nature genetics

Evolutionary toggling of the *MAPT* 17q21.31 inversion region

Michael C Zody[1,2,9], Zhaoshi Jiang[3,9], Hon-Chung Fung[4,5], Francesca Antonacci[3], LaDeana W 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[8], Mario Ventura[7], John Hardy[4], Wesley C Warren[6] & Evan E Eichler[3]

Using comparative sequencing approaches, we investigated the evolutionary history of the European-enriched 17q21.31 *MAPT* inversion polymorphism. We present a detailed, BAC-based sequence assembly of the inverted human H2 haplotype and compare it to the sequence structure and genetic variation of the corresponding 1.5-Mb region for the noninverted H1 human haplotype and that of chimpanzee and orangutan. We found that inversion of the *MAPT* region is similarly polymorphic in other great ape species, and we present evidence that the inversions occurred independently in chimpanzees and humans. In humans, the inversion breakpoints correspond to core duplications with the *LRRC37* gene family. Our analysis favors the H2 configuration and sequence haplotype as the likely great ape and human ancestral state, with inversion recurrences during primate evolution. We show that the H2 architecture has evolved more extensive sequence homology, perhaps explaining its tendency to undergo microdeletion associated with mental retardation in European populations.

It has become clear that a large proportion of genetic variability among humans and between humans and chimpanzees involves large-scale genomic structural changes such as deletions, insertions and inversions[1–5]. In this regard, the ∼970-kb inversion of the *MAPT* (microtubule-associated protein tau) locus on human chromosome 17 represents one of the most structurally complex and evolutionarily dynamic regions of the genome[6–8]. This locus occurs in humans as two haplotypes, H1 (direct orientation) and H2 (inverted orientation)[6,9], which show no recombination between them over a region of ∼1.5 Mb[10]. The two haplotypes have different functional impacts. Consistent differences in cortical gene expression have been observed between the two[11]. Specific H1 haplotypes are associated with Alzheimer's disease, amyotrophic lateral sclerosis and parkinsonism dementia complex of Guam, corticobasal degeneration and progressive supranuclear palsy[9,10,12–15]. The H2 haplotype is predisposed to recurrent microdeletions associated with the 17q21.31 microdeletion syndrome[16–18].

The H1 haplotype occurs in all populations and shows a normal pattern of genetic variability and recombination. In contrast, the H2 haplotype occurs predominantly in populations of European ancestry[19], where it shows limited sequence diversity but extensive diversity (0.3%) compared to H1, suggesting an ancient coalescence ∼3 million years ago[6,8,20]. Both the ancient inversion and the microdeletion event are thought to have arisen as the result of nonallelic homologous recombination between large blocks of segmental duplications (200–500 kb in length). The goal of this study was to reconstruct the evolutionary history of this region by conducting detailed analysis of its sequence organization and assessing variation in its structure within and between human and nonhuman primate populations.

RESULTS

Duplication analysis

Given the central role of the duplications in both the microdeletion and the evolution of the inversion, we began our analysis by comparing the duplication architecture among primate species. According to the H1 haplotype organization in the genome assembly, the inversion is flanked by two duplication blocks 203 kb (proximal) and 484 kb (distal) in length. We estimated the evolutionary timing of various segmental duplications by comparing the duplication architecture in

[1]Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA. [2]Department of Medical Biochemistry and Microbiology, Uppsala University, Box 597, Uppsala SE-751 24, Sweden. [3]Department of Genome Sciences, Howard Hughes Medical Institute, University of Washington, Seattle, Washington 98195, USA. [4]Department of Molecular Neuroscience and Reta Lila Weston Laboratories, Institute of Neurology, University College London, London WC1N 3BG, UK. [5]Department of Neurology, Chang Gung Memorial Hospital and College of Medicine, Chang Gung University, 199 Tung Hwa North Road, Taipei 10591, Taiwan. [6]Genome Sequencing Center, Washington University School of Medicine, Campus Box 8501, 4444 Forest Park Avenue, St. Louis, Missouri 63108, USA. [7]Department of Genetics and Microbiology, University of Bari, Via Amendola 165/A, Bari 70126, Italy. [8]Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, Maryland 20892, USA. [9]These authors contributed equally to this work. Correspondence should be addressed to E.E.E. (eee@gs.washington.edu) or W.C.W. (wwarren@watson.wustl.edu).

Received 4 March; accepted 27 May; published online 10 August 2008; corrected online 17 August 2008 (details online); doi:10.1038/ng.193

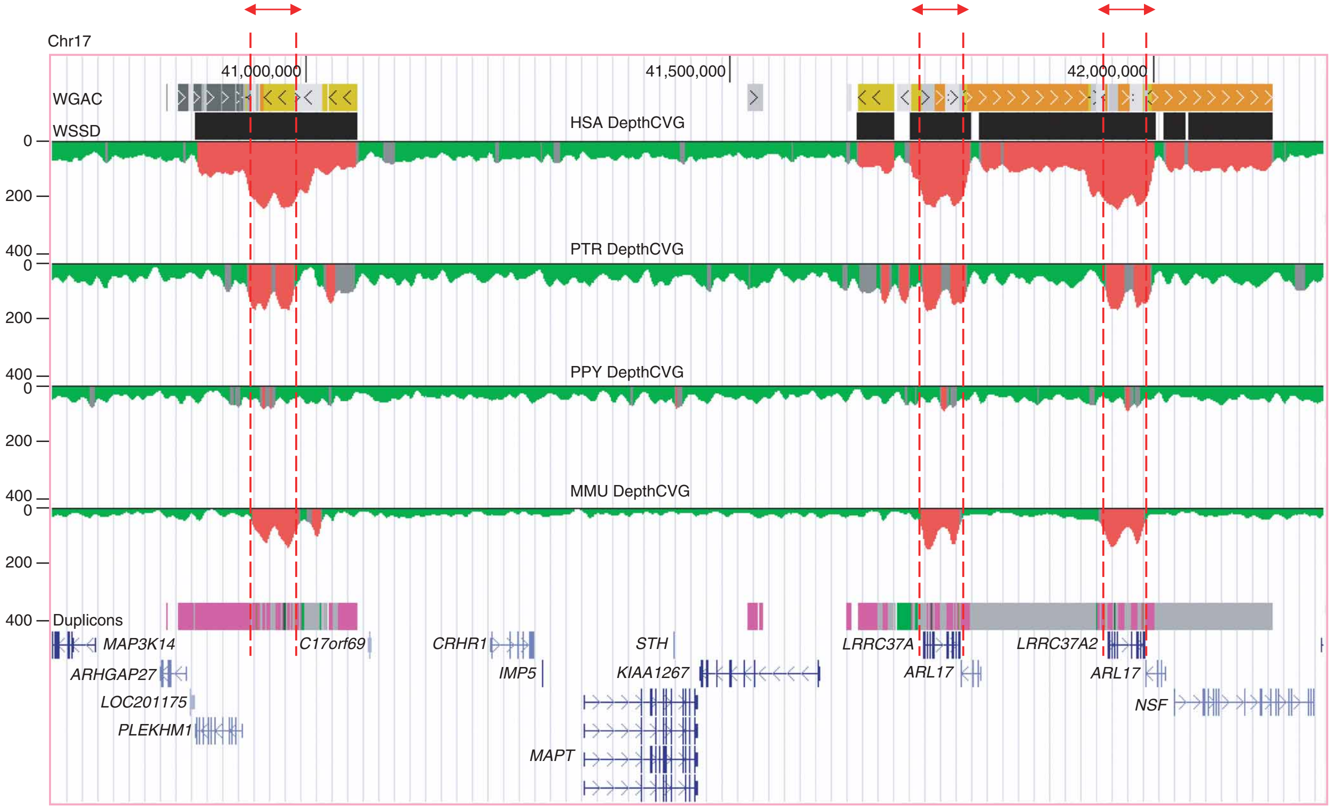

Figure 1 Comparative segmental duplication analysis of the 17q21.31 region. Regions of excess (≥mean + 3 s.d.; red) WGS depth of coverage are shown for human (HSA), chimpanzee (PTR), orangutan (PPY) and macaque (MMU) mapped against the human reference genome (build36). This approach detects ≥90% of all segmental duplications that are longer than 10 kb and have ≥94% sequence identity[5]. The analysis suggests that the majority (~71%) of the duplication architecture is human specific, except for a core duplicated segment corresponding to the *LRRC37A* gene family (highlighted by red dashed lines)[32]. DepthCVG, depth coverage of WGS reads within 5-kb gap-free window; WGAC, whole-genome assembly comparison[38]; WSSD, WGS sequence detection[21].

human, chimpanzee, orangutan and macaque (**Fig. 1**). Using whole-genome shotgun (WGS) sequence data from each species (see Methods), we mapped regions of excess read depth and sequence divergence against the human reference genome assembly (build36; **Fig. 1**). This approach may be used to accurately predict large (≥10 kb), high-identity segmental duplications within[21] and between[5] species. We found that 71% (486 of 687 kb) of the duplication architecture is specific to the human species (that is, not detected as duplicated in the chimpanzee, orangutan or macaque genome). The analysis predicts that most (≥87%) of the segmental duplications emerged after the divergence of the chimpanzee and human lineage from the orangutan <12 million years ago; this was subsequently confirmed by a more detailed examination of the chimpanzee and orangutan sequence assemblies, which show limited evidence of duplications within the orangutan sequence assembly for this locus (**Supplementary Note** online). Notably, a core segmental duplication of ~40 kb, corresponding to the *LRRC37* (leucine-rich repeat–containing 37A) gene family, is distributed throughout chromosome 17 and predicted to be one of the few duplications common to chimpanzee, human and macaque.

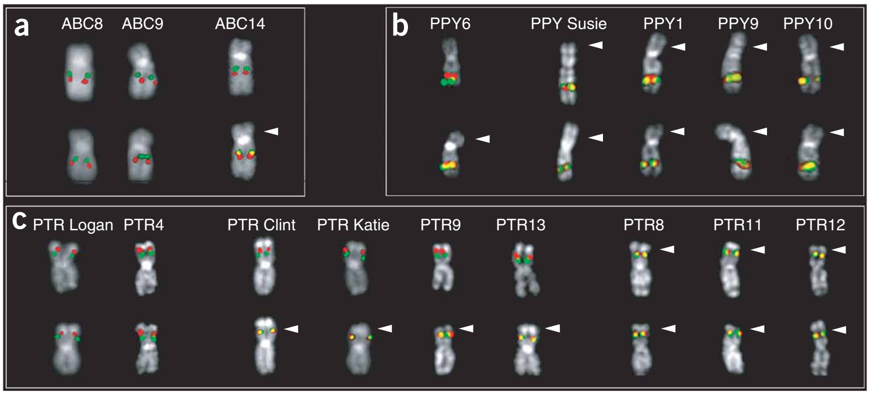

Figure 2 Inversion polymorphism among primates. A metaphase FISH assay distinguishes between the H2 orientation (merged yellow) and H1 orientation (distinct green and red) signals based on proximity of two unique probes (**Supplementary Note**). Shown are extract chromosome 17 from three human (**a**), five orangutan (**b**) and nine chimpanzee (**c**) lymphoblast cell lines. The H2 orientation (arrowhead) predominates in orangutan and chimpanzee samples. In humans, the H2 haplotype is restricted to Middle Eastern and European populations.

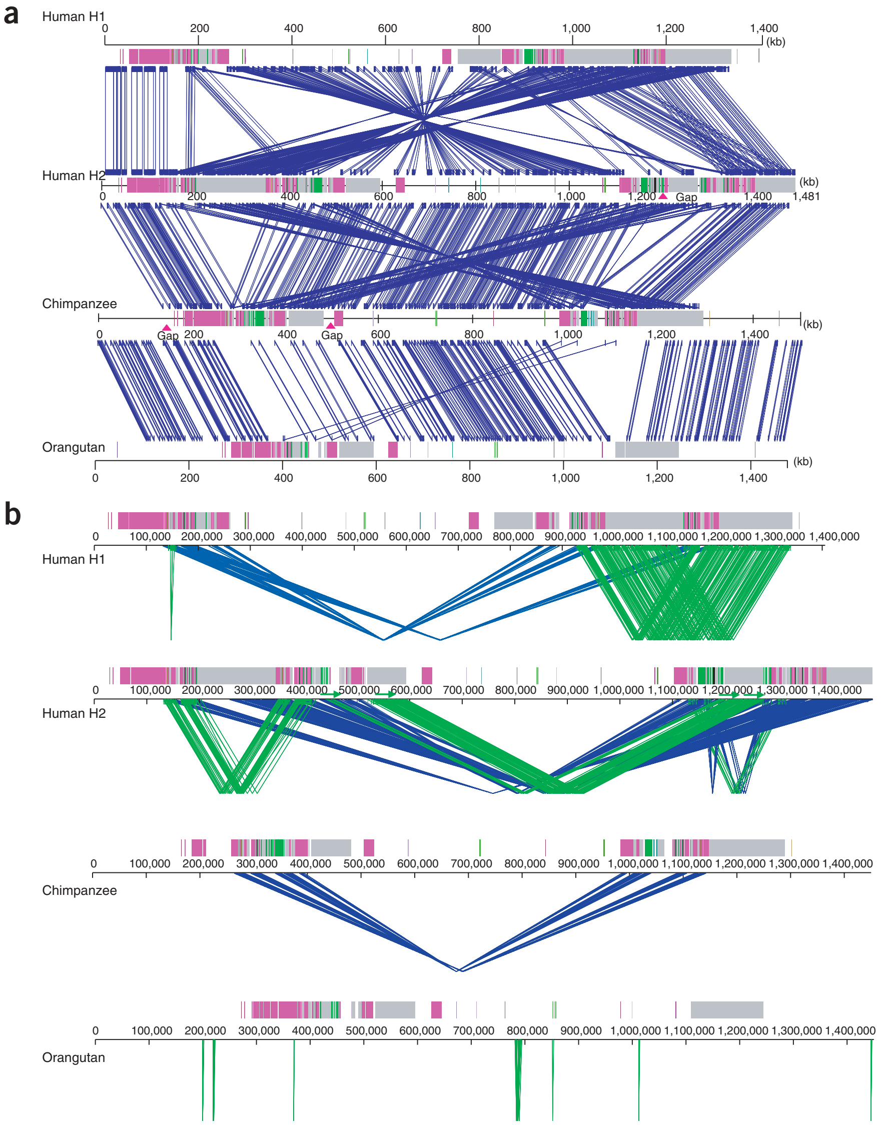

Figure 3 Sequence comparison of the human H1, H2, chimpanzee and orangutan 17q21.31 region. (**a**) BAC-based sequence assemblies of the human inverted (H2) and noninverted haplotype (H1) were compared using Miropeats[39] to a BAC-based assembly of the chimpanzee (PTR) and a WGS-based assembly of the orangutan (PPY; Miropeats sequence similarity threshold –s 1,000). Regions of homology are shown with blue lines joining the corresponding sequence above. Duplicon architecture based on human segmental duplications is overlaid as colored or gray bars. H1 shows a ~970-kb inverted segment compared to H2, chimpanzee and orangutan sequence assemblies. The H2 sequence assembly shows a relocation of large (~200 kb) high-identity duplications on either side of the unique interval compared to chimpanzee (crisscross pattern). Comparison of orangutan and chimpanzee shows evidence of a ~100-kb segmental duplication from proximal to distal duplication block, which probably occurred in the common ancestor of chimpanzee and human (6–12 million years ago). (**b**) The extent of local direct (green) and inverted (blue) intrachromosomal segmental duplications (SDs) flanking the inversion are shown for human H1 and H2 haplotypes, chimpanzee and orangutan (Miropeats sequence similarity threshold –s 300). We examined the duplication content (whole-genome assembly comparison) within each assembly and computed the number of nonredundant duplicated base pairs for each assembly (**Supplementary Table 1**). No homologous SDs (sequence identity ≥ 90%, size ≥ 1 kb) were found in orangutan genome flanking the inversion region, whereas in chimpanzee and H1 haplotype, 292 kb and 227 kb were identified, respectively. H2 shows the most extensive duplication architecture flanking the inversion, including 95 kb in direct orientation.

Inversion analysis

We next developed a reciprocal FISH assay to characterize the orientation of the region by taking advantage of the physical limits of metaphase chromosomes to resolve distinct signals (**Supplementary Note**). We tested for the presence of the inversion by examining lymphoblastoid cell lines from a diverse panel of hominoids and macaque Old World monkey species. Although the H1 and H2 haplotypes are specific to humans, for simplicity, we will refer to the H1 and H2 orientations when describing the configuration in other nonhuman primate species. All three macaque species tested (*Macaca fascicularis, Macaca arctoides* and *Macaca mulatta*) and orangutan showed FISH signatures consistent with the H2 orientation, suggesting that this orientation represents the ancestral configuration (**Supplementary Note**). Surprisingly, examination of a single individual from each of the two chimpanzee species (*Pan paniscus* and *Pan troglodytes*) showed that they were heterozygous for the inversion. We examined a larger population of unrelated chimpanzees ($n = 9$, *Pan troglodytes*) and found the inversion to be highly polymorphic (**Fig. 2**). Unlike the human population, the H2 configuration represents the major allele (56% allele frequency) in chimpanzee. All Sumatran orangutans were homozygous for the H2 orientation; however, analysis of a single Bornean orangutan (PPY6) showed that it was heterozygous, indicating that the inversion is likely to be polymorphic within this subspecies (**Fig. 2b**). Combined, these data argue that the H2 orientation represents the ancestral state and that this region of the genome has been subject to inversion polymorphisms for the last 12 million years of hominoid evolution.

Sequence analysis

Breakpoint refinement of the human inversion is complicated by extensive structural variation within the flanking duplication blocks[6,8,16]. Because the current genome assembly is based on the sequence of multiple individuals, we constructed and sequenced a BAC-based assembly corresponding to the human H2 haplotype (1,481 kb) and the human H1 haplotype (1,406 kb) from a donor that was heterozygous for the inversion (RPCI-11). Requiring 100% sequence identity overlap between overlapping BAC clones ensured that two distinct sequence haplotypes could be constructed. A subsequent examination of 79 diagnostic SNPs confirmed that the H1 and H2 haplotypes had been successfully resolved. We also developed a BAC-based assembly of the chimpanzee (1,852 kb) and the orangutan

Table 1 Duplication alignments flanking the *MAPT* inversion

Sequence	Alignments[a]	Length (bp)	% Identity	K2M	SE
PTR	2	110279	98.73	0.012814	0.000471
H1	6	169796	98.34	0.016892	0.000920
H2	8	441832	99.30	0.007002	0.000517

K2M, Kimura two-parameter model genetic distance estimates; PTR, chimpanzee; SE, standard error.

[a]Only pairwise sequence alignments >5 kb mapping within the two duplication blocks flanking the inversion were considered. In H1 and PTR, all pairwise alignments are in an inverted orientation with respect to one another; only within the H2 haplotype were three alignments identified in a direct orientation (corresponding to 97,301 bp with 99.53% sequence identity).

(1,859 kb) in H2 orientation, requiring haplotype contiguity specifically over the breakpoint regions. There was a paucity of segmental duplications in orangutan, so the WGS and clone-based assemblies were virtually identical (see **Supplementary Note** for details regarding the sequence and assembly of these regions).

We compared the sequence organizations of the human H1 and H2 haplotypes, and we compared both human haplotypes to the nonhuman primate sequence assemblies (**Fig. 3** and **Supplementary Fig. 1** online). We identified all regions of segmental duplication based on a variety of independent analyses (**Supplementary Table 1** online). The analysis revealed several important features. First, sequence alignments confirmed an 'H2 orientation' for the *CRHR1* (corticotropin-releasing hormone receptor-1)-*MAPT* region in chimpanzees and orangutan compared to the H1 haplotype. After the inversion, the largest genomic structural difference seems to have occurred within the shared human and chimpanzee lineage, where a duplicative transposition (≥100 kb) placed two inverted copies of the core segmental duplication on either side of the inversion region (**Fig. 3b**). Second, although it was impossible to precisely delineate the breakpoints at the single base-pair level because of the high degree of sequence identity, it was possible to identify the inversion H1-H2 breakpoint intervals based on alignment of flanking sequences. We estimated the inversion to be ~970 kb in length and found that each of the four breakpoints map to a *LRRC37* core duplication (**Supplementary Fig. 1a** and **Supplementary Table 2** online).

Sequence comparison with nonhuman primates revealed that more extensive and complex duplication architecture has emerged in the evolutionary lineage leading to humans (**Table 1** and **Fig. 3b**). Focusing only on those duplications that align between the proximal and distal blocks (**Supplementary Table 3** online), we found that the H1 duplication organization is slightly (59.5 kb) larger than that of chimpanzee. In contrast, the H2 haplotype shows the greatest duplication complexity. We found a total of 441 kb of homologous sequence flanking either side of the inverted region in H2, compared to only 169 kb for H1. Similarly, we found that the average sequence identity for the H2 sequences is substantially greater (99.3%) than that of the H1 sequences (98.3%). We constructed a series of phylogenetic trees from a multiple sequence alignment of shared duplication (40 kb) common to human H1, H2, chimpanzee and orangutan (**Supplementary Note**). In most cases, duplicated sequences from H2 grouped separately from H1, suggesting that the H2 segmental duplications have been markedly homogenized by gene conversion or secondary duplication events.

In addition to greater sequence identity, we found important differences in the orientation of the duplications. Within the sequenced H2 contig, there are 95 kb of segmental duplication in direct orientation (**Fig. 3b**). This contrasts with the H1 and chimpanzee sequence, where none of the alignments between the proximal and distal duplication blocks are in direct H1 orientation. Among the H2 alignments, we identified in particular a 73-kb 'H2-only' segmental duplication noted previously as a copy-number polymorphism in the human population[6,22]. To test whether this large direct repeat might have a role in the predisposition of H2 to microdeletions, we compared the evolutionary inversion breakpoints with the predicted microdeletion breakpoints associated with the H2 and 17q21.31 microdeletion (**Supplementary Fig. 2** online)[16,17]. We found that the inversion breakpoints and microdeletion breakpoints are not identical. Notably, one of the microdeletion breakpoints maps within the largest H2-specific segmental duplications, suggesting that the

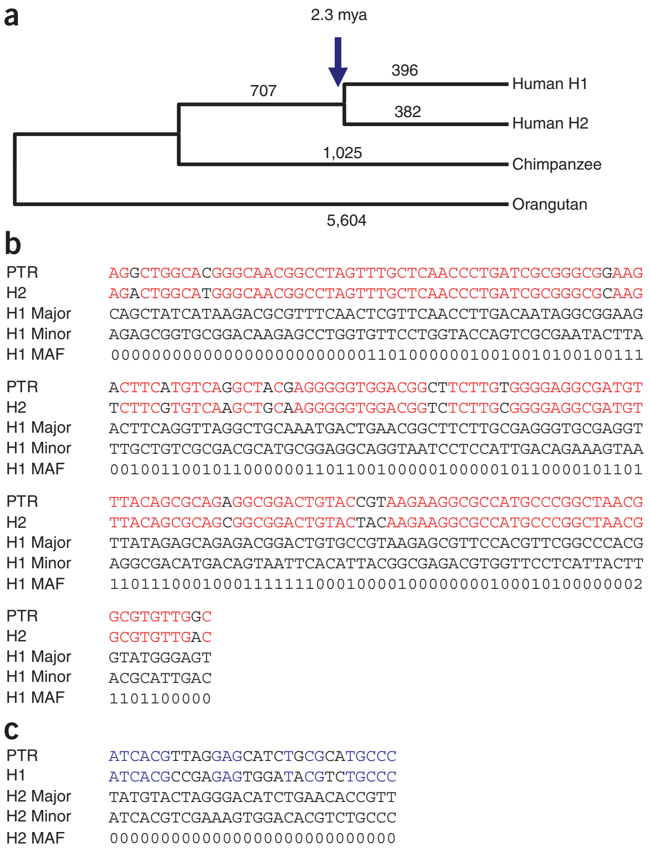
a
2.3 mya
707
396
382
Human H1
Human H2
1,025
Chimpanzee
Orangutan
5,604
b
PTR AGGCTGGCACGGGCAACGGCCTAGTTTGCTCAACCCTGATCGCGGGCGGAAG
H2 AGACTGGCATGGGCAACGGCCTAGTTTGCTCAACCCTGATCGCGGGCGCAAG
H1 Major CAGCTATCATAAGACGCGTTTCAACTCGTTCAACCTTGACAATAGGCGGAAG
H1 Minor AGAGCGGTGCGGACAAGAGCCTGGTGTTCCTGGTACCAGTCGCGAATACTTA
H1 MAF 0000000000000000000000001101000000100100101001001111
PTR ACTTCATGTCAGGCTACGAGGGGGTGGACGGCTTCTTGTGGGGAGGCGATGT
H2 TCTTCGTGTCAAGCTGCAAGGGGGTGGACGGTCTCTTGCGGGGAGGCGATGT
H1 Major ACTTCAGGTTAGGCTGCAAATGACTGAACGGCTTCTTGCGAGGGTGCGAGGT
H1 Minor TTGCTGTCGCGACGCATGCGGAGGCAGGTAATCCTCCATTGACAGAAAGTAA
H1 MAF 0010011001011000000110110010000010000010110000101101
PTR TTACAGCGCAGAGGCGGACTGTACCGTAAGAAGGCGCCATGCCCGGCTAACG
H2 TTACAGCGCAGCGGCGGACTGTACTACAAGAAGGCGCCATGCCCGGCTAACG
H1 Major TTATAGAGCAGAGACGGACTGTGCCGTAAGAGCGTTCCACGTTCGGCCCACG
H1 Minor AGGCGACATGACAGTAATTCACATTACGGCGAGACGTGGTTCCTCATTACTT
H1 MAF 1101110001000111111100010000100000000100010100000002
PTR GCGTGTTGGC
H2 GCGTGTTGAC
H1 Major GTATGGGAGT
H1 Minor ACGCATTGAC
H1 MAF 1101100000
c
PTR ATCACGTTAGGAGCATCTGCGCATGCCC
H1 ATCACGCCGAGAGTGGATACGTCTGCCC
H2 Major TATGTACTAGGGACATCTGAACACCGTT
H2 Minor ATCACGTCGAAAGTGGACACGTCTGCCC
H2 MAF 0000000000000000000000000000

Figure 4 Phylogenetic and SNP haplotype analysis. (**a**) An unrooted neighbor-joining phylogram was constructed (MEGA pairwise deletion option, sum of branch lengths = 0.0414) based on 219,165 aligned base pairs from unique sequence within the inverted region. H1 and H2 sequence taxa clustered together with 100% bootstrap support (*n* = 500 replicates). The number of single-nucleotide variants specific for each branch in the tree is assigned above each branch. We estimate that the H1 and H2 haplotypes diverged 2.3 million years ago (mya; **Table 2**). (**b**,**c**) We treated H1 and H2 haplotypes as separate populations in the analysis and 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 (**b**), the allele found in the H2 haplotypes matched the chimpanzee allele 90% of the time (150 of 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 of 28 considered positions) (**c**). Red indicates alleles shared between chimpanzee and H2; blue indicates shared alleles between H1 and chimpanzee. The major and minor alleles are denoted, with the minor allele frequency represented by a single digit (for example, '2' refers to a minor allele frequency (MAF) of ≥20%). PTR, chimpanzee.

Table 2 Sequence divergence of orthologous sequences

	H1	H2	PTR	PPY
H1	—	0.000090	0.000140	0.000260
H2	0.004170	—	0.000140	0.000250
PTR	0.010930	0.010890	—	0.000260
PPY	0.034090	0.033920	0.033790	—

Kimura two-parameter model genetic distance estimates (left diagonal) and standard error (right diagonal). There were 219 kb of four-way alignment of unique sequence within the inversion interval. Tajima's relative rate test showed that the genomic sequence is evolving neutrally ($P = 0.22$–0.81). PTR, chimpanzee; PPY, orangutan.

large direct repeats that emerged specifically within the H2 lineage predispose to rearrangement. Further experimentation will be required to define the microdeletion breakpoints more precisely.

To determine the most likely ancestral state in humans, we constructed a 219-kb multiple sequence alignment of human H1, H2, chimpanzee and orangutan from unique sequences mapping to the inversion interval (**Fig. 4**). Similar to previous analyses[6,8], the phylogenetic tree of all single-nucleotide variants did not distinguish H1 or H2 as ancestral. Rather, the analysis revealed that human H1 and H2 arose from an intermediate ancestral haplotype, with a large and approximately equal number of haplotype-specific single-nucleotide variants ($n = 382$ and 396, respectively) mapping to each of the H1 and H2 lineages. Assuming the chimpanzee and human lineages diverged 6 million years ago, the diversity between H1 and H2 (0.476%) predicts that the two human haplotypes diverged 2.3 million years ago (**Table 2**). Furthermore, if we assume that the inversion was a unique event in human evolutionary history and that the inversion has been an effective barrier to recombination, we can treat the H1 and H2 regions as nonmixing populations. All modern copies of the derived population must be descended from a single founder, so all variants present only in the derived population must have arisen since the inversion. This means that for all SNPs segregating in the derived population, the allele found in the ancestral population would be more likely to match the chimpanzee variant. To reduce the impact of genotype error caused by paralogous sequences, we limited our consideration to HapMap SNPs that can be uniquely mapped onto both sequenced haplotypes (**Supplementary Note**). Dividing SNPs into those only variant within H1 haplotypes (fixed in H2) and those only variant in H2 haplotypes (fixed in H1), we found that 90% (150 of 166) of SNPs polymorphic in H1 have an H2 allele matching the chimpanzee allele, whereas for those variant only in H2, only 60% (17 of 28) have an H1 allele matching the chimpanzee allele (**Fig. 4b,c**, **Table 3** and **Supplementary Note**). This significant result ($P = 0.0002332$, Fisher's exact test) is consistent with an ancestral H2 state in humans and inconsistent with an ancestral H1 state. Notably, we found a small fraction of shared polymorphic sites in H1 and H2, which represent either recurrent CpG mutations or, possibly, gene flow between the H1 and H2 regions, perhaps as a result of gene conversion within the inversion loop.

The fact that the inversion is polymorphic in human, chimpanzee, bonobo and orangutan may be the result of evolutionary recurrence[23,24] or lineage-specific sorting of an ancient polymorphism[25]. To assess the reciprocal event within a nonhuman primate lineage, we took advantage of the fact that the sequenced chimpanzee (Clint) was heterozygous for the inversion (**Fig. 2c**). We aligned all end-sequences derived from Clint against the BAC-based chimpanzee haplotype (H2 orientation at the breakpoints) (**Supplementary Note**). Excluding duplicated sequences, the level of sequence divergence (1 of 336 bp, or 0.30%) confirmed that the two chimpanzee haplotypes had emerged recently (within the last 1–2 million years of chimpanzee evolution). These values are consistent with global estimates of chimpanzee diversity[26] but slightly less than diversity between chimpanzee and bonobos (0.354%)[27]. On the basis of sequence divergence between the human H1 and H2 haplotypes, we calculated a more ancient origin for the divergence of human H1 and H2 lineages (1.9–2.7 million years ago based on uncertainty in the chimpanzee-human divergence, which is the largest contributor to error in time estimates), but still clearly within the *Homo* lineage of evolution. Combined, these data strongly argue that the H1 orientation emerged independently in both lineages. Taken together with the observation of both chromosomal configurations in bonobo (*Pan paniscus*) and Bornean orangutan, the data also suggest that this particular region has been prone to recurrent inversion events within multiple primate lineages (**Fig. 5** and **Supplementary Table 4** online). We propose that this inversion 'toggling' has contributed, in part, to the complex duplication architecture that emerged in this region over the last 12 million years of evolution[28].

DISCUSSION

Our analysis establishes the H2 orientation as the most likely great ape and human ancestral state. Notably, we found that inversion of the *CRHR1-MAPT* region is similarly polymorphic in other extant great ape populations, where it represents the major allele. Despite the fact that the inverted configuration occurs in only 20% of European chromosomes, both SNP haplotype analysis and comparative FISH analysis point to an inverted H2-like ancestor. It was previously assumed that H1 was the ancestral sequence because >99% of sub-Saharan African haplotypes are variants of the H1 clade[6]. Based on analysis of the CEPH-HGDP sample collection[29], the few Mbuti and Biaka pygmies with an H2 allele (HGdp980, HGdp985, HGdp463 and HGdp474) have a haplotype and SNP architecture identical to that of the European H2 allele, making it difficult to distinguish an ancient origin from recent admixture[19,30]. We propose that an H2-like allele

Table 3 Analysis of various SNP classes

			Among H1 chromosomes			Among H2 chromosomes		
Category	Number of SNPs	PTR unknown	Equal frequency	Maj = PTR	Maj ≠ PTR	Equal frequency	Maj = PTR	Maj ≠ PTR
H2 fixed, H1 polymorphic	178	12	0	108	58	0	150	16
H1 fixed, H2 polymorphic	29	1	0	17	11	0	12	16
H1 fixed, H2 fixed	381	39	0	164	178	0	178	164
Polymorphic among both H1 and H2	23	0	1	18	8	0	19	4

Shown are total number of SNPs in each category, number of SNPs where corresponding chimpanzee allele could not be confidently determined, and ancestral classification among H1 and H2 chromosomes. Maj = PTR, major allele in class matches chimpanzee sequence; Maj ≠ PTR, major allele in class is different from chimpanzee allele; PTR unknown, SNPs where chimpanzee allele could not be determined because no high-identity BLAT alignment could be found. We identified 23 SNPs that were polymorphic in both H1 and H2 chromosomes; most are single occurrences and are likely to reflect genotyping errors (**Supplementary Note**).

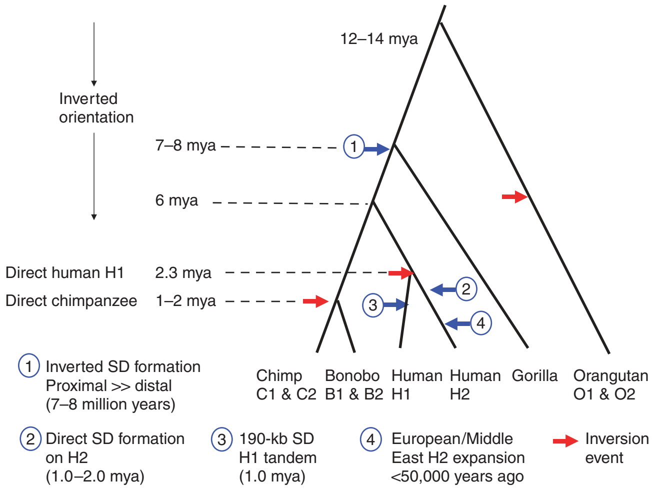
12–14 mya
Inverted orientation
7–8 mya
6 mya
Direct human H1
2.3 mya
Direct chimpanzee
1–2 mya
1
2
3
4
Chimp C1 & C2
Bonobo B1 & B2
Human H1
Human H2
Gorilla
Orangutan O1 & O2
1 Inverted SD formation Proximal >> distal (7–8 million years)
2 Direct SD formation on H2 (1.0–2.0 mya)
3 190-kb SD H1 tandem (1.0 mya)
4 European/Middle East H2 expansion <50,000 years ago
Inversion event

Figure 5 Evolutionary model of inversion toggling, segmental duplication formation and relationship with disease susceptibility. We estimated the evolutionary age of various duplication, gene conversion and rearrangement events by establishing a local molecular clock for single-nucleotide substitution (**Supplementary Table 4**) and superimposed these estimates over a generally accepted hominoid phylogeny[45]. Because of uncertainty in the chimpanzee-human divergence, timing of events should be considered an approximation. We propose that the ancestral *CRHR1-MAPT* region was inverted but has toggled to an H1 orientation multiple times within the evolution of various great ape and human lineages (red arrows). Large blocks (≥100 kb) of inverted segmental duplications were formed in the common ancestor of chimpanzee and human, further predisposing the region to recurrent inversion. An inversion of the predominant H2 allele created the H1 allele ~2.3 million years ago (mya) within the human lineage. Subsequently, larger blocks of directly oriented segmental duplications (SDs) emerged within the H2 lineage, predisposing it to microdeletion and disease. As a result of this negative selection against H2, the H1 haplotype rose in frequency and became the predominant allele in all human populations, with a subsequent polymorphic tandem duplication occurring on some haplotypes. In the out-of-Africa European founding population, however, the H2 allele resurged in frequency as a result of a partial selective sweep or a population bottleneck in the founding population.

was the predominant allele among ancestral populations (*Homo heidelbergensis*)[20], but its frequency was subsequently reduced and nearly eliminated in ancestral African *Homo sapiens* populations. The lack of diversity among extant human H2 haplotypes and its apparent ancient origin (2–3 million years ago) could be the result of either a founder effect[20]or a partial selective sweep of a particular H2 haplotype within the European population, as has been posited[6].

Here we present evidence that the inversion occurred independently in both chimpanzees and humans. Although the data are limited, the finding of both orientations in the Bornean orangutan argue strongly that this particular region of chromosome 17 is prone to recurrent inversions and has toggled multiple times between the inverted and noninverted state during the course of hominoid evolution (**Fig. 5**). In humans and chimpanzees, these changes occurred in concert with the evolution of a more complex duplication architecture flanking the inverted region in humans. These findings are strikingly reminiscent of an evolutionary survey of the human *FLNA-EMDA* X chromosome inversion[24]. The X chromosome inversion has been shown to have occurred independently ten times in 27 Eutherian lineages, and in each case the region was flanked by duplications in an inverted orientation. If taken as a general principle of genome evolution, these data suggest extraordinary breakpoint reuse for inversions and predict that some apparently fixed inversions between species may actually be polymorphic as a result of recurrence.

In our study, we similarly found large, inverted segmental duplications flanking the inversion region in humans and chimpanzees, and we showed that the H1-H2 inversion breakpoints map in close proximity to the *LRRC37* core duplicon sequence within these inverted segmental duplications. Cores represent some of the most abundant and rapidly evolving duplicated sequences in the human genome, perhaps because they are prone to double-strand breakage and/or positive selection[23,31,32]. Moreover, such sequences have been shown to be associated with recurrent evolutionary events[23]. Although there are more than 11 copies of the *LRRC37* core duplicon on chromosome 17, evolutionary reconstruction in primates shows that the proximal (H1) 17q21.31 locus corresponds to the ancestral position[32]. Comparative analyses of the mouse sequence[8] and macaque genome[33] reveal that this region of the genome has been a hotbed for multiple inversions and other rearrangements during mammalian evolution—long before most of the hominoid duplication architecture emerged. We propose that inversion toggling is a longstanding evolutionary property of the 17q21.31 region, promoted, in part, because of its association with the *LRRC37* core duplicon sequence. Most of the human and chimpanzee large segmental duplications flanking the inversions are, themselves, within an inverted orientation, and it is possible that such structures were created as part of the double-strand DNA repair process[28,34]. Such inverted segmental duplications, once formed, would reinforce and continue to promote recurrent inversion events through nonallelic homologous recombination.

Although our analysis of the unique sequences identified an H2-like sequence as the likely ancestral allele, a detailed comparison of the genomic architecture of the segmental duplications suggests that the extant H2 sequence is much more highly derived than the H1 or chimpanzee (**Table 1** and **Fig. 3b**). Phylogenetic analysis of the duplicated sequences supports extensive H2-specific sequence homogenization, perhaps as a result of gene conversion between proximal and distal segmental duplication blocks. Consequently, there are three times as many duplicated base pairs in H2 compared to chimpanzee or H1; these duplicated bases show higher sequence identity, and ~95 kb are in direct orientation on either side of the inversion. Orientation, length and degree of sequence identity between duplicated sequences are the most important parameters for nonallelic homologous recombination[35]. In the case of H2, the orientation, proportion and sequence identity would all favor microdeletion on this chromosome haplotype compared to H1. We showed that at least one of the microdeletion breakpoints associated with developmental delay and mental retardation in children corresponds to a recently evolved H2-specific segmental duplication. We propose that it is not the inversion *per se* that promotes microdeletion and disease; rather, the configuration and structure of the segmental duplications favors nonallelic homologous recombination on the particular inverted haplotype. Marked changes in copy number, structure and homology of flanking segmental duplications may explain why inversion haplotypes predispose to other microdeletion syndromes[36,37].

METHODS

Segmental duplication detection. Segmental duplication content of the *CRHR1-MAPT* region was initially assessed by mapping WGS sequence assembly reads from human, chimpanzee, orangutan and macaque against human chromosome 17 (build 36, chr17:40799295–42204344) and identifying regions of excess (≥mean + 3 s.d.) depth of coverage and divergence, as described previously[5]. For all hominoids, sequence identity alignment thresholds were set at ≥94%, with the exception of macaque, where a ≥88% identity threshold was used to capture more divergent macaque sequence reads aligned to the human genome. Three independent approaches were used to analyze the segmental duplication content of each clone-based sequence assembly. A BLAST-based whole-genome assembly comparison method[38] identified all sequence alignments ≥1 kb in length and with ≥90% identity[38] among the four sequence assemblies. The WGS sequence detection approach identified regions ≥10 kb in length with a significant excess of high-quality WGS reads[21] within overlapping 5-kb windows. WGS sequence detection analysis was based on an alignment of 22,590,543 chimpanzee WGS reads and 18,355,056 orangutan WGS reads against their BAC-based sequence assemblies. Finally, we annotated all human duplications by WU-BLAST alignments of a non-redundant data set of human duplicons[32] against each assembly. High-identity sequence alignments were generated using Miropeats[39] and visualized using a Perl script of two-way-mirror.pl (J.A. Bailey, Case Western Reserve University, personal communication).

FISH inversion assay. Metaphase spreads were obtained from lymphoblast cell lines from two human HapMap individuals (YRI NA18507 and CEU NA12156, Coriell Cell Repository), nine chimpanzees (Clint, Katie, Logan, PTR14, PTR8, PTR9, PTR11, PTR12 and PTR13), four Sumatran orangutans (Susie, PPY1, PPY9 and PPY10), one Bornean orangutan (PPY6), two bonobos (PPA1 and PPA2) and three subspecies of macaque: MMU (*Macaca mulatta*), MAR (*Macaca arctoides*) and MFA (*Macaca fascicularis*). Inversions were detected using a two-color FISH assay (fosmid probes WIBR2-634F12 and WIBR2-1948K20), and inversion genotype status was confirmed using a reciprocal assay (fosmids WIBR2-634F12 and ABC9_41289800G20). Inversion genotyping accuracy was tested by comparing FISH genotypes to a previously designed molecular assay (**Supplementary Note**). All 24 human samples (three H2 and 45 H1 chromosomes) were concordant between FISH and molecular assays. Probes were directly labeled by nick-translation with Cy3-labeled dUTP (Perkin-Elmer) and labeled with fluorescein-dUTP (Enzo). Each hybridization used 300 ng of labeled probe, 5 μg of COT1 DNA (Roche) and 3 μg of sonicated salmon sperm DNA at 37 °C in 10 μl of 2× saline–sodium citrate buffer, 50% formamide and 10% dextran sulfate. This was followed by three washes at 60 °C in 0.1× saline–sodium citrate buffer. Nuclei were stained with DAPI, and digital images were obtained using a Leica DMRXA2 epifluorescence microscope equipped with a cooled charge-coupled device camera (Princeton Instruments).

Sequence and assembly. We constructed, sequenced and assembled minimal tiling paths of large-insert genomic clones for both human haplotypes (H1 and H2), an H2-oriented chimpanzee chromosome and an H2-oriented orangutan chromosome (see **Supplementary Note** for detailed clone order, sequence assembly and annotation). In humans, this entailed disentangling existing H1 and H2 RPCI-11 BACs and generating an additional 1.7 Mb of high-quality finished sequence. In chimpanzee and orangutan, a minimum tiling path of BAC clones (chimpanzee, CHORI-251; orangutan, CHORI-276) was sequenced to derive a consensus assembly (∼2 Mb) that identified BACs containing inversion breakpoints. Orangutan consensus sequence was also extracted from the *Pongo pygmaeus* draft assembly 2.0.2. To verify the *MAPT* locus orientation, analyze flanking duplication architecture and measure evolutionary distance of haplotypes in chimpanzee and orangutan, we used the corresponding regions (human build36 chr17:40799295-42204344) from both whole-genome and BAC-based consensus sequence assemblies. Sequences were compared using Miropeats, and inversion breakpoint intervals were defined based on a consistent orientation shift between the aligned sequence assemblies.

Phylogenetic and haplotype analyses. An unrooted neighbor-joining[40] phylogram was constructed (MEGA pairwise deletion option)[41] based on a multiple sequence alignment (ClustalW)[42] of 219,165 bp within the inverted region. Genetic distances were computed using the Kimura two-parameter method[43], and Tajima's relative rate test (PPY-H1-H2; PTR-H1-H2) was used to assess branch length neutrality ($P = 0.22$–0.81). Using chimpanzee as the outgroup, an estimated local substitution rate (9.0916×10^{-4} substitutions per site per million years) and the uncertainty in chimpanzee-human divergence (5–7 million years ago), we calculated that the human H1 and H2 haplotypes diverged 1.9–2.7 million years ago. We compared 123 chromosomes (120 CEPH HapMap chromosomes, and the sequence of the H1, H2 and PTR haplotypes) using HapMap SNPs (phase II HapMap release 21 phased consensus)[44]. SNP genotypes were assigned to the H1, H2 and PTR sequences using BLAT, and regions of segmental duplication (including H2-specific duplications) were excluded. Haplotypes were assigned to the H1 or H2 class based on two diagnostic SNPs (rs1800547 and rs9468), as described previously[6]. Errors in the inferred SNP-phased haplotypes were manually corrected (**Supplementary Note**). We assessed haplotype diversity within the chimpanzee based on alignment of Clint fosmid end-sequence pairs to the BAC-based chimpanzee assembly (**Supplementary Note** and **Supplementary Table 5** online).

URLs. The *Pongo pygmaeus* draft assembly 2.0.2 is available at http://genome.wustl.edu/genome_group_index.cgi/. The phase II HapMap release 21 phased-consensus is available at http://hapmap.org/.

Accession codes. The BAC clones that were used for sequence and assembly of the *MAPT* region in this study have been submitted to GenBank under the following accession numbers. Human H1 assembly: AC091132, AC126544, AC217774, AC217771, CR936218, AC217773, AC005829, AC217777, AC138645 and AC217780. Human H2 assembly: AC217778, AC217769, AC138688, AC127032, AC217772, AC217779, BX544879, AC217770, AC225613, AC217768, AC139677 and AC217775. Chimpanzee assembly: AC185328, AC185293, AC187127, AC186740, AC185975, AC186440, AC185329, AC185979, AC186439, AC187126, AC185346, AC186739 and AC185985. Orangutan assembly: AC205775, AC206340, AC206276, AC207288, AC206550, AC206558, AC205859, AC206353, AC216075, AC206444, AC207097, AC216102, AC216058 and AC216103.

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

ACKNOWLEDGMENTS

We thank T. Marques, D. Reich, A. Navarro, N. Patterson, S. McCarroll, T. Brown and K. Augustyn for critical comments and valuable discussions in the preparation of this manuscript; C. Alkan for providing computational assistance; and members of the Broad Institute Sequence Platform (BISP) and Washington University Genome Sequencing Center (WUGSC) for generating clone-based sequencing data for this project. M.C.Z., A.A., W.C.W., L.W.H., T.A.G., R.K.W., A.D.R., J.W., J.G. and the BISP and WUGSC were supported by grants from the National Human Genome Research Institute. This work was supported in part by the Intramural Research Program of the National Institute on Aging, National Institutes of Health, Department of Health and Human Services, project number Z01 AG000957-05. This work was supported by NIH grants GM058815 and HG002385 to E.E.E. and a Rosetta Inpharmatics Fellowship (Merck Laboratories) to Z.J. E.E.E. is an investigator of the Howard Hughes Medical Institute.

AUTHOR CONTRIBUTIONS

Z.J., M.C.Z. and E.E.E. analyzed and annotated sequence organization; J.M.K. did the haplotype analyses; F.A., M.F.C. and M.V. developed the FISH inversion assay and typed all primate metaphase chromosomes; L.C. and Z.C. did the segmental duplication analyses; M.C.Z., W.C.W., A.A., T.A.G., L.W.H., A.D.R., H.C.F., J.W., J.G. and J.D. generated, sequenced and analyzed the BAC clone assembly; R.K.W. oversaw sequence production; E.E.E., M.C.Z., Z.J., W.C.W., J.H. and J.M.K. drafted the manuscript; W.C.W., J.H. and E.E.E. designed the study; and E.E.E. finalized the manuscript.

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

1. Iafrate, A.J. *et al.* Detection of large-scale variation in the human genome. *Nat. Genet.* **36**, 949–951 (2004).
2. Sebat, J. *et al.* Large-scale copy number polymorphism in the human genome. *Science* **305**, 525–528 (2004).
3. Tuzun, E., Bailey, J.A. & Eichler, E.E. Recent segmental duplications in the working draft assembly of the brown Norway rat. *Genome Res.* **14**, 493–506 (2004).
4. Chimpanzee Sequence and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69–87 (2005).
5. Cheng, Z. *et al.* A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature* **437**, 88–93 (2005).
6. Stefansson, H. *et al.* A common inversion under selection in Europeans. *Nat. Genet.* **37**, 129–137 (2005).
7. Gijselinck, I. *et al.* Visualization of *MAPT* inversion on stretched chromosomes of tau-negative frontotemporal dementia patients. *Hum. Mutat.* **27**, 1057–1059 (2006).

8. Cruts, M. *et al.* Genomic architecture of human 17q21 linked to frontotemporal dementia uncovers a highly homologous family of low-copy repeats in the tau region. *Hum. Mol. Genet.* **14**, 1753–1762 (2005).
9. Baker, M. *et al.* Association of an extended haplotype in the tau gene with progressive supranuclear palsy. *Hum. Mol. Genet.* **8**, 711–715 (1999).
10. Pittman, A.M. *et al.* The structure of the tau haplotype in controls and in progressive supranuclear palsy. *Hum. Mol. Genet.* **13**, 1267–1274 (2004).
11. Myers, A.J. *et al.* A survey of genetic human cortical gene expression. *Nat. Genet.* **39**, 1494–1499 (2007).
12. Conrad, C. *et al.* Genetic evidence for the involvement of tau in progressive supranuclear palsy. *Ann. Neurol.* **41**, 277–281 (1997).
13. Rademakers, R. *et al.* High-density SNP haplotyping suggests altered regulation of tau gene expression in progressive supranuclear palsy. *Hum. Mol. Genet.* **14**, 3281–3292 (2005).
14. Sundar, P.D. *et al.* Two sites in the *MAPT* region confer genetic risk for Guam ALS/PDC and dementia. *Hum. Mol. Genet.* **16**, 295–306 (2007).
15. Myers, A.J. *et al.* The H1c haplotype at the *MAPT* locus is associated with Alzheimer's disease. *Hum. Mol. Genet.* **14**, 2399–2404 (2005).
16. Sharp, A.J. *et al.* Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. *Nat. Genet.* **38**, 1038–1042 (2006).
17. Koolen, D.A. *et al.* A new chromosome 17q21.31 microdeletion syndrome associated with a common inversion polymorphism. *Nat. Genet.* **38**, 999–1001 (2006).
18. Shaw-Smith, C. *et al.* Microdeletion encompassing *MAPT* at chromosome 17q21.3 is associated with developmental delay and learning disability. *Nat. Genet.* **38**, 1032–1037 (2006).
19. Evans, W. *et al.* The tau H2 haplotype is almost exclusively Caucasian in origin. *Neurosci. Lett.* **369**, 183–185 (2004).
20. Hardy, J. *et al.* Evidence suggesting that *Homo neanderthalensis* contributed the H2 *MAPT* haplotype to *Homo sapiens*. *Biochem. Soc. Trans.* **33**, 582–585 (2005).
21. Bailey, J.A. *et al.* Recent segmental duplications in the human genome. *Science* **297**, 1003–1007 (2002).
22. Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444**, 444–454 (2006).
23. Johnson, M.E. *et al.* Recurrent duplication-driven transposition of DNA during hominoid evolution. *Proc. Natl. Acad. Sci. USA* **103**, 17626–17631 (2006).
24. Caceres, M., Sullivan, R.T. & Thomas, J.W. A recurrent inversion on the eutherian X chromosome. *Proc. Natl. Acad. Sci. USA* **104**, 18571–18576 (2007).
25. Ebersberger, I. *et al.* Mapping human genetic ancestry. *Mol. Biol. Evol.* **24**, 2266–2276 (2007).
26. Gagneux, P. The genus *Pan*: population genetics of an endangered outgroup. *Trends Genet.* **18**, 327–330 (2002).
27. Yu, N. *et al.* Low nucleotide diversity in chimpanzees and bonobos. *Genetics* **164**, 1511–1518 (2003).
28. Kehrer-Sawatzki, H., Sandig, C.A., Goidts, V. & Hameister, H. Breakpoint analysis of the pericentric inversion between chimpanzee chromosome 10 and the homologous chromosome 12 in humans. *Cytogenet. Genome Res.* **108**, 91–97 (2005).
29. Rosenberg, N.A. *et al.* Genetic structure of human populations. *Science* **298**, 2381–2385 (2002).
30. Jakobsson, M. *et al.* Genotype, haplotype, and copy number variation in worldwide human populations. *Nature* **451**, 998–1003 (2008).
31. Zody, M.C. *et al.* Analysis of the DNA sequence and duplication history of human chromosome 15. *Nature* **440**, 671–675 (2006).
32. Jiang, Z. *et al.* Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution. *Nat. Genet.* **39**, 1361–1368 (2007).
33. Cardone, M.F. *et al.* Hominoid chromosomal rearrangements on 17q map to complex regions of segmental duplication. *Genome Biol.* **9**, R28 (2008).
34. Casals, F. & Navarro, A. Chromosomal evolution: inversions: the chicken or the egg? *Heredity* **99**, 479–480 (2007).
35. Lupski, J.R. Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends Genet.* **14**, 417–422 (1998).
36. Osborne, L.R. *et al.* A 1.5 million-base pair inversion polymorphism in families with Williams-Beuren syndrome. *Nat. Genet.* **29**, 321–325 (2001).
37. Gimelli, G. *et al.* Genomic inversions of human chromosome 15q11-q13 in mothers of Angelman syndrome patients with class II (BP2/3) deletions. *Hum. Mol. Genet.* **12**, 849–858 (2003).
38. Bailey, J.A., Yavor, A.M., Massa, H.F., Trask, B.J. & Eichler, E.E. Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res.* **11**, 1005–1017 (2001).
39. Parsons, J. Miropeats: graphical DNA sequence comparisons. *Comput. Appl. Biosci.* **11**, 615–619 (1995).
40. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425 (1987).
41. Tamura, K., Dudley, J., Nei, M. & Kumar, S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol. Biol. Evol.* **24**, 1596–1599 (2007).
42. Higgins, D.G. CLUSTAL V: multiple alignment of DNA and protein sequences. *Methods Mol. Biol.* **25**, 307–318 (1994).
43. Kimura, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**, 111–120 (1980).
44. International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
45. Goodman, M. The genomic record of Humankind's evolutionary roots. *Am. J. Hum. Genet.* **64**, 31–39 (1999).