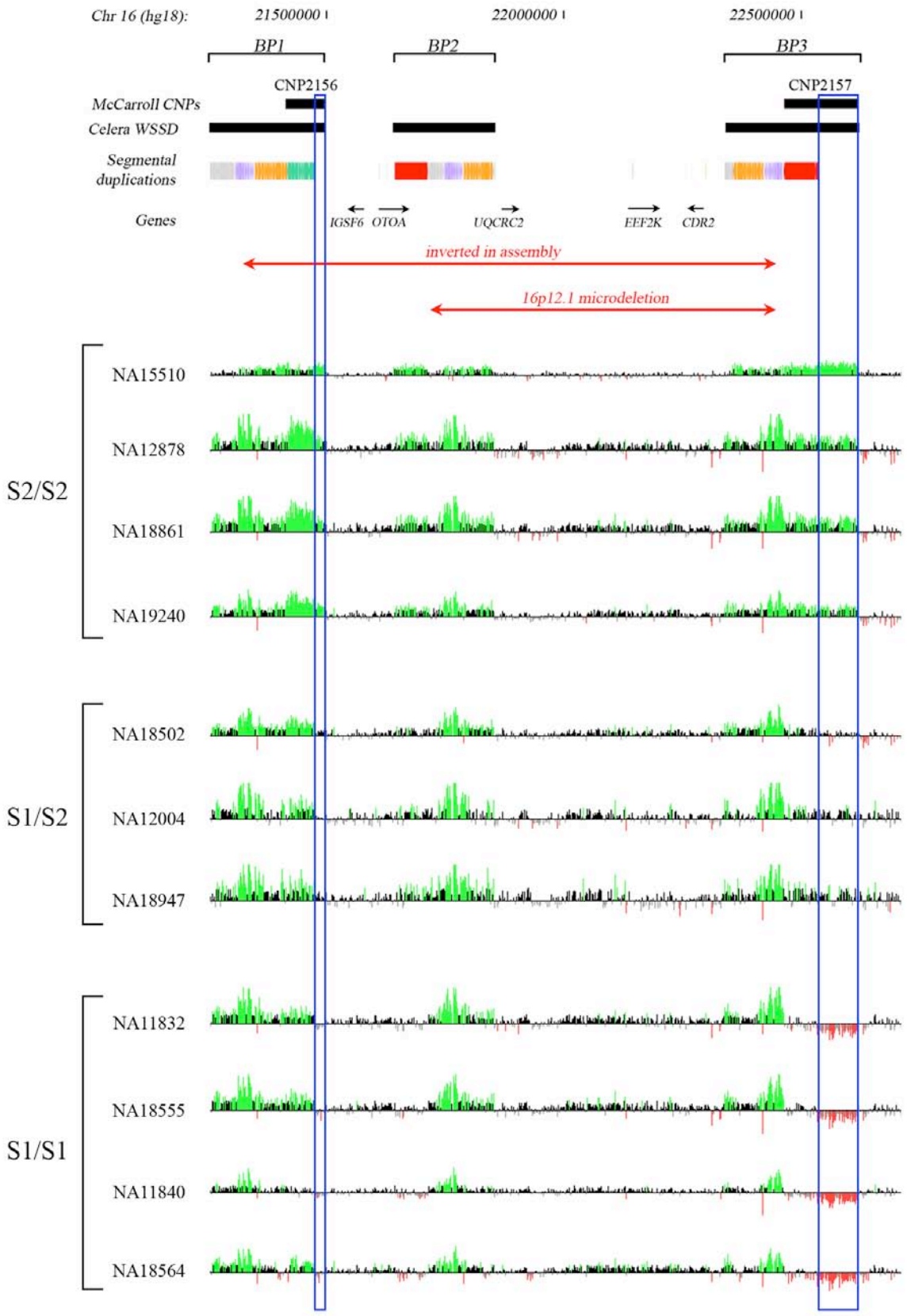

chr16:
21500000
22000000
22500000
BP1
BP2
BP3
CNP2156
CNP2157
McCarroll CNPs
Celera WSSD
Segmental duplications
Genes
IGSF6
OTOA
UQCRC2
EEF2K
CDR2
inverted in assembly
16p12.1 microdeletion
SG01
SG02
SG04
SG05
SG06
SG07
SG08
SG09
SG10
SG11
SG12
SG13
SGA3
SGA5
SGA6
SGA7

Supplementary Note Figure 12 ArrayCGH data from 16 microdeletion samples are shown for the 16p12.1 region. Segmental duplications were annotated using SegDupMasker [5]. The location of duplications identified by read-depth is also depicted (Celera WSSD). Shown are the positions of copy-number polymorphisms (CNP2156 and CNP2157) from McCarroll *et al.* [4]. Empty red boxes highlight the breakpoints of the 16p12.1 microdeletion mapping at the 68 kbp of polymorphic duplicated sequence present in direct orientation only on S2. The empty blue box highlights the S2-specific duplication that has a diploid copy number of 2 in S1/S1 individuals, 3 in S1/S2 heterozygotes, and 4 in S2/S2 homozygotes. The HapMap sample NA15724 with S2/S2 genotype was used as reference.

Based on our analysis of the region, only the S2 structural configuration would possess segmental duplications in direct orientation and therefore be predisposed to microdeletion. If this was the case, we would expect an enrichment of S2 chromosomes among patient samples. In contrast, parental S1/S1 homozygotes should be protective and for that reason should rarely be observed among patients where the deletion has recently emerged. We determined the structural genotype present in each of the 16p12.1 microdeletion cases using arrayCGH. It was possible to distinguish between these two structural configurations based on differences in the segmental duplication architecture (see above). In particular the S2-specific duplication block, corresponding to the distal segment of CNP2157 (empty blue box in Figure 2; Supplementary Note Figure 13; Supplementary Note Figure 14; Figure S3 is an higher resolution version of Supplementary Note Figure 14), has a diploid copy number of 2 in S1/S1 individuals, 3 in S1/S2 heterozygotes, and 4 in S2/S2 homozygotes. 35 microdeletion samples were hybridized against two control samples with known genotypes (NA15724 that is S2/S2

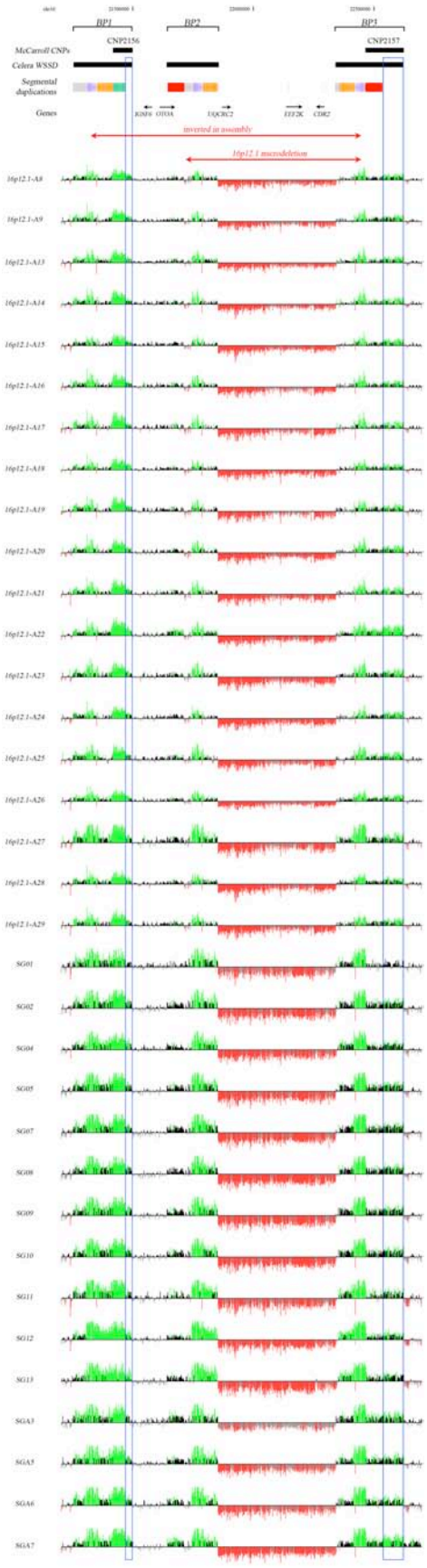
and NA18956 that is S1/S2). Based on the observed mean $\log_2$ values for the S2-specific duplication block, the genotype of each sample was determined (see Figure 2; Supplementary Note Figure 12; Supplementary Note Figure 14). Using HapMap population allele frequency estimates [4], we expected 13.7% S1 frequency (9 chromosomes out of 70) after controlling for the ethnicity of our cases (78% Caucasian vs. 22% African descent). We identified a single S1 allele (S1/S2 heterozygote) or an allele frequency of 1.4% (Supplementary Note Table 6). Among cases, there is a significant deficit of S1 alleles (p-value = 0.0088) and we can reject the Hardy-Weinberg equilibrium. This four-fold enrichment of the S2 haplotype among cases suggests that this structural polymorphism predisposes to the 16p12.1 microdeletion and subsequent neuropsychiatric disease (see Table 1).

Nature Genetics: doi:10.1038/ng.643


Chr 16 (hg18):
21500000
22000000
22500000
BP1
BP2
BP3
CNP2156
CNP2157
McCarroll CNPs
Celera WSSD
Segmental duplications
Genes
IGSF6
OTOA
UQCRC2
EEF2K
CDR2
inverted in assembly
16p12.1 microdeletion
S2/S2
NA15510
NA12878
NA18861
NA19240
S1/S2
NA18502
NA12004
NA18947
S1/S1
NA11832
NA18555
NA11840
NA18564

Supplementary Note Figure 13 ArrayCGH data from 11 HapMap samples are shown for the 16p12.1 region. The positions of copy-number polymorphisms (CNP2156 and CNP2157) from McCarroll *et al.* [4] are indicated. Segmental duplications were annotated using SegDupMasker [5]. The location of duplications identified by read-depth is also depicted (Celera WSSD). Empty blue 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. Based on the observed mean $\log_2$ values the genotype of each sample was determined.

Nature Genetics: doi:10.1038/ng.643


BP1
BP2
BP3
CNP2156
CNP2157
McCarroll CNPs
Celera WSSD
Segmental duplications
Genes
KIAA1468
OTOA
UQCRC2
ITFG2
CDR2
inverted in assembly
16p12.1 microdeletion
16p12.1-A8
16p12.1-A9
16p12.1-A13
16p12.1-A14
16p12.1-A15
16p12.1-A16
16p12.1-A17
16p12.1-A18
16p12.1-A19
16p12.1-A20
16p12.1-A21
16p12.1-A22
16p12.1-A23
16p12.1-A24
16p12.1-A25
16p12.1-A26
16p12.1-A27
16p12.1-A28
16p12.1-A29
SG01
SG02
SG04
SG05
SG07
SG08
SG09
SG10
SG11
SG12
SG13
SGA3
SGA5
SGA6
SGA7

Supplementary Note Figure 14 ArrayCGH data from 34 microdeletion samples are shown for the 16p12.1 region. Segmental duplications were annotated using SegDupMasker [5]. The location of duplications identified by read-depth is also depicted (Celera WSSD). Shown are the positions of copy-number polymorphisms (CNP2156 and CNP2157) from McCarroll *et al.* [4]. Empty blue 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. The HapMap sample NA18956 with S1/S2 genotype was used as reference. Based on the observed mean $\log_2$ values the genotype of each sample was determined.

Nature Genetics: doi:10.1038/ng.643

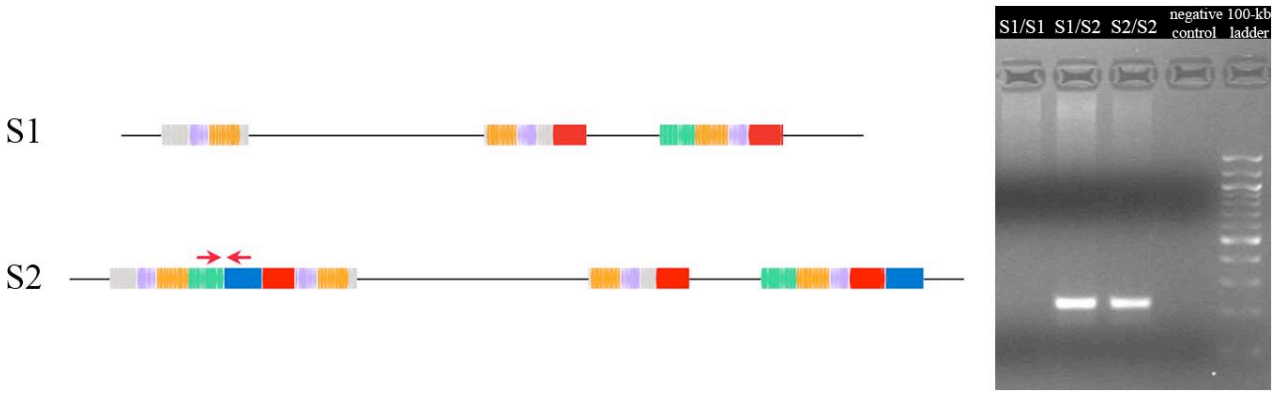
Sample	Ethnicity	ArrayCGH I (NA15724 S2/S2 reference)	ArrayCGH II (NA18956 S1/S2 reference)
SG01	Caucasian	S1/S2	S1/S2
SG02	Caucasian	S2/S2	S2/S2
SG04	Caucasian	S2/S2	S2/S2
SG05	Caucasian	S2/S2	S2/S2
SG06	Caucasian	S2/S2	n.a.
SG07	Caucasian	S2/S2	S2/S2
SG08	African american	S2/S2	S2/S2
SG09	African american	S2/S2	S2/S2
SG10	Caucasian	S2/S2	S2/S2
SG11	Caucasian	S2/S2	S2/S2
SG12	n.a.	S2/S2	S2/S2
SG13	Caucasian	S2/S2	S2/S2
SGA3	African american	S2/S2	S2/S2
SGA5	Caucasian	S2/S2	S2/S2
SGA6	Caucasian	S2/S2	S2/S2
SGA7	Caucasian	S2/S2	S2/S2
16p12.1-A8	African american	n.a.	S2/S2
16p12.1-A9	Caucasian	n.a.	S2/S2
16p12.1-A13	African american	n.a.	S2/S2
16p12.1-A14	Caucasian	n.a.	S2/S2
16p12.1-A15	n.a.	n.a.	S2/S2
16p12.1-A16	n.a.	n.a.	S2/S2
16p12.1-A17	n.a.	n.a.	S2/S2
16p12.1-A18	Caucasian	n.a.	S2/S2
16p12.1-A19	n.a.	n.a.	S2/S2
16p12.1-A20	Caucasian	n.a.	S2/S2
16p12.1-A21	Caucasian	n.a.	S2/S2
16p12.1-A22	Caucasian	n.a.	S2/S2
16p12.1-A23	Caucasian	n.a.	S2/S2
16p12.1-A24	n.a.	n.a.	S2/S2
16p12.1-A25	n.a.	n.a.	S2/S2
16p12.1-A26	African american	n.a.	S2/S2
16p12.1-A27	Caucasian	n.a.	S2/S2
16p12.1-A28	Caucasian	n.a.	S2/S2
16p12.1-A29	n.a.	n.a.	S2/S2

Supplementary Note Table 6 Inferred genotypes by arrayCGH for 35 analyzed samples with 16p12.1 microdeletion. NA15724 (S2/S2) and NA18956 (S1/S2) were used as reference in two sets of samples.

Nature Genetics: doi:10.1038/ng.643

3.2 PCR analysis

We designed a PCR-based assay to genotype S1 and S2 alleles. PCR primers were designed to specifically amplify an S2-specific segmental duplication junction (Supplementary Note Figure 15), namely (S2F = GCCAAGGAAGCTGCATTTTA, S2R = CTTAGCACAGGGCAGACCAT). Twenty-seven HapMap samples, whose genotypes were known based on McCarroll calls [4], were used as controls to validate the method (Supplementary Note Figure 15; Supplementary Note Table 7) and showed perfect correspondence to inferred SNP microarray CNP genotypes. Note the absence of an amplification product would indicate an S1/S1 genotype. All of the 31 microdeletion samples tested gave an amplification product by PCR (Supplementary Note Table 8), indicating the presence of at least one S2 haplotype, therefore supporting our hypothesis that the S2 structure is predisposing to the microdeletion.



Supplementary Note Figure 15 PCR genotyping using primers designed for an S2-specific segmental duplication junction (red arrows) shows an amplification product for NA18956 (S1/S2) and NA12878 (S2/S2) HapMap samples with known genotypes. No amplification product is shown for NA18555 (S1/S1).

HapMap ID	Population	Genotype based on MCCarroll calls	PCR
NA10847	CEU	S2/S2	+
NA11840	CEU	S1/S1	-
NA11832	CEU	S1/S1	-
NA11993	CEU	S2/S2	+
NA12156	CEU	S2/S2	+
NA12004	CEU	S1/S2	+
NA12813	CEU	S2/S2	+
NA12878	CEU	S2/S2	+
NA18502	YRI	S1/S2	+
NA18507	YRI	S2/S2	+
NA18517	YRI	S2/S2	+
NA18523	YRI	S2/S2	+
NA18861	YRI	S2/S2	+
NA19102	YRI	S2/S2	+
NA19172	YRI	S2/S2	+
NA19116	YRI	S2/S2	+
NA19129	YRI	S2/S2	+
NA19132	YRI	S2/S2	+
NA19240	YRI	S2/S2	+
NA18552	CHB	S2/S2	+
NA18555	CHB	S1/S1	-
NA18564	CHB	S1/S1	-
NA18573	CHB	S2/S2	+
NA18942	JPT	S1/S2	+
NA18947	JPT	S1/S2	+
NA18956	JPT	S1/S2	+
NA18980	JPT	S2/S2	+

Supplementary Note Table 7 PCR genotyping experiments were performed on 27 HapMap control samples whose genotypes were inferred based on McCarroll calls [4].

Sample	PCR
SG01	+
SG02	+
SG04	+
SG05	+
SG06	+
SG07	+
SG08	+
SG09	n.a.
SG10	+
SG11	+
SG12	+
SG13	+
SGA3	n.a.
SGA5	n.a.
SGA6	n.a.
SGA7	+
16p12.1-A8	+
16p12.1-A9	+
16p12.1-A13	+
16p12.1-A14	+
16p12.1-A15	+
16p12.1-A16	+
16p12.1-A17	+
16p12.1-A18	+
16p12.1-A19	+
16p12.1-A20	+
16p12.1-A21	+
16p12.1-A22	+
16p12.1-A23	+
16p12.1-A24	+
16p12.1-A25	+
16p12.1-A26	+
16p12.1-A27	+
16p12.1-A28	+
16p12.1-A29	+

Supplementary Note Table 8 PCR genotyping experiments were performed on 31 microdeletion samples. No DNA was available to test four patients' samples (n.a.). All samples tested gave an amplification product by PCR (+).

3.3 ArrayCGH genotyping in six HapMap populations

We assessed the population frequency for the S1 and S2 configurations analyzing 357 additional individuals from different populations including 118 Maasai individuals and

60 Luhya individuals using a custom Agilent 4x180K microarray targeted to copy-number polymorphic regions of the human genome (Campbell *et al.*, unpublished). This microarray contains 50 probes in the CNP2157 at chr16:22533636-22618896. Based on the observed mean log2 values for the S2-specific duplication block corresponding to the distal segment of CNP2157, the genotype of each sample was determined (Supplementary Note Table 9). These data confirmed the low frequency of the protective S1 haplotype in African populations when compared to Asians.

Population	Number of individuals	S1 frequency	S2 frequency
Maasai (MKK)	118	0.08	0.92
Luhya (LWK)	60	0.02	0.98
Yorubans (YRI)	56	0.03	0.97
Europeans (CEU)	54	0.11	0.89
Chinese (CHB)	34	0.13	0.87
Japanese (JPT)	35	0.33	0.67

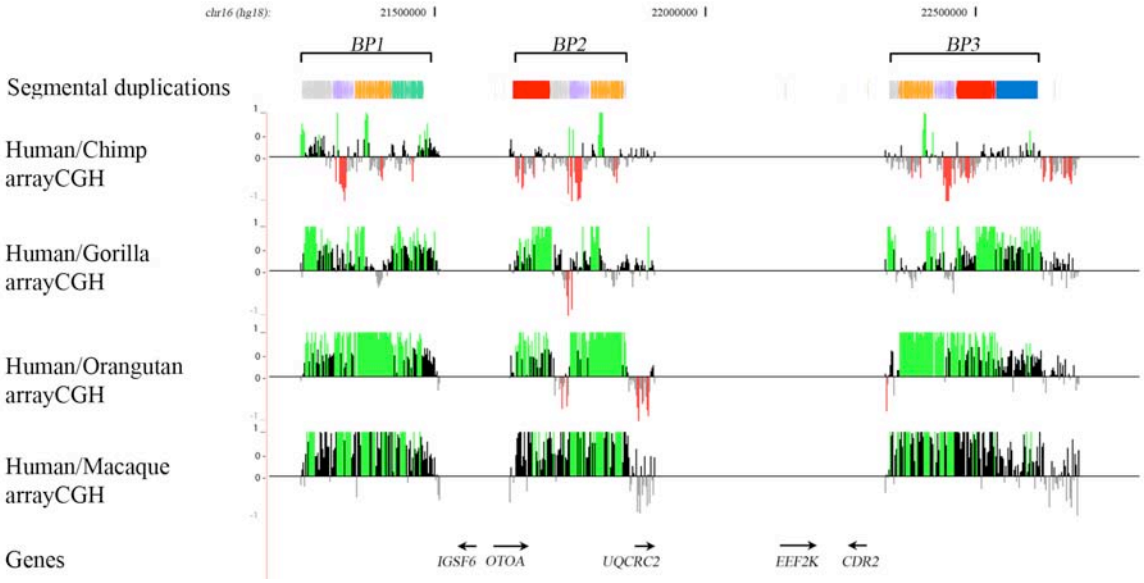
Supplementary Note Table 9 The frequencies of S1 and S2 haplotypes in six HapMap populations are shown.

4. Evolutionary origin

4.1 Segmental duplications analysis

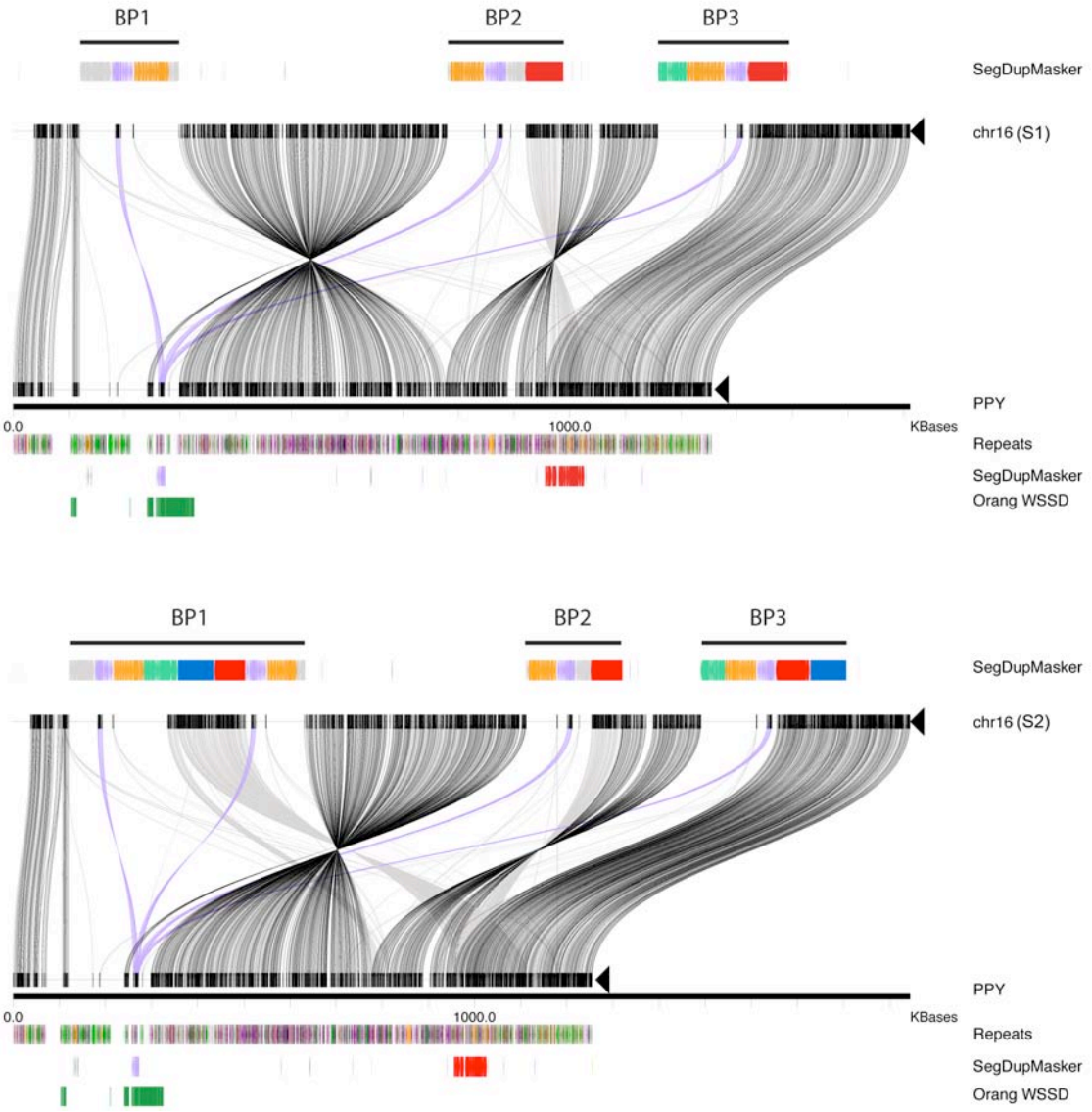
We compared the extent of segmental duplications within the 16p12.1 region among human, chimpanzee, gorilla, orangutan, gibbon and macaque using whole-genome shotgun sequences (WGS) [15] and comparative genomic hybridization using a targeted oligonucleotide array (arrayCGH). These analyses showed an expansion of segmental duplications in the African great apes (human, chimpanzee, gorilla) with respect to orangutan, gibbon and macaque (see Figure 4; Supplementary Note Figure 16). This particular region of 16p12.1 has increased in size from 726 kbp to 1,671 kbp during the

last 10 million years primarily as a result of a duplicative transposition of segmental duplications in the region. Our analysis of the great apes suggests that the region has become increasingly complex in human leading to the addition of another polymorphic 333 kbp specifically in the human lineage.



Supplementary Note Figure 16 The figure shows a comparison of the extent of segmental duplications in the 16p12.1 region among four different primate species using comparative genomic hybridization with a targeted oligonucleotide array (arrayCGH) [16]. Also shown are the segmental duplications in human annotated using SegDupMasker [5].

Interestingly, sequence comparison of the orangutan sequence (WUGSC 2.0.2/ponAbe2) and human S1 and S2 haplotypes at 16p12.1 using the program miropeats [7] clearly shows an expansion of the 16p12.1 region due to segmental duplications formation accompanied by two local inversions of 481 kbp and 142 kbp (Supplementary Note Figure 17).



Supplementary Note Figure 17 Sequence comparison of the orangutan sequence (WUGSC 2.0.2/ponAbe2) and human S1 (top) and S2 (bottom) haplotypes at 16p12.1 using the program miropeats [7] shows an expansion of the 16p12.1 region due to segmental duplication formation accompanied by two local inversions of 481 kbp and 142 kbp. Black lines connect matching segments between the orangutan sequence and S1 and S2 sequence while light gray lines indicate lower-identity matches to distinct paralogs. Purple lines connect paralogous sequences of a particular segmental duplication (LCR16a), carrying the *NPIP* gene, that might have been the source locus for the formation of the other segmental duplications at 16p12.1 in great apes.

Segmental duplications were annotated using SegDupMasker [5]. The location of duplications identified by read-depth is also depicted (WSSD).

4.2 Non-human primate BAC clone sequence analysis

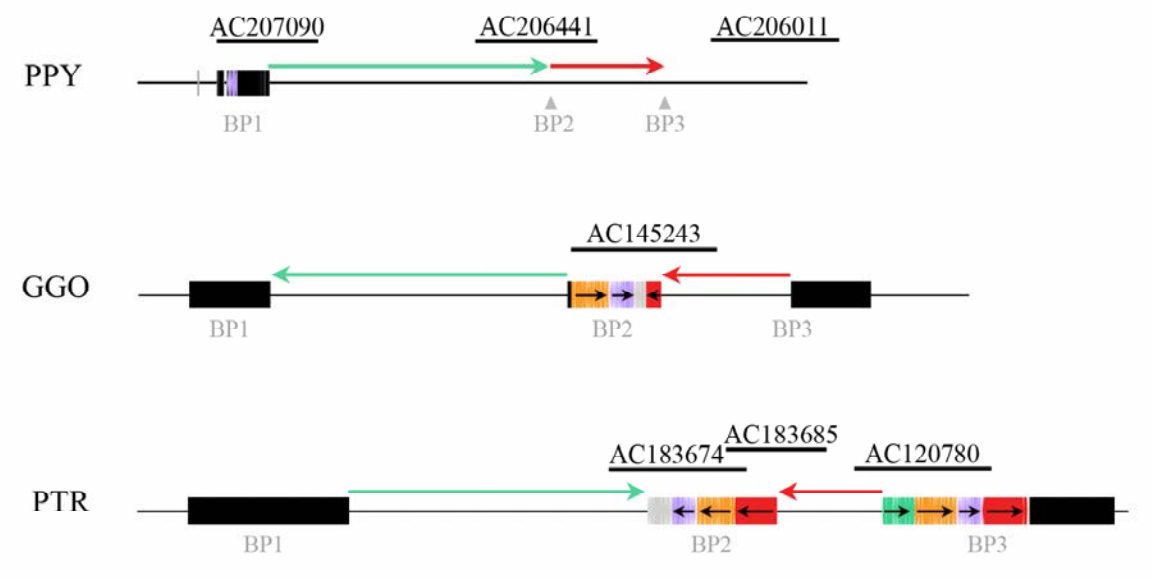
In order to investigate the ancestral configuration of the 16p12.1 region, we compared the orientation of the region in human with four outgroup non-human primate species. We selected nine BAC clones from the libraries of chimpanzee (CH251), orangutan (CH276), and gorilla (CH255) genomes mapping to the 16p12.1 segmental duplications in human (Supplementary Note Table 10). Each clone was high quality, fully sequenced (we also retrieved a previously published gorilla clone AC145243 [17]) and aligned to the human genome and to the S1 haplotype that we reconstructed with miropeats [7]. Final annotation with common repeats and DupMasker output [5] describing the composition of segmental duplications was also included with customized Perl scripts.

Clone	Insert Size (bp)	Species	Informative
AC183674	198084	PTR	yes
AC183685	164817	PTR	yes
AC120780	193047	PTR	yes
AC183619	178822	PTR	no
AC183100	173550	PTR	no
AC145243	227052	GGO	yes
AC206441	206416	PPY	yes
AC206011	199647	PPY	yes
AC207090	221568	PPY	yes

Supplementary Note Table 10 The table shows nine clones from the chimpanzee (PTR), gorilla (GGO), and orangutan (PPY) genomes that were high quality, fully sequenced and aligned to the human genome with blast [18] and miropeats [7].

Nature Genetics: doi:10.1038/ng.643

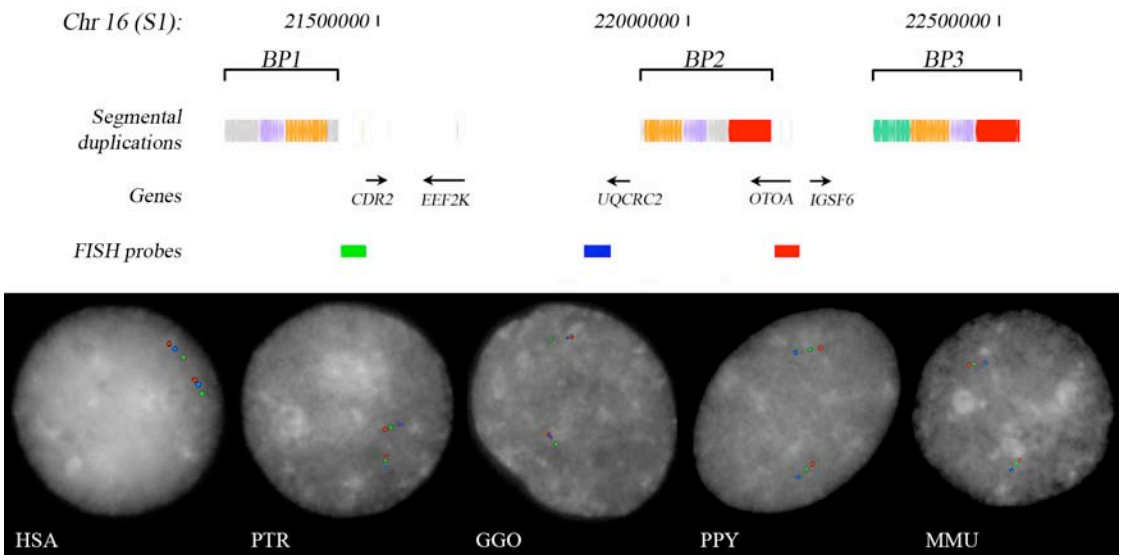
Notably, analysis of seven informative sequenced chimpanzee, orangutan and gorilla BAC clones (Figure S4; Supplementary Note Figure 18) indicated that orangutan and chimpanzee are inverted with respect to human for the region spanning from BP1 to BP2. Furthermore, orangutan is also inverted for the region spanning from BP2 to BP3 (see Figure 3).



Supplementary Note Figure 18 The figure shows the location of seven informative orangutan (PPY), gorilla (GGO) and chimpanzee (PTR) high quality, fully sequenced clones mapping at the 16p12.1 locus (see Supplementary Note Table 9).

We interrogated the *Macaca mulatta* genome assembly (rheMac2) and found that macaque is inverted for both the regions spanning from BP1 to BP2 and from BP2 to BP3 as in orangutan. We tested for the presence of the bigger inversion between BP1 and BP2 (481 kbp) by examining lymphoblastoid cell lines from three chimpanzee (*Pan troglodytes*), three orangutan (*Pongo pygmaeus*), two gorilla (*Gorilla gorilla*) and one

macaque (*Macaca mulatta*) individuals by FISH and confirmed the orientation of the region in all of them as shown by sequence analysis (Supplementary Note Figure 19).



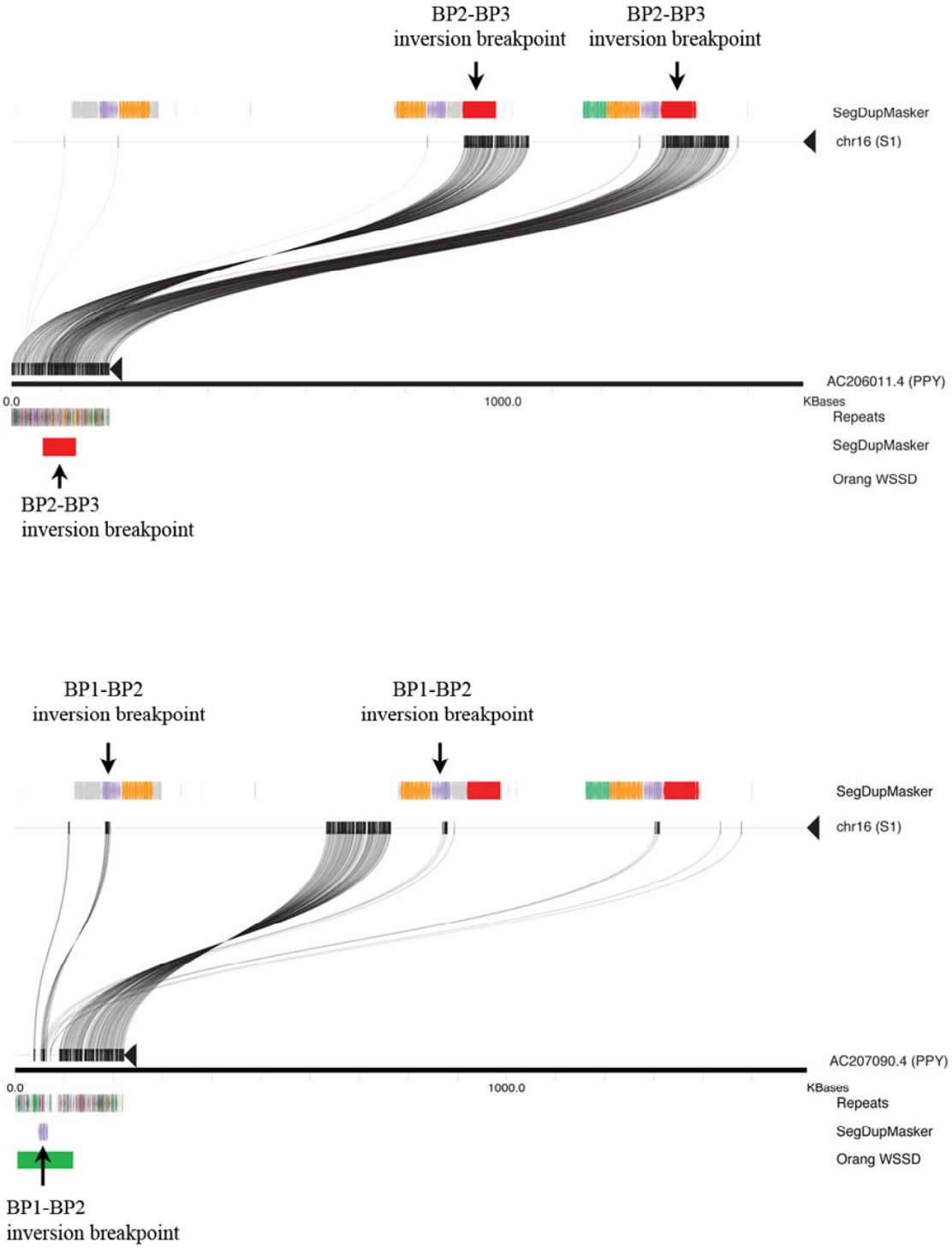
Supplementary Note Figure 19 The figure shows a cohybridization experiment using probes WIBR2-2735O20 (red), WIBR2-3632J22 (green) and WIBR2-3790H20 (blue) on human (HSA, GM18956), chimpanzee (PTR), gorilla (GGO), orangutan (PPY), and macaque (MMU) in order to determine the orientation of the region spanning from BP1 to BP2 in non-human primate species. The inversion changes the order of the green and blue probes and their relative position with respect to the red probe. Also shown are the segmental duplications in the human S1 haplotype annotated using SegDupMasker [5].

These data indicate that the inverted configuration as found in orangutan and macaque is likely the ancestral state. Both inversions are likely to have occurred in the African great ape ancestor after formation and expansion of the segmental duplications (see Figure 3). Subsequently, the region spanning from BP1 to BP2 might have flipped back to the direct orientation in the chimpanzee lineage.

Based on the high-quality BAC-based non-human primate sequence data (Supplementary Note Figure 20) and on the comparison of the extent of segmental duplications using whole-genome shotgun sequences (WGS) [15], we attempted to reconstruct the organization of the 16p12.1 region in the African great ape ancestor (Supplementary Note Figure 21). Analysis of the sequenced non-human primate BACs using the program miropeats [7] allowed us to refine the location of the breakpoints of the smaller inversion (BP2-BP3) to two paralogous copies of the same segmental duplication found at the breakpoints of the 16p12.1 microdeletion (BP1-BP2) in human (red box). This duplication block is oriented in inverted configuration in relation to its paralog at the inversion proximal breakpoint (BP3), and in direct configuration to its paralog at the microdeletion distal breakpoint (BP1). We propose that NAHR between the directly oriented segmental duplication blocks at BP1 and BP2 in human results in the 16p12.1 microdeletions, whereas NAHR between the inverted duplication blocks at BP2 and BP3 led to an inversion of the intervening sequence in the ancestor of the African great apes during evolution.

We determined that the breakpoints of the larger inversion (BP1-BP2) map to the core duplicon LCR16a [19] carrying the gene *NPIP*. Paralogous copies of this segmental duplication are located in inverted orientation at the breakpoints of the inversion, between BP1 and BP2, and in direct orientation at the breakpoints of a non-pathogenic microdeletion of the region spanning from BP2 to BP3 found in 20/6712 controls [1]. Moreover, two additional paralogous copies of this core duplicon are located in direct

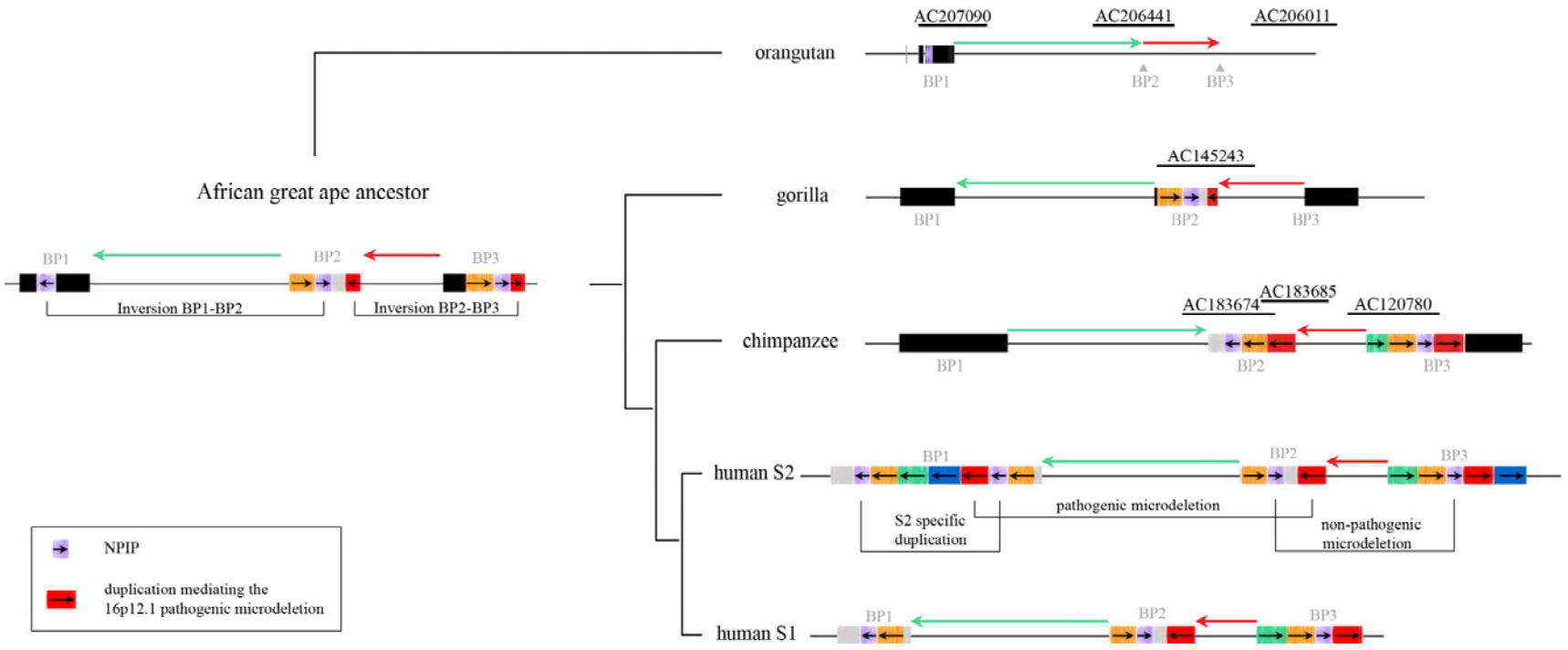
orientation at the boundaries of the complex 333-kbp copy-number polymorphism at BP1 specific of the S2 configuration.



Supplementary Note Figure 20 Sequence alignment of two representative orangutan BAC clones with the S1 reconstructed haplotype that were used to refine the location of the breakpoints of the two inversions in the African great ape ancestor. The purple rectangle represents the core duplicon LCR16a [19] carrying the gene *NPIP*, found at the breakpoints of the larger inversion (BP1-BP2). The red rectangle represents the same 68-kbp segmental duplication mediating the 16p12.1 microdeletion in human and found at the breakpoints of the smaller inversion (BP2-BP3). Black lines connect matching segments between the clone and the S1 haplotype sequence while light gray lines indicate lower identity matches to distinct paralogs. Segmental duplications were annotated using SegDupMasker [5]. The location of duplications identified by read-depth is also depicted (WSSD).

All these data indicate that segmental duplications at the 16p12.1 region have a complex structure consisting of both direct and inverted subunits that serve as NAHR substrates leading to pathogenic microdeletions, non-pathogenic rare CNVs and evolutionary inversions. It is interesting that all of the 16p12.1 changes are associated with the spread of the human-great ape gene family *morpheus* (*NPIP*) [19]. The core duplicon carrying this gene, LCR16a, maps to each of the breakpoint regions, including the boundaries of the 333-kbp CNP ant BP1 specific of the S2 configuration. The same gene family appears to be at the breakpoints of other recurrent microdeletions on chromosome 16 [20-26].

Nature Genetics: doi:10.1038/ng.643

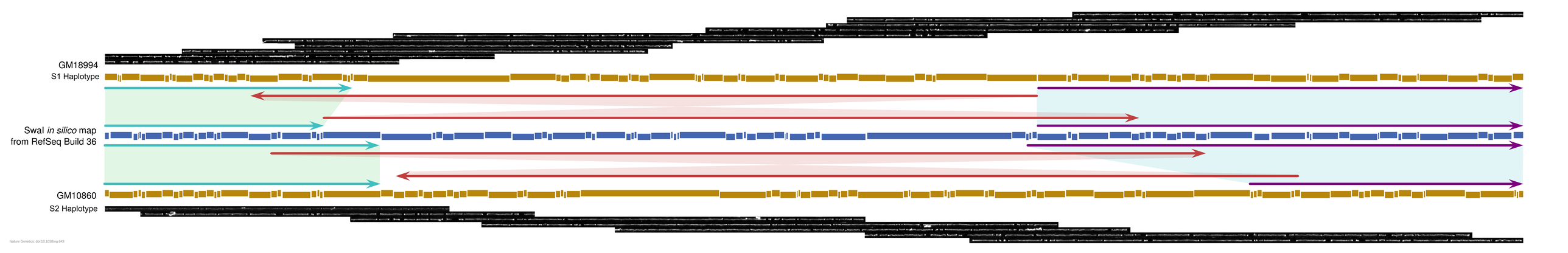


Supplementary Note Figure 21 Reconstruction of the evolutionary history of the 16p12.1 region. The figure shows the location of seven informative orangutan (PPY), gorilla (GGO) and chimpanzee (PTR) high quality, fully sequenced clones mapping at the 16p12.1 locus (see Supplementary Note Table 9) that were used to reconstruct the organization of the 16p12.1 region in the African great ape ancestor. Segmental duplications at the 16p12.1 region have a complex structure consisting of both direct and inverted subunits that serve as NAHR substrates leading to pathogenic microdeletions, non-pathogenic rare CNVs and evolutionary inversions.

References

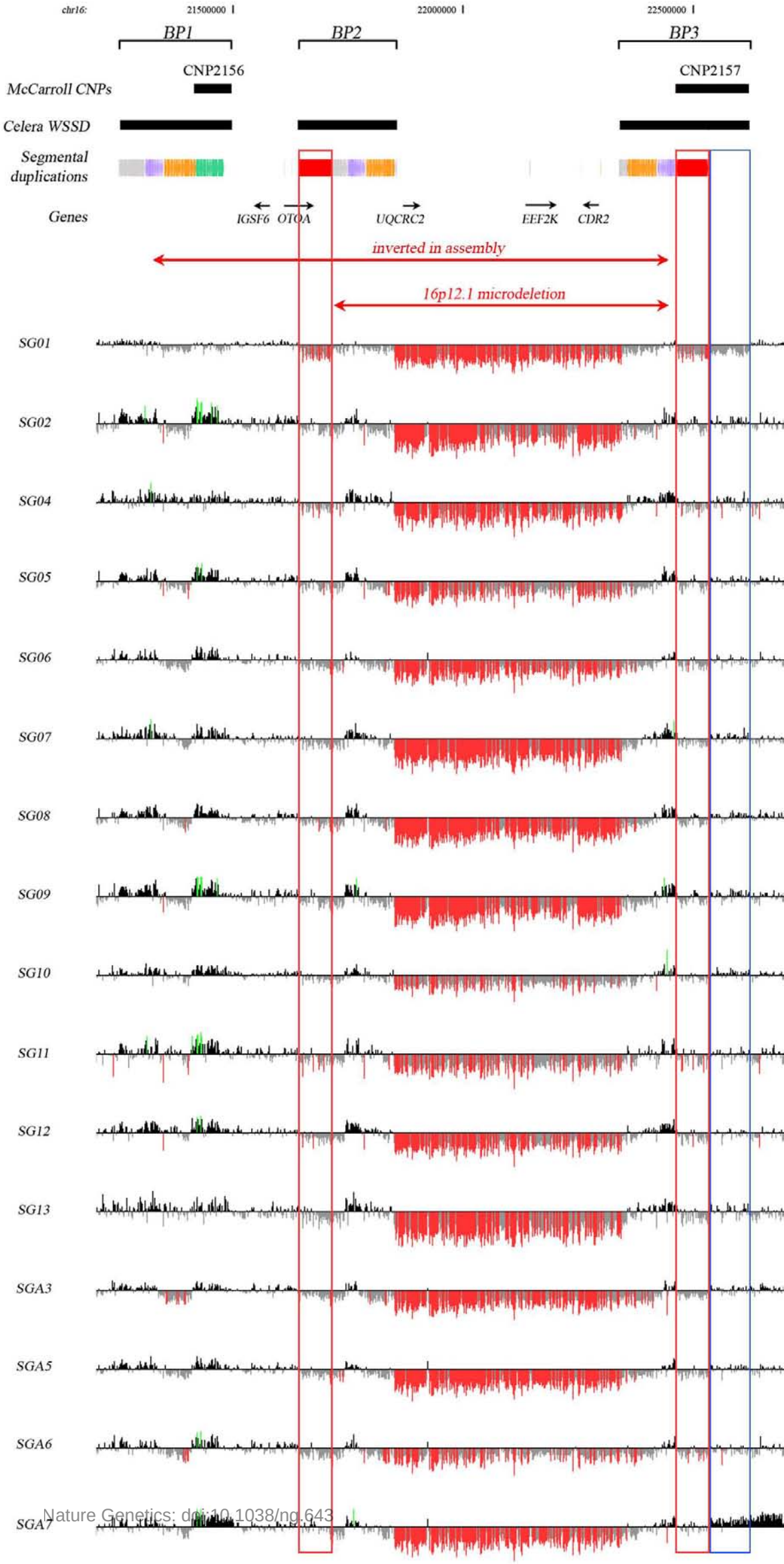
1. Itsara, A. et al. Population analysis of large copy number variants and hotspots of human genetic disease. *Am J Hum Genet* **84**, 148-61 (2009).
2. Girirajan, S. et al. A recurrent 16p12.1 microdeletion supports a two-hit model for severe developmental delay. *Nat Genet*.
3. Levy, S. et al. The Diploid Genome Sequence of an Individual Human. *PLoS Biol* **5**, e254 (2007).
4. McCarroll, S.A. et al. Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genet* **40**, 1166-74 (2008).
5. Jiang, Z., Hubley, R., Smit, A. & Eichler, E.E. DupMasker: a tool for annotating primate segmental duplications. *Genome Res* **18**, 1362-8 (2008).
6. Kidd, J.M. et al. Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56-64 (2008).
7. Parsons, J.D. Miropeats: graphical DNA sequence comparisons. *Comput Appl Biosci* **11**, 615-9 (1995).
8. Zhou, S. et al. A single molecule scaffold for the maize genome. *PLoS Genet* **5**, e1000711 (2009).
9. Teague, B. et al. High-resolution human genome structure by single-molecule analysis. *Proc Natl Acad Sci U S A* **107**, 10848-53.
10. Surti, U., Szulman, A.E. & O'Brien, S. Complete (classic) hydatidiform mole with 46,XY karyotype of paternal origin. *Hum Genet* **51**, 153-5 (1979).
11. Craig, D.W. et al. Identification of genetic variants using bar-coded multiplexed sequencing. *Nat Methods* **5**, 887-93 (2008).
12. Quail, M.A. et al. A large genome center's improvements to the Illumina sequencing system. *Nat Methods* **5**, 1005-10 (2008).
13. Ng, S.B. et al. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272-6 (2009).
14. Alkan, C. et al. Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat Genet* **41**, 1061-7 (2009).
15. Marques-Bonet, T., Girirajan, S. & Eichler, E.E. The origins and impact of primate segmental duplications. *Trends Genet* **25**, 443-54 (2009).
16. Marques-Bonet, T. et al. A burst of segmental duplications in the genome of the African great ape ancestor. *Nature* **457**, 877-81 (2009).
17. Johnson, M.E. et al. Recurrent duplication-driven transposition of DNA during hominoid evolution. *Proc Natl Acad Sci U S A* **103**, 17626-31 (2006).
18. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J Mol Biol* **215**, 403-10 (1990).
19. Johnson, M.E. et al. Positive selection of a gene family during the emergence of humans and African apes. *Nature* **413**, 514-9 (2001).
20. Weiss, L.A. et al. Association between microdeletion and microduplication at 16p11.2 and autism. *N Engl J Med* **358**, 667-75 (2008).
21. Kumar, R.A. et al. Recurrent 16p11.2 microdeletions in autism. *Hum Mol Genet* **17**, 628-38 (2008).

22. Ullmann, R. et al. Array CGH identifies reciprocal 16p13.1 duplications and deletions that predispose to autism and/or mental retardation. *Hum Mutat* **28**, 674-82 (2007).
23. 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**, 223-32 (2009).
24. Ballif, B.C. et al. Discovery of a previously unrecognized microdeletion syndrome of 16p11.2-p12.2. *Nat Genet* **39**, 1071-3 (2007).
25. Bochukova, E.G. et al. Large, rare chromosomal deletions associated with severe early-onset obesity. *Nature* **463**, 666-70.
26. Girirajan, S. et al. A recurrent 16p12.1 microdeletion supports a two-hit model for severe developmental delay. *Nat Genet* **42**, 203-9.



Nature Genetics: doi:10.1038/ng.643

Figure S1 Optical mapping data for the 16p12.1 locus. We examined the 16p12.1 locus in two whole-genome optical mapping analyses, those of the HapMap panel members GM10860 and GM18994. The arrows demonstrate the alignment between the optical mapping consensus restriction maps (yellow) and a restriction map created *in silico* from the hg18 reference sequence (build36) (blue). Places where multiple arrows overlap indicate apparent duplications in the optical map as compared to the reference. A montage of representative single-DNA molecule micrographs is provided for each consensus restriction map.



Nature Genetics: doi:10.1038/ng.643

Figure S2 ArrayCGH data from 16 microdeletion samples are shown for the 16p12.1 region. Segmental duplications were annotated using SegDupMasker [5]. The location of duplications identified by read-depth is also depicted (Celera WSSD). Shown are the positions of copy-number polymorphisms (CNP2156 and CNP2157) from McCarroll *et al.* [4]. Red empty boxes highlight the breakpoints of the 16p12.1 microdeletion mapping at the 68-kbp polymorphic duplicated sequence present in direct orientation only on S2. The blue empty box highlights the S2-specific duplication that has a diploid copy number of 2 in S1/S1 individuals, 3 in S1/S2 heterozygotes, and 4 in S2/S2 homozygotes. The HapMap sample NA15724 with S2/S2 genotype was used as reference.

Nature Genetics: doi:10.1038/ng.643

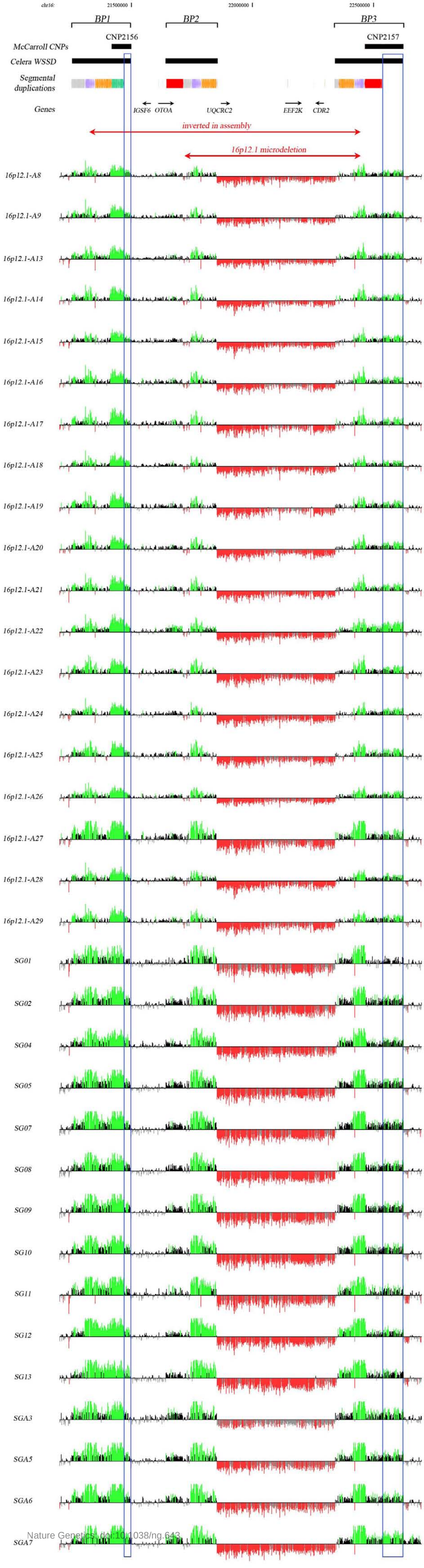
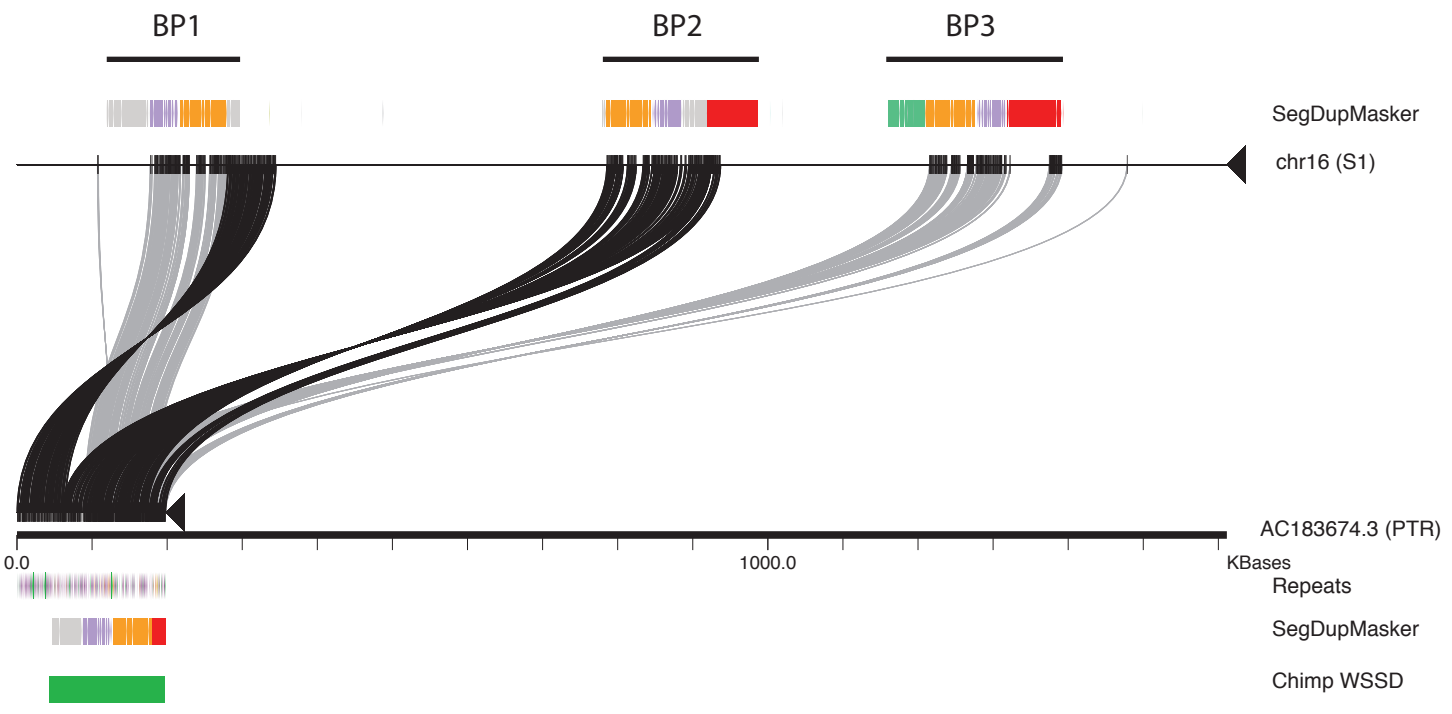
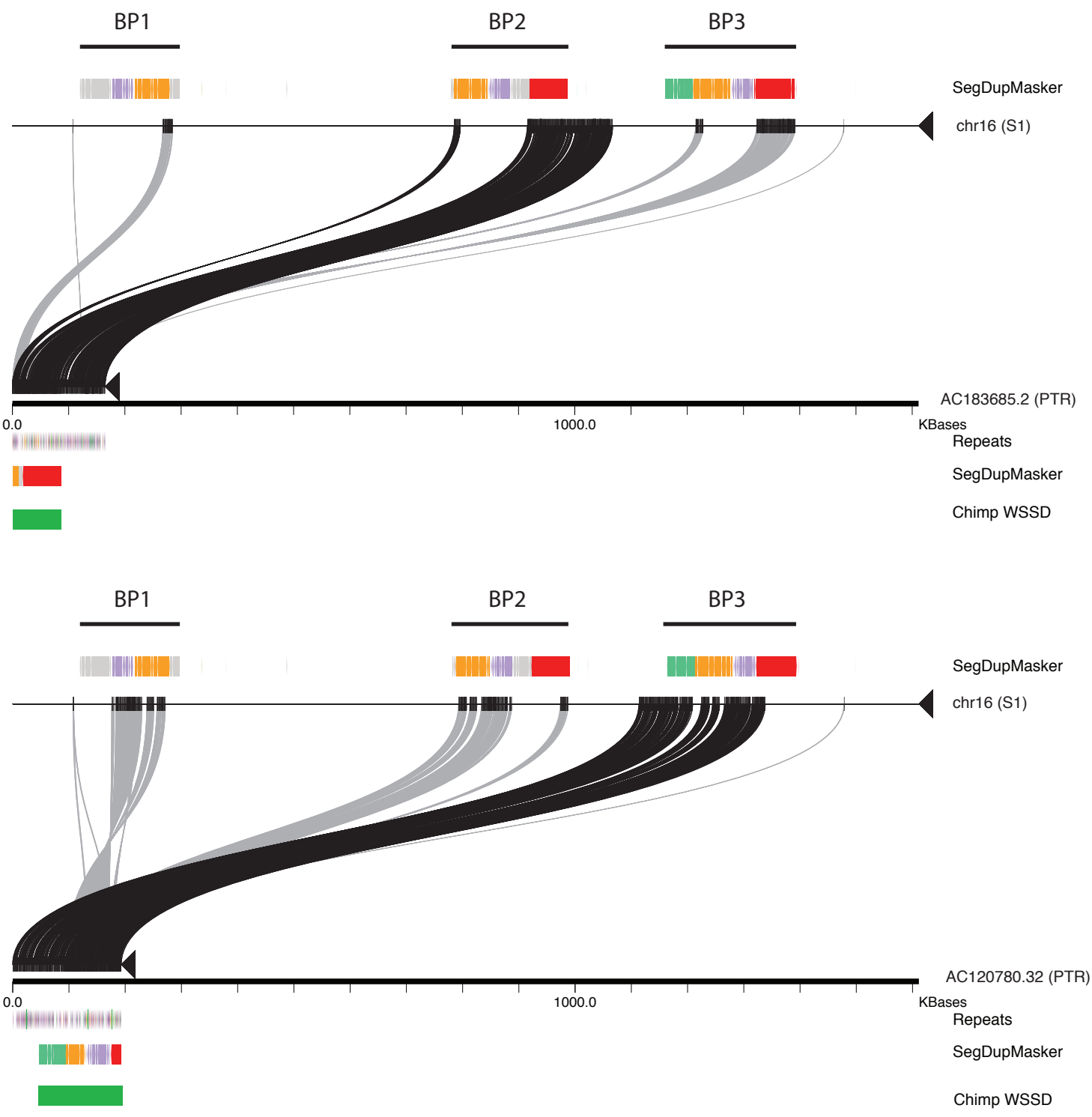
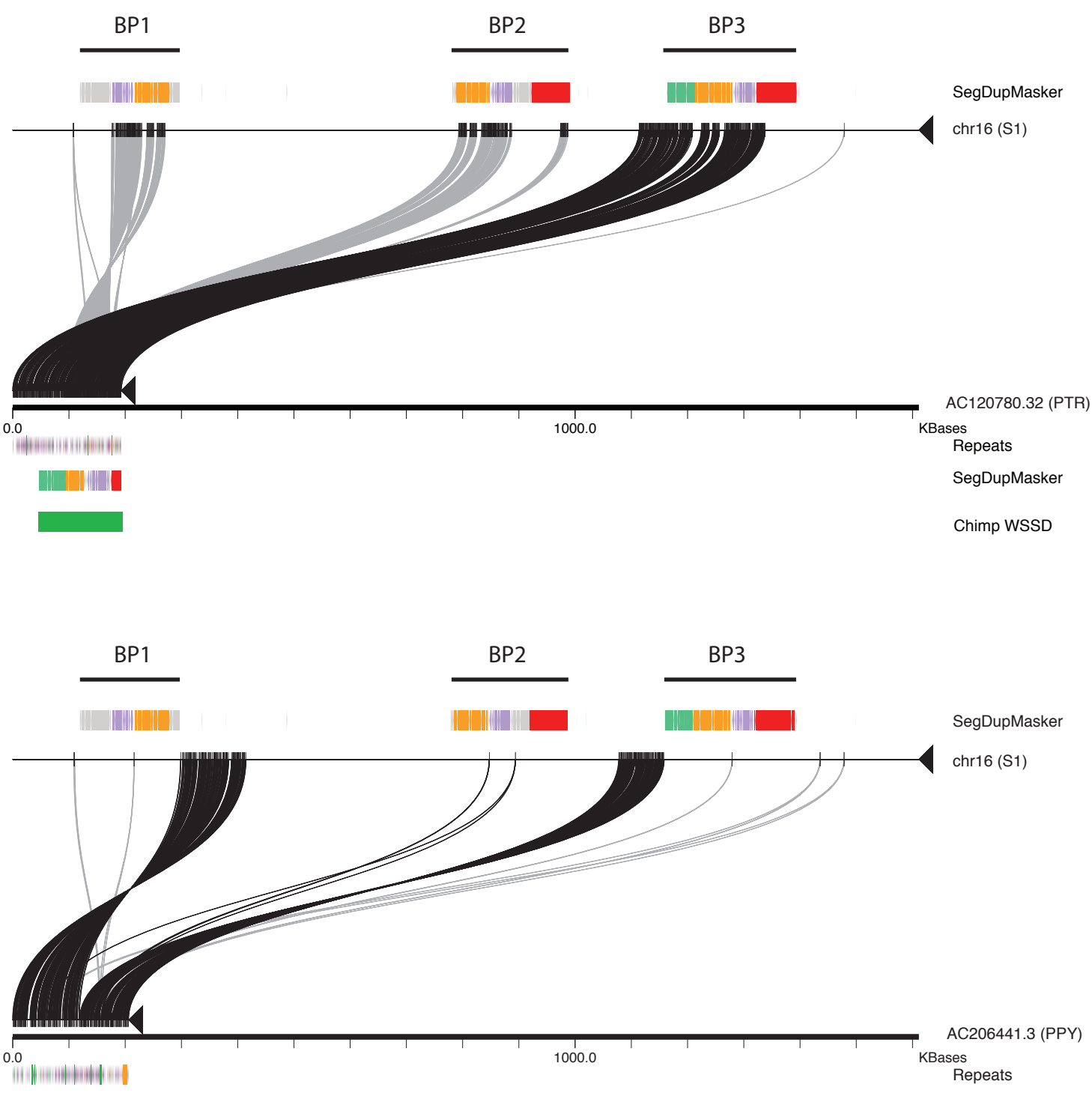


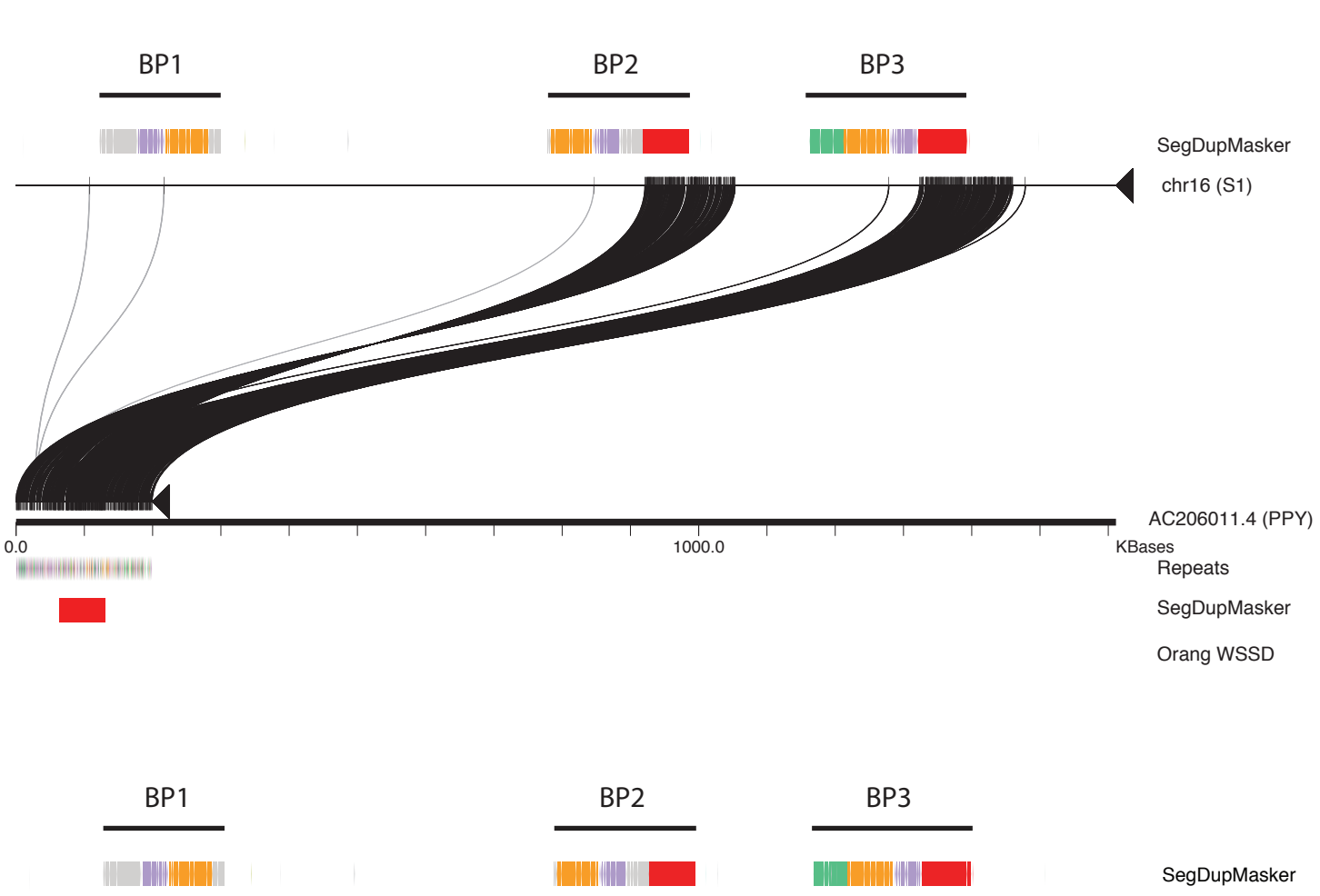
Figure S3 ArrayCGH data from 34 microdeletion samples are shown for the 16p12.1 region. Segmental duplications were annotated using SegDupMasker [5]. The location of duplications identified by read-depth is also depicted (Celera WSSD). Shown are the positions of copy-number polymorphisms (CNP2156 and CNP2157) from McCarroll *et al.* [4]. 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. The HapMap sample NA18956 with S1/S2 genotype was used as reference. Based on the observed mean $\log_2$ values, the genotype of each sample was determined.

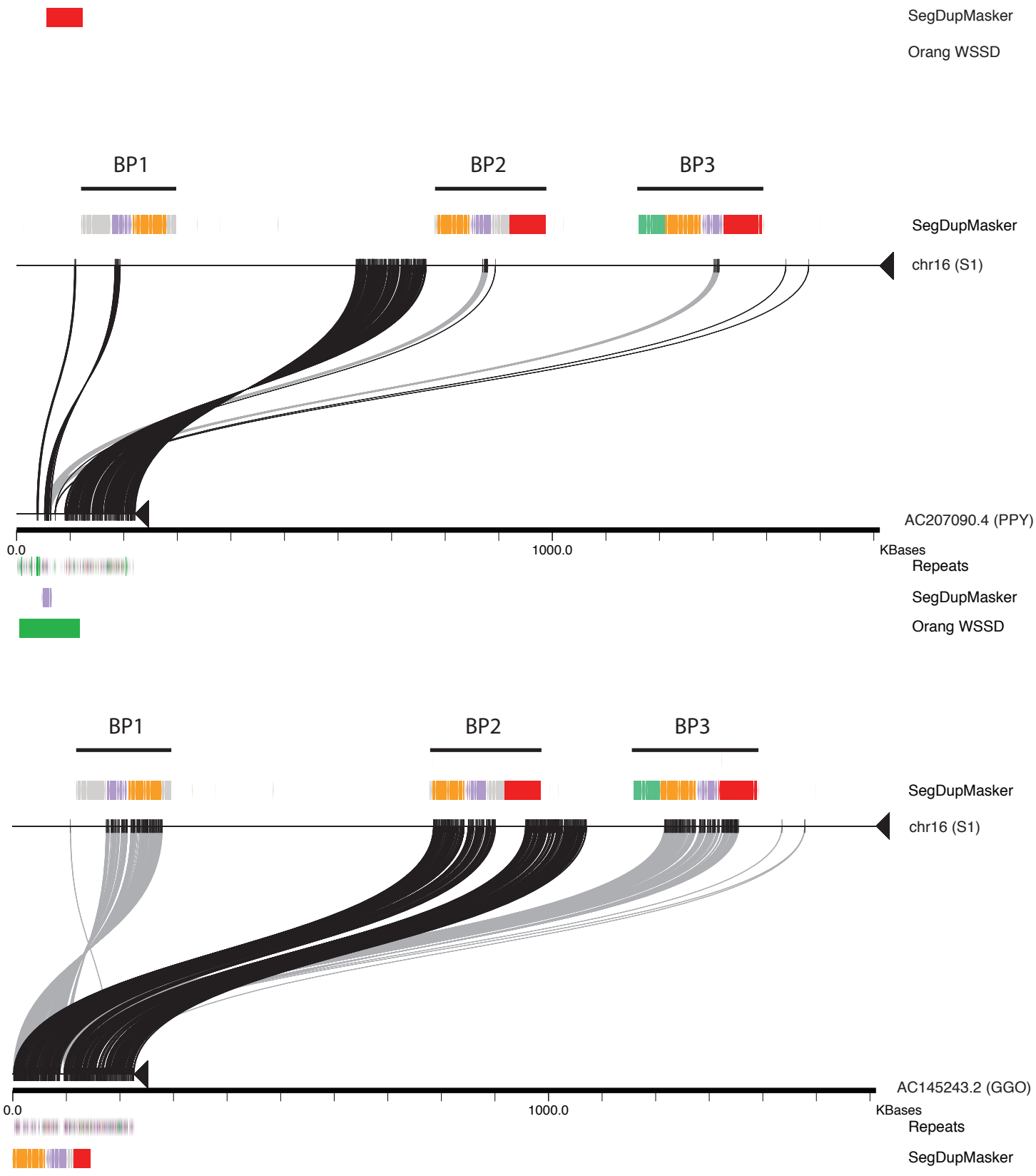
Nature Genetics: doi:10.1038/ng.643











Nature Genetics: doi:10.1038/ng.643

Figure S4 Sequence alignment of three chimpanzee, one gorilla and three orangutan BAC clones with the S1 reconstructed haplotype. Black lines connect matching segments between the clone and the S1 haplotype sequence while light-gray lines indicate lower-identity matches to distinct paralogs. Segmental duplications were annotated using SegDupMasker [5]. The location of duplications identified by read-depth is also depicted (WSSD).

Nature Genetics: doi:10.1038/ng.643