nature genetics

A common inversion under selection in Europeans

Hreinn Stefansson[1,3], Agnar Helgason[1,3], Gudmar Thorleifsson[1], Valgerdur Steinthorsdottir[1], Gisli Masson[1], John Barnard[2], Adam Baker[1], Aslaug Jonasdottir[1], Andres Ingason[1], Vala G Gudnadottir[1], Natasa Desnica[1], Andrew Hicks[1], Arnaldur Gylfason[1], Daniel F Gudbjartsson[1], Gudrun M Jonsdottir[1], Jesus Sainz[1], Kari Agnarsson[1], Birgitta Birgisdottir[1], Shyamali Ghosh[1], Adalheidur Olafsdottir[1], Jean-Baptiste Cazier[1], Kristleifur Kristjansson[1], Michael L Frigge[1], Thorgeir E Thorgeirsson[1], Jeffrey R Gulcher[1], Augustine Kong[1,3] & Kari Stefansson[1,3]

A refined physical map of chromosome 17q21.31 uncovered a 900-kb inversion polymorphism. Chromosomes with the inverted segment in different orientations represent two distinct lineages, H1 and H2, that have diverged for as much as 3 million years and show no evidence of having recombined. The H2 lineage is rare in Africans, almost absent in East Asians but found at a frequency of 20% in Europeans, in whom the haplotype structure is indicative of a history of positive selection. Here we show that the H2 lineage is undergoing positive selection in the Icelandic population, such that carrier females have more children and have higher recombination rates than noncarriers.

Though important for evolution, large chromosomal rearrangements such as deletions, duplications and inversions are generally thought to be deleterious. These large-scale polymorphisms contribute substantially to genomic variation among humans and account for much of the genomic difference between humans and other primates[1–3]. The architecture of large inversion polymorphisms suggests they may occur through nonallelic homologous recombination assisted by low-copy repeats positioned in the genome in an inverted orientation[4,5].

Genotype analysis may show whether a segment is duplicated or deleted, by means of an apparent gain or loss of heterozygosity, respectively. In contrast, inversions are difficult to detect, particularly those of moderate size. Genotypes of markers inside inverted regions are consistent among relatives and, unless inversions are several megabases in size, they are not easily detected with standard cytogenetic assays. A few large and common inversion polymorphisms have been detected in the human genome, the most notable being a large inversion on chromosome 8p (ref. 5). These may be only the tip of the iceberg.

Here we describe, for the first time to our knowledge, a 900-kb inversion polymorphism at 17q21.31, a region that contains several genes, including those encoding corticotropin releasing hormone receptor 1 (*CRHR1*) and microtubule-associated protein tau (*MAPT*). Previous studies have characterized two highly divergent *MAPT* haplotypes, H1 and H2, and noted the existence of strong linkage disequilibrium (LD) across a 1.6-Mb region containing the gene[6–12]. We provide a detailed description of the unusual haplotype structure in this inverted region, evaluate the impact of natural selection in the past and present, and discuss the implications for our understanding of human evolutionary history.

RESULTS

Discovery of a 900-kb inversion polymorphism

The Build 34 assembly of chromosome 17q21.31 is chimeric, constructed to a large extent from clones representing different *MAPT* haplotypes of type H1. We used a set of chromosome-specific BAC contigs to show that there is a 900-kb inversion polymorphism in this region (**Fig. 1**). We generated the chromosome-specific assembly from RP11 BAC clones (Roswell Park Cancer Institute Human BAC Library) originating in a DNA sample from one individual. The two RP11 chromosomes represent *MAPT* haplotypes of type H1 and H2 based on the characteristic alleles for a dinucleotide marker in intron nine[6] (DG17S142) and a characteristic 238-bp deletion in the same intron on the H2 background[6]. By genotyping RP11 clones from 17q21.31 for 60 microsatellite markers and assembling the clones into chromosome-specific contigs, we found that the H2 haplotype was structurally different from the Build 34 assembly. The segment from 44.1 to 45.0 Mb is inverted on the H2 background compared with the H1 background and Build 34. Furthermore, a 127-kb tandem duplication containing exons 1–13 of the N-ethylmaleimide-sensitive factor gene (*NSF*) is located upstream of a full-length copy of *NSF* in the H1 variant in Build 34. The same *NSF* exons are also duplicated on the H2 chromosome, but the H2 duplication is larger (280 kb), spanning the H1 duplication and extending to the 5′ end of the gene *LOC284058*. The two *NSF* copies (the partial copy with exons 1–13 and the full-length copy) are separated by only 100 kb on the H1 chromosome in the RP11 library, whereas on the H2 chromosome in the RP11 library, the partial copy of *NSF* is inverted and located 1 Mb upstream of the full-length copy of *NSF*.

[1]deCODE Genetics, Sturlugata 8, 101 Reykjavík, Iceland. [2]Department of Biostatistics and Epidemiology, Cleveland Clinic Foundation, Cleveland, Ohio, USA. [3]These authors contributed equally to this work. Correspondence should be addressed to A.K. (kong@decode.is) or K.S. (kari.stefansson@decode.is).

Published online 16 January 2005; doi:10.1038/ng1508

H2 chromosomes seem to have an identical 280-kb *NSF* duplication. H1 chromosomes have a complex arrangement of duplications close to *NSF*. In one common H1 variant, the first 13 exons of *NSF* are found in three copies, and in another H1 variant, a duplicated region is immediately upstream of *NSF*, not including exons 1–13. Surrogate haplotypes have been identified for four different H1 variants, and these variants have been characterized with respect to size, location of the duplication (**Supplementary Fig. 1** online) and gene dosage (**Supplementary Fig. 2** online). Furthermore, pulsed-field Southern-blot analysis of *NSF* identified at least five chromosomal variants with considerable variation in size at the 17q21.31 locus (**Supplementary Fig. 3** online). The duplicated, truncated form of *NSF* is expressed, and its expression is strongly correlated with gene dosage ($P = 1.4 \times 10^{-7}$; **Supplementary Fig. 4** online). The truncated copy is not polyadenylated (data not shown) and is probably unstable. Furthermore, expression of the truncated gene does not seem to affect expression of full-length *NSF* ($P = 0.84$, $n = 85$).

Figure 2a shows a maximum parsimony tree for sequences from four human clones (including the two RP11 chromosomes with known inversion status) and a chimpanzee clone spanning a 77-kb region that contains the first two exons of *MAPT*. The CTD-2524N8 and 17q21.1 clones share SNP alleles characteristic of the H1 variant with the RP11 H1 chromosome (**Supplementary Table 1** online), with negligible divergence (0.0243%) between these three sequences. In comparison, the scale of mutational divergence between the RP11 H2 chromosome and the three H1 lineage clones is vast (0.3444%). Assuming that the split between humans and chimpanzees occurred ~6 million years ago[13] and that the rates of evolution for this region

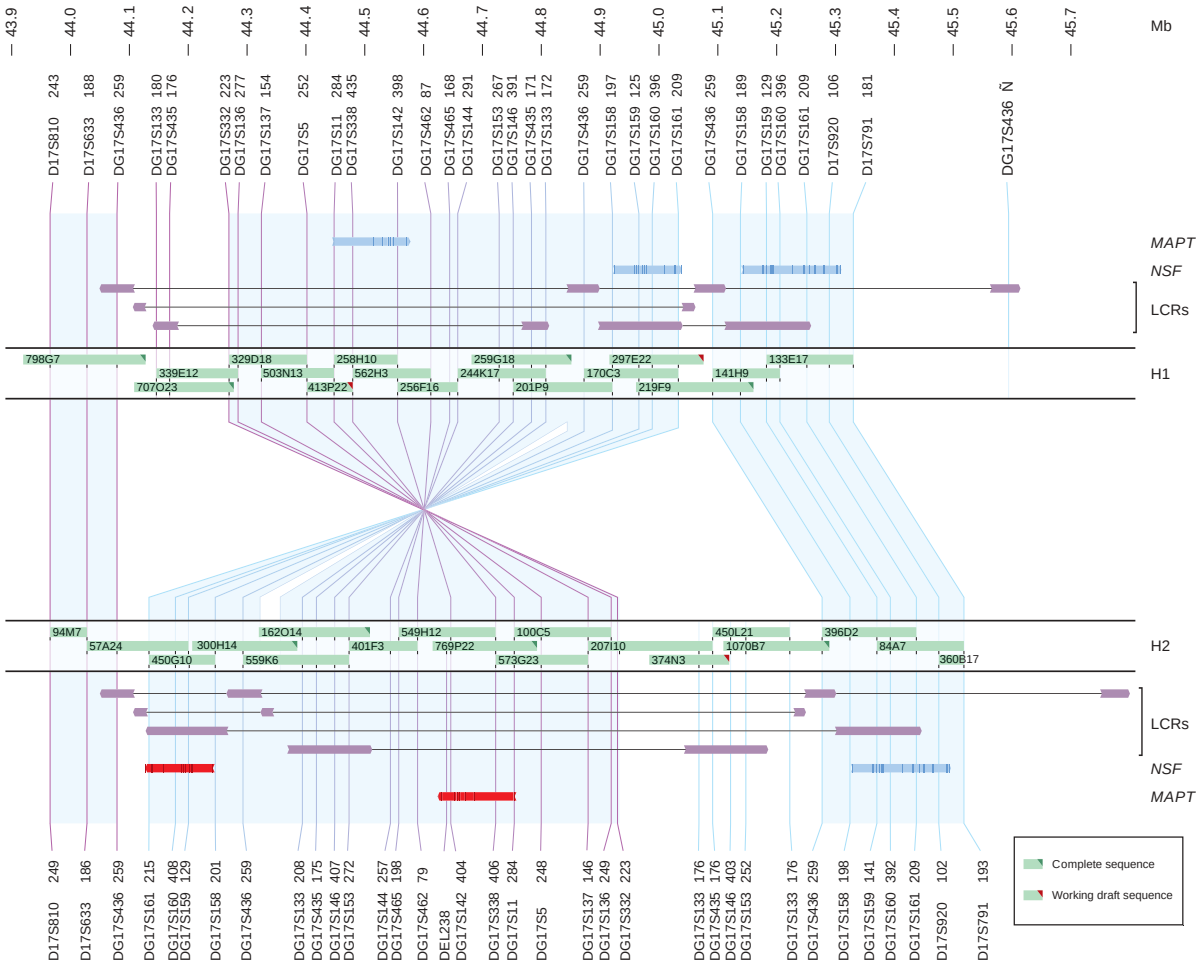

Figure 1 A 900-kb inversion polymorphism was detected on one of the RP11 chromosomes at the 17q21.31 locus. The breakpoints reside in 100-kb palindrome low-copy repeats (LCRs), 900 kb apart. BACs from the two chromosomes in the library have been separated by genotyping microsatellite markers on respective clones and assembling the clones into two chromosome specific contigs. The BAC clones are represented by green bars and genotypes for each clone shown on the side for respective markers (primer sequences are given in **Supplementary Table 5** online). The upper chromosome is in keeping with the Build 34 assembly and is of type H1 on the basis of repeat numbers for a dinucleotide marker (DG17S142) in intron 9 of *MAPT*. The allele size 398 refers to 11 repeats (allele a0 in the original references[6,45]) representing the H1 lineage. The lower chromosome has 14 repeats on the same marker and a characteristic 238-bp H2 deletion in clone 769P22, which contains *MAPT*. The H2 chromosome is structurally different. In particular, the region between markers DG17S332 and DG17S161 is inverted on the H2 chromosome relative to the H1 chromosome. Furthermore, the H2 chromosome carries two copies of the low-copy repeat spanning from DG17S146 to DG17S153, whereas the H1 chromosome has only one copy of this segment. *NSF* and a truncated copy of *NSF*, exons 1–13, are present on both chromosomes. But the truncated *NSF* copy (blue) is on the same strand and 100 kb upstream of the full-length copy of the gene on the H1 variant, whereas the truncated *NSF* copy (red) does not map to the same strand and is located 900 kb upstream of the full-length copy in the H2 variant. The use of microsatellite genotypes for the RP11 clones was the key to the chromosome-specific assemblies. The assemblies are in keeping with available sequences from RP11 clones mapping to this locus (**Supplementary Table 1** online).

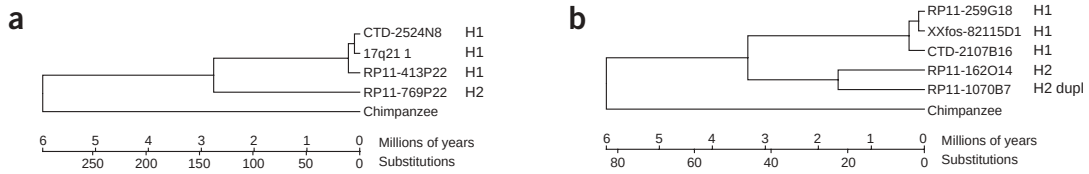

Figure 2 The sequence divergence of the H1 and H2 lineages. (**a**) A maximum parsimony tree based on substitutions identified in full sequences from clones spanning a 77-kb region containing the first two exons and introns of *MAPT* (44.444005–44.520949 Mb in Build 34). (**b**) A maximum parsimony tree based on 32-kb clone sequences for a region that is present on both the H1 and H2 lineages but is duplicated (dupl) in the RP11 H2 chromosome (44.730628–44.762592 Mb in Build 34). The sequence divergence between the H1 and H2 lineages in **b** is 0.2894%, indicating a coalescent age of close to 3 million years, with the age of the duplication event in the H2 lineage estimated at ~1.5 million years and a divergence among H1 sequences of 0.0167%. Both trees show an unusually deep divergence between the H1 and H2 lineages relative to the overall sequence divergence between humans and chimpanzees in this region (see **Supplementary Fig. 5** online for a 5-kb sliding-window analysis of sequence divergence in the two sequence regions). The age estimation for the split between the H1 and H2 lineages is dependent on a number of assumptions, such as neutral evolution and equivalent mutation rates along all branches. The roughly equal divergence of the H1 and H2 lineages from the chimpanzee sequence indicates that selection has at least not differentially affected the accumulation of mutations in the H1 and H2 lineages.

from the common ancestor of both species were equal, the human H1 and H2 lineages probably diverged as much as 3 million years ago (**Fig. 2a**). The probability of such an ancient divergence assuming a standard neutral coalescent model with an ancestral human effective population size of 10,000 and a generation interval of 30 years is 0.006737. We obtained a similar result for a second 32-kb sequence that is duplicated in the RP11 H2 chromosome but found in only single copies on the H1 chromosomes and in the chimpanzee (**Fig. 2b**). For both the 77-kb sequence (**Fig. 2a**) and the 32-kb sequence (**Fig. 2b**), the divergence between H1 chromosomes is ~250,000–300,000 years, a coalescent age more typical for human autosomal genes[14].

We were not able to establish the ancestral orientation of the 17q21.31 inversion region in humans, as the orientation of the homologous chimpanzee chromosome 19 sequence is not known and the human H1 and H2 chromosomes share a similar number of ancestral allele states with the chimpanzee. Thus, whereas the divergence between the H1 variant and the chimpanzee is 0.7759% for the 77-kb sequence (**Fig. 2a**) and 0.5193% for the 32-kb sequence (**Fig. 2b**), the corresponding divergence between the H2 variant and the chimpanzee is 0.763% and 0.5303%, respectively (**Supplementary Fig. 5** online).

An unusual pattern of haplotype diversity

The haplotype formed by the alleles of six microsatellite loci, DG17S137 [146], DG17S7 [206], DG17S338 [406], DG17S10 [177], DG17S340 [108] and DG17S142 [404 or 406] (the H2 microsatellite haplotype), has a frequency of 17.5% in Icelanders (n = 1,880). All these microsatellites had a bimodal distribution of allele sizes, with the H2 allele differentiated from H1 alleles.

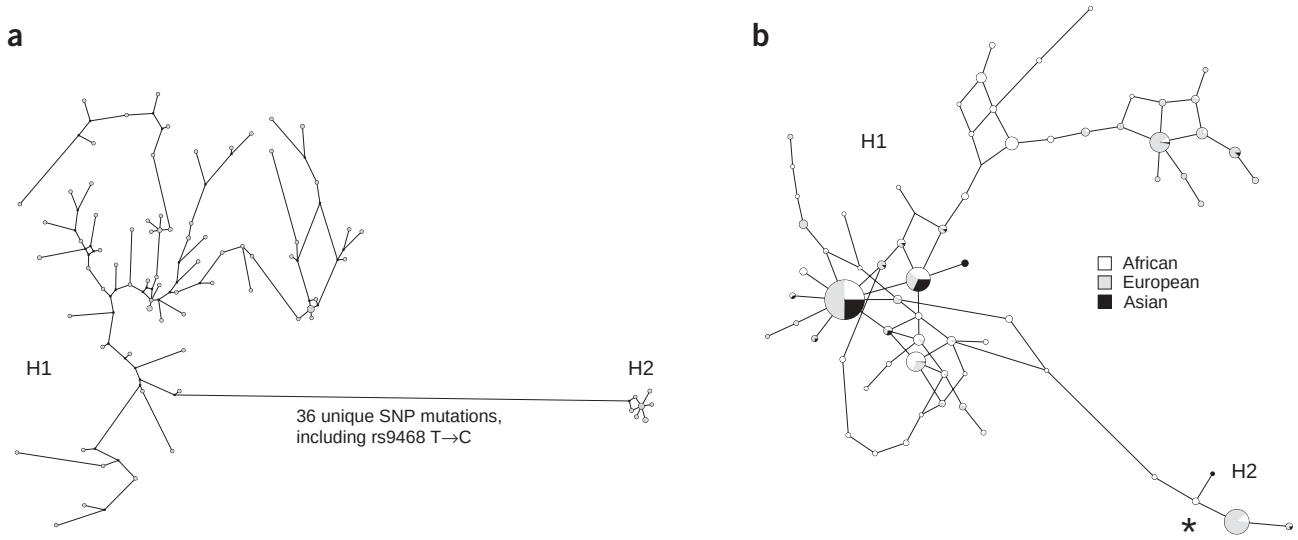

Figure 3 Networks depicting the haplotype structure of the inversion. (**a**) A median-joining network describing the genealogical relationship between 77 haplotypes from 60 individuals from Utah with Northern and Western European ancestry included in the Centre d'Etude de Polymorphisme Humane database. Haplotypes were constructed using 95 SNPs from the HapMap project database and six microsatellites covering the part of the inversion between megabases 44.201106 and 44.625694 according to Build 34. The network describes the inferred genealogical relationships between haplotypes in terms of mutational distance, which in this case can be the result of either substitution or recombination events. Haplotypes are represented by circles, whose area reflects the number of copies observed in the sample. Lines represent differences between the allelic states of haplotypes, with length proportional to the number of differences. A stepwise mutational model was assumed for microsatellites. (**b**) A median-joining network describing the genealogical relationship between 68 haplotypes from 484 individuals of African (n = 177), European (n = 241) and Asian (n = 66) ancestry produced from six microsatellites. The asterisk marks the H2 founder haplotype, all five copies of which are found in individuals of African ancestry, three of whom are Mbuti.

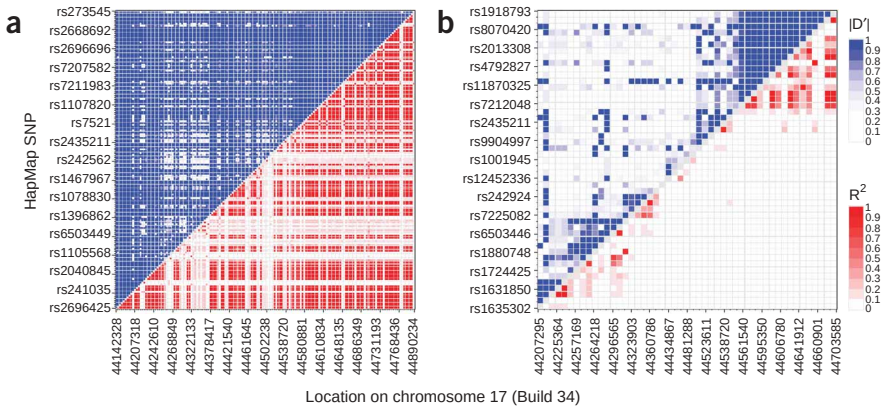

Figure 4 The impact of the inversion on LD patterns. LD between SNPs spanning the inversion at 17q21.31 (**a**) in the Utah sample and (**b**) in a Yoruba sample. The names of every fifth SNP for the Utah sample and every third SNP for the Yoruba sample are given on the *y* axis, and the corresponding Build 34 base location are given on the *x* axis. The LD plots were generated using data from the HapMap project database (release 13, megabases 44.1–45.0 in Build 34). The SNPs used for the two plots overlap for the most part, but some SNPs available for the Utah sample have not yet been typed for the Yoruba sample and some SNPs are not polymorphic in either the Utah or Yoruba samples; in particular, the large number of SNPs that differentiate the H1 and H2 lineages are not polymorphic in the Yoruba sample. Only those markers that were polymorphic in >1% of the sample chromosomes being studied were used in the LD analysis. In the Yoruba sample, 37 SNPs that tag H1 or H2 status in the Utah sample are not included in the LD picture because they are not polymorphic in that sample. In the Utah sample, there are 165 polymorphic SNP markers (159 polymorphic in >1% of the sample chromosomes) in the region spanning megabases 44.1–45.0 in the HapMap database. Of these, 79 are perfectly correlated with the tagging microsatellite for the inversion status. For the Yoruba sample, 112 SNP are typed, of which 53 are polymorphic (49 polymorphic in >1% of the sample chromosomes). Of the 79 tagging SNPs in the Utah sample, 41 are typed in the Yoruba sample, only four of which are polymorphic.

The H2 microsatellite haplotype (with two variant haplotypes carrying one-step mutations in DG17S142 and DG17S10, respectively) is found in 20% (24 of 120) of the independent chromosomes in the Utah HapMap project sample[15,16]. We generated haplotypes for these individuals by combining alleles from the six microsatellites with alleles of 95 polymorphic SNPs from the HapMap database spanning a 424-kb nonduplicated segment inside the inverted region. The H2 microsatellite haplotypes are differentiated on a long branch defined by numerous SNP and microsatellite mutations (**Fig. 3a**). Among the 95 SNPs, 36 separate the H1 variant from the H2 variant, such that one allele is fixed in the H1 variant and the other allele is fixed in the H2 variant (a magnitude of mutational divergence broadly consistent with the results shown in **Fig. 2**). This indicates that no recombination has taken place between the H1 and H2 chromosomes and supports the use of the H2 microsatellite haplotype and SNP allele states as surrogate markers for inversion status.

The H2 haplotypes are extremely homogeneous relative to the diverse H1 haplotypes (**Fig. 3a**), which, when added to the ancient divergence between the H2 and H1 lineages and the 20% frequency of H2 haplotypes in the Utah sample, creates an unusual pattern of diversity, which could reflect a history of ancient balancing selection (maintaining ancient lineages in the human gene pool) superseded by an episode of strong positive selection[17,18]. To test whether such patterns of diversity could have arisen through neutral evolution, we compared the relative diversity of microsatellites on the H1 and H2 backgrounds to multiple sets of haplotypes generated through coalescent simulations under four different demographic scenarios (**Supplementary Table 2** online). Our simulations indicate that the extreme homogeneity of the H2 lineage is incompatible with the expectation from neutral evolution for a lineage at 20% frequency under the tested demographic scenarios, including a recent expansion after a severe bottleneck. An additional twist in the evolutionary history of the 17q21.31 inversion region is that the apparent signature of positive selection of the H2 variant is limited to European populations. Thus, the frequency of H2 microsatellite haplotypes is 21% in individuals of European ancestry ($n = 241$), but only 6% and 1% in individuals of African ($n = 177$) and Asian ($n = 66$) ancestry, respectively (**Fig. 3b**).

Figure 4 shows the marked effect of the inversion on LD patterns. In individuals of European origin, here represented by the Utah sample, LD across the inverted region is consistently strong. In contrast, the LD structure in the Africans in the Yoruba HapMap sample is weaker and broken up into several smaller blocks. The LD structure in the Han Chinese and Japanese HapMap samples is analogous to that in the Yoruba sample (**Supplementary Fig. 6** online), indicating that the strong LD in the Utah sample is attributable to the relatively high frequency of the extremely homogeneous and divergent H2 lineage, a result of the inversion hindering recombination and, probably, the impact of positive selection.

Figure 5 presents H2 lineage frequencies in various human populations using two *MAPT* SNPs from the ALFRED database[19]. Both are among the 36 SNPs that differentiate H1 from H2 haplotypes in the Utah HapMap sample and have minor alleles that indicate H2 lineage status (**Fig. 3a**). They are also present in the H2 founder haplotype (**Fig. 3b**). The high frequency of H2 chromosomes throughout Europe is notable, with a maximum in the Near Eastern Druze and Samaritan populations and a diminishing gradient eastward into Asia (see also ref. 20).

An African origin for H2 chromosomes is indicated by the greater diversity of H2 haplotypes in individuals of African ancestry (despite their apparent low frequency) and the observation that the H2 founder haplotypes are present in only this group (**Fig. 3b**). This raises the possibility of finding mutational differences between African and European H2 chromosomes that contributed to their selective expansion in the latter group.

Evidence for positive selection of H2 chromosomes in Iceland

To determine whether positive selection is presently acting on H2 chromosomes, we genotyped 29,137 Icelanders, 16,959 women and 12,178 men, born between 1925 and 1965, with the marker DG17S142. In Icelanders (and in the Utah sample), the alleles 404 and 406 of this marker are perfect surrogates for the H2 lineage. We regressed the number of offspring on the number of copies of H2 a person carries (*i.e.*, an additive model), adjusted for year of birth and sex. We used weighted regression to adjust for the fact that people who have more children are over-represented in the genotyped samples (**Supplementary Table 3** online). Because the individuals are related and their genotypes are not independent, we determined standard errors and *P* values empirically by carrying out 10,000

simulations of the H2 lineage through the entire Icelandic genealogy. The estimated effect of H2 chromosomes for both sexes combined, assuming an additive model, is an increase of 0.0623 children per copy with a two-sided P value of 0.0056. Upon closer examination, we found that homozygous carriers of H2 have, on average, fewer children than heterozygous carriers, although the difference is not statistically significant. Hence, the additive model, which assumes that homozygous carriers have more children than heterozygous carriers, is not a good fit. We refitted the data assuming a dominant effect for the H2 and assuming that heterozygous and homozygous carriers have the same number of children on average (**Table 1**). With both sexes combined, a carrier of H2 is estimated to have 0.0796 more children than a noncarrier ($P = 0.0025$). For women, the estimated effect is 0.0907 ($P = 0.0068$), which translates to an increase of ~3.5%, as women in this era had an average of 2.584 children. For men, the estimated effect is 0.0679 ($P = 0.0719$) or ~2.9% of 2.369 (the average number of children for men in this era). The estimated effect for men is insignificant and smaller than that for women but probably real. Results from fitting a full model for H2 status are presented in **Supplementary Table 4** online. Because of the small number of homozygous carriers, we can neither confirm nor reject the possible action of balancing selection, with the heterozygous carriers having more children than either type of homozygotes.

The sample frequencies of H2 for women who have zero, one, two, three, four or five or more children show a monotonically increasing trend (**Supplementary Fig. 7** online). Women who have six or more children, however, do not have a higher frequency of H2 than those who have five children. For women who already have five children, fertility may not be a main factor in determining whether they would have more. For men, we also see a tendency for the frequency of H2 to increase as the number of children increases (**Supplementary Fig. 7** online), but the trend is more jagged.

Apart from the carriers having more children, another way that the inversion could have increased its frequency is through transmission disequilibrium, when children are more likely to inherit H2 chromosomes than the more common H1 chromosome from heterozygous parents. In 5,529 parent-offspring trios, in which we genotyped all three individuals for DG17S142, we found no evidence of transmission disequilibrium; of the 3,286 cases in which a parent was heterozygous, the H2 chromosome was transmitted 1,641 times ($P = 0.96$).

Mothers carrying H2 have higher recombination rates

Given that there is support for the hypothesis that positive selection is acting on H2 chromosomes, we wondered through what mechanisms selection manifests itself. Mothers who have a higher recombination

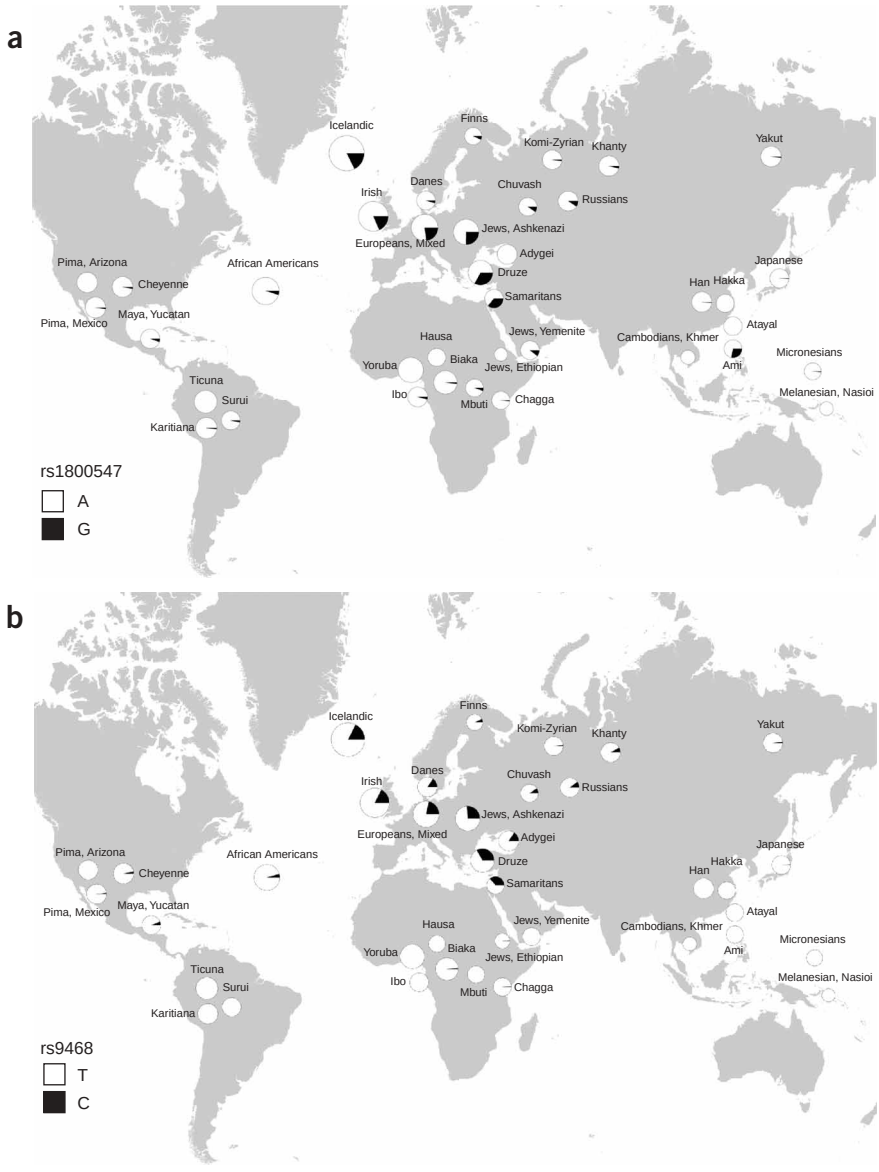

Figure 5 Worldwide distribution of the H2 lineage based on allele frequencies of two diagnostic SNPs obtained from the ALFRED database[19]. (**a**) rs1800547 (ALFRED name C7563692), located downstream of *MAPT* exon 4. (**b**) rs9468 (ALFRED name C7563752), located in exon 9 of *MAPT*. On the basis of the clones used to construct the maximum parsimony tree in **Figure 2a**, the rs1800547 G allele and the rs9468 C allele are on an H2 haplotype background, whereas the A and T alleles are on an H1 background. The allele frequencies for Icelanders were obtained through genotyping in the deCODE laboratories. There is a discrepancy between the allele frequencies of rs1800547 and rs9468 in the Ami population from Taiwan and the Adygei from Southwest Russia, which may be due to either a database error or an unknown complication in the haplotype structure of H1 and H2 chromosomes. Allele C of rs9468 is rarer than allele G of rs1800547 for some population samples, particularly in the African Ibo, Mbuti and Biaka groups. This may be indicative of greater mutational diversity of the H2 lineage in these populations, or it could be accounted for by the fact that sample sizes for the two SNPs differ in the ALFRED database for most of the populations. The frequency of H2 chromosomes based on six microsatellite haplotypes in a small sample of Mbuti pygmies is three of ten (the H2 founder haplotypes marked by the asterisk in **Fig. 3b**). But the frequency of H2 chromosomes is substantially lower if allele G of rs1800547 (5%, $2N = 78$) or allele C of rs9468 (0%, $2N = 78$) is used as a surrogate allele in the sample for this population in the ALFRED database[19].

Table 1 Relationship between number of children and H2 carrier status

Cohort	Predictor	Estimate	Standard error	*P* value
Combined	Year of birth	−0.0340	0.0008	0.0000
	Carrier of H2	0.0796	0.0259	0.0025
	Sex	0.2360	0.0190	0.0000
Female	Year of birth	−0.0331	0.0010	0.0000
	Carrier of H2	0.0907	0.0338	0.0068
Male	Year of birth	−0.0349	0.0012	0.0000
	Carrier of H2	0.0679	0.0375	0.0719

Results from multiple regression analyses for a cohort of 16,959 females and 12,178 males born between 1925 and 1965, provided for the entire cohort and for the females and males analyzed separately. The regression uses weights that are inversely proportional to the probability that a person will be included in the regression. Standard errors and *P* values are empirically determined based on 10,000 simulations. The estimated effects presented here are statistical; they do not imply causation, as the results could be a consequence of the H2 chromosomes being correlated with some unknown functional variant(s).

Table 2 Relationship between recombination rates and the number of H2 chromosomes carried

Response (recombinations)	Predictor	Estimate	Standard error	*P* value
Total	Number of H2	0.4721 (1.04)	0.1180	0.0002
	Year of birth	−0.0024	0.0064	0.7037
	Average age at birth of offspring	0.0838	0.0152	0.0000
Telomeric	Number of H2	0.0257 (2.24)	0.0092	0.0047
	Year of birth	−0.0006	0.0005	0.2130
	Average age at birth of offspring	0.0077	0.0012	0.0000
Nontelomeric	Number of H2	0.4465 (1.01)	0.1160	0.0004
	Year of birth	−0.0018	0.0062	0.7724
	Average age at birth of offspring	0.0760	0.0149	0.0000

Results from multiple regressions using inversion count, mother's year of birth and mother's average age at birth of her offspring as predictors, and using the total genome and the genome divided into telomeric and nontelomeric regions as the three different responses. The telomeric region for each chromosome arm is the region including all genotyped markers within 6 cM of the most telomeric marker for that arm. In aggregate, the telomeric regions account for 2.8% of the total genomic length. Data is based on 5,012 mothers born between 1925 and 1965 with at least two children genotyped. For H2, the estimated percentage increase per copy is indicated in parentheses.

rate tend to have more children[21]. Therefore, we examined whether the inversion has an impact on recombination rate. Of the 16,959 women genotyped with the DG17S142 marker, 5,012 have also been typed with our set of ~1,000 genome-wide linkage markers, as have two or more of their children and often their spouses, allowing us to estimate recombination rates[21]. We studied a total of 20,955 individuals with 17 million genotypes. Regressing the estimated recombination rate on the number of copies of inversion the mother carries, adjusting for her year of birth and the average age of the mother at the birth of her genotyped children (**Table 2**), the recombination rate is estimated to increase by 0.472 Morgans per copy of H2 (two sided $P = 0.0002$), or ~1.0% of the average female genetic length for the genome. Here, the additive model fits the data nearly perfectly. Fitting a full model for H2 status, the increase in recombination rates of homozygous carriers relative to noncarriers is estimated to be 1.006 Morgans, close to double the increase in the heterozygous carriers.

We previously determined that the recombination counts of live-born children increases with mother's age and, in percentage increase, the effect is approximately four times larger for the telomeric regions[21]. Here we find the same trend; the estimated effect of the inversion for the telomeric regions is ~2.2%, twice that of the rest of the genome. We further investigated the 23 chromosomes individually (**Fig. 6**). Although the estimates vary, the inversion has a broad impact on the recombination rate over the genome; its effect is not limited to a few chromosomes. This is expected, because recombination counts are highly correlated across chromosomes[22].

We did not find significant correlation between recombination rate of fathers and inversion status, which is not unexpected, given that difference in paternal recombination rates has not been detected previously with family data[22,23].

These results support our previous finding that fertility of women is in part affected by their recombination rates. But is the impact of inversion status on the number of children completely explained by its effect on maternal recombination rate? Though significant, the inversion explains only ~0.3% of the variance in maternal recombination rates among mothers, which translates to ~1% of the genetic component, as the heritability of maternal recombination rate was estimated to be ~30% (ref. 21). Based on our estimate of the effect of the H2 variant on recombination (0.472 Morgans per copy) and the effect of recombination rate on the number of children (0.0109 child per Morgan), the effect of the H2 variant on number of children through recombination is only 0.00514 children per copy or, on average, 0.00564 more children for a carrier relative to a noncarrier, which is less than one-tenth of the total estimated effect of the inversion (0.0907 children for females). Hence, H2 chromosomes probably influence the number of children through pathways other than recombination rates of women. This conclusion is not definitive, however, because we are able to estimate recombination rates only for women who have at least two children, leading to a possible underestimation of the effects of recombination rate.

DISCUSSION

Previous studies identified extensive LD in the 17q21.31 region and considerable divergence between the H1 and H2 lineages[6–11]. Our discovery of a 900-kb inversion accounts for and expands on these observations and provides insights into the organization of the human genome (*e.g.*, that there are multiple genetic and physical maps of the genome) and the phenotypic and reproductive impact of large-scale genetic variation. The inversion is a new genetic variant for association testing and a potential source of variation in expression patterns that could contribute to disease phenotypes. Our study identifies an uncommonly ancient divergence between the H1 and H2 lineages and indicates that positive selection has probably shaped the unusual haplotype structure of this genomic region in European populations. We furthermore identify, for the first time to our knowledge in humans, a sequence variant associated with recombination rate and an association with an increased number of children, a direct observation of natural selection taking place in a human population.

It remains to be determined how abundant inversion polymorphisms are in the human genome and what impact they have on phenotypic variation. One promising indirect approach to identifying inversion polymorphisms is through patterns of LD, which can be unusually strong and extended at inverted regions[4,24] owing to restriction of crossovers in individuals who are heterozygous

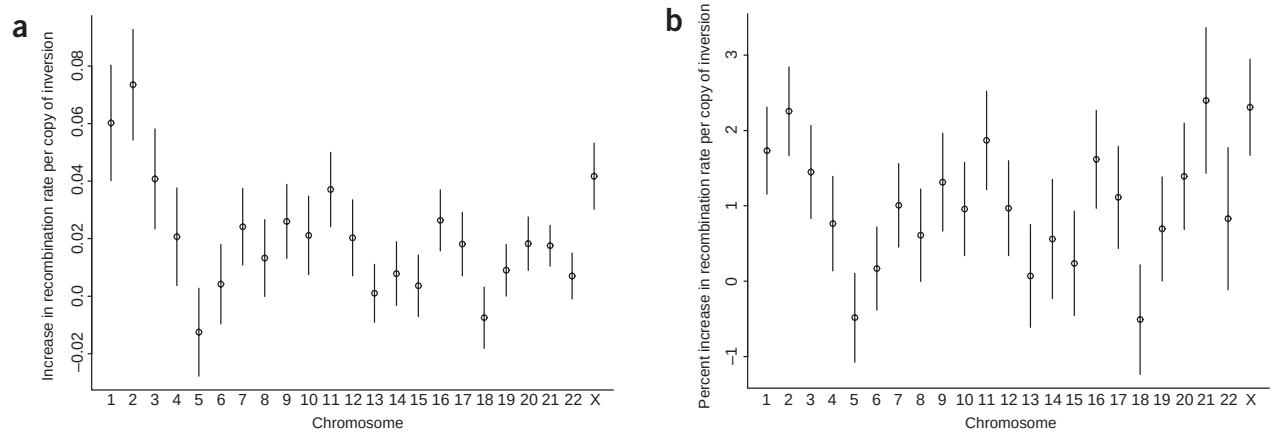

Figure 6 The genome-wide impact of the number of H2 copies on the rate of recombination in mothers. (**a**) Increase in Morgans. (**b**) Percentage increase. Error bars represent s.e.

with respect to inversions. This not only reduces the overall recombination rate but also causes the two orientations to accumulate mutations independently, mimicking the evolution of lineages in isolated populations and making it possible for variants on the same orientation to develop symbiotic relationships. When numerous alleles are in perfect LD, it may be difficult to identify those associated with disorders, and functional approaches may be required to identify the causative variants. Although our results implicate a higher recombination rate as a potential source for the reproductive advantage conferred by H2 chromosomes, an explicit functional link has yet to be established to a particular variant on the H2 chromosomal background.

The evolutionary history of the 17q21.31 inversion polymorphism is puzzling. First, we are faced with the question of how to account for the presence of two such ancient lineages in the ancestral human gene pool, with sequence divergence indicating a coalescent age of up to three million years. This predates the emergence of anatomically modern *Homo sapiens* in Africa (~150,000 years ago) and may even predate the origin of the genus *Homo* (~2.5 million years ago)[25]. One possible explanation is that both lineages were maintained in the ancestral human gene pool from the time that the inversion occurred, with some kind of balancing selection preventing either lineage from being lost through genetic drift[17,26]. Another possibility is that the two lineages diverged in isolated gene pools for an extensive period, and the H2 lineage was introduced into the ancestral human gene pool before the migration(s) of anatomically modern humans from Africa ~60,000 years ago. This would imply admixture from contemporaneous hominin species, like *Homo heidelbergensis* or *Homo erectus*. Two studies provide indirect evidence that anatomically modern *H. sapiens* probably encountered other hominin species in Asia[27,28]. It is possible that such encounters took place in Africa, leading to the introduction of the H2 lineage into our ancestral gene pool. Time will tell whether we find other genomic regions with similarly ancient coalescent ages (particularly in regions with low recombination rates) as the admixture hypothesis predicts.

Whether retained by balancing selection or introduced by another hominin species, the H2 lineage in contemporary humans seems to have an African origin (see **Fig. 3b**). H2 chromosomes were probably rare in the groups that left Africa more than 60,000 years ago and gave rise to the populations of Europe and Asia[29], on the basis of the observation that H2 chromosomes are presently rare in Africans and almost nonexistent in Asians. The higher frequency of H2 chromosomes in European populations, coupled with extreme homogeneity, is consistent with a rapid expansion from a few founder chromosomes, probably owing to positive selection. This selective episode could reflect exposure to some local environmental pressures in Europe and the Near East. A number of studies have inferred the action of positive selection on haplotype structure[18,30–33], and some have identified geographic differences in selection histories similar to that reported here for 17q21.31 (ref. 34).

Sequencing genes in a modest number of individuals or using densely distributed SNPs such as those from the HapMap project, preferably from multiple populations, can provide evidence for past episodes of selection[18,31,33,34]. But dissecting and reconstructing the potential range of selection scenarios for such data is tricky with the methodological tools currently available, particularly for inverted regions like 17q21.31. Conclusions are sensitive to model assumptions that are difficult to validate and can be affected by subtle ascertainment biases. Although our results suggest that positive selection is the most likely determinant of the unusual haplotype structure observed in European populations at 17q21.31, a more complicated selection history for the region cannot be ruled out. Neither can the possibility that an extreme founder event occurred in the ancestral European gene pool. But our finding that positive selection is presently operating to increase the frequency of the H2 lineage in Icelanders supports the hypothesis that positive selection is a key determinant in its evolutionary history in European populations. We predict that the selection observed in Iceland is also taking place in other European populations, because they share not only ancestry but also environmental conditions and demographic features.

In this study, we established a general methodology to investigate regions in the genome for signatures of selection in the present. By genotyping a large number of individuals with a surrogate marker, with the help of a detailed genealogy, one can assess the selective impact of genetic variants by examining the number of children carriers have relative to noncarriers. This is arguably the most direct approach to studying selection, regardless of the functional mechanisms behind the selective forces. Effects such as those we detected for H2 chromosomes can be small in each generation but can have profound consequences on an evolutionary timescale. We believe that this approach can be used to assess both the deleterious and advantageous impact of genetic variation on human lives, thereby providing a new basis for increasing our understanding of the molecular nature of disease and positively selected adaptations.

METHODS

Subjects and genotyping. We obtained all the biological samples used in this study with informed consent in accordance with protocols approved by the Data Protection Commission of Iceland and the National Bioethics Committee of Iceland. All personal identifiers were encrypted using a code that is held by the Data Protection Commission of Iceland[35]. The genealogy database provides the birth years of the individuals, rounded up to the nearest five years to protect privacy. Details concerning genotyping, allele calling and genotype quality control were previously published[22]. Of the 5,012 nuclear families that we used to study the correlation between the H2 chromosomes and recombination rates of mothers, 4,805 are a subset of the 5,463 families used in another study[21], and more information regarding these families can be found in that report. The 207 new families in this study have mothers who were born between 1955 and 1965, a period not included in the previous study.

BAC genotyping and assembly. We isolated RP11 clones by hybridizing membranes containing DNA from the RP11 library using markers from the region as probes. We then genotyped positive clones for 60 microsatellite markers and used the genotypes to generate the chromosome specific assemblies. We cultured BACs in Luria broth and chloramphenicol (25 mg ml^{-1}) overnight and deposited 3 µl of the culture into 100 µl of cold Tris-EDTA buffer. To release DNA from the clones, we incubated samples at 98 °C in a PCR thermocycler for 10 min. We added 2 µl of the sample to a PCR mix and carried out PCR in a Peltier Thermal Cycler (PTC-225 from MJ Research). Cycling conditions were as follows: 95 °C for 12 min, followed by 40 cycles of 95 °C for 30 s and annealing for 30 s at 55 °C, and 1 min extension at 72°C.

Phylogenetic analyses. We generated a maximum parsimony tree using MEGA version 3 (ref. 36) for full sequence data spanning the 77-kb region containing *MAPT*. We aligned three clone sequences representing the H1 lineage (CTD-2524N8 [AC010792.4], 17q21.1 [AC091628.2] and RP11-413P22 [AC036218.3]), one clone sequence representing the H2 lineage (RP11-769P22 [BX544879.6]) and a chimpanzee sequence (National Center for Biotechnology Information Build 1 Version 1, chr19:44,570,001–44,950,000) using the software BlastZ[37]. We used only those sites that were present in all sequences, where base substitution was indicated by the alignment, to score differences between sequences.

We also generated a maximum parsimony tree for the 32-kb region that is duplicated in the H2 lineage. We obtained the original H2 copy from the clone RP11-162O14 (AC127032.8) and the duplicate H2 copy from the clone RP11-1070B7 (AC139677.4). We also included in the analysis three versions of this sequence from the H1 lineage, clones RP11-259G18 (AC005829.1), XXfos-82115D1 (AC139095.1) and CDT-2107B16 (AC090419.7), and a chimpanzee sequence (National Center for Biotechnology Information Build 1 Version 1, chr19:44,934,906–45,026,647).

We obtained SNP genotypes from the HapMap project database (release 10) for 30 parent-offspring trios of European ancestry for the 17q21.31 region spanning megabases 44.20–44.63 in the sequence assembly of the human genome (Build 34) from the National Center for Biotechnology Information. We constructed haplotypes for the 60 parents in the 30 trios by direct inference from the children's genotypes. We used the NEMO software[38] to estimate the phase and character state of alleles that could not be directly inferred in this way. We used the median-joining algorithm[39] to construct a network describing the genealogical relationships between the parental haplotypes in terms of mutation events. We used the NEMO software to generate haplotypes from alleles at six microsatellites (DG17S137, DG17S7, DG17S338, DG17S10, DG17S340 and DG17S142; see **Supplementary Table 5** online) from a diverse collection of 484 individuals of African, European and Asian ancestry from the Coriell cell repositories. Most of the individuals of African ancestry are African Americans.

Testing for past positive selection of inversion orientation. Selection tests typically exploit the relationship expected under neutral evolution between the frequency of an allele and the extent of variation at closely linked loci given a particular demographic scenario[40]. Recently proposed LD-based tests[21,24] were not appropriate for assessing the impact of positive selection on the chromosome 17 inversion, because the two orientations have different genetic maps even outside the inverted region. Complex modification of these methods would be required to untangle the impact of the inversion itself on LD from any putative signal deriving from natural selection. Instead, we devised a simpler test based on the relative mutational diversity of microsatellites from the inverted region on the background of the two inversion orientation states, H1 and H2. The microsatellite diversity of one orientation state was defined as the average across loci of the average squared difference (ASD) between the repeat sizes of all pairs of alleles, or

$$\mathrm{ASD}_{\mathrm{avg}} = \frac{\sum_{l=1}^{L} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}(r_{il}-r_{jl})^2}{n^2}}{L}, \qquad (1)$$

where L is the number of loci, n is the number of alleles sampled and r_{il} and r_{jl} are the number of repeat units of the ith and jth alleles of locus l. We calculated the difference in the diversity of microsatellites on the H1 and H2 backgrounds in the Utah sample and presented it as a ratio: $\mathrm{ASD}_{\mathrm{avgH2}}/\mathrm{ASD}_{\mathrm{avgH1}}$.

We then used the SIMCOAL2 software[41] to simulate multiple sets of haplotypes assuming neutral evolution and various demographic scenarios. In these coalescent simulations, we used a single surrogate SNP to represent the inversion status of haplotypes. We simulated five microsatellites to represent DG17S137, DG17S7, DG17S338, DG17S10 and DG17S142 (we did not simulate DG17S340 because it shows no diversity in the H1 or H2 lineages). We set mutation rates for the simulated microsatellites at 0.0000288, 0.0018814, 0.0000237, 0.0000347 and 0.000004, respectively (reflecting the varying diversities of the real microsatellites among the H1 haplotypes from the Utah sample). We ran simulations without recombination to ensure independent mutational histories of microsatellites on the different inversion orientation backgrounds until we obtained 5,000 sets of haplotypes in which the frequency of the minor allele for the inversion surrogate SNP was 20% (the frequency of H2 chromosomes in the Utah sample). For each such simulated set of haplotypes, we calculated the $\mathrm{ASD}_{\mathrm{avg}}$ ratio for the different inversion orientations. We estimated the probability of obtaining the observed $\mathrm{ASD}_{\mathrm{avg}}$ ratio under neutrality and a given demographic model as the proportion of 5,000 simulated data sets with a lower $\mathrm{ASD}_{\mathrm{avg}}$ ratio. For four of the five microsatellites, the H2 alleles have fewer repeats than the H1 alleles. As the mutation rate of microsatellites can depend on the number of repeats, a faster mutation rate of H1 microsatellite alleles might contribute to the greater diversity in this lineage. Although our estimates of mutation rates for these five microsatellites in Icelandic parent-offspring trios do not indicate statistically significant differences on H1 and H2 backgrounds (**Supplementary Table 6** online), they have power only to detect differences due to an unusually fast mutation rate in H1 rather than an unusually slow rate in H2.

A more efficient test for positive selection on an inverted region would exploit the full haplotype diversity instead of the average diversity of individual loci. But this would require carrying out coalescent simulations with recombination, where recombination events are restricted between chromosomes depending on their orientation. Such situations have not been explored in detail in the literature, and we know of no publicly available software that can handle inversions for this purpose, although some theoretical steps have been taken in this direction[42].

Detecting association between H2 carrier status, recombination rate and offspring numbers. We used a likelihood approach, using the expectation-maximization algorithm[43] implemented in the NEMO software[38], to estimate haplotype frequencies and identify haplotype carriers. The method used to estimate the recombination rates for mothers and fathers is the same as the likelihood method used in our previous study[21] and was described there in detail.

By comparing the genotyped subjects with everyone born in the same era recorded in the genealogy database (**Supplementary Table 3** online), we noted that people who have more children are over-represented in the genotyped samples. For example, a female with six or more children is nearly five times more likely to be included in the sample than a female who has no children. This is to be expected, as most of our samples were recruited through family studies. Men are also under-represented. Hence, when studying the effects of the H2 chromosomes on number of children and recombination rates, we

carried out weighted regressions in which subjects are weighted inversely to their probability of being included.

Although the estimates from standard weighted regression software are valid, the standard errors and *P* values provided are both smaller than they should be because the individuals in our samples are related and their genotypes are not independent. We determined the presented standard errors and two-sided *P* values empirically. We simulated 10,000 sets of genotypes for H1 and H2 through the whole Icelandic genealogy of 708,683 people. With each simulated set, we carried out all the relevant regressions assuming the simulated genotypes are the actual genotypes of the subjects under study and thereby obtained reference distributions for all the estimated effects under the null hypothesis. We took the standard deviation of an effect estimate in these 10,000 simulations as the empirical standard error, and the two-sided *P* value is the fraction of times the absolute value of the simulated effect is bigger than the actual observed effect. These simulations are necessary because the subjects in the study are related. In particular, the 29,137 individuals used in the number-of-children study often come from genotyping plates in which many members of a nuclear family are included. This leads to an adjustment of the standard error that is an increase between 17% (for the males) and close to 30% (for the females and the combined set) relative to the standard outputs of the multiple regressions. The adjustment for the recombination rate study is much smaller (only ~1% increase) because the set of mothers studied is substantially smaller and much less closely related. Finally, the adjustment factor depends not only on how closely related the subjects are but also on how correlated the phenotype is between close relatives.

The transmission disequilibrium test we carried out for heterozygous parents is mathematically identical to that previously described[44], but in our study, the children are not selected for having a specific trait. Among the 5,529 trios studied, 1,213 of the children were born after 1965, increasing the number of samples genotyped for H1 or H2 status to 30,350.

URLs. We obtained a diverse sample of 484 individuals of African, European and Asian ancestry from the Coriell cell repositories (http://locus.umdnj.edu/ccr/). We obtained SNP genotypes for four populations from the HapMap project database (http://www.hapmap.org/) and genotypes for two SNPs located in *MAPT* from the ALFRED database (http://alfred.med.yale.edu/alfred/index.asp).

GenBank accession number. DG17S142, L77209.

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

ACKNOWLEDGMENTS

We thank D. Reich, N. Patterson and D. Donnelly for constructive comments regarding this work.

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests (see the *Nature Genetics* website for details).

Received 11 November; accepted 17 December 2004
Published online at http://www.nature.com/naturegenetics/

1. Sebat, J. *et al.* Large-scale copy number polymorphism in the human genome. *Science* **305**, 525–528 (2004).
2. Iafrate, A.J. *et al.* Detection of large-scale variation in the human genome. *Nat. Genet.* **36**, 949–951 (2004).
3. Jin, H. *et al.* Structural evolution of the BRCA1 genomic region in primates. *Genomics* **84**, 1071–1082 (2004).
4. Shaw, C.J. & Lupski, J.R. Implications of human genome architecture for rearrangement-based disorders: the genomic basis of disease. *Hum. Mol. Genet.* **13** Spec No 1 R57–R64 (2004).
5. Giglio, S. *et al.* Olfactory receptor-gene clusters, genomic-inversion polymorphisms, and common chromosome rearrangements. *Am. J. Hum. Genet.* **68**, 874–883 (2001).
6. Baker, M. *et al.* Association of an extended haplotype in the tau gene with progressive supranuclear palsy. *Hum. Mol. Genet.* **8**, 711–715 (1999).
7. 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).
8. Skipper, L. *et al.* Linkage Disequilibrium and Association of MAPT H1 in Parkinson Disease. *Am. J. Hum. Genet.* **75**, 669–677 (2004).
9. Oliveira, S.A. *et al.* Linkage disequilibrium and haplotype tagging polymorphisms in the Tau H1 haplotype. *Neurogenetics* **5**, 147–155 (2004).
10. Farrer, M. *et al.* The tau H1 haplotype is associated with Parkinson's disease in the Norwegian population. *Neurosci. Lett.* **322**, 83–86 (2002).
11. Conrad, C. *et al.* Molecular evolution and genetics of the Saitohin gene and tau haplotype in Alzheimer's disease and argyrophilic grain disease. *J. Neurochem.* **89**, 179–188 (2004).
12. Kwok, J.B. *et al.* Tau haplotypes regulate transcription and are associated with Parkinson's disease. *Ann. Neurol.* **55**, 329–334 (2004).
13. Chen, F.C. & Li, W.H. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am. J. Hum. Genet.* **68**, 444–456 (2001).
14. Excoffier, L. Human demographic history: refining the recent African origin model. *Curr. Opin. Genet. Dev.* **12**, 675–682 (2002).
15. The International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796 (2003).
16. The International HapMap Consortium. Integrating ethics and science in the International HapMap Project. *Nat. Rev. Genet.* **5**, 467–475 (2004).
17. Bamshad, M. & Wooding, S.P. Signatures of natural selection in the human genome. *Nat. Rev. Genet.* **4**, 99–111 (2003).
18. Toomajian, C., Ajioka, R.S., Jorde, L.B., Kushner, J.P. & Kreitman, M. A method for detecting recent selection in the human genome from allele age estimates. *Genetics* **165**, 287–297 (2003).
19. Osier, M.V. *et al.* ALFRED: An allele frequency database for anthropology. *Am. J. Phys. Anthropol.* **119**, 77–83 (2002).
20. Evans, W. *et al.* The tau H2 haplotype is almost exclusively Caucasian in origin. *Neurosci. Lett.* **369**, 183–185 (2004).
21. Kong, A. *et al.* Recombination rate and reproductive success in humans. *Nat. Genet.* **36**, 1203–1206 (2004).
22. Kong, A. *et al.* A high-resolution recombination map of the human genome. *Nat. Genet.* **31**, 241–247 (2002).
23. Broman, K.W., Murray, J.C., Sheffield, V.C., White, R.L. & Weber, J.L. Comprehensive human genetic maps: individual and sex-specific variation in recombination. *Am. J. Hum. Genet.* **63**, 861–869 (1998).
24. Schaeffer, S.W. *et al.* Evolutionary genomics of inversions in Drosophila pseudoobscura: evidence for epistasis. *Proc. Natl. Acad. Sci. USA* **100**, 8319–8324 (2003).
25. Carroll, S.B. Genetics and the making of Homo sapiens. *Nature* **422**, 849–857 (2003).
26. Andolfatto, P., Depaulis, F. & Navarro, A. Inversion polymorphisms and nucleotide variability in Drosophila. *Genet. Res.* **77**, 1–8 (2001).
27. Brown, P. *et al.* A new small-bodied hominin from the Late Pleistocene of Flores, Indonesia. *Nature* **431**, 1055–1061 (2004).
28. Reed, D.L., Smith, V.S., Hammond, S.L., Rogers, A.R. & Clayton, D.H. Genetic analysis of lice supports direct contact between modern and archaic humans. *PLoS Biol.* **2**, 340 (2004).
29. Lewin, R. & Foley, R.A. *Principles of Human Evolution* (Blackwell, Oxford, 2004).
30. Sabeti, P.C. *et al.* Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**, 832–837 (2002).
31. Nachman, M.W. & Crowell, S.L. Contrasting evolutionary histories of two introns of the duchenne muscular dystrophy gene, Dmd, in humans. *Genetics* **155**, 1855–1864 (2000).
32. Gilad, Y., Rosenberg, S., Przeworski, M., Lancet, D. & Skorecki, K. Evidence for positive selection and population structure at the human MAO-A gene. *Proc. Natl. Acad. Sci. USA* **99**, 862–867 (2002).
33. Thompson, E.E. *et al.* CYP3A variation and the evolution of salt-sensitivity variants. *Am. J. Hum. Genet.* **75**, 1059–1069 (2004).
34. Akey, J.M. *et al.* Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS Biol.* **2**, e286 (2004).
35. Gulcher, J.R., Kristjansson, K., Gudbjartsson, H. & Stefansson, K. Protection of privacy by third-party encryption in genetic research in Iceland. *Eur. J. Hum. Genet.* **8**, 739–742 (2000).
36. Kumar, S., Tamura, K. & Nei, M. MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Brief Bioinform.* **5**, 150–163 (2004).
37. Schwartz, S. *et al.* Human-mouse alignments with BLASTZ. *Genome Res.* **13**, 103–107 (2003).
38. Gretarsdottir, S. *et al.* The gene encoding phosphodiesterase 4D confers risk of ischemic stroke. *Nat. Genet.* **35**, 131–138 (2003).
39. Bandelt, H.J., Forster, P. & Rohl, A. Median-joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evol.* **16**, 37–48 (1999).
40. Slatkin, M. & Bertorelle, G. The use of intraallelic variability for testing neutrality and estimating population growth rate. *Genetics* **158**, 865–874 (2001).
41. Laval, G. & Excoffier, L. SIMCOAL 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history. *Bioinformatics* **20**, 2485–2487 (2004).
42. Navarro, A., Barbadilla, A. & Ruiz, A. Effect of inversion polymorphism on the neutral nucleotide variability of linked chromosomal regions in Drosophila. *Genetics* **155**, 685–698 (2000).
43. Dempster, A., Laird, N. & Rubin, D. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. B* **39**, 1–38 (1977).
44. Spielman, R.S., McGinnis, R.E. & Ewens, W.J. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet.* **52**, 506–516 (1993).
45. Conrad, C. *et al.* Genetic evidence for the involvement of tau in progressive supranuclear palsy. *Ann. Neurol.* **41**, 277–281 (1997).