Master Thesis

UPPSALA UNIVERSITET

Identifying genetic signatures of recent local adaptations in people from Ibiza

Submitted for the degree of Master of Science

Diego Alejandro Londoño Correa

Supervised by

Elena Bosch (Institute of Evolutionary Biology, CSIC-UPF, Barcelona)
Carina Schlebusch (internal supervisor at Uppsala University)

March 2021

Table of Contents

List of tables

List of Figures

ABSTRACT

Islands have been considered natural laboratories to study evolutionary processes. Ibiza is a small island in Spain whose population stands out from other Spanish populations due to its particular demographic and historical processes. War, famine, and several epidemics have affected Ibizans, and these phenomena could have left signatures of positive selection in their genomes. Here, we used three different methodologies to detect positive selection: The Population Branch Statistic (PBS), the Integrated Haplotype Score (iHS), and the Cross-Population Extended Haplotype Homozygosity (XP-EHH). We used a sliding windows approach to control for spurious results. The candidate windows for selection were chosen using three different criteria for each test: maximum and mean score within each window, and proportion of high scores in each window. Only the windows being simultaneously on the top of each of the three criteria were selected for annotation and enrichment analyses. The most common traits associated with the SNPs present in the candidate windows were blood function, cardiovascular diseases, body mass measures, lipid metabolism, renal function, and skin diseases. We suggest some hypotheses to explain the selection signatures related to some of these traits and some recommendations for further studies to overcome the present research's limitations.

INTRODUCTION

Islands are important places to study evolutionary processes. They have been considered laboratories of evolution, suitable to study the mechanisms of species formation and adaptive radiation (Losos & Ricklefs, 2009). Some studies suggest that the evolutionary rate in islands is faster compared to mainland species (Garcia-Porta et al., 2016; Millien, 2006), whereas others have found the opposite (Salazar et al., 2019). One of the most studied evolutionary processes in islands is the change in body size in insular species compared to their mainland relatives. Some authors have suggested a pattern of giantism in smaller species and dwarfism in larger species (Lomolino, 1985). This pattern has been also seen in humans, particularly in the Flores island where a member of the genus *Homo*, allegedly *Homo erectus,* suffered from insular dwarfism becoming *Homo floresiensis* (Brown et al., 2004). This insular dwarfism is suggested to have occurred again in contemporary *Homo sapiens* when genomes from current Flores inhabitants in the village of Rampasa were found to have signatures of recent positive polygenic selection acting on standing variation (Tucci et al., 2018) that contributed to their short-stature phenotype (average 145 cm). However, evidence of recent evolution in humans on islands is scarce. Another evolutionary process that commonly happens on islands, is the genetic drift experienced by founder populations, with a consequent reduction in the genetic diversity (Barton & Mallet, 1996). This reduction in diversity is correlated with island size, with smaller islands harboring lower allelic diversity (Jordan & Snell, 2008).

Ibiza (Eivissa in Catalan) is the smallest island of the three main Balearic Islands in the Mediterranean Sea, and it is located approximately 70 km from mainland Spain. A recent study found that the genetic differentiation of Ibiza from the rest of Spain is at least of the same order of magnitude as that of the Basque Country (Biagini et al., 2019). This differentiation was particular for Ibiza and was not found for the other Balearic Islands. Furthermore, the Ibizan differentiation cannot be attributed to genetic influences from other populations that have settled on the island: ancient Phoenicians who first settled there, or contemporary Middle Easterners, Northern Africans, or other Spanish populations (Biagini et al., 2019). Ibizans seem to have suffered their own evolutionary processes for 700 years since the arrival of the Catalans who conquered the island in 1235 CE until the extensive immigration from the hippy community in the second half of the 20th century. When using genetic data to estimate the variation in population size (Ne), there is evidence of a collapse of Ne between the 16th and 17th century and a recovery that started in the 18th century. This is consistent with the historical records of famine after the Franco-Ottoman attack in 1536, the global bubonic plague epidemic in 1652 (Kohn, 2007), and malaria epidemics for many generations (Picornell et al., 1996; Ramon et al., 2008). These events could have significantly contributed to its current genetic differentiation and could have represented a major opportunity for natural selection to have altered the frequencies of adaptive alleles in key regions.

In the present study, we investigated whether natural selection could have acted specifically in Ibiza in response to such selective pressures by searching for signatures of adaptive evolution in current Ibizans and localizing specific signals when compared to the other Balearic Islands and continental Spain. Our hypothesis is thus that the events that led to the reduction of effective population size in Ibiza may have also offered an opportunity for adaptive selection to act. To test this hypothesis, we performed three different tests for positive selection: the Population Branch Statistic (PBS), a population differentiation-based method, and two methods focused on detecting long unbroken haplotypes (or extended haplotype homozygosity), the Integrated Haplotype Score (iHS), and the Cross-Population Extended Haplotype Homozygosity (XP-EHH). These methods were chosen as they represent the present cutting-edge techniques to detect recent selection in human populations (Vitti et al., 2013).

PBS is a population differentiation-based method, which has been shown to have a strong power to detect local genetic adaptation (Yi et al., 2010). As a result of a geographically restricted selective pressure, positive selection may cause a particular adaptive allele to rise to high frequency in a population where it results in a beneficial phenotypic outcome but not in other populations from distant regions where it confers no selective advantage and thus remains neutral. This situation creates a distinctive pattern of extreme population differentiation for these genomic regions harboring adaptive alleles that have facilitated local adaptation. These alleles stand out from the remaining presumably neutral (i.e., non-selected) alleles of the genome, whose pattern of population differentiation will be determined only by demographic processes such as migration and genetic drift (Vitti et al., 2013). In particular, PBS is based on the estimation of the amount of allele frequency change that occurred at a given locus in the history of a test population since its divergence from two other populations (Cardona et al., 2014), one closely related and another more distantly related.

iHS and XP-EHH are two linkage disequilibrium-based methods exploring unusually large haplotypes at high frequency. Such pattern is only expected after a recent selective sweep since each adaptive variant, and all its neighboring variants are rapidly raised to high frequency during a sweep and subsequently observed as long unbroken haplotypes at high frequencies until recombination breaks their association with time (Vitti et al., 2013). Both statistics are based on the measure of the extended haplotype homozygosity (EHH) proposed by Sabeti et al. (2002). For each SNP (Single Nucleotide Polymorphism), EHH is measured as the probability that two randomly chosen haplotypes are identical over a range next to the tested SNP.

For iHS, the extended haplotype background for the derived allele is assessed against the haplotype background of the ancestral allele at each site (Voight et al., 2006). A slower decay in the haplotype homozygosity for the derived alleles compared to ancestral ones is an indicator of selection. The iHS has good power to detect incomplete selective sweeps, at a moderate frequency (~50%–80%), but low power to detect sweeps that have reached high frequency (>80%) or fixation (Pickrell et al., 2009). Contrary to the iHS, in the XP-EHH, the extended haplotype background is assessed against the haplotype background of another population, where longer haplotypes in one population are indicators of selection. The XP-EHH statistic compares haplotype lengths between populations under differential selective pressures to control for geographical variation in recombination rates (Pickrell et al., 2009; Vitti et al., 2013), as recombination affects the haplotype homozygosity scores. The XP-EHH and has been shown especially useful to detect selective sweeps near fixation (above 80% frequency) within one population.

For each test, we used an overlapping windows approach to minimize spurious scores of individual SNPs and considered as candidates for positive selection those genomic windows identified as outliers according to three different genome-wide empirical criteria. We then annotated all candidate regions analysis to link the top candidate regions to putatively selected traits at the phenotypic level. Additionally, performed an enrichment analysis with the software traseR to link all candidate regions to putatively selected traits, using cataloged trait-associated variants. Our results show that even though we use tests that detect different patterns of selection, some putative adaptive traits appear consistently across the three tests. Moreover, our approaches of gene annotation, and SNP-based enrichment analysis, are also consistent between them, and each one helped to disentangle some of the complex associations between the candidate SNPs, genes and traits.

MATERIALS AND METHODS

Data

We compiled genomic data for 489 individuals from 8 populations. 189 of them were Spanish from the Iberian Peninsula and the Balearic Islands including 13 from Ibiza (EIV) and 176 from other populations in Spain (IBE). Genotyping data for these Spanish samples were retrieved from a previous study (Biagini et al., 2019) and comprised ~629 K SNPs genotyped using the Axiom® Genome-Wide Human Origins Array (Patterson et al., 2012) that resulted in 519,297 variants when merged as described in (Biagini et al., 2019). Additionally, we retrieved reference genotypes from the 1000 Genomes Project (Auton et al., 2015) including 300 samples from 6 different populations and 50 individuals per population: China (CHB), Yoruba (YRI), Finland (FIN), Great Britain (GBR), Toscani in Italy (TSI) and Utah Residents with Northern and Western European ancestry (CEU).

Quality control

We excluded variants where more than 5% of the genotype calls were missing with the command –*geno* 0.05 in PLINK (Purcell et al., 2007) and samples where more than 10% of the genotype calls were missing with the command –*mind* 0.1.

This was done separately for the set of 6 reference populations from the 1000 Genomes Project (1000G samples: CHB, YRI, FIN, GBR, TSI, CEU) and the Spanish populations (SP samples: IBE and EIV), respectively. The reason was that the 1000G sample genotypes result from whole-genome sequencing whereas the genotypes in the SP samples result from a genotyping custom array.

This filtering left 80,855,722 SNPs (no variants or samples removed) for the 1000G sample set and 519,297 SNPs (35 variants removed, no samples removed) for the SP sample set. Before merging the data from both groups, we further corrected for strand flipping issues and removed multiallelic SNPs. This resulted in 80,845,346 variants for the 1000G samples, and 518,608 variants for the SP samples, which became 498,132 variants when merged. The PLINK integrated KING-robust (Manichaikul et al., 2010) pairwise relatedness estimator was used to prune for related pairs, using a cutoff of 0.125 to exclude second-degree relations or duplicate samples. All the samples remained after this filter.

Pruning linked variants, PCA and Admixture

Before Principal Component Analysis (PCA) and Admixture analyses, we removed variants with low MAF (Minor Allele Frequency) and LD-pruned the resulting merged data to better capture its global structure. In particular, we first removed variants with MAF < 0.02 in the 8 populations, which resulted in 403,489 SNPs and then looked for SNPs that were in linkage equilibrium with the –indep-pairwise 200 25 0.5 PLINK command, so that no pair of SNPs within 200 kilobases (windows overlapping by 25 kilobases) had a squared-allele-count-correlation (r^2) greater than 0.5, which resulted in a total of 189,744 independent SNPs to be used as input for the PCA and Admixture analyses.

PCA was performed in PLINK and plotted in R (R Core Team, 2017) at four geographic scales: (1) At the global level, including 8 populations from different continental areas (CHB, YRI, FIN, CEU, GBR, TSI, IBE, and EIV); (2) at the European-ancestry level (FIN, CEU, GBR, TSI, IBE, EIV); (3) in South Europe (including the TSI, IBE, and EIV populations); as well as (4) considering only two focal populations (IBE, with sub-populations for better resolution, and EIV). The Admixture Analysis was performed in Admixture 1.3 (Alexander & Lange, 2011). This analysis is based on the assumption that each individual has ancestry from one or more of K genetically distinct sources (Lawson et al., 2018). For each dataset we tested 2 to 12 K's, doing 10 runs for the same K, with different fixed seeds for each run to make the results replicable. The results were plotted with the software pong (Behr et al., 2016).

Selection Tests

Genomic signatures of positive natural selection were explored genome-wide through the use of the PBS (Population Branch Statistic), a population differentiation-based method, and two methods focused on detecting long unbroken haplotypes (or extended haplotype homozygosity), the iHS (Integrated Haplotype Score), and the XP-EHH (Cross-Population Extended Haplotype Homozygosity). All tests were applied on the 403,489 SNP genotyping data that remained after the quality control and with a MAF > 0.02.

PBS

We computed the PBS as suggested in Yi et al. (2010) from the F_{ST} values calculated pairwise between three different populations: Ibiza (EIV), Spain without Ibiza (IBE), and Great Britain (GBR). GBR was chosen as an outgroup, as this is also an European population, and the samples from this population are considered as representative of the geographical region, unlike the next closest population TSI, that is not considered to be representative of Italy. F_{ST} values were computed with the software PLINK v1.9 (Purcell et al., 2007), according to the method developed by Weir and Cockerham (1984), for each pair of populations using 403,489 SNPs. The PBS was then calculated with in-house scripts in R, transforming each corresponding F_{ST} in a T value:

$$T = -log(1 - F_{st})$$

The T values were used to calculate the specific PBS for EIV as follows:

$$PBS = \frac{T^{EI} + T^{EG} - T^{IG}}{2}$$

where E is EIV, I denotes IBE and G is GBR.

iHS and XP-EHH

Genotyping data was first phased with Shape-IT (Delaneau et al., 2008) using GRCh37 as the reference genome and EHH was subsequently calculated for each chromosome with the function *scan_hh* from the software rehh (Gautier & Vitalis, 2012).

Having EHH, iHS can be calculated. The statistic requires polarized SNPs (i.e. knowledge of the ancestral and derived alleles of each SNP), as it compares the area under the curve defined by EHH between the derived and the ancestral alleles, as one travels further in genetic distance from the core region. Extreme long iHS values are suggestive of positive selection (Vitti et al., 2013). The information about ancestral and derived alleles was taken from the 1000 genomes project (Auton et al., 2015). It was integrated into our data using the function annotate from bcftools (Danecek et al., 2011), and taken by rehh with the function *polarize_vcf = TRUE.* After the polarization, only 222,927 (55%) out of the initial 403,489 SNPs had this information. To calculate the iHS for these 222,927 SNPs, the function *ihh2ihs* from the software rehh was executed. The *freqbin* option was set to 0 as the software suggests using this value when there is a large number of markers and few different haplotypes, as is the case in EIV. The analysis was also done for IBE. The R scripts can be found in https://github.com/dielondono/Eivissa_selection/blob/master/rehh_eivissa.R for EIV and https://github.com/dielondono/Eivissa_selection/blob/master/rehh_iberia.R for IBE. As the iHS statistic is performed in one population without any external comparison, it could be the case that what it is detected in EIV is not specific to EIV but a selection signal already present in the ancestral population that originated IBE and EIV. Thus, to identify iHS signals specific to EIV, we filtered for those iHS signals not shared with IBE.

As the XP-EHH statistic does not require information about the ancestral or derived allele, the analysis could be performed for the 403,489 SNPs that remained after the quality control and with a MAF > 0.02. The two compared populations were EIV and IBE, and the rehh was used as for iHS but without polarizing ancestral and derived alleles. EHH was retrieved using the script: https://github.com/dielondono/Eivissa_selection/blob/master/transform_IHH.R and the XP-EHH between EIV and IBE was retrieved using: https://github.com/dielondono/Eivissa_selection/blob/master/xp-ehh.R

Windows approach

To minimize spurious scores of individual SNPs for PBS, iHS, and XP-EHH, all subsequent interpretations of signatures of positive selection and enrichment analyses were done using an overlapping windows approach. After testing different sizes of overlapping windows with steps of 1/5 or 1/4 of the window size, we ended up selecting a window size of 100 Kb overlapping every 25 Kb. This size was chosen because in smaller windows sizes (50, 25 and 10 Kb) more than 40% of the windows were removed after selecting windows with at least 5 SNPs. In the case of our selected size, for PBS and XP-EHH, ~87% (~89,000 out of ~102,000) of the considered windows remained after removing those with less than 5 SNPs. The mean number of SNPs in the remaining windows was 15 with a standard deviation of 10. As for the iHS statistic, due to the polarization process, only 71.5% (69,320 out of 96,890) of the windows remained when considering a 100 Kb window size overlapping every 25 Kb and removing those with less than 5 SNPs. The mean number of SNPs in the remaining windows for iHS was 9.11 with a standard deviation of 6.53.

Three different criteria were used to identify regions with signatures of recent positive selection. For each selection score, we used the maximum (Max) and mean (Mean) value of all SNPs within a window. Then, we took the top 1% (99th percentile) windows for the maximum and mean genome-wide distributions. Additionally, we considered the top 1% of the SNPs in the genome-wide distribution for each statistic score, looked for the proportion (Prop) of top 1% SNPs within each window and subsequently, took the top 5% windows genome-wide with the highest proportion of top 1% SNPs within each window and score. For iHS, we considered the absolute values to obtain the statistic distribution and identify the corresponding top scores. Finally, we considered as candidates for positive selection those genomic regions detected as outliers for a given statistic according to the three criteria (Max, Mean, and Prop).

Annotation and Enrichment Analysis

The SNPs present in the candidate regions for positive selection were annotated with Annovar (Wang et al., 2010). These included 1,006 SNPs for PBS, 1,297 for XP-EHH, and 530 for iHS. Moreover, all genes present in the top 5 windows with the highest Max value (and meeting the three intersection criteria) were further investigated using Gene Cards (https://www.genecards.org/) to explore for any putative adaptive trait or phenotype that could be associated with them.

In addition, we used an R Bioconductor package named traseR (Trait-Associated SNP EnRichment analysis) to perform an enrichment analysis of the genes and functions annotated in all candidate regions for recent positive selection. TraseR uses trait-associated SNP (taSNPs) taken from different GWAS databases with information about variants that are significantly associated with common diseases and traits. With this software, the databases are exploited to indicate whether the genomic intervals of a particular query are likely to be functionally connected with certain traits (Chen & Qin, 2016). Unlike other enrichment resources like Gene Ontology, based on genes and functional categories, in TraseR functional categories are replaced by traits and genes by genomic intervals. This allows capturing associations of traits within intergenic regions, differentially methylated regions, or putative enhancers. TraseR also allows comparing how enriched are our query windows compared to the genome-wide SNPs associated with the analyzed trait. This allows avoiding over/underestimations of the number trait-hits in our significant SNPs. We used a binomial test where the null hypothesis states that the chance of observing a SNP being a taSNP is the same in our query genomic windows as in the whole genome. We included the LD-taSNP option in TraseR to include all the SNPs that are in tight linkage disequilibrium (r2 > 0.8) with any of the taSNPs. This extended set contains 78,247 unique SNPs. We only took those SNPs with a p-value < 0.001, and only analyzed those traits that passed the Bonferroni correction for multiple testing.

Control for iHS signals in Iberia

As the iHS statistic is performed in one population without any external comparison, it could be the case that what we find as selected in EIV is not specific to EIV but was already selected in the ancestral population that originated IBE and EIV. We performed on IBE the same analyses done previously for iHS in EIV, to see if the signals were specific to EIV or shared with IBE.

RESULTS

PCA and Admixture

Biagini et al. (2019) described that Ibizans are genetically distinct from an ancient Phoenician sample found in the Island, and from modern populations in North Africa, the Middle East, and the rest of Spain. We set to replicate these findings as a quality control check of our newly compiled dataset before performing the positive selection tests.

When using all populations from the compiled dataset, the first two principal components explain 59.66% (PC1 41%, PC2 18.66%) of the variation and separate Africans (YRI), East Asians (CHB) and Europeans (Fig.1). When only populations of European ancestry were analyzed, PC1 (explaining 2.77% of the variation) separates FIN, and PC2 (explaining 1.46% of the variation) separates CEU and GBR from the other populations (IBE, TSI and EIV) (Fig. 2)

We then analyzed only populations of South Europe including TSI, IBE, and EIV, but increasing the resolution for IBE populations by using labels for different localities. At this point, PC2 (explaining 1.29% of the variation) makes an EIV cluster apart from the other populations, and

PC1 (explaining 1.30% of the variation) does so for TSI at one extreme, Basques at the other, and the rest of Spain in between (Fig. 3). Finally, the PCA for IBE and EIV replicates the results from Biagini et al. (2019), where the PC1 (explaining 2.77% of the variation) separates EIV and PC2 (explaining 1.46% of the variation) separate Basques from other Spanish populations (Fig. 4).

Additionally, we performed an ADMIXTURE analysis with the 8 populations, for which K = 3 and K = 4 have the lowest cross-validation error (Cv-error) (Fig. 5). The four major components detected corresponded to YRI, CHB, North Europe and South Europe (Fig. 6). A Basque component appears at K = 6, and an EIV component appears at K = 7 (Fig. 7), before the TSI component that appears at K = 8.

Both PCA and Admixture analyses replicate previous results from Biagini et al (2019) and confirm that the EIV data to be used in the subsequent analyses have been correctly integrated with the new external datasets (Iberian and GBR).

Selection Tests

To identify the genomic signatures of adaptations in the Ibiza population resulting from recent historical events impacting the population (such as episodes of famine, malaria or the plague), three different selection tests were performed on EIV as the focal population.

From an initial dataset of 403,489 SNPs, the computational performance and requirements of each test — as for example, the need of allele frequency information for a given SNPs in each population to be compared pairwise in the F_{st} calculation for PBS or the polarization of the ancestral and derived alleles for the iHS statistic — determined different final numbers of SNPs with values in each statistic (Figure 8). After considering a window size of 100 Kb and overlapping windows every 25 Kb with statistic values for at least 5 SNPs, we selected as candidates for positive selection only those windows whose maximum, mean and proportion values were found on the top 1% (99th percentile) of each corresponding genome-wide distribution. The corresponding unique SNPs in these outlier windows matching the three criteria (maximum, mean and proportion) comprised 1,006 SNPs for PBS, 1,297 for XP-EHH, and 530 for iHS (Figure 9, Supplementary Tables S1-3). Additionally, we found SNPs shared between the top windows for some tests: 84 SNPs shared between PBS and XP-EHH and 37 between XP-EHH and iHS (Supplementary Tables S4-5).

Control for iHS signals in Iberia

When the windows approach is also used for iHS in IBE, no shared SNPs are found between EIV and IBE for the top windows selected with the three different criteria. However, we did a more restrictive control selecting the top 1% values for iHS in IBE without using the windows approach and found 30 SNPs that went in the same direction for iHS in IBE as the SNPs in the top windows in EIV. These SNPs were removed for the annotation, the top 5 regions selection and the enrichment analysis with traseR.

Annotation of candidate genes for positive selection in EIV

Candidate SNPs were annotated to identify putative functional polymorphisms and to locate genes potentially associated with them. As many genes were found, we only took those genes present in the top 5 regions (Figure 10) as well as those shared across the different signatures shown by each statistic to explore for any putative adaptive trait or phenotype that could be associated with them. If two windows were consecutive, they were taken as part of the same region, until finding 5 independent regions for our analysis. The most common traits or phenotypes annotated for the top 5 regions are highlighted in bold font and per each statistic in Tables 1-3.

The 84 candidate SNPs shared between the PBS and XP-EHH analyses and the 37 common candidate SNPs identified by XP-EHH and iHS were also annotated. The complete list of genes and traits associated with these SNPs are shown in Tables 4 and 5.

Notably, several similar biological functions are found across the top 5 candidate regions for positive selection identified by each statistic as well as across the shared candidate SNPs found across the three analyses:

Blood function. XP-EHH: *LINC01506*, *PIP5K1B*, *ACOXL* (with the rs56088557, linked to erythrocyte count, among the top 1% scores) and *NEPRO*. iHS: *HCG18*, *HCG17*, *TRIM39*, *TRIM39-RPP21*, *HLA-E* and *GATB*. Shared between PBS/XP-EHH: *GLIS3*. Shared between XP-EHH/iHS: *C3ORF38*, *GATB*, *OR4D2* and *EPX.* **Body mass measures**. PBS: *AGFG1* (with the rs6741427, linked to body height, among the top 1% scores) XP-EHH: *NEPRO* and

LINC02044 (with the rs2700201, linked to body height, among the top 1% scores). iHS: *NSUN3* and *LINC02273*. PBS/XP-EHH: *ARL4A, PTPRZ1* and *FEZF1*. XP-EHH/iHS: *C3ORF38, EPHA3, LINC02273* and *OR4D2.* **Lipid metabolism**. PBS: *APOA* family in the top 2 windows. XP: *ACOXL*, *LINC02044* and *FMO5*. **Coronary heart disease**. PBS*: APOA1-AS, APOC3, APOA5, LRRTM4*. XP-EHH*: PIP5K1B, LINC02044*. iHS*: NSUN3*. XP-EHH/iHS *MSX2P1.* **Renal function**. PBS: *LOC654841*, *C2ORF83*. XP-EHH: *LINC01506*, *PIP5K1B* and *NBPF19* . IHS: *HLA-E* (with the rs1265159, linked to membranous glomerulonephritis, among the top 5% scores). **Skin diseases**. PBS*: APOA4*. XP-EHH*: OCA2*. iHS: *HCG18, TRIM39, HLA-E*. PBS/XP-EHH: *ETV1.* **Eye diseases**. PBS: *MFF, LOC654841*. PBS/XP-EHH: *GLIS3.*

Even if a given gene can be related to several functions or phenotypes, the identification of common functions across the most promising candidate regions for positive selection allows identifying putative adaptive functions specific of the EIV population,

EIV Enrichment Analysis

Genomic windows of all candidate SNPs were subsequently analyzed with TraseR to identify the biological functions or traits overrepresented in the candidate regions for positive selection in Ibizans. Results are shown in Table 6 for individual traits that passed the Bonferroni correction (q value in the table). Colors are used to show the most common categories when the traits associated with the three tests are combined and summarized in Table 7.

DISCUSSION

Taken together, our results show some consistencies at the trait level for the three different approaches used to detect signatures of selection on the island of Ibiza. Body mass measures, lipid metabolism, blood function, cardiovascular diseases, skin diseases, kidney function, and immune system diseases were very common categories of traits that constantly appeared across the tests, using both gene annotation and enrichment analysis.

When we looked at the top genes for each test and shared genes between tests, some of them were associated with many of these categories at the same time. *LINC02044* was associated with body mass measures, lipid metabolism, and cardiovascular diseases. *OR4D2*, *C3ORF38*, and *NEPRO* were related to body mass measures and blood function. *NSUN3* was associated with body mass measures and cardiovascular diseases. *APOA* family with lipid metabolism and cardiovascular diseases. *LINC*01506 and *PIP5K1B* with cardiovascular diseases, blood, and renal function. *HLA-E* was associated with blood, renal function, and skin diseases. *LOC654841* was related to renal function and eye diseases. Finally, *HCG18* was associated with blood function and skin diseases. This does not mean that other genes in our candidate windows are not related to many traits as well. Still, we only analyzed those genes in the top 5 candidate regions and those that were shared between tests. To find more relations, a closer inspection is required, using al the genes after the annotation, and not only the top 5.

The enrichment analysis with traseR based on SNPs is consistent with the analysis based on genes in the top regions and shared SNPs across tests. Blood function, body mass measures, cardiovascular and skin diseases are traits overrepresented in our windows compared to what is expected across the genome. TraseR classifies traits related to erectile dysfunction, prostatic neoplasms, uric acid, and kidney diseases in the same category: male urogenital diseases. We grouped these categories with the label “kidney function” given that in other analyses using the NCBI database for phenotypes, we found that the SNPs and genes were usually associated with terms like “urinary bladder”, “renal cell” and “kidney failure” but not with other urogenital diseases not related to the kidney function.

Since no phenotypical data of Ibizans has been specifically compiled, it is hard to link these results to observed traits in the population. However, three important patterns can be linked to the geographical and historical conditions in the Island. Firstly, Ibiza suffers from frequent droughts (Lorenzo-Lacruz & Morán-Tejeda, 2016), which could have historically affected the availability of water for human consumption. It is known that hydration directly affects the glomerular filtration rate (Anastasio et al., 2001) and is linked to kidney disease (Roncal-Jimenez et al., 2015). Hence, the signatures of selection for SNPs and genes related to the glomerular filtration rate and kidney function in EIV might suggest that historical water availability could have been a selective pressure on Ibizans. Secondly, Ibiza is the smallest island of the three main Balearic Islands. It is known that body size tends to be affected in many species after some generation when new populations arrive in small islands. One common case in humans is the insular dwarfism that appeared at least two times in *Homo* history: *Homo floresiensis* dwarfism (Brown et al., 2004) and contemporary *Homo sapiens* dwarfism in the current Flores inhabitants (Tucci et al., 2018). Interestingly, the same study (Tucci et al., 2018) found genetic signatures of selection on genes related to lipid metabolism in contemporary inhabitants. We also found signals of selection on SNPs and genes related to lipid metabolism and body mass. Moreover, some of the genes that showed signatures of selection were associated with both traits at the same time. This might suggest that Ibizans could have suffered a process of positive selection on their body size, at least partially influenced by selection on lipid metabolism genes. Finally, Ibiza has historically been impacted by epidemics of malaria (Picornell et al., 1996; Ramon et al., 2008) and bubonic plague (Kohn, 2007). We found many SNPs and genes related to immune function and blood cells. At the top genes level in iHS, we found some contiguous genes in the chromosome 6, linked to the HLA complex group, related to the innate immune system and class I MHC mediated antigen processing and presentation. *EPX,* found with a strong signal of selection by both XP-EHH and iHS, encodes for a peroxidase enzyme that is released at sites of parasitic infection or allergen stimulation to mediate lysis of protozoa or parasitic worms and plays a role in eosinophil activation when protecting the body against malaria by induction of parasite killing (Kurtzhals et al., 1998). Moreover, some of the most common traits associated with the genes in the top regions were eosinophil count, erythrocyte count, and platelet count, cells that play an important role in malaria infection. The traits found with traseR: receptors, transferrin for PBS, and blood coagulation factor Inhibitors for XP-EHH may also be linked to this type of infection. These hypothesized roles in the adaptation of Ibizans to the Island need to be tested with current and historical phenotypic data in further studies.

It is also important to consider that our sample size for EIV was very small (13 individuals), and this can impact the statistical power of our analyses. Simulations using the YRI demography and a selection coefficient of 1% have shown that while iHS power to detect selected alleles at a frequency of ~0.70 ranged between 0.20 and 0.40 for sample sizes of 20 and 40 chromosomes, respectively, power for XP-EHH was close to 0.60 for both samples' sizes for alleles at a frequency greater than 0.90, and up to 0.80 for detecting fixed selected alleles (Pickrell et al., 2009). Thus, some adaptive signals in our data could have been missed in our analyses due to the small sample size available. Moreover, the recent and severe bottlenecks experienced in the recent demography of Ibizans might have also decreased the power to detect selection. It has been shown that bottlenecks shift the distribution of iHS toward negative values; the more severe the bottleneck, the greater the increase in LD (Macpherson et al., 2008).

On the other hand, EIV is a population with a high level of inbreeding evidenced by a high number of long runs of homozygosity (Biagini et al., 2019). This could have impacted the variability of the haplotypes. Moreover, it has been shown that sometimes the signatures of inbreeding and selection using LD methods can be confused (Ablondi et al., 2019; Ghoreishifar et al., 2020; Islam et al., 2019). However, the haplotype-based methods that we used here (iHS and XP-EHH) consider haplotype lengths across the whole genome and take the outliers of the distributions. Moreover, our overlapping windows approach that selected windows based on three different criteria is another control for false positives of selection that could appear given the demographic history of the population. In further studies, we suggest analyzing more samples from Ibiza and collect phenotypic data to test whether our findings and potential evolutionary explanations are correct.

REFERENCES

Ablondi, M., Viklund, Å., Lindgren, G., Eriksson, S., & Mikko, S. (2019). Signatures of selection in the genome of Swedish warmblood horses selected for sport performance. BMC Genomics, 20(1), 717. https://doi.org/10.1186/s12864-019-6079-1

Alexander, D. H., & Lange, K. (2011). Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics*, *12*(1), 246. https://doi.org/10.1186/1471-2105-12-246

Anastasio, P., Cirillo, M., Spitali, L., Frangiosa, A., Pollastro, R. M., & De Santo, N. G. (2001). Level of hydration and renal function in healthy humans. Kidney International, 60(2), 748-756. https://doi.org/10.1046/j.1523-1755.2001.060002748.x

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., Korbel, J. O., Lander, E. S., Lee, C., … National Eye Institute, N. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68-74. https://doi.org/10.1038/nature15393

Barton, N. H., & Mallet, J. (1996). Natural Selection and Random Genetic Drift as Causes of Evolution on Islands [and Discussion]. Philosophical Transactions: Biological Sciences, 351(1341), 785-795.

Behr, A. A., Liu, K. Z., Liu-Fang, G., Nakka, P., & Ramachandran, S. (2016). pong: Fast analysis and visualization of latent clusters in population genetic data. *Bioinformatics*, *32*(18), 2817-2823. https://doi.org/10.1093/bioinformatics/btw327

Biagini, S. A., Solé-Morata, N., Matisoo-Smith, E., Zalloua, P., Comas, D., & Calafell, F. (2019). People from Ibiza: An unexpected isolate in the Western Mediterranean. *European Journal of Human Genetics*, *27*(6), 941-951. https://doi.org/10.1038/s41431-019-0361-1

Cardona, A., Pagani, L., Antao, T., Lawson, D. J., Eichstaedt, C. A., Yngvadottir, B., Shwe, M. T. T., Wee, J., Romero, I. G., Raj, S., Metspalu, M., Villems, R., Willerslev, E., Tyler-Smith, C., Malyarchuk, B. A., Derenko, M. V., & Kivisild, T. (2014). Genome-Wide Analysis of Cold Adaptation in Indigenous Siberian Populations. *PLOS ONE*, *9*(5), e98076. https://doi.org/10.1371/journal.pone.0098076

Chen, L., & Qin, Z. S. (2016). traseR: An R package for performing trait-associated SNP enrichment analysis in genomic intervals. *Bioinformatics (Oxford, England)*, *32*(8), 1214-1216. https://doi.org/10.1093/bioinformatics/btv741

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., & Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156-2158. https://doi.org/10.1093/bioinformatics/btr330

Delaneau, O., Coulonges, C., & Zagury, J.-F. (2008). Shape-IT: New rapid and accurate algorithm for haplotype inference. *BMC Bioinformatics*, *9*(1), 540. https://doi.org/10.1186/1471-2105-9-540

Gautier, M., & Vitalis, R. (2012). rehh: An R package to detect footprints of selection in genome-wide SNP data from haplotype structure. *Bioinformatics*, *28*(8), 1176-1177. https://doi.org/10.1093/bioinformatics/bts115

Ghoreishifar, S. M., Moradi-Shahrbabak, H., Fallahi, M. H., Jalil Sarghale, A., Moradi-Shahrbabak, M., Abdollahi-Arpanahi, R., & Khansefid, M. (2020). Genomic measures of inbreeding coefficients and genome-wide scan for runs of homozygosity islands in Iranian river buffalo, Bubalus bubalis. BMC Genetics, 21(1), 16. https://doi.org/10.1186/s12863-020-0824-y

Islam, R., Li, Y., Liu, X., Berihulay, H., Abied, A., Gebreselassie, G., Ma, Q., & Ma, Y. (2019). Genome-Wide Runs of Homozygosity, Effective Population Size, and Detection of Positive Selection Signatures in Six Chinese Goat Breeds. Genes, 10(11). https://doi.org/10.3390/genes10110938Jordan, M. A., & Snell, H. L. (2008). Historical fragmentation of islands and genetic drift in populations of Galápagos lava lizards (Microlophus albemarlensis complex). Molecular Ecology, 17(5), 1224-1237. https://doi.org/10.1111/j.1365-294X.2007.03658.x

Kurtzhals, J. A. L., Reimert, C. M., Tette, E., Dunyo, S. K., Koram, K. A., Akanmori, B. D., Nkrumah, F. K., & Hviid, L. (1998). Increased eosinophil activity in acute Plasmodium falciparum infection—Association with cerebral malaria. Clinical and Experimental Immunology, 112(2), 303-307. https://doi.org/10.1046/j.1365-2249.1998.00586.x

Lawson, D. J., van Dorp, L., & Falush, D. (2018). A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. Nature Communications, 9(1), 3258. https://doi.org/10.1038/s41467-018-05257-7Lorenzo-Lacruz, J., & Morán-Tejeda, E. (2016). Spatio-temporal patterns of meteorological droughts in the Balearic Islands (Spain). Cuadernos de Investigación Geográfica, 42(1), 49. https://doi.org/10.18172/cig.2948

Macpherson, J. M., González, J., Witten, D. M., Davis, J. C., Rosenberg, N. A., Hirsh, A. E., & Petrov, D. A. (2008). Nonadaptive Explanations for Signatures of Partial Selective Sweeps in Drosophila. Molecular Biology and Evolution, 25(6), 1025-1042. https://doi.org/10.1093/molbev/msn007

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., & Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics*, *26*(22), 2867-2873. https://doi.org/10.1093/bioinformatics/btq559

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., & Reich, D. (2012). Ancient Admixture in Human History. *Genetics*, *192*(3), 1065-1093. https://doi.org/10.1534/genetics.112.145037

Pickrell, J. K., Coop, G., Novembre, J., Kudaravalli, S., Li, J. Z., Absher, D., Srinivasan, B. S., Barsh, G. S., Myers, R. M., Feldman, M. W., & Pritchard, J. K. (2009). Signals of recent positive selection in a worldwide sample of human populations. *Genome Research*, *19*(5), 826-837. https://doi.org/10.1101/gr.087577.108

Picornell, A., Miguel, A., Castro, J. A., Ramon, M. M., Arya, R., & Crawford, M. H. (1996). Genetic Variation in the Population of Ibiza (Spain): Genetic Structure, Geography, and Language. Human Biology, 68(6), 899-913.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, *81*(3), 559-575. https://doi.org/10.1086/519795

Ramon, M., Picornell, A., & Castro, J. A. (2008). Human population of the Balearic Island: The case of Chuetas and Ibizans. Contributions to Science, 4, 85-91. https://doi.org/10.2436/20.7010.01.39

Roncal-Jimenez, C., Lanaspa, M. A., Jensen, T., Sanchez-Lozada, L. G., & Johnson, R. J. (2015). Mechanisms by Which Dehydration May Lead to Chronic Kidney Disease. Annals of Nutrition and Metabolism, 66(Suppl. 3), 10-13. https://doi.org/10.1159/000381239

R Core Team. (2017). *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.*

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z. P., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., McDonald, G. J., Ackerman, H. C., Campbell, S. J., Altshuler, D., Cooper, R., Kwiatkowski, D., Ward, R., & Lander, E. S. (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature*, *419*(6909), 832-837. https://doi.org/10.1038/nature01140

Vitti, J. J., Grossman, S. R., & Sabeti, P. C. (2013). Detecting Natural Selection in Genomic Data. *Annual Review of Genetics*, *47*(1), 97-120. https://doi.org/10.1146/annurev-genet-111212-133526

Voight, B. F., Kudaravalli, S., Wen, X., & Pritchard, J. K. (2006). A Map of Recent Positive Selection in the Human Genome. *PLoS Biology*, *4*(3). https://doi.org/10.1371/journal.pbio.0040072

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, *38*(16), e164-e164. https://doi.org/10.1093/nar/gkq603

Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z. X. P., Pool, J. E., Xu, X., Jiang, H., Vinckenbosch, N., Korneliussen, T. S., Zheng, H., Liu, T., He, W., Li, K., Luo, R., Nie, X., Wu, H., Zhao, M., Cao, H., … Wang, J. (2010). Sequencing of Fifty Human Exomes Reveals Adaptation to High Altitude. *Science (New York, N.Y.)*, *329*(5987), 75-78. https://doi.org/10.1126/science.1190371

TABLES

Table 1

Genes present in the top 5 candidate windows for positive selection after analysis with PBS when comparing EIV versus IBE and using GBR as external reference. Colors indicate different windows where the genes are found. Terms in bold indicate common traits found across tests.

PBS								
	Name	**Traits**						
APOA1-AS Chr 11 - Window 61728	Apoa1 antisense rna - lncrna	Tangier disease: reduced levels of hdl in the blood.	Risk of **cardiovascular disease**	Elevated amount of **fat** in the blood (mild hypertriglyceridemia)			
APOA4 Chr 11 - Window 61727	Apolipoprotein a4	Function unknown	Lecithin-**cholesterol** acyltransferase in vitro	Carotenemia: yellow **pigmentation of the skin** and increased beta-carotene levels in the blood	**High levels of cholesterol (ldl) in the blood**		
APOC3 Chr 11 - Window 61727	Apolipoprotein c3	Protein component of triglyceride (tg)-rich **lipoproteins** (trls)	Inhibit lipoprotein lipase enzyme activity	**Coronary heart disease 1**			
APOA5 Chr 11 - Window 61727	Apolipoprotein a5	Important role in regulating the plasma **triglyceride** levels, a major risk factor for coronary artery disease	Hypertriglyceridemia and hyperlipoproteinemia type 5				
TM4SF20 Chr 2 - Window 15543	Transmembrane 4 l six family member 20	Member of the four-transmembrane l6 superfamily - function in various cellular processes	Cell proliferation, motility, and adhesion via their interactions with integrins	Knockout of the homologous gene in mice results in neurobehavioral enhanced **motor coordination**. A deletion mutation in the human gene is associated with specific language impairment-5	Collagen synthesis	reading and spelling ability in gwas	
MIR5703 Chr 2 - Window 15543	Mir5703 gene mirna						
MFF Chr 2 - Window 15540	Mitochondrial fission factor	mitochondrial and peroxisomal fission	May be involved in regulation of synaptic vesicle membrane dynamics	**Intraocular pressure measurement**	Astigmatism	Reading and spelling ability	**Acute myeloid leukemia**
LOC654841 Chr 2 - Window 15540	Ncrna associated with mff	Mff-dt gene lncrna	**Intraocular pressure measurement**	**Central corneal thickness**	Resting heart rate	Type i diabetes mellitus, type 1 diabetes nephropathy	**Albuminuria**, type i diabetes mellitus, type 1 diabetes nephropathy
GCFC2 Chr 2 - Window 10439	GC-Rich Sequence DNA-Binding Factor 2	Chronotype measurement	Acute myeloid leukemia	Hippocampal atrophy	Isocitrate measurement	Mental deterioration	

LRRTM4 Chr 2 - Window 10439	Leucine Rich Repeat Transmembrane Neuronal 4	Chronotype measurement	Acute myeloid leukemia	**Peripheral arterial disease, traffic air pollution measurement**	Risky sexual behaviour measurement	Insomnia measurement	
AGFG1 Chr 2 - Window 15548	ArfGAP With FG Repeats 1	Childhood onset asthma	**Body height – rs6741427 pbs: 0.36 (> perc 99)**	Ulcerative colitis			
C2ORF83 Chr 2 - Window 15548	Chromosome 2 Open Reading Frame 83	**Urinary albumin to creatinine ratio**	**Albuminuria**	Gut microbiome measurement, taxonomic microbiome measurement	Asthma	Response to vaccine, cytokine measurement	

Table 2

Genes present in the top 5 candidate windows for positive selection after XP-EHH analysis in EIV when compared to IBE. Colors indicate different windows where the genes are found. Terms in bold indicate common traits found across tests.

XP-EHH							
	Name	**Traits**					
LINC01506 Chr 9 – Window 51105	Long intergenic non-protein coding rna 1506	**Glomerular filtration rate**	**Erythrocyte count**				
PIP5K1B Chr 9 – Window 51106	Phosphatidylinositol-4-phosphate 5-kinase type 1 beta	Friedreich ataxia: affects the nervous system and causes movement problems. People with this condition develop impaired **muscle coordination** (ataxia)	Loss of strength and sensation in the arms and legs; muscle stiffness (spasticity); and impaired speech, hearing, and vision.	Individuals with friedreich ataxia often have a form of heart disease called hypertrophic cardiomyopathy, which enlarges and weakens the heart muscle and can be life-threatening.	**Glomerular filtration rate**	**Creatinine measurement, chronic kidney disease**	**Erythrocyte count**
OCA2 Chr 15 – Windows 73338-73339-73340	Oca2 melanosomal transmembrane protein	Believed to be an integral membrane protein involved in small molecule transport, specifically tyrosine, which is a precursor to melanin synthesis.	It is involved in **mammalian pigmentation,** where it may control **skin color variation** and act as a determinant of brown or blue eye color	Albinism, oculocutaneous, type ii and **skin/hair/eye pigmentation**, variation	**Hair color, eye color, suntan, sunburn**		
ACOXL Chr 2 – Windows 11492-11493	Acyl-coenzyme a oxidase-like protein	retinitis pigmentosa. Among its related pathways are metabolism and peroxisomal **lipid metabolism**	Retinal pigment deposits visible on fundus examination and primary loss of rod **photoreceptor cells** followed by secondary loss of cone photoreceptors. Patients typically have night vision **blindness** and loss of midperipheral visual field	Mean corpuscular hemoglobin - mean corpuscular volume	**Eosinophil count**	**Erythrocyte count - rs56088557 present – xp-ehh:** 3.7857 **(> perc 99)**	Monocyte percentage of leukocytes

NEPRO Chr 3 – Windows 20333-20334	Nucleolus and neural progenitor protein	diseases associated with nepro include anauxetic dysplasia 3 and anauxetic dysplasia 1.	**Severe short stature,** brachydactyly, **skin laxity**, joint hypermobility, and joint dislocations.	Radiographs show short metacarpals, broad middle phalanges, and metaphyseal irregularities. Most patients also exhibit **motor and cognitive delays**	**Eosinophil count**	Diabetic nephropathy	Economic and social preference
LINC02044 Chr 3 – Windows 20333-20334	Lncrna	**Body height – rs2700201 xp-ehh: 4.95 (> perc 99)**	Sleep duration, low-density **lipoprotein cholesterol measurement**	Idiopathic dilated cardiomyopathy	Alzheimer's disease, psychotic symptoms		
FMO5 Chr 1 - Window 3979	Flavin containing dimethylaniline monoxygenase 5	**Regulation of cholesterol metabolic process**	Mitochondrial dna measurement	Aldehyde oxidase activity	Monooxygenase activity		
NBPF19 Chr 1 - Window 3979	Nbpf member 19	**Uric acid measurement**					

Table 3

Genes present in the top 5 candidate windows for positive selection after iHS analysis in EIV. Colors indicate different windows where the genes are found. Terms in bold indicate common traits found across tests.

IHS								
	Name	**Traits**						
NSUN3 Chr 3 – Windows 15236-15237-15238	Nop2/sun rna methyltransferase 3	**Cardiomyopathy**, familial restrictive, 1 and **neuropathy, hereditary motor and sensory**, type via, with **optic atrophy**	**Body mass index**	**Visceral adipose tissue measurement**	Mathematical ability	Alcohol consumption measurement	Smoking status measurement	
MIR6730 Chr 3 – Windows 15236-15237-15238	Mirna	Bitter beverage consumption measurement						
HCG18 Chr 6 – Windows 26390-26391	Incrna	**Psoriasis**	**Cutaneous psoriasis measurement, psoriasis**	Psoriatic arthritis	**Myeloid white cell count**	**Leukocyte count**		
HCG17 Chr 6 – Windows 26390-26391	Incrna	**Reticulocyte count**	Ankylosing spondylitis	Complement c4 measurement	**Blood protein measurement**	**Schizophrenia, autism spectrum disorder**		
TRIM39 Chr 6 – Window 26391	Tripartite motif containing 39	Trim motif includes three zinc-binding domains	The function of this protein has not been identified.	This gene lies within the major histocompatibility complex class i region on chromosome 6	**Psoriasis rs3130453 his: 1.8 (> perc 95)**	Cutaneous psoriasis measurement, psoriasis	**Reticulocyte count**	**Eosinophil count, basophil count**
TRIM39-RPP21 Chr 6 – Window 26391	Rpp21 domain-containing trim protein	Related pathways are innate immune system and class i mhc mediated	**Eosinophil count**	**Basophil count, eosinophil count**	Eosinophil percentage of leukocytes	Eosinophil percentage of granulocytes		**Schizophrenia, autism**

		antigen processing and presentation						**spectrum disorder**
HLA-E Chr 6 – Windows 26391	Major histocompatibility complex, class i, e	**Eosinophil percentage of leukocytes**	**Membranous glomerulonephritis – rs1265159 iHS 1.65 (> perc 95)**	Eosinophil percentage of granulocytes	Psoriatic arthritis	**Psoriasis**		
PIP5K1B chr9 - Windows 39329-39331-.39334	Phosphatidylinositol-4-phosphate 5-kinase type 1 beta	Friedreich ataxia: affects nervous system and causes movement problems. People with this condition develop impaired muscle coordination (ataxia)	Loss of strength and sensation in the arms and legs; muscle stiffness (spasticity); and impaired speech, hearing, and vision.	Individuals with friedreich ataxia often have a form of heart disease called hypertrophic cardiomyopathy, which enlarges and weakens the heart muscle and can be life-threatening.	Glomerular filtration rate	**Creatinine measurement, chronic kidney disease**	**Erythrocyte count**	
GATB chr4 - Windows 21164-21165-21166	Glutamyl-trna amidotransferase subunit b	Cognitive function measurement	Type ii diabetes mellitus	Intelligence	**Platelet count**	**Leukocyte count**		
LINC02273 chr4 - Windows 21164-21165-21166	Long intergenic non-protein coding rna 2273	Alcohol consumption measurement	**Body mass index**	Smoking status measurement	Acute myeloid leukemia	**Diet measurement**		

Table 4

Genes shared between candidate windows for positive selection identified in EIV with PBS and XP-EHH. Colors indicate different windows where the genes are found. Terms in bold indicate common traits found across tests.

Shared PBS/XP-EHH									
	Name	**Traits**							
ARL4A Chr 7	Adp ribosylation factor like gtpase 4a	**Body height**	Adolescent idiopathic scoliosis **rs12538763 xp=2.6**	Type II diabetes mellitus	Tea consumption measurement	Response to diisocyanate, asthma			

ETV1 Chr 7	Ets variant transcription factor 1	Ets proteins regulate many target genes that modulate biological processes like cell growth, angiogenesis, migration, proliferation and differentiation	Diseases associated with etv1 include ewing sarcoma and gastrointestinal stromal tumor. Among its related pathways are transcriptional misregulation in cancer	Acute myeloid leukemia	**Hair colour measurement**	**Cup-to-disc ratio measurement**	Atrial fibrillation	Hair color	
PTPRZ1 Chr 7	Protein tyrosine phosphatase receptor type z1	Expression of this gene is restricted to the central nervous system (cns), and it may be involved in the regulation of specific developmental processes in the cns	Diseases associated with ptprz1 include generalized epilepsy with febrile seizures plus, type 1 and oligodendroglioma.	Among its related pathways are Insulin receptor recycling and PAK Pathway.	Heel bone mineral density	**Body height**	Smoking status measurement, high **density lipoprotein cholesterol** measurement	Bone mineral content measurement	Grip strength measurement
AASS Chr 7	Aminoadipate-semialdehyde synthase	Bifunctional enzyme that catalyzes the first two steps in the mammalian lysine degradation pathway	Heel bone mineral density	**Schizophrenia**	Reaction time measurement				
FEZF1 Chr 7	Fez family zinc finger 1	Role in the embryonic migration of gonadotropin-releasing hormone neurons into the brain. Mutations in this gene are associated with hypogonadotropic hypogonadism-22 with anosmia	Age at menarche	Chronotype measurement	Smoking status measurement	**Body height**	Smoking behavior		
GLIS3 Chr 9	Glis family zinc finger 3	Mutations in this gene have been associated with neonatal diabetes and congenital hypothyroidism (ndh	Diabetes mellitus, neonatal, with congenital hypothyroidism and congenital hypothyroidism	**Intraocular pressure measurement**	FEV/FEC ratio	**Erythrocyte count**	Forced expiratory volume	Chin morphology measurement	

Table 5

Genes shared between candidate windows for positive selection identified in EIV with XP-EHH and iHS. Colors indicate different windows where the genes are found. Terms in bold indicate common traits found across tests.

Shared XP-EHH/iHS							
	Name	**Traits**					
VWA3B Chr 2	Von Willebrand Factor A Domain Containing 3B	Diseases associated with VWA3B include Spinocerebellar Ataxia, Autosomal Recessive 22 and Pylorospasm	Self reported educational attainment	**Triglyceride measurement**	Adverse effect, response to drug	Gestational age, birth measurement	Bipolar disorder
C3ORF38 Chr 3	Chromosome 3 Open Reading Frame 38	**Mean corpuscular hemoglobin**	**Body mass index**	Age at menarche	**Cup-to-disc ratio measurement**	Grip strength measurement	
EPHA3 Chr 3	EPH Receptor A3	**Cup-to-disc ratio measurement**	**Waist-hip ratio**	Mathematical ability	Atrial fibrillation	Brain volume measurement	
GATB Chr 4	Glutamyl-TRNA Amidotransferase Subunit B	Cognitive function measurement	Type II diabetes mellitus	Intelligence	**Platelet count**	**Leukocyte count**	
LINC02273 Chr 4	Long Intergenic Non-Protein Coding RNA 2273	Alcohol consumption measurement	**Body mass index**	Smoking status measurement	**Acute myeloid leukemia**	**Diet measurement**	
MSX2P1 Chr 17	Msh Homeobox 2 Pseudogene 1	Intelligence	**Cardiovascular disease**	Prostate carcinoma	Agents acting on the renin-angiotensin system use measurement	**Schizophrenia**	
OR4D2 Chr 17	Olfactory Receptor Family 4 Subfamily D Member 2	**Monocyte count**	**Platelet crit**	Alzheimer's disease	**Red blood cell distribution width**	**Waist-hip ratio**	
EPX Chr 17	Eosinophil Peroxidase	Neutrophil count	Neutrophil count, basophil count	**Neutrophil count, eosinophil count**	Myeloid white cell count	Granulocyte count	

Table 6

Results from TraseR for traits associated with the candidate windows for positive selection. LD means that SNPs in LD with the taSNPs are also considered for the enrichment.

PBS					
Trait_LD					
Trait	p.value	q.value	odds.ratio	taSNP.hits	taSNP.num
Pulse	1.26E-14	7.22E-12	16.37	17	266
Respiratory Function Tests	5.13E-10	1.47E-07	5.20	24	1115
Lupus Erythematosus, Systemic	7.84E-09	1.50E-06	8.59	14	410
Antidepressive Agents	3.32E-07	4.75E-05	29.46	6	59
Forced Vital Capacity	1.08E-06	1.24E-04	8.77	10	294
Receptors, Transferrin	3.82E-05	3.64E-03	681.88	2	2
Triglycerides	9.67E-05	7.91E-03	2.69	22	1968
Tunica Media	2.31E-04	1.65E-02	3.85	12	779

Erectile Dysfunction	3.74E-04	2.38E-02	15.78	4	75
Body Weight Changes	7.85E-04	4.50E-02	6.74	6	241
XP-EHH					
Trait_LD					
Trait	p.value	q.value	odds.ratio	taSNP.hits	taSNP.num
Blood Coagulation Factor Inhibitors	3.11E-17	1.78E-14	181.37	10	22
Body Weight and Body Measures	8.28E-09	2.37E-06	6.11	18	683
Parietal Lobe	8.26E-07	1.58E-04	41.49	5	35
Psoriasis	1.86E-06	2.28E-04	11.83	8	170
Prostatic Neoplasms	1.99E-06	2.28E-04	5.33	14	616
Vaccination	9.61E-05	8.24E-03	321.51	2	3
Leukemia, Lymphocytic, Chronic, B-Cell	1.01E-04	8.24E-03	22.80	4	50
iHS					
Trait_LD					
Trait	p.value	q.value	odds.ratio	taSNP.hits	taSNP.num
Behcet Syndrome	4.15E-57	2.38E-54	82.84	39	274
Psoriasis	2.19E-16	6.27E-14	52.95	12	131
Lupus Erythematosus, Systemic	2.88E-12	5.50E-10	27.53	11	223
Multiple Sclerosis	2.51E-09	3.59E-07	21.44	9	236
Uric Acid	7.58E-09	8.69E-07	34.91	7	118
Stevens-Johnson Syndrome	9.03E-08	8.62E-06	977.45	3	4

Table 7

Common categories found with TraseR across the tests.

- Blood Function, Cardiovascular Diseases
- Skin and Connective Tissue Diseases
- Body Weight and Body Measures
- Kidney Function / Urogenital System
- Immune System Diseases
- Nervous System and Mental Diseases

FIGURES

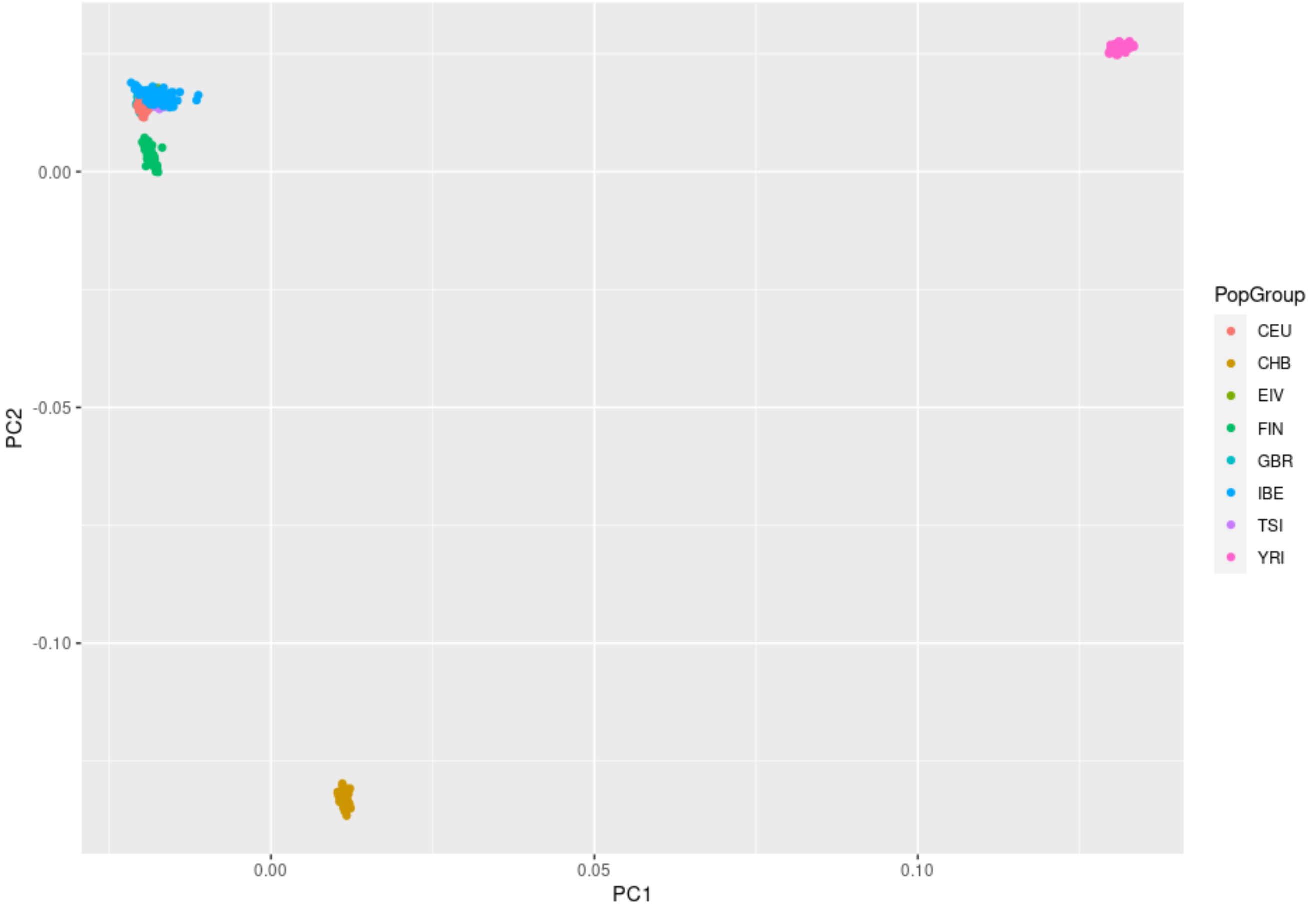


Fig. 1

Principal Component Analysis for all populations. PC1 explains 41% of the variation, whereas PC2 explains 18.66%. CEU, Utah Residents with Northern and Western European ancestry; CHB, Han Chinese; EIV, Ibizans; FIN, Finns; GBR samples from Great Britain; IBE, Iberians; TSI, Toscani from Italy; YRI, Yoruba from Nigeria.

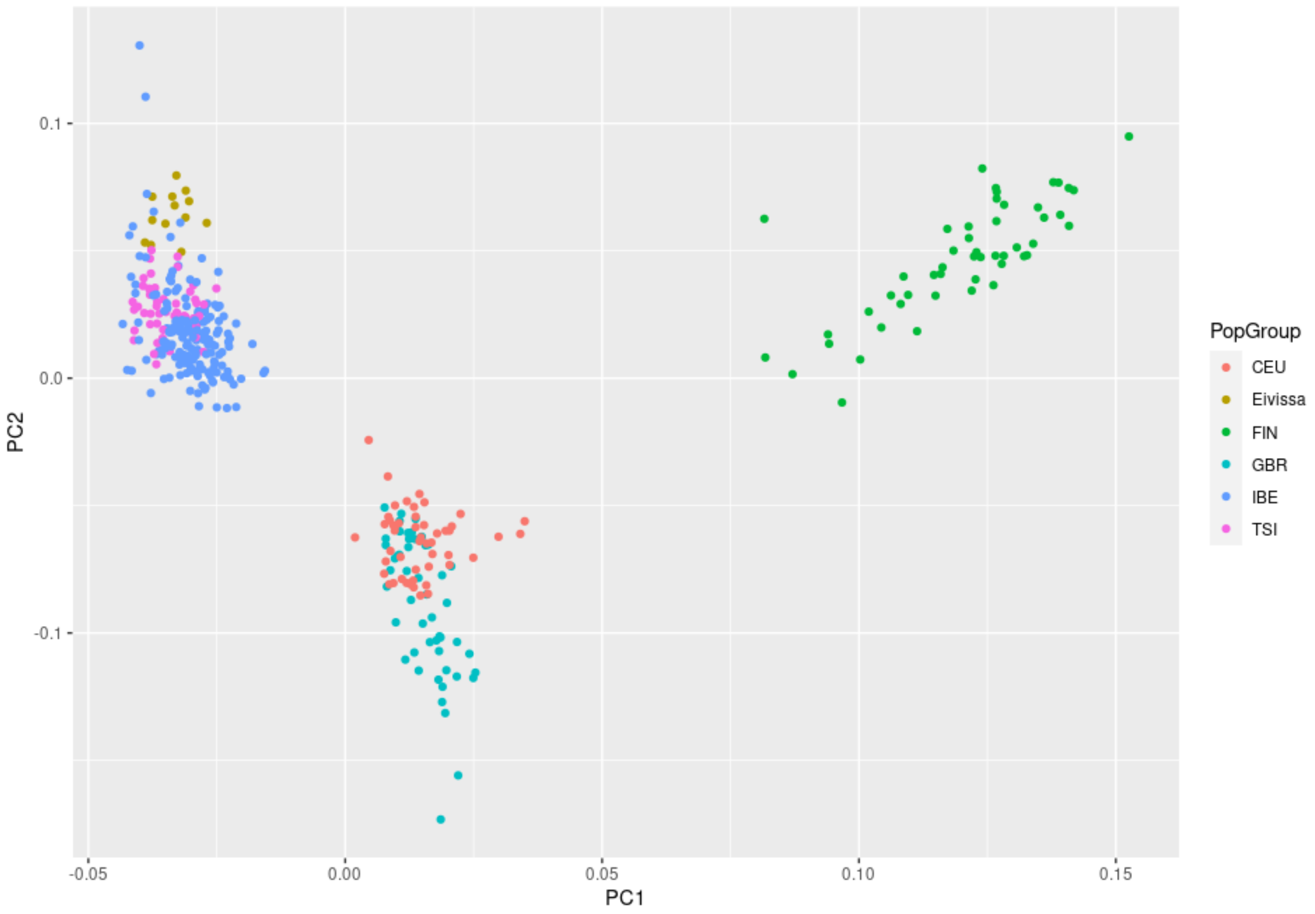


Fig. 2

Principal Component Analysis for populations of European Ancestry. PC1 explains 2.77% of the variation, whereas PC2 explains 1.46%. CEU, Utah Residents with Northern and Western European ancestry; FIN, Finns; GBR samples from Great Britain; IBE, Iberians; TSI, Toscani from Italy.

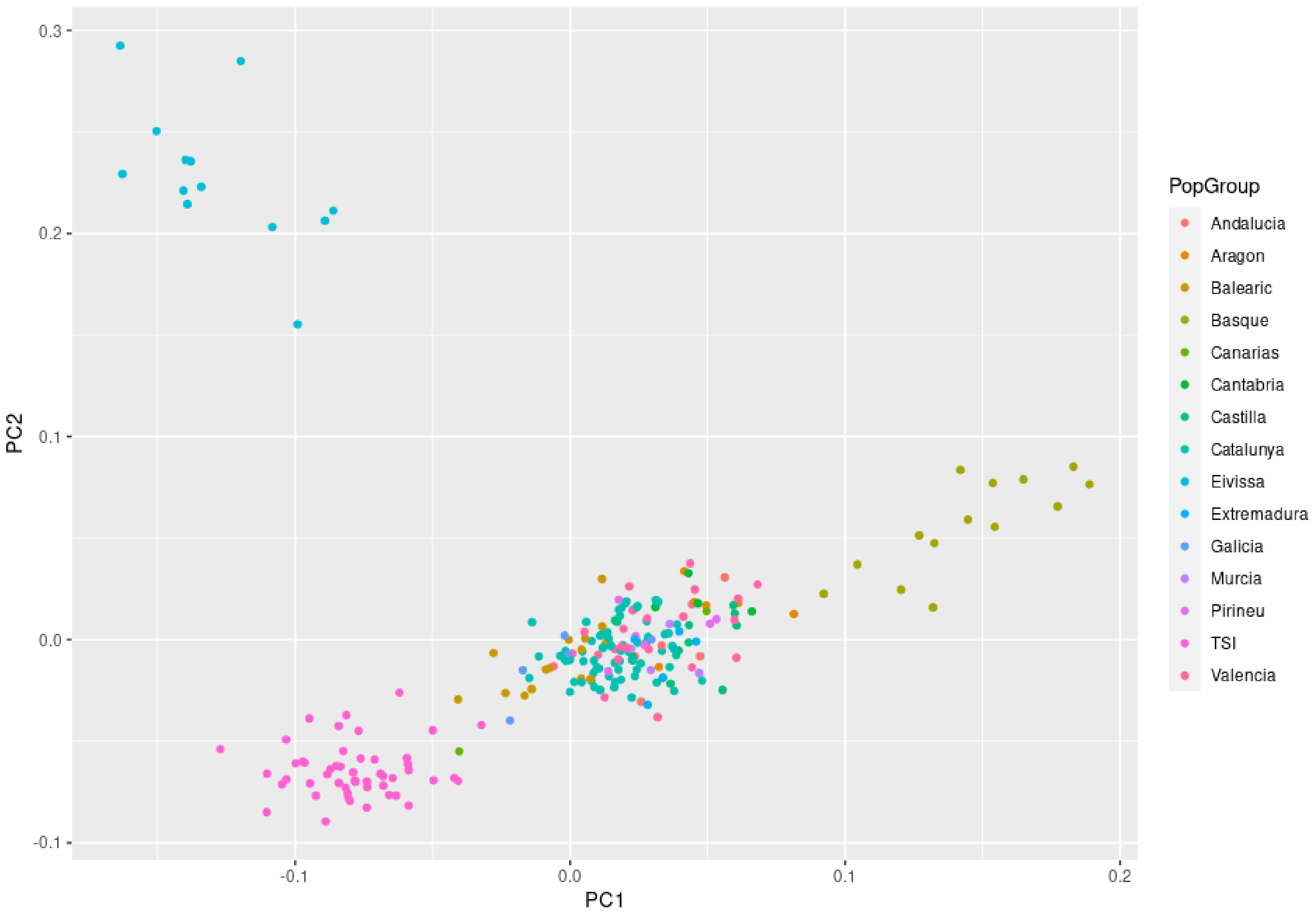


Fig. 3

Principal Component Analysis for South European populations. PC1 explains 1.30% of the variation, whereas PC2 explains 1.29%. TSI, Toscani from Italy.

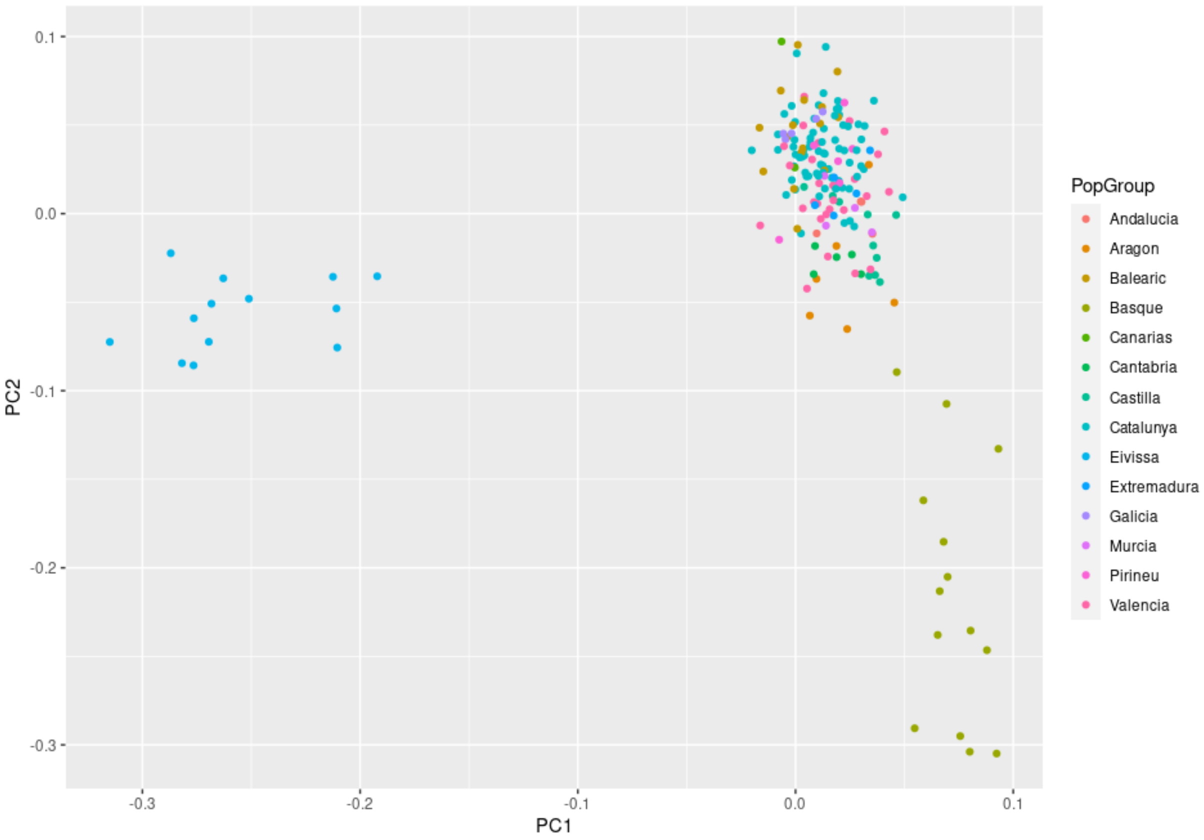


Fig. 4

Principal Component Analysis for Spanish populations (EIV, and IBE divided by localities). PC1 explains 2.77% of the variation, whereas PC2 explains 1.46%.

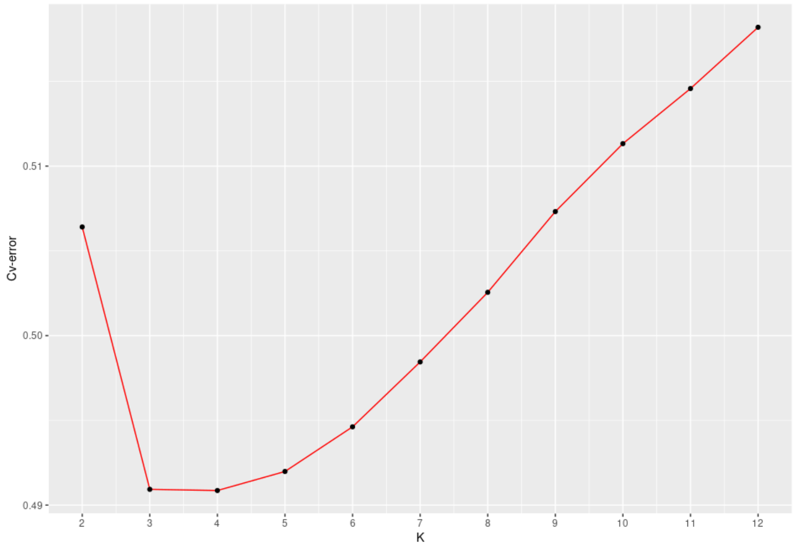


Fig. 5

Cross-validation error as a function of the tested Ks. K2 and K3 are the best models to explain the patterns of admixture in the populations. The difference between the minimum and maximum cross-validation error is less than 0.03, suggesting that the different models of admixture are not too different in power.

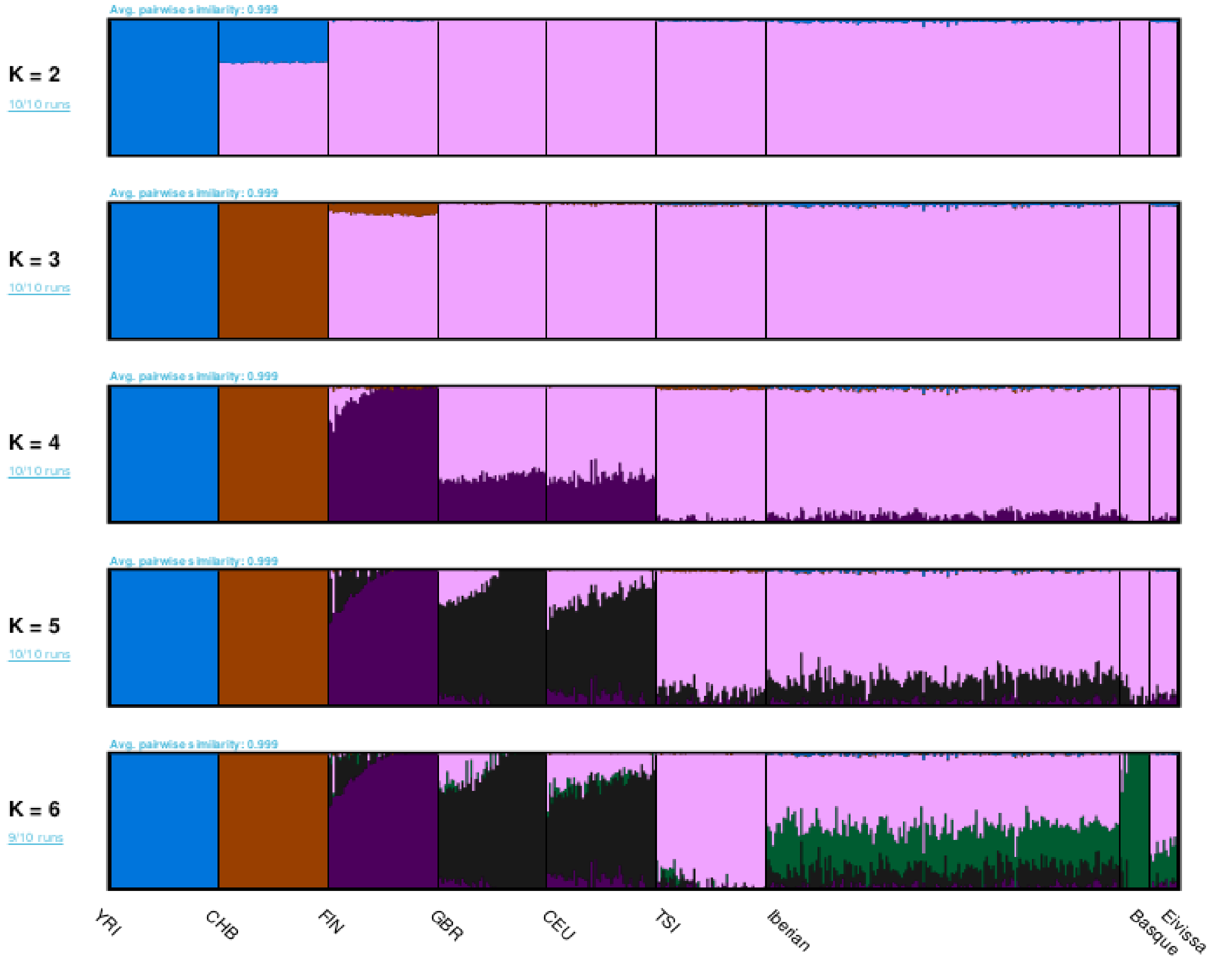


Fig. 6

Admixture analyses for K = 2-6 considering the structure found in the PCA (Iberian subpopulations clustered together). The software pong was used to visualize the maximum-weight alignments between 10 runs.

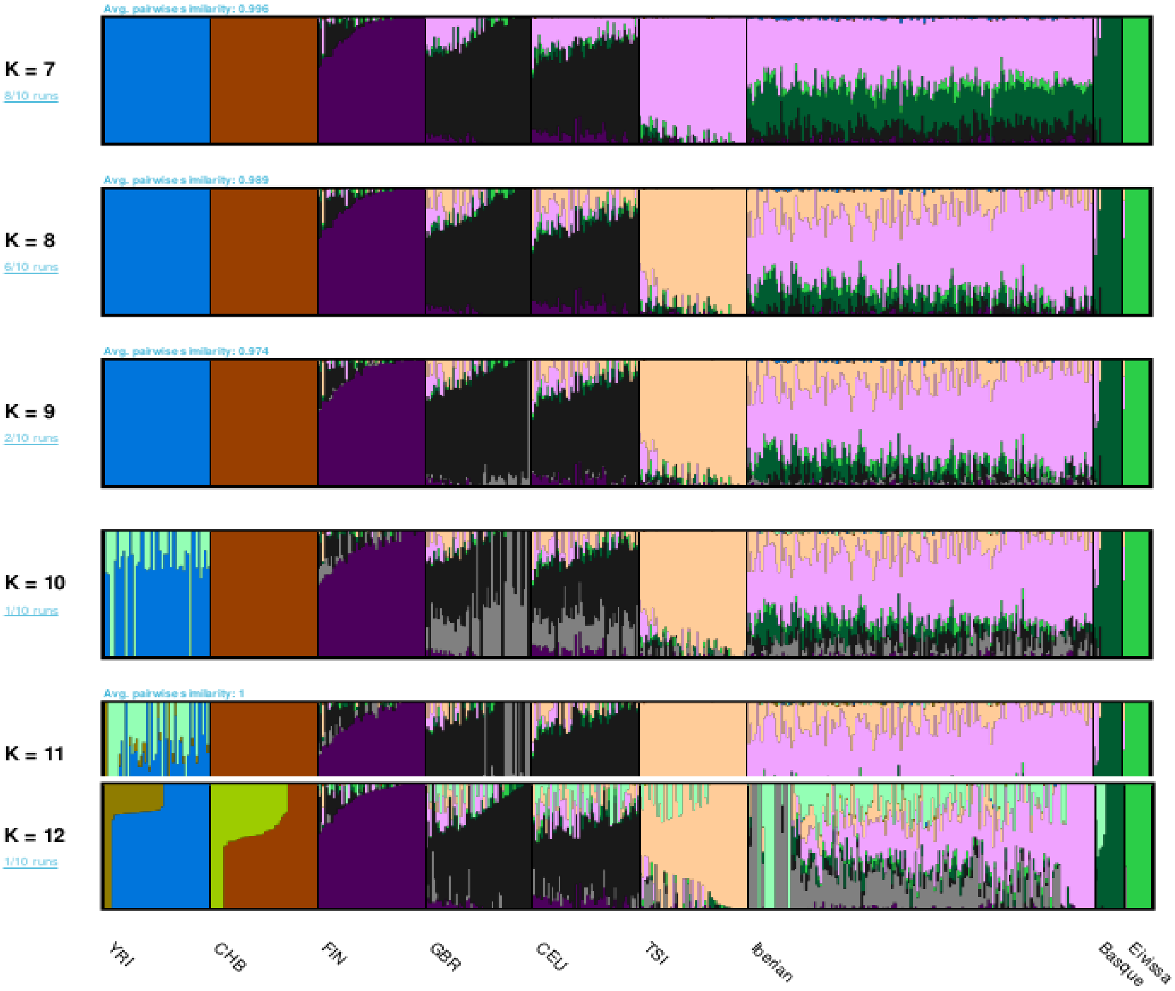


Fig. 7

Admixture analyses for K = 7-12 considering the structure found in the PCA (initial 8 + Basques who also clustered apart). The software pong was used to visualize the maximum-weight alignments between 10 runs.

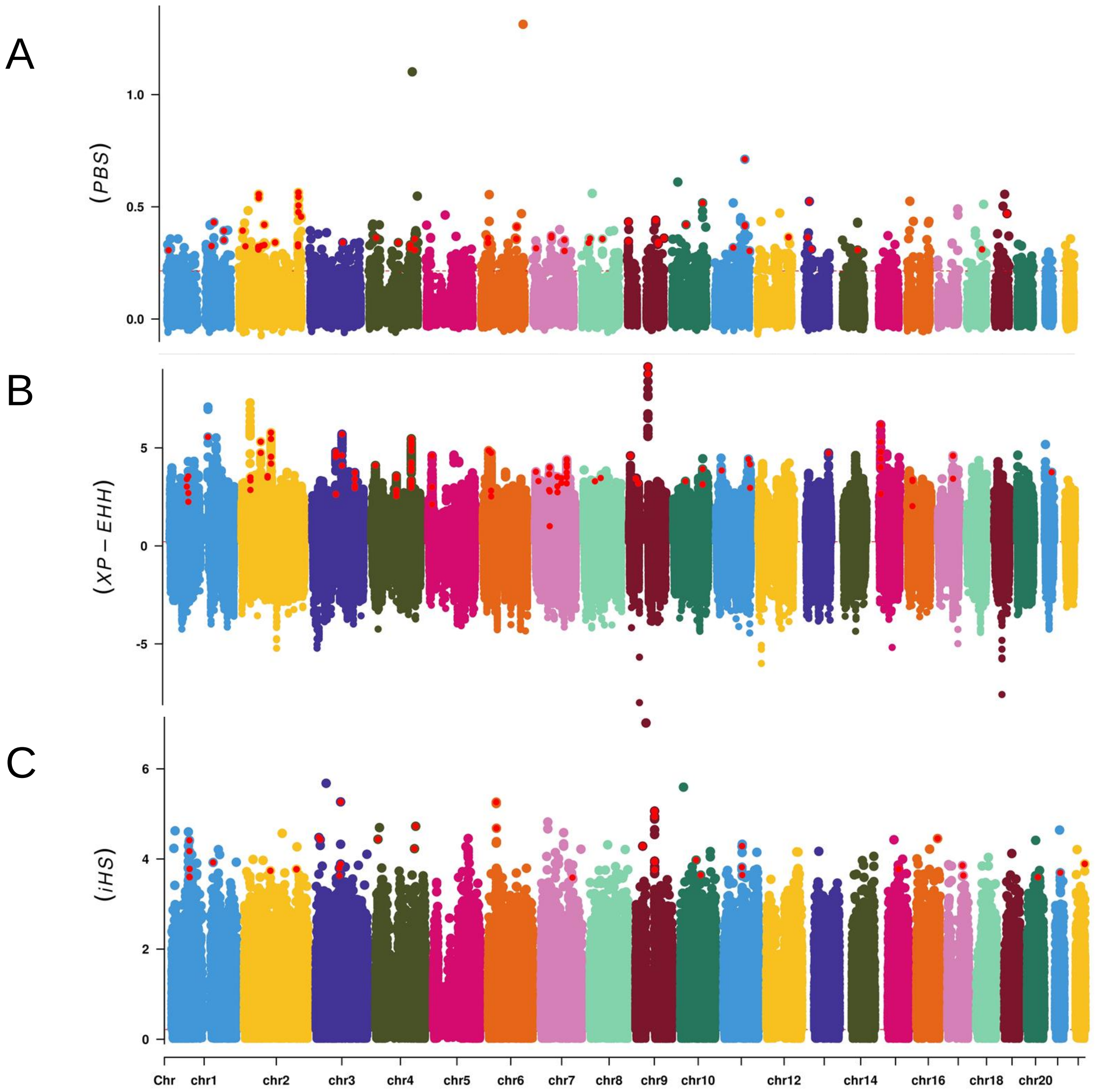
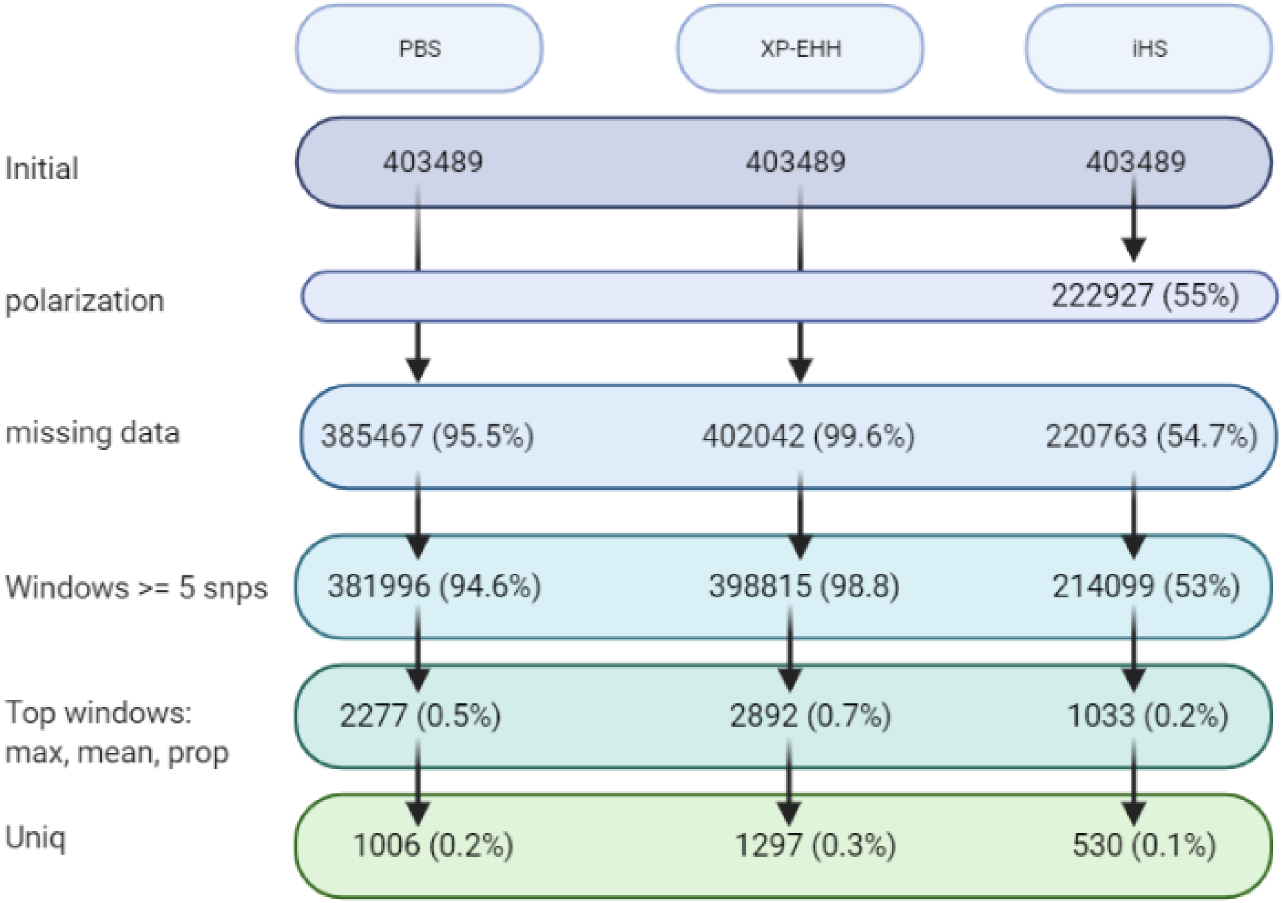


Fig. 8

Manhattan plots of genome-wide signatures of positive selection in EIV. SNPs with the highest score within each window are showed in red for the candidate top windows for positive selection identified by using the intersection of the three filtering criteria: Max, Mean, and Prop**.** **(A)** From a total of 385,467 SNPs across the genome, 1,006 candidate SNPs were identified with the PBS when comparing EIV to IBE and using GBR as external population. **(B)** From a total of 402,042 genome-wide SNPs, 1,297 candidate SNPs were identified when comparing EIV (positive values in plot) versus IBE (negative values in plot) with XP-EHH **(C)** From a total of 220,763 genome-wide SNPs, 530 candidate SNPs were identified when analyzing the iHS in EIV.



	PBS	XP-EHH	iHS
Initial	403489	403489	403489
polarization			222927 (55%)
missing data	385467 (95.5%)	402042 (99.6%)	220763 (54.7%)
Windows >= 5 snps	381996 (94.6%)	398815 (98.8)	214099 (53%)
Top windows: max, mean, prop	2277 (0.5%)	2892 (0.7%)	1033 (0.2%)
Uniq	1006 (0.2%)	1297 (0.3%)	530 (0.1%)

Fig. 9

For each test, number of SNPs that remained after each step of the analysis after starting with the same number of SNPs as input (403,489). Percentages correspond to the comparison between the number at each particular step and the initial number of SNPs. Polarization was only necessary for iHS. The missing data step corresponds to the step where the tests were performed but information from some SNPs was lost. As the windows were overlapping, some SNPs were repeated, after taking all the selected windows together; these duplicates were removed at the Uniq step.

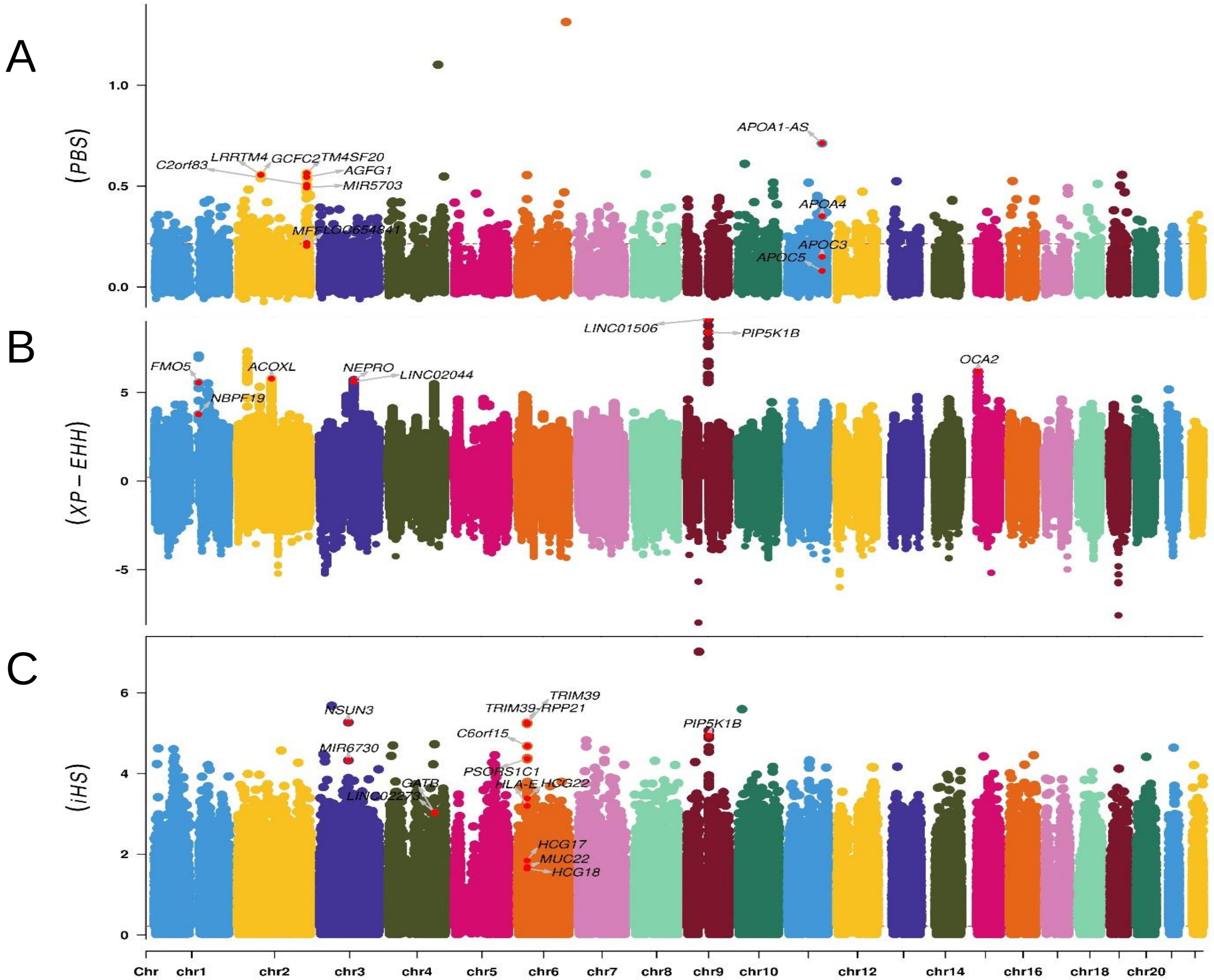


Fig. 10

Manhattan plots of genome-wide signatures of positive selection in EIV with genes in the top 5 regions. SNPs from the top 5 candidate regions are showed in red together with the genes associated with them. **(A)** 12 genes associated with the top 5 regions found with PBS **(B)** 8 genes associated with the top 5 regions found with XP-EHH **(C)** 10 genes associated with the top 5 regions found with his.