Supplementary Note

Evolutionary Toggling of the MAPT 17q21.31 Inversion Region

Michael C. Zody[1,2]*, Zhaoshi Jiang[3]*, Hon Chung Fung[4,5], Francesca Antonacci[3], LaDeana Hillier[6], Maria Francesca Cardone[7], Tina A. Graves[6], Jeffrey M. Kidd[3], Ze Cheng[3], Amr Abouelleil[1], Lin Chen[3], John Wallis[6], Jarret Glasscock[6], Richard K. Wilson[6], Amy Denise Reily[6], Jaime Duckworth[7], Mario Ventura[8], John Hardy[4]†, Wesley Warren[6]†, Evan E. Eichler[3]†

1.1) Human H2 haplotype sequence assembly

The sequence and orientation of the 17q21 region within the current genome assembly (build36) is consistent with the H1 haplotype, however, the underlying clones were derived from different donors. We outline below the steps taken in resolving/confirming the sequence of a single H1 haplotype, constructing a corresponding H2 minimal tiling path, and its ultimate sequencing to create an alternate haplotype for this region of the human genome. A critical aspect in this effort was the observation from Stefansson and colleagues [1] that the RPCI-11 BAC library was derived from a heterozygous donor. The availability of large-insert clones (~150 kbp) from a heterozygous donor was necessary to construct a complete tiling path across the region from both haplotypes (i.e. the high depth of the coverage of the RPCI-11 library and the large inserts allowed contiguity to be established in both haplotypes, despite the extensive duplication and copy-number variation associated within this region of the genome).

To completely encompass the region, we examined both the inverted region bounded by the large inverted segmental duplications (chr17:40,866,797 to 42,139,903 bp, identical coordinates on both NCBI build35 and 36) as well as 300 kb of sequence flanking either side of the inversion. We then sampled 1 kb of non-repeatmasked sequence approximately every 50 kb along this sequence and searched GenBank by BLAST, recovering a total of 62 finished and unfinished clones (not including non-human or non-genomic sequences). Within the inverted region, we identified 12 clones (11 finished and one draft) from RPCI-11 that contributed at least some unique sequence. In addition, we identified several clones from other libraries, most of which currently constitute the reference sequence.

Using a panel of 79 single nucleotide polymorphisms (SNPs) that differentiate the H1/H2 haplotype (HapMap) [2], we assigned 10 of these RPCI-11 clones to either H1 or H2. The proximal breakpoint clone AC091132 was assigned to H1 based on overlap with other assigned clones. AC019319 lies outside the distal breakpoint and remains unassigned, but is inferred to come from the chromosome carrying the H2 haplotype based on partial

overlap sequence data. Using this same SNP panel, we also determined that all other sequenced GenBank clones from other libraries (RPCI-5, RPCI-13, Cal Tech B & D, Genome Systems, WIBR-1 [Fosmid] and an unidentified PAC library) were of H1 origin, although not necessarily identical to the RPCI-11 H1 variant. Consequently, we decided to replace the H1 path within the genome assembly (build36) along with generating an H2-specific tiling path. Note that due to the sparse sampling of these other libraries it is impossible to determine whether the other BAC libraries are derived from H1 homozygous or H1/H2 heterozygous donors.

From this data, we were initially able to construct four sequence contigs consisting of six finished clones in H1 and three sequence contigs consisting of four finished clones and one draft clone on H2. We then proceeded to fill gaps and extend to the unique sequence outside the breakpoints using a method we term "haplotype walking". We aligned all existing BAC end sequences for RPCI-11 to all the sequenced clones in the region (in some cases including non-RPCI-11 clones where they overlapped portions of the H2 sequence that was not covered by RPCI-11 H1 clones). Due to the high sequence divergence between H1 and H2, for most BAC ends hitting both H1 and H2 (including within segmentally duplicated regions) we were able to find at least one position where the haplotypes or segmental duplications differed by at least 1 base and the end sequence matched one of the two haplotypes (mismatches due to BAC end sequencing error most frequently appeared as mismatches against both haplotypes). Because the segmental duplications of this region map to other regions on chromosome 17, some high quality BAC end sequences mapped to several other locations. In these cases, we examined all possible matches within the genome assembly as well as all sequenced BAC clones, selecting only those end-placements that had no better hit elsewhere in the genome. By this method, we were able to select clones of known haplotype that spanned gaps or extended sequence within either H1 or H2 haplotypes. Subsequent sequencing (100% identity of overlap of the complete clone sequences) confirmed haplotype contiguity for both H1 and H2.

The final paths for both H1 and H2 (Table 1) begin at 40,847,865 on NCBI build36 (coincidentally, both proximal clones share the same proximal restriction site). The H1 path consists of 11 non-redundant finished clones, contains no gaps and joins into AC019319 on the build36 path. The H2 path consists of 11 non-redundant finished clones and one gap currently spanned by several unfinished clones. It does not link out to AC019319, as end sequence probing of the RPCI-11 library has not revealed any clones that appear to span this region on the H2 haplotype. Sequence comparisons between H1 and H2, however, suggest that the distal breakpoint is captured. The remaining gap region contains a large inverted duplication unique to the H2 haplotype with >99.95% identity between the arms that has not yet been adequately resolved.

Table 1. Human H1 and H2 clone sequence assembly.

Assembly	Start	End	Status	Clone ID	Clone Start	Clone End	Orientation
H1	1	207611	F	AC091132.16	1	207611	-
H1	207612	361933	F	AC126544.5	18619	172940	+
H1	361934	367166	F	AC217774.1	34243	39475	-
H1	367167	515232	F	AC217771.1	24510	172575	-
H1	515233	728971	F	CR936218.6	1	213739	-
H1	728972	764751	F	AC217773.1	97752	133531	-
H1	764752	934781	F	AC005829.1	1	170030	-
H1	934782	1094428	F	AC217777.1	21173	180819	+
H1	1094429	1294848	F	AC138645.3	1	200420	+
H1	1294849	1446099	F	AC217780.1	1	151251	-
H2	1	176270	F	AC217778.1	1	176270	-
H2	176271	284196	F	AC217769.1	86955	194880	+
H2	284197	467289	F	AC138688.2	1	183093	-
H2	467290	591112	F	AC127032.8	63863	187685	+
H2	591113	664772	F	AC217772.1	89488	163147	+
H2	664773	815539	F	AC217779.1	30096	180862	+
H2	815540	876284	F	BX544879.6	116408	177152	+
H2	876285	1027216	F	AC217770.1	1	150932	-
H2	1027217	1042766	F	AC225613.2	154971	170520	-
H2	1042767	1197876	F	AC217768.1	1	155110	+
H2	1197877	1198876	N	1000	clone no captured		
H2	1198877	1378440	F	AC139677.4	1	179564	-
H2	1378441	1481050	F	AC217775.1	74484	177093	+

F=Finished clone; N=Gap. Gap size, gap type and capture state were indicated where there is a gap.

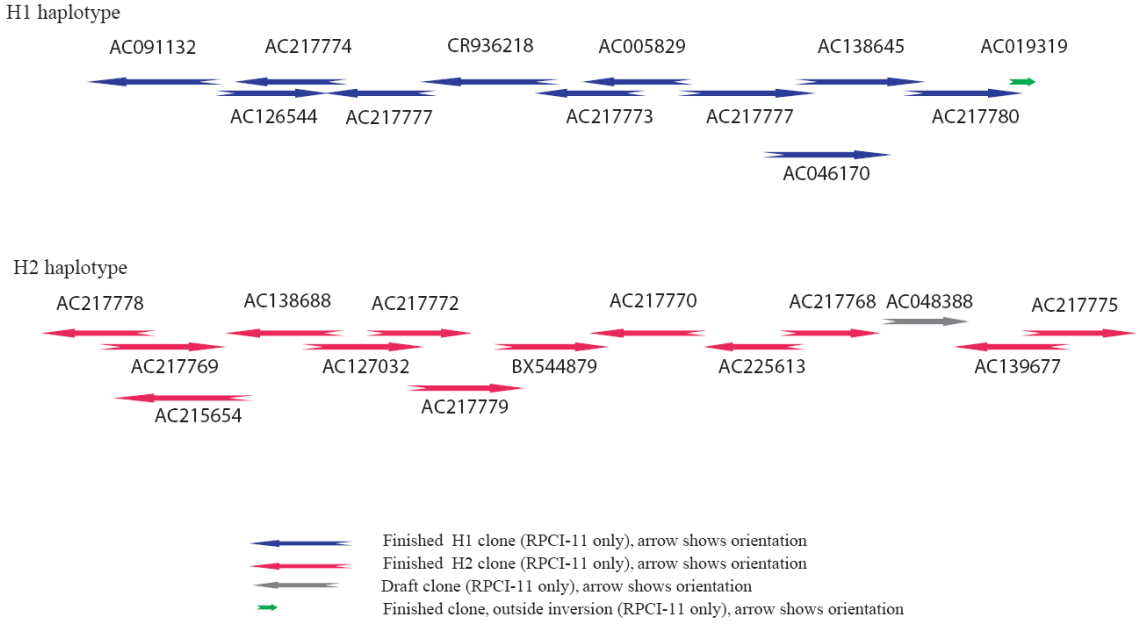


Figure 1: Human H1 and H2 RPCI-11 clone sequence assembly

1.2) Comparison of finished paths to those from Stefansson et al.

In comparing our final finished path to those from Stefansson et al., we note that our H1 path spans approximately the same distance (starting with the same proximal clone) and contains only 10 clones compared to the 17 on the Stefansson path. We incorporated three of the four finished clones on the previous path (798-G-7 [AC091132], 707-O-23 [AC126544], and 259-G-18 [AC005829]), with the fourth (219-F-9) rendered redundant by a new clone that was needed to close the adjacent sequence gap. We also incorporated two already finished BACs (669-E-14 [CR936218] and 995-C-19 [AC138645]) not on the Stefansson path, rendering the two clones they show in working draft status redundant (413-P-22 [AC036218] and 297-E-22 [AC138687]).

We then proceeded to close gaps using walking based on BAC end sequence overlaps, to guarantee both optimality of the tiling path and correct haplotype placement of clones. Of the 11 Stefansson clones on H1 with no sequence data, five had no existing BAC ends (329-D-18, 503-N-13, 258-H-10, 562-H-3, and 201-P-9), one had low quality ends (256-F-16), and five had highly repetitive ends that could not be placed uniquely (339-E-12, 244-K-17, 170-C-3, 141-H-9, and 133-E-17).

We used a similar process on H2, incorporating all four finished clones from the Stefansson path (300-H-14 [AC138688], 162-O-14 [AC127032], 769-P-22 [BX544879], and 1070-B-7 [AC139677]) and the working draft clone (374-N-3 [AC048388]; this is the final gap closer and remains unfinished as of this writing despite two new subclone libraries). Our final path contains 14 clones (with an additional redundant clone sequenced to confirm a join) compared to the 18 in the Stefansson path and is longer on the proximal end but shorter on the distal end. Of 13 clones with no sequence data in the Stefansson H2 path, one (57-A-24) was identified by end sequence and used, a second (207-I-10) was sequenced and assembled as a backup for the gap region but proved redundant, three had no end sequences (401-F-5, 549-H-12, and 573-G-23), one had only one end sequence (84-A-7), one was redundant to two finished clones (even on the Stefansson map, 559-K-6), four were repetitive (94-M-7, 450-G-10, 450-L-21, and 396-D-2), one was discarded for a more efficient spanner (100-C-5), and one (360-B-17) actually appears from end sequence to belong to H1, although this is based on only a single end. In the end, the construction of the H2 haplotype proved much more difficult and required more redundant sequencing; unlike H1, the distal and proximal repeat copies of H2 are so similar as to often be indistinguishable from a 500-800 bp of end read sequence."

1.3) Chimpanzee sequence and assembly of the MAPT region

Due to the less extensive duplication architecture in non-human primates, the development of a clone tiling path for chimpanzee was less complicated. We initially constructed a region-specific chimp assembly using a combination of the whole genome BAC fingerprint map and revised sequence assembly of the chimp genome (both located at our chimpanzee genome web page: http://www.genome.wustl.edu). Independent from fingerprints, the same order was confirmed from the mapping of end sequences of each clone to the human reference assembly, with the exception of the flanking duplications where discordant BAC end sequences suggested the presence of an inversion.

Our objective was to determine if the MAPT locus in the chimp could establish the most likely orientation of the H1/H2 haplotype in the last common ancestor of chimps and humans. Using BAC clone order from the chimp fingerprint map and BAC end sequence discordant pair analyses we were able to localize the putative points of inversion. Unfortunately, alignment of this chimp assembly sequence to the human genomic sequence (build36) did not allow for the unambiguous inversion orientation of the chimp genome assembly in this region. To confidently verify inversion orientation we selected and sequenced several candidate chimp BAC clones, using the 6X draft sequence assembly and fingerprint map coordinates that potentially span the predicted inversion and its breakpoints. At the inversion breakpoints, we required that there be 100% overlap between overlapping clones in order to ensure a single haplotype at each breakpoint. We sequenced the haplotypes corresponding to the inverted orientation—it was subsequently determined by FISH that chimpanzee Clint was heterozygous for the inversion.

We constructed a minimum tiling path across approximately 1.8 Mb from 15 BAC clones. The clone assembly order is outlined according to clone accession numbers (Figure 2). In addition to this BAC-based assembly, we established a primer pair set corresponding to known human H1/H2 SNPs. A subset of these chimp SNPs are characterized in Hardy et al. [3]. Despite our attempts to use these SNPs to differentiate chimp H1 and H2 clones, the high degree of sequence similarity in the duplicated regions and the coverage of the chimpanzee BAC library limited our ability to derive two distinct haplotype tiling-paths across the region.

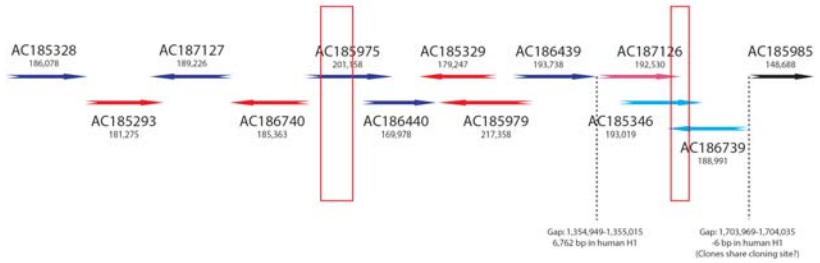


Figure 2: Chimpanzee MAPT locus clone assembly. Colors indicate clone chromosome of origin. In the first contig, clones alternate blue and red. In the second, light red and light blue. The third contig contains a single black clone. As yet, there is no ability to link the chromosomes of origin across the single gap (dotted line). The orientation is distal to proximal based on the alignment outside the inversion. The red boxes represent the posited location of the inversion breakpoints. Over these specific regions single haplotype continuity was maintained.

Table 2 Chimpanzee MAPT region clone assembly.

Assembly	Start	End	Status	Accession	Clone Start	Clone End	Orienation
PTR_MAPT	1	186078	F	AC185328	1	186078	+
PTR_MAPT	186079	361176	F	AC185293	6178	181275	+
PTR_MAPT	361177	519738	F	AC187127	1	158562	-
PTR_MAPT	519739	698518	F	AC186740	1	178780	-
PTR_MAPT	698519	887061	F	AC185975	12616	201158	+
PTR_MAPT	887062	986262	F	AC186440	70778	169978	+
PTR_MAPT	986263	1126775	F	AC185329	1	140513	-
PTR_MAPT	1126776	208850	F	AC185979	1	82075	-
PTR_MAPT	1208851	1354948	F	AC186439	47641	193738	+
PTR_MAPT	1354949	1361710	N	6762	clone	yes	
PTR_MAPT	1361711	1554240	F	AC187126	1	192530	+
PTR_MAPT	1554241	1604891	F	AC185346	142369	193019	+
PTR_MAPT	1604892	1710663	F	AC186739	1	105772	-
PTR_MAPT	1710664	1859345	F	AC185985	7	148688	+

F=Finished clone; N=Gap. Gap Size, gap type,and caputre state were indicated where there is a gap.

1.4) Orangutan sequence and assembly of the MAPT region

We developed two consensus sequences for the corresponding region in orangutan: one from whole genome shotgun sequence data and another from BAC clones sequenced to span the region. PCAP [4] software was used to assemble *Pongo pygmaeus abelii* whole genome shotgun data (donor=Susie, a female sumatran orangutan housed at the Gladys Porter Zoo, Brownsville, TX). FISH analysis showed that a cell line derived "Susie" was homozygous for the inversion. To determine chromosomal order and organization, the WGS assembly data were compared to the human genome utilizing BLASTZ [5] and Miropeats [6] and only "reciprocal best" alignments were retained. The ordered and oriented list of overlapping clones that form a minimal tiling path through the region (AGP) were generated from these alignments as described previously [7]. The primary inversion in the orangutan genome with respect to the human genome reference (H1) assembly is predicted and contained within a single PCAP supercontig (Supercontig339).

For the clone-based assembly, orangutan BAC clones were selected for sequencing and initial estimates of clone order were obtained based on BAC end sequence alignment to the corresponding region of the human genome (build36). After sequencing, the BAC

clone sequences were each aligned against all others requiring topological consistency to determine order, orientation and overlap in the orangutan genome (Fig. 3). The minimum tiling path across the 2.0 Mb of orangutan sequence consists of 14 clones (Table 3). A comparison between the clone-based assembly and sequence-based assembly found few differences (Fig. 3).

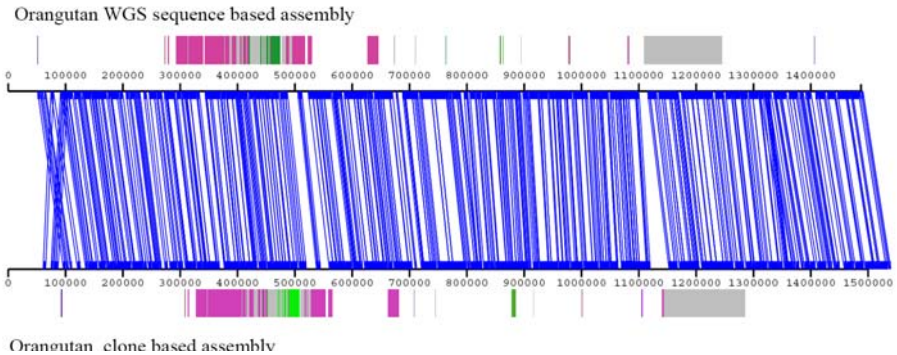


Figure 3: Comparison between sequence based assembly and clone based assembly of the MAPT region in orangutan. Parallel blue joining-lines show consistency in sequence structure and orientation (Miropeats –s 2000). The largest discrepancy was an 18 kbp segment that was missing from the WGSA assembly (located between 1.1 and 1.2Mbp of clone assembly). Regions that correspond to human duplicons were annotated as color-coded boxes; however, the majority of this sequence is not duplicated within the orangutan based on WSSD analysis.

Figure 4: Orangutan MAPT locus clone assembly. A minimum tiling path of BACs selected across 2 Mb of the orangutan sequence assembly in correspondence with the human MAPT locus (chr17:40.46-42.85Mb). The red boxes contain the inversion breakpoints as determined by alignment with human. Sequence overlaps between clones (AC206558/AC205859 and AC207097/AC216102/AC216058) are >99.9% identical.

Table 3 Orangutan MAPT region clone assembly.

Assembly	Start	End	Status	Clone	Clone Start	Clone End	Orientation
PPY_MAPT	1	86813	F	AC205775	1	86813	-
PPY_MAPT	86814	285708	F	AC206340	1	198895	+
PPY_MAPT	285709	285709	D	AC206276	51188	51188	+
PPY_MAPT	285710	497493	F	AC207288	1	211784	-
PPY_MAPT	497494	560358	D	AC206550	68688	131552	+
PPY_MAPT	560359	765828	F	AC206558	1	205470	+
PPY_MAPT	765829	879982	F	AC205859	1	114154	-
PPY_MAPT	879983	1060659	F	AC206353	20275	200951	+
PPY_MAPT	1060660	1154228	D	AC216075	83428	176996	-
PPY_MAPT	1154229	1361422	F	AC206444	1	207194	-
PPY_MAPT	1361423	1536151	D	AC207097	1	174729	-
PPY_MAPT	1536152	1646374	D	AC216102	30062	140284	-
PPY_MAPT	1646375	1834875	D	AC216058	1	188501	+
PPY_MAPT	1834876	2008578	D	AC216103	1	173703	-

F=Finished clone. D=high quality draft clones.

1.4) Non-human primate segmental duplication analysis

We analyzed duplication content using the WSSD method [8,9] for both the chimpanzee (Figure 5 a) and orangutan (Figure 5 b) 17q21.31 MAPT region. Regions of excess sequence read coverage (per 5kb window) are flagged in red and concatenated (light blue WSSD intervals) to identify recent duplications in each species. For comparison, human segmental duplications [10] are annotated (colored blocks) on the chimpanzee and orangutan sequences—although these are not necessarily duplicated within the non-human primate species. A comparison of the predicted duplications and detected duplication suggest that most of the duplication has occurred subsequent to the separation of the human/Great ape lineage from the Asian ape lineage (<12 mya).

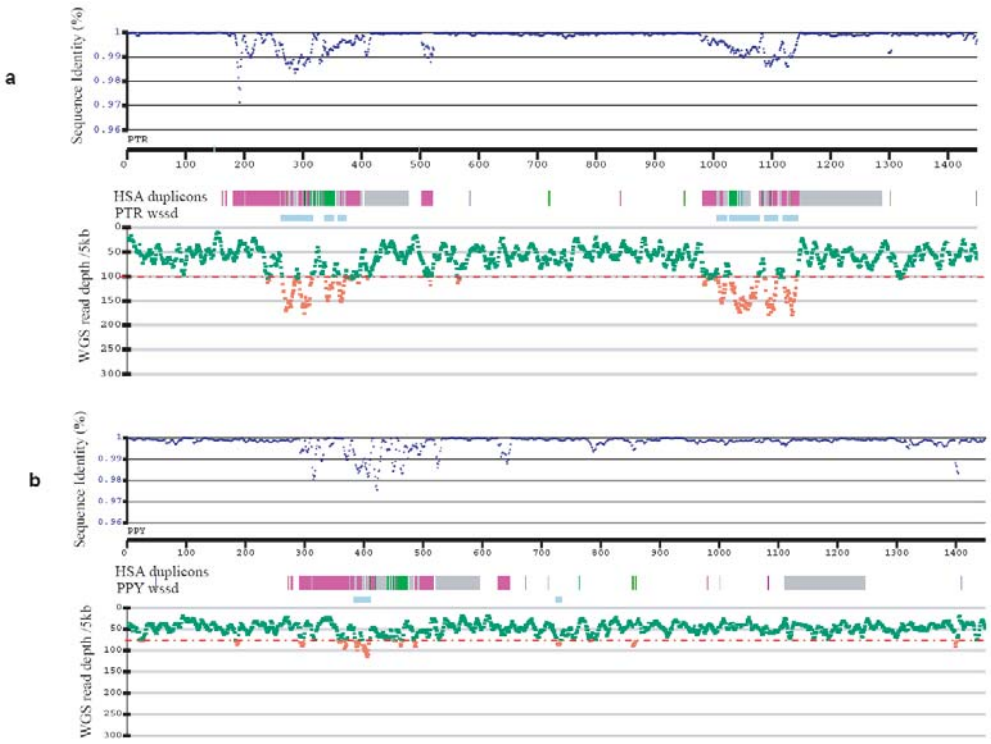


Figure 5: Non-human primate segmental duplication analysis.

2) Human haplotype analysis

Using the diagnostic SNP markers (rs1800547, rs9468), we partitioned the CEU HapMap haplotypes (Phase II HapMap release 21 phased-consensus available at http://hapmap.org) into 96 H1-chromosomes and 24 H2-chromosomes (after correcting for genotyping phasing errors, see below). We treated H1 and H2 haplotypes as separate populations in the analysis and limited our consideration to 611 SNP positions which could be uniquely mapped to non-duplicated portions of the sequenced H1 and H2 haplotypes. We identified 381 SNPs whose alleles are fixed differences between the H1 and H2 haplotypes. In addition, we identified a total of 207 SNPs that were fixed in one haplotype but polymorphic in the other. We assessed the likely ancestral state of each SNP through a comparison with the sequenced chimpanzee haplotype. For SNPs that are monomorphic among H2 haplotypes but polymorphic among the H1s, we found that the allele found in the H2 haplotypes matched the chimpanzee allele 90% of the time (150/166 considered positions). For SNPs that are monomorphic among H1 haplotypes but polymorphic among the H2s, the allele found in the H1 haplotypes matched the chimpanzee 60% of the time (17/28 considered positions). This suggests that the ancestral haplotype was H2-like.

This analysis of SNP ancestral state is based on a comparison against a single chimpanzee chromosome (the sequenced Clint haplotype). In order to assess possible biases introduced by this approach, we selected 10 SNPs that are polymorphic among CEU H1 chromosomes but are fixed among all CEU H2 chromosomes. Based on sequencing of PCR products, we genotyped seven chimpanzees (Clint plus six additional chimps, corresponding to a total of 14 chromosomes) at these SNP positions. The examined chimpanzees had a mixture of H1 and H2 orientations, but all of the chimpanzees are homozygous for the allele found among the H2 chromosomes.

Table 4. Assessing SNP ancestral state in multiple chimpanzees

			17q21 Orientation						
			H1/H1		H1/H2			H2/H2	
SNP ID	H1 Alleles	H2 Allele	Logan	PTR4	Clint	PTR13	Katie	PTR8	PTR12
rs417968	A/G	G	G/G	G/G	G/G	G/G	G/G	G/G	G/G
rs1724409	G/T	G	G/G	G/G	G/G	G/G	G/G	G/G	G/G
rs1635291	A/G	G	G/G	G/G	G/G	G/G	G/G	G/G	G/G
rs1635289	A/G	G	G/G	G/G	G/G	G/G	G/G	G/G	G/G
rs10451282	C/T	C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
rs1880756	C/T	C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
rs110402	A/G	G	G/G	G/G	G/G	G/G	G/G	G/G	G/G
rs242939	C/T	T	T/T	T/T	T/T	T/T	T/T	T/T	T/T
rs242943	C/T	C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
rs1158660	A/G	G	G/G	G/G	G/G	G/G	G/G	G/G	G/G

We genotyped 10 SNPs which are polymorphic among CEU H1 chromosomes but are fixed among H2 chromosomes in 7 chimpanzees. As indicated, the sampled chimpanzees containd both H1 and H2 orientations (see Figure 2)

While most SNPs represented fixed differences between H1 and H2, we did identify 23 SNPs that are polymorphic in both H1 and H2; in addition, we find 16 SNPs where H2 is fixed derived allele when compared to chimpanzee. For these, we reanalyzed the SNPs considering both CpG status, frequency and the quality of the underlying data as possible sources for the discrepancy. Of the 16 SNPs that are polymorphic among H1s but do not have an H2 allele matching PTR, 9/16 (56%) are at potential CpG sites, corresponding to likely recurrent mutation events. Four of the remaining seven positions are found on two or fewer H1 chromosomes and may be expected to have a higher genotype miscall rate because of their low frequency. Such variants may be positions where a derived allele became fixed before the split of the H1/H2 lineages, and subsequently the same position mutated again among the H1 chromosomes. Three of the remaining positions are found at a 5% frequency or greater among the H1 chromosomes and are without a clear explanation. Of the 23 SNPs polymorphic in both lineages, 12/23 (52%) are at potential CpGs. Of the remaining 11 positions, five are polymorphic because of a single H1 or H2 chromosome. The most "problematic" positions are four sites that are not CpG and have a minor allele frequency >10% among both H1s and H2s. Based on their frequency it is unlikely that these represent low-quality SNP genotypes. Rather, this minority of SNPs may represent gene flow between the H1 and H2 regions perhaps by gene conversion processes within the inversion loop.

Note: Initial analysis of the HapMap SNPs (internal to the inversion chr17:40974015-41926692 and excluding SNPs that could not be mapped to the H1 or H2 sequences or that mapped into duplicated sequences) indicated that 15% (95/611) of the SNPs were polymorphic among both the H1 and H2 haplotypes. Such a pattern would suggest a substantial degree of gene flow among H1 and H2 haplotypes, an unlikely result given the impact of the inversion on recombination between H1 and H2. In order to investigate this pattern more carefully, we visualized the distribution of the 611 SNPs across this interval that could be uniquely mapped onto the sequenced H1 and H2 haplotypes. Figure 6 summarizes the alleles present at each of these positions in the H1 and H2 sequences as well as the 24 H2 haplotypes inferred from the HapMap data. We observed clear stretches of H1-like haplotypes (compare yellow squares in Fig. 5) on an otherwise H2-background—accounting for the majority (72/95) of the SNPs that were polymorphic in both haplotypes.

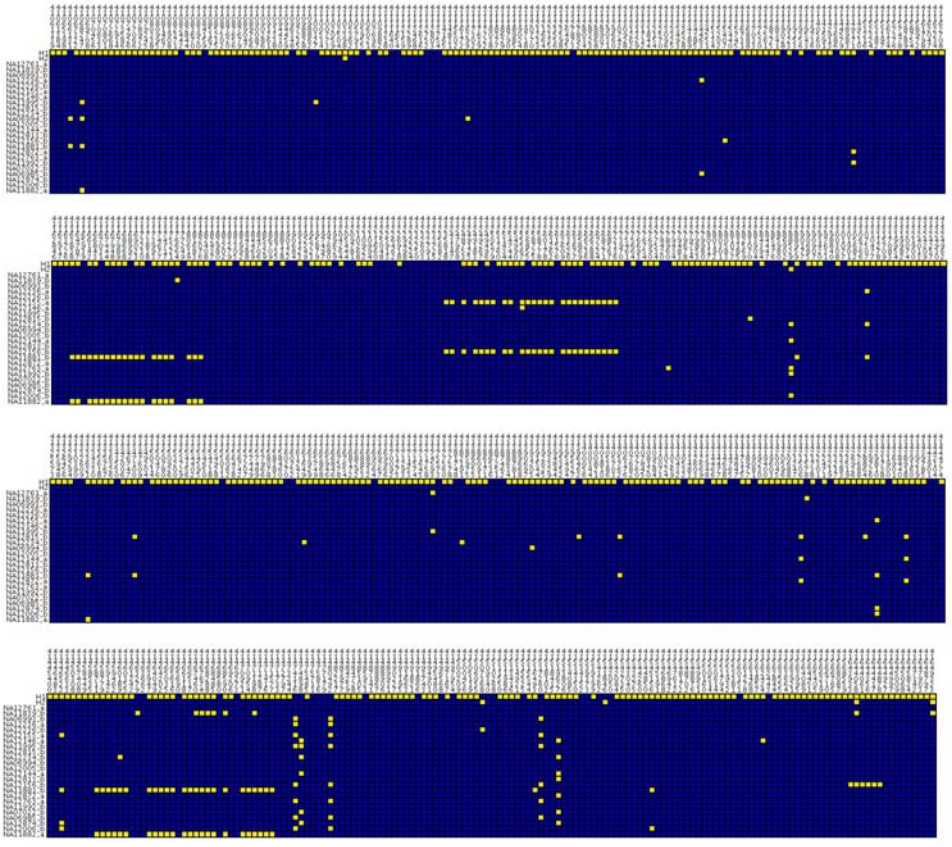


Figure 6: Observed H2 haplotypes. The alleles present at 611 HapMap SNPs are depicted for the sequenced H1 and H2 haplotypes and for all 24 inferred CEU H2 chromosomes within the HapMap sample set.

These stretches are derived from four H2 haplotypes inferred from four individuals. An examination of the other haplotype present in these four individuals (Fig. 6) indicates the presence of alternative alleles over these intervals that match the H2 haplotype. Blue square: major allele among 26 chromosomes depicted, yellow square: minor allele.

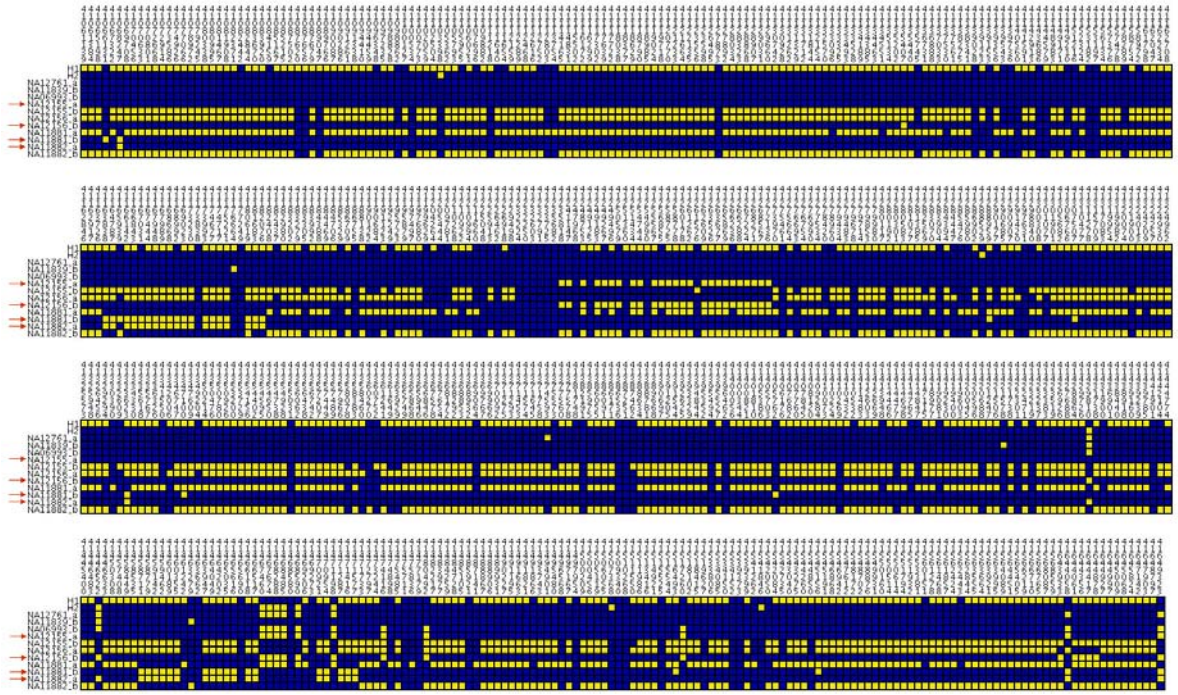


Figure 7: Identification of phasing errors. The four inferred H2 haplotypes containing unusual stretches of H1-like genotypes are depicted (NA11882_a, NA11881_b, NA12156_b, NA12155_a, highlighted by red arrows) along with the other haplotype from the same samples. The H1 and H2 sequenced haplotypes as well as three inferred H2 haplotypes are included for reference (top rows).

The four haplotypes represent two transmitted and two untransmitted chromosomes derived from the four parents of two CEU trios. Figure 8 indicates that for each of the four samples the two independent haplotypes show reciprocal phasing patterns (i.e. reciprocal H1-H2 hybrid haplotypes). Therefore, we conclude that the observed pattern is an artifact caused by phasing errors in the HapMap data. To fix the phase errors, we switched the haplotypes for four samples in the following intervals:

NA11881: 41163838-41182076; 41458711-41471577
NA11882: 41163838-41182076; 41458711-41471577
NA12155: 41235818-41272136
NA12156: 41235818-41272136; 41643933-41644878

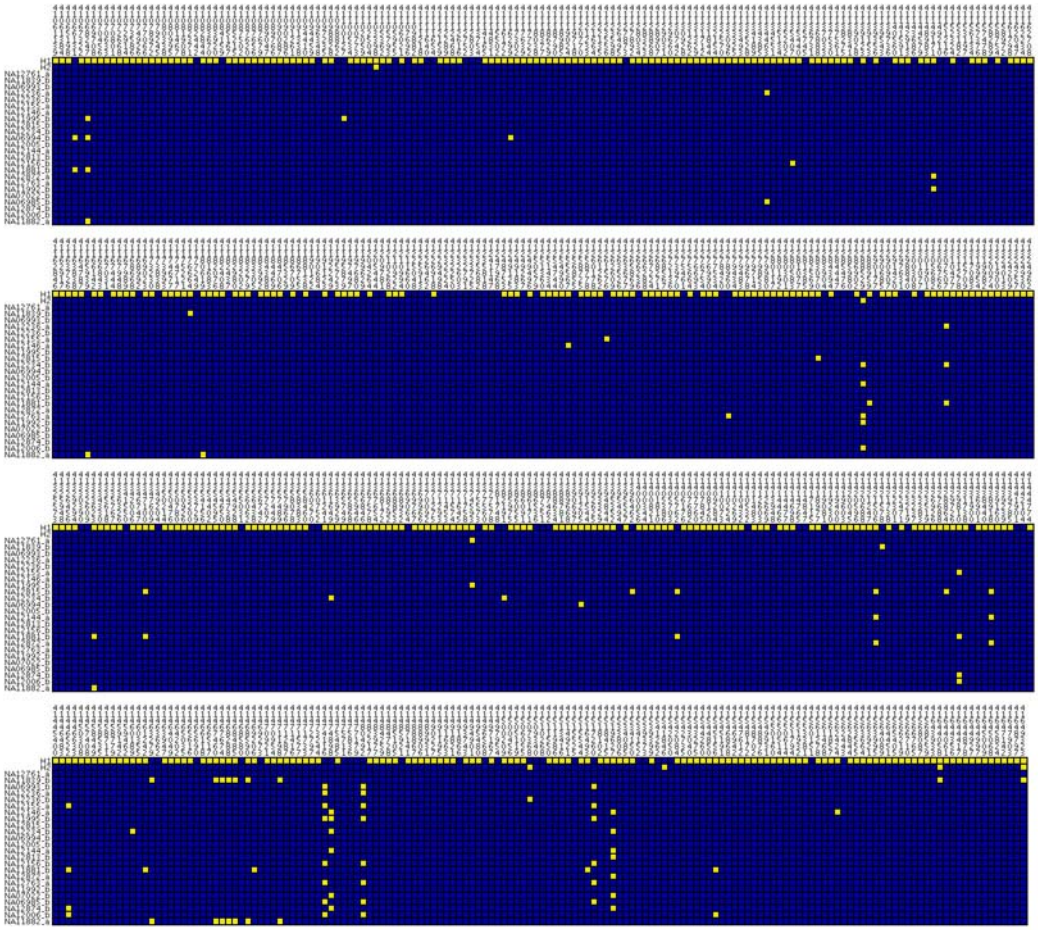


Figure 8: Sequence variation among H2 haplotypes. Variation among H2 haplotypes (depicted as in Fig. 5), following the correction of likely haplotype phasing errors.

3) Haplotype analysis by FISH and paired-end mapping

We developed a FISH assay to distinguish the orientation of the 17q21.31 region on metaphase chromosomes (Figure 9). Human genomic fosmid probes A and B map >1.5 Mb apart in the non-inverted state and appear as 2 distinct signals (red and green) on chromosomal metaphase spreads. In contrast, in the inverted state probes A and B map ~1 Mb apart and appear as a merged (red +green =yellow) signal. A reciprocal assay on the same samples using probes A and D (non-inverted=red + green; inverted=yellow) confirm the specificity of the assay. An analysis of 25 HapMap cell lines using this assay showed 100% correspondence between the H1/H2 haplotype and the non-inverted/inverted status (data not shown).

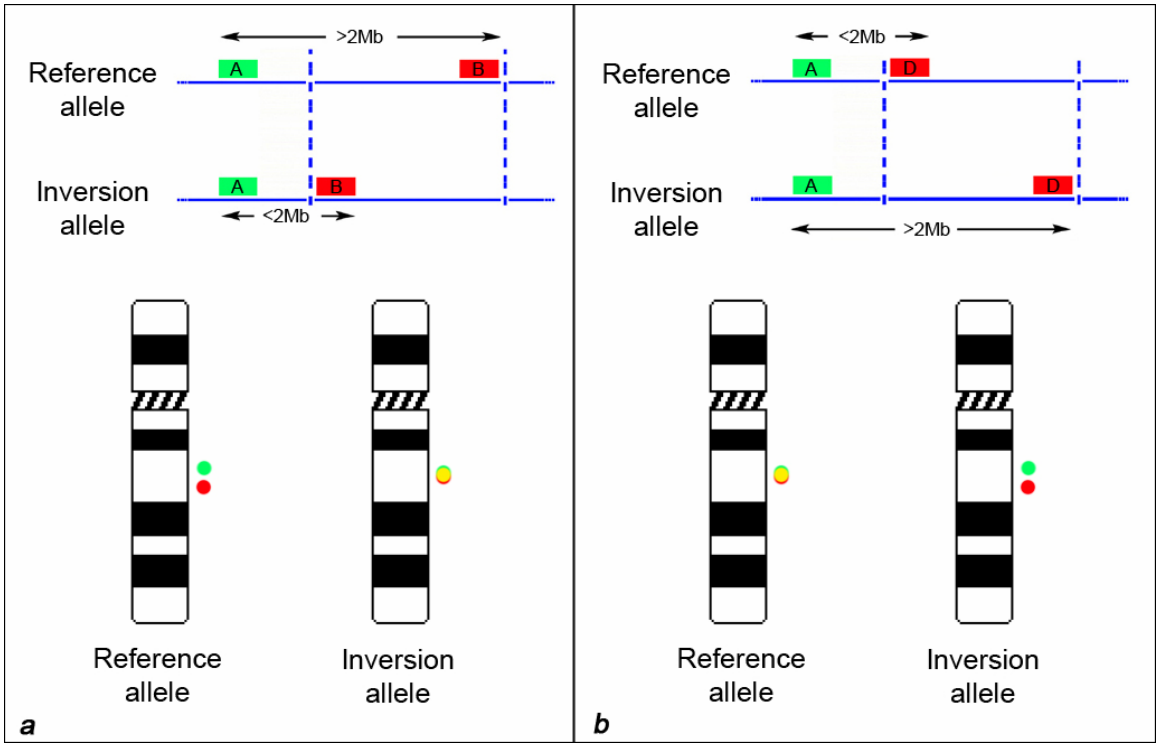


Figure 9: Chr17q21.31 reciprocal inversion FISH assay.

We applied this reciprocal FISH assay to other non-human primate metaphases, such as PPA (Pan paniscus); MMU (Macaca mulatta); MAR (Macaca arctoides); MFA (Macaca fascicularis). We found PPA2 is heterozygous for the inversion while PPA1 and all other non-human primates are homozygous for the inversion (Figure 10).

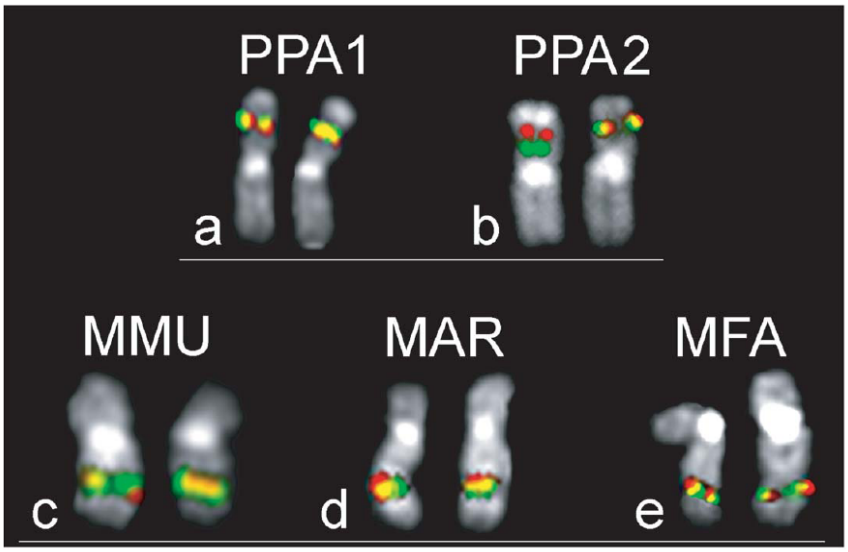


Figure 10: Primate FISH analysis of 17q21.31 inversion.

We then assessed haplotype diversity within the chimpanzee Clint (heterozygous for the inversion by FISH mapping) by mapping fosmid end-sequence pairs (ESPs) to unique portion of the BAC-based chimpanzee assembly. A total of 171 fosmid ESPs (top panel) showed perfect sequence identity to the unique region of the assembly (after quality rescoring, Phred Q≥30) and 53 ESPs (bottom panel) showed at least one high quality single basepair discrepancy and were assigned the alternate haplotype (at that position). We considered all ESPs with sequence identity ≥ 95% and only clones which mapped to a "best" location [11]. Based on the aligned sequence, we computed the sequence divergence between the two haplotypes as 0.297% (144 difference /48408 aligned basepairs) or 99.70% sequence identity. The distribution of sequence-identical and sequence different clones based on ESP placement is shown (Figure 11)

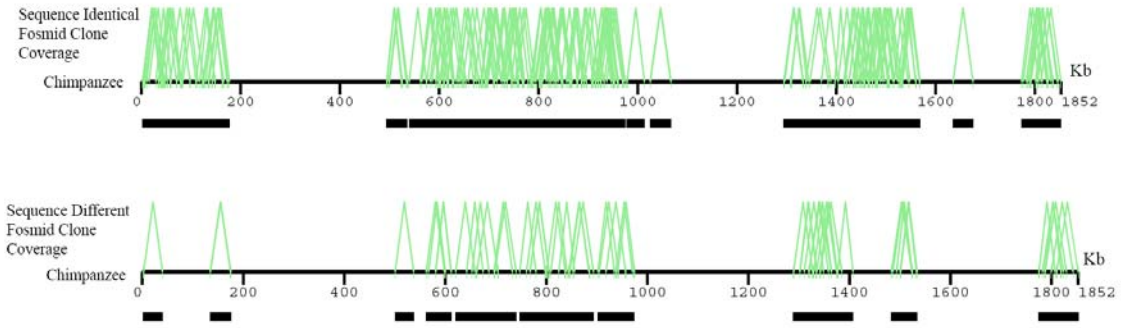
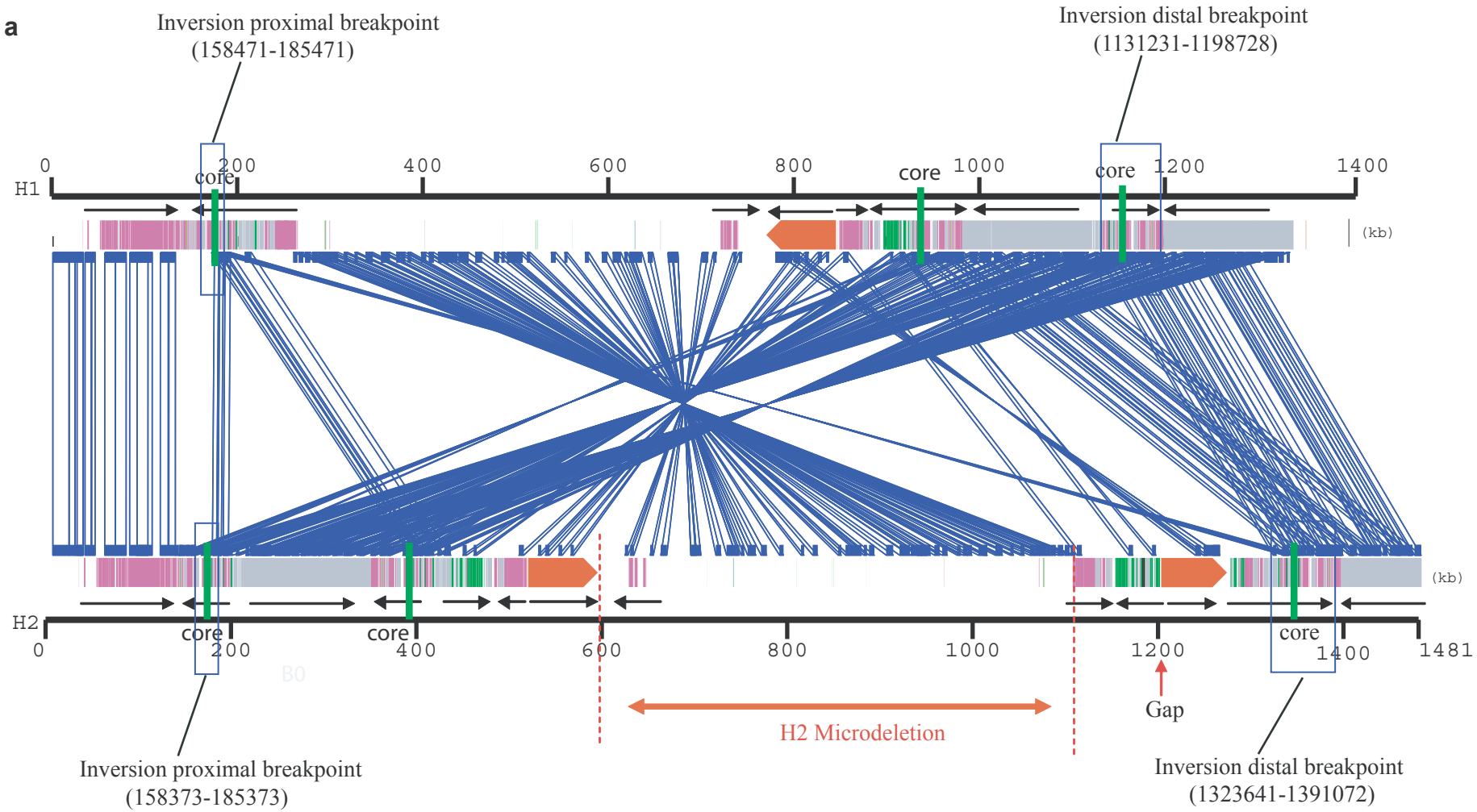


Figure 11: Chimpanzee haplotype analysis.

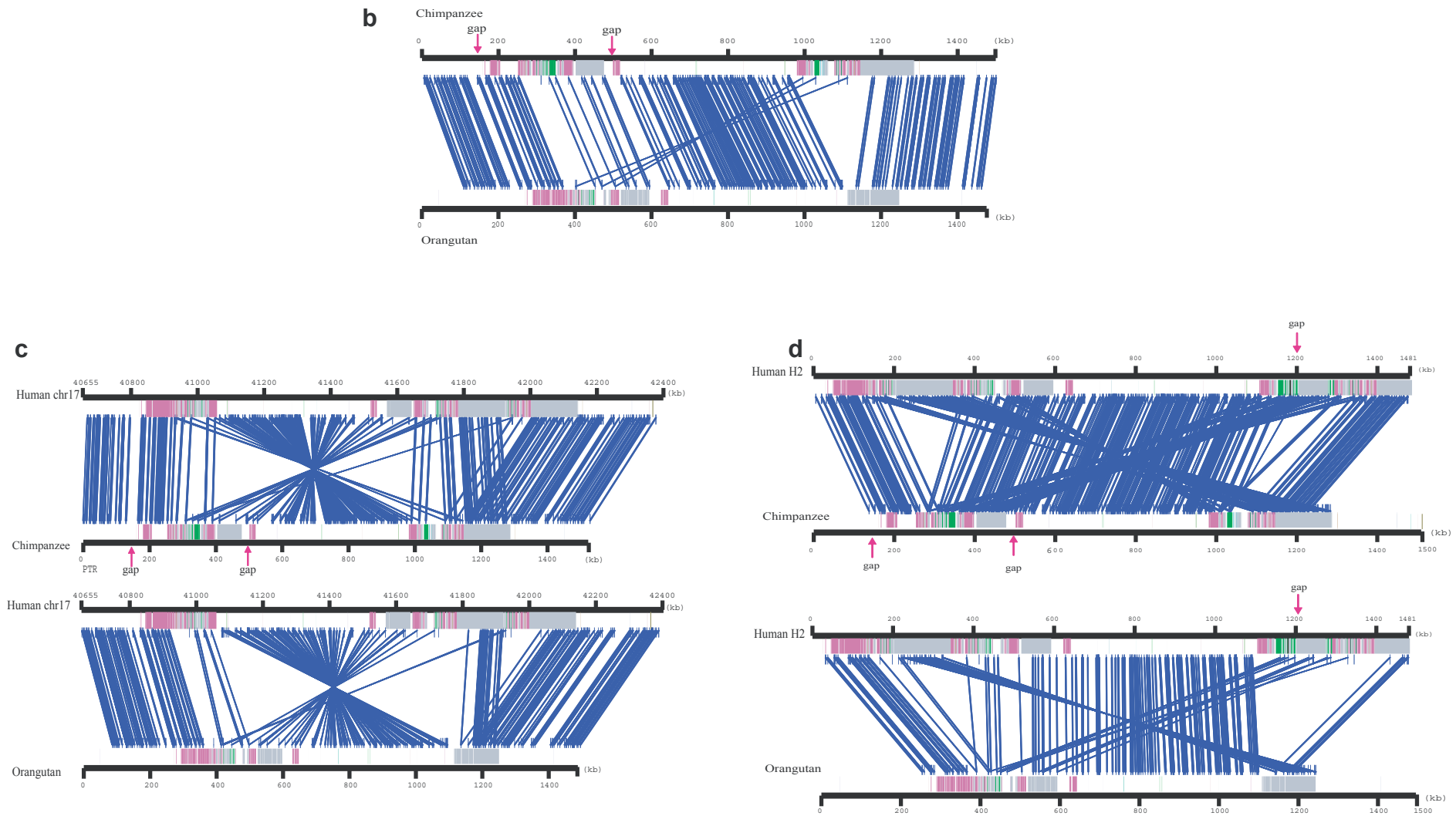
REFRENCES

1. Stefansson, H. et al. A common inversion under selection in Europeans. *Nat Genet* **37**, 129-37 (2005).
2. McCarroll, S.A. et al. Common deletion polymorphisms in the human genome. *Nat Genet* **38**, 86-92 (2006).
3. Hardy, J. et al. Evidence suggesting that Homo neanderthalensis contributed the H2 MAPT haplotype to Homo sapiens. *Biochem Soc Trans* **33**, 582-5 (2005).
4. Huang, X., Wang, J., Aluru, S., Yang, S.P. & Hillier, L. PCAP: a whole-genome assembly program. *Genome Res* **13**, 2164-70 (2003).
5. Schwartz, S. et al. Human-mouse alignments with BLASTZ. *Genome Res* **13**, 103-7 (2003).
6. Parsons, J.D. Miropeats: graphical DNA sequence comparisons. *Comput. Applic. Biosci* **11**, 615-619 (1995).
7. CSAC. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69-87 (2005).
8. Bailey, J.A. et al. Recent segmental duplications in the human genome. *Science* **297**, 1003-7 (2002).

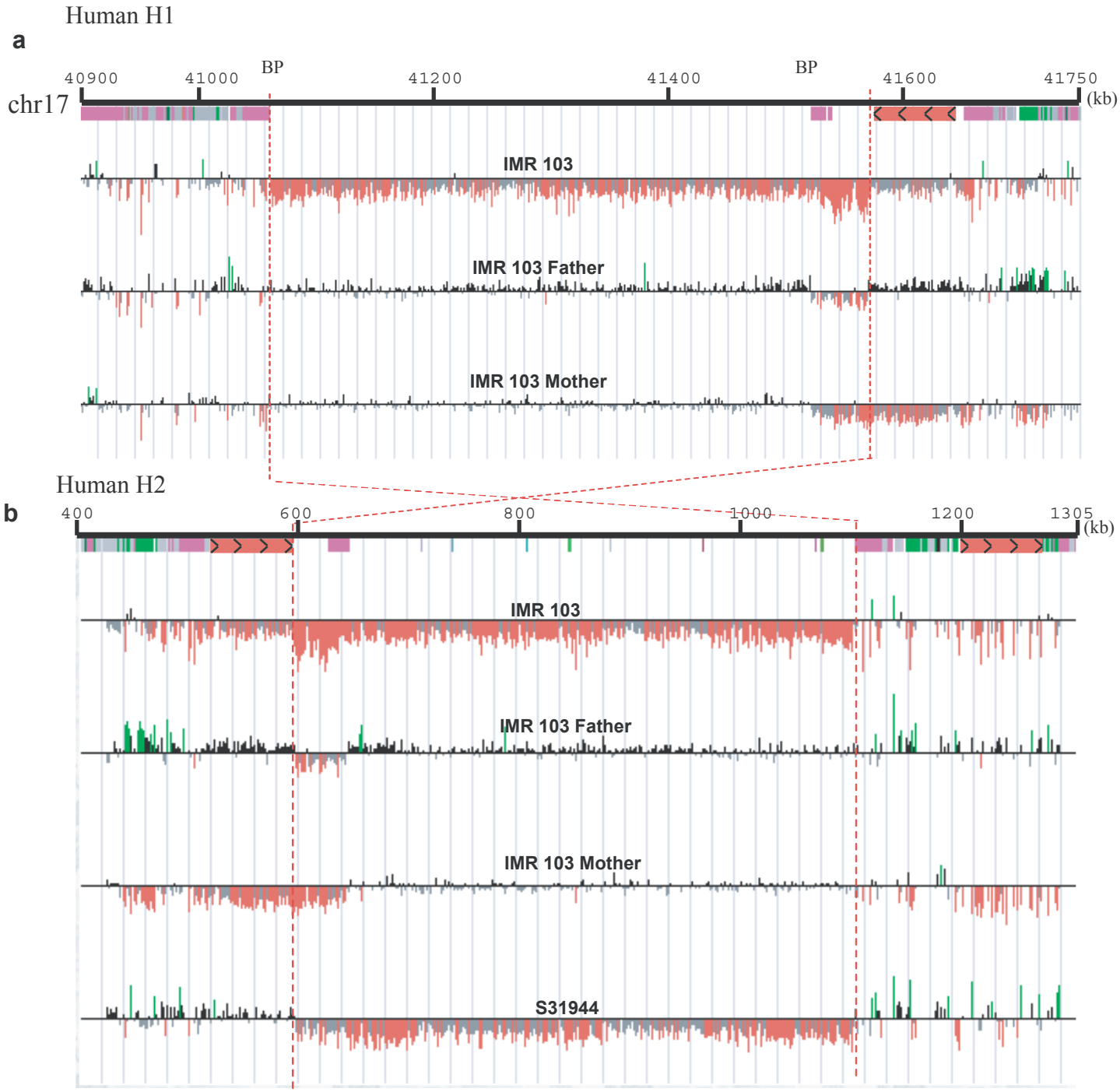
9. Cheng, Z. et al. A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature* **437**, 88-93 (2005).
10. Jiang, Z. et al. Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution. *Nat Genet* **39**, 1361-8 (2007).
11. Tuzun, E. et al. Fine-scale structural variation of the human genome. *Nat Genet* **37**, 727-32 (2005).



Supplementary Figure 1: Pairwise sequence comparison of chr17q21.31 region among primates.
Sequences (~1.5 Mbp) from human H1, human H2, chimpanzee and orangutan orthologous regions were compared using miropeats (Parsons, 1999) and two-way_mirror.pl (Bailey et al., unpublished). Human duplication structures (colored blocks) were annotated according to a database of curated duplicons (Jiang et al., 2007).
(a) Human H1 vs. H2 sequence comparison revealed that the core segmental duplications (green bars) mapped within the inversion breakpoints (blue box) between H1 and H2 haplotypes. In order to define the inversion breakpoints, we repeatmasked both H1 and H2 haplotype assemblies and compared the sequences using default settings (s -30). We examined each of the four breakpoints in turn by examining the furthest extent of directly oriented sequence between the two haplotypes and the furthest extent of inverted oriented sequence. For each comparison, we required that the majority of consecutive seed alignments be collinear. We defined the breakpoint intervals based on this intersection for each breakpoint. Once intervals had been identified, we further refined the breakpoint by constructing 4 pairwise alignments based on the intervals and requiring at least 99.5% sequence identity. The intervals, thus defined, are indicated by blue boxes. The inversion places larger blocks of sequence homology flanking the 17q21.31 microdeletion region, including H2-specific segmental duplications (red block arrows) in direct orientation (which are single copy in H1). High-density array CGH maps one of the microdeletion breakpoints (red broken lines) to the H2-specific segmental duplication.



Supplementary Figure 1: Pairwise sequence comparison of chr17q21.31 region among primates (continue).
(b) Chimpanzee (top) vs. Orangutan (bottom) sequence comparison compared as described above. We observe a 170 kb segmental duplication from proximal to distal duplication block within chimpanzee but not in orangutan creating a proximal and distal copy of the core. The data suggest that the event occurred in the common ancestor of chimpanzee and human (6-12 million years ago). (c) Human assemblies of non-inverted (H1) 17q21.31 region compared against BAC-based assembly of chimpanzee (PTR) and WGS-based assembly of orangutan. (d) BAC-based sequence assembly of inverted 17q21.31 region compared against BAC-based assembly of chimpanzee (PTR) and WGS-based assembly of orangutan shows ~200 kbp of higher identity homology between analogous segmental duplications (crisscross pattern).



Supplementary Figure 2: Chr17q21.31 microdeletion breakpoints.
We remapped the oligonucleotides used microarray comparative genomic hybridization experiments (Sharp et al, 2006) to both the H1 and H2 haplotypes. Array CGH data are shown for parents and a child (IMR103) with mental retardation against both an a) H1 and b) H2 sequence haplotype. The microdeletion was shown to have occurred on the H2 chromosome of the father of IMR103. The pair of "H2-specific" duplications (red color with internal arrows) are shown with respect to the microdeletion breakpoints (red dashed lines). Precise localization of breakpoints is confounded by segmental duplications and copy-number polymorphism, but an examination another de novo microdeletions consistently places one of the breakpoints in or within 1 kbp of the H2-specific segmental duplication (e.g. patient S31944). Although we have fairly clear evidence that the microdeletion breakpoint occurs within proximal breakpoint near the "H2-specific" segmental duplication, we have less mapping precision at the distal end due to the CNP that exists between H2 coordinates 1103477-1197879. So even though the distal breakpoint signal ends abruptly by the distal duplication block, due to the CNP of that region the distal breakpoint could still map closer to the internal H2-specific duplicon (i.e. a reduction in copy would not necessarily be expected to show a detectable difference when compared to an H1/H1 haplotype reference).

Supplementary Table 1: Pairwise sequence alignments (≥90 % identity, ≥1 kbp) among 4 assemblies

QNAME	QB	QE	QLEN	SNAME	SB	SE	per_SE	SE_sim	K2p	SE_kimura
H1	129729	150021	1405709	H1	981035	963600	0.9816757	NA	0.018573	0.001046
H1	129729	150021	1405709	H1	1198970	1181462	0.9823263	NA	0.0179037	0.0010243
H1	146032	150021	1405709	H1	150025	154014	0.9989975	0.000501	0.0010033	0.0005018
H1	150025	207448	1405709	H1	1185446	1123304	0.983541	NA	0.0166546	0.0005443
H1	150025	187318	1405709	H1	967593	925941	0.9803974	NA	0.0198783	0.0007393
H1	158243	159594	1405709	H1	232620	233843	0.9010152	0.0086864	0.1070585	0.0101912
H1	219599	226634	1405709	H1	894677	887698	0.9835141	0.0015313	0.0166946	0.0015704
H1	227453	261693	1405709	H1	886258	850934	0.9884582	0.0005826	0.011645	0.000593
H1	230482	233843	1405709	H1	958106	954620	0.9095907	0.0050118	0.0969868	0.0057809
H1	230482	233843	1405709	H1	1175961	1172475	0.9098962	0.0050041	0.096629	0.0057683
H1	879862	883229	1405709	H1	954620	958106	0.911084	0.0049667	0.0953039	0.0057189
H1	879862	883229	1405709	H1	1172475	1175961	0.9113886	0.004959	0.0949479	0.0057064
H1	925941	1123304	1405709	H1	1143518	1341017	0.9981153	NA	0.0018875	9.80E-05
H1	1	150021	1405709	H2	1	147152	0.99632	NA	0.0036902	0.0001588
H1	129729	150021	1405709	H2	347549	365072	0.9818818	NA	0.0183623	0.0010374
H1	129729	150021	1405709	H2	1391378	1373864	0.9820425	NA	0.0181954	0.0010328
H1	150025	207448	1405709	H2	143159	205298	0.9834128	NA	0.0167853	0.0005465
H1	150025	219717	1405709	H2	361079	435489	0.9841496	NA	0.0160331	0.0004847
H1	150025	205626	1405709	H2	1377857	1317537	0.9826352	NA	0.0175814	0.0005687
H1	212314	219747	1405709	H2	1186182	1175766	0.9865229	0.0013386	0.0136112	0.0013654
H1	219599	261693	1405709	H2	469149	513195	0.9873231	0.0005497	0.0127997	0.0005605
H1	219599	926076	1405709	H2	1146505	437775	0.9948377	NA	0.0051828	8.62E-05
H1	219599	247714	1405709	H2	1316783	1289117	0.9875109	0.0006691	0.0126098	0.0006822
H1	230482	233843	1405709	H2	370571	374065	0.9107852	0.0049826	0.0956032	0.0057345
H1	230482	233843	1405709	H2	152652	156148	0.9086465	0.005036	0.098019	0.005811
H1	230482	233843	1405709	H2	1368369	1364872	0.9099237	0.0050027	0.0966123	0.0057683
H1	769304	844580	1405709	H2	1273881	1198880	0.9955597	NA	0.0044555	0.0002448
H1	850934	886258	1405709	H2	1103477	1138618	0.99032	0.0005271	0.0097523	0.000535
H1	865792	886258	1405709	H2	1289117	1308890	0.9929406	0.0005967	0.0070983	0.0006033
H1	879862	883229	1405709	H2	374065	370571	0.9128846	0.0049218	0.0932312	0.0056493
H1	879862	883229	1405709	H2	156148	152652	0.911057	0.0049681	0.0953052	0.0057171
H1	879862	883229	1405709	H2	1364872	1368369	0.9123288	0.0049344	0.0938953	0.0056725
H1	887698	911684	1405709	H2	1139437	1163543	0.9865985	0.0007441	0.0135392	0.0007595
H1	887698	895435	1405709	H2	1309709	1317540	0.9831278	0.0014672	0.0170986	0.001507
H1	907113	911684	1405709	H2	1197879	1193288	0.9901445	0.0014619	0.0099302	0.0014842
H1	910905	926076	1405709	H2	1289120	1273938	0.9849156	0.0009914	0.015261	0.0010148
H1	916037	926076	1405709	H2	1163440	1173502	0.9902332	0.0009818	0.0098355	0.0009956
H1	916037	923242	1405709	H2	1193391	1186179	0.9895833	0.0011965	0.010495	0.0012146
H1	925941	1198970	1405709	H2	403035	129631	0.9978491	NA	0.0021544	8.90E-05
H1	925941	1071136	1405709	H2	1335915	1481052	0.997421	0.0001334	0.002584	0.0001339

H1	954620	958106	1405709	H2	484902	481538	0.9093407	0.0050165	0.0973116	0.0057933
H1	954620	958106	1405709	H2	1132239	1135600	0.9095907	0.0050118	0.0970019	0.0057828
H1	954620	958106	1405709	H2	1302509	1305872	0.9098962	0.0050041	0.096644	0.0057702
H1	1123304	1146564	1405709	H2	423242	400003	0.9978908	0.000301	0.0021123	0.0003019
H1	1124829	1288708	1405709	H2	1317451	1481052	0.9976384	NA	0.0023658	0.0001205
H1	1172475	1175961	1405709	H2	484902	481538	0.9096459	0.0050089	0.096969	0.0057827
H1	1172475	1175961	1405709	H2	1132239	1135600	0.9098962	0.0050041	0.0966592	0.0057721
H1	1172475	1175961	1405709	H2	1302509	1305872	0.9102016	0.0049965	0.0963014	0.0057596
H1	1200001	1341017	1405709	H2	346518	205298	0.9982524	0.0001113	0.00175	0.0001116
H1	1	27116	1405709	PPY	242371	269809	0.9618917	0.0011819	0.0392346	0.0012533
H1	28340	94329	1405709	PPY	270586	338746	0.9647576	0.0007357	0.0362236	0.0007775
H1	48679	50223	1405709	PPY	291385	292941	0.9485677	0.0056358	0.0534512	0.006091
H1	95460	150021	1405709	PPY	341139	389891	0.9644631	NA	0.0365232	0.0009151
H1	150025	161007	1405709	PPY	385872	399999	0.9589368	NA	0.0423496	0.0020216
H1	178527	193979	1405709	PPY	403650	416881	0.960122	NA	0.0411769	0.0018227
H1	194330	219717	1405709	PPY	417738	442977	0.9662602	0.0011416	0.0346294	0.001203
H1	219599	241814	1405709	PPY	476554	500383	0.9550969	0.0014119	0.046436	0.0015107
H1	230481	233852	1405709	PPY	395134	398623	0.9044858	0.0051345	0.1028433	0.0059678
H1	242896	259027	1405709	PPY	500537	517298	0.9524689	0.0017276	0.0492153	0.0018531
H1	269938	393250	1405709	PPY	1110628	988615	0.9659031	NA	0.0350294	0.0005573
H1	393057	417488	1405709	PPY	987380	962841	0.9666539	NA	0.0342258	0.0012355
H1	418796	524876	1405709	PPY	962753	853399	0.963311	0.0005828	0.0377592	0.0006175
H1	523579	585990	1405709	PPY	852374	788701	0.9643768	NA	0.0366245	0.0007976
H1	579487	598093	1405709	PPY	788496	769490	0.9668041	0.0013194	0.0340955	0.0013924
H1	599331	653468	1405709	PPY	768796	714977	0.9666287	0.0007891	0.0342661	0.0008322
H1	654057	758862	1405709	PPY	714223	607753	0.9680784	0.0005499	0.0327108	0.0005776
H1	696651	786621	1405709	PPY	669911	578769	0.9675895	0.0005982	0.0332297	0.000629
H1	787663	799693	1405709	PPY	577961	565887	0.9715227	0.001529	0.0291076	0.0015978
H1	800373	849598	1405709	PPY	565090	519425	0.9711416	NA	0.0295062	0.0008344
H1	854471	886258	1405709	PPY	517298	484236	0.9508487	0.0012608	0.0509674	0.0013565
H1	879853	883230	1405709	PPY	398623	395134	0.9056317	0.0051006	0.1015106	0.0059165
H1	887699	926076	1405709	PPY	483426	445272	0.9628095	0.000998	0.0382701	0.0010572
H1	925941	934731	1405709	PPY	410255	403650	0.9491292	NA	0.0530191	0.0029526
H1	953206	981031	1405709	PPY	400000	376680	0.9548497	NA	0.0467066	0.0014705
H1	991321	1009100	1405709	PPY	1112419	1130773	0.9630238	0.00142	0.0380001	0.0015002
H1	1010319	1123300	1405709	PPY	1131744	1245883	0.9709245	NA	0.0297354	0.0005278
H1	1123304	1136475	1405709	PPY	430770	417738	0.9683437	0.0015403	0.0324677	0.0016208
H1	1136826	1152307	1405709	PPY	416881	403650	0.9594265	NA	0.0419203	0.0018403
H1	1171061	1198966	1405709	PPY	400000	376680	0.9550944	NA	0.0464385	0.0014631
H1	1209243	1227021	1405709	PPY	1112419	1130773	0.9629105	0.0014221	0.0381177	0.0015025
H1	1228240	1373285	1405709	PPY	1131744	1277187	0.9718322	NA	0.0287915	0.0004593
H1	1373708	1405131	1405709	PPY	1277986	1309453	0.973362	0.0009117	0.0271933	0.0009503
H1	1	150021	1405709	PTR	132503	280180	0.9889582	NA	0.0111353	0.0002772

H1	129729	150021	1405709	PTR	1144414	1124389	0.9789284	0.0010173	0.0214039	0.0010498
H1	150025	186204	1405709	PTR	276184	314603	0.9819312	0.0007017	0.0183149	0.0007211
H1	150025	194310	1405709	PTR	1128352	1078907	0.9800813	NA	0.0202191	0.0006862
H1	158243	159594	1405709	PTR	368933	370157	0.9003378	0.0087055	0.1077732	0.0102097
H1	211222	219747	1405709	PTR	314519	323032	0.9841326	0.0013548	0.0160585	0.0013877
H1	219599	261693	1405709	PTR	355882	398285	0.9852015	0.0005942	0.0149638	0.0006076
H1	219599	393264	1405709	PTR	1024042	850371	0.9873945	0.000269	0.0127274	0.0002742
H1	230482	233843	1405709	PTR	1118905	1115425	0.9098411	0.005007	0.0967535	0.0057799
H1	393057	744538	1405709	PTR	849129	497776	0.9892911	NA	0.0107986	0.0001771
H1	751301	886258	1405709	PTR	497708	363754	0.9907694	NA	0.0092962	0.0002656
H1	850934	886258	1405709	PTR	981545	1016187	0.9889726	0.0005633	0.0111202	0.0005729
H1	879862	883229	1405709	PTR	1115425	1118905	0.9110299	0.0049696	0.095444	0.0057326
H1	887698	926076	1405709	PTR	362932	325293	0.9862257	0.0006022	0.0139223	0.0006152
H1	887698	910935	1405709	PTR	1017009	1040257	0.9871057	0.0007421	0.0130242	0.0007572
H1	925941	1123304	1405709	PTR	1085926	1286401	0.9894673	NA	0.0106166	0.0002343
H1	927052	981035	1405709	PTR	314603	261788	0.9794679	NA	0.0208515	0.0006648
H1	954620	958106	1405709	PTR	370157	366794	0.9117827	0.0049551	0.0944544	0.0056929
H1	954620	958106	1405709	PTR	1009808	1013165	0.9105617	0.0049859	0.0958809	0.0057431
H1	1136495	1405709	1405709	PTR	1078907	1350934	0.9899471	NA	0.0101296	0.0001957
H1	1144631	1198970	1405709	PTR	314603	261788	0.9791427	NA	0.0211872	0.0006679
H1	1172475	1175961	1405709	PTR	370157	366794	0.9120879	0.0049473	0.0941128	0.0056822
H1	1172475	1175961	1405709	PTR	1009808	1013165	0.9108669	0.0049782	0.0955239	0.0057306
H2	129631	205298	1481053	H2	347549	423242	0.9988895	NA	0.0011115	0.0001213
H2	129631	203770	1481053	H2	1391378	1317451	0.9975226	0.0001829	0.002482	0.0001836
H2	152652	156148	1481053	H2	481538	484902	0.9090076	0.0050255	0.0976711	0.005804
H2	152652	156148	1481053	H2	1135600	1132239	0.9092576	0.0050208	0.0973618	0.0057935
H2	152652	156148	1481053	H2	1305872	1302509	0.9095631	0.0050132	0.0970037	0.0057809
H2	257472	421714	1481053	H2	1481053	1317451	0.9985446	NA	0.001457	9.45E-05
H2	370571	374065	1481053	H2	481538	484902	0.911145	0.004972	0.0952586	0.0057276
H2	370571	374065	1481053	H2	1135600	1132239	0.9113963	0.0049671	0.0949485	0.0057169
H2	370571	374065	1481053	H2	1305872	1302509	0.9117018	0.0049594	0.0945915	0.0057043
H2	428107	447822	1481053	H2	1186182	1163440	0.9908462	0.0006792	0.0092177	0.0006887
H2	437714	452954	1481053	H2	1273877	1289120	0.9859785	0.000954	0.0141746	0.000975
H2	440620	447822	1481053	H2	1186179	1193391	0.9922233	0.0010352	0.0078234	0.0010476
H2	452165	513195	1481053	H2	1163543	1103477	0.9902065	0.0004047	0.009866	0.0004107
H2	452165	456744	1481053	H2	1193288	1197879	0.9925699	0.0012695	0.0074713	0.0012836
H2	468394	498328	1481053	H2	1317540	1289117	0.991574	0.0005439	0.0084787	0.0005507
H2	481538	484902	1481053	H2	1368369	1364872	0.9102838	0.0049921	0.0962372	0.0057576
H2	519560	594627	1481053	H2	1198880	1273881	0.9974903	NA	0.0025143	0.0001835
H2	1118360	1147260	1481053	H2	1289117	1317540	0.9936843	0.0004706	0.0063468	0.0004752
H2	1132239	1135600	1481053	H2	1364872	1368369	0.9105344	0.0049874	0.0959894	0.0057548
H2	1158952	1170649	1481053	H2	1197879	1186179	0.9990596	0.0002834	0.0009412	0.0002839
H2	1162742	1173563	1481053	H2	1289120	1273877	0.9870998	0.0010871	0.0130276	0.0011087

H2	1186179	1193391	1481053	H2	1276778	1283979	0.9849833	0.0014341	0.0151886	0.0014672
H2	1302509	1305872	1481053	H2	1364872	1368369	0.9108397	0.0049797	0.095616	0.0057402
H2	1	27119	1481053	PPY	242371	269809	0.9618946	0.0011818	0.0392315	0.0012532
H2	28343	94226	1481053	PPY	270586	338746	0.9645905	0.0007374	0.0363999	0.0007795
H2	48647	50193	1481053	PPY	291385	292941	0.9472656	0.0057028	0.0548563	0.0061753
H2	95357	157562	1481053	PPY	341139	399999	0.9618711	NA	0.0392511	0.00086
H2	176315	191792	1481053	PPY	403650	416881	0.9599024	NA	0.0414089	0.0018277
H2	192143	205298	1481053	PPY	417738	430770	0.967722	0.0015549	0.0331087	0.0016365
H2	205302	318301	1481053	PPY	1245883	1131744	0.9707526	NA	0.0299152	0.0005294
H2	319524	337317	1481053	PPY	1130768	1112419	0.9627533	0.0014247	0.0382831	0.0015056
H2	347553	375479	1481053	PPY	376680	399999	0.9547336	NA	0.0468345	0.0014697
H2	394249	397671	1481053	PPY	403650	407056	0.9571133	0.0034843	0.0444301	0.0037419
H2	400003	409727	1481053	PPY	407223	416881	0.961667	0.0019622	0.0395154	0.0020861
H2	410078	509656	1481053	PPY	417738	517298	0.9601617	0.0006397	0.0410552	0.0006796
H2	481537	484911	1481053	PPY	395134	398623	0.9057352	0.0051035	0.1013548	0.0059144
H2	514533	576579	1481053	PPY	519425	577961	0.9707427	NA	0.0299186	0.000745
H2	577621	669291	1481053	PPY	578769	669911	0.9677103	0.0005972	0.0331038	0.0006279
H2	655436	711717	1481053	PPY	656418	714133	0.969311	0.0007363	0.0314167	0.0007718
H2	712392	766254	1481053	PPY	714977	768796	0.9669806	0.0007866	0.0338974	0.0008292
H2	767496	786090	1481053	PPY	769494	788496	0.9658653	0.0013376	0.0350961	0.0014146
H2	779566	841910	1481053	PPY	788701	852374	0.9650727	NA	0.0358865	0.0007892
H2	840613	926695	1481053	PPY	853399	942648	0.9629533	0.0006508	0.0381403	0.0006901
H2	842430	946620	1481053	PPY	856807	962753	0.9630982	0.0005898	0.037987	0.0006253
H2	947928	972416	1481053	PPY	962841	987380	0.9665317	0.0011744	0.034369	0.0012389
H2	972223	1095235	1481053	PPY	988615	1110628	0.9662128	NA	0.0347046	0.0005546
H2	1107023	1123171	1481053	PPY	517298	500537	0.951512	0.0017434	0.050229	0.0018719
H2	1124254	1163533	1481053	PPY	500383	459761	0.9591985	0.0010253	0.0420747	0.0010908
H2	1132230	1135601	1481053	PPY	398623	395134	0.9060403	0.0050961	0.1010402	0.0059076
H2	1163440	1186181	1481053	PPY	455351	435638	0.9638186	0.0013562	0.0371922	0.0014335
H2	1186179	1193391	1481053	PPY	448127	455351	0.9623648	0.0022553	0.0387401	0.0023906
H2	1193298	1197879	1481053	PPY	459761	464573	0.9687361	0.0025914	0.0320338	0.0027214
H2	1199592	1255860	1481053	PPY	522000	577961	0.9704927	NA	0.030178	0.0007627
H2	1256902	1273881	1481053	PPY	578769	596514	0.9655931	0.0014051	0.0353287	0.0014819
H2	1273877	1289120	1481053	PPY	445211	460525	0.9617852	0.0015923	0.0393591	0.0016898
H2	1289117	1291076	1481053	PPY	505786	503791	0.9627551	0.0042772	0.0382763	0.0045189
H2	1291610	1317540	1481053	PPY	502937	475791	0.9536172	0.0013374	0.0480056	0.0014334
H2	1302500	1305873	1481053	PPY	398623	395134	0.9063453	0.0050887	0.100696	0.005897
H2	1317537	1328870	1481053	PPY	428921	417738	0.9672381	0.0016865	0.0336086	0.0017754
H2	1329221	1344702	1481053	PPY	416881	403650	0.95975	NA	0.0415767	0.001832
H2	1363466	1391374	1481053	PPY	399991	376680	0.9546545	NA	0.046912	0.0014712
H2	1401594	1419391	1481053	PPY	1112419	1130773	0.9630426	0.0014193	0.0379792	0.0014993
H2	1420610	1481053	1481053	PPY	1131744	1193032	0.9691186	0.0007116	0.0316272	0.0007466
H2	1	183999	1481053	PTR	132503	314603	0.9868346	NA	0.013298	0.0002761

H2	129631	192123	1481053	PTR	1144414	1078907	0.9828339	NA	0.0173866	0.0005358
H2	152652	156148	1481053	PTR	366794	370157	0.9108397	0.0049797	0.0955115	0.0057269
H2	152652	156148	1481053	PTR	1013165	1009808	0.9096183	0.0050103	0.0969253	0.0057752
H2	205298	410058	1481053	PTR	1286401	1078907	0.9891775	NA	0.0109111	0.0002332
H2	347549	401934	1481053	PTR	261788	314603	0.9790659	NA	0.0212662	0.0006691
H2	370571	374065	1481053	PTR	366794	370157	0.9129771	0.0049254	0.093105	0.00565
H2	370571	374065	1481053	PTR	1013165	1009808	0.9117557	0.0049565	0.0945146	0.0056985
H2	427014	612632	1481053	PTR	314519	497708	0.9891334	NA	0.0109566	0.0002467
H2	452924	513195	1481053	PTR	1040257	981545	0.9874117	0.0004617	0.0127074	0.0004705
H2	481538	484902	1481053	PTR	1118905	1115425	0.9093407	0.0050165	0.0973428	0.0057973
H2	619419	972416	1481053	PTR	497776	849129	0.9892877	NA	0.0108018	0.0001772
H2	972209	1162772	1481053	PTR	850371	1040257	0.9873555	0.000257	0.0127674	0.000262
H2	1103477	1186182	1481053	PTR	398285	315609	0.9856409	0.0004243	0.0145156	0.0004336
H2	1132239	1135600	1481053	PTR	1115425	1118905	0.9095907	0.0050118	0.0971125	0.0057967
H2	1186179	1197879	1481053	PTR	328142	343521	0.986637	0.0010627	0.0135007	0.0010848
H2	1194059	1197879	1481053	PTR	1040257	1036434	0.9887257	0.0017096	0.0113764	0.0017408
H2	1198880	1273881	1481053	PTR	404658	479773	0.9914605	NA	0.0085943	0.0003407
H2	1273877	1289120	1481053	PTR	325232	339740	0.9822899	0.001097	0.0179521	0.0011273
H2	1289117	1317540	1481053	PTR	383423	355147	0.9858885	0.0007032	0.0142654	0.0007187
H2	1289117	1317540	1481053	PTR	996420	1024778	0.9865225	0.0006876	0.0136153	0.0007018
H2	1302509	1305872	1481053	PTR	1115425	1118905	0.9098962	0.0050041	0.0967375	0.005782
H2	1328890	1481053	1481053	PTR	1078907	1234192	0.9880938	NA	0.0120128	0.0002839
H2	1337017	1391378	1481053	PTR	314603	261788	0.9791056	NA	0.0212226	0.0006683
H2	1364872	1368369	1481053	PTR	370157	366794	0.9121147	0.0049459	0.0940966	0.0056821
H2	1364872	1368369	1481053	PTR	1009808	1013165	0.9108941	0.0049768	0.0955226	0.0057324
PPY	199253	201421	1450000	PPY	201686	203878	0.9935395	0.0017211	0.0064903	0.001737
PPY	218814	221070	1450000	PPY	221081	223353	0.9924612	0.0018215	0.0075798	0.0018414
PPY	781566	788500	1450000	PPY	788701	795440	0.9900062	0.001243	0.0100697	0.001262
PPY	851072	852374	1450000	PPY	853399	854701	0.9923195	0.0024194	0.0077261	0.0024483
PPY	1010628	1012303	1450000	PPY	1012377	1014051	0.9928358	0.0020607	0.0072119	0.0020883
PPY	103703	131823	1450000	PTR	1	27203	0.9699659	0.0010557	0.0307334	0.0011058
PPY	132555	201357	1450000	PTR	28121	95958	0.9700295	0.0006597	0.0306969	0.0006923
PPY	201691	218321	1450000	PTR	93836	110265	0.9693167	0.0013661	0.0314284	0.0014336
PPY	218814	221203	1450000	PTR	111231	114118	0.9636671	0.003846	0.0374463	0.0040872
PPY	221081	269809	1450000	PTR	111231	159659	0.9616596	0.0008919	0.0394808	0.0009461
PPY	270586	338746	1450000	PTR	160885	226661	0.9649428	0.0007344	0.0360241	0.0007757
PPY	291385	292941	1450000	PTR	181204	182740	0.9509804	0.0055198	0.0508403	0.005941
PPY	341139	400000	1450000	PTR	227789	287175	0.9641067	NA	0.0368966	0.0008569
PPY	376680	399999	1450000	PTR	1144410	1114034	0.9552048	NA	0.0463342	0.0014703
PPY	395134	398623	1450000	PTR	366793	370166	0.9066789	0.0050798	0.100319	0.0058845
PPY	395134	398623	1450000	PTR	1013166	1009799	0.9041514	0.0051433	0.1032538	0.0059842
PPY	403650	409165	1450000	PTR	306982	314603	0.9511568	NA	0.0508351	0.0031663
PPY	403650	416881	1450000	PTR	1094702	1079238	0.9598902	NA	0.0413901	0.0018254

PPY	434537	474687	1450000	PTR	314519	353259	0.9622813	0.0009847	0.0388348	0.0010442
PPY	459662	517298	1450000	PTR	338843	394742	0.955081	0.0009036	0.0464541	0.000967
PPY	460495	517298	1450000	PTR	1040257	985090	0.9559691	0.0009024	0.0454952	0.0009639
PPY	519425	577961	1450000	PTR	399621	461679	0.9710717	NA	0.0295742	0.0007406
PPY	578769	615335	1450000	PTR	462722	497708	0.9666705	0.0009646	0.0342089	0.0010165
PPY	621805	714223	1450000	PTR	497776	587896	0.9685643	0.0005876	0.0322011	0.0006167
PPY	714977	763543	1450000	PTR	588482	637892	0.9655646	0.0008384	0.0353944	0.000886
PPY	717996	768796	1450000	PTR	592464	642125	0.9650628	0.00083	0.0359209	0.0008777
PPY	769486	788496	1450000	PTR	643357	662010	0.9658611	0.0013367	0.0350985	0.0014134
PPY	788701	825374	1450000	PTR	655473	691736	0.9645258	NA	0.0364796	0.00104
PPY	826170	852374	1450000	PTR	692334	717597	0.9672581	0.00113	0.0335935	0.00119
PPY	853399	893251	1450000	PTR	716300	753908	0.9623322	0.0009915	0.0387897	0.0010519
PPY	894076	962753	1450000	PTR	754433	823358	0.9641042	NA	0.036928	0.0007603
PPY	962841	1110628	1450000	PTR	824666	973271	0.9661971	NA	0.034718	0.0005046
PPY	1112419	1130773	1450000	PTR	1154751	1172537	0.9636045	0.0014089	0.0373943	0.0014878
PPY	1131744	1155292	1450000	PTR	1173756	1197011	0.9722748	0.0010892	0.0283301	0.0011375
PPY	1155706	1277186	1450000	PTR	1197430	1318598	0.9724457	NA	0.0281512	0.000496
PPY	1277986	1309453	1450000	PTR	1319021	1350361	0.9731759	0.0009152	0.0273846	0.000954
PPY	1310383	1337243	1450000	PTR	1351013	1376929	0.9629851	0.001193	0.0381427	0.0012674
PPY	1337285	1379266	1450000	PTR	1378506	1420133	0.9651755	0.000907	0.0357989	0.0009588
PPY	1380332	1382491	1450000	PTR	1421309	1423469	0.9693878	0.00371	0.0314109	0.0039073
PPY	1380337	1386786	1450000	PTR	1421314	1427885	0.9655849	0.0024343	0.0353631	0.0025711
PPY	1388100	1400795	1450000	PTR	1429052	1441605	0.9648806	0.0017058	0.0361356	0.0018066
PPY	1403409	1410348	1450000	PTR	1443147	1449999	0.9620409	0.002369	0.039139	0.0025197
PTR	261788	314603	1450000	PTR	1144414	1087024	0.9803464	NA	0.0199441	0.0006275
PTR	339710	398285	1450000	PTR	1040257	981545	0.9942617	0.0003126	0.0057638	0.0003154
PTR	366794	370157	1450000	PTR	1118905	1115425	0.9126451	0.0049346	0.0935511	0.0056722
PTR	1009808	1013165	1450000	PTR	1115425	1118905	0.9114233	0.0049657	0.0949792	0.0057228

A BLAST-based comparison (WGAC) method identified all pairwise sequence alignments (size ≥1 kb and ≥90% identity) among the 4 sequence assemblies: human H1, human H2, chimpanzee (PTR) and orangutan (PPY). Per_SE: % of sequence identity; K2p: Kimura 2 parameter. QNAME=query name; SNAME=subject name. All coordinates are based on the clone-based sequence assembly (supplementary file).

Supplementary Table 2 a: Inversion breakpoint intervals

Sequence	Inversion proximal BP		Inversion distal BP	
	start	end	start	end
H1	158471	185471	1128131	1198728
chr17*	40957727	40984727	41926919	41997516
H2	158373	185373	1320541	1391072

*Inversion breakpoints (BP) were identified based on comparison of H1 and H2 sequence assembly.
* The breakpoints of H1 were also mapped to human genome assembly (build36) at chr17.

Supplementary Table 2 b: Core segmental duplication intervals (10 kb chaining)

sequence	start	end	length
H1	144187	180281	36095
H1	932976	969438	36463
H1	1150544	1187293	36750
H2	141313	178076	36764
H2	359233	396011	36779
H2	1342939	1379705	36767
PTR	300915	308751	7837
PTR	1092945	1130188	37244
PTR	1595520	1603326	7807
PPY	403545	405392	1848

Position of core segmental duplications based on 10 kbp chaining method as described by Jiang et al., 2007.

Supplementary Table 2 c: Core segmental duplication intervals (5 kb chaining)

core	start	end	length
H1	172804	180281	7478
H1	932976	940455	7480
H1	1150544	1158030	7487
H2	170595	178076	7482
H2	388526	396011	7486
H2	1342939	1350427	7489
PTR	300915	308751	7837
PTR	1092945	1100348	7404
PTR	1595520	1603326	7807
PPY	403545	405392	1848

More stringent core segmental duplication interval as defined by chaining core duplicons within 5kb.

Supplementary Table 3: Pairwise sequence alignments (>5 kb) flanking inverted region

Human H1 vs. H1

QNAME	QB	QE	QLEN	SNAME	SB	SE	Length	Per_sim	K2p	Orientation
H1	129729	150021	1405709	H1	981035	963600	17354	0.981675694	0.018572994	Inverted
H1	129729	150021	1405709	H1	1198970	1181462	17427	0.982326275	0.017903657	Inverted
H1	150025	207448	1405709	H1	1185446	1123304	57294	0.983541034	0.016654597	Inverted
H1	150025	187318	1405709	H1	967593	925941	37189	0.980397429	0.019878259	Inverted
H1	219599	226634	1405709	H1	894677	887698	6915	0.9835141	0.016694636	inverted
H1	227453	261693	1405709	H1	886258	850934	33617	0.988458221	0.011645017	Inverted
sum							169796	0.983318792	0.016891527	

Human H1 vs. H1

QNAME	QB	QE	QLEN	SNAME	SB	SE	Length	Per_sim	K2p	Orientation
H2	129631	203770	1481053	H2	1391378	1317451	73867	0.997522574	0.002482041	Inverted
H2	257472	421714	1481053	H2	1481053	1317451	163531	0.998544618	0.001456983	Inverted
H2	428107	447822	1481053	H2	1186182	1163440	19664	0.990846216	0.009217729	Inverted
H2	**437714**	**452954**	**1481053**	**H2**	**1273877**	**1289120**	**15191**	**0.98597854**	**0.014174647**	**Direct**
H2	**440620**	**447822**	**1481053**	**H2**	**1186179**	**1193391**	**7201**	**0.992223302**	**0.007823391**	**Direct**
H2	452165	513195	1481053	H2	1163543	1103477	59223	0.990206508	0.009866043	Inverted
H2	468394	498328	1481053	H2	1317540	1289117	28246	0.991574028	0.008478729	Inverted
H2	**519560**	**594627**	**1481053**	**H2**	**1198880**	**1273881**	**74909**	**0.997490288**	**0.002514289**	**Direct**
sum							441832	0.993048259	0.007001732	

Chimpanzee vs. Chimpanzee Clone Assembly

QNAME	QB	QE	QLEN	SNAME	SB	SE	Length	Per_sim	K2p	Orientation
PTR	261788	314603	1450000	PTR	1144414	1087024	51899	0.980346442	0.019944123	Inverted
PTR	339710	398285	1450000	PTR	1040257	981545	58380	0.994261733	0.00576378	Inverted
sum							110279	0.987304088	0.012853952	

Orangutan vs. Orangutan Clone Assembly

QNAME	QB	QE	QLEN	SNAME	SB	SE	Length	Per_sim	K2p	Orientation
PPY	0	0	0	PPY	0	0	0	0	0	NA
sum										

The sequence identity and orientation of segmental duplications that flanking the chr17q21.31 inversion were examined (see Supplementary Table 1 for whole data set). We found that H2 when compared to chimpanzee or H1, shows higher sequence identity and ~95 kbp of duplications in the direct orientation on either side of the inversion. Q/SNAME=Query/Subject name along with coordinates within assembly, length of alignment (Length), percent sequence identity (Per_sim), genetic distance Kimura 2-parameter (K2p) model and orientation of alignment.

Supplemental Table 4: Evolutionary age estimation

	Sequence Divergence (K2p)				
Type	Length	Mismatch	%Div	K (kimura)	SE
H1tandem Dup	195264	353	0.180	0.001810	0.000096
H2-specific Dup	73733	183	0.242	0.002486	0.000184
PTR Dup	98145	1243	1.22	0.013088	0.000501
H1 vs. H2 Inversion				0.004170	0.000090

	Evolutionary Age T=K/2R	
	PTR outgroup	PPY outgroup
H1 tandem Dup	0.995± 0.053 mya	0.745 ± 0.040 mya
H2 Dup	1.367 ± 0.101 mya	1.024 ± 0.076 mya
PTR Dup	7.198 ± 0.275 mya	5.388 ± 0.206 mya
H1 vs. H2	2.293 ± 0.049 mya	1.717 ± 0.037 mya

We first estimated the average sequence divergence distance based on a 219 kb multiple alignment of Human H1, H2, chimpanzee and orangutan. Human (H1, H2) vs. chimpanzee distance K= (0.010930+0.010890)/2= 0.010910 (Table 2). Using chimpanzee as outgroup (6 mya), we plug in the equation R=K/2T and calculated the average substitution rate = 0.010910/12 mya = 9.0916×10^{-4} per site/mya. If we use orangutan as outgroup, we estimated the average substitution rate=0.034005/28 mya =1.214×10^{-3} per site/mya. Using above substitution rate (R) and sequence divergence (Kimura-2 parameter, K), we then estimated the evolutionary age (T=K/2R) of each event.

Supplementary Table 5: Chimpanzee sequence divergence in chr17q1.31 region

Library	Chimpanzee	Geographic Location	Total Number of Traces	Analyzed Traces	Aligned Positions	Differences	Heterozygosity
S221	Yvonne	Western	504,000	172	81,216	92	0.11%
S222	Karlien	Western	504,576	174	83,203	140	0.17%
S216	Mauku	Central	507,840	106	41,042	137	0.33%
S217	Noemie	Central	592,577	170	83,201	318	0.38%
S215	Clara	Central	495,456	90	41,163	69	0.17%
Fosmids*	Clint		1,493,254	106	48,408	144	0.30%

In order to assess the extent of chimpanzee nucleotide diversity of this region, we analyzed diversity of this region from five other chimpanzees by aligning paired-end sequences to the unique portion of the Clint 17q21.31 BAC assembly (See Methods). *For Clint, only ESPs with at least one end mapping with <100% identity were considered. Thus, the true divergence between the two chromosomes found in Clint will be less than the calculated value of 0.3%.