European Journal of Human Genetics (2008) 16, 252–260

www.nature.com/ejhg

ARTICLE

Haplotype patterns in cancer-related genes with long-range linkage disequilibrium: no evidence of association with breast cancer or positive selection

Gloria Ribas*[,1], Roger L Milne[2], Anna Gonzalez-Neira[2] and Javier Benítez[1,2]

[1]*Human Genetics Group, Human Cancer Genetics Programme, Spanish National Cancer Centre (CNIO), Madrid, Spain;* [2]*National Genotyping Centre (CeGen), Human Cancer Genetics Programme, Spanish National Cancer Centre (CNIO), Madrid, Spain*

The average length of linkage disequilibrium (LD) blocks in European populations is about 22 kb. In this study, we have selected 20 genes with LD blocks larger than 60 kb (with a median length of 88 kb) from a total of 121 cancer-related genes. We observed limited haplotype diversity, with an average of three haplotypes per gene accounting for more than 90% of the diversity, two of these being a Yin–Yang pair in 95% of the LD blocks. The mean frequency of the most common haplotype in the Spanish population was just below 50%, similar to those for the HapMap CEU and African samples, but lower than the 60% observed in Asian samples. Genes involved in the regulation of nucleobases and nucleic acid metabolism were overrepresented among these 20 genes with long LD blocks (eight genes *ATM*, *BRCA1*, *BRCA2*, *ERCC6*, *MLH1*, *MSH3*, *RAD54B* and *XRCC4*) relative to the other 101 cancer-related genes studied ($P = 1.23 \times 10^{-6}$). The ancestral haplotype was observed at a frequency greater than 3 in 67% of the genes either in the Spanish or one of the HapMap sampled populations. When observed, the ancestral haplotype had an average 15% frequency in the Spanish sample, less than half that observed in Asian and African samples. The Spanish Yin–Yang haplotype pair represented over 35% of haplotypes in African samples and over 65% in non-African samples. We detected differences in SNP frequencies between populations for five genes (*ALDH2*, *APC*, *PIK3CB*, *RB1* and *XRCC4*, all with Fst > 0.4); however, these genes did not show evidence of positive selection. Finally, we found no evidence that the haplotypes formed by SNPs in the 20 genes are associated with breast cancer.

European Journal of Human Genetics (2008) **16,** 252–260; doi:10.1038/sj.ejhg.5201953; published online 14 November 2007

Keywords: cancer-related genes; haplotypes; ancestral haplotype; positive selection; association study

Introduction

The data generated by the HapMap project have determined the common patterns of DNA sequence variation in the human genome from populations across four geographical regions.[1,2] This information is providing an unprecedented view of human genetic diversity that is used primarily in association studies but will give insights into many other areas of research such as studies of linkage disequilibrium, haplotype block distributions, the localisation of recombination hotspots, effects of natural selection and how these have shaped human genetic variation. On top of that, the scientific community now has access to a draft of the chimpanzee genome (*Pan troglodyte*), which was recently released.[3] At nearly all SNP locations in human genes, chimps have a nucleotide identical to one of

*Correspondence: Dr G Ribas, Human Cancer Genetics Programme, Spanish National Cancer Centre (CNIO), Melchor Fernández Almagro, 3, Madrid E-28029, Spain.
Tel: +34 91 224 6950; Fax: +34 91 224 6923;
E-mail: gribas@cnio.es
Received 19 May 2007; revised 25 September 2007; accepted 10 October 2007; published online 14 November 2007

the human nucleotides at nearly all SNP (single-nucleotide polymorphism) locations in human genes which means that our common ancestor almost certainly had the same nucleotide. The search for ancestral and derived nucleotides has recently been the object of attention in the scientific community and may uncover 'footprints' of positive selection that have occurred recently in humans and may explain different susceptibilities to disease. One example of this is the work of Puente *et al*,[4] who have suggested that small differences in cancer genes might influence the difference in cancer susceptibility between the two species.

Although some reviews have reported linkage disequilibrium (LD) extending over distances greater than 100 kb[5–8] the average length of LD blocks in European populations is about 22 kb, although at least 50% of the European human genome exists in blocks of around 44 kb.[9] Besides, it has been suggested that some of these regions of extended LD may play an important role in determining the genetic bases of human phenotypic differences.[10] Regions of LD are characterised by strong association between alleles, low haplotype diversity and low recombination rates.[11] In addition, some of the larger LD blocks have recently been associated with positive selection through human evolution.[8] Several authors have described that regions with limited haplotype diversity have at least one pair of high-frequency haplotypes composed of completely mismatching SNP alleles, also referred to as a Yin–Yang pair, and these pairs are suspected to be of a very ancient origin.[12,13]

We have recently reported that only 12% of a set of cancer-related genes contained at least one LD block larger than 60 kb.[14] In this present study, we aimed to further test whether such genes with longer LD blocks in the Spanish population were subject to some sort of selection and make some contribution to disease aetiology. We first examined whether 20 cancer-related genes with LD blocks larger than 60 kb fell into any particular category of function. Second, we studied the haplotype block structure in each of the genes, including the frequency distribution, the presence of Yin–Yang pairs, and whether the ancestral haplotype was present in Spanish controls and then compared all these factors across the four sampled HapMap populations (CEU, YRI, JPT and CHB). Third, we looked for positive selection and finally, we study whether these genes were associated with breast cancer by comparing their haplotype frequency distributions among Spanish breast cancer cases and controls.

Materials and methods

Study population

The recruitment of cases and controls has been previously described.[15] Briefly, cases were 864 women with breast cancer and mean age at diagnosis of 50 years (range: 23–86 years) recruited between 2000 and 2004. Of these, 574 were consecutively recruited via three public hospitals in Spain: Hospital La Paz, the Fundación Jiménez Díaz, Hospital Monte Naranco, while 290 were cases attending our family cancer clinic for genetic testing who had at least one affected first-degree relative. Controls were 845 Spanish women free of breast cancer at ages ranging from 23 to 86 years (mean = 53 years), recruited between 2000 and 2005 via the following sources: the Menopause Research Centre at the Instituto Palacios, the College of Lawyers; the National Blood Transfusion Centre, the Catalan Institute of Oncology (ICO); and from the Centre for the Investigation of Cancer (CIC). Informed consent was obtained from all participants, and the study was approved by the Institutional Review Board of Hospital La Paz, Madrid.

Candidate gene choice, SNP selection and haplotype analysis

The 121 genes and SNPs were selected according to previously published criteria:[14,15] genes previously reported to be associated with or known to be involved in cancer; genes involved in cell cycle pathways; DNA repair; cell communication; hormone metabolism; apoptosis; carcinogen metabolism; cell adhesion; cell proliferation and differentiation; nucleoside, nucleotide and nucleic acid metabolism; oncogenesis; developmental processes; and/or signal transduction. The main criterion for SNP selection was marker density as a function of LD with priority given to tag-SNPs defining common haplotypes.[15] The 20 genes with LD blocks larger than 60 kb and their corresponding SNPs studied are detailed in Supplementary Table 1. The final average SNP density was one SNP for every 8.7 kb.

Genotyping

Genomic DNA from subjects was isolated from peripheral blood lymphocytes using automatic DNA extraction (Magnapure; Roche, Mannhein, Germany) according to the manufacturer's recommended protocols. This DNA was quantified using picogreen and diluted to a final concentration of 50 ng/μl for genotyping.

Genotyping of SNPs was carried out using the Illumina Bead Array System (Illumina Inc., San Diego, CA, USA) according to the manufacturer's protocols.[16] At least one duplicate and one negative control were included per 96-well plate, and six samples were duplicated across plates. The total number of duplicates across all plates was 35 (15 cases, 17 controls and a nonstudy child–parents' triad).

Assignment of ancestral alleles

We obtained FASTA sequences surrounding each SNP from the dbSNP database (build 35 of the human genome) and aligned those to the draft build of the chimp genome sequence, (http://genome.ucsc.edu/cgi-bin/hgBlat). For each SNP, we selected the best overall alignment, preferring

alignments mapping to a unique chimp chromosome. We then inferred the ancestral state as the chimp allele at the corresponding position in the sequence, provided that the sequence quality score was greater than 20 at that site, and that it matched one of the human alleles.

Block definition and haplotype distribution

The LD blocks within genes were determined among controls using an R^2 threshold of 0.8. among Spanish controls (Haploview v3.1.1).[17] The LD structure of the 20 genes with LD blocks larger than 60 kb is shown in Supplementary Figure 1. Haplotypes (the combinations of variants along chromosomes) were inferred using PHASE 2.1. Haplotype blocks determined in the Spanish controls were applied to all four HapMap samples (CEU, YRI, JPT and CHB), and further analyses were restricted to SNPs in these blocks. The LD structure of each gene is shown in Supplementary Figure 1, and a full list of these 20 genes and selected SNPs is provided in Supplementary Table 1.

We identified Yin–Yang haplotype pairs within LD blocks according to the following criteria: at least five SNPs, each with a minor allele frequency (MAF) of at least 10%, or less SNPs meeting this frequency criterion but spanning more than 22 kb; and the least frequent of the Yin–Yang haplotype pair having a frequency greater than 3%.[13] The ancestral haplotype was inferred for each LD block by combining the ancestral allele in each SNP per block considered.

Statistical analysis (haplotype association study)

Deviations from Hardy–Weinberg equilibrium were tested using the *genhwi* command in STATAv8.0.[18] Differences in the haplotype distributions between cases and controls were tested using the χ^2-test. PHASEv2.1 software[19,20] was used to impute haplotypes and compare their frequency distributions in cases and controls. Odds ratios (ORs), their 95% confidence intervals (CIs) and Wald's statistic *P*-values were estimated, via unconditional logistic regression (STATAv8.0), for haplotypes with frequency greater than 0.01, using the most frequent haplotype among controls as reference and assuming, for each subject, that the most likely imputed haplotypes were observed. Analysis of haplotypes was repeated using the *haplo.stats* library implemented in R, which compares haplotype frequencies in cases and controls in an unbiased way by including haplotype uncertainty in the estimation of ORs.

HapMap project data

We used Phase I data from the HapMap project, which comprises samples of Utah residents with ancestry from northern and western Europe (CEU); Han Chinese in Beijing, China (CHB); Japanese in Tokyo, Japan (JPT); and Yoruba in Ibadan, Nigeria (YRI). In some analysis, the CHB and JPT samples have been pooled and are referred to as the ASN sample. Haplotype blocks determined in the Spanish controls were applied to all four HapMap samples (CEU, YRI, JPT and CHB). Haplotype phase estimation for all the data was performed by the HapMap consortium using Phase 2.0. The phasing procedure also imputed all missing genotypes at SNPs with less than 20% missing data.

Gene ontology analysis

Genes were classified into Gene ontology (GO) categories[21] using DAVID.[22] Differences in frequency of GO categories among the 20 genes with long LD blocks compared to other 101 cancer-related genes were tested for using Fisher's exact test.

Recent positive selection

We assessed evidence of recent positive selection in the 20 candidate genes using the online browser Haplotter[23] (http://pritch.bsd.uchicago.edu/data.html). The web page displays the results for positive selection in genes or genomic regions of the human genome using the HapMap data. This program provides plots of two parameters: iHS, the integrated haplotype score, which measures positive selection on the ancestral and derived alleles via the decay of extended haplotype homozygosity[24] and Fst, a measure of the degree of population differentiation based on pairwise SNP frequency comparisons. Haplotter also identifies iHS and Fst scores considered to be statistically significant.

Results

A total of 191 SNPs were successfully genotyped, and of those, a final 159 SNPs (83%) with a MAF of at least 10% were included in the haplotype analysis of the 20 genes with LD blocks larger than 60 kb. We observed a total of 21 blocks complying with these criteria, two in *CDK6* and one in each of the other genes. These 20 genes are located across 12 different chromosomes (2, 3, 4, 5, 6, 7, 8, 10, 11, 12, 13 and 17), have an average size of 101 kb (range: 61–199 kb) and together span 2.2 Mb of the genome (Table 1). The average density of SNPs genotyped in these genes with an allele frequency higher than 10% was one SNP for every 9.7 kb. The LD structure of each gene is shown in Supplementary Figure 1, and a full list of these 20 genes and selected SNPs is provided in Supplementary Table 1.

To understand more about these 20 genes with large LD blocks, we looked into their Gene Ontology (GO) classifications,[21,25] and observed that they cover a broad range of biological processes. Nevertheless, the most overrepresented category was DNA repair with eight (*ATM*, *BRCA1*, *BRCA2*, *ERCC6*, *MLH1*, *MSH3*, *RAD54B* and *XRCC4*) of the 20 genes with long LD blocks classified as being involved in 'regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism' *vs* none of the 101 other genes. This difference was highly statistically significant (unadjusted $P = 1.2 \times 10^{-6}$), even after the most conservative

Table 1 Summary of the haplotype data in the 20 candidate genes

Gene	*Total no. of SNPs genotyped*	*Total no. of SNPs studied**	*SNPs/block*	*Block size (bp)*	*Haplotype frequencies > 10%*	*% of haplotypes*	*Total % of haplotypes*
ALDH2	4	4	4	61 066	3	68/16/10	94.44
APAF1	4	4	4	73 712	5	33/22/17/15/11	98.12
APC	7	5	5	66 546	2	52/32	84.10
ATM	11	10	10	144 408	4	38/36/15/10	98.06
BRAF	5	5	5	191 899	2	72/16	88.41
BRCA1	6	5	5	80 842	3	42/34/22	97.53
BRCA2	7	7	5	61 124	4	22/21/20/20	83.80
CDK6_1	18	11	6	95 546	3	37/23/16	75.41
CDK6_2			5	63 760	2	64/16	80.42
EGF	11	9	8	75 906	2	58/30	88.00
ERCC6	12	11	11	74 947	3	43/22/15	80.00
MAP2K4	10	8	8	89 507	4	38/23/22/12	95.23
MAPK14	8	7	7	82 035	3	58/26/12	96.06
MLH1	5	5	5	64 407	3	49/26/21	95.90
MSH3	18	16	13	199 200	6	27/16/12/11/11/10	85.91
NFKB1	11	10	9	107 580	3	47/22/15	84.79
PIK3CB	9	8	8	115 631	2	50/43	93.76
RAD54B	12	12	12	115 167	2	55/28	82.77
RB1	13	7	7	170 142	3	57/21/13	91.45
SOS1	3	3	3	87 421	2	58/30	88.38
XRCC4	16	13	7	93 710	2	45/36/11	92.19
Mean	9.50	10.40	7.00	100 693	3.0		89.3

Note: *Excluding SNPs with minor allele frequency <10%, as well those with genotyping errors or evidence of departure from Hardy–Weinberg equilibrium.
SNPs/block: number of polymorphisms analysed per LD block. % of haplotypes: represents the frequencies of each of the haplotypes with frequency greater than 10%. Total % of haplotypes: the percentage of the haplotypes represented by those with frequency greater than 10%.
Total length analysed 2 192 605 bp.

corrections for multiple testing ($P = 2.8 \times 10^{-4}$, assuming all 228 categories observed are independent).

Haplotype structure

We observed a very high correlation in haplotype frequencies between the Spanish control sample ($N = 845$), and HapMap CEU sample ($R^2 = 0.96$, Supplementary Figure 2) as previously reported[14] and further comparisons were therefore not made between these two samples. The correlation was much lower when the Spanish sample was compared to that from the other two HapMap populations, $R^2 = 0.38$ for JPT, 0.32 for CHB and 0.038 for YRI (Supplementary Figure 2).

An average of three haplotypes per block represented 89% of all haplotypes with frequencies greater than or equal to 10% (Table 1). We compared the individual haplotypes observed in the Spanish, European – CEU, Asian and Yoruban samples considering only those observed in at least one sample with a frequency greater than 5% (Figure 1).

The ancestral allele for each of the SNPs in this study is shown in Supplementary Table 1. We identified 14 ancestral haplotypes among the 21 LD block fragments (67%) present in either the Spanish sample or the HapMap samples. The average frequency of this ancestral haplotype, when observed, was 15% (SD = 6.8%) among Spanish controls, and varied in the three non-CEU HapMap samples, being 36% (SD = 25%) in YRI, 36% (SD = 25%) in CHB and 32% (SD = 26%) in JPT. This information is summarised in Table 2a and detailed information highlighted in Supplementary Table 2.

The average frequency of the most common haplotype in each block in the Spanish control sample was 48% (SD = 13%) and that for the same haplotype in YRI, JPT and CHB was 48% (SD = 19%), 62% (SD = 19%) and 60% (SD = 18%), respectively. Results for the Spanish and CEU samples were so similar (data not shown) that only the former are reported here. The most common Spanish haplotype was also the most common in only 19% (4/21) of YRI blocks. In the case of the JPT and CHB samples, the most common haplotype coincided with the European counterpart 62% (13/21) and 67% (14/21) of the time, respectively (Table 2b and Supplementary Table 2).

The Yin–Yang haplotype pair was observed in the Spanish sample in 20 of the 21 blocks (90%). We did not observe the Yin–Yang pair in the *APAF1* gene. We included these 20 blocks in further analyses of Yin–Yang haplotypes (Table 3a and b). The Yin haplotype was generally the most common haplotype among the Spanish controls (16 (80%) of the 20 blocks) with an average frequency of 45% (SD = 15%). The mean Yin haplotype frequency was 13% (SD = 12%) in the YRI sample and 39% (SD = 28%) and 37% (SD = 28%) in the JPT and CHB samples, respectively.

The Yang haplotype was the second most common in Spaniards for 12 (60%) of the 20 blocks, and the third in frequency for 7 (35%). The Yin–Yang pair was made up by the second and fourth haplotype in frequency for one gene (*BRCA2*). The average frequency of the Yang haplotype in Spaniards was 23% (SD = 9.0%) very similar to that in the other sampled populations, 14% (SD = 16%) in YRI, and 19% (SD = 21%) and 20% (SD = 21%) in JPT and CHB, respectively. However, no Yin–Yang pairs were observed in blocks or fragments therein of nine (45%) genes among YRI; five (25%) among JPT and four (20%) among CHB (Table 3a and b).

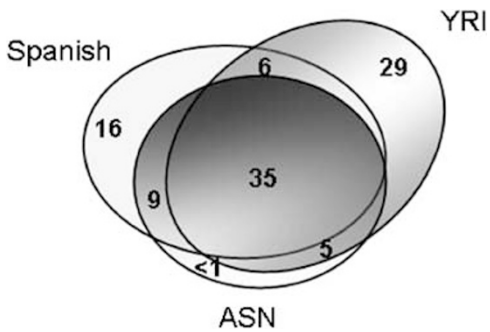

Figure 1 Distribution of haplotypes with frequency >5% across populations.

Where Spanish Yin–Yang haplotypes were present, they accounted for an average of 68% (SD = 19%) of all haplotypes in the three non-African samples. These same two haplotypes comprised an average of 57% (SD = 31%) and 57% (SD = 30%) in JPT and CHB, respectively. They accounted for only 27% of all YRI haplotypes, on average (SD = 17%). For only three genes (*ALDH2*, *BRAF* and *SOS1*), the ancestral haplotype was identical to one of the Spanish Yin–Yang haplotype pair, in all cases, it was the Yang haplotype.

Recent positive selection

We did not observe significant evidence of positive selection (|iHS| > 2.0. across a substantial portion of the gene) for any of the 20 genes with long LD blocks. However, some genes had small areas with elevated iHS scores: *XRCC4* (|iHS| values from 1.9 up to 2.4 in YRI), *APC* (|iHS| values from 1.7 to 2.1 in ASN), *CDK6* (|iHS| values from 1.5 up to 2.0 in YRI), *MSH3* (|iHS| values from 1.0 to 1.9 in CEU), *RAD54B* (|iHS| values from 1.1 up to 1.7 in YRI) and *RB1* (|iHS| values from 1.1 up to 1.75 in YRI).

The average Fst level of autosomal SNPs is 0.15,[26] which is within the range of previously published Fst estimates (0.05–0.15) for neutral (nonselected) genes and SNPs.[26,27] We found evidence of differences between population samples for *ALDH2* (Fst = 0.6 for CEU *vs* ASN), *PICK3CB* (Fst = 0.6 for CEU *vs* ASN), *RB1* (Fst = 0.5 for both CEU *vs*

Table 2 Summary information for each LD block regarding the ancestral haplotype (based on the chimpanzee sequence) (a) and the most common haplotype in Spaniards and its frequency in the other populations of HapMap (b).

		a					*b*			
Gene	*Ancestral*	*Freq. Spanish*	*Freq. YRI_HM*	*Freq. JPT_HPM*	*Freq. CHB_HM*	*Most common haplotype*	*Freq. Spanish*	*Freq. YRI_HM*	*Freq. JPT_HPM*	*Freq. CHB_HM*
ALDH2	H2	16.43	38.70	79.60	88.70	H Yi	67.58	52.80	79.60	79.60
APAF1	H4	14.94	21.00	30.20	42.20	H1	32.67	36.30	46.50	44.30
APC	H6	—	9.60	4.80	7.80	H Yi	51.90	31.10	72.60	66.70
ATM	—	—	—	—	—	H Yi	38.12	39.20	57.00	60.00
BRAF	H2	16.20	33.30	7.00	22.00	H Yi	72.22	35.00	91.90	76.70
BRCA1	—	—	—	—	—	H1	41.56	90.80	72.10	68.90
BRCA2	H3	20.45	37.50	5.80	23.00	H1	22.14	37.50	47.70	33.70
CDK6_1	H3	15.66	66.40	71.50	61.10	H Yi	36.80	73.50	78.90	61.00
CDK6_2	H3	6.02	4.20	—	—	H1	64.40	62.50	91.90	86.80
EGF	—	—	5.10	27.10	24.40	H1	58.15	37.50	27.10	33.50
ERCC6	H5	8.31	38.30	18.60	18.90	H Yi	42.86	38.30	44.20	40.00
MAP2K4	—					H1	38.20	41.40	62.70	60.40
MAPK14	H2	26.38	85.80	72.10	76.10	H Yi	57.61	85.80	72.10	76.10
MLH1	H2	25.96	41.30	28.30	43.20	H Yi	49.34	41.30	50.70	51.20
MSH3	—	—	—	—	—	H1	26.66	16.70	46.50	30.70
NFKB1	—	—	5.40	—	—	H Yi	47.33	40.00	50.00	45.60
PIK3CB	—	—	—	—	—	H Yi	50.43	39.10	96.60	97.30
RAD54B	—	—	—	—	—	H Yi	54.92	43.90	34.90	43.30
RB1	H4	6.56	74.20	12.80	10.00	H Yi	57.47	74.20	47.70	57.80
SOS1	H3	9.23	47.70	23.10	15.60	H Yi	58.00	47.70	62.50	67.80
XRCC4	—	—	—	—	—	H Yi	44.56	35.00	69.70	72.20
Mean		15.10	36.32	31.74	36.08	Mean	48.23	47.60	62.04	59.70

Note: H1%: the most common haplotype, H2%: the second most frequent haplotypes and H3, H4, H5 and H6%: the third, fourth, fifth and sixth most frequent haplotypes, respectively. H Yi: Yin haplotype, H Ya: Yang haplotype, YRI_HM%: haplotype frequency among the Yoruba sample, JPT_HM%: haplotype frequency among the Japanese sample and CHB_HM%: haplotype frequency among the Chinese sample.

Table 3 Summary information regarding Yin–Yang haplotypes in the different populations sampled: Spanish and YRI (Yoruba), JPT (Japanese) and CHB (Chinese) from HapMap. (a) Yin pair (b) Yang pair

	a					*b*				
Gene	*H Yi*	*Freq. Spanish*	*Freq. YRI_HM*	*Freq. JPT_HPM*	*Freq. CHB_HM*	*H Ya*	*Freq. Spanish*	*Freq. YRI_HM*	*Freq. JPT_HPM*	*Freq. CHB_HM*
ALDH2	H1	67.58	—	3.50	1.10	H2	16.43	38.70	79.60	88.70
APAF1	—	—	—	—	—	—	—	—	—	—
APC	H1	51.90	5.80	72.60	66.70	H2	32.20	31.10	7.90	10.00
ATM	H1	38.12	30.00	57.00	60.00	H2	35.75	1.70	37.20	31.00
BRAF	H1	72.22	7.50	91.90	76.70	H2	16.20	33.30	7.00	22.00
BRCA1	H2	33.88	9.20	27.90	31.30	H3	22.09	—	—	—
BRCA2	H2	20.90	16.70	47.70	33.70	H4	20.33	—	17.40	24.40
CDK6_1	H1	36.80	—	2.30	5.60	H2	22.95	1.00	—	2.20
CDK6_2	H2	16.02	15.00	—	—	H3	8.10	24.90	4.70	11.10
EGF	H1	58.15	13.30	20.70	33.50	H2	29.85	5.80	12.80	5.40
ERCC6	H1	42.86	—	30.20	31.30	H2	21.98	9.20	1.20	—
MAP2K4	H2	22.74	41.40	14.00	12.50	H3	22.26	3.30	10.50	15.60
MAPK14	H1	57.67	9.20	26.10	22.20	H3	12.01	5.00	—	—
MLH1	H1	49.34	27.40	20.70	5.60	H3	20.60	—	50.70	51.20
MSH3	H1	26.66	16.70	14.00	13.30	H2	15.65	3.30	13.90	13.30
NFKB1	H1	47.33	16.80	50.00	45.60	H2	22.26	—	38.40	37.60
PIK3CB	H1	50.43	19.20	96.60	97.80	H3	43.33	39.20	2.30	2.20
RAD54B	H1	54.92	—	20.90	14.40	H2	27.85	—	34.90	41.10
RB1	H1	57.47	—	47.70	57.80	H2	20.72	13.30	30.20	23.30
SOS1	H1	58.00	3.30	62.75	67.80	H3	9.23	47.70	23.10	15.60
XRCC4	H1	44.56	35.00	69.70	72.20	H2	36.39	12.50	—	5.60
Mean		45.38	17.77	40.86	39.43	Mean	22.81	18.00	23.24	23.55

H Yi: Yin haplotype. H Ya: Yang haplotype.

Table 4 Summary of the most significant findings from the case–control association study with haplotypes in 20 studied genes

Gene	*No. of SNPs*	*No. of Hap >1%*	*Hap REF/Hap SIG*	*Freq (%) cases*	*Freq (%) controls*	P-*value*
APAF1	4	5	TGCG/CGTA	13.7	17.3	0.007
CDK6_1	6	7	ACCTTG/ACGCGA	6.1	4.4	0.0457
EGF	8	4	AGGATAGC/GAAGCTGC	3.4	2	0.0154
ERCC6	11	6	GAATTGGTGGG/CAACTGGTGGG	3.3	2	0.0171
MAPK14	7	4	TCACTAG/CTGTCGA[a]	9.8	12	0.0385
MSH3	13	8	CCCCTGGGAGATG/TCCCCAAGGCATG	9.4	11.8	0.0122
SOS1	3	4	CCC/TCC	3.2	1.7	0.0249

[a]Yin–Yang pair. Hap REF, reference haplotype; Hap SIG, potentially associated haplotype; Freq, haplotype frequency.

YRI and for YRI *vs* ASN,), *XRCC4* (Fst = 0.4 for CEU *vs* ASN) and *APC* (Fst = 0.4 for YRI *vs* ASN). Fst and iHS data for all 20 extended LD genes is provided in Supplementary Figure 3.

Association study

Comparison of haplotype frequencies in the 864 breast cancer patients and 845 healthy controls gave some evidence of association with breast cancer (unadjusted *P*-value <0.05) for haplotypes in seven genes (*APAF1*, *CDK6_1*, *EGF*, *ERCC6*, *MAPK14*, *MSH3_1* and *SOS1*) (Table 4). However, none of these associations would be statistically significant after consideration of the multiple tests performed. All except APAF1 (for which Yin–Yang haplotypes were not observed) had the Yin haplotype as the most common (and therefore the reference) haplotype. The putative associated haplotype was the Yang haplotype for only one of these (*MAPK14*). The ancestral haplotype was neither the most common haplotype, nor that putatively associated with breast cancer risk for any of these seven genes. A full list of haplotypes per gene and their frequencies in Spanish cases and controls is given in Supplementary Table 2.

Discussion

The aim of our study was to test whether cancer-related genes with long LD block structure are subject to some sort of selection and could contribute to breast cancer aetiology. To address this, we selected and analysed the 20 genes with

LD blocks larger than 60 kb among 121 cancer-related genes.

Basic functions such as 'regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism' related to DNA repair were enriched in these 20 genes compared to the other 101 cancer-related genes ($P = 2.8 \times 10^{-4}$). We believe that these functions in the cell had to be carried out by highly 'regulated' and 'controlled' proteins, and that selection would therefore act against variation in the genes that encode them. Such genes would be presumably under purifying selection and not under the type of positive selection that can be detected by this method. The few haplotype combinations that are present should be fully functional. This would explain the low recombination rates observed in these genes.

The results of our comparison of haplotype distributions across populations are generally in concordance with those obtained by Gabriel.[9] The highest percentage of population-specific haplotypes was present in African samples (HapMap_YRI) which is in agreement with Africans being the most diverse population sampled. Moreover, a great similarity in haplotypes and their frequencies was observed between the Spanish and Asian populations; however, both populations had a greater proportion of unique population-specific haplotypes than observed in Gabriel.[9]

We found that an average of three haplotypes per gene represented over 90% of the total haplotype distribution. That is, the majority of these haplotypes have high frequencies. The most common haplotype in each block had a frequency of over 50% in all the sampled populations. Our results are consistent with those of other groups in terms of finding that genes with long LD blocks have reduced diversity of haplotypes.[28]

Since the sequence of the chimpanzee genome was published in 2005, genetic comparisons between chimp and human have become widely possible.[3] We were able to determine the ancestral allele for all the SNPs in the 20 genes included in this study, which is not surprising considering that human and chimp genomes are 99% identical.[4] The ancestral haplotype was inferred for 21 long (>60 kb) LD blocks in the 20 genes. However, it was not observed in about a quarter of the LD blocks. It was most often observed, and with higher frequency, in YRI, next most often in Asians (JPT and CHB) and least often in Spaniards. This finding is consistent with Africans being the most genetically diverse population. It also corroborates the out-of-Africa hypothesis of human populations[29] as well as human demographic history in which the ancestral African population has maintained a larger effective population size and has had more time for recombination and mutations to reduce LD. On the contrary, the HapMap CEU and Spanish samples had the highest number of derived haplotypes.

We found that a Yin–Yang haplotype pair was present in more genes in the non-African samples than in the YRI sample, being highest in the European-Spanish population. The Yin–Yang pairs constituted a substantial fraction of the total haplotype diversity. The average combined frequency of the Yin–Yang pair in the four populations sampled constituted an overall 62% (SD = 25.08%) of the haplotype diversity seen in these blocks. This percentage is almost double that reported by Zhang.[13] That is, it appears that in general, Yin–Yang haplotypes are more prevalent, and haplotypic diversity is lower in genes with large LD blocks compared with nonselected genomic regions although this is less the case in older (African) populations. One possible explanation for this is that regions of high LD were naturally enriched with this phenomenon and follow a neutral evolutionary model, suggesting that Yin–Yang haplotypes are genetic signatures that emerged prior to the African diaspora.[13] Another possible explanation is that they represent a selection bias, thus, when selecting candidate SNPs with high pairwise R^2 (>0.8), the Yin–Yang pairs naturally appear more often in the specific population from which they have been selected.

To evaluate the sensibility to detect iHS and Fst with HapMap data, we obtained these values for the lactase (*LCT*) and *SCA2* (*ATXN2*) genes, both in regions with high LD and both known to be positively selected, but not related to cancer.[8,30] Using Haplotter,[23] we observed significant values of iHS over a large portion of the lactase gene (iHS > 3 for about 1.5 MB) and elevated Fst for SCA2 (*ATXN2*) (Fst > 0.6 for about 1.5 MB).

When the genes of this study were screened using the same program, we detected differences in SNPs frequency across populations (Fst > 0.4) for four genes (*APC*, *CDK6*, *RB1* and *XRCC4*. It was also detected for *ALDH2*, but this gene is located in the same genomic region as SCA2 (*ATXN2*).[8] However, for each gene, the elevated Fst was observed as a single peak that did not extend across the genomic region. Such long regions of high LD could have been subjected to evolutionary forces such as selection in humans. However, in the present study, we found no clear evidence of positive selection having acted on our 'high-LD' candidate cancer genes using the method of Sabeti *et al*.[24]

Voight *et al*[23] used the same method to identify positive selection acting on genes involved in chemosensory perception, olfaction and fertilisation. However, these functions are very different to those of our 20 genes, which are involved in more processes such as basic cellular signal transduction, DNA repair and cell cycle. It may be that these latter functions are too basic or that the role of these genes may tend to act later in life (ie, after reproduction), so that any positive selection does not act upon them.

We hypothesised that the genes with long LD blocks, and Yin–Yang and ancestral haplotypes contained therein in particular, might be more likely to be involved in breast cancer predisposition. An overrepresentation of mutations

in BRCA1 has been observed in the Yang haplotype relative to the most common Ying haplotype.[31] In addition, a protective effect against breast cancer risk has been shown for the minor allele at an SNP, which occurs on the Yang haplotype in *ERCC4*.[15] Furthermore, a putative role of the ancestral allele in six cancer-susceptibility SNPs has been suggested based on a review of selected association studies.[4] For one of the SNPs, R72P in *TP53*, the most common human allele, Arg72, is the derived allele whose frequency ranges from 55 to 92% among different human populations, and the ancestral allele is Pro72, both alleles have been associated with cancer risk in different studies.[32–37] In our study, we found no evidence of association with breast cancer risk for haplotypes in any of the 20 genes studied after adjustment for multiple testing. For just one of the seven blocks with haplotypes that had unadjusted *P*-values less than 0.05 (*MAPK14*), the reference (most common) and putative risk-associated haplotypes constituted the Yin and Yang, respectively in the Yin–Yang pair. The ancestral haplotype was neither the reference nor the best candidate associated haplotype for any of these blocks. In summary, we found no evidence that Yin–Yang haplotypes nor ancestral haplotypes are more likely to be associated with breast cancer risk. This may be due to our study lacking power to detect association.

In conclusion, we detected a reduced haplotype diversity in genes with elevated LD over a long distance (>60 kb), with an average of three haplotypes per gene accounting for >90% of the diversity, two of those being the pair Yin–Yang in most of the cases. Moreover, the most common haplotype (most of the time the Yin) had an average frequency of around 50%. In addition, we observed the ancestral haplotype in the Spanish, JPT, CHB and YRI populations for 65% of the genes at a mean frequency of 20% in the Spanish and about 40% in the other HapMap populations. Finally, we found no evidence that positive selection has acted on these 20 genes nor that haplotypes formed by SNPs in them are associated with breast cancer.

Acknowledgements

GR conceived the study, participated in its design, in the genotyping and analysis of the data, organised the coordination and drafted the manuscript. RLM participated in the design of the study, performed the statistical analysis and drafted the manuscript. AGN performed the genotyping experiments, participated in its design and helped to draft the manuscript. JB participated in design of study, coordination and helped to draft the manuscript. All authors read and approved the final manuscript. We thank JI Arias (Hospital Monte Naranco), P Zamora (Hospital la Paz), A Ruibal (Fundación Jiménez Díaz), S Palacios (Instituto Palacios), S de Sanjose (ICO) and R González (CIC) for the use of Spanish samples of cases and controls. Emilio Gonzalez and Rosario Alonso for their technical support with the Illumina Platform, and Fatima Mercadillo, Victoria Fernandez, Alicia Barroso and Rocio Letón for their technical assistance. This study was partially funded by the Genome Spain Foundation and BFI2003-03852.

References

1 The International HapMap Consortium: The International HapMap project. *Nature* 2003; **426**: 789–796.
2 The International HapMap Consortium: A haplotype map of the human genome. *Nature* 2005; **437**: 1299–1320.
3 Chimpanzee Sequencing and Analysis Consortium: Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 2005; **437**: 69–87.
4 Puente XS, Velasco G, Gutierrez-Fernandez A, Bertranpetit J, King MC, Lopez-Otin C: Comparative analysis of cancer genes in the human and chimpanzee genomes. *BMC Genomics* 2006; **7**: 15.
5 Collins A, Lonjou C, Morton NE: Genetic epidemiology of single-nucleotide polymorphisms. *Proc Natl Acad Sci USA* 1999; **96**: 15173–15177.
6 Huttley GA, Smith MW, Carrington M, O'Brien SJ: A scan for linkage disequilibrium across the human genome. *Genetics* 1999; **152**: 1711–1722.
7 Jorde LB: Linkage disequilibrium and the search for complex disease genes. *Genome Res* 2000; **10**: 1435–1444.
8 Yu F, Sabeti PC, Hardenbol P *et al*.: Positive selection of a pre-expansion CAG repeat of the human SCA2 gene. *PLoS Genet* 2005; **1**: e41.
9 Gabriel SB, Schaffner SF, Nguyen H *et al*: The structure of haplotype blocks in the human genome. *Science* 2002; **296**: 2225–2229.
10 Hinds DA, Stuve LL, Nilsen GB *et al*: Whole-genome patterns of common DNA variation in three human populations. *Science* 2005; **307**: 1072–1079.
11 Goldstein DB: Islands of linkage disequilibrium. *Nat Genet* 2001; **29**: 109–111.
12 Clark AG, Weiss KM, Nickerson DA *et al*: Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. *Am J Hum Genet* 1998; **63**: 595–612.
13 Zhang J, Rowe WL, Clark AG, Buetow KH: Genomewide distribution of high-frequency, completely mismatching SNP haplotype pairs observed to be common across human populations. *Am J Hum Genet* 2003; **73**: 1073–1081.
14 Ribas G, Gonzalez-Neira A, Salas A *et al*: Evaluating HapMap SNP data transferability in a large-scale genotyping project involving 175 cancer-associated genes. *Hum Genet* 2006; **118**: 669–679.
15 Milne RL, Ribas G, Gonzalez-Neira A *et al*: ERCC4 associated with breast cancer risk: a two-stage case-control study using high throughput genotyping. *Cancer Res* 2006; **66**: 9420–9427.
16 Oliphant A, Barker DL, Stuelpnagel JR, Chee MS: BeadArray technology: enabling an accurate, cost-effective approach to high-throughput genotyping. *Biotechniques* 2002; (Suppl): 56–58, 60–61.
17 Barrett JC, Fry B, Maller J, Daly MJ: Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 2005; **21**: 263–265.
18 StataCorp: *Stata Statistical Software: Release 8.0TX.*. College Station, TX: Stata Corporation, 2003.
19 Stephens M, Smith NJ, Donnelly P: A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* 2001; **68**: 978–989.
20 Stephens M, Donnelly P: A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet* 2003; **73**: 1162–1169.
21 The Gene Ontology: Gene ontology: tool for the unification of biology. *Nat Genet* 2000; **426**: 789–796.
22 Dennis Jr G, Sherman BT, Hosack DA *et al*: DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol* 2003; **4**: P3.
23 Voight BF, Kudaravalli S, Wen X, Pritchard JK: A map of recent positive selection in the human genome. *PLoS Biol* 2006; **4**: e72.
24 Sabeti PC, Reich DE, Higgins JM *et al*: Detecting recent positive selection in the human genome from haplotype structure. *Nature* 2002; **419**: 832–837.

25 Ashburner M, Ball CA, Blake JA *et al*: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000; **25**: 25–29.
26 Shriver MD, Mei R, Parra EJ *et al*: Large-scale SNP analysis reveals clustered and continuous patterns of human genetic variation. *Hum Genomics* 2005; **2**: 81–89.
27 Kidd KK, Pakstis AJ, Speed WC, Kidd JR: Understanding human DNA sequence variation. *J Hered* 2004; **95**: 406–420.
28 Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES: High-resolution haplotype structure in the human genome. *Nat Genet* 2001; **29**: 229–232.
29 Cann RL, Stoneking M, Wilson AC: Mitochondrial DNA and human evolution. *Nature* 1987; **325**: 31–36.
30 Bersaglieri T, Sabeti PC, Patterson N *et al*: Genetic signatures of strong recent positive selection at the lactase gene. *Am J Hum Genet* 2004; **74**: 1111–1120.
31 Osorio A, de la Hoya M, Rodriguez-Lopez R *et al*: Over-representation of two specific haplotypes among chromosomes harbouring BRCA1 mutations. *Eur J Hum Genet* 2003; **11**: 489–492.
32 Osorio A, Martinez-Delgado B, Pollan M *et al*: A haplotype containing the p53 polymorphisms Ins16 bp and Arg72Pro modifies cancer risk in BRCA2 mutation carriers. *Hum Mutat* 2006; **27**: 242–248.
33 Goodman JE, Mechanic LE, Luke BT, Ambs S, Chanock S, Harris CC: Exploring SNP–SNP interactions and colon cancer risk using polymorphism interaction analysis. *Int J Cancer* 2006; **118**: 1790–1797.
34 Sul J, Yu GP, Lu QY *et al*: P53 Codon 72 polymorphisms: a case–control study of gastric cancer and potential interactions. *Cancer Lett* 2006; **238**: 210–223.
35 Ohayon T, Gershoni-Baruch R, Papa MZ, Distelman Menachem T, Eisenberg Barzilai S, Friedman E: The R72P P53 mutation is associated with familial breast cancer in Jewish women. *Br J Cancer* 2005; **92**: 1144–1148.
36 Schabath MB, Wu X, Wei Q, Li G, Gu J, Spitz MR: Combined effects of the p53 and p73 polymorphisms on lung cancer risk. *Cancer Epidemiol Biomarkers Prev* 2006; **15**: 158–161.
37 Siddique M, Sabapathy K: Trp53-dependent DNA-repair is affected by the codon 72 polymorphism. *Oncogene* 2006; **25**: 3489–3500.

Supplementary information accompanies the paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)