nature genetics

Structural diversity and African origin of the 17q21.31 inversion polymorphism

Karyn Meltz Steinberg[1,11], Francesca Antonacci[1,11], Peter H Sudmant[1], Jeffrey M Kidd[1,10], Catarina D Campbell[1], Laura Vives[1], Maika Malig[1], Laura Scheinfeldt[2], William Beggs[2], Muntaser Ibrahim[3], Godfrey Lema[4], Thomas B Nyambo[4], Sabah A Omar[5], Jean-Marie Bodo[6], Alain Froment[7], Michael P Donnelly[8,10], Kenneth K Kidd[8], Sarah A Tishkoff[2] & Evan E Eichler[1,9]

The 17q21.31 inversion polymorphism exists either as direct (H1) or inverted (H2) haplotypes with differential predispositions to disease and selection. We investigated its genetic diversity in 2,700 individuals, with an emphasis on African populations. We characterize eight structural haplotypes due to complex rearrangements that vary in size from 1.08–1.49 Mb and provide evidence for a 30-kb H1-H2 double recombination event. We show that recurrent partial duplications of the *KANSL1* gene have occurred on both the H1 and H2 haplotypes and have risen to high frequency in European populations. We identify a likely ancestral H2 haplotype (H2′) lacking these duplications that is enriched among African hunter-gatherer groups yet essentially absent from West African populations. Whereas H1 and H2 segmental duplications arose independently and before human migration out of Africa, they have reached high frequencies recently among Europeans, either because of extraordinary genetic drift or selective sweeps.

Chromosomal rearrangements occur in many species and can contribute to phenotypic variability and genomic evolution[1–5]. Compared to other structural variants, inversions may be under different selective pressures, because recombination is suppressed between heterokaryotypes[6–9]. The 17q21.31 inversion locus represents one of the most dynamic and complex regions of the human genome. Two haplotypes exist, in direct (H1) and inverted (H2) orientation, which previous studies have shown do not recombine over nearly 2 Mb, resulting in extended linkage disequilibrium (LD)[10]. The H2 haplotype is enriched in Europeans, and carriers are predisposed to the 17q21.31 microdeletion syndrome as a result of non-allelic homologous recombination (NAHR) between directly oriented segmental duplications present on the inverted chromosome[11–14]. A recent study of copy-number variation in the 1000 Genomes Project showed that a 205-kb duplication is associated with 30% of European H1 haplotypes, whereas a smaller 155-kb duplication in the same region is fixed in European H2 haplotypes[15]. The latter predisposes to NAHR and, thus, to the 17q21.31 microdeletion syndrome.

Using short tandem repeats, a recent study[16] estimated that the time to the most recent common ancestor (TMRCA) with the H2 haplotype was between 16,000–108,000 years ago and that the H2 haplotype originated in Africa; however, sequence divergence between H1 and H2 indicates a more ancient coalescence of 2.3 million years ago. The discovery of an H2 haplotype without duplication from the genome sequence of a Khoisan Bushman[17] suggested recent structural changes in the evolution of the H2 lineage. Given the importance of the H2-specific duplication in disease and its substantial population stratification, we explored the architecture of this region in more detail using a combination of next-generation sequencing (NGS), array–comparative genomic hybridization (aCGH) and FISH in a total of 2,700 individuals from diverse geographic populations. We specifically surveyed the distribution of the H2 haplotype in African ancestry groups with variable modes of subsistence, focusing on hunter-gatherer populations, to capture potentially ancient structural and nucleotide diversity.

RESULTS

Duplication architecture of the 17q21.31 region

Using NGS of 620 individuals from 3 major continental groups (Africans, Asians and Europeans; 1000 Genomes Project) and 185 admixed individuals (total n = 805), we estimated copy-number variation in the 17q21.31 region using sequence read depth as

[1]Department of Genome Sciences, University of Washington, Seattle, Washington, USA. [2]Department of Genetics and Biology, University of Pennsylvania, Philadelphia, Pennsylvania, USA. [3]Department of Molecular Biology, Institute of Endemic Diseases, University of Khartoum, Khartoum, Sudan. [4]Department of Biochemistry, Muhimbili University of Health and Allied Sciences, Dar es Salaam, Tanzania. [5]Kenya Medical Research Institute, Center for Biotechnology Research and Development, Nairobi, Kenya. [6]Unité Mixte de Recherche (UMR) 208, Institut de Recherche pour le Développement (IRD)–Muséum National d'Histoire Naturelle (MNHN), Musée de l'Homme, Paris, France. [7]Ministère de la Recherche Scientifique et de l'Innovation, Yaoundé, Cameroon. [8]Department of Genetics, Yale University, New Haven, Connecticut, USA. [9]Howard Hughes Medical Institute, University of Washington, Seattle, Washington, USA. [10]Present addresses: Department of Human Genetics, University of Michigan, Ann Arbor, Michigan, USA (J.M.K.), Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan, USA (J.M.K.) and Department of Human Genetics, University of Chicago, Chicago, Illionois, USA (M.P.D.). [11]These authors contributed equally to this work. Correspondence should be addressed to E.E.E. (eee@gs.washington.edu).

Received 4 October 2011; accepted 1 June 2012; published online 1 July 2012; corrected online 11 July 2012 (details online); doi:10.1038/ng.2335

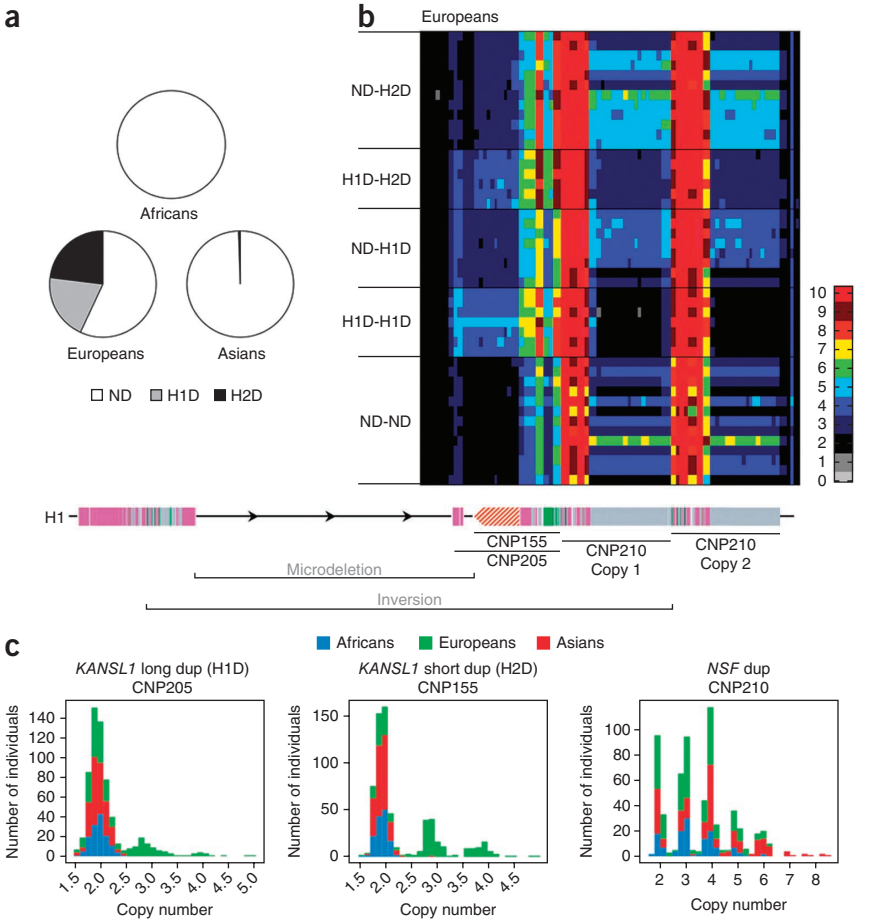

Figure 1 Duplication architecture of 17q21.31. (**a**) Frequency of haplotypes (H2D, H1D) carrying duplications (CNP155 and CNP205) and those not carrying duplications (ND) are shown for three major continental groups (Africans, Asians and Europeans) on the basis of analysis of 620 individuals. (**b**) Read depth–based copy-number estimates of the 17q21.31 region from 46 representative European genomes show different patterns of duplication for the *KANSL1* and *NSF* regions. Colors indicate the absolute copy number across the genome for each given segment[15]. The heatmap is aligned to the H1 haplotype structure from the reference genome (bottom), with colored boxes indicating segmental duplications[13] and black lines representing single-copy regions. The heatmap distinguishes genotypes for CNP205 associated with H2D, CNP155 associated with H1D and CNP210, which ranges from 2–8 copies. Note that CNP205 and CNP155 have a diploid copy number of two in the reference genome assembly, whereas CNP210 has a diploid copy number of four. (**c**) Population stratification of duplicated alleles. Dup, duplication.

previously described[15] (**Supplementary Tables 1** and **2**). The region consists of three large copy-number polymorphic (CNP) segmental duplications (**Fig. 1**), which include short (155-kb) and long (205-kb) duplications corresponding to the promoter and first exon of *KANSL1* (previously called *KIAA1267*) associated with the H2 and H1 haplotypes, respectively. For simplicity, we refer to these duplications as CNP155 and CNP205, respectively. We found that almost 60% of Europeans carry at least one of these duplications (**Fig. 1a**,**b**); however, they are virtually nonexistent in African and Asian populations (**Fig. 1c**). The third polymorphism is 210 kb in length and spans most of the *NSF* gene upstream of *KANSL1* (CNP210)[15]. Asian populations show higher copy number of CNP210 compared to European and African populations. In fact, individuals with four haploid copies of this duplication—an estimated 800 kb of tandem repeats—are exclusively of Asian descent (**Fig. 1c**).

Alternative structural configurations of 17q21.31

To further investigate the genomic organization of the region, we performed FISH experiments on 30 individuals of diverse ancestry (**Fig. 2a** and **Supplementary Table 3**). In this sample, the region on chromosome 17 was inverted on 35 of the 60 chromosomes analyzed, and the results from all 35 were concordant with those from the PCR assay diagnostic for the H2 haplotype inversion (**Supplementary Fig. 1**)[18]. We did not observe any non-inverted H2 chromosomes, in contrast to what has recently been reported[19]. FISH and aCGH experiments confirmed three haploid copy-number states for the *KANSL1* locus (copy number = 1, 2 or 3) and three haploid copy-number states for the *NSF* locus (copy number = 1, 2 or 3). We analyzed 21 samples of African descent (**Supplementary Fig. 2** and **Supplementary Table 4**) and found that the H2-specific duplication (CNP155) was highly polymorphic among Africans.

Using two constructed BAC-based assemblies corresponding to one direct and one inverted haplotype[13] together with read depth–based copy-number estimates, BAC pool sequencing (F.A., P.H.S., J. Kitzman, C.T. Amemiya and E.E.E., unpublished data) and FISH, we characterized at least eight alternate structural configurations of the 17q21.31 region, which are markedly different in their organization and duplication content (**Fig. 2b**). Four main structural haplotypes can be defined on the basis of inversion status and the copy-number status (CNP205 and CNP155) of *KANSL1* duplications: H1′ (direct) and H2′ (inverted) carry no duplications of *KANSL1*, H1D (direct) has two copies of CNP205, and H2D (inverted) has two copies of CNP155. We further identified configurations with three copies of CNP205 as H1D.3, whereas H1′ configurations with no *NSF* duplications were defined as H1.1, with two copies of CNP210 as H1.2 and with three copies as H1.3. Similarly, H2′ configurations with one copy of *NSF* were defined as H2.1 and with two copies as H2.2. Notably, all H1D haplotypes analyzed showed a fixed copy number of 1 for *NSF*, and all H2D haplotypes had a fixed copy number of 2 (CNP210). The H2.1 haplotype was among the simplest, carrying single copies of all CNPs, including CNP210. The *NSF* CNP was highly variable among H1′ haplotypes, ranging from 2–8 diploid copies. FISH experiments using a probe mapping to both CNP155 and CNP205 and a probe mapping uniquely to CNP205 showed that the proximal breakpoints of the duplications were different and that the duplications mapped to different locations (tandem duplication on H1D and interspersed on H2D, respectively), strongly suggesting that the two duplications occurred in independent events (**Supplementary Fig. 3**). An independent, parallel study by Boettger *et al.*[20] reports multiple distinct haplotypes also defined on the basis of duplication content and organization (**Supplementary Table 5**).

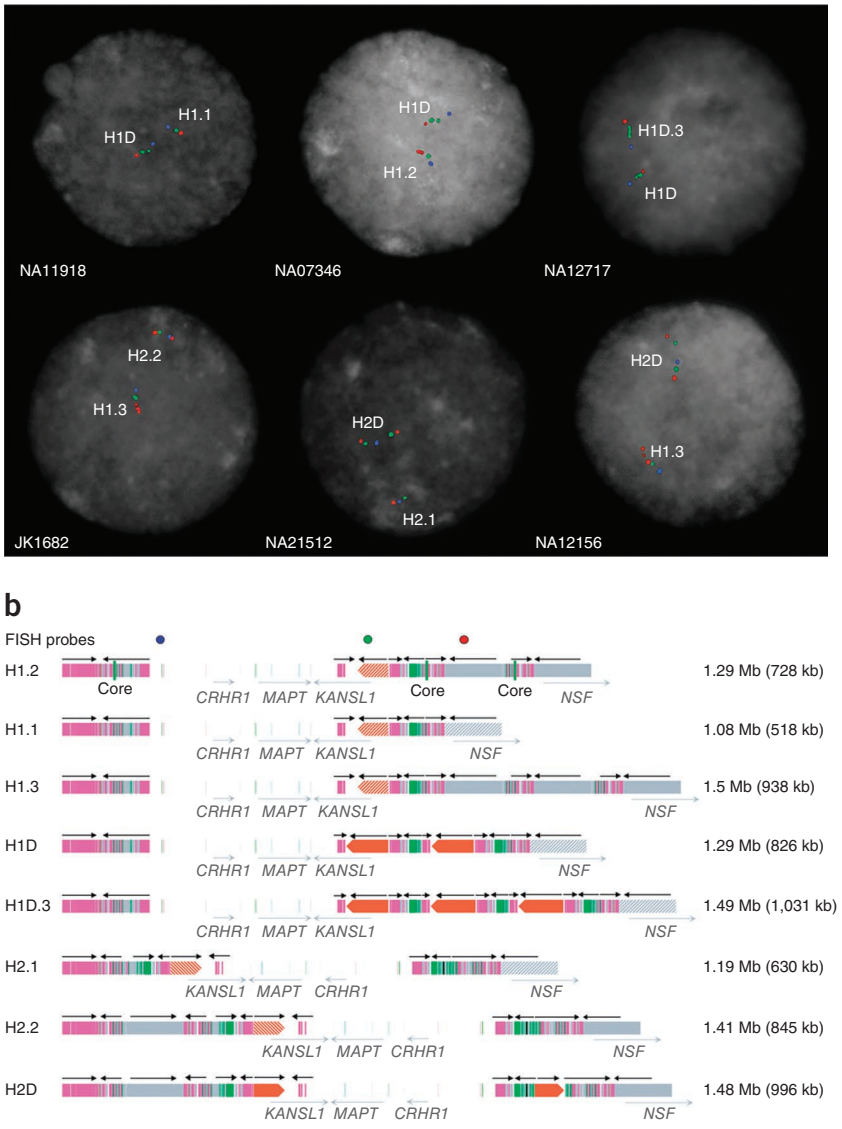

Figure 2 Alternative structural haplotypes of 17q21.31. (**a**) FISH co-hybridization experiments using probes mapping to CNP155 and CNP205 (WIBR2-2342H02, green), *NSF* CNP210 duplication (WIBR2-1321L07, red) and the single-copy region (WIBR2-3237D21 in blue). (**b**) Eight distinct structural haplotypes (five H1 and three H2) ranging in size from 1.08–1.49 Mb. Colored boxes indicate segmental duplications, as determined by complete sequencing of large-insert BAC clones[13]. Hashed boxes correspond to regions present in single copy in that specific haplotype but duplicated in others. Dots, FISH probes; arrows, direction of repeat. The locations of three core duplicons are shown. These represent some of the most abundant and rapidly evolving duplicated sequences in the human genome[50]. The duplication content for each haplotype is indicated in parentheses. Four main haplotypes are defined on the basis of *KANSL1* copy number and the length of the duplication (Boettger *et al.*[20] nomenclature in parentheses): H1′ (direct) and H2′ (inverted) with one copy each of *KANSL1*, H1D (H1.β2.γ1) with a long duplication of the gene and H2D (H2.α2.γ2) with a short duplication. H1′ configurations with one copy of *NSF* are defined as H1.1 (H1.β1.γ1), with two copies as H1.2 (H1.β1.γ2) and with three copies as H1.3 (H1.β1.γ3). H1D configurations with three copies of the long duplication are defined as H1D.3 (H1.β3.γ1). Similarly, H2′ configurations with one copy of *NSF* are defined as H2.1 and with two copies as H2.2 (H2.α1.γ2).

Inversion and duplication frequencies in Africa

Previous studies of this locus in African populations have suggested that the inverted haplotype was rare or nonexistent in most of Africa. Diversity sample surveys, however, have been biased toward populations primarily of western or southern African descent. We sought to explore diversity more systematically by analyzing genetic data from a larger collection of African samples, including from the HapMap Project, the 1000 Genomes Project, the Human Genome Diversity Project (HGDP), the African Diversity Panel (S.A.T., unpublished data), the Hunter-Gatherer Panel[21] and the Bushman Panel[17] (**Supplementary Fig. 4** and **Supplementary Tables 6** and **7**). We used previously published SNPs mapping to the inversion[10,16] and our copy-number estimates in combination with publicly available phased SNP data to identify additional inversion- and haplotype-specific duplication-tagging SNPs (**Supplementary Table 8**).

We were able to accurately type 818 African individuals from 23 diverse ancestry groups for the H1, H2′ and H2D haplotypes. We tested the H2 orientation of nine samples for which cell lines were available and confirmed the presence of the inverted orientation in all these individuals (**Supplementary Table 3**). We were unable to identify common SNPs that could distinguish H1′ from H1D haplotypes. We found that the H2 haplotypes are almost absent in western African populations but are much more prevalent in eastern African populations (**Fig. 3**) than was originally estimated[10,16]. For example, we examined 286 Maasai individuals and found the H2 haplotype frequency to be approximately 7%, although the H2 haplotype was thought to be absent from this population[16] (**Table 1**). The highest inversion frequency was 13%, found in the Beja from Sudan—a population that has experienced substantial gene flow from Middle Eastern populations[16] where the frequency of H2 haplotypes is enriched. The inversion was nonexistent in populations speaking Niger-Kordofanian languages, with the major exception of the Biaka and Bakola Pygmies, in whom the inversion was found at a frequency of ~5%. The H2′ haplotype was primarily found in hunter-gatherer populations—the San, Hadza, Bakola, Biaka, Mbuti and Sengwer. The highest frequency of H2D was found in the Boni, Maasai and Sandawe.

Evolutionary age of duplication events

To understand the evolutionary history of this region without the complication of recombination, we used the alignment of phased SNPs from a 136-kb LD block within the inversion interval to build a maximum-likelihood tree for each haplotype and ancestry group in the HapMap Project (**Fig. 4a**). On the basis of this analysis, we made four important observations. First, there was strong bootstrap

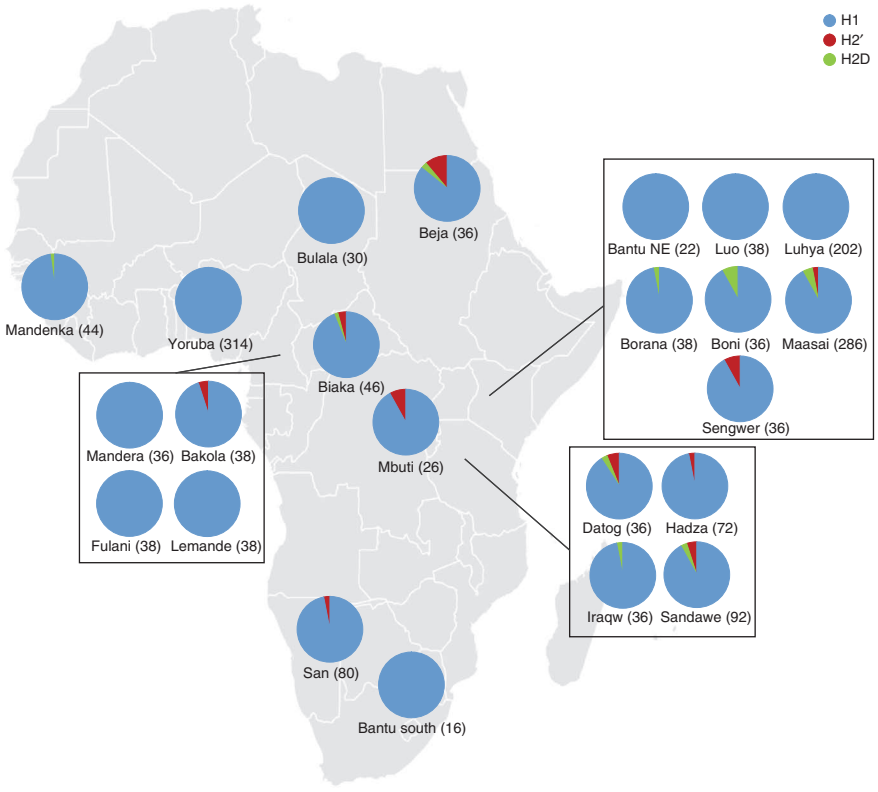

Figure 3 Haplotype frequency of 17q21.31 inversion in Africa. Frequency of direct (H1), inverted (H2′) and inverted with duplication (H2D) haplotypes in 818 individuals (1,636 chromosomes) from 23 African populations. The H2′ haplotype was absent from virtually all western African individuals except for the Pygmy populations (Bakola, Biaka and Mbuti). The H2′ haplotype frequency was highest in the Beja from Sudan, probably due to admixture from neighboring Middle Eastern countries. The inversion was also at appreciable frequencies in other hunter-gatherer populations (San, Hadza, Sandawe, Boni and Sengwer). NE, northeast.

support indicating that the H1 and H2 haplotype clades are completely distinct. Second, there was strong support for the hypothesis that the H1′ haplotype is ancestral to the H1D haplotype. Third, the H2 and H2D haplotypes showed little to no variation. Finally, the analysis strongly suggested that the H1- and H2-specific duplications of the *KANSL1* locus were separate, derived events.

To overcome SNP ascertainment biases, we obtained complete genomic sequence from the representative haplotypes, including all possible single-nucleotide variants (SNVs). We sequenced H2D haplotype–resolved fosmids derived from a European individual (NA12156) carrying this duplication[22], analyzed a European H2′ homozygote (NA20589) and a Maasai H2D homozygote (NA21599) using NGS and used publicly available sequence from a San Bushman (KB1)[17] carrying the H2′ haplotype and an H1D homozygous individual (NA12878) from the 1000 Genomes Project. We also included previously published assemblies of the H1′ and H2D haplotypes from the RP11 BAC assemblies[13] and used these references to assist in sequence alignment for the other genomes.

We constructed an unrooted neighbor-joining tree (**Fig. 4b**) using Kimura two- parameter distance estimates based on sequence alignments to the unique 204,447-bp portion of the inversion region. Consistent with previous analyses, we estimated that the H1 and H2 haplotypes coalesced approximately 2.3 million years ago. There was, however, a notable dearth of genetic diversity on the H2 lineage (**Table 2**). We expect nucleotide diversity (π) between any two chromosomes from a constant population size that is evolving neutrally to be approximately 0.001 (ref. 23). For the H1 lineage, π was equal to 0.00047 but was nearly four times lower for the H2 lineage (π = 0.00012). Although our sample size was small, we note that π was lowest for the H2D haplotype in comparison to the H2′ haplotype (0.00004 versus 0.00025). We observed virtually no sequence differences in the inversion region between the genomic sequences of individuals with the H2D haplotype. This is unlikely to represent cryptic ancestry between these individuals, as whole-genome comparison of RP11 and NA12156 suggests an average heterozygosity of 0.000943 ± 0.000597. The topology of the tree, as well as the lack of diversity on the H2 haplotypes, is suggestive of a recent bottleneck followed by population expansion or selective sweep.

We estimated the coalescent time of the H2 and H2D haplotypes in African populations to be approximately 136,000 ± 19,000 years ago and the coalescent time between African and European H2 haplotypes to be 48,000 ± 11,000 years ago (**Supplementary Table 9**). The European H2′ and H2D haplotypes were more similar than the African and European H2D haplotypes, suggesting that the European H2′ haplotype has possibly undergone homogenization with the more predominant H2D haplotypes. The H1′ and H1D haplotypes had a much older date of coalescence of approximately 250,000 ± 26,000 years ago, consistent with published data[10]. We present these dates with the caveat that they represent an average coalescent time over the entire interval, given that H1′ and H1D haplotypes and H2′ and H2D haplotypes can freely recombine, and these segments may represent sequences from multiple common ancestors.

We compared the sequence in the duplicated regions for each haplotype clade to obtain a more accurate evolutionary age for the duplication events that was not biased by recombination events across the inversion interval. We aligned the NA12878 H1D sequence to the RP11 H1′ sequence at CNP205 to estimate the age of the H1 duplication, which we found to have occurred 247,000 ± 20,000 years ago. We repeated this analysis for the H2D sequence (CNP155) from NA21599, aligning it to the reference H2D sequence from RP11, and estimated that this duplication occurred 1.3 million ± 106,000 years ago. This finding is consistent with the range of values estimated for the duplication in previous analyses[10,13]. We note that the H2 duplication was much older than coalescence determined using the unique sequence of the sampled H2 haplotypes, which was estimated to have occurred from approximately 48,000 to 136,000 years ago, depending on the pairs of haplotypes chosen for analysis and the segment of DNA analyzed. These more recent coalescent times for H2 haplotypes are consistent with the recent TMRCA of H2 haplotypes observed in a previous study[16] that analyzed short tandem repeat polymorphisms within the inversion interval. These discrepancies are noteworthy and suggest that selection on the H2 haplotype resulted in the recent coalescence of extant H2 haplotypes.

Table 1 Frequencies of H1′, H2′ and H2D in 23 diverse African ancestry groups

Population	Country	Number of individuals	Frequency of H1′ (%)	Frequency of H2′ (%)	Frequency of H2D (%)	Subsistence pattern	Language family	Language major subgrouping	Reference
Bakola	Cameroon	19	94.74	5.26	0	Hunter-gatherer	Niger-Kordofanian	Bantoid	African Diversity Panel
Bantu-northeast	Kenya	11	100	0	0	Farmer	Niger-Kordofanian	Bantoid	HGDP
Bantu-south	South Africa	8	100	0	0	Farmer	Niger-Kordofanian	Bantoid	HGDP
Beja	Sudan	18	86.11	11.11	2.78	Herder	Afroasiatic	Cushitic	African
Biaka	Central African Republic	23	93.48	4.35	2.17	Hunter-gatherer	Niger-Kordofanian	Adamawa-Ubangi	HGDP, H2 Diversity Panel
Boni	Kenya	18	92.11	0	7.89	Hunter-gatherer	Afroasiatic	Cushitic	African Diversity Panel
Borana	Kenya	19	97.37	0	2.63	Herder	Afroasiatic	Cushitic	African Diversity Panel
Bulala	Chad	15	100	0	0	Farmer	Nilo-Saharan	Central Sudanic	African Diversity Panel
Datog	Tanzania	18	91.67	5.56	2.78	Herder	Nilo-Saharan	Eastern Sudanic	African Diversity Panel
Fulani	Cameroon	19	100	0	0	Herder	Niger-Kordofanian	Senegambian	African Diversity Panel
Hadza	Tanzania	36	97.22	2.78	0	Hunter-gatherer	Khoesan	Hadza	African Diversity Panel, Hunter-Gatherer
Iraqw	Tanzania	18	97.22	0	2.78	Mixed farmer	Afroasiatic	Cushitic	African Diversity Panel
Lemande	Cameroon	19	100	0	0	Farmer	Niger-Kordofanian	Bantoid	African Diversity Panel
Luhya	Kenya	101	100	0	0	Farmer	Niger-Kordofanian	Bantoid	HapMap, 1000 Genomes Project
Luo	Kenya	19	100	0	0	Herder	Niger-Kordofanian	Bantoid	African Diversity Panel
Maasai	Kenya	143	92.66	2.45	4.90	Farmer	Nilo-Saharan	Eastern Sudanic	HapMap
Mandenka	Senegal	22	97.73	0	2.27	Herder	Niger-Kordofanian	Mande	HGDP
Mandera	Cameroon	18	100	0	0	Farmer	Niger-Kordofanian	Mande	African Diversity Panel
Mbuti	Democratic Republic of Congo	13	92.31	7.69	0	Hunter-gatherer	Nilo-Saharan	Central Sudanic	African Diversity Panel, HGDP
San	Namibia, South Africa	40	97.50	2.50	0	Hunter-gatherer	Khoesan	Southern	HGDP, Bushman Collection, Hunter-Gatherer
Sandawe	Tanzania	46	91.30	5.43	3.26	Hunter-gatherer	Khoesan	Sandawe	African, Hunter-Gatherer Panel
Sengwer	Kenya	18	91.67	8.33	0	Hunter-gatherer	Nilo-Saharan	Eastern Sudanic	African Diversity Panel
Yoruba	Nigeria	157	100	0	0	Farmer	Niger-Kordofanian	Defoid	African Diversity Panel, HGDP, HapMap, 1000 Genomes Project

H1 and H2 haplotype exchange

In general, inversions are predicted to result in complete suppression of recombination; therefore, sequence divergence is expected to be higher than in freely recombining chromosomal segments. We examined the sequence divergence between the H1′ and H2D haplotype sequences from the RP11 BAC assembly. The average value of π between the two haplotypes was 0.00416; however, we identified a 30-kb stretch of sequence over which the average value of π was 0.0005. This level of divergence was significantly different from the distribution of nucleotide diversity over the entire inversion interval (Kolmogorov-Smirnov $D = 0.9345$; $P = 0$) (**Fig. 5a**). The region of relatively high sequence identity overlapped the 5′ region of the *CRHR1* gene (encoding corticotropin-releasing hormone receptor 1), including the promoter and first two exons. *CRHR1* is involved in anxiety-related behavior and stress adaptation[24–28].

To study the history of this region in greater detail, we constructed a series of median-joining haplotype networks (see URLs) using

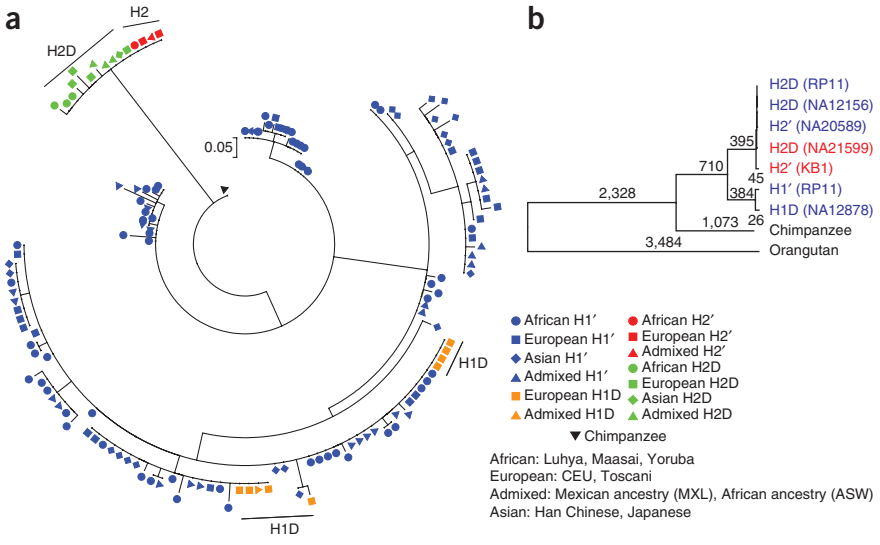

Figure 4 Phylogenetic relationship between H1 and H2 haplotypes. (**a**) Alignment of 43 SNPs from a 136-kb LD block within the inversion region (chr. 17: 41,466,118–41,602,794, NCBI Build 36) from HapMap individuals (total $n = 728$ individuals, 1,456 chromosomes) were used to build a maximum-likelihood tree with 1,000 bootstrap replicates (all branches with 100% bootstrap support). (**b**) An unrooted neighbor-joining tree was constructed using MEGA4 complete deletion option[51] on the basis of 204,447 aligned base pairs from unique sequence within the inversion. The number of mutations for each branch is indicated above the branch. Red, Africans; blue, Europeans.

Table 2 Nucleotide diversity between haplotype groups

Population	*n*	Nucleotide diversity (π)
All H1	2	0.00047
All H2	5	0.00012
H2D	3	0.00004
H2′	2	0.00025
Human (all haplotypes)	7	0.00207
Nonhuman primates[a]	2	0.03281

[a]Nonhuman primates include chimpanzee and orangutan.

HapMap phase 3 SNPs for 728 unrelated individuals in the region of reduced diversity, as well as proximal and distal loci for comparison (**Supplementary Note**). Over the proximal and distal intervals, the H1- and H2-containing chromosomes were cleanly divided into distinct haplotype clades (**Fig. 5b**,**c**). In contrast, over the homogenized *CRHR1* region, we found that 15 H2′- and 123 H2D-containing chromosomes as well as 197 H1′- and 27 H1D-containing chromosomes grouped together in a single haplogroup (**Fig. 5d**). Thus, over the 5′ segment of *CRHR1*, some H1 haplotypes have a sequence unusually similar to that found in H2 haplotypes. As this region is too large for a gene conversion event, this likely represents a historical double recombination event between the H1 and H2 haplotypes. This haplogroup configuration was found in all major continental groupings of HapMap, suggesting that the double recombination event predated the dispersal of modern humans out of Africa.

DISCUSSION

On the basis of our survey of structural genetic diversity from 2,700 diverse population samples, we conclude that the H1- and H2-specific duplications evolved independently and that an absence of duplication was ancestral in both the H1 and H2 lineages. We have resolved eight distinct structural haplotypes that vary in size from 1.08–1.49 Mb. Five of these haplotypes belong to the H1 lineage, whereas three belong to the H2 lineage. The least complex haplotype with regard to duplication architecture was the H2.1 haplotype, which is consistent with the H2 haplotype reported for the San Bushman. European and Mediterranean populations showed a marked enrichment of duplicated haplotypes (60% frequency) compared to any other worldwide population group. In comparison to the inversion, these duplications showed greater population stratification.

These population differences have important implications for disease with respect to the 17q21.31 microdeletion syndrome[11,14,29]. Because the H2D haplotype is the only one out of eight possible configurations with homologous segmental duplications in direct orientation flanking the disease-critical region, only carriers of the H2D haplotype are predisposed to the 17q21.31 microdeletion through NAHR. Thus, European populations are at much higher risk for this syndrome than are Asians and Africans. The H2′ inversion haplotype (enriched among Africans and southern Europeans) does not carry the predisposing duplication, and, therefore, populations with this haplotype are not at risk for this recurrent deletion. We found that 97% of the 17q21.31 microdeletion syndrome cases reported in the literature occurred in individuals of European descent[30]. A screen of 1,084 samples from African-American individuals with developmental delay by TaqMan assay found no occurrence of the 17q21.31 microdeletion[31]. The only known African-American 17q21.31 microdeletion reported[30] had breakpoints mapping outside of the segmental duplications and, thus, occurred by a mechanism other than NAHR.

Our analyses show that either the H1′ or H2′ haplotype is ancestral; however, combined with previous analyses, the results presented here favor H2′ as the ancestral haplotype of the genus *Homo*. First, 90% of the SNPs that are monomorphic in H2 haplotypes but polymorphic in H1 haplotypes matched the chimpanzee allele, but only 60% of SNPs that are monomorphic in H1 haplotypes and polymorphic in H2 haplotypes matched the chimpanzee allele[13]. Second, the idea that the inverted configuration is the ancestral state is supported by the results of an analysis of Old World and New World monkeys[13]. Finally, our phylogenetic and coalescent analyses provide strong evidence for an African origin of the H2 haplotype. We found the H2′ haplotype among populations thought to map near the root of human phylogeny, such as the San Bushman and other hunter-gatherers, including the click-speaking populations of Tanzania and Pygmies. Previous studies suggest an ancient genetic affinity and shared ancestry among these groups[21,32]. Additionally, our analyses indicate that the San H2′ haplotype has an older evolutionary age than the African H2D and European H2′ haplotypes.

Despite its high frequency among European populations, the H2 haplotypes showed extraordinary homogeneity. The ancient coalescence of H1 and H2 and the excess of rare polymorphisms in H2 haplotypes indicate a recent bottleneck or selective sweep, particularly in the European H2D lineage, where nucleotide diversity was found to be the lowest. Recent analyses support the original observation that the H2 haplotype is associated with increased mean rates of recombination in females[33]. It is also known that females with increased mean rates of recombination have more offspring[34,35], strengthening the evidence for a selective advantage for H2D carriers. This observation remains intriguing, as the duplication architecture associated with H2D carriers clearly predisposes to microdeletion[14] and must therefore be subjected to weak purifying selection.

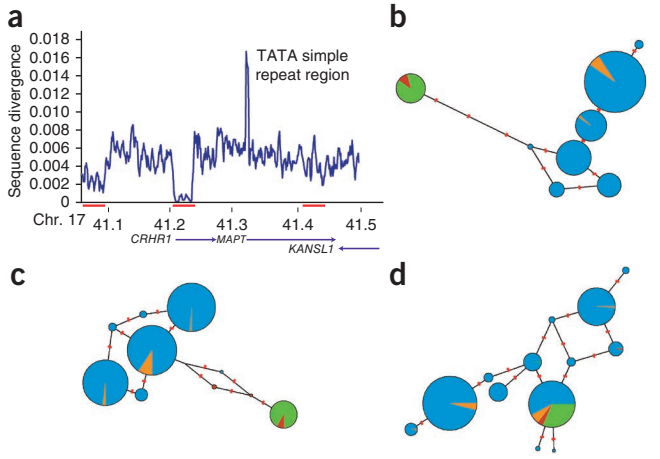

Figure 5 Historical exchange between H1 and H2 haplotypes. (**a**) Divergence plotted in 5-kb sliding windows. A 30-kb region (chr. 17: 41,213,364–41,248,960, middle red bar) of reduced divergence over the 5′ end of *CRHR1* is shown. The spike in divergence at 41.35 Mb corresponds to a simple TATA repeat tract. (**b**–**d**) Median-joining haplotype networks based on data from the HapMap collection for the region proximal to the 5′ end of *CRHR1* (chr. 17: 41,011,056–41,091,056; left red bar in **a**) (**b**), the region distal to the region of reduced divergence (chr. 17: 41,410,073–41,425,073; right red bar in **a**) (**c**) and the region of reduced divergence (**d**). The proportion of H1′ (blue), H1D (orange), H2′ (red) and H2D (green) haplotypes are shown. The haplotypes form distinct clades proximal and distal to the *CRHR1* region, whereas over the region of reduced divergence, the haplotypes are mixed, creating a large haplogroup where H1 and H2 chromosomes have similar sequence. Red tick marks represent the number of mutations separating each haplogroup.

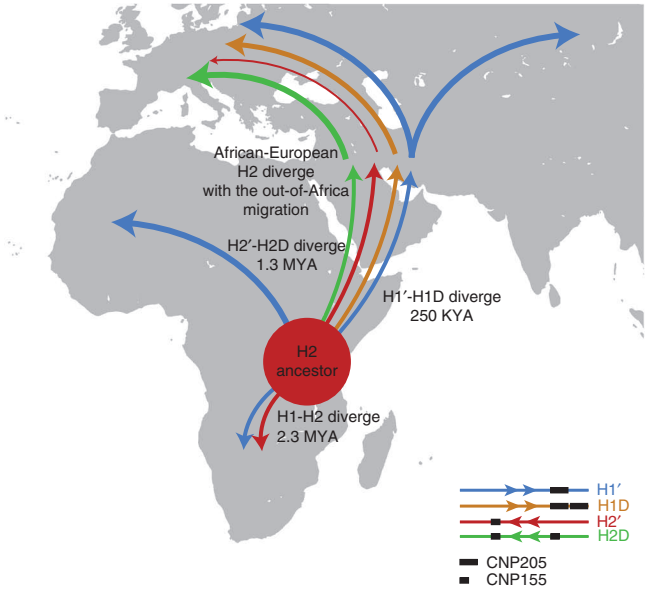

Figure 6 Evolutionary history of 17q21.31 haplotypes. We propose a model where the H2′ haplotype represents the ancestral configuration of the 17q21.31 region in humans. Approximately 2.3 million years ago (MYA), the inversion toggled back to the direct orientation and spread to southern Africa before the emergence of modern humans. The H2D duplication arose in Africa 1.3 MYA, and the H1D duplication independently arose much more recently, approximately 250,000 years ago (250 KYA). The H1′ haplotype spread throughout western Africa, and all haplotypes spread to the Middle East and Europe as part of the out-of-Africa migration.

Among the 887 informative SNPs distinguishing the H1 and H2 haplotypes, there are 9 missense, 9 synonymous and 22 UTR mutations on the H2 haplotype (**Supplementary Table 10).** Seven of the nine missense mutations are in *IMP5* (encoding intramembrane protease 5), and four of these are predicted to alter the protein structure by PolyPhen[36]. The H1 alleles containing two of the four mutations predicted to alter amino acids have previously been associated with Parkinson's disease[37], and the missense mutations on the H2 haplotype may therefore represent nucleotide substitutions that are under positive selection. Notably, the H2 haplotypes carry the derived allele for these four SNPs, whereas the other SNPs that are not predicted to alter the protein retain the ancestral allele with respect to the chimpanzee. The H2 haplotype has potentially changed functionally, while the H1 haplotype has kept a structure more similar to that found in chimpanzee, lending some support to the hypothesis of selection on the H2 haplotype. The *IMP5* gene is predicted to be under purifying selection using maximum-likelihood analyses of eight mammalian species (data not shown), although nonsynonymous mutations are not uncommon during evolution. The two remaining missense alterations are in *KANSL1*—a gene that, when mutated, results in a phenotype similar to the 17q21.31 microdeletion syndrome[38,39]. Both variants are not predicted to alter the protein structure by PolyPhen but are highly conserved (genomic evolutionary rate profiling (GERP) scores between 2 and 4).

It is also not clear why the H1D haplotypes have risen to such high frequencies in European populations. Given that certain H1 haplotypes are associated with neurological disorders, such as Parkinson's disease[40,41], Alzheimer's disease[42] and progressive supranuclear palsy[18,43,44], we examined association between the duplication and these disease-predisposing haplotypes in individuals from the 1000 Genomes Project (**Supplementary Table 11**). We found that the duplication was present less often than predicted by linkage equilibrium for the H1c haplotype ($P < 0.0001$), which is a risk haplotype for Alzheimer's disease and progressive supranuclear palsy[42,43]. The duplication was extremely rare on Parkinson's disease risk haplotypes[40,43,45]; however, given the sample size and low frequencies of the Parkinson's disease risk haplotypes in the population, we were unable to assess whether these associations were significant. Nevertheless, these observations suggest that the disease risk haplotypes probably arose on H1′ haplotypes, indicating that further investigation into the possible protective role of the duplication in these diseases is warranted. The H2 haplotype, which almost always bears the duplication in Europeans, is protective against many of these diseases[40,45] but predisposes to microdeletion associated with intellectual disability.

If there is a selective advantage to both the H1D and H2D haplotypes, one possible reason for this may be the recurrent duplications involving both the promoter and first exon of *KANSL1*. *KANSL1* encodes a chromatin modifier that is thought to have a role in complex brain function and has been associated with the 17q21.31 microdeletion syndrome[11,38,39,46]. We note that *KANSL1* gene expression is increased in the brains of individuals with Parkinson's disease compared to controls[40], suggesting that dysregulation of *KANSL1* expression may have phenotypic consequences of disease relevance.

Another notable observation was the striking absence of genetic diversity within a 30-kb stretch of *CRHR1* between H1 and H2 haplotypes despite the deep evolutionary divergence of the haplotypes 2.3 million years ago. The observed decrease in divergence at the *CRHR1* locus is reminiscent of diversity patterns observed among some highly divergent human leukocyte antigen (HLA) haplotypes—a group of haplotypes that, like the 17q21.31 interval, are otherwise characterized by high sequence diversity and reduced recombination between divergent clades[47]. It seems implausible that this 30-kb stretch has been maintained at such a high degree of sequence identity relative to the rest of the region since the coalescence of the two haplotypes. We propose that the observed pattern results from a classical double crossover event via an inverted loop structure during meiosis, resulting in the transfer of sequence between these two haplotypes sometime after their initial separation. We can find no evidence of the reciprocal event in sequence data, suggesting that it may have been lost from the human population.

In conclusion, we propose that the ancestral H2′ haplotype arose in eastern or central Africa and spread to southern Africa before the emergence of anatomically modern humans (**Fig. 6**). Approximately 2.3 million years ago, the inversion rearranged to what we now refer as the direct orientation haplotype (H1′). This haplotype spread throughout the *Homo* ancestral populations in the African continent, virtually replacing the H2′ haplotype and becoming the predominant haplotype. We note that both the Denisova and Neandertal sister groups are predicted to have H1′ haplotypes[48,49]. These early haplotypes were much simpler in their duplication architecture, similar to the patterns seen in great apes. We find that the more complex duplication architectures are particularly enriched in populations that migrated out of Africa. On the basis of sequence at the duplication loci, we estimate that the H2-specific duplication event occurred approximately 1.3 million years ago. Independent of the H2 duplication, the H1-specific duplication event occurred much more recently, approximately 250,000 years ago. Notably, we did not observe this haplotype in any of the African or Asian populations studied, suggesting that it may have been lost in these groups as a result of genetic drift. The H2D haplotype has risen to frequencies of 10–25% in European populations with virtually no genetic variation, suggesting an extremely recent and rapid expansion of this haplotype. High-coverage sequencing

of more individuals along with fecundity data will likely shed further light on whether the high frequency of the haplotype-specific duplication in Europeans is due to selection or the effects of demographic history specific to this locus.

URLs. Network 4.610, http://www.fluxus-engineering.com/; HapMap Phase 3, http://hapmap.ncbi.nlm.nih.gov/; Human Genome Diversity Project (HGDP), http://www.cephb.fr/en/hgdp/; 1000 Genomes Project, http://www.1000genomes.org/; SNP data for Hunter-Gatherer Panel, http://www-evo.stanford.edu/repository/paper0002/; San Bushman (KB1) sequence data, ftp://ftp.bx.psu.edu/data/bushman.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. All sequence data have been submitted to the Short Read Archive (SRA) under accession SRA046964.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

We thank J. Akey, M. Dennis and B. Dumont for helpful discussions and C. Alkan for computational assistance. We thank Z. Jiang for his initial work on the H1-H2 alignments. We are grateful to T. Brown for assistance with manuscript preparation, to C. Lee for technical assistance and to the anonymous reviewers of this paper who provided insightful comments. We thank the 1000 Genomes Project Consortium for access to unpublished sequence data for the 17q21.31 locus. K.M.S. was supported by a Ruth L. Kirschstein National Research Service Award (NRSA) training grant to the University of Washington (T32HG00035) and an individual NRSA Fellowship (F32GM097807). C.D.C. was supported by an individual NRSA Fellowship (F32HG006070). P.H.S. was supported by a Natural Sciences and Engineering Research Council of Canada Fellowship. J.M.K. was supported by a Ruth L. Kirschstein NRSA training grant to Stanford University (T32HG000044). This work was supported by the US National Institutes of Health (grants HG002385 and HG004120 to E.E.E.). E.E.E. is an Investigator of the Howard Hughes Medical Institute.

AUTHOR CONTRIBUTIONS

K.M.S., F.A. and E.E.E. designed the study. K.M.S. performed aCGH, genotyping and sequence analysis. F.A. performed FISH experiments and fosmid shotgun sequencing library construction. P.H.S. performed read depth–based copy-number analysis. J.M.K. performed sequence analysis on the double recombination region. C.D.C. performed aCGH analysis. L.V. and M.M. performed whole-genome shotgun sequencing library construction and PCR genotyping. L.S. and W.B. performed PCR genotyping and SNP array genotyping. M.I., G.L., T.B.N., S.A.O., J.-M.B. and A.F. contributed to African sample collection. M.P.D. and K.K.K. contributed to H2 Diversity Panel sample collection and genotyping. S.A.T. contributed to African sample collection and SNP array data. K.M.S., F.A., J.M.K., S.A.T. and E.E.E. contributed to data interpretation. K.M.S., F.A. and E.E.E. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

Published online at http://www.nature.com/doifinder/10.1038/ng.2335.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Dobzhansky, T. The genetics of natural populations. *Genetics* **35**, 288–302 (1950).
2. Dobzhansky, T. & Sturtevant, A.H. Inversions in the chromosomes of *Drosophila pseudoobscura*. *Genetics* **23**, 28–64 (1938).
3. Lowry, D.B. & Willis, J.H. A widespread chromosomal inversion polymorphism contributes to a major life-history transition, local adaptation, and reproductive isolation. *PLoS Biol.* **8**, e1000500 (2010).
4. Tuzun, E. *et al.* Fine-scale structural variation of the human genome. *Nat. Genet.* **37**, 727–732 (2005).
5. Kidd, J.M. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56–64 (2008).
6. Bailey, J.A. *et al.* Recent segmental duplications in the human genome. *Science* **297**, 1003–1007 (2002).
7. Sharp, A.J. Emerging themes and new challenges in defining the role of structural variation in human disease. *Hum. Mutat.* **30**, 135–144 (2009).
8. Lupski, J.R. Genome structural variation and sporadic disease traits. *Nat. Genet.* **38**, 974–976 (2006).
9. Antonacci, F. *et al.* Characterization of six human disease-associated inversion polymorphisms. *Hum. Mol. Genet.* **18**, 2555–2566 (2009).
10. Stefansson, H. *et al.* A common inversion under selection in Europeans. *Nat. Genet.* **37**, 129–137 (2005).
11. 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).
12. Koolen, D.A. *et al.* A new chromosome 17q21.31 microdeletion syndrome associated with a common inversion polymorphism. *Nat. Genet.* **38**, 999–1001 (2006).
13. Zody, M.C. *et al.* Evolutionary toggling of the *MAPT* 17q21.31 inversion region. *Nat. Genet.* **40**, 1076–1083 (2008).
14. Koolen, D.A. *et al.* Clinical and molecular delineation of the 17q21.31 microdeletion syndrome. *J. Med. Genet.* **45**, 710–720 (2008).
15. Sudmant, P.H. *et al.* Diversity of human copy number variation and multicopy genes. *Science* **330**, 641–646 (2010).
16. Donnelly, M.P. *et al.* The distribution and most recent common ancestor of the 17q21 inversion in humans. *Am. J. Hum. Genet.* **86**, 161–171 (2010).
17. Schuster, S.C. *et al.* Complete Khoisan and Bantu genomes from southern Africa. *Nature* **463**, 943–947 (2010).
18. Baker, M. *et al.* Association of an extended haplotype in the *tau* gene with progressive supranuclear palsy. *Hum. Mol. Genet.* **8**, 711–715 (1999).
19. Rao, P.N., Li, W., Vissers, L.E., Veltman, J.A. & Ophoff, R.A. Recurrent inversion events at 17q21.31 microdeletion locus are linked to the *MAPT* H2 haplotype. *Cytogenet. Genome Res.* **129**, 275–279 (2010).
20. Boettger, L.M., Handsaker, R.E., Zody, M.C. & McCarroll, S.A. Structural haplotypes and recent evolution of the human 17q21.31 region. *Nat. Genet.* published online, doi:10.1038/ng.2334 (1 July 2012).
21. Henn, B.M. *et al.* Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. *Proc. Natl. Acad. Sci. USA* **108**, 5154–5162 (2011).
22. Kidd, J.M. *et al.* Haplotype sorting using human fosmid clone end-sequence pairs. *Genome Res.* **18**, 2016–2023 (2008).
23. Cargill, M. *et al.* Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.* **22**, 231–238 (1999).
24. Heim, C. *et al.* Effect of childhood trauma on adult depression and neuroendocrine function: sex-specific moderation by CRH receptor 1 gene. *Front. Behav. Neurosci.* **3**, 41 (2009).
25. Liu, Z. *et al.* Association of corticotropin-releasing hormone receptor 1 gene SNP and haplotype with major depression. *Neurosci. Lett.* **404**, 358–362 (2006).
26. Liu, Z. *et al.* Association study of corticotropin-releasing hormone receptor 1 gene polymorphisms and antidepressant response in major depressive disorders. *Neurosci. Lett.* **414**, 155–158 (2007).
27. Bradley, R.G. *et al.* Influence of child abuse on adult depression: moderation by the corticotropin-releasing hormone receptor gene. *Arch. Gen. Psychiatry* **65**, 190–200 (2008).
28. Polanczyk, G. *et al.* Protective effect of *CRHR1* gene variants on the development of adult depression following childhood maltreatment: replication and extension. *Arch. Gen. Psychiatry* **66**, 978–985 (2009).
29. 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).
30. Cooper, G.M. *et al.* A copy number variation morbidity map of developmental delay. *Nat. Genet.* **43**, 838–846 (2011).
31. Mefford, H.C. *et al.* A method for rapid, targeted CNV genotyping identifies rare variants associated with neurocognitive disease. *Genome Res.* **19**, 1579–1585 (2009).
32. Tishkoff, S.A. *et al.* The genetic structure and history of Africans and African Americans. *Science* **324**, 1035–1044 (2009).
33. Fledel-Alon, A. *et al.* Variation in human recombination rates and its genetic determinants. *PLoS ONE* **6**, e20321 (2011).
34. Kong, A. *et al.* Recombination rate and reproductive success in humans. *Nat. Genet.* **36**, 1203–1206 (2004).
35. Coop, G., Wen, X., Ober, C., Pritchard, J.K. & Przeworski, M. High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans. *Science* **319**, 1395–1398 (2008).
36. Ramensky, V., Bork, P. & Sunyaev, S. Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.* **30**, 3894–3900 (2002).
37. Simón-Sánchez, J. *et al.* Genome-wide association study reveals genetic risk underlying Parkinson's disease. *Nat. Genet.* **41**, 1308–1312 (2009).
38. Koolen, D.A. *et al.* Mutations in the chromatin modifier gene *KANSL1* cause the 17q21.31 microdeletion syndrome. *Nat. Genet.* **44**, 639–641 (2012).
39. Zollino, M. *et al.* Mutations in *KANSL1* cause the 17q21.31 microdeletion syndrome phenotype. *Nat. Genet.* **44**, 636–638 (2012).
40. Tobin, J.E. *et al.* Haplotypes and gene expression implicate the *MAPT* region for Parkinson disease: the GenePD Study. *Neurology* **71**, 28–34 (2008).
41. Skipper, L. *et al.* Linkage disequilibrium and association of *MAPT* H1 in Parkinson disease. *Am. J. Hum. Genet.* **75**, 669–677 (2004).

42. Myers, A.J. *et al.* The *MAPT* H1c risk haplotype is associated with increased expression of tau and especially of 4 repeat containing transcripts. *Neurobiol. Dis.* **25**, 561–570 (2007).
43. 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).
44. Höglinger, G.U. *et al.* Identification of common variants influencing risk of the tauopathy progressive supranuclear palsy. *Nat. Genet.* **43**, 699–705 (2011).
45. Setó-Salvia, N. *et al.* Dementia risk in Parkinson disease: disentangling the role of *MAPT* haplotypes. *Arch. Neurol.* **68**, 359–364 (2011).
46. Dubourg, C. *et al.* Clinical and molecular characterization of 17q21.31 microdeletion syndrome in 14 French patients with mental retardation. *Eur. J. Med. Genet.* **54**, 144–151 (2011).
47. Dawkins, R. *et al.* Genomics of the major histocompatibility complex: haplotypes, duplication, retroviruses and disease. *Immunol. Rev.* **167**, 275–304 (1999).
48. Reich, D. *et al.* Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468**, 1053–1060 (2010).
49. Green, R.E. *et al.* A draft sequence of the Neandertal genome. *Science* **328**, 710–722 (2010).
50. Jiang, Z. *et al.* Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution. *Nat. Genet.* **39**, 1361–1368 (2007).
51. 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).

ONLINE METHODS

Samples. Genomic DNA and lymphoblast cell lines from HapMap individuals were obtained from the Coriell Cell Repository. Genomic DNA and lymphoblast cell lines from African individuals were obtained as described[16,32] (**Supplementary Tables 1** and **2** for samples used). Publicly available SNP and sequence data for the 17q21.31 region were downloaded from HapMap Phase 3, HGDP, the 1000 Genomes Project and Stanford University.

aCGH. We designed a custom-targeted microarray (Roche NimbleGen) for the 17q21.31 region (12,468 probes tiled across 1.9 Mb). DNA was labeled using the NimbleGen Dual-Color DNA Labeling Kit (Roche NimbleGen), using NA19240 as a reference sample. Hybridization reactions were performed at 42 °C for 60 h using the NimbleGen Wash Buffer kit as described previously[52]. Scanned images (obtained with a GenePix 4000B Scanner) were analyzed using NimbleScan v2.5, and copy-number variants were called using the segMNT algorithm.

FISH. Metaphase spreads and interphase nuclei were obtained from human lymphoblast cell lines. FISH experiments were performed using fosmid clones (**Supplementary Table 8**) directly labeled by nick translation with Cy3-dUTP (PerkinElmer), Cy5-dUTP (PerkinElmer) and fluorescein-dUTP (Enzo) as described previously[52]. Digital images were obtained using a Leica DMRXA2 epifluorescence microscope equipped with a cooled charge-coupled device (CCD) camera (Princeton Instruments). DAPI (4′,6-diamidino-2-phenylindole), Cy3, Cy5 and fluorescein fluorescence signals, detected with specific filters, were captured separately as grayscale images. Pseudocoloring and merging of images were performed using Adobe Photoshop software. A minimum of 50 interphase nuclei were scored for each inversion to statistically determine the orientation of the examined region.

Illumina sequencing of H2D-haplosorted fosmid clones. Fosmid clones from NA12156 were assigned to haplotypes using previously described methods[22], and DNA was isolated by a modified alkaline lysis miniprep procedure. In this procedure, a cell pellet was resuspended in 250 µl of Qiagen buffer P1 with RNase and lysed with 250 µl of 0.2 M NaOH–1% SDS solution for 5 min. Lysis was neutralized by the addition of 250 µl of 3 M sodium acetate, pH 4.8. Neutralized lysate was incubated on ice for 40 min, and DNA was collected by centrifugation for 15 min at 15.7*g* at 4 °C, concentrated by standard ethanol precipitation and resuspended in 50 µl of 10 mM Tris-HCl, pH 8.5. Libraries were prepared from fosmid clone DNA using Illumina-compatible Nextera DNA sample prep kits (Epicentre, GA09115). The manufacturer's protocol was followed with modifications, including the use of a set of barcoded oligonucleotides as described[53]. Barcoded libraries were combined for size selection using E-Gel SizeSelect 2% (Invitrogen, G6610-02). The band spanning 600–700 bp in size was amplified via limited-cycle PCR with iProof High-Fidelity polymerase (Bio-Rad) with the following program: initial denaturation at 98 °C for 30 s and 6–12 cycles of denaturation at 98 °C for 10 s, annealing at 64 °C for 30 s and extension at 72 °C for 40 s. Amplified, size-selected libraries were then purified with the QIAquick PCR Purification kit (Qiagen, 28104) and quantified on an Invitrogen Qubit Fluorometer, and paired-end sequencing (of 101-bp reads) was performed on an Illumina HiSeq 2000. Sequence reads were aligned to the chr17_ctg5_hap1 reference sequence (GRCh37) using the Burrows-Wheeler Aligner (BWA; version 0.5.9), and variants were called using SAMtools mpileup (version 0.1.16).

Illumina sequencing of H2 and H2D homozygous genomes. Genomic DNA (3 µg) from NA20589 (H2′/H2′) and NA21599 (H2D/H2D) was sheared and end-repaired, an A-tail was added, and adaptors were ligated to the fragments as described[54]. After ligation, samples were run on a 6% precast polyacrylamide gel (Invitrogen, EC6265BOX). The portion of the gel corresponding to the band at 400 bp was excised, diced and incubated as described above. Size-selected fragments were amplified with 0.5 µl of primers, 25 µl of 2× iProof, 0.25 µl of SYBR Green and 8.25 µl of distilled H_2O under the following conditions: an initial denaturation at 98 °C for 30 s, 30 cycles of 98 °C for 10 s, 60 °C for 30 s, 72 °C for 30 s and 72 °C for 15 s, and a final incubation at 72 °C for 2 min. Fluorescence was assessed between the 30-s and 15-s steps at 72 °C. Amplified, size-selected libraries were quantified using an Agilent 2100 Bioanalyzer, and paired-end sequencing (of 101-bp reads) was performed on an Illumina HiSeq 2000. We generated a total of 13–14 fold sequence coverage.

Haplotype assignment and coalescent and phylogenetic analyses. We assigned haplotypes to 728 phased HapMap samples (1,456 chromosomes) using previously ascertained inversion-marking SNPs[10,16] and SNPs specifically tagging the H2 duplication that were identified in the present study in combination with aCGH[55] copy-number estimates. We used a three-SNP haplotype (rs1800547, rs2957297 and rs199451) to assign phased haplotypes to H1, H2 and H2D haplogroups. We required that all three SNPs match the expected haplotype. Phase-switch errors were manually corrected as described[13]. We confirmed H1-H2 status with PCR genotyping results and were able to resolve all but one haplotype; this individual was excluded from further analysis. We used aCGH[55] copy-number estimates to assign genotypes for these 728 individuals. Finally, we combined these sources of information to assign the final haplotype. Six out of 728 individuals showed discordance between the SNP genotype– and aCGH-based copy-number estimates. Three of these discrepancies were resolved by FISH experiments, and three were resolved by PCR genotyping. Fifty-seven individuals were heterozygous for H1 and H1D and carried both haplotypes, and, therefore, the haplotypes could not be assigned with confidence; these individuals were included in haplotype frequency data but not in phylogenetic analyses.

We combined the read depth–based copy-number estimates with the phased SNP data from the 1000 Genomes Project using a four-SNP haplotype (rs1396862, rs17650901, rs1052553 and rs199448) to assign phased haplotypes to H1, H2 and H2D haplogroups. We required that all four SNPs match the expected haplotype; 0.5% of haplotypes had one discordant SNP, and these haplotypes were flagged for manual inspection for phase-switch and genotyping errors. Phase-switch errors were manually corrected as described[13]. We then used read depth–based copy-number estimates to assign phased haplotypes to H1, H1D, H2 and H2D haplogroups. Finally, we combined these sources of information to assign the final haplotype. Sixteen out of 805 individuals showed discordance between SNP genotype– and read depth–based copy-number estimates. For five of these individuals, we also had aCGH and/or FISH data; for 4 out of 5 of these discrepancies, the SNP haplotype was concordant with aCGH and/or FISH data, whereas for the other, the read depth–based haplotype was consistent with aCGH and/or FISH data. We excluded the remaining 11 individuals with discordance from the rest of the analysis, as we could not confidently assign the haplotype. Sixty-seven individuals had the H1 or H1D haplotype, and, therefore, the haplotypes could not be assigned with confidence; these individuals were included in haplotype frequency data but not in phylogenetic analyses.

We also used the Illumina 650Y SNP genotyping data from 936 unrelated individuals from the HGDP collection. We phased the 650Y data using Beagle[56,57] and assigned haplotypes (H1, H2 and H2D) on the basis of a four-SNP haplotype (rs175635986, rs19871997, rs2668692 and rs199533). We required that all four SNPs match the expected haplotype; 0.5% of haplotypes had one discordant SNP or missing data, and these were excluded from further analysis. Illumina 1M SNP data from the African Diversity Panel (S.A.T., unpublished data) were used to assign haplotypes (H1, H2′ and H2D) on the basis of a three-SNP haplotype (rs1800547, rs2957297 and rs199451). Four individuals from the African Diversity Panel showed discordance between SNP haplotype and PCR genotyping results and were removed from further analysis. Illumina 550K SNP data from the Hunter-Gatherer Panel[21] were used to assign haplotypes (H1, H2′ and H2D) on the basis of a five-SNP haplotype (rs17563986, rs1800547, rs1981997, rs2668692 and rs199533). We required that all five SNPs match the expected haplotype; two haplotypes had one SNP that did not match the expected haplotype and were excluded from further analysis. We were able to type a total of 351 African samples from the African Diversity Panel and the Hunter-Gatherer Panel. We were unable to find any tag SNPs for the H1D haplotype; thus, analyses of the HGDP, African Diversity Panel and Hunter-Gatherer Panel did not include any chromosomes assigned to that haplotype, as we did not have independent copy-number information for these individuals.

We used PHYLIP[58] to build a maximum-likelihood tree on the basis of the alignment of 43 SNPs from the 136-kb LD block identified in the HapMap Project populations. BAC-based assemblies of the RP11 H1′ and H2D haplotypes and chimpanzee and orangutan haplotypes were aligned to whole-genome sequence from KB1, NA12878, NA20589 and NA21599 and haploresolved fosmid sequence from NA12156 with CLUSTALW[59].

We constructed a neighbor-joining phylogeny using Kimura two-parameter distance (complete deletion option) using MEGA4 (ref. 50). Nucleotide diversity and other population genetics analyses were performed using DnaSp v5 (ref. 60).

52. Antonacci, F. *et al.* A large and complex structural polymorphism at 16p12.1 underlies microdeletion disease risk. *Nat. Genet.* **42**, 745–750 (2010).
53. Adey, A. *et al.* Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density *in vitro* transposition. *Genome Biol.* **11**, R119 (2010).
54. Igartua, C. *et al.* Targeted enrichment of specific regions in the human genome by array hybridization. *Curr. Protoc. Hum. Genet.* Chapter 18, Unit 18 3 (2010).
55. Campbell, C.D. *et al.* Population-genetic properties of differentiated human copy-number polymorphisms. *Am. J. Hum. Genet.* **88**, 317–332 (2011).
56. Browning, B.L. & Browning, S.R. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.* **84**, 210–223 (2009).
57. Browning, S.R. & Browning, B.L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81**, 1084–1097 (2007).
58. Felsenstein, J. PHYLIP—Phylogeny Inference Package (Version 3.2). *Cladistics* **5**, 164–166 (1989).
59. Larkin, M.A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947–2948 (2007).
60. Librado, P. & Rozas, J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25**, 1451–1452 (2009).

doi:10.1038/ng.2335