Título artículo / Títol article:	**Simultaneous purifying selection on the ancestral MC1R allele and positive selection on the melanoma-risk allele V60L in South Europeans**
Autores / Autors	**Martínez Cadenas, Conrado ; López, Saioa ; Ribas, Gloria ; Flores, C. ; García, Oscar ; Sevilla, Arrate ; Smith-Zubiaga, I. ; Ibarrola-Villava, Maider ; Pino-Yanes, M . ; Gardeazabal, Jesús ; Boyano, Dolores ; García de Galdeano, Alicia ; Izagirre, Neskuts ; de la Rúa, Concepción ; Alonso, Santos**
Revista:	**The Journal of Physical Chemistry Letters**
Versión / Versió:	**Post-print**
Cita bibliográfica / Cita bibliogràfica (ISO 690):	**MARTÍNEZ-CADENAS, Conrado, et al. Simultaneous purifying selection on the ancestral MC1R allele and positive selection on the melanoma-risk allele V60L in South Europeans. Molecular biology and evolution, 2013, vol. 30, no 12, p. 2654-2665.**
url Repositori UJI:	http://hdl.handle.net/10234/94790

OXFORD JOURNALS
MOLECULAR BIOLOGY AND EVOLUTION

Simultaneous purifying selection on the ancestral MC1R allele and positive selection on the melanoma-risk allele V60L in South Europeans

Journal:	*Molecular Biology and Evolution*
Manuscript ID:	MBE-13-0413.R1
Manuscript Type:	Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Martinez-Cadenas, Conrado; University of Castellón Jaume I, Dept. Medicine López, Saioa; University of the Basque Country UPV/EHU, Dept. Genetics, Physical Anthropology and Animal Physiology Ribas, Gloria; INCLIVA, Biomedical Research Institute Flores, Carlos; Universitary Hospital N.S. de Candelaria, Research Unit; University Hospital N. S. Candelaria, Research Unit García, Oscar; Forensic Science Unit, Basque Country Police, Forensic Genetics Section Sevilla, Arrate; University of the Basque Country UPV/EHU, Dept. Genetics, Physical Anthropology and Animal Physiology Smith-Zubiaga, Isabel; University of the Basque Country UPV/EHU, Dept. Zoology and Animal Cell Biology Ibarrola-Villaba, Maider; INCLIVA, Biomedical Research Institute Pino-Yanes, Maria del Mar; University Hospital N. S. Candelaria, Research Unit Gardeazabal, Jesús; Cruces Hospital, Dermatology Service Boyano, Dolores; University of the Basque Country, Cell Biology and Histology García de Galdeano, Alicia; University of the Basque Country, Cell Biology and Histology Izagirre, Neskuts; University of the Basque Country, Genetics, Physical Anthopology and Animal Physiology de la Rúa, Concepción; University of the Basque Country, Genetics, Physical Anthropology and Animal Physiology Alonso, Santos; University of the Basque Country, Genetics, Physical Anthropology and Animal Physiology
Key Words:	skin pigmentation, mc1r, adaptive evolution, melanoma

Simultaneous purifying selection on the ancestral *MC1R* allele and positive selection on the melanoma-risk allele V60L in South Europeans

Martínez-Cadenas C[1]*, López S[2]*, Ribas G[3], Flores C[4,5], García O[6], Sevilla A[2], Smith-Zubiaga I[7], Ibarrola-Villaba M[3], Pino-Yanes M[4,5,8], Gardeazabal J[9], Boyano MD[10], García de Galdeano A[10], Izagirre N[2], de la Rúa C[2], Alonso S[2]

*both authors contributed equally to the realization of this work

Manuscript intended as Article (Discoveries)

Affiliations:

[1]Dept. Medicine, University of Castellón Jaume I. Avenida Sos Baynat 12071, Castellón, Spain. [2]Dept. Genetics, Physical Anthropology and Animal Physiology. University of the Basque Country UPV/EHU, Barrio Sarriena s/n. 48940 Leioa, Bizkaia, Spain. [3]INCLIVA Biomedical Research Institute, Valencia, Spain. [4]Research Unit, University Hospital N.S. de Candelaria, Carretera del Rosario s/n, 38010 Santa Cruz de Tenerife, Spain. [5]CIBER on Respiratory Diseases, ISCIII, Madrid, Spain. [6]Forensic Science Unit, Forensic Genetics Section, Basque Country Police, Erandio, Bizkaia, Spain. [7]Dept. Zoology and Animal Cell Biology, University of the Basque Country UPV/EHU, Barrio Sarriena s/n. 48940 Leioa, Bizkaia, Spain. [8]Department of Medicine, University of California, San Francisco, CA, USA. [9]Dermatology Service, Cruces Hospital, Plaza de Cruces s/n, 48903, Cruces-Barakaldo, Bizkaia, Spain. [10]Dept. Cell Biology and Histology. University of the Basque Country UPV/EHU, Barrio Sarriena s/n. 48940 Leioa, Bizkaia, Spain.

Corrresponding author

Santos Alonso. santos.alonso@ehu.es

Abstract

In humans, the geographical apportionment of the coding diversity of the pigmentary locus *MC1R* is, unusually, higher in Eurasians than in Africans. This atypical observation has been interpreted as the result of purifying selection due to functional constraint on *MC1R* in high UVB radiation environments. By analyzing 3,142 human *MC1R* alleles from different regions of Spain in the context of additional haplotypic information from the 1000 Genomes (1000G) Project data, we show that purifying selection is also strong in Southern Europe, but not so in Northern Europe. Furthermore, we show that purifying and positive selection act simultaneously on *MC1R*. Thus, at least in Spain, regions at opposite ends of the incident UV-B radiation distribution show significantly different frequencies for the melanoma-risk allele V60L (a mutation also associated to red hair and fair skin and even blonde hair), with higher frequency of V60L at those regions of lower incident UV-B radiation. Besides, using the 1000G South-European data, we show that the V60L haplogroup is also characterized by an EHH pattern indicative of positive selection. We, thus, provide evidence for an adaptive value of human skin depigmentation in Europe and illustrate how an adaptive process can simultaneously help maintain a disease-risk allele. In addition, our data support the hypothesis proposed by Jablonski and Chaplin (2010), which posits that habitation of middle latitudes involved the evolution of partially depigmented phenotypes that are still capable of suitable tanning.

Introduction

The human melanocortin-1 receptor (*MC1R*) is an integral membrane G-protein-coupled receptor (GPCR) of skin and hair follicle melanocytes. GPCRs communicate the cells with their environment and upon receiving external stimuli trigger an adaptive response. In this regard, the binding of a small peptide, the α-melanocyte stimulating hormone (α-MSH) to MC1R is coupled to a cAMP signaling pathway that stimulates eumelanogenesis (synthesis of brown/black melanins). In the absence of MC1R signaling, pheomelanogenesis (synthesis of yellow/red melanins) is the default pathway (Ito 2003). Therefore, *MC1R* is a regulator of the amount and type of pigment production, and has thus been referred to as a major determinant of skin phototype (García-Borrón, Sánchez-Laorden and Jimenez-Cervantes 2005). In this regard, some non-functional *MC1R* variants lead to phenotypes characterized by red hair, fair skin, freckles and poor tanning ability (the red hair and fair skin phenotype, RHC) in a dominant manner. These have been called ***R*** alleles, and include D84E, R151C, R160W, and D294H. Variants that do not lead to a total loss of function and that have weak or no association to the RHC phenotype (NRHC phenotype) are called ***r*** alleles, and include for instance V60L, V92M and R163Q (Valverde et al. 1995; Box et al. 1997; Sturm et al. 2003; Duffy et al. 2004; Beaumont et al. 2007; Dessinioti et al. 2011).

As skin phototype is a risk factor for melanoma (Abdel-Malek and Ito, 2013), considerable effort has been dedicated to find association between variants at *MC1R* (and other pigmentation genes) to melanoma risk (Valverde et al. 1996; Ichii-Jones et al. 1998; Palmer et al. 2000; Kennedy et al. 2001; Fernandez et al. 2007; Raimondi et al. 2008; Scherer et al 2009; Williams et a. 2011; Davies et al. 2012). Less effort has been dedicated to understand

pigmentation and melanoma risk from an evolutionary perspective. Interestingly, *MC1R* is highly diverse in humans: a recent literature review catalogued up to 57 non-synonymous and 25 synonymous variants in ~1kb of coding DNA (Gerstenblith et al., 2007). However, the geographical apportionment of this coding diversity shows that there is an increased diversity in Eurasian populations in comparison to African populations (Rana et al., 1999; Harding et al., 2000), which is an unusual observation, given that Africans typically exhibit greater genomic diversity than Eurasians. This is even more intriguing as the upstream region containing the promoter of *MC1R* does show the expected, opposite pattern, i.e. a higher diversity in Africans compared to Asians and Europeans (Makova et al. 2001).

These patterns of diversity in the coding region of *MC1R* led both Rana et al (1999) and Harding et al (2000) to posit that *MC1R* is under strong functional constraint in Africa. However, their interpretation for the diversity patterns in Eurasia differed. Thus, while Rana et al (1999) proposed local selection for mutations that result in lighter pigmentation (diversifying selection) in Eurasia, Harding et al (2000) proposed instead that the level of *MC1R* polymorphism simply reflects the expectations of a neutral evolution for this locus in Eurasians.

In this work, we propose an explanation for the high diversity at *MC1R* and assess the role of selection in one large Spanish population-sample in the European context. We conclude that purifying and positive selection can act simultaneously on the same locus. More interestingly, *MC1R* illustrates how an adaptive process can simultaneously help maintain a disease-risk allele.

Results

After resequencing *MC1R* in 1,217 healthy Spanish individuals, we detected 38 SNPs, 6 of which have not been previously reported: V193M, A285T, R229H, Q233X, S302N and W317C. We did not observe the R307G Neanderthal *MC1R* mutation described by Lalueza-Fox et al (2007), which adds support to their findings. We also found one new deletion, 520_522delGTC and one new insertion 539insC. However, for the subsequent diversity analyses, indel polymorphisms were not considered due to their different nature to SNPs. The most frequent variant was V60L, consistent with other studies on European populations (Flanagan et al., 2000; Fargnoli et al., 2006; Latreille et al., 2009). To this sample, we added those sequences previously published by some of us (Fernández et al., 2007; Ibarrola-Villava et al., 2012), which included 354 individuals from Madrid (Spain). The global sample showed a total of 42 SNPs, with a total of 42 different haplotypes observed (see Table 1 for a comparative diversity analysis).

We further resequenced *MC1R* in 127 melanoma patients and detected 15 SNPs. The most frequent variant was also V60L. To this sample, we added those melanoma sequences previously published by some of us (Fernández et al., 2007; Ibarrola-Villava et al., 2012), which corresponded to 595 individuals. The global melanoma sample showed then a total of 37 different haplotypes formed from 28 SNPs, 7 of which did not show up in the control samples (S41F, S41C, G89R, I180I, R213W, P268R and N279K).

In our healthy sample the most likely root is the most common haplotype. Its likelihood was ~2 orders of magnitude higher than the one corresponding to the next most likely rooted tree. The most common haplotype in our sample is also the most common haplotype in the 1000G African sample. Mutation T314T, which lies in a CpG doublet, was removed because it seems to have arisen more than once (within different lineages). Its frequency in the healthy

samples is 0.07 (total 2N = 3,142 alleles). Mutation R151C was also discarded because it seems to have arisen several times (within different lineages) in the melanoma samples (frequency in the healthy samples: 0.02).

Diversity analysis

The coding region of *MC1R* shows a very high nucleotide diversity for such a short coding region: we observed an estimated nucleotide diversity π of 0.08%, whereas in a small set of European human exomes the average π was 0.038% (Tennessen, Madeoy and Akey, 2010). This increased nucleotide diversity seems to be the result of a high mutation rate derived from a high CpG content. Figure 1 shows the CpG content of the coding region of *MC1R*. For this region, but not for the flanking regions, the observed vs. expected CpG ratio is higher than 0.65, which qualifies this region as a CpG island (Takai and Jones, 2002). This high CpG content is also correlated with the highest π, but the lowest Tajima´s D values.

Evidence for purifying selection at MC1R

There are many mutations at coding *MC1R*, but these do not reach in general a substantial frequency in our sample (see Supplementary Information 1). An exception could be V60L, an ***r*** allele with a frequency of ~0.15, which will be dealt with below. Neutrality tests indicate that the low frequency of new mutants can be due to the effect of purifying selection that tends to remove these new mutations from the population. Thus, Table 2 shows that Tajima's D and Fu and Li's D and F tests are highly significant even under models that incorporate a recent Out-of-Africa demographic history (Gutenkunst et al., 2009; Laval et al., 2010), whereas the combined DHEW test, which is specific of recent positive selection (due to the incorporation of the H test) is not. This is in agreement with the observation that

many of the polymorphisms at this region are non-synonymous and that the most common allele is also the ancestral one.

Using data from the 1000 Genomes Project (1000G) we observe that this scenario of purifying selection in our Spanish population could be shared with other Southern European countries (see Table 1). Thus, for instance, Tajima´s D test applied to the TSI sample shows a D value of -1.437 (p-value = 0.0438 under standard coalescent simulations). As expected, a similar pattern of purifying selection can be observed for the pooled 1000G African sample, for whom Tajima´s D is -1.913 (p-value = 0.001 under standard coalescent simulations), and also for the individual African populations (see Table 1). However, the pooled Northern European sample (or the North European populations individually; see Table 1) does not share this pattern (Tajima´s D = -0.870, p-value = 0.21 under standard coalescent simulations) The South European-African purifying selection pattern is not observed in the pooled East Asian sample either (or in any of the individual East Asian populations, see Table 1), which is characterized by the high frequency (~0.64) of R163Q, an ***r*** allele (Tajima´s D -= 0.393, p-value = 0.411 under standard coalescent simulations).

Evidence for positive selection at V60L in South Europeans

The strong pattern of purifying selection maintaining the ancestral *MC1R* allele contrasts with a substantial frequency of V60L in our sample. However, the relatively high frequency of V60L is not homogeneously distributed within Spain, as there seems to be some spatial heterogeneity in the frequency of this allele (Table 3; Supplementary Information 2). In order to explore if this heterogeneity in the frequency of the V60L allele is Spanish populations could be explained by a general underlying demographic process in the region, we then assessed the existence of population stratification by means of principal component

analyses based on data derived from 93 European ancestry informative markers (termed EuroAIMs), that are well-known to recover the largest features of the North-Northwest (NNW) to South-Southeast (SSE) axis of genetic differentiation in Europe (Price et al. 2008). In fact, analyzing 1,187 individuals from different European populations including the Spanish, EIGENSOFT found significant only the first principal component (PC) (Tracy-Widom test, p-value = 2.06×10^{-34}), differentiating NNW-Europeans from SSE-Europeans, and leaving the Spanish individuals at intermediate positions in the NNW-SSE axis of European differentiation. However, using this set of EuroAIMs, EIGENSOFT was unable to identify a significant PC (Tracy-Widom test, p-value = 0.646 for the first PC) when restricting the analysis to the Spanish populations (including all the Spanish provinces, the Spanish Basque and the Canary Islanders). To check if this heterogeneity could be reflecting the geographical UV-B irradiation pattern, we selected two population groups, one including those provinces for which UV-B irradiation values (see Figure 2) lie in the lower 5th percentile in the Iberian Peninsula (Group 1: Asturias, Bizkaia, Gipuzkoa; all of them Northern populations), an another one including those provinces above the 95th percentile of UVB in the Iberian Peninsula (the Canary Islands, Cádiz and Granada; all of them Southern populations) plus (Group 2), Fisher exact test shows that the frequency of V60L in Group 1 (0.17) is significantly higher than that in Group 2 (0.09) (1-tailed p–value = 0.042). This is confirmed if we consider also other percentiles (Table 3). Nevertheless, the distribution of the V60L allele frequency is far from closely following a strictly linear latitudinal pattern. Thus, when we group samples according to their UV-radiation percentile (at non-inclusive 0.05 quantiles from the lower and upper ends respectively), there is actually a negative correlation between the V60L allele frequency and UV radiation (Pearson's correlation coefficient = -0.165), but this is not significant (T-test p-value = 0.25). This suggests that other evolutionary

processes, including demographic history, may interfere. However, we also have to take into account the low sample sizes for many of the bins, which result in imprecise estimates of allele frequencies (see Supplementary Information 2).

Notwithstanding, the relative high frequency of V60L in the Spanish population could be indicative of some selective advantage. Therefore, we then investigated if recent positive selection could explain this pattern. For that, we applied the EHH test to the 1000G Southern European sample, the best proxy available, as the TSI population has similar latitude to North Spain. Figure 3 shows that the extent of homozygosity from V60L is in fact longer than expected under the neutral demographic models considered. It is also longer than that observed for the 1000G North European sample set. See also Supplementary Information 3 for EHH analysis in single populations. We performed an additional test of selection based on a haplogroup-specific Tajima's D test on a region of 10kb 5' of rs1805005 for the set of European rs1805005-T (V60L) haplotypes (see Material and Methods). Tajima's D for the rs1805005-T haplogroup in this population was -2.045, whereas for the alternative rs1805005-G haplogroup Tajima's D was 1.693. The p-value under the Gutenkunst et al. (2009) demographic model for the T haplogroup is 0.021 (0.032 under the Laval et al. (2010) model). This indicates that positive selection can be held responsible for the substantial frequency of V60L at least in some areas of Europe.

The collateral damage of positive selection

It is interesting to note that individuals with melanoma show a significantly higher frequency of V60L than their corresponding controls (Fernández et al., 2007; Ibarrola-Villava et al., 2012; see also http://www.melgene.org/). We have improved this data set with new melanoma cases and compared it with our Spanish reference set as the control sample.

Cochran-Armitage test also shows an association between V60L and melanoma risk in an additive model (Table 4) only when considering as controls the set of populations belonging to the provinces with highest UV-B irradiation (above the 95th, 90th, 85th or 80th percentiles). This suggests that the adaptive evolution of skin depigmentation is concurrently responsible for an increased health risk in, most typically, post-reproductive age.

On the other hand, there were no significant differences when using as controls either the set of populations with the lowest UV-B radiation or the whole set of Spanish provinces. This is illustrative of how population structure in allele frequencies can be an issue in genetic association tests, which can lead to false negatives if not properly accounted for.

Discussion

We have shown that *MC1R* shows high coding nucleotide diversity in the (South European) population analyzed. We have proposed that methylation-mediated deamination of 5-methylcytosine is ultimately responsible for the high nucleotide diversity observed at *MC1R* The following observations add support to the hypothesis that: a) CpGs within genes, contrary to regulatory CpG islands, tend to be highly methylated (Jones 1999), b) methylome analysis of H1 human embryonic stem cells and IMR90 fetal lung fibroblasts shows that *MC1R* is highly methylated (Lister et al. 2009) (http://neomorph.salk.edu/human_methylome/browser.html), c) our own bisulfite sequencing data indicates that *MC1R* is methylated in sperm, d) in mammalian genomes, CpG sites are hotspots for transition mutations because of methylation-mediated deamination of 5-methylcytosine (Cooper and Krawczak 1993), e) 88% (44 out of 50) of the SNPs observed in our sample (considering both controls and melanomas) are transitions, and

f) CpG sites show a mutation rate 10 times higher than non-CpGs (Nachman and Crowell, 2000).

This high variability is however shaped by contrasting evolutionary forces. It has been argued that environmental pressure is responsible for maintaining the ancestral *MC1R* allele at high frequency in Africa, whereas in Eurasia instead, this functional constraint is absent (Rana et al., 1999; Harding et al., 2000). We, however, given the conspicuous pattern of purifying selection detected in our sample, show that strong functional constraint on the ancestral *MC1R* haplotype also takes place at least in some areas of Southern Europe. This observation might subsequently lead us to think that *MC1R* has played no role in the adaptive lightening of the human skin in Eurasia. Actually, *MC1R* variability has been found not to contribute to light pigmentation in mice from the Florida's Atlantic Coast (Steiner et al. 2009), although it had been highlighted as an adaptive locus in beach mice from Florida´s Gulf Coast (Hoekstra et al. 2006). Similarly, in non-human primates, *MC1R* variability does not seem either to show a simple relationship to coat color (Mundy and Kelly, 2003; Haitina et al 2007). Besides, the presence of a functional, MSH-responsive *MC1R* in orangutan shows that alternative mechanisms of coat color generation can exist (Haitina et al., 2007).

However, our results indicate that two selective forces shape human *MC1R* diversity in Sothern Europe. Similarly to Africa, purifying selection is strong maintaining the ancestral allele in Sothern Europe. But simultaneously, positive selection is responsible for an increased frequency of the V60L allele, a low-penetrance "red hair and fair skin" allele (also associated to blond hair; Box et al. 1997). The fact that V60L is under positive selection constitutes evidence that *MC1R* is also playing an active adaptive role in human skin depigmentation, at least for some areas of South Europe. Actually, the world frequency

distribution of V60L shows that this allele is circumscribed to Europe and the Near East (Gerstenblith et al., 2007; Acar, Bozkurt and Görgülü, 2012; Galore-Haskel et al., 2009; Eskandani et al., 2010). However, we show that this increase in fitness for the population has collateral damaging consequences for the individual´s health. In this regard, the same variant (V60L) that is under positive selection is also a melanoma-risk allele. As melanoma is typically a post-reproductive disease (Sáenz, Conejo-Mir and Cayuela, 2005), we can consider that we are postponing the payment due to our adaptation to a reduced-sunlight environment. Does this mean that melanoma, or more generally, skin cancer risk has no fitness consequences itself? We cannot answer that question with our data, because although the ancestral *MC1R* allele is under strong purifying selection (in our sample, but also in more demanding regions like Africa), we would also need to assess other protective roles that have been proposed for dark skin, most notably photoprotection of circulating folate (Jablonski and Chaplin, 2010). However, the role of *MC1R* in DNA repair and apoptosis suggests that *MC1R* is not simply encoding a passive filter (melanin) (García-Borrón, Sánchez-Laorden and Jimenez-Cervantes, 2005). In any case, whatever the fitness forces at play, the final fitness balance seems to favor depigmentation in Europe.

On the other hand, despite proving evidence that depigmentation can be adaptive in relation to scarce sunlight, we however, cannot explicitly say which is the underlying physiological mechanism involved. The vitamin D hypothesis (Loomis, 1967) would constitute a straightforward connection. Nevertheless, there is still a gap that needs to be further explored in order for the above hypothesis to be proven.

Finally, it also remains to be known why other **R/r** mutations are not so common in our population. One hypothesis is that **r** mutations result in a lighter phenotype but in a way that

does not compromise tanning during seasonal increases in UVB, thus enabling protection against skin cancer (Robins 1972) or against folate destruction (Jablonski and Chaplin, 2010). In this regard, V60L has been shown to achieve significantly lower cAMP levels after agonist stimulation than the wild type MC1R protein, which indicates substantial impaired function, similar to that of the **R** variants (Schiöth et al, 1999; Beaumont et al. 2007; Beaumont, Liu, Sturm 2009). This result is supported by genetic data, as the skin color of individuals homozygous for V60L is significantly fairer than wild-type individuals, to the same extent, for instance, as homozygotes for the **R** allele R160W (Beaumont et al, 2007). However, in heterodimers with the wild-type receptor, V60L does not seem to show (but most **R** mutants do) dominant-negative effect on wild-type receptor's cell-surface expression. A dominant-negative effect would result in a decreased ability of the wild-type receptor (the ancestral allele) to elevate intracellular cAMP levels (Beaumont et al, 2007; Beaumont, Liu, Sturm 2009), and would add to the decreased ability of the V60L allele itself. Thus, generally speaking, V60L would preserve the ability to increase eumelanin production (tanning) during seasonal increases in UVB better than the **R** alleles. This would be useful particularly in Southern European regions, where summer UV-B radiation is higher than in Northern Europe. Thus, in conclusion, our results would support the hypothesis proposed by Jablonski and Chaplin (2010) that "habitation of middle latitudes between approximately 23° and 46° involved the evolution of partially depigmented phenotypes capable of tanning".

Material and Methods

Population samples sequenced

We analyzed 3,142 human *MC1R* alleles (1,571 individuals) from different regions of Spain through direct DNA sequencing of the coding region. Of the 1,571 individuals, 962 came

from the Spanish National DNA Bank (http://www.bancoadn.org/en/home.htm) and represent different regions in Spain. Of these, 670 individuals were known to have all grandparents from a single specific province of Spain. In addition, we sequenced a further sample of 172 alleles from the Basque Country (86 individuals), 60 alleles (30 individuals) from the Canary Islands (a subset of those described in de Pino-Yanes et al., 2011, selected based on the lower ancestry to Northern Africa as deduced from 93 EuroAIMs), and 278 alleles (139 individuals) from the province of Castellón. Sequences have been deposited in GenBank with accession numbers KC981760 - KC984193. Finally, 708 already sequenced *MC1R* alleles (354 individuals) from Madrid were recruited from previously published work (Fernández et al., 2007; Ibarrola-Villava et al., 2012).

Data compiled from other sources

We obtained full *MC1R* sequence information as strings of SNP genotypes from the 1000G (http://www.1000genomes.org/) by means of **SPSmart v. 5.1.1 (**http://spsmart.cesga.es/, Amigo et al. 2008) which uses the information 1000 Genomes Phase I data (data freeze as of May 2011) for the following populations:

European populations: 268 individuals from Northern Europe, comprising Utah residents with N. and W. European ancestry from the CEPH collection (CEU, N=87), Finnish from Finland (FIN, N=93) and British from England and Scotland (GBR, N=88); and 112 individuals from Southern Europe, comprising Tuscans from Italy (TSI, N=98) and Iberian Spanish (IBS, N=14).

African populations: 246 individuals of African ancestry from Southwest USA (ASW, N=61), Luhya in Webuye – Kenya (LWK, N=97), and Yoruba in Ibadan – Nigeria (YRI, N=88).

East Asian populations: 268 individuals from China (CHB), Han Chinese South (CHS), and Japanese in Tokyo – Japan (JPT)

Resequencing

All DNA samples were analyzed for the coding sequence of the *MC1R* gene by direct automated DNA sequencing. PCR primer sequences used to frame a 1,150-bp fragment containing the entire sequence of the *MC1R* gene single exon were: 5'- CAG CAC CAT GAA CTA AGC AGG ACA CCT -3' (sense) and 5'- AAG GGT CCG CGC TTC AAC ACT TTC AGA G -3' (antisense). PCR reactions were carried out using 100 ng of DNA as template, and the PCR reaction mixture also included 2µl of 10X PCR buffer, 25 mM $MgCl_2$, 5 µM each of dGTP, dATP, dCTP and dTTP, 25 pmol of each PCR primer, and 1 unit of AmpliTaq Gold DNA polymerase (Life Technologies). Thermocycling conditions included a denaturation step at 95ºC for 8 minutes, followed by 35 cycles of 95ºC for 30 seconds, 56ºC for 45 seconds and 72ºC for 1 minute, plus a final extension cycle of 72ºC for 10 minutes. Amplified products were purified using ExoSAP-IT (USB Corporation) according to the manufacturer's instructions. In order to obtain both forward and reverse sequences of the entire *MC1R* exon, purified PCR products were subjected to four sequencing reactions using forward and reverse sequencing primers: MC1R-F1 5'- CAG CAC CAT GAA CTA AGC AGG ACA CCT -3'; MC1R-F2 5'- TGG GTG GCC AGT GTC GTC TTC AGC A -3'; MC1R-R1 5'- AAG GGT CCG CGC TTC AAC ACT TTC AGA G -3'; and MC1R-R2 5'- CCA GCA TAG CCA GGA AGA AGA CCA CGA -3'. Gaps were filled using primers, MC1R-F3-5'-ATGCCAGGAGGTGTCTGGACTG-3', MC1R-F4-5'-GATCACCTGCAGCTCCATGCTG-3', MC1R-R3 5'-AAGGGTCCGCGCTTCAACACTTTCAGAG-3', MC1R-R4 5'-CAACACTTTCAGAGATCATTTAGTCCATC-3', and MC1R-R5 5'-CAGCACGGCCATGAGCACCAG-3'

Sequencing reactions were performed using 10 ng of PCR purified product (MicroSpin S-400 HR Columns, GE Healthcare) and BigDye Terminator Cycle Sequencing kits (Life Technologies). Sequencing reaction products were then washed with Performa DTR V3 96-well plates (EdgeBio), and they were finally run on an ABI 3130 Genetic Analyzer (Life Technologies). Sequences were edited with Genalys v2.8 (Takahashi et al. 2003) and further checked manually afterwards.

Bisulfite methylation analysis

Methylation was assessed using a bisulfate treatment protocol based on Kaneda et al. (2004). In brief, 3 micrograms of DNA from sperm were digested with *Hinf*I and incubated with 6N NaOH for 15 min at 37ºC. To this solution we added 120 microlitres of the conversion solution: sodium bisulfite 4.04M (Sigma), Hydroquinone 10mM (Sigma) and NaOH 6N (pH 5), and performed 20 cycles of denaturing for 30 sec at 95ºC and incubation for 15 min at 50ºC. DNA was then purified using the QIAquick PCR purification kit (Qiagen), desulfonated by incubation with NaOH 6N for 5 min at room temperature, and precipitated with ethanol 100% and ammonium acetate 6M (pH 7) (Sigma).

Converted DNA was amplified by PCR with the following primers: Fw 5'-TGTTGGTTCGGGTTTGTTA-3' and Rev 5'-ACCTCCTTAAACGTCCTACG-3', which were designed to anneal specifically to the bisulfite converted sequence (considering methylation in CpG). PCR conditions were the following: 3 min at 96ºC; 45 cycles of 30 seconds at 96ºC, 20 seconds at 59ºC and 1 min at 70ºC; and a final step of 10 min at 70ºC.

Sequencing reactions were performed using the BigDye Terminator kit 3.1 (Life Technologies) and run on an ABI 310 using the same primers that were used for the PCR.

Probabilistic and experimental phasing

Haplotypes were estimated with PHASE (downloadable from http://stephenslab.uchicago.edu/software.html#phase). For those pairs of SNPs whose phase could not be statistically inferred with a p-value greater than 0.9, we proceeded to infer their phase experimentally by cloning (TOPO-TA kit, Invitrogen) and resequencing again.

Population Genetics analysis of the MC1R sequences

Population diversity and neutrality tests parameters were obtained using **DnaSP 5.10** (Librado and Rozas, 2009) (http://www.ub.edu/dnasp). For the neutrality tests, p-values were initially obtained from simulations run under a neutral, constant-sized population model using the program **ms** (Hudson 2002). Twenty thousand simulations were obtained fixing on θ_w. To take into consideration the possible effect of demography on the neutrality tests, we carried out additional simulations incorporating optimized parameters that attempt to reflect the human evolutionary history. These parameters were obtained from the models of Gutenkunst et al. (2009) and that of Laval et al. (2010). The **ms** command lines for the Gutenkunst et al. (2009) model were:

./ms 3142 20000 -t 1.018 -I 3 0 3142 0 -n 1 1.68202 -n 2 3.73683 -n 3 7.29205 -g 2 116.010723 -g 3 160.246047 -ma x 0.881098 0.561966 0.881098 x 2.79746 0.561966 2.79746 x -ej 0.028985 3 2 -en 0.028985 2 0.287184 -ema 0.028985 3 x 7.29314 x 7.29314 x x x x x -ej 0.197963 2 1 -en 0.303501 1 1

and for the Laval et al. (2010) parameters were:

./ms 3142 20000 -t 1.922 -I 3 0 3142 0 -n 1 1666.667 -n 2 2.261 -n 3 1.051 -g 1 372.278 -g 2 380.217 -g 3 259.366 -ma x 0.7176 0.7176 0.7176 x 0.7176 0.7176 0.7176 x -ej 0.0163 3 2 -en 0.0435 2 0.2029 -ej 0.0435 2 1

Using these simulations, the p-values for the Tajima´s D, Fay and Wu's H (normalized) and Fu and Li's D and F tests were obtained using the software **msstats** (K. Thornton's lab, http://www.molpopgen.org/software/msstats/). The DHEW combined test, which includes Tajima´s D , Fay and Wu´s H (normalized) and the Ewens and Waterson (EW) test were performed with software kindly provided by K. Zeng and modified by us to allow for larger sample sizes. This compound test is more robust against the presence of recombination, is insensitive to background selection and demography and it is thus more powerful in detecting positive selection (Zeng, Shi and Wu, 2007).

The genealogical relationships among *MC1R* haplotypes were inferred by means of **Network 4.6** (http://www.fluxus-engineering.com). We used **genetree** v9.0 (by RC Griffiths, downloadable at http://www.stats.ox.ac.uk/~griff/software.html) to infer the root of the tree.

Structure analysis

In order to explore the genetic structure in the Spanish populations, data from 93 EuroAIMs recovering the largest features of the European north-northwest to south-southeast axis of differentiation (Price et al. 2008), from a total of 1,187 individuals, was utilized to assess a principal component analysis by means of EIGENSOFT 4.2 (Price et al. 2006). Data finally included in this analysis was derived from individuals known to have all grandparents from a single specific province of Spain, and from reference populations, satisfying a genotyping

completion rate >90%. We restricted the genotyping and analyses to those 93 SNPs that overlapped with samples used as reference (see Pino-Yanes et al. [2011] for a full description of methods). Briefly, genotyping was conducted utilizing the iPLEX™ Gold assay on MassARRAY system (Sequenom, San Diego, CA) by the Spanish National Genotyping Center, Santiago de Compostela Node (CeGen, http://www.cegen.org), and was performed in:

-Samples from 553 individuals from 14 Spanish provinces from the Spanish National DNA Bank (www.bancoadn.org): Andalusia (n=105), Murcia (n=27), Extremadura (n=25), Castile-La Mancha (n=50), Valencia (n=38), Castile and Leon (n=93), Madrid (n=6), Catalonia (n=55), Galicia (n=69), Cantabria (n=18), Navarre (n=16), La Rioja (n=21), Asturias (n=14), and Aragon (n=16).

-Samples from 24 Spanish Basques satisfying that all their four parental surnames were of Basque origin.

In addition, data from 80 Canary Islanders taken from previous studies (Pino-Yanes et al. 2011), selected based on the lower ancestry to Northern Africa, were used for comparisons, along with data from 144 Swedish, 52 Polish, 69 English, 117 Italians, 64 Greeks, and 28 French from Price et al. (2008), as well as from 56 Utah residents with ancestry from northern and western Europe (CEU), all of them also satisfying a genotyping completion rate >90%.

Extended Haplotype Homozygosity (EHH) analysis of the 1000G data

To assess the possibility of positive selection acting on single alleles we used our own implementation of the EHH method (Sabeti et al., 2002), which uses simulations obtained by

means of Hudson´s **ms** program (Hudson 2002) to estimate the p-values. This process was implemented in the following way:

We selected two 100kb regions, one on each side of the SNP we wanted to test for recent positive selection, and for each region we estimated divergence to the orthologous region in the chimpanzee. From divergence (assuming a 6 million year divergence time and the Jukes and Cantor model, Li 1997) we estimated the mutation rate μ, which in its turn was used to estimate the parameter θ for both the Gutenkunst et al. (2009) and Laval et al (2010) models. The recombination parameter ρ was estimated from the genome-wide recombination rate from Phase 2 HapMap estimated from phased haplotypes in HapMap Release 22 (NCBI 36) (The International HapMap Consortium, 2007). In each simulation, the specific θ and ρ values were randomly sampled from a normal distribution with a standard deviation equal to the mean. The demographic models used for the simulations were the above-mentioned of Gutenkunst et al. (2009) and Laval et al. (2010). For each region, the first SNP in the set of simulations was assumed to represent the SNP of interest. Only those simulations for which the frequency pattern of this SNP, taking into account the ancestral or derived state, was equal (±10% allowance) to the observed values in the 1000G European population, were considered for p-value estimations. In addition, only those simulations showing a total number of SNPs per 100kb equal to the observed value (±10% allowance), were considered for EHH estimations. EHH profiles from the set of simulated haplotypes carrying the (derived) allele under consideration (at the first SNP) were then used to infer p-values. Typically, for each region several tens of thousands simulations were needed to obtain a set of 400-500 simulations satisfying these conditions. After this, the EHH values at points distant from the first SNP 0.1, 0.2, 0.3, 0.4 and 0.5 units (in a scale of 0 to 1, being 1

equivalent to 100kb) were recorded and the individual simulations showing the EHH value at the 95th percentile at each point were selected. All simulation selected were lumped together. Next, duplicated points were removed, and then we forced all points to satisfy the condition $EHH_{i+1} \leq EHH_i$ (where *i* increases from SNP number 1 in any direction). The resulting distribution of EHH values was used as the 95% cut-off line.

Tajima's D test for specific haplogroups

As an additional test for positive selection on V60L, we carried out Tajima's D test on that subset of haplotypes defined by the derived state T at rs1805005 (V60L), using the information from the 1000G project. We initially tried to carry out this test on the South European 1000G samples, but within the haplogroup defined by allele rs1805005-T there is no variability at least 10kb on both sides of rs1805005.Thus, we decided to test the whole 1000G European sample (GBR, CEU, FIN, TSI and IBS pooled). Note that it is unlikely that pooling will result in more extreme negative Tajima's D values, so this is *a priori*, a conservative approach. Besides, there is some evidence (although weaker) from the EHH test that the V60L allele is also under some positive selection in the 1000G North European populations (see figure 3). Note that evidence for selection on the pooled 1000G European population is only apparent when we analyze the internal diversity of the rs1805005-T haplogroup and not when we analyze the global diversity, i.e., including both rs1805005-T and -G haplogroups.

As regards the region analyzed, we decide to analyze a region comprising 10kb of the *upstream* flanking region to V60L. Using just one flanking region of V60L in which one SNP at the beginning or the end is the V60L SNP is more efficient when calculating the p-values (see below). Even after pooling 1000G data from North and South European samples, there is just

one polymorphism within the T haplogroup in the *downstream* 10kb region to V60L, which does not give us power to perform Tajima's D test. However, there are 6 SNPs in the upstream 10kb region within the T haplogroup.

To estimate the p-values for the rs1805005-T haplogroup we proceeded as follows:

a) First, run coalescent simulations using parameters from the whole European population:

For the Gutenkunst et al. (2009) model:

./ms 760 690 -t tbs -r tbs 10000 -I 3 0 760 0 -n 1 1.68202 -n 2 3.73683 -n 3 7.29205 -g 2 116.010723 -g 3 160.246047 -ma x 0.881098 0.561966 0.881098 x 2.79746 0.561966 2.79746 x -ej 0.028985 3 2 -en 0.028985 2 0.287184 -ema 0.028985 3 x 7.29314 x 7.29314 x x x x x -ej 0.197963 2 1 -en 0.303501 1 1

where the –t (theta) option is a random number from a normal distribution with mean 9.75 and variance equal to the mean, and the–r (recombination parameter) option is a random number from a normal distribution with mean 3.25 and variance equal to the mean.

And for the Laval et al (2010) model:

./ms 760 252 -t tbs -r tbs 10000 -I 3 0 760 0 -n 1 1666.667 -n 2 2.261 -n 3 1.051 -g 1 372.278 -g 2 380.217 -g 3 259.366 -ma x 0.7176 0.7176 0.7176 x 0.7176 0.7176 0.7176 x -eg 0.0063 2 0 -eg 0.0063 3 0 -ej 0.0163 3 2 -eg 0.0199 1 0 -ej 0.0435 2 1 -en 0.0435 2 0.2029

where the –t (theta) option is a random number from a normal distribution with mean 18.4 and variance equal to the mean, and the–r (recombination parameter) option is a random number from a normal distribution with mean 6.14 and variance equal to the mean.

b) Then, for each simulation, we check that at a certain "predefined position" allele frequencies are close (allowance of 5%) to those observed at rs1805005 at that population. This predefined position is normally the first or the last SNP, because this facilitates steps b), c) and d).Similarly, we check that the number of segregating sites is close to the observed value (allowance of 10%) for the whole sample. If these conditions are satisfied, we keep that simulation, otherwise we discard it.

c) Then, for each accepted simulation we extract the set of haplotypes that are of a desired allelic state at the predefined position (0=ancestral, 1=derived), emulating the rs1805005-T (derived) allele in that population.

d) We check that the number of segregating sites within this extracted set of haplotypes is "similar" (allowance 10%) to that observed for that haplogroup in the 1000G population. For 10kb upstream rs1905005-T alleles 6 segregating sites are observed, so the allowed range was between 5 and 7 segregating sites. Note that several million simulations are needed to end up with 100-200 simulations satisfying our conditions. For each accepted simulation, we calculate its corresponding Tajima's D value (for the specified set of haplotypes forming a haplogroup). The set of the so-calculated Tajima's D values forms our distribution from which to infer the p-value for the observed Tajima D of the rs1805005-T haplogroup of the 1000G European sample.

Association analysis

To detect the association of some *MC1R* variants to melanoma susceptibility we sequenced *MC1R* from 127 DNA samples from melanoma patients attending the outpatient clinic at the Castellón Province Hospital (Castellón, Spain) for a 5-year period, between 2007 and 2011. In

addition, already published *MC1R* sequence data on 595 melanoma patients from Madrid were also recruited (Fernández et al., 2007; Ibarrola-Villava et al., 2012). Both sets of samples are registered at the National Health Institute Carlos III with reference numbers C.0000274 and C.0000293.

For the association analyses, we used the R libraries *MaXact* and *HardyWeinberg* (http://cran.r-project.org/web/packages/).

UV radiation maps

We elaborated a UV radiation map of Spain to correlate the daily annual average of incident UV-B radiation (J/m^2) with *MC1R* haplotype frequencies. Surface UV-B data (from 1995-2010, although not for all the provinces) were supplied to us by AEMET (the Spanish National Meteorological Agency), who used 16 meteorological stations spread over the Spanish territory. We also obtained data on the satellite-inferred global solar irradiation (annual mean; irradiation in the range from 0.2 -4μm) per province from the period 1983-2005, obtained from CM-SAF (Climate Satellite Application Facilities), EUMETSAT, and elaborated by AEMET (http://www.aemet.es/documentos/es/serviciosclimaticos/datosclimatologicos/atlas_radiacion_solar/atlas_de_radiacion_24042012.pdf). Because UV-B data and global data across the Iberian Peninsula (i.e. excluding the Canary Islands) correlated well (Pearson´s r > 0.9) for those provinces lacking a meteorological station recording UV-B data, UV-B values were inferred by linear regression from the global irradiation.

Irradiation maps were elaborated with QGIS v 1.8.0 (http://www.qgis.org/)

Acknowledgements

This works was supported by the former Spanish Ministerio de Ciencia e Innovación, project CGL2008-04066/BOS to S.A.; by the Dpt. Educacion, Universidades e Investigación of the Basque Government, project IT542-10; by program UFI11/09 by the University of the Basque Country, by "Programa de Investigacion Cientifica de la Universidad de La Laguna" (boc-a-2010-255-7177), and by grants from the Health Institute "Carlos III" (FIS PI08/1383, FIS PI11/00623) to C.F. and co-financed by the European Regional Development Funds, "A way of making Europe" from the European Union. M.P.Y. was supported by a postdoctoral fellowship from Fundación Ramón Areces. We thank the Spanish Banco Nacional de ADN (BNADN) (http://www.bancoadn.org/) for providing us with DNA samples from all over Spain. We also thank the Spanish Agencia Estatal de Meteorología (AEMET) (http://www.aemet.es/) for kindly providing us with the UV-B radiation data.

Literature

Abdel-Malek ZA, Ito S. 2013. Being in the red: a no-win situation with melanoma. Pigment Cell Mel Res. 26:164–166.

Acar E, Bozkurt G, Görgülü A. 2012. Polymorphisms in the Melanocortin-1 Receptor (Mc1R) Gene in Vitiligo. Balkan Med J. 29: 26-31.

Amigo J, Salas A, Phillips C, Carracedo A (2008) SPSmart: adapting population based SNP genotype databases for fast and comprehensive web access. BMC Bioinformatics 9:428.

Beaumont KA, Liu YY, Sturm RA. 2009. The melanocortin-1 receptor gene polymorphism and association with human skin cancer. In: Progress in molecular biology and translational science Vol. 88. G Protein-coupled receptors in health and disease (A). Tao YX Ed. London. Academic Press.

Beaumont KA, Shekar SL, Newton RA, James MR, Stow JL, Duffy DL, Sturm RA. 2007. Receptor function, dominant negative activity and phenotype correlations for MC1R variant alleles. Hum Mol Genet. 16:2249-2260.

Box NF, Wyeth JR, O'Gorman LE, Martin NG, Sturm RA. 1997. Characterization of melanocyte stimulating hormone receptor variant alleles in twins with red hair. Hum Mol Genet. 6:1891–1897.

Cooper DN, Krawczak M. 1993. Human Gene Mutation. Oxford. BIOS Scientific Publishers.

Davies JR, Randerson-Moor J, Kukalizch K, Harland M, Kumar R, Madhusudan S, Nagore E, Hansson J, Höiom V, et al. (30 co-authors). 2012. Inherited variants in the MC1R gene

and survival from cutaneous melanoma: a BioGenoMEL study. Pigment Cell Mel Res. 25:384-394.

Dessinioti C, Antoniou C, Katsambas A, Stratigos AK. 2011. Melanocortin 1 receptor variants: functional role and pigmentary associations. Photochem Photobiol. 87:978-987.

Duffy DL, Box NF, Chen W, Palmer JS, Montgomery GW, James MR,Hayward NK, Martin NG, Sturm RA. 2004. Interactive effects of MC1Rand OCA2 on melanoma risk phenotypes. Hum Mol Genet. 13:447–461.

Eskandani M, Hasannia S, Vandghanooni S, Pirooznia N, Golchai J. 2010. Assessment of MC1R and α-MSH gene sequences in Iranian vitiligo patients. Indian J Dermatol. 55:325–328.

Fargnoli MC, Altobelli E, Keller G, Chimenti S, Höfler H, Peris K. 2006. Contribution of melanocortin-1 receptor gene variants to sporadic cutaneous melanoma risk in a population in central Italy: a case-control study. Melanoma Res. 16:175-182.

Fernandez L, Milne R, Bravo J, Lopez J, Aviles J, Longo M, Benitez J, Lazaro P, Ribas G. 2007. MC1R: three novel variants identified in a malignant melanoma association study in the Spanish population Carcinogenesis 28:1659–1664.

Flanagan N, Healy E, Ray A, Philips S, Todd C, Jackson IJ, Birch-Machin MA, Rees JL. 2000. Pleiotropic effects of the melanocortin 1 receptor (MC1R) gene on human pigmentation. Hum Mol Genet. 9:2531-2537.

Galore-Haskel G, Azizi E, Mohamdi H, Scope A, Chaudruc V, Laitman Y, Barak F, Pavlotsky F, Demenais F, Friedman E. 2009. MC1R variant alleles and malignant melanoma risk in Israel. Eur J Cancer 4 5:2015-2022.

Garcia-Borron JC, Sanchez-Laorden BL, Jimenez-Cervantes C. 2005. Melanocortin-1 receptor structure and functional regulation. Pigment Cell Mel Res. 18:393–410.

Gerstenblith MR, Goldstein AM, Fargnoli MC, Persi K, Landi MT. 2007. Comprehensive evaluation of allele frequency differences of MC1R variants across populations. Hum Mutat. 28:495-505.

Gutenkunst RY, Hernandez RD, Williamson SH and Bustamante CD. 2009. Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. PLoS Genetics 5:e1000695.

Haitina T, Ringholm A, Kelly J Mundy NI and Helgi B. Schiöth HB. 2007. High Diversity in Functional Properties of Melanocortin 1 Receptor (MC1R) in Divergent Primate Species Is More Strongly Associated with Phylogeny than Coat Color. Mol Biol Evol. 24:2001-2008.

Harding RM, Healy E, Ray AJ, Ellis NS, Flanagan N, Todd C, Dixon C,Sajantila A, Jackson IJ, Birch-Machin MA, et al. (11 co authors). 2000. Evidence forvariable selective pressures at MC1R. Am J Hum Genet. 66:1351–1361.

Hoekstra HE, Hirschmann RJ, Bundey RA, Insel PA, Crossland JP (2006) A single amino acid mutation contributes to adaptive beach mouse color pattern. Science 313:101-104.

Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18:337-338.

Ibarrola-Villava M, Martin-Gonzalez M, Lazaro P, Pizarro A,Lluch A, Ribas G. 2012. Role of glutathione S-transferases in melanoma susceptibility: association with GSTP1 rs1695 polymorphism. Br J Dermatol. 166:1176–1183.

Ichii-Jones F, Lear JT, Heagerty AH, Smith AG, Hutchinson PE, Osborne J, Bowers B, Jones PW, Davies E, Ollier WE, et al. (15 co-authors). 1998. Susceptibility to melanoma: influence of skin type and polymorphism in the melanocyte stimulating hormone receptor gene. J Invest Dermatol. 111:218-221.

Ito S. 2003. A Chemist's View of Melanogenesis. Pigment Cell Mel Res. 16:230–236.

Jablonski NG, Chaplin G. 2010. Human skin pigmentation as an adaptation to UVB radiation. Proc Natl Acad Sci USA. 107:8962-8968.

Jones PA. 1999. The DNA methylation paradox. Trends Genet. 15:34–37.

Kaneda A, Tsukamoto T, Takamura-Enya T, Watanabe N, Kaminishi M, Sugimura T, Tatematsu M, Ushijima T. 2004. Frequent hypomethylation in multiple promoter CpG islands is associated with global hypomethylation, but not with frequent promoter hypermethylation. Cancer Sci. 95:58-64.

Kennedy C, ter Huurne J, Berkhout M, Gruis N, Bastiaens M, Bergman W, Willemze R, Bavinck JN. Melanocortin 1 receptor (MC1R) gene variants are associated with an increased risk for cutaneous melanoma which is largely independent of skin type and hair color (2001) J Invest Dermatol. 117:294–300.

Lalueza-Fox C, Römpler H, Caramelli D, Stäubert C, Catalano G, Hughes D, Rohland N, Pilli E, Longo L, Condemi S, et al (17 co-authors). 2007. A Melanocortin 1 Receptor Allele Suggests Varying Pigmentation Among Neanderthals. Science 318:1453-1455.

Latreille J, Ezzedine K, Elfakir A, Ambroisine L, Gardinier S, Galan P, Hercberg S, Gruber F, Rees J, Tschachler E, et al. (11 co-authors). 2009. MC1R gene polymorphism affects skin color and phenotypic features related to sun sensitivity in a population of French adult women. Photochem Photobiol. 85:1451-1458.

Laval G, Patin E, Barreiro LB, Quintana-Murci L (2010) Formulating a Historical and Demographic Model of Recent Human Evolution Based on Resequencing Data from Noncoding Regions. PLoS One 5:e10284.

Li WH. 1997. Molecular Evolution. Sunderland (MA). Sinauer Associates.

Librado P and Rozas J (2009) DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25**:** 1451-1452.

Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM, et al . (18-coauthors). 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 462:315-322.

Loomis WF. 1967. Skin-pigment regulation of vitamin-D biosynthesis in Man. Science 157:501-506.

Makova KD, Ramsay M, Jenkins T, Li WH. 2001. Human DNA sequence variation in a 6.6-kb region containing the melanocortin 1 receptor promoter. Genetics 158:1253-1268.

Mundy NI, Kelly J. 2003. Evolution of a pigmentation gene, the melanocortin-1 receptor, in primates. Am J Phys Anthropol. 121:67–80.

Nachman MW, Crowell SL. 2000. Estimate of the Mutation Rate per Nucleotide in Humans. Genetics 156:297-304.

Palmer JS, Duffy DL, Box NF, Aitken JF, O'Gorman LE, Green AC, Hayward NK, Martin NG, Sturm RA. 2000. Melanocortin-1 receptor polymorphisms and risk of melanoma: is the association explained solely by pigmentation phenotype? Am J Hum Genet. 66:176–186.

Pino-Yanes M, Corrales A, Basaldúa S, Hernández A, Guerra L, Villar J, Flores C. 2011. North African influences and potential bias in case-control association studies in the Spanish population. PLoS One 6:e18389.

Price AL, Butler J, Patterson N, Capelli C, Pascali VL, Scarnicci F, Ruiz-Linares A, Groop L, Saetta AA, Korkolopoulou P, et al. (20 co-authors). 2008. Discerning the ancestry of European Americans in genetic association studies. PLoS Genet 4: e236.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38: 904-909.

Raimondi S, Sera F, Gandini S, Iodice S, Caini S, Maisonneuve, Fargnoli MC. 2008. MC1R variants, melanoma and red hair color phenotype: A meta-analysis. Int. J. Cancer 122:1753–1760.

Rana BK, Hewett-Emmett D, Jin L, Chang BH, Sambuughin N, Lin M,Watkins S, Bamshad M, Jorde LB, Ramsay M, et al. (12 co-authors). 1999. High polymorphism at the human melanocortin 1 receptor locus. Genetics 151:1547–1557.

Robins AH. 1991. Biological perspectives on human pigmentation. Cambridge Studies in Biological Anthropology 7. Cambridge. Cambridge University Press.

Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ, et al. (17 co-authors). 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature 419:832-837.

Sáenz S, Conejo-Mir J, Cayuela A. 2005. Epidemiología del melanoma en España. Actas Dermosifiliogr. 96:411-418.

Scherer D, Nagore E, Bermejo JL, Figl A, Botella-Estrada R, Thirumaran RK, Angelini S, Hemminki K, Schadendorf D, Kumar R. 2009. Melanocortin receptor 1 variants and melanoma risk: a study of 2 European populations. Int J Cancer. 125:1868-1875.

Schiöth HB, Phillips SR, Rudzish R, Birch-Machin MA, Wikberg JE, Rees JL. 1999. Loss of function mutations of the human melanocortin 1 receptor are common and are associated with red hair. Biochem Biophys Res Commun. 260:488-491.

Steiner CC, Römpler H, Boettger LM, Schöneberg T, Hoekstra HE. 2009. Mol Biol Evol. 26:35-45.

Sturm RA, Duffy DL, Box NF, Newton RA, Shepherd AG, ChenW, MarksLH, Leonard JH, Martin NG. 2003. Genetic association and cellularfunction of MC1R variant alleles in human pigmentation. Ann NY Acad Sci. 994:348–358.

Takai D, Jones PA. 2002. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. Proc Natl Acad Sci USA. 99: 3740–3745.

Takahashi M, Matsuda F, Margetic N, Lathrop M. 2003. Automated identification of single nucleotide polymorphisms from sequencing data. J Bioinform Comput Biol. 1:253-265.

Tennessen JA, Madeoy J and Akey M. 2010. Signatures of positive selection apparent in a small sample of human exomes. Genome Res. 20:1327-1334.

The International HapMap Consortium. 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851-861.

Valverde P, Healy E, Jackson I, Rees JL, Thody AJ. 1995. Variants of themelanocyte-stimulating hormone receptor gene are associated with redhair and fair skin in humans. Nat Genet. 11:328–330.

Valverde P, Healy E, Sikkink S, Haldane F, Thody AJ, Carothers A, Jackson IJ, Rees JL. 1996. The Asp84Glu variant of the melanocortin 1 receptor(MC1R) is associated with melanoma. Hum Mol Genet. 5:1663–1666.

Williams PF, Olsen CM, Hayward NK, Whiteman DC. 2011. Melanocortin 1 receptor and risk of cutaneous melanoma: a meta-analysis and estimates of population burden. Int J Cancer 129:1730-1740.

Zeng K, Shi S and Wu CI. 2007. Compound tests for the detection of hitchhiking under positive selection. Mol Biol Evol. 24:1898-1908.

Legends to Figures

Figure 1.- Sliding window graph of polymorphism and CpG values around *MC1R*. The graph shows the values obtained for different parameters in sliding windows (of size 500pb, shifted 100pb) across the region. On top, the coding region (broad gray box) and UTRs (gray lines) of *MC1R*. Below this, the continuous line indicates that the GC content across the entire region is high (~65%). However, the observed to expected CpG ratio (continuous black line with regularly interspersed dots) substantially increases in the coding region (~0.75). A high content of CpGs, if methylated, could result in a high number of new variants (see Theta W, continuous line with regularly interspersed asterisks), which because of purifying selection result in highly negative Tajima's D values (~-2; discontinuous gray line).

Figure 2.- UVB radiation map of Spain. Map showing the annual average daily UVB radiation (J/m^2) per province (see Materials and Methods). Note that for aesthetical purposes the Canary Islands are displayed closer to the Iberian Peninsula than they actually are (i.e. off the cost of the Southern end of Morocco). Dots correspond to the centroids of the polygons representing each province and not the location of the capital of each province. Map representation and color interpolation has been carried out with QGIS v 1.8.0. Numeric codes for provinces or regions indicated in Tables 3 and 4 are as follows: Provinces with the lowest UV values: Asturias (1), Bizkaia (2), Gipuzkoa (3), Cantabria (4), Araba (5), Lugo (6), A Coruña (7), Navarra (8), Pontevedra (9), Ourense (10) and La Rioja (11); provinces with the highest UV values: the Canary Islands (12), Cádiz (13), Granada (14), Almería (15), Huelva (16), Cáceres (17), Jaén (18), Sevilla (19), Málaga (20), Melilla (21) and Ciudad Real (22), Melanoma samples were taken from Castellón (23).

Figure 3. EHH analysis of the 1000G European samples. On top, results of the simulated EHH values for the extended V60L haplotype A) using the demographic model for Europeans of Laval et al (2010) and B) using the demographic model for Europeans of Gutenkunst et al (2009). Shades of blue correspond to a smoothed color density representation of the scatterplot of all EHH simulations for each region, obtained through a kernel density estimate, using the smoothScatter function in R. Blue darkness is proportional to the density of points in that vicinity. The continuous blue line corresponds to the 95th percentile of the distribution (see Materials and Methods), smoothed using the loess function in R. Orange dots are the observed values from the 1000G data for South Europeans. For comparative purposes, black dots corresponding to the EHH observed profile of 1000G data from N. Europeans (GBR, FIN, CEU) are also shown. Evidence for selection in N. Europeans is not so evident as in S. Europeans, and in any case, much weaker.

Table 1.- Diversity parameters for the coding region of *MC1R* in several populations

Sample[a]	2N[b]	Length[c]	S[d]	#Haps[e]	h[f]	π[g]	$\hat{\theta}_w$[h]	k[i]	Taj's D (p-value)[j]
This work	3142	953	42	42	0.517	6.3	4.867	0.599	-2.143 (~0)
TSI	196	953	10	11	0.562	6.9	1.709	0.659	-1.437 (0.048)
ALL-AFR	492	950	11	12	0.184	2	1.624	0.192	-1.913 (0.001)
ASW	122	950	10	11	0.325	3.7	1.860	0.353	-2.014 (0.001)
LWK	194	950	4	5	0.071	0.8	0.685	0.072	-1.560 (0.010)
YRI	176	950	5	6	0.203	2.2	0.870	0.212	-1.477 (0.048)
ALL-NEU	536	951	10	11	0.674	9	1.458	0.854	-0.870 (0.210)
CEU	174	951	8	9	0.671	9	1.395	0.852	-0.873 (0.215)
FIN	186	951	7	8	0.663	8.7	1.207	0.827	-0.676 (0.282)
GBR	176	951	8	9	0.682	11	1.393	0.863	-0.852 (0.224)
ALL-ASI	572	953	6	8	0.525	9.1	0.866	0.913	0.097 (0.614)
CHB	194	953	4	5	0.555	10.2	0.685	0.967	0.736 (0.801)
CHS	200	953	4	5	0.538	10.6	0.681	1.009	0.856 (0.828)
JPT	178	953	4	6	0.435	7.3	0.695	0.664	0.081 (0.541)

[a] Samples labeled "This work" correspond to healthy individuals from Spain. The rest correspond to the 1000 Genomes Project populations: the South-European (TSI) population (the South-European IBS population from 1000G was not considered due to its low sample size; 2N=28); the African populations (individually: ASW, LWK, YRI, and all three pooled: ALL-AFR), the North-European populations (individually: CEU, FIN, GBR; the three pooled: ALL-NEU) and the Asian populations (individually: CHB, CHS, JPT, and all three pooled: ALL-ASI), [b] 2N=Number of haplotypes. [c] nucleotides considered after removing coding synonymous polymorphisms. In the case of South European samples we excluded T314T, caused by a polymorphism in nucleotide 942; in the case of the North European samples we excluded I264I, T314T and S316S, caused by polymorphisms at nucleotides 318, 531 and 942 respectively. In the African samples we excluded L106L, T177T, F300F and T314T, caused by polymorphisms at nucleotides 318, 531, 900 and 942 respectively. In the Asian samples we excluded A103A, Y298Y, and T314T, caused by polymorphisms at nucleotides 309, 894, and 942 respectively [d] Number of segregating sites. [e]Number of different haplotypes observed in the sample. [f]Haplotype diversity. [g]Nucleotide diversity ($*10^{-4}$). [h]Theta (per sequence) estimated from the number of segregating sites. [i]Average number of nucleotide differences between haplotypes. [j]Tajima's D value (significance inferred after 5.000 standard coalescent simulations with DnaSP 5.1).

Table 2. Neutrality tests for the observed *MC1R* diversity in the Spanish sample

	Taj's D[a]	F&W´s nH[b]	F&L´s D[c]	F&L´s F[d]	DHEW[e]
Observed values	-2.143	0.161	-7.300	-6.000	NA[f]
p-values, under models:					
Single, constant-size pop[g]:	~0.000*	0.396	0.000*	0.000*	~0.112
Guntekunst et al. (2009)	~0.0002*	~0.597	~0.0001*	~0.0003*	~0.325
Laval et al. (2010)	~0.001*	~0.942	~0.0007*	~0.002*	~0.992

[a] Tajima´s D test. [b]Fay and Wu's normalized H test. [c]Fu and Li´s D test. [d]Fu and Li´s F test.

[e]Zeng et al's combined DHEW test. [f]Not applicable. [g] no recombination assumed

*Significant after Bonferroni correction

Table 3. Fisher exact test of V60L (rs1805005) allele frequency

	2N alleles		Frequency V60L allele		Fisher exact
Group 1 vs Group 2[a]	Group 1	Group 2	Group 1	Group 2	test (p-value)[b]
a) UVB ≤ 5th vs. UVB ≥95th	188	134	32	13	0.042
b) UVB ≤ 10th vs. UVB ≥ 90th	266	170	41	16	0.046
c) UVB ≤ 15th vs. UVB ≥ 85th	348	270	55	28	0.057
d) UVB ≤ 20th vs. UVB ≥ 80th	466	314	72	37	0.088

[a]Population groups compared: in a) Group 1 includes those provinces whose daily (annual mean) received UV-B radiation is lower than or equal to the 5th percentile of the distribution of UV-B radiation in the Iberian Peninsula. Group 1: Asturias (1), Bizkaia (2) and Gipuzkoa (3) (all of them northern populations); Group 2 includes those provinces whose daily (annual mean) received UV-B radiation is greater than or equal to the 95th percentile of the distribution of UV-B radiation in the Iberian Peninsula: the Canary Islands (12), Cádiz (13) and Granada (14). Similarly for b), c) and d) but considering the percentiles indicated. In b) Group 1 as in a) plus Cantabria (4), Araba (5) and Lugo (6); Group 2 as in a) plus Almería (15), Huelva (16) and Cáceres (17). In c) Group 1 as in b) plus A Coruña (7), and Navarra (8); Group 2 as in b) plus Sevilla (18) and Jaén (19). In d) Group 1 as in c) plus Pontevedra (9), Ourense (10) and La Rioja (11); Group 2 as in c) plus Ciudad Real (17), Málaga (20), Melilla (21) and Ciudad Real (22). Only individuals whose grandparents were also born in these provinces were considered. See Figure 2 to localize provinces or regions numeric codes on the map.

[b] Fisher exact test 1-tailed p–values.

Table 4. Cochran-Armitage association test for V60L (rs1805005) and melanoma risk

		Genotypes[a]			Exact HW test	Cochran-Armitage
Sample	N (inds)	GG	GT	TT	(p-value)	trend test (p-value)[b]
Melanoma[c]	722	494	210	18	0.54	
UVB (upper 95%)[d]	72	52	8	1	0.66	0.0134
UVB (upper 90%)[e]	85	62	11	1	0.87	0.0089
UVB (upper 85%)[f]	139	101	21	2	0.70	0.0059
UVB (upper 80%)[g]	157	115	28	3	0.60	0.0222

[a]The T allele is responsible for the V60L aminoacid change. [b]Tested against the melanoma sample under an additive model. [c] Individuals from our own sequenced melanoma samples from Castellón (22) plus those of Fernández et al. (2007) and Ibarrola-Villava et al. (2012). [d]Includes individuals from the provinces of the Canary Islands (12), Cádiz (13) and Granada (14) and whose grandparents are also from the same province; [e]as in [d] plus Almería (15), Huelva (16) and Cáceres (17); [f] as in [e] plus Sevilla (18) and Jaén (19); [g] as in [f] plus Málaga (20), Melilla (21) and Ciudad Real (22). See Figure 2 for localizing the provinces or regions on the map by using the corresponding numeric codes between parentheses.

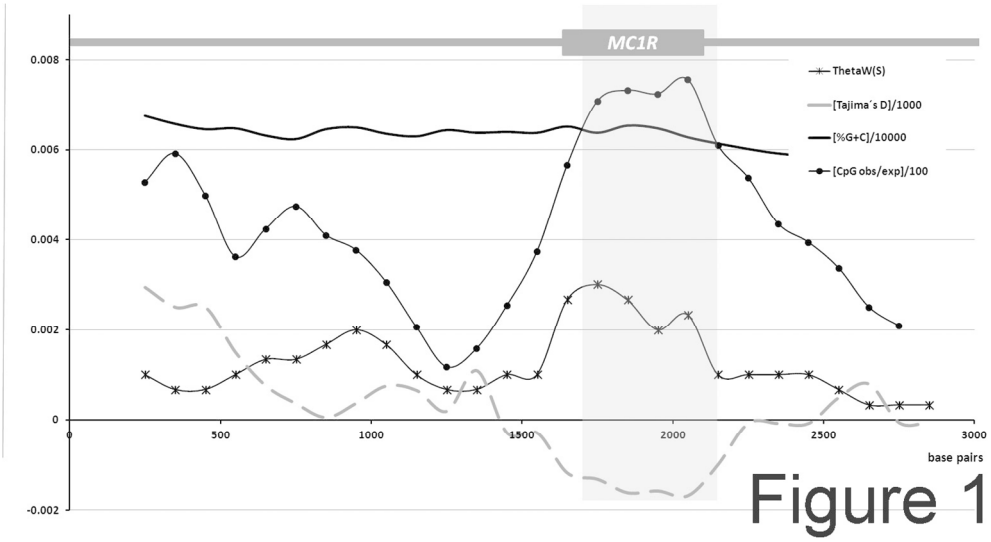

Figure 1

Sliding-window graph of polymorphism and CpG values around MC1R

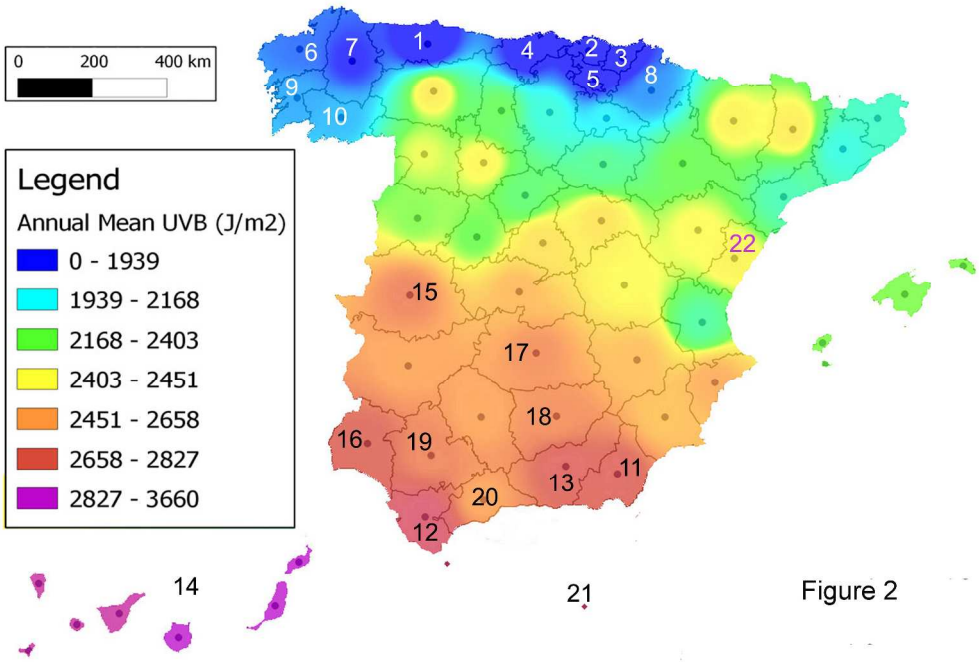

Figure 2

UV radiation map of Spain

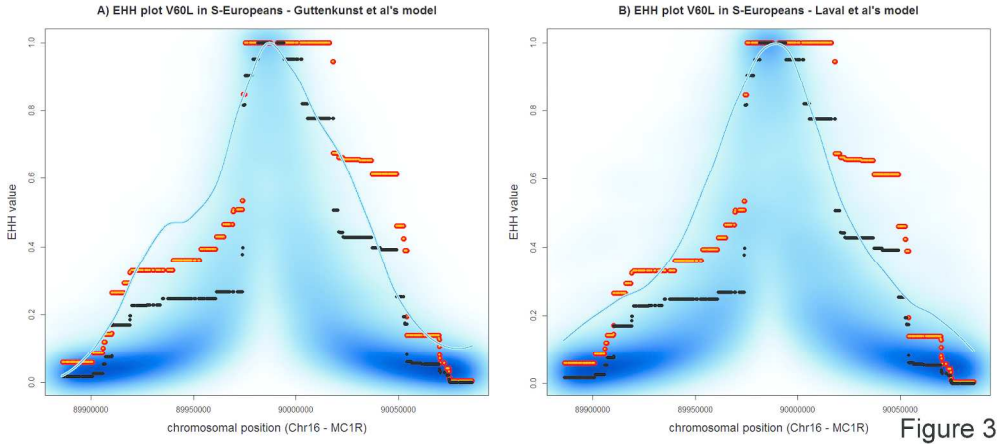

Figure 3

EHH analysis of the 1000G European samples