Structural Diversity and African Origin of the 17q21.31 Inversion Polymorphism

Karyn Meltz Steinberg[1*], Francesca Antonacci[1*], Peter H. Sudmant[1], Jeffrey M. Kidd[1,9], 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], Kenneth K. Kidd[8], Sarah A. Tishkoff[2], Evan E. Eichler[1,10,11]

SUPPLEMENTARY NOTE

1. Haplotype frequencies in HapMap/1000 Genomes Project and HGDP

We combined the read-depth-based copy number estimates, published array CGH results [1], and publicly available phased SNP genotype information for 1293 nonredundant HapMap and 1000 Genomes Project individuals (99 of the 1392 nonredundant HapMap/1000 Genomes Project sample collection did not have both copy number and genotype data and were excluded from this analysis). We could then assign each phased haplotype to H1', H1D, H2', and H2D haplogroups (Supplementary Figure 4; Supplementary Table 6). Individuals with the H1D haplotype carrying three copies of the H1-specific duplication were assigned to a separate haplogroup designated H1D.3.

The results confirm that the H1D, H2', and H2D haplotypes are nearly absent from Asian and West African populations. In African populations, we only observe the H2' and H2D haplotypes in the Maasai. Previous analyses suggested that all H2' haplotypes from Europe carry CNP155. In this larger cohort, we do observe H2' in Northern Europeans (CEU and British); however, the frequency is much lower when compared to the frequency of H2D with the exception of the Toscani who have appreciable frequencies of H2'. The H1D haplotypes are found exclusively in European and admixed American populations.

We used haplotype-informative SNPs on the phased Illumina 650Y from 936 HGDP samples to assign each haplotype to H1', H2', and H2D haplogroups (Supplementary Table 7). We did not find any common SNPs that could distinguish H1' from H1D in previous analyses and since we did not have copy number estimates we could not separate the H1' from H1D haplotypes. The highest frequencies of H2' are found in the Italians from Tuscany, the Adygei from Russia and the Mbuti Pygmies from the Democratic Republic of Congo; however, the sample sizes are low. The highest frequencies of H2D are found in Europeans and Middle Easterners, although it is found in Central and Southern Asia at lower frequencies. The H2 haplotype is virtually absent from Eastern Asians, consistent with the HapMap and 1000 Genomes Project observations, with the notable exceptions of the Mongolians and the Yakut. The

H2 haplotype is also absent from the sub-Saharan Africans with the exception of the pygmy populations who were found to only carry the H2' and not the H2D haplotype in this sample set.

2. Double recombination event

We aligned the H1 and H2D haplotypes from the RP11 BAC assembly and analyzed sequence divergence in 5 kbp sliding windows across 240 kbp of aligned sequence. The region of 30 kbp where sequence divergence is significantly reduced is at these genomic coordinates: NCBI build36: chr17:41213364-41248960; GRCh37: chr17:43857600-43893180. When building the haplotype networks we used 10 polymorphic SNPs in the region of reduced diversity, and we chose 10 additional polymorphic SNPs approximately 100 kbp upstream and a region of 10 polymorphic SNPs approximately 100 kbp downstream as a comparison.

SNP array data may not accurately reflect whether the reciprocal event is present in the human population due to ascertainment biases. We therefore examined the phased genotype data from whole-genome sequence from 627 individuals from the 1000 Genomes Project to determine if the reciprocal event could be detected. We did not detect any individual who showed greater genetic diversity over the 30 kbp interval that would indicate the presence of the reciprocal event. One caveat is that the 1000 Genomes Project reference sequence does not contain the alternative haplotype (GRCh37: chr17_ctg5_hap1) that is represents the H2 haplotype. Hence if the donor haplotype of the double recombination event was an H1 we would not be able to detect the reciprocal event if the H2 haplotype is not represented in the assembly.

3. Diversity between RP11 and ABC14

We used publicly available end sequence data from the NA12156 fosmid library [2] to calculate the genome-wide heterozygosity and percent identity to the reference sequence (RP11). We filtered reads that fell within segmental duplications. We then calculated the number of SNPs found in each read

and calculated heterozygosity from unique SNPs in 100 kbp windows. For the autosomes the average heterozygosity is 0.000943 ± 0.000597 and the percent identity is 99.91%.

4. Whole-genome alignment of 136 kbp LD block

It is possible that there are still recombination events not captured by HapMap SNPs so we took the whole-genome sequence alignment from the 136 kbp LD block and tested for recombination using the 4 Gamete Test. Rm values for the H1 clade, the H2 clade and the H1/H2 clades together were zero suggesting that there were no recombination events within this interval consistent with the HapMap observation. We used all sequences except the NA12156 fosmid sequence and the orangutan for this analysis. Overall the sequence identity of this region in humans is 99.6% and pi is equal to 0.00210. Among the H1 sequences π was 0.0002 and among the H2 sequences π was 0.00005. We then estimated evolutionary age using the Kimura 2 parameter distances based on these sequences. We estimate that the H1 and H2 haplotypes are 2.70 ± 0.13 million years old, the KB1/NA21599 haplotypes are 77,000 ± 22,000 years, the H1 and H1D haplotypes are 121,000 ± 36,000 years old.

5. Estimation of sequence divergence and coalescence times from phylogenetic trees

The number of mutations per branch was calculated by multiplying the branch length by the number of total sites. We also calculated Kimura 2-parameter model genetic distances and standard errors using MEGA4 (Supplementary Table 9). We estimated the average sequence divergence of human (H1,H2) vs. chimpanzee distance K=(0.010752+0.010847)/2=0.0107995. Using chimpanzee as the outgroup, we calculated the average substitution rate using the equation R=K/2T (0.0107995/12 mya)=8.99958×10^{-4} per site/mya. Using the above substitution rate (R) and sequence divergence (Kimura-2 parameter, K), we estimated the coalescence time (T=K/2R) of the haplotypes.

For CNP205 we identified the approximate duplication coordinates based on read-depth analysis (GRCh37: chr17:44183556-44294406) excluding segmental duplications for a total of 103,945 bp. We

then aligned the NA12878 H1D sequence to the RP11 H1' sequence. There are 48 differences for an overall sequence identity of 99.954%. We aligned the chimpanzee sequence for this region estimate the coalescence times. For CNP155, the region is 76,745 bp excluding segmental duplications (GRCh37: chr17:44210855-44294624; chr17_ctg5_hap1: 571621-655190). When we compare the NA21599 H2D sequence to the reference H2D sequence with the derived duplication masked out, there are 153 differences for an overall sequence identity of 99.8%. We aligned the chimpanzee sequence and estimated the evolutionary age of the H2 duplication.

6. Functional differences between H1 and H2 haplotypes

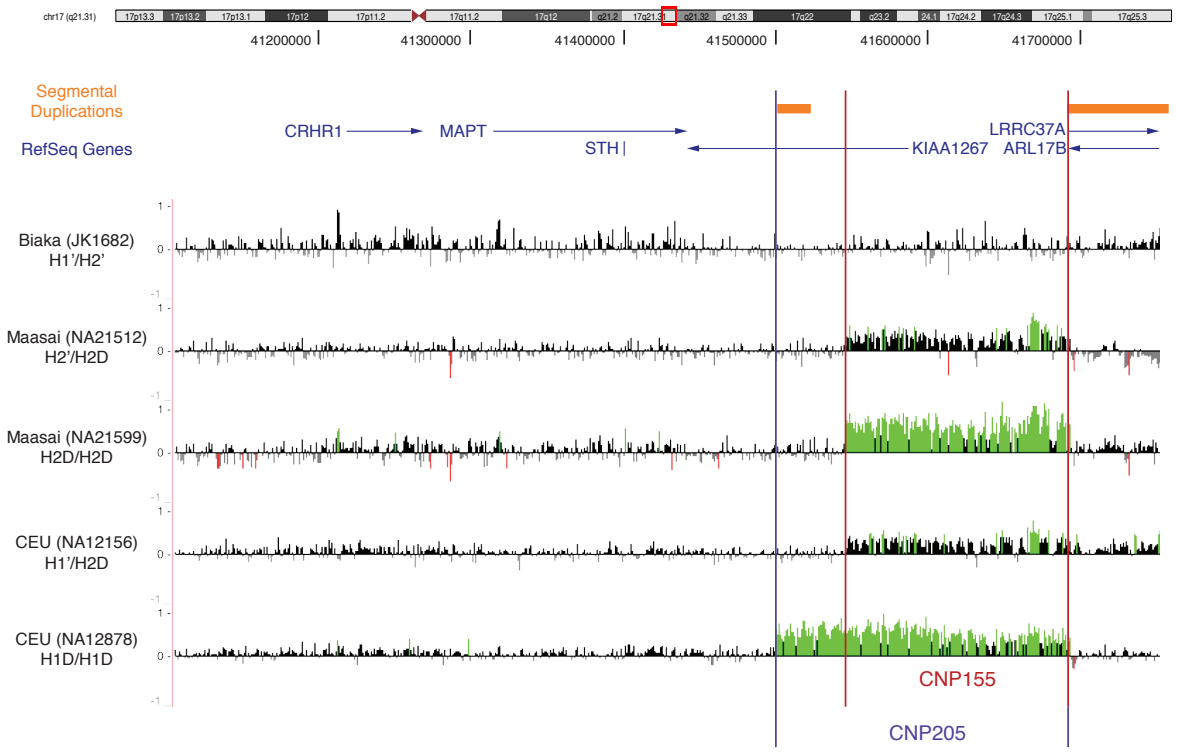
We identified 887 parsimony-informative sites between the H1 and H2 haplotypes; there are 9 missense, 9 synonymous, 22 UTR, and 269 intronic mutations on the H2 haplotype (Supplementary Table 10). Seven of the nine missense mutations are in the gene *IMP5* (intramembrane protease 5), and four of these are predicted to alter the protein structure by PolyPhen. The H1 alleles of two of these four amino acid altering mutations have been previously associated with Parkinson Disease [3] and therefore the H2 missense mutations may represent substitutions that could be under positive selection. The two remaining missense mutations are in *KIAA1267*. Both variants are not predicted to alter the amino acid as predicted by PolyPhen, but they are highly conserved (GERP scores between 2 and 4). We tested whether the variant alleles were in LD with the H2 haplotype in the 1000 Genomes Project individuals using the plink ld test on the phased genotype data. The seven *IMP5* and one of the *KIAA1267* (44108906) variants are in LD with the H2 haplotype (R^2 = 0.995, D'=1 for all eight variants). The remaining *KIAA1267* variant at position 44144993 is found in the Maasai H2D homozygote NA21599 and not found in any other 1000 Genomes Project samples. As it is not in dbSNP or HapMap, we could not assess its frequency in the Maasai HapMap population. We did not find any gene disruptive indels.

7. H1 haplotype disease associations

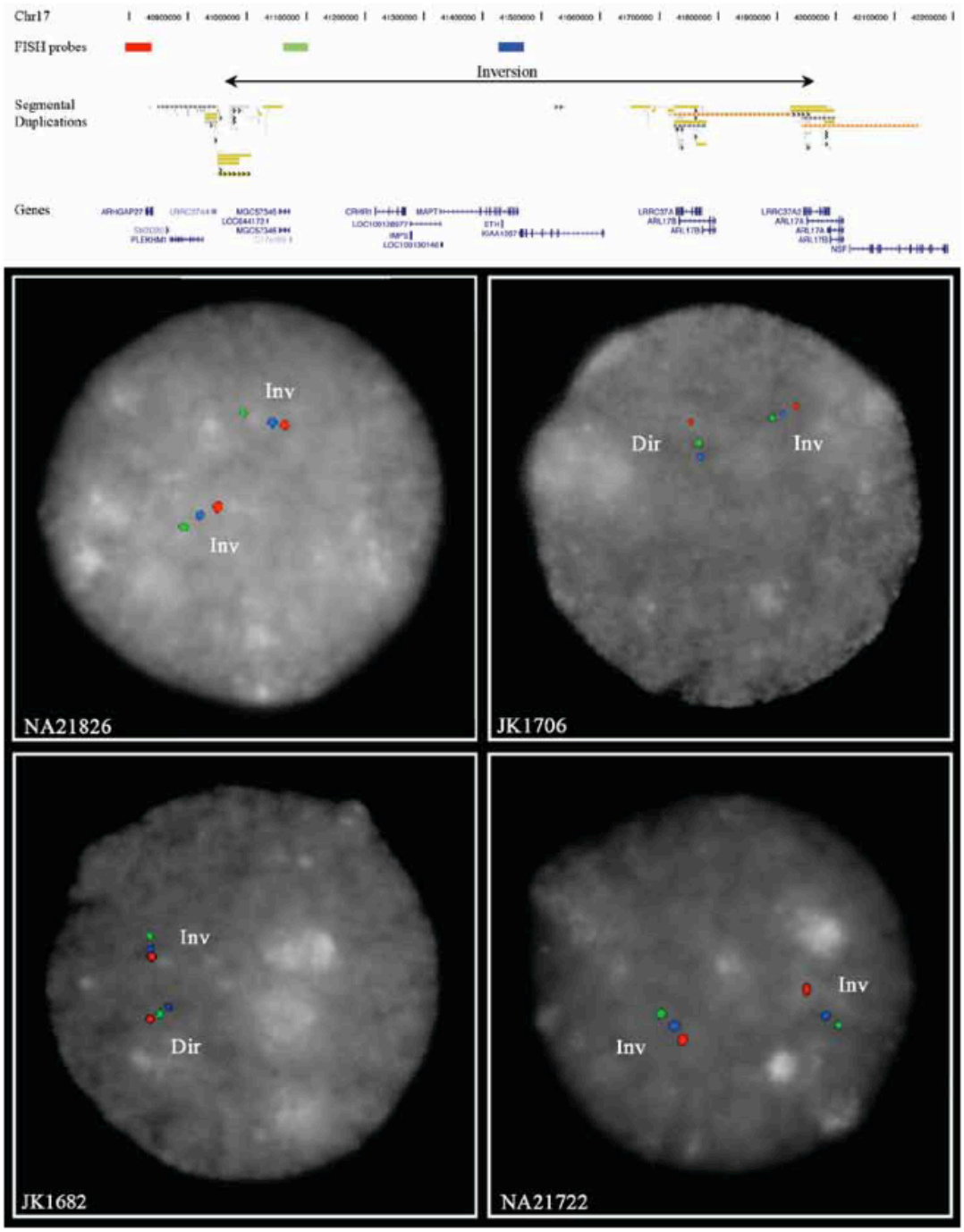
Previous studies have implicated extended 17q21.31 haplotypes with various neurodegenerative diseases. We performed a literature search and gathered published haplotype associations with Parkinson's disease (PD) [4,5], progressive supranuclear palsy (PSP) and corticobasal degeneration (CBD), [6,7] and Alzheimer's disease (AD) [4]. Of the 11 published haplotypes we examined, two H2 haplotypes were protective, two H1 haplotypes were PD risk haplotypes, one H1 haplotype was an AD/PSP risk haplotype, and the rest were common H1 haplotypes (frequencies higher than 5% in the original studies) that were not associated with risk or protection from disease.

We then utilized the phased genotype data from the 1000 Genomes Project to assign risk/protective haplotypes to the European individuals. The frequencies of each haplotype are presented in Supplementary Table 11. We assigned a genotype to each individual based on the presence or absence of the haplotype as well as a genotype based on the presence or absence of the haplotype specific duplications. We then searched for associations between particular haplotypes and duplications. As expected, the H2 duplication was in LD with the H2a and H2 haplotypes identified in Tobin *et al.* [5] more often than predicted by linkage equilibrium (LE) (Fisher's exact test two-tailed p=0). The H1 duplication was in LD with the H1b (non-risk) haplotype more often than predicted by LE (Fisher's exact test two-tailed $p<0.0001$) and less often than predicted by LE in the H1c (AD/PSP risk) haplotype (Fisher's exact test two tailed $p<0.0001$). Notably, the H1c haplotype has been shown to increase expression of total *MAPT* transcripts in the brain[8] and tau levels in cerebrospinal fluid [9]. In addition, the H1 duplication was never found on the H1p (PD risk) haplotype and rarely found on the risk haplotype identified in Tobin *et al.* [5]; however, given the sample size and low frequencies of risk haplotypes, we were unable to assess whether these associations were significant. These observations suggest that the PD and AD risk haplotypes likely arose on unduplicated H1 haplotypes and warrant further investigation on the possible protective role of the H1 duplication in these diseases.

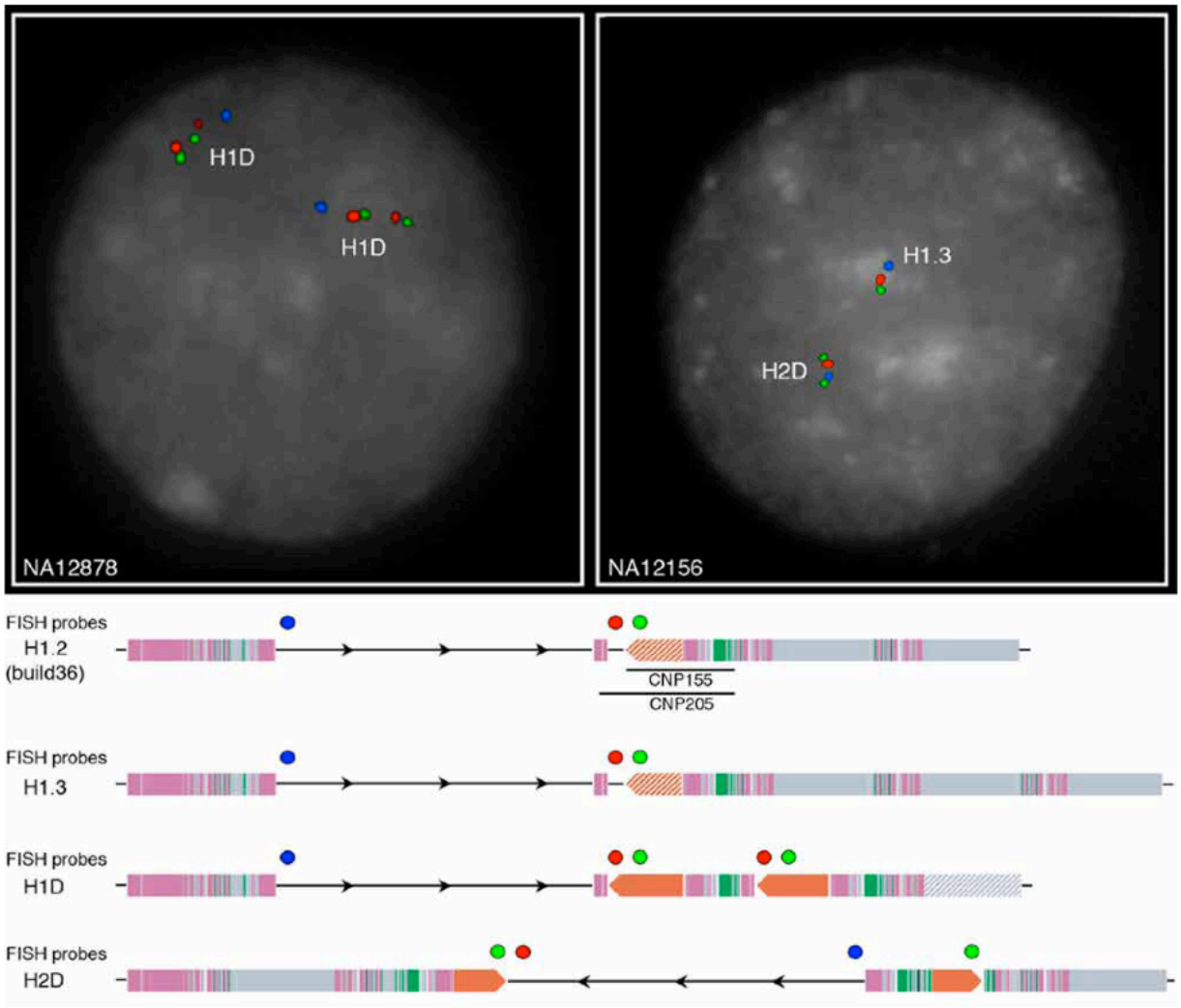
SUPPLEMENTARY FIGURES



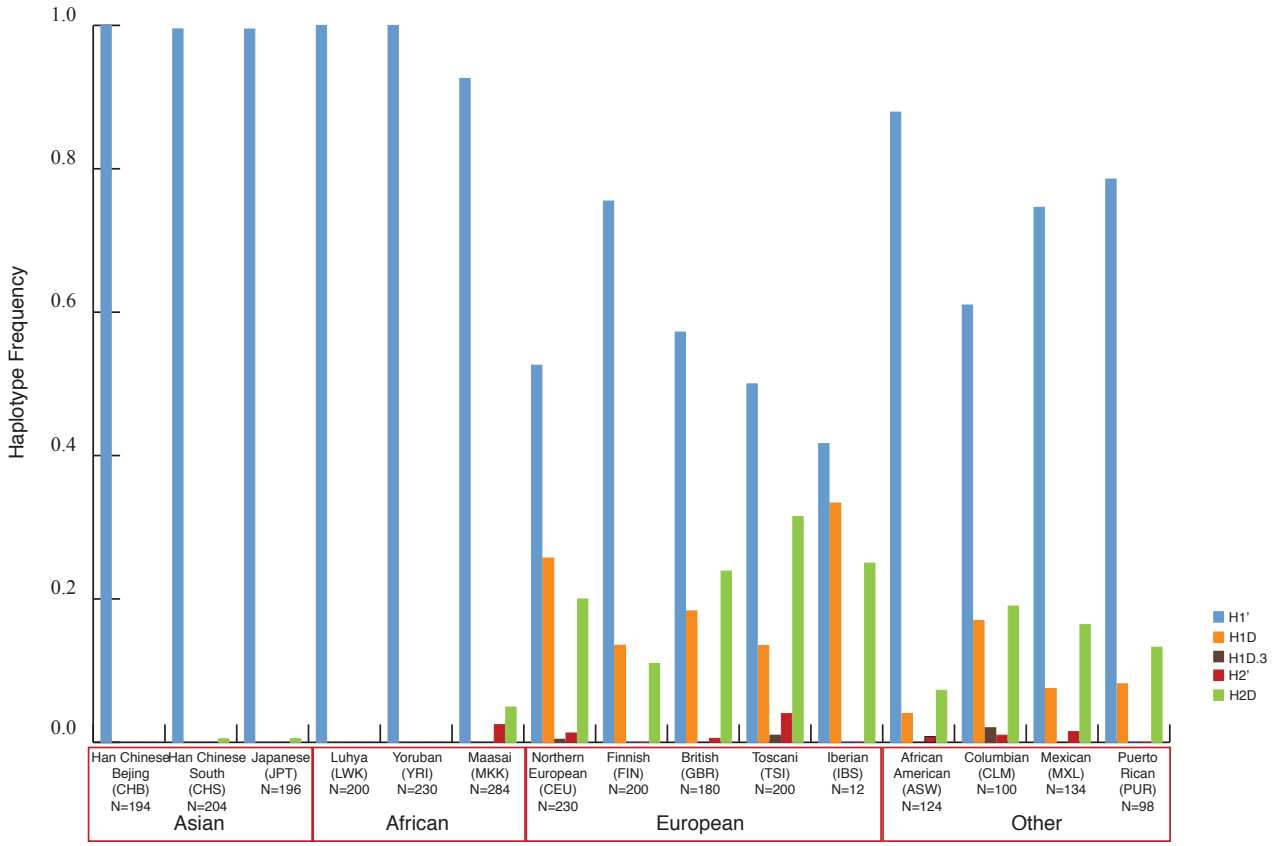
Supplementary Figure 1. Inversion analysis by FISH. Interphase triple-color FISH using two probes inside the 17q21.31 inversion (WIBR2-2095G22 in red; WIBR2-3237D21 in green) and one outside (WIBR2-2567E17 in blue) shows the inversion of the region in two heterozygous Biaka individuals (JK1706 and JK1682) and two homozygous Maasai individuals (NA21826 and NA21722). Inv = inverted orientation; Dir = direct orientation. The top panel shows a screenshot of the UCSC Genome Browser (build36) and the location of the probes used by FISH.



Supplementary Figure 2. Array CGH results. Previous analyses indicated that the duplication seen on all European H2 chromosomes was absent from the African population [10]. However, these analyses only included Yorubans in which the H2 haplotype is absent. Results from this array CGH experiment demonstrate that the H2-specific duplication is polymorphic in Africans with the H2 haplotype. Also shown is the H1-specific duplication seen only in Europeans.



Supplementary Figure 3. FISH confirmation of proximal breakpoint. FISH cohybridization experiments using a probe tagging both CNP155 and CNP205 (green probe) and a probe tagging uniquely CNP205 (red probe) show that the proximal breakpoints of the duplications are different and the duplications map in different locations (tandem duplication on H1D and interspersed on H2D), suggesting that the two duplications are independent events.



Supplementary Figure 4. Distribution of 17q21.31 haplotypes. A total of 1059 individuals (2118 chromosomes) from three major continental groups (Africa, Asia, and Europe) and 234 admixed individuals (468 chromosomes) were assigned haplotypes based on copy number estimates and phased genotype data. The populations are followed by the number of chromosomes in parentheses. The H1D haplotypes are found exclusively in the European populations, and although the H2' and H2D haplotypes are found predominately in these populations, they are also present in the Maasai population and at extremely low frequencies (<0.5%) in Asian populations.

SUPPLEMENTARY TABLES (Supplementary Table 2 in separate Excel file)

Supplementary Table 1. Sample collections

Sample Collection	Total	African	European	Asian	Other	Source	Reference
Non-redundant, unrelated HapMap and 1000 Genomes	1392	361	421	303	307		
HapMap Phased SNP*	728	372	154	159	43	Illumina 1M and Affymetrix 6.0	HapMap Project
HapMap ArrayCGH*	1209	350	388	340	131	Custom arrayCGH	Campbell et al (2011) and present study
1000 Genomes Phased SNP*	1211	204	446	302	259	Illumina Omni2.5M	1000 Genomes Project
1000 Genomes Genotypes*	627	145	260	177	45	Whole Genome Sequence	1000 Genomes Project
1000 Genomes Read Depth*	805	139	261	220	185	Whole Genome Sequence	Sudmant et al (2010) and present study
HGDP Phased SNP	936	101	317	427	91	Illumina 650Y	HGDP
African Diversity Panel	275	275	0	0	0	Illumina 1M	Tishkoff et al (2009) and present study
Hunter-Gatherer Panel	76	76	0	0	0	Illumina 550K	Henn et al (2011)
Bushman Panel	4	4	0	0	0	Whole Genome Sequence	Schuster et al (2010)
H2 Diversity Panel	17	3	9	2	3	TaqMan/Illumina Bead Array/Custom arrayCGH	Donnelly et al (2010) and present study
Total Non-redundant	**2700**	**820**	**747**	**732**	**401**		

Supplementary Table 3. FISH experiments (a) Samples and FISH cohybridizations (b) Clones used for FISH cohybridizations

Sample ID	Ethnicity	Genotype	FISH cohybridizations	Sample Collection
HG01113	CLM	H2D/H1D.3	(1,2,3), (1,2,4)	HapMap/1000 Genomes Project
HG01465	CLM	H2D/H1D.3	(1,2,3), (1,2,4)	HapMap/1000 Genomes Project
JK1600	Maya	H2D/H2D	(1,2,3), (1,2,4), (1,4,6), (2,4,6), (4,6,7)	H2 Diversity Panel
JK1682	Biaka	H1.3/H2.2	(1,2,3), (1,2,4), (1,4,6), (2,4,6), (4,6,7)	H2 Diversity Panel
JK1706	Biaka	H1.3/H2D	(1,2,3), (1,2,4), (1,4,6), (2,4,6), (4,6,7)	H2 Diversity Panel
JK2356	Quechua	H1'/H2'	(1,2,3)	H2 Diversity Panel
JK3035	TWChinese	H1.3/H2D	(1,2,3), (1,2,4), (1,4,6), (2,4,6), (4,6,7)	H2 Diversity Panel
JK3768	PrimaMx	H1'/H2D	(1,2,3)	H2 Diversity Panel
JK4155	Chagga	H1'/H2D	(1,2,3)	H2 Diversity Panel
NA07346	CEU	H1.2/H1D	(1,2,3), (2,4,6)	HapMap/1000 Genomes Project
NA11839	CEU	H1D/H2'	(1,2,3), (1,2,4)	HapMap/1000 Genomes Project
NA11918	CEU	H1.1/H1D	(1,2,3), (2,4,6)	HapMap/1000 Genomes Project
NA12156	CEU	H1.3/H2D	(1,2,3), (1,2,4), (1,4,6), (2,4,6), (4,6,7)	HapMap/1000 Genomes Project
NA12717	CEU	H1D/H1D.3	(1,2,3), (2,4,6)	HapMap/1000 Genomes Project
NA12878	CEU	H1D/H1D	(1,2,3), (1,2,4), (1,4,6), (2,4,6), (4,6,7)	HapMap/1000 Genomes Project
NA19314	LWK	H1'/H1'	(1,2,3), (1,2,4)	HapMap/1000 Genomes Project
NA19436	LWK	H1'/H1'	(1,2,3), (1,2,4)	HapMap/1000 Genomes Project
NA19723	MEX	H1'/H2'	(1,2,3), (1,2,4)	HapMap/1000 Genomes Project
NA19749	MEX	H2'/H2D	(1,2,3), (1,2,4)	HapMap/1000 Genomes Project
NA19828	ASW	H2D/H2D	(1,2,3)	HapMap/1000 Genomes Project
NA20509	TSI	H2'/H2D	(1,2,3), (1,2,4)	HapMap/1000 Genomes Project
NA20516	TSI	H1'/H2'	(1,2,3), (1,2,4)	HapMap/1000 Genomes Project
NA20589	TSI	H2'/H2'	(1,2,3), (1,2,4)	HapMap/1000 Genomes Project
NA20765	CEU	H2D/H1D.3	(1,2,3), (1,2,4), (1,2,5)	HapMap/1000 Genomes Project
NA20811	TSI	H2D/H2D	(1,2,3), (1,2,4)	HapMap/1000 Genomes Project
NA20828	TSI	H2'/H2D	(1,2,3), (1,2,4)	HapMap/1000 Genomes Project
NA21512	MKK	H2.1/H2D	(1,2,3), (1,2,4), (1,4,6), (2,4,6), (4,6,7)	HapMap/1000 Genomes Project
NA21599	MKK	H2D/H2D	(1,2,3), (1,2,4), (1,4,6), (2,4,6), (4,6,7)	HapMap/1000 Genomes Project
NA21722	MKK	H2.2/H2D	(1,2,3), (1,2,4), (1,4,6), (2,4,6), (4,6,7)	HapMap/1000 Genomes Project
NA21826	MKK	H2D/H2D	(1,2,3)	HapMap/1000 Genomes Project

Clone number	Clone name	Chr	Start	End
1	WIBR2-2095G22	chr17	40794845	40837752
2	WIBR2-3237D21	chr17	41061443	41102662
3	WIBR2-2567E17	chr17	41427862	41469027
4	WIBR2-2342H02	chr17	41576607	41613815
5	ABC8_40982400_I3	chr17	41537304	41572792
6	WIBR2-1321L07	chr17	41810516	41854071
7	WIBR2-2606H15	chr17	42142766	42180535

Supplementary Table 4. ArrayCGH experiments

Sample ID	Population	Genotype	Sample Collection
JK1682	Biaka	H1'/H2'	H2 Diversity Panel
KB1	San	H1'/H2'	Bushman Panel
NA21339	Maasai	H1'/H2'	HapMap/1000 Genomes Project
NA21438	Maasai	H1'/H2'	HapMap/1000 Genomes Project
NA21580	Maasai	H1'/H2'	HapMap/1000 Genomes Project
NA21685	Maasai	H1'/H2'	HapMap/1000 Genomes Project
NA21741	Maasai	H1'/H2'	HapMap/1000 Genomes Project
JK1706	Biaka	H1'/H2D	H2 Diversity Panel
JK4155	Chagga	H1'/H2D	H2 Diversity Panel
NA21353	Maasai	H1'/H2D	HapMap/1000 Genomes Project
NA21408	Maasai	H1'/H2D	HapMap/1000 Genomes Project
NA21417	Maasai	H1'/H2D	HapMap/1000 Genomes Project
NA21615	Maasai	H1'/H2D	HapMap/1000 Genomes Project
NA21632	Maasai	H1'/H2D	HapMap/1000 Genomes Project
NA21683	Maasai	H1'/H2D	HapMap/1000 Genomes Project
NA21742	Maasai	H1'/H2D	HapMap/1000 Genomes Project
NA21743	Maasai	H1'/H2D	HapMap/1000 Genomes Project
NA21512	Maasai	H2'/H2D	HapMap/1000 Genomes Project
NA21722	Maasai	H2'/H2D	HapMap/1000 Genomes Project
NA21599	Maasai	H2D/H2D	HapMap/1000 Genomes Project
NA21826	Maasai	H2D/H2D	HapMap/1000 Genomes Project

Supplementary Table 5. Haplotype nomenclature for the two independent studies

Steinberg et al	Boettger et al
H1.1	H1.β1.γ1
H1.2	H1.β1.γ2
H1.3	H1.β1.γ3
	H1.β1.γ4
H1D	H1.β2.γ1
H1D.3	H1.β3.γ1
H2.1	
H2.2	H2.α1.γ2
	H2.α2.γ1
H2D	H2.α2.γ2

CNP155 = α, CNP205 = β, CNP210 = γ

Supplementary Table 6. Frequencies of H1', H1D, H1D.3, H2' and H2D in 1000 Genomes Project and HapMap Samples

Continent	Country	Population	Frequency of H1'	Frequency of H1D	Frequency of H1D.3	Frequency of H2'	Frequency of H2D	TOTAL (Number of Chromosomes)
Asia	China	Han Chinese Bejing	100.0%	0.0%	0.0%	0.0%	0.0%	194
Asia	China	Han Chinese South	99.5%	0.0%	0.0%	0.0%	0.5%	204
Asia	Japan	Japanese	99.5%	0.0%	0.0%	0.0%	0.5%	196
Africa	Kenya	Luhya	100.0%	0.0%	0.0%	0.0%	0.0%	200
Africa	Nigeria	Yoruban	100.0%	0.0%	0.0%	0.0%	0.0%	230
Africa	Kenya	Maasai	92.6%	0.0%	0.0%	2.5%	4.9%	284
Europe	Northern Europe	CEU	52.6%	25.7%	0.4%	1.3%	20.0%	230
Europe	Finland	Finnish	75.5%	13.5%	0.0%	0.0%	11.0%	200
Europe	England/Scotland	British	57.2%	18.3%	0.0%	0.6%	23.9%	180
Europe	Italy	Toscani	50.0%	13.5%	1.0%	4.0%	31.5%	200
Europe	Spain	Iberian	41.7%	33.3%	0.0%	0.0%	25.0%	12
Other	Southwest, US	African American SW	87.9%	4.0%	0.0%	0.8%	7.3%	124
Other	Columbia	Columbian	61.0%	17.0%	2.0%	1.0%	19.0%	100
Other	Los Angeles, US	Mexican	74.6%	7.5%	0.0%	1.5%	16.4%	134
Other	Puerto Rico	Puerto Rican	78.6%	8.2%	0.0%	0.0%	13.3%	98

Supplementary Table 7. Frequencies of H1', H2' and H2D in HGDP samples

Population	Geographic Origin	Region	Number of H1' Haplotypes	Haplotype Frequency of H1'	Number of H2' Haplotypes	Haplotype Frequency of H2'	Number of H2D Haplotypes	Haplotype Frequency of H2D	Total
Bantu N.E.	Kenya	Subsaharan Africa	22	100.0%	0	0.0%	0	0.0%	22
Bantu S.E.	South Africa	Subsaharan Africa	10	100.0%	0	0.0%	0	0.0%	10
Bantu S.W.	South Africa	Subsaharan Africa	6	100.0%	0	0.0%	0	0.0%	6
Biaka Pygmies	Central African Republic	Subsaharan Africa	41	97.6%	1	2.4%	0	0.0%	42
Mandenka	Senegal	Subsaharan Africa	42	100.0%	0	0.0%	0	0.0%	42
Mbuti Pygmies	Democratic Republic of Congo	Subsaharan Africa	24	92.3%	2	7.7%	0	0.0%	26
San	Namibia	Subsaharan Africa	10	100.0%	0	0.0%	0	0.0%	10
Yoruba	Nigeria	Subsaharan Africa	42	100.0%	0	0.0%	0	0.0%	42
Colombians	Colombia	America	14	100.0%	0	0.0%	0	0.0%	14
Karitiana	Brazil	America	28	100.0%	0	0.0%	0	0.0%	28
Maya	Mexico	America	40	95.2%	0	0.0%	2	4.8%	42
Pima	Mexico	America	28	100.0%	0	0.0%	0	0.0%	28
Surui	Brazil	America	16	100.0%	0	0.0%	0	0.0%	16
Balochi	Pakistan	Central/Southern Asia	44	91.7%	1	2.1%	3	6.3%	48
Brahui	Pakistan	Central/Southern Asia	47	94.0%	2	4.0%	1	2.0%	50
Burusho	Pakistan	Central/Southern Asia	47	94.0%	0	0.0%	3	6.0%	50
Hazara	Pakistan	Central/Southern Asia	43	97.7%	1	2.3%	0	0.0%	44
Kalash	Pakistan	Central/Southern Asia	42	91.3%	0	0.0%	4	8.7%	46
Makrani	Pakistan	Central/Southern Asia	43	86.0%	3	6.0%	4	8.0%	50
Pathan	Pakistan	Central/Southern Asia	40	90.9%	0	0.0%	4	9.1%	44
Sindhi	Pakistan	Central/Southern Asia	44	91.7%	1	2.1%	3	6.3%	48
Uygur	China	Central/Southern Asia	20	100.0%	0	0.0%	0	0.0%	20
Cambodians	Cambodia	East Asia	20	100.0%	0	0.0%	0	0.0%	20
Dai	China	East Asia	20	100.0%	0	0.0%	0	0.0%	20
Daur	China	East Asia	18	100.0%	0	0.0%	0	0.0%	18
Han	China	East Asia	87	98.9%	0	0.0%	1	1.1%	88
Hezhen	China	East Asia	16	100.0%	0	0.0%	0	0.0%	16
Japanese	Japan	East Asia	56	100.0%	0	0.0%	0	0.0%	56
Lahu	China	East Asia	16	100.0%	0	0.0%	0	0.0%	16
Miaozu	China	East Asia	20	100.0%	0	0.0%	0	0.0%	20
Mongola	China	East Asia	19	95.0%	0	0.0%	1	5.0%	20
Naxi	China	East Asia	16	100.0%	0	0.0%	0	0.0%	16
Oroqen	China	East Asia	18	100.0%	0	0.0%	0	0.0%	18
She	China	East Asia	20	100.0%	0	0.0%	0	0.0%	20
Tu	China	East Asia	20	100.0%	0	0.0%	0	0.0%	20
Tujia	China	East Asia	20	100.0%	0	0.0%	0	0.0%	20
Xibo	China	East Asia	18	100.0%	0	0.0%	0	0.0%	18
Yakut	Siberia	East Asia	47	97.9%	0	0.0%	1	2.1%	48
Yizu	China	East Asia	20	100.0%	0	0.0%	0	0.0%	20
Adygei	Russia Caucasus	Europe	27	79.4%	4	11.8%	3	8.8%	34
French	France	Europe	45	80.4%	1	1.8%	10	17.9%	56
French Basque	France	Europe	35	72.9%	1	2.1%	12	25.0%	48
North Italian	Italy (Bergamo)	Europe	15	68.2%	1	4.5%	6	27.3%	22
Orcadian	Orkney Islands	Europe	22	73.3%	0	0.0%	8	26.7%	30
Russian	Russia	Europe	45	90.0%	0	0.0%	5	10.0%	50
Sardinian	Italy	Europe	40	71.4%	3	5.4%	13	23.2%	56
Tuscan	Italy	Europe	11	68.8%	3	18.8%	2	12.5%	16
Bedouin	Israel (Negev)	Middle East	69	76.7%	2	2.2%	19	21.1%	90
Druze	Israel (Carmel)	Middle East	60	73.2%	2	2.4%	20	24.4%	82
Palestinian	Israel (Central)	Middle East	69	75.0%	2	2.2%	21	22.8%	92
Mozabite	Algeria (Mzab)	Middle East	50	86.2%	2	3.4%	6	10.3%	58
NAN Melanesian	Bougainville	Oceania	20	100.0%	0	0.0%	0	0.0%	20
Papuan	New Guinea	Oceania	34	100.0%	0	0.0%	0	0.0%	34

Supplementary Table 8. Inversion and duplication tagging SNPs

SNP	Genomic position (b36)	Genomic position (b37)	H1'	H2'	H2D	Reference
rs241039	41070456	43714673	A	T	T	Donnelly et al (2010)
rs434428	41081467	43725684	G	A	A	Donnelly et al (2010)
rs241027	41091261	43735478	A	G	G	Donnelly et al (2010)
rs2049515	41117639	43761856	C	T	T	Donnelly et al (2010)
rs10491144	41128907	43773124	A	C	C	Donnelly et al (2010)
rs10514879	41158754	43802971	C	T	T	Donnelly et al (2010)
rs2902662	41162708	43806925	G	A	A	Donnelly et al (2010)
rs17563599	41163726	43807955	A	C	C	Present study
rs11079718	41195723	43839951	A	T	T	Donnelly et al (2010)
rs1396862	41258778	43902997	G	A	A	Donnelly et al (2010)
rs1078830	41301901	43946112	T	C	C	Donnelly et al (2010)
rs916793	41310477	43954686	G	A	A	Donnelly et al (2010)
rs17563986	41347100	43991272	A	G	G	Present study
rs17650901	41395527	44039691	T	C	C	Donnelly et al (2010)
rs1800547	41407682	44051846	A	G	G	Stefansson et al (2005)
rs17651213	41407760	44051924	G	A	A	Donnelly et al (2010)
rs1981997	41412603	44056767	G	A	A	Present study
rs1052553	41429726	44073889	A	G	G	Donnelly et al (2010)
rs8070723	41436901	44081064	A	G	G	Present study
rs9468	41457408	44101563	T	C	C	Stefansson et al (2005)
rs12150447	41483977	44128125	A	C	C	Donnelly et al (2010)
rs2838	41497167	44141347	A	G	G	Donnelly et al (2010)
rs1468241	41551932	44196153	A	C	C	Donnelly et al (2010)
rs1528075	41576231	44220454	T	G	G	Donnelly et al (2010)
rs1528072	41592502	44236725	C	A	A	Donnelly et al (2010)
rs2668692	41648797	44293020	C	T	T	Present study
rs2957297	41723989	44368212	A	G	G	Present study
rs199457	42150653	44795469	C	C	T	Present study
rs199456	42153103	44797919	C	C	T	Present study
rs199451	42156968	44801784	G	G	A	Present study
rs199448	42164185	44809001	A	A	G	Present study
rs199533	42184098	44828931	C	C	T	Present study

Supplementary Table 9. Kimura 2-parameter distances based on 204,447 bp alignment. Distances in lower left, standard errors in upper right

	RP11 (H1')	NA12878 (H1D)	NA12156 (H2D)	RP11 (H2D)	NA20589 (H2')	NA21599 (H2D)	KB1 (H2')	Chimpanzee	Orangutan
RP11 (H1')		0.000047	0.000142	0.000142	0.000142	0.000142	0.000144	0.000231	0.000417
NA12878 (H1D)	0.00045		0.000142	0.000142	0.000143	0.000143	0.000144	0.000232	0.000418
NA12156 (H2D)	0.004097	0.004112		0.000005	0.000015	0.000017	0.000032	0.000231	0.000416
RP11 (H2D)	0.004093	0.004117	0.000005		0.000015	0.000017	0.000032	0.000231	0.000417
NA20589 (H2')	0.004122	0.004137	0.000044	0.000049		0.000021	0.000035	0.000231	0.000417
NA21599 (H2D)	0.004122	0.004137	0.000054	0.000059	0.000088		0.000035	0.000231	0.000417
KB1 (H2')	0.004191	0.004206	0.00021	0.000215	0.000254	0.000245		0.000232	0.000417
Chimpanzee	0.010752	0.010827	0.010743	0.010738	0.010758	0.010768	0.010847		0.000415
Orangutan	0.034019	0.034092	0.033866	0.033861	0.033892	0.033892	0.03397	0.033674	

Supplementary Table 10. Gene Disruptive Mutations on the H2 haplotype

GRCh37 Position	Reference Allele	Variant Allele	Chimpanzee Allele	rsID	Gene	Amino Acid Change	Protein Position	PolyPhen Prediction	PhastCons Score	GERP Score
43922942	T	C	C	62621252	*IMP5*	SER,PRO	224/685	benign	0.006	0.325
43923266	G	A	A	62054815	*IMP5*	ALA,THR	332/685	benign	0	-8.31
43923654	G	C	G	12185233	*IMP5*	ARG,PRO	461/685	probably-damaging	1	4.74
43923683	A	G	G	12185268	*IMP5*	ILE,VAL	471/685	benign	0.302	-1.22
43924073	T	C	T	12185268	*IMP5*	SER,PRO	601/685	probably-damaging	0.275	4.7
43924130	G	A	G	12373123	*IMP5*	GLY,ARG	620/685	probably-damaging	0	-4.2
43924200	C	G	C	12373139	*IMP5*	PRO,ARG	643/685	probably-damaging	0	-1.44
44108906	A	G	G	34579536	*KIAA1267*	ILE,THR	1085/1106	benign	0.001	2.79
44144993	C	G	G	NA	*KIAA1267*	ARG,PRO	525/1106	benign	0.948	3.99

Supplementary Table 11. Disease Associated Haplotype Analysis

HAPLOTYPE	DISEASE ASSOCIATION	FREQUENCY IN 1000 GENOMES EUROPEANS	REFERENCE
H2a	PD/PSP/CBD Protective	23.0%	4,6
H1b		15.1%	4,6
H1c	AD/PSP risk	14.7%	4,6,7
H1d		7.0%	4,6
H1e		0.0%	4,6
H1m		2.2%	4,6
H1p	PD risk	1.0%	4,6
H1 Major Allele		30.6%	5
H1 Common		20.8%	5
Risk	PD risk	3.6%	5
H2	PD Protective	22.8%	5

REFERENCES

1. Campbell, C.D. *et al.* Population-genetic properties of differentiated human copy-number polymorphisms. *Am J Hum Genet* **88**, 317-32 (2011).
2. Kidd, J.M. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56-64 (2008).
3. Simon-Sanchez, J. *et al.* Genome-wide association study reveals genetic risk underlying Parkinson's disease. *Nat Genet* **41**, 1308-12 (2009).
4. Seto-Salvia, N. *et al.* Dementia risk in Parkinson disease: disentangling the role of MAPT haplotypes. *Arch Neurol* **68**, 359-64 (2011).
5. Tobin, J.E. *et al.* Haplotypes and gene expression implicate the MAPT region for Parkinson disease: the GenePD Study. *Neurology* **71**, 28-34 (2008).
6. Pittman, A.M. *et al.* The structure of the tau haplotype in controls and in progressive supranuclear palsy. *Hum Mol Genet* **13**, 1267-74 (2004).
7. 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**, 561-70 (2007).
8. Myers, A.J. *et al.* A survey of genetic human cortical gene expression. *Nat Genet* **39**, 1494-9 (2007).
9. Laws, S.M. *et al.* Association of the tau haplotype H2 with age at onset and functional alterations of glucose utilization in frontotemporal dementia. *Am J Psychiatry* **164**, 1577-84 (2007).
10. Sudmant, P.H. *et al.* Diversity of human copy number variation and multicopy genes. *Science* **330**, 641-6 (2010).