The Diversity Profile of *TP53* Is Influenced by Positive Selection on the Immediately Upstream Locus *WDR79*

Santos Alonso^a Neskuts Izagirre^a Saioa López^a Isabel Smith-Zubiaga^b

Montse Hervella^a María Dolores Boyano^c Yoana Arroyo-Berdugo^c

Jesús Gardeazabal^d José Luís Díaz-Ramón^d Ana Sánchez Díez^e

Jesus María Careaga^e Concepción de la Rúa^a

Departments of ^aGenetics, Physical Anthropology and Animal Physiology, ^bZoology and Animal Cell Biology, and ^cCell Biology and Histology, University of the Basque Country, Leioa, ^dDermatology Service, Hospital de Cruces, Cruces-Barakaldo, and ^eDermatology Service, Hospital de Basurto, Bilbao, Spain

Key Words

TP53 • *WDR79* • Arg72Pro • rs1042522 • rs2287499 •
Positive selection • HapMap

Abstract

Background/Aim: *TP53* is an efficient central node in a signal transduction network that responds to minimize cancer. However, over 50% of tumors show some mutation in *TP53*. Thus, one might argue that this single central node network lacks robustness. Therefore, we wanted to investigate if natural selection has played a role in shaping the genomic region containing *TP53*. **Methods:** We have analyzed the HapMap data for evidence of selection using F_{ST} pairwise comparisons and the extended haplotype homozygosity test on a 200-kb region encompassing *TP53*. We have also resequenced 4 kb upstream *TP53* in Europeans (including melanoma patients), Asians, Australian Aborigines and Africans. **Results:** Genetic hitchhiking by a linked, positively selected allele at the nearby gene *WDR79* may be partly responsible for the sequence diversity profile of *TP53*. It can help explain why the *TP53* Arg72 allele is the major allele in Europeans

even when the alternative allele, 72Pro, has been reported to offer an increased longevity after disease. **Conclusions:** Despite the important role of *TP53*, a complex interplay with other evolutionary forces, which are extrinsic to *TP53* function, may have driven the genetic diversity pattern of this locus, and, as a consequence, its structure and function.

Copyright © 2009 S. Karger AG, Basel

Introduction

The tumor suppressor gene *TP53* responds to a wide variety of stress signals by regulating a set of genes that can mediate cell cycle arrest and apoptosis, facilitate DNA repair, or alter other cellular processes. Thus, *TP53* functions as a central node in a signal transduction network that responds against any damage to the cell that can lead to cancer or other pathologies [1, 2]. However, given the fact that over 50% of a wide variety of cancers show some mutation in *TP53* [3], one might also argue that this single central node network lacks robustness [1]. The presence of polymorphisms in the *TP53* network

genes may further involve that individuals could have higher or lower efficient *TP53* pathways, which may influence cancer risk or treatment [3]. Thus, an Arg72Pro non-synonymous polymorphism in *TP53* (rs1042522) has been associated to susceptibility to human papillomavirus [4], to skin cancer [5–7] or to breast cancer [8] among others, although these results have been challenged as inconsistent [9–12]. However, one of the largest association studies provides strong evidence for a dose-dependent association of the 72Pro allele to increased longevity after diagnosis of any life-threatening disease [12]. Interestingly, in Europeans the major allele is, however, the alternative Arg allele. Thus, in the HapMap CEU sample, the frequency of the Arg allele reaches 0.767; whereas in the CHB sample it is 0.511; in the JPT sample 0.591; and in the African Yoruba sample 0.331. This raises the question of how this derived allele (Pro seems to be ancestral) has reached such a high frequency in this population.

In this regard, the study of the forces that maintain genetic diversity has always been a topic of utmost importance, as it is related to evolutionary adaptation to the environment and function novelty. Thus, for instance, in the last few years a substantial amount of work has been dedicated to scan the human genome for signatures of positive selection ([13–25], among others). Regarding *TP53*, it has been proposed that the *TP53* Arg72Pro polymorphism is maintained by natural selection as an ecological adaptation to UV irradiation or other factors linked to climate [26, 27].

In this context, we wanted to investigate if the genomic region containing *TP53* shows evidence of selection in human populations. However, we have found that other evolutionary forces, extrinsic to *TP53* function, may have also exerted a strong influence on the sequence diversity profile of this locus, and as a consequence, on its structure and function.

Method

F_{ST} Pairwise Comparisons

SNP frequencies for single SNPs across a 200-kb region encompassing *TP53*–*WDR79* were downloaded from the HapMap web page for each of the three populations: Yorubans, Caucasians and Asians (Chinese + Japanese). All *F_{ST}* pairwise values were obtained as described in [13]. To obtain the neutral distribution of the *F_{ST}* statistic we used the data reported previously in [20]. Briefly, we selected 43 regions distributed across the autosomal genome that belong to broader regions of low gene density, and which are separated at least 150 kb from the closest exon. For each region we downloaded the SNP frequency information available

from the HapMap browser (data Rel #20/phase II on NCBI B35 assembly, dbSNP b125) for the 3 major populations (Caucasians: 153,339 SNPs, Yorubans: 123,798 SNPs and Chinese: 33,190 SNPs). We further filtered the number of SNPs to include only those SNPs that: (a) were separated at least 100 kb from each other; (b) have been genotyped in all three populations, and (c) in which at least one of the three major populations had a minor allele frequency (MAF) higher than 0.1. A final list of 546 SNPs satisfied these criteria. The mean and 95% upper limits (in parentheses) of the observed *F_{ST}* distributions were: Caucasians-Chinese: 0.08 (0.33); Caucasians-Yorubans: 0.14 (0.47) and Chinese-Yorubans: 0.16 (0.45).

Extended Haplotype Homozygosity Test

For the extended haplotype homozygosity (EHH) test, we initially used the HapMap haplotype data in combination with Sweep 1.1 (downloadable from <http://www.broad.mit.edu/mpg/sweep/index.html>) to fast-scan and select, in a visual way, the best candidate two-SNP core. We initially focused on those HapMap haplotypes that showed both substantial frequency and long, high EHH (i.e. high EHH values from the core SNPs, in both 5' and 3' directions). We chose a core region formed by SNPs rs2287499, a non-synonymous C/G SNP mapping at chr17:7532893 (first CDS of *WDR79*, an immediately 5' flanking gene to *TP53*) and rs2287498, a synonymous A/G SNP mapping at chr17:7533285 (second CDS of *WDR79*). Choosing two SNPs as a core region is arbitrary. The only condition was that there was not obvious recombination across the core SNPs. By means of a Perl script, we then calculated the EHH values as in [14] for a region extending about 100 kb from the closer core SNP. To test the significance of the test, we used the approach described in [28]. Briefly, by means of msHOT [29], we run coalescent simulations including heterogeneity in recombination rates across the region. Thus, in each case we discretized the recombination rates using the following approach: we obtained the average and standard deviation (SD) of the distribution of recombination rates and first considered all recombination values greater than the average plus 3 SD. If these outlier regions were consecutive, a local average was calculated for each extended region, otherwise single point estimates were used for a region flanked between the previous and the next recombination value. Next, we excluded these outliers and repeated the process. Recombination rates that were higher than the new global average plus 2 SD were considered again as a class each, except if they were consecutive, in which case a local average was estimated for that extended region. The average of the rest of recombination rates was considered as the background recombination rate, which was used as a reference to estimate the relative intensity of recombination for the rest of recombination classes. See the Appendix for the full command line.

The demographic parameters used in the simulations were obtained as follows. We used as an initial reference for the human demographic parameters those described in [30]. Thus, we used an ancestral population size of 24,000 for the African population (population 1) and 7,700 for both Asians and Caucasians (populations 2 and 3, respectively). We further slightly fine-tuned the demographic parameters by an iterative process, in which the optimization criterion was non-departure from the *F_{ST}* pairwise distributions obtained from the HapMap data (see '*F_{ST}* pairwise comparisons' above). We finally assumed a migration rate matrix $M_{ij} = \{0, 0.05, 0.4, 0.1, 0.3, 0.8, 2.5, 0\}$ for *i* and *j* values from 1 to

3. Looking back in time, we assumed two bottlenecks with instant population reduction, each followed by a population fusion: one approximately 40,000 years ago, in which the Chinese population reduced its size to approximately one sixth. About 2,000 years after this episode, the Chinese population fuses with the European population. Assuming a generation time of 20 years, this represents an F value of 0.04 for this bottleneck. A second population bottleneck takes place about 90,000 years ago. On this occasion, the Eurasian population suffers a reduction in size to one sixth of its previous size. About 10,000 years after this bottleneck ($F = 0.21$), the Eurasian population fuses with the African population. Using these parameters, the mean and 95% upper limits (between brackets) of simulated F_{ST} distributions were: Caucasians-Chinese: 0.08 (0.31); Caucasians-Yorubas: 0.15 (0.44) and Chinese-Yorubas: 0.15 (0.45). The optimized, simulated F_{ST} distribution and the observed distribution of pairwise F_{ST} values between the HapMap populations (see Method, pairwise F_{ST} comparison) recorded Kolmogorov-Smirnov D values of 0.0356 for Caucasians vs. Asians (p value 0.891), 0.0748 for Caucasians vs. Africans (p value 0.104), and 0.0441 for Asians vs. Africans (p value 0.683). As both simulated and real data are not statistically different, we used those demographical parameters described above to estimate the neutral distribution of EHH values by simulation. No further refinements were pursued.

In these simulations, the 'core' consisted in the set of the first n SNPs, where n is the number of SNPs in the original (HapMap data) core. One of the core haplotypes in the simulations had to show the same ancestral/derived states for the SNPs composing the original core. For this, the ancestral information was obtained by comparing the corresponding orthologous regions from the Chimpanzee genome sequence and the *Macaca mulatta* genome sequence obtained from the Golden Path genome browser (<http://genome.ucsc.edu/cgi-bin/hgGateway>). We discarded those simulations that (a) did not result in a number of distinct core haplotypes (as defined from the simulated core set of SNPs) identical to that observed in the original HapMap data using the corresponding observed core SNPs, and where (b) the frequency of the chosen core haplotype did not match the observed core haplotype (plus a 2% allowance). We typically ran the program until about 500 simulations satisfying all the conditions were obtained. For each set of simulations satisfying these conditions, we obtained the 95% upper percentile of the expected EHH distribution for each of a series of consecutive 1-kb-long windows spanning the region. We finally declared a core haplotype as under selection if the distance for which the EHH values were equal to 1 for the SNPs in the HapMap data was longer than that observed from the 95% upper percentile distribution of the EHH values obtained in the simulations.

Multiple Testing Corrections

To control for multiple testing we used the approach by [31], which sets the false discovery rate (FDR) at a level α , by first, ranking the initial p values in ascending order $P_{(1)} \leq P_{(2)} \leq \dots \leq P_{(m)}$, being m the number of tests, and second, by specifying $P_{(i)} \leq \alpha^* i/m$ as the point below which there is no rejection at a FDR α .

DNA Samples and Resequencing

In an initial sample of 116 human chromosomes we have resequenced over 4 kb of a genomic region including (a) the 5' UTR of *TP53* (224 bp long); (b) the 3' adjacent 1,346 bp of the first in-

tron of *TP53*, and (c) 2,630 bp of the immediately 5'-flanking region of *TP53*, which included the 5' UTR, coding exon 1, intron 1, coding exon 2, intron 2, coding exon 3 and 568 bp of intron 3 of *WDR79* (also known as *WRAP53*). The individuals sequenced are an arbitrary representation of human world populations and include: 20 chromosomes from Biaka and Mbuti Pygmies (DNA purchased from the European Collection of Cell Cultures, ECACC), 20 chromosomes from Senegalese individuals residing in Spain, 42 European (Basque) chromosomes (including 20 chromosomes from melanoma patients), 20 Asian chromosomes including Chinese samples from Coriell Cell Repositories and Chinese residents in Spain, and 14 chromosomes from Australian aborigines (DNA purchased from the ECACC).

DNA was PCR-amplified in overlapping segments ~ 1 kb long and these were resequenced using BigDye 3.1 chemistry and an ABI PRISM 3730 and ABI 310 DNA analyzers. Sequences were edited with Genalys v3.3.45a (M. Takahashi). Primers and PCR conditions are available on request. Sequences have been deposited in GenBank, accession numbers EU876946–EU877061.

Estimation of the Haplotypes from Resequencing Data

To solve the phase of haplotypes obtained from our resequencing data we first run PHASE [32]. For those pairs of SNPs that did not get a probability greater than 0.95 we solved experimentally their phase by ARMS-PCR and/or cloning (using the TOPO-TA kit from Invitrogen) plus resequencing.

Population Parameters and Neutrality Tests from Resequencing Data

Population diversity parameters and neutrality statistics like Tajima's D [33], were obtained by means of DnaSP 4.5 [34]. The population differentiation S_{nn} test [35] is also included in DnaSP 4.5. Fay and Wu's H , Ewens-Watson's EW , and the DH , HEW and $DHEW$ tests were performed using software provided by Kai Zeng [36].

Genealogical Relationships among the Haplotypes Obtained from Resequencing

Graphical representations of the genealogical relationships among the haplotypes obtained from our resequencing data were estimated by the Median-Joining (MJ) algorithm implemented in Network 4.5.0.0 [37]. The ancestral haplotype was inferred using parsimony by comparison to the chimp and rhesus macaque sequences.

SNP Genotyping

We genotyped the coding non-synonymous SNP rs1042522 at *TP73* codon 72 (Arg72Pro) by PCR amplifying genomic DNA from a further set of 101 control samples plus 115 melanoma patients, not included in the resequencing study, with primers 5'-GCAAGTCACAGACTTGGCTG-3' (Fw) and 5'-GAAGACCCAGGTCCAGATG-3' (Rev), and digesting the PCR products with *Bst*UI. The coding non-synonymous SNP rs2287499 from *WDR79* was also genotyped in 106 controls and 112 melanoma patients using primers 5'-TGAAGACTTTGGAGACTCAAC-3' (Fw) and 5'-CCAGAGGTATCTTCCATGGC-3' (Rev), and digesting the PCR products with *Bfa*I.

Fig. 1. Pairwise F_{ST} values for single SNPs across a 200-kb region encompassing *TP53*–*WDR79*. Circled SNPs (rs8064946, rs2287499 and rs7640) present significant values for the CEU vs. YRI comparison even after multiple test correction ([31]; false discovery rate of 5%). On top, the approximate location of the loci within the region is indicated. The arrows indicate the sense of transcription.

Results

An Exploratory Scan for Selection Using the HapMap Data

To check for possible departures from neutrality, we have analyzed the HapMap data using the F_{ST} and the EHH statistics. We are aware that the HapMap data suffers from ascertainment bias and that this can complicate the detection of selection when using F_{ST} tests [38, 39], although it is less likely to bias the analyses of long haplotypes (i.e. EHH) [40]. Therefore, we consider these analyses based on the HapMap data (F_{ST} and EHH) only as exploratory. Final proof for selection will be provided by the resequencing study.

Significantly low or high pairwise F_{ST} values can give us a clue on the presence of selection. The pairwise F_{ST} distribution of SNPs across the *TP53*–*WDR79* region (covering ~200 kb) shows a cluster of SNPs with significantly high F_{ST} values for the HapMap CEU vs. YRI comparison (even after multiple test correction) upstream of *TP53*, among which rs2287499 is included (fig. 1). Actually, both the CEU vs. YRI comparison and ASI vs. YRI comparison show high F_{ST} values at rs2287499 (and surrounding SNPs), although we only find statistically significant those for CEU vs. YRI. This could indicate a differentiation of YRI from both Asians and Caucasians, or a differentiation of both Caucasians and Asians from YRI. Given the EHH profiles and the sequencing data (see below) we rather support the latter. The *WDR79* rs2287499-G allele is a derived allele which seems to have substantially increased its frequency particularly in Europeans (frequency in the HapMap CEU sample 0.842;

CHB sample 0.67; JPT sample 0.761; whereas in the African Yoruban sample it is 0.083). To confirm the possibility of positive selection, we applied the EHH test [14] to the HapMap Caucasian sample around the *TP53* gene region. The EHH test is particularly suited to detect recent positive selection (less than 30,000 years old; [41]) and it has been shown to have high power to detect selection when the sites under selection are part of the core set of SNPs used to measure the extent of linkage disequilibrium (LD) [42]. In this regard, one core haplotype formed by alleles rs2287499-G and rs2287498-G reaches a frequency of 84.17% in the HapMap Caucasian sample and shows a longer than expected haplotype homozygosity (fig. 2). These two SNPs belong to the immediately *TP53* 5'-flanking locus *WDR79*, and the haplotype under selection is one mutational step away from the inferred ancestral core haplotype rs2287499-C, rs2287498-G (by comparison to the chimp and macaque sequence; HapMap nomenclature). Taken together, these results are suggestive of recent, possibly ongoing, positive selection on *WDR79*, an upstream locus immediately flanking the *TP53* gene.

Evidence for Selection from Sequence Haplotype Diversity

To confirm this observation more robustly, we decided to resequence approximately 4 kb of the *TP53* upstream genomic region in a sample of 116 chromosomes, including healthy Europeans, European melanoma patients, Africans, Asians and Australian Aborigines. In this region we have detected 16 SNPs. We also found three polymorphic indels, one of them involving 2 bp at

Fig. 2. Extended haplotype homozygosity (EHH) tests. EHH test for a ~200-kb region, encompassing *TP53*–*WDR79* in the HapMap Caucasian, Asian and Yoruba samples. After scanning the HapMap data with Sweep (<http://www.broad.mit.edu/mpg/sweep/>), candidate core haplotypes were tested for the significance of their EHH. Circles represent the observed EHH values extending about 100 kb from the core haplotype rs2287499–G rs2287498–G in both directions, white for 3' and gray for 5'. Dots represent the corresponding 95% upper percentiles of the simulated EHH distributions. For Caucasians, the expected EHH values decay earlier than the observed values on both ends. On top, the approximate location of *TP53*, *WDR79* and other genes with respect to the core haplotypes is indicated. The arrows indicate the sense of transcription.

Fig. 3. Median-Joining network describing the genealogical relationships among the haplotypes found in the resequenced 5' region of *TP53*. Areas are proportional to relative frequency. Blue circles: Chinese; black circles: Europeans; red circles: European melanoma patients; pink circles: Senegalese; green circles: Australian Aborigines, and yellow circles: Pygmies. The ancestral allele is one mutational step away from SE7A and has not been detected in the global sample resequenced.

approximately chr17:7531870 and the other two involving 1 bp each, both within a run of 'C's located at chr17:7532244. A microsatellite polymorphism derived from a run of 'A's at chr17:7530358 was also found, but it was not further considered, partly because of its different mutational nature to SNPs and partly because of the uncertainty in the reads due to the intrinsic difficulty in sequencing through these regions. Out of the 16 SNPs, five were newly discovered SNPs: of these, three SNPs were found within the third intron of *WDR79*, a close 5'-flanking gene to *TP53* (two of them within an AluSx element); a fourth SNP was found within the first coding exon of *WDR79* and creates a non-synonymous polymorphism whose derived allele is a singleton. Finally, a fifth SNP was found within the 5' UTR of *WDR79*. The rest of the detected SNPs, eleven in total, have already been described in dbSNP build 128 (rs6503051, rs6503050, rs34289079, rs17886268, rs2287497, rs2287498, rs17885803, rs2287499, rs17886079, rs17886250 and rs9905505). Overall, two SNPs (rs17886250 and rs9905505) were found within the 1,346 bp long region resequenced of the first intron of *TP53*, no SNP was found within the 5' UTR of *TP53* (224 bp long), and 14 SNPs were found within the proximal 5'-flanking region of *TP53* (2,630 bp long; includes part of *WDR79*). The number of SNPs in the 5'-flanking region, 5' UTR and the *TP53* intronic region showed no significant differences (by pairwise Fisher's exact test, data not

shown). These polymorphisms assemble into 23 different haplotypes, whose genealogical relationships are shown in figure 3.

For the subsequent diversity analyses we did not consider the indel polymorphisms due to its different nature to SNPs. Diversity measures for all of the samples, are summarized in table 1. As the healthy Europeans sample and the European melanoma patients sample did not show significant population differentiation ($S_{nn} = 0.479$, $p = 0.526$) we pooled both samples together. This lack of differentiation between healthy Europeans and melanoma patients was further confirmed by genotyping two SNPs (rs1042522 at *TP73* codon 72, responsible for the Arg72Pro polymorphism, and rs2287499, from the EHH core haplotype above described) in a different, broader European sample set of over 100 controls and over 100 melanoma patients (see below).

Overall, the 4-kb region resequenced shows higher haplotype diversity in the African populations analyzed (table 1), as expected under an out-of-Africa model of human evolution. Initially, none of the populations showed significant departure from neutrality (as measured by standard Tajima's D). One possible exception might be the European sample, which shows a Tajima's D value with a p value below 0.05, but this value is not longer significant after Bonferroni correction (table 1). A similar situation is observed after calculating Tajima's D p value

Table 1. Diversity parameters and neutrality test for the resequenced region of *TP53*

	Pygmies	Senegalese	Australian Ab.	Chinese	Europeans ^a
Number of sequences	20	20	14	20	42
Number of segregating sites (S)/haplotypes (h)	12/10	12/12	5/3	6/6	10/8
Haplotype diversity, Hd (SD)	0.92 (0.03)	0.93 (0.04)	0.56 (0.12)	0.78 (0.06)	0.54 (0.01)
Nucleotide diversity, Pi (SD) ($\times 10^{-4}$)	8.7 (0.8)	8.1 (0.8)	3.5 (1.3)	5.8 (0.7)	2.6 (0.7)
Theta (per site) from S, Theta-W (SD ^b) ($\times 10^{-4}$)	8.4 (3.6)	8.4 (3.6)	3.9 (2.2)	4.2 (2.1)	5.8 (2.4)
Average number of nucleotide differences, k	3.511	3.237	1.418	2.316	1.028
Tajima's D value (seg. sites)/[p(D $\leq$ D-obs)]	0.136/0.613 ^c	-0.154/0.503 ^c	-0.341/0.417 ^c	1.177/0.901 ^c	-1.648/0.028 ^c [0.035] ^d
Fay and Wu's H [p(H-neutral $\leq$ H-obs)]				-0.159 [0.274]	-2.481 [0.025]
Ewens and Watterson EW [p(EW-neutral $\leq$ EW-obs)]				0.255 [0.639]	0.468 [0.034]
p values DH test/HEW test				0.799/0.417	0.005 ^c /0.002 ^c
p value DHEW test				0.746	0.0013 ^c

^a Healthy individuals plus melanoma patients; ^b assuming no recombination; ^c using the standard neutral model in DnaSP, with no recombination; ^d using a demography-corrected neutral model, with 1 cM/Mb; ^e also significant after Bonferroni multiple test correction (assuming 5 tests).

using coalescent simulations including population growth and recombination (a homogeneous recombination rate of 1 cM/Mb). Similarly for the Fay and Wu H statistic, or the Ewens-Waterson EW statistic. However, when it comes to detect positive selection, the use of compound tests could result in a more specific and powerful approach [36, 42]. In particular, Tajima's D may be powerful after the fixation of the selected allele, particularly when the ratio between the recombination rate to the mutation rate is approximately 1, as it is the case here. However, if the selected allele is still segregating, compound tests will be more powerful [36, 42]. For instance, for a sample size of 50, $\theta \approx 2.5$, a present-day frequency of the selected allele of ≈ 0.7 (conditions similar to those for the European sample in this study) and selection coefficients $\alpha = 300$ and $h = 0.5$, the power of the HEW test to detect selection is approximately 0.5, whereas for Tajima's D this figure is less than 0.1 [36]. The compound statistic DH seems to show high specificity to positive selection in comparison to Tajima's D or the EW test [36]. Other compound statistics like HEW and DHEW are more robust than DH to the presence of recombination and are even more powerful in detecting positive selection. Furthermore, DHEW is relatively insensitive to background selection and demography [36]. Finally, in general, the rejection rates of DH and DHEW are not affected by subdivision when $F_{ST} \leq 0.25$. Usually HEW's rejection rate is reasonably low when $F_{ST} \leq 0.2$ [36, 42]. When these compound tests were applied to our data only the European sample showed a consistent significant p value, even after Bonferroni correction (table 1), which indicates that positive selection is ongoing in this genomic region. It is

likely that, as a consequence of pooling, the increase in the sample size of the European pooled sample compared to the rest of the samples is providing an edge in the power to detect selection in Europeans. However, it seems unlikely that it is responsible for the wide differences in the neutrality tests when compared for instance to the Asian sample (see table 1).

Positive selection on *WDR79* might thus help explain the unexpected high frequency of the derived allele of the Arg72Pro polymorphism in Europeans. LD between *WDR79* rs2287499 and *TP53* Arg72Pro for HapMap CEU is, however, modest: $D' = 0.48$; $r^2 = 0.14$; $\chi^2 = 8.54$, two-tailed $p = 0.0013$; using CubeX [43] (fig. 4). Nevertheless, this can be explained by considering that we are not likely dealing with a standard selective sweep, in which a newly arisen allele is selected and reaches fixation. The fact that the selected *WDR79* allele is also present in Africans and Asians indicates that the selection process in Europeans may likely have started from standing variation. In addition, the selected allele has not reached fixation yet in Europeans. Both factors can explain why neutrality tests like Tajima's D can not detect selection (Tajima's D would have been more powerful just after fixation). It can also explain why the detected LD values are moderate, as in standard LD measures all the diversity is taken into account. In these cases, internal LD measures within the selected haplotype (i.e. EHH) can be more informative. The EHH value at *TP53* Arg72 from *WDR79* rs2287499 is ~ 0.7 for the selected haplotype, whereas it is ~ 0.55 in Asians. Thus, there is an increase in the association between the *TP53* Arg72 allele and the selected *WDR79* rs2287499 allele in Euro-

Fig. 4. Linkage disequilibrium between *TP53* Arg72Pro and *WDR79* rs2287499 in HapMap Caucasians. D' and r^2 values are indicated within the squares. The p values were obtained from a $3 \times 3 \chi^2$ test using CubeX [43].

peans compared to Asians, which can explain the proportional increase in the frequency of the *TP53* Arg72 allele in Europeans.

Exploring for Association to Melanoma Susceptibility

Some studies have proposed that the Arg72Pro polymorphism is maintained by natural selection as an ecological adaptation to UV irradiation [26, 27]. In this regard, [44] showed that UV induction of the melanogenic α -melanocyte-stimulating hormone (α -MSH) in skin is directly controlled by *TP53*. Other studies have pointed to a relationship between the *TP53* Arg72Pro polymorphism and susceptibility to different cancer types, including melanoma [6], although this association however, has not gone unchallenged. Thus, a large association study found no association of Arg72Pro to six subtypes of cancer, including melanoma and non-melanoma skin cancer [12, 45]. On the other hand, a relationship between estrogen-negative breast cancer and the *WDR79* rs2287499 SNP has also been described [46].

It could be argued that because in present times death from melanoma is particularly more frequent in the elderly, the evolutionary impact of melanoma is null. However, the melanoma-associated death rate seems not to be negligible in people in the reproductive age. Thus, among young people (aged 25–49) the rate is increasing in a magnitude of 2–3% per year in some north and west European countries and up to 8% in Spain [47]. Besides, [48]

reviews how works done during the 70s and 80s point out that in Nigeria and Tanzania all the ‘Negroid’ albinos studied exhibited skin cancers or pre-malignant lesions by the age of 20. Only 6% of Nigerian albinos were in the age-range of 31–60 years compared to 20% of the non-albinos. Finally, the evolutionary impact of melanoma could have been higher in our evolutionary past. We, in fact, ignore what was the effect of prolonged sun exposure in prehistoric human hunter-gatherer societies.

Thus, given the evidence found for positive selection at *WDR79* (or the 5' flanking region of *TP53*), we decided to explore if this role of selection could be straightforwardly related to a possible protective effect against cutaneous melanoma in our population. Thus we genotyped rs2287499 (from the EHH core haplotype at *WDR79*) and rs1042522 (responsible for the *TP53* Arg72Pro polymorphism) in a second European sample of over 100 controls and over 100 melanoma patients (see Method). However, in our samples, the European control sample and the European melanoma patients sample did not show significant allele frequency differences. Thus, the frequency of *TP53* Arg72 is 0.787 in controls, and 0.769 in melanoma patients (two-sided Fisher exact test p value = 0.728). For *WDR79* rs2287499, the frequency of the G allele is 0.847 and 0.857 in controls and melanoma patients, respectively (two-sided Fisher exact test p value = 0.79). It is important to note that because this is a C/G SNP, allele nomenclature can be confusing if the reference strand is not stat-

ed. We have followed the HapMap nomenclature, which does not agree with that used by [46]. The frequency comparison implies that, in our populations, both SNPs are not strongly associated to predisposition to or protection against cutaneous melanoma. However this observation should be considered only as exploratory, as it is not our purpose in this work to conduct a thorough test of association.

Discussion

In this work, by using HapMap data-based F_{ST} and EHH tests as well as a battery of neutrality tests based on our own resequencing data, we have found that a region immediately upstream *TP53* shows a robust signature of positive selection in Europeans. The use of F_{ST} tests for ascertaining positive selection has been controversial. Detractors argue that ascertainment bias in the HapMap SNPs [38–40] or demography [49] may interfere with the detection of selection by means of F_{ST} tests. Supporters instead argue that because the proportion of genic SNPs with high F_{ST} is about 2.8-fold larger than the proportion of non-genic SNPs with equally large F_{ST} values, then the F_{ST} test is still a useful way to detect the signature of selection despite the demographic effects related to the colonization of Eurasia by modern humans [50]. In our case, nevertheless, all tests (F_{ST} , EHH and several tests based on unbiased resequencing) converge in showing a signature of selection in Europeans.

The EHH analysis shows that this region is centred on *WDR79* (WD-repeat domain 79, aka FLJ10385 or *WRAP53*), a locus that contains six tryptophan-aspartate (WD) repeat domains. WD repeats may serve as sites of protein-protein interaction for adaptor proteins and facilitate multiprotein complex formation. Unfortunately, there is not much literature on the function of *WDR79*, although we know that the protein has been reported to be upregulated in parathyroid [51] and brain tumors [52]. Also, the discovery that silencing of this locus results in massive cell apoptosis has led to the patenting of a new method of treatment of hyperproliferative conditions (i.e., cancer) (International Application No. PCT/SE2007/050692). *WDR79* is transcribed in an antisense orientation relative to *TP53*, and thus, it has been argued that proximal promoter elements may be juxtaposed [46]. Therefore, the region in *WDR79* under selection might include regulatory elements for *TP53*.

In particular, both the F_{ST} and EHH analyses indicate that the region under selection seems to be centered on

WDR79 rs2287499, a SNP that has been associated to breast cancer susceptibility in estrogen receptor-negative tumors [46]. The selected haplotype seems to include the G allele at this locus (following the HapMap nomenclature, but the C allele in the nomenclature used by [46]), which according to [46] is the normal allele. We did not, however, find an association between these SNPs and melanoma predisposition, but maybe larger and better characterized samples would allow us to bring up more subtle associations.

Given that little is known about the function of *WDR79*, therefore, until we have further information any hypothesis on why it seems to be under positive selection would be simply speculative. Nevertheless, the fact is that positive selection on *WDR79* and genetic hitchhiking might help explain the diversity profile of *TP53* in Europeans. Thus, internal LD between the positively selected *WDR79* allele and the *TP53* Arg72 allele may explain, at least partly, why *TP53* Arg72 is the major allele in Europeans, even when the alternative allele, *TP53* 72Pro, seems to offer the allure of an increased longevity after diagnosis of any life-threatening disease [12]. It can be argued, however, that this may not be a trait that confers a selective advantage, but, nevertheless, it can serve to illustrate how evolutionary factors that are in principle unrelated to *TP53* function can shape the diversity profile of *TP53*, which in its turn may have an influence on the function of this gene. It remains possible, however, that further work can demonstrate a regulatory role of *WDR79* rs2287499 on *TP53* gene expression. Until then, it seems that the sequence diversity profile at *TP53* may be the result of a complex trade-off with other evolutionary forces extrinsic to the key role of *TP53* in minimizing cancer risk, at least in Europeans. This may have consequences for the structure and function of *TP53*.

Appendix

The command line for Caucasians (for the 3' end) was:

```
msHOT 120 100 -s 103 -r 81 120000 -v 7 2465 31629 2.4 31630
40121 12.3 40122 42776 4.7 42777 54157 2.3 54158 54736 3.6 54737
57133 2.4 97405 100000 1.9 -I 3 0 0 120 -n 2 0.32 -n 3 0.32 -ma ×
0.05 0.4 0.1 × 3 0.8 2.5 × -en 0.02083 2 0.15 -ej 0.02093 2 3 -en
0.047 3 0.15 -ej 0.0525 3 1
```

and for the 5' end:

```
msHOT 120 100 -s 110 -r 75 140000 -v 5 6008 10049 2.5 10050
14408 3.5 14409 26572 3 26573 33089 3.7 52845 53881 2.25 -I 3 0
0 120 -n 2 0.32 -n 3 0.3 2 -ma × 0.05 0.4 0.1 × 3 0.8 2.5 × -en
0.02083 2 0.15 -ej 0.02093 2 3 -en 0.047 3 0.15 -ej 0.0525 3 1.
```

Simulations for the two other populations were obtained by modifying correspondingly the $-I$ parameters.

Acknowledgements

The SGI/IZO-SGIker UPV/EHU (supported by the National Program for the Promotion of Human Resources within the National Plan of Scientific Research, Development and Innovation – Fondo Social Europeo, MCyT and Basque Government) is gratefully acknowledged for the generous allocation of computational

resources. The authors also wish to thank Garret Hellenthal for kindly providing a modified version of msHOT, and Kai Zeng for the software to compute the DH, HEW and DHEW tests. Work was funded by grants GIC07/43 (Basque Government) and CGL2008-04066/BOS (Spanish Ministry of Science and Innovation).

References

- Vogelstein B, Lane D and Levine AJ: Surfing the p53 network. *Nature* 2000;408:307–310.
- Levine AJ, Finlay CA, Hinds PW: P53 is a tumor suppressor gene. *Cell* 2004;116:S67–S69.
- Levine AJ, Hu W and Feng Z: The P53 pathway: what questions remain to be explored? *Cell Death Differ* 2006;13:1027–1036.
- Thomas M, Kalita A, Labrecque S, Pim D, Banks, Matlashewski G: Two polymorphic variants of wild-type p53 differ biochemically and biologically. *Mol Cell Biol* 1999;19:1092–1100.
- Han J, Cox DG, Colditz GA, Hunter DJ: The p53 codon 72 polymorphism, sunburn and risk of skin cancer in US Caucasian women. *Mol Carcinog* 2006;45:694–700.
- Stefanaki I, Stratigatos AJ, Dimisianos G, Nikolau V, Papadopoulos O, Polydorou D, Gogas H, Tsoutsos D, Panagiotou P, Kanavakis E, Antoniou C, Katsambas AD: p53 codon 72 Pro homozygosity increases the risk of cutaneous melanoma in individuals with dark skin complexion and among noncarriers of melanocortin 1 receptor red hair variants. *Br J Dermatol* 2007;156:357–362.
- Nan H, Qureshi AA, Hunter DJ, Han J: Interaction between p53 codon 72 polymorphism and melanocortin 1 receptor variants on suntan response and cutaneous melanoma risk. *Br J Dermatol* 2008;159:314–321.
- Cavallone L, Arcand SL, Maugard C, Ghadirian P, Mes-Masson AM, Provencher D, Tonin PN: Haplotype analysis of TP53 polymorphisms, Arg72Pro and Ins16, in BRCA1 and BRCA2 mutation carriers of French Canadian descent. *BMC Cancer* 2008;8:96.
- Matakidou A, Eisen T, Houlston RS: TP53 polymorphisms and lung cancer risk: a systematic review and meta-analysis. *Mutagenesis* 2003;18:377–385.
- Koushik A, Platt RW, Franco EL: p53 codon 72 polymorphism and cervical neoplasia: a meta-analysis review. *Cancer Epidemiol Biomarkers Prev* 2004;13:11–22.
- Zhou Y, Li N, Zhuang W, Liu GJ, Wu TX, Yao X, Du L, Wei ML, Wu XT: P53 codon 72 polymorphism and gastric cancer: a meta-analysis of the literature. *Int J Cancer* 2007;121:1481–1486.
- Bojesen SE, Nordestgaard BG: The common germline Arg72Pro polymorphism of p53 and increased longevity in humans. *Cell Cycle* 2008;7:158–163.
- Akey JM, Zhang G, Zhang K, Jin L, Shriver MD: Interrogating a high-density SNP map for signatures of natural selection. *Genome Res* 2002;12:1805–1814.
- Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ, Ackerman HC, Campbell SJ, Altshuler D, Cooper R, Kwiatkowski D, Ward R, Lander ES: Detecting recent positive selection in the human genome from haplotype structure. *Nature* 2002;419:832–837.
- Kayser M, Brauer S, Stoneking M: A genome scan to detect candidate regions influenced by local natural selection in human populations. *Mol Biol Evol* 2003;20:893–900.
- Storz JF, Payseur BA, Nachman MW: Genome scans of DNA variability in humans reveal evidence for selective sweeps outside of Africa. *Mol Biol Evol* 2004;21:1800–1811.
- Bustamante CD, Fedel-Alon A, Williamson S, Nielsen R, Hubisz MT, Glanowski S, Tanenbaum DM, White TJ, Sninsky JJ, Hernandez RD, Civello D, Adams MD, Cargill M, Clark AG: Natural selection on protein-coding genes in the human genome. *Nature* 2005;437:1153–1157.
- Carlson CS, Thomas DJ, Eberle MA, Swanson JE, Livingston RJ, Rieder RJ, Nickerson DA: Genomic regions exhibiting positive selection identified from dense genotype data. *Genome Res* 2005;15:1553–1565.
- Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C: Genomic scans for selective sweeps using SNP data. *Genome Res* 2005;15:1566–1575.
- Izagirre N, García I, Junquera C, de la Rúa C, Alonso S: A scan for signatures of positive selection in candidate loci for skin pigmentation in humans. *Mol Biol Evol* 2006;23:1697–1706.
- Voight BF, Kudaravalli S, Wen X, Pritchard JK: A map of recent positive selection in the human genome. *PLoS Biol* 2006;4:e72.
- Zhang C, Bailey DK, Awad T, Liu G, Xing G, Cao M, Valmeekam V, Retief J, Matsuzaki H, Taub M, Seielstad M, Kennedy GC: A whole genome long-range haplotype (WGLRH) test for detecting imprints of positive selection in human populations. *Bioinformatics* 2006;22:2122–2128.
- Kimura R, Fujimoto A, Tokunaga K, Ohashi J: A practical genome scan for population-specific strong selective sweeps that have reached fixation. *PLoS ONE* 2007;2:e286.
- Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xiu X, Byrne EH, McCarroll SA, Gaudet R, Schaffner SF, Lander ES: Genome-wide detection and characterization of positive selection in human populations. *Nature* 2007;449:913–918.
- Tang K, Thornton KR, Stoneking M: A new approach for using genome scans to detect recent positive selection in the human genome. *PLoS Biol* 2007;5:e171.
- Beckman G, Birgander R, Sjölander A, Saha N, Holmberg PA, Kivelä A, Beckman L: Is p53 polymorphisms maintained by natural selection? *Hum Hered* 1994;44:266–270.
- Khrunin AV, Tarskaia LA, Spitsyn VA, Lylova OI, Bebyakova NA, Milkulich AI, Limborska SA: p53 polymorphisms in Russia and Belarus: correlation of the 2-1-1 haplotype frequency with longitude. *Mol Gen Genomics* 2005;272:666–672.
- Alonso S, Izaguirre N, Smith-Zubiaga I, Gardeazabal J, Díaz-Ramón JL, Díaz-Pérez JL, Zelenika D, Boyano MD, Smit N, de la Rúa C: Complex signatures of selection for the melanogenic loci *TYR*, *TYRP1* and *DCT* in humans. *BMC Evol Biol* 2008;8:74.
- Hellenthal G, Stephens M: msHOT: modifying Hudson's ms simulator to incorporate crossover and gene conversion hotspots. *Bioinformatics* 2007;23:520–521.
- Schaffner SF, Foo C, Gabriel S, Reich D, Daly D, Altshuler D: Calibrating a coalescent simulation of human genome sequence variation. *Genome Res* 2006;15:1576–1583.
- Benjamini Y, Hochberg Y: Controlling the false discovery rate – a practical and powerful approach to multiple testing. *J Roy Stat Soc B* 1995;57:289–300.

- 32 Stephens M, Smith N, Donnelly P: A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* 2001;68:978–989.
- 33 Tajima F: Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 1989;123:585–595.
- 34 Rozas J, Sanchez-Del Barrio JC, Messeguer X, Rozas R: DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* 2003;19:2496–2497.
- 35 Hudson RR: A new statistic for detecting genetic differentiation. *Genetics* 2000;155:2011–2014.
- 36 Zeng K, Shi S, Wu CI: Compound tests for the detection of hitchhiking under positive selection. *Mol Biol Evol* 2007;24:1898–1908.
- 37 Bandelt HJ, Forster P, Röhl A: Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol* 1999;16:37–48.
- 38 Clark AG, Hubsiz MJ, Bustamante CD, Williamson SH, Nielsen R: Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res* 2005;15:1496–1502.
- 39 Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante CD: Genomic scans for selective sweep using SNP data. *Genome Res* 2005;15:1566–1575.
- 40 Conrad DF, Jakobsson M, Coop G, Wen X, Wall JD, Rosenberg NA, Pritchard JK: A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nat Genet* 2006;38:1251–1260.
- 41 Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, Palma A, Mikkelsen TS, Altshuler D, Lander ES: Positive natural selection in the human lineage. *Science* 2006;312:1614–1620.
- 42 Zeng K, Mano S, Shi S, Wu CI: Comparisons of site- and haplotype-frequency methods for detecting positive selection. *Mol Biol Evol* 2007;24:1562–1574.
- 43 Gaunt TR, Rodríguez S, Day INM: Cubic exact solutions for the estimation of pairwise haplotype frequencies: implications for linkage disequilibrium analyses and a web tool ‘CubeX’. *Bioinformatics* 2007;8:428.
- 44 Cui R, Widlund H, Feige E, Lin JY, Wilensky DL, Igras VE, D’Orazio J, Fung CY, Schanbacher CF, Granter SR, Fisher DE: Central role of p53 in the suntan response and pathologic hyperpigmentation. *Cell* 2007;128:853–864.
- 45 Ørsted DD, Bojesen SE, Tybhaerg-Hansen A, Nordestgaard BG: Tumor suppressor p53 Arg72Pro polymorphism and longevity, cancer survival and risk of cancer in the general population. *J Exp Med* 2007;204:1295–1301.
- 46 García-Closas M, Kristensen V, Langerød A, Qi Y, Yeager M, Burdett L, Welch R, Lissowska J, Peplonska B, Brinton L, Gerhard DS, Gram IT, Perou CM, Børresen-Dale AL, Chanock S: Common genetic variation in *TP53* and its flanking genes, *WDR79* and *ATP1B2*, susceptibility to breast cancer. *Int J Cancer* 2007;121:2532–2538.
- 47 de Vries E, Coebergh JWW: Melanoma incidence has risen in Europe. *BMJ* 2005;331:698.
- 48 Robins AH: Biological perspectives on human pigmentation. *Cambridge Studies in Biological Anthropology* 7. Cambridge, Cambridge University Press, 1991.
- 49 Hofer T, Ray N, Wegmann D, Excoffier L: Large allele frequency differences between human continental groups are more likely to have occurred by drift during range expansions than by selection. *Ann Hum Genet* 2009;73:95–108.
- 50 Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L: Natural selection has driven population differentiation in modern humans. *Nat Genet* 2008;40:340–345.
- 51 Velázquez-Fernández D, Laurell C, Saqui-Salces M, Pantoja JP, Candanedo-Gonzalez F, Reza-Albarrán A, Gamboa-Dominguez A, Herrera MF: Differential RNA expression profile by cDNA microarray in sporadic primary hyperparathyroidism (pHPT): primary parathyroid hyperplasia versus adenoma. *World J Surg* 2006;30:705–713.
- 52 Zhang Y, Eberhard DA, Frantz GD, Dowd P, Wu TD, Zhou Y, Watanabe C, Luoh SM, Polakis P, Hillan KJ, Wood WI, Zhang Z: GE-PIS-quantitative gene expression profiling in normal and cancer tissues. *Bioinformatics* 2004;20:2390–2398.