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.

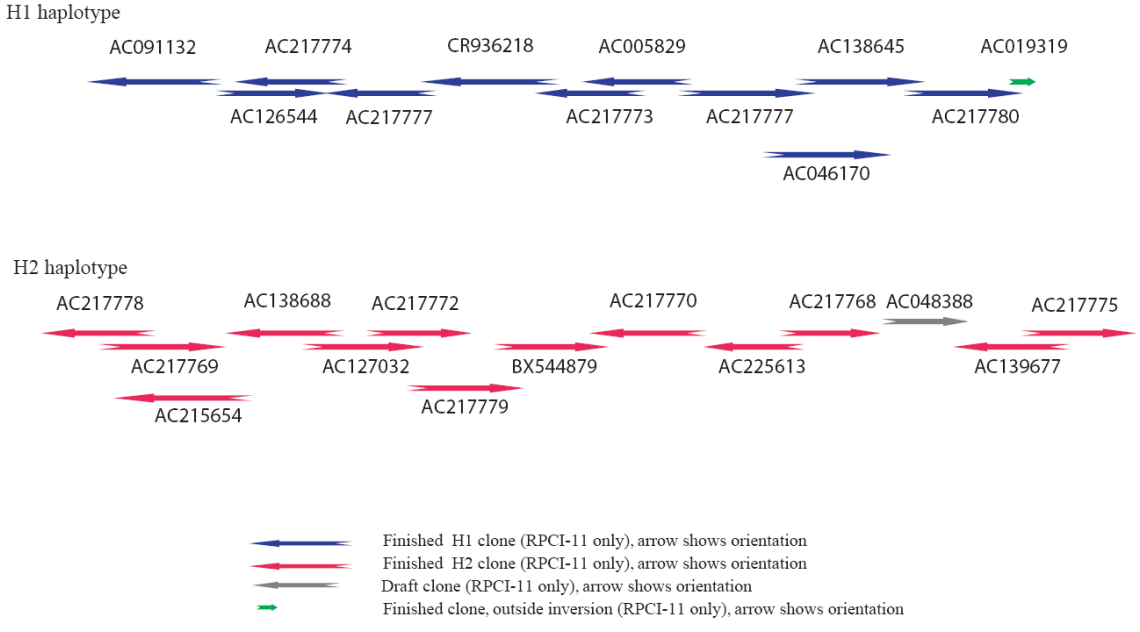

Figure 1: Human H1 and H2 RPCI-11 clone sequence assembly

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.

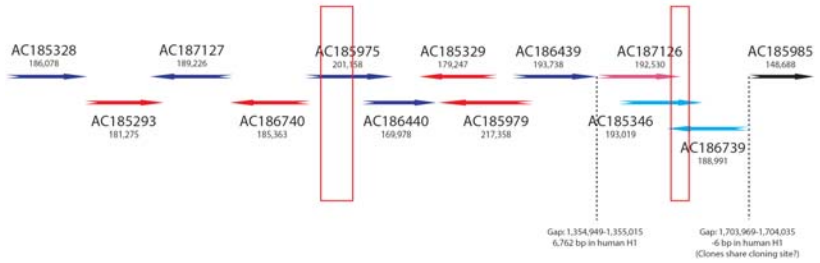

Figure 2: Chimpanzee MAPT locus clone assembly. Colors indicate clone chromosome of origin. In the first contig, clones alternate blue and red. In the second, light red and light blue. The third contig contains a single black clone. As yet, there is no ability to link the chromosomes of origin across 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).

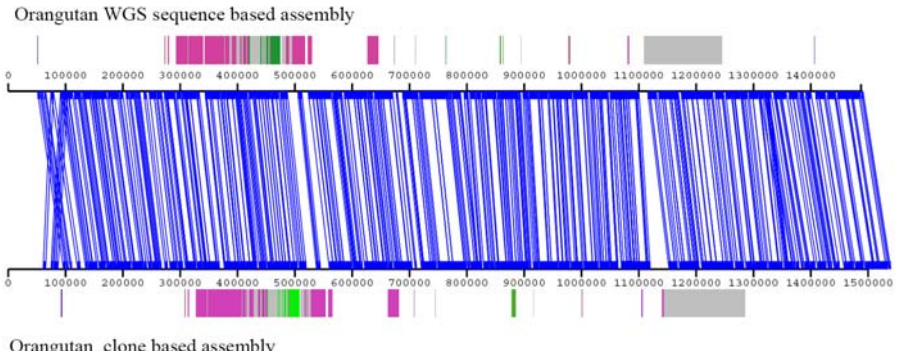

Figure 3: Comparison between sequence based assembly and clone based assembly of the MAPT region in orangutan. Parallel blue joining-lines show consistency in sequence structure and orientation (Miropeats –s 2000). The largest discrepancy was an 18 kbp segment that was missing from the WGSA assembly (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).

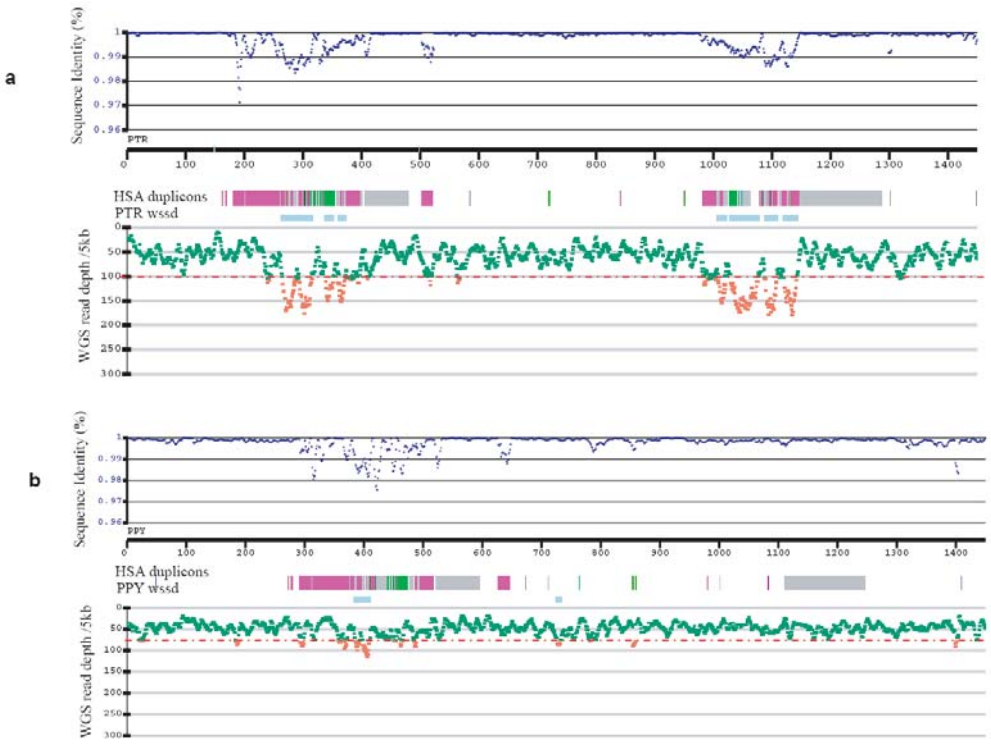

Figure 5: Non-human primate segmental duplication analysis.

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.

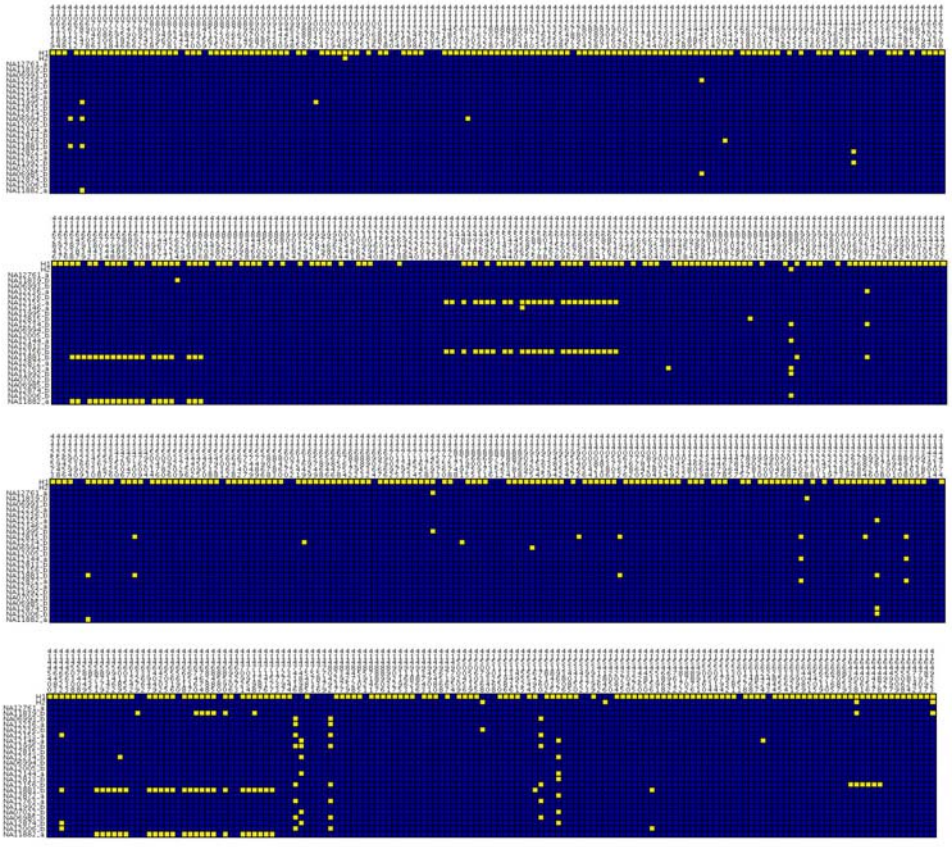

Figure 6: Observed H2 haplotypes. The alleles present at 611 HapMap SNPs are depicted for the sequenced H1 and H2 haplotypes and for all 24 inferred CEU H2 chromosomes within the HapMap sample set.

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.

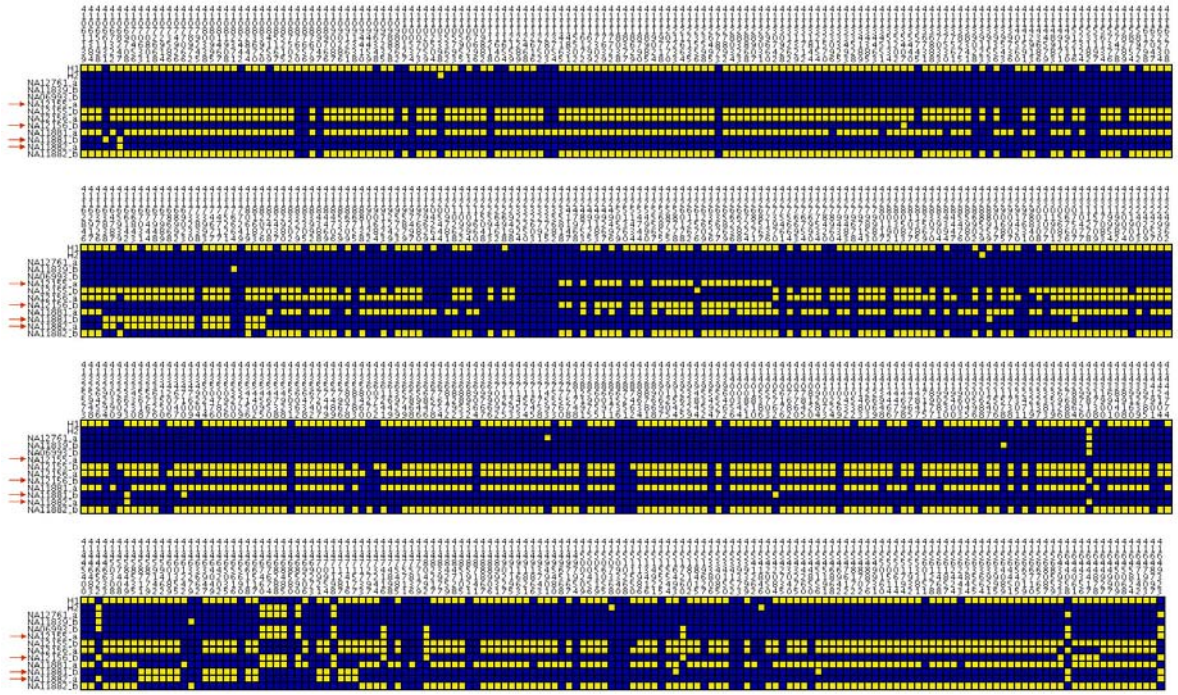

Figure 7: Identification of phasing errors. The four inferred H2 haplotypes containing unusual stretches of H1-like genotypes are depicted (NA11882_a, NA11881_b, NA12156_b, NA12155_a, highlighted by red arrows) along with the other haplotype from the same samples. The H1 and H2 sequenced haplotypes as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

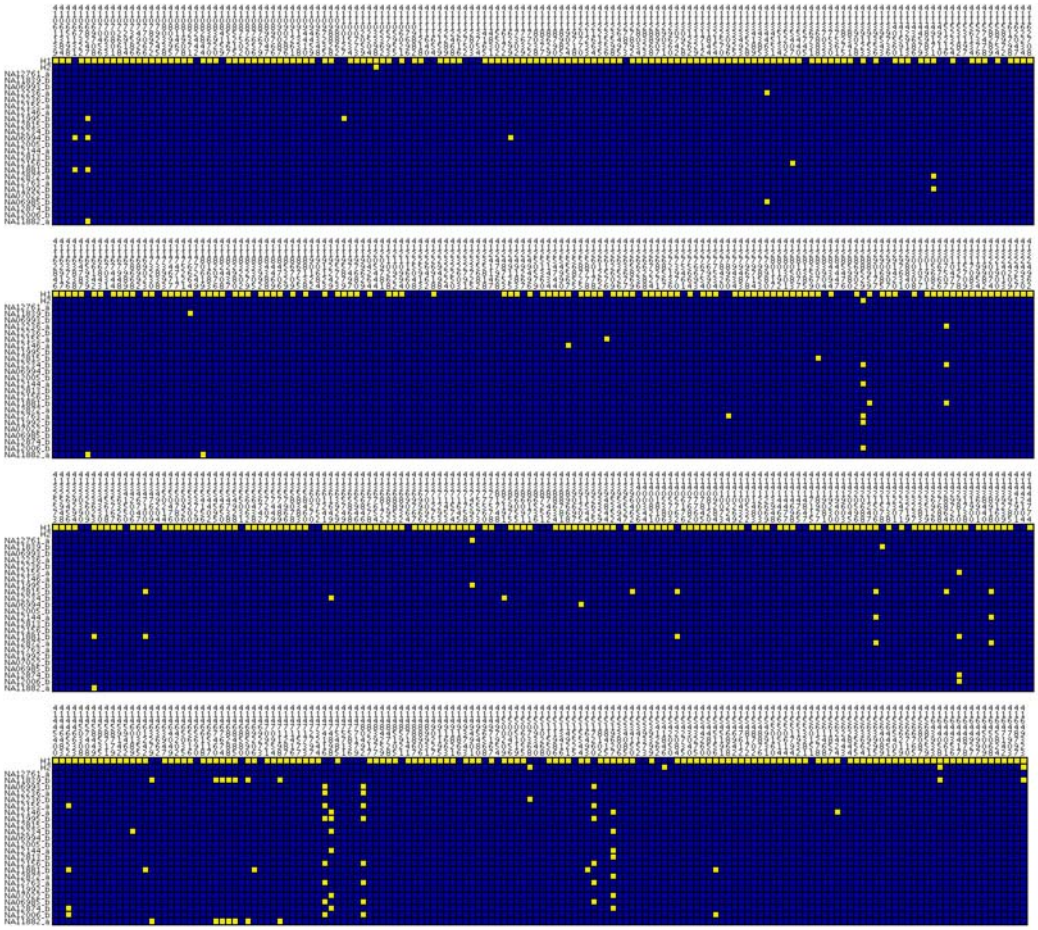

Figure 8: Sequence variation among H2 haplotypes. Variation among H2 haplotypes (depicted as in Fig. 5), following the correction of likely haplotype phasing errors.

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).

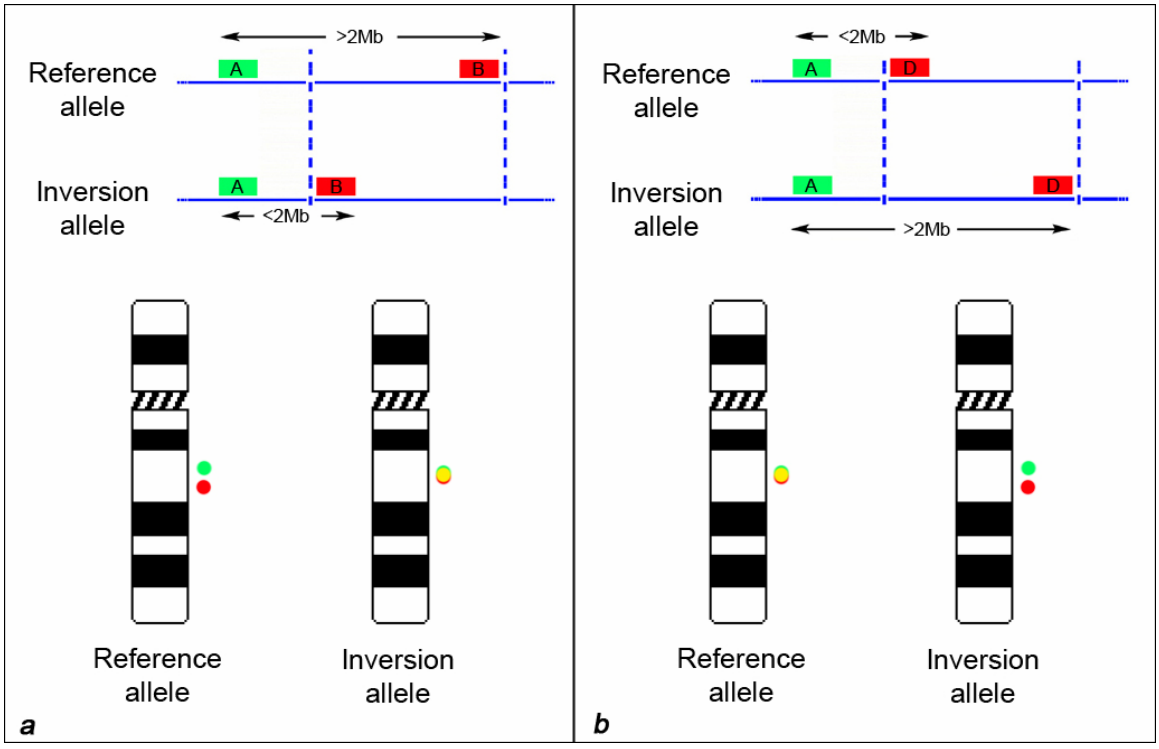

Figure 9: Chr17q21.31 reciprocal inversion FISH assay.

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).

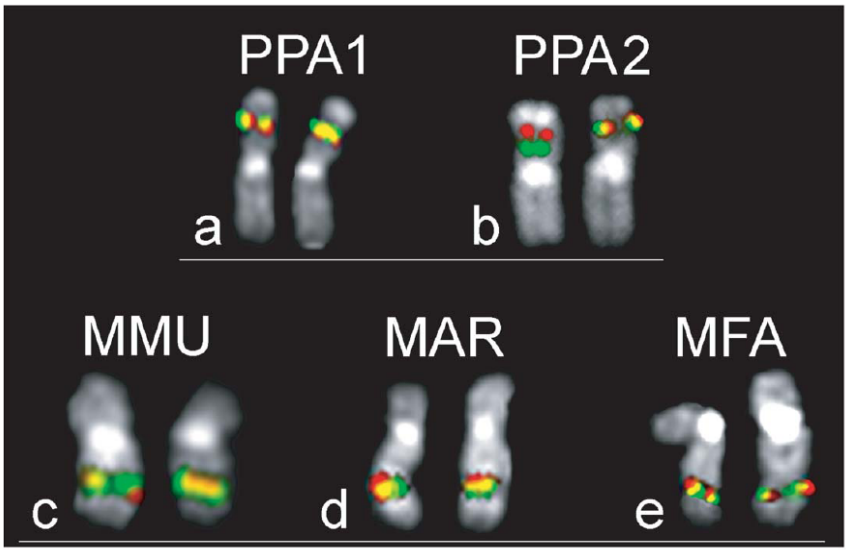

Figure 10: Primate FISH analysis of 17q21.31 inversion.

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)

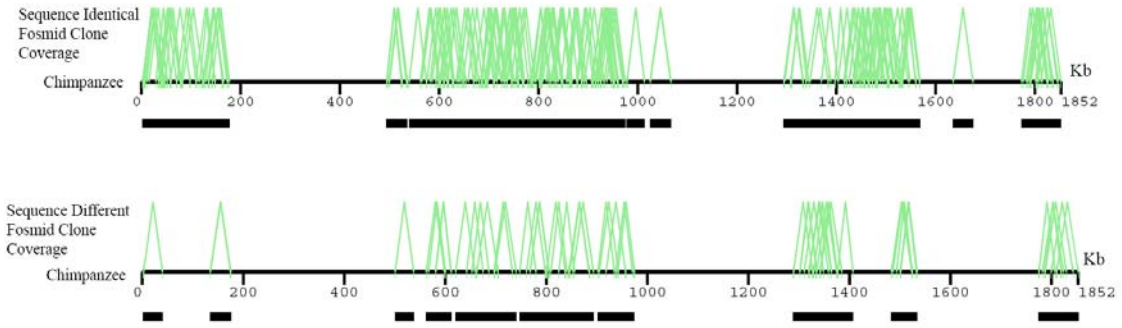

Figure 11: Chimpanzee haplotype analysis.

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).