Interrogating eleven fast-evolving genes for signatures of recent positive selection in worldwide human populations

Submitted as Research Article

Andrés Moreno-Estrada[1], Kun Tang[2,3], Martin Sikora[1], Tomàs Marquès-Bonet[1,4], Ferran Casals[5], Arcadi Navarro[1,6], Francesc Calafell[1,7], Jaume Bertranpetit[1,7], Mark Stoneking[3] and Elena Bosch[1,7]

[1] Institut de Biologia Evolutiva (UPF-CSIC), Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Barcelona, Spain.

[2] CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China

[3] Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany.

[4] Department of Genome Sciences, University of Washington, Seattle, USA.

[5] Ste Justine Hospital Research Centre, Department of Pediatric, Faculty of Medicine, University of Montreal, Montreal, Quebec H3T 1C5, Canada

[6] Institució Catalana de Recerca i Estudis Avançats (ICREA) i UPF

[7] Centro de Investigación Biomédica en Red de Epidemiología y Salud Pública (CIBERESP), Barcelona, Spain.

Corresponding author: Elena Bosch, Institut de Biologia Evolutiva (UPF-CSIC), Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra,
C/ Dr. Aiguader 88, 08003 Barcelona, Spain.
Tel. +34 93 316 0841
Fax. +34 93 316 0901
E-mail: elena.bosch@upf.edu

Key words: Accelerated Evolution, Recent Positive Selection, SNP Data, Extended Haplotype Homozygosity, Population Differentiation, Human Genome Diversity Panel.

Running head: Recent selection on fast-evolving genes

Abstract

Different signatures of natural selection persist over varying time scales in our genome, revealing possible episodes of adaptative evolution during human history. Here, we identify genes showing signatures of ancestral positive selection in the human lineage and investigate whether some of those genes have been evolving adaptatively in extant human populations. Specifically, we compared more than 11,000 human genes with their orthologs in chimpanzee, mouse, rat and dog and applied a branch-site likelihood method to test for positive selection on the human lineage. Among the significant cases, a robust set of 11 genes were then further explored for signatures of recent positive selection using SNP data. We genotyped 223 SNPs in 39 worldwide populations from the HGDP Diversity panel and supplemented this information with available genotypes for up to 4,814 SNPs distributed along 2 Mb centered on each gene. After exploring the allele frequency spectrum, population differentiation and the maintainance of long unbroken haplotypes, we found signals of recent adaptative phenomena in only one of the 11 candidate gene regions. However, the signal of recent selection in this region may come from a different, neighbouring gene (*CD5*) rather than from the candidate gene itself (*VPS37C*). For this set of positively-selected genes in the human lineage, we find no indication that these genes maintained their rapid evolutionary pace among human populations. Based on these data, it therefore appears that adaptation for human-specific and for population-specific traits may have involved different genes.

Introduction

Identifying traits that have undergone positive selection during human evolution is essential to understand the adaptive events that have shaped our genomes. Whereas comparative genomics of closely-related species has shed light on the species-specific traits that set us apart from our closest living relatives, the genomic signature of recent adaptations can be directly detected from human population genetic data. The standard tool to detect signatures of selection in comparative data is the K_a/K_s ratio (also called d_N/d_S or ω in different analysis contexts) which expresses the ratio of nonsynonymous to synonymous substitutions in a given protein coding sequence (see a review in Yang and Bielawski 2000). Genetic variants that were selected during the process of hominization are common to all humans and are detected by comparison with sequences from other primates. Moreover, methods have been developed to examine variation in the K_a/K_s ratio among lineages (Zhang, Kumar, and Nei 1997), among codon sites (Nielsen and Yang 1998; Yang et al. 2000), and to identify selection on individual codons along specific lineages (Zhang, Nielsen, and Yang 2005). Recently, several genomewide efforts for identifying positively selected genes and/or functional categories enriched for such genes in the human and chimpanzee lineages have been conducted, yielding valuable insights into understanding human-specific traits (Clark et al. 2003; Bustamante et al. 2005; Nielsen et al. 2005; Arbiza, Dopazo, and Dopazo 2006; Kosiol et al. 2008).

The divergence that has accumulated in the human lineage since our separation from the chimpanzee occurred over the past 5 to 7 million years, and hence does not necessarily reflect recent selection that occurred after the origin of our species some 150,000 years ago. It is precisely within this evolutionary time scale that our ancestors dispersed from Africa to colonize most of the globe, and were challenged by many new selective pressures. Nonetheless, intraspecific diversity patterns within human populations may indeed reflect

such modern adaptations, since distinctive signatures of recent selective sweeps (such as an overall reduction in genetic diversity, excess of high frequency derived alleles or long range haplotypes) stay imprinted in the genome for many tens to a few hundreds of thousands of years (Sabeti et al. 2006). Identifying genes affected by recent selective sweeps in human populations has gained great interest during the last few years, as they may help to explain population-specific adaptations. Along with the tremendous increase in the availability of SNPs in public databases (Hinds et al. 2005; Frazer et al. 2007) and the high-throughput methodologies that currently exist, new analytical methods aimed at detecting the footprint of selection from SNP data have been developed and widely applied in a number of candidate-gene (Sabeti et al. 2002; Bersaglieri et al. 2004; Hughes et al. 2008), path-related genes (Walsh et al. 2006; Han et al. 2007; Sikora et al. 2008) and genome-scanning studies (Akey et al. 2002; Kimura et al. 2007; Sabeti et al. 2007; Tang, Thornton, and Stoneking 2007; Williamson et al. 2007; Barreiro et al. 2008; Myles et al. 2008; Pickrell et al. 2009). Signatures of selection such as increased levels of population differentiation, unusual allele frequency spectra and elevated levels of linkage disequilibrium (LD) are usually examined and identified in comparison to a genome-wide empirical distribution or simulated data.

Recently, a remarkable amount of evidence for targets of recent selection in humans has been gained from a set of relatively new statistics especially design to detect long range haplotypes through the measure of the extended haplotype homozygosity (EHH), whose first implementation was introduced by the long range haplotype (LRH) test (Sabeti et al. 2002). These methods rely on the principle that positively selected alleles are expected to rise to high frequency rapidly enough that long range association with alleles at nearby loci will not have time to be erased by recombination. Different strategies have recently been developed in order to capture the extended LD around a putatively selected allele (or core haplotype) in a given population (Sabeti et al. 2002; Voight et al. 2006), and on particular alleles (Kimura et al.

2007; Sabeti et al. 2007; Hughes et al. 2008) or sites (Tang, Thornton, and Stoneking 2007) when comparing pairs of populations.

Here, we address whether a sample of genes that were positively-selected during the evolution of the *Homo* lineage present any molecular signature of recent positive selection among human populations; that is, whether they continued to evolve at a non-neutral pace. We first identified fast-evolving human genes by comparing more than 11,000 coding sequences from the human genome with chimpanzee, mouse, rat and dog orthologs. For a robust subset of 11 significant cases, we analysed SNP data along 2 Mb genomic regions centered on each of the 11 genes. SNP variation was investigated in a worldwide sample of 39 human populations belonging to the HGDP-CEPH diversity panel (Cann et al. 2002). Signals of recent positive selection were interrogated through population differentiation, allele frequency threshold analyses, and by applying two complementary EHH-based tests especially designed to detect both fixed (or nearly so) and intermediate frequency selected variants. The variety of methods applied allowed us to investigate different signatures of recent selection persisting over varying evolutionary time scales (see Sabeti et al. 2006).

Materials and Methods

DNA samples

We analyzed the H971 standardized subset of the Human Genome Diversity Cell Line Panel (HGDP-CEPH) (Cann et al. 2002) recommended by Rosenberg (2006), which contains no atypical, duplicated or deduced first-degree related individuals. In order to maximize sample sizes and after considering geographic and ethnic criteria, populations from the original panel were re-grouped into 39 population samples. In particular, we grouped Tuscans and Bergamese into North Italians; Dai, Lahu, Miaozu, Naxi, She, Tujia, and Yiku as South Chinese; Daur, Hezhen, Mongolian, and Oroqen as Northeast Chinese; and Tu, Uygur, and

Xibo as Northwest Chinese. For some analyses, populations were further grouped into 7 major geographical regions as in Moreno-Estrada et al. (2008): Sub-Saharan Africa (SSAFR), Middle East-North Africa (MENA), Europe (EUR), Central-South Asia (CSASIA), East Asia (EASIA), Oceania (OCE) and America (AME). Individual samples for which genotypes in any gene region analyzed failed for at least 50% of the SNPs were not considered. Two chimpanzee samples provided by the Barcelona Zoological Park were also genotyped and if their alleles matched one of the human states, were considered ancestral.

Selection of genes

We selected 11 genes (Table 1) that exhibited evidence of accelerated evolution in the human lineage after applying the following process. Around 11,000 human, chimpanzee, mouse, rat, and dog orthologous genes were retrieved from Ensembl (March 2006) and subsequently checked to have a unique best-reciprocal BLAST match in all 5 species. For the remaining 9,170 orthologous genes, we then performed multiple sequence alignments with ClustalW (Thompson, Higgins, and Gibson 1994) and applied a branch-site likelihood method (Zhang, Nielsen, and Yang 2005) to test for positive selection on the human branch in the underlying phylogeny of the five mammal species. A likelihood ratio test (LRT) was performed by comparing the likelihood values of 2 hypotheses, allowing variation in omega (ω) among lineages and sites at the same time. As the alternative hypothesis, we used the modified Branch-Site Model A as described in Zhang, Nielsen, and Yang (2005), in which ω is estimated for all branches of the phylogeny, allowing for sites with $\omega > 1$ only in the human (foreground) branch. As the null hypothesis, we used the same Branch-Site Model A but with $\omega_2 = 1$ fixed in the human branch. The null and alternative hypotheses for this improved branch-site test 2 of positive selection (Zhang, Nielsen, and Yang 2005) and the corresponding Bayesian empirical inference of amino acid sites under positive selection

(Yang, Wong, and Nielsen 2005) were performed using the codeml program implemented in the PAML package (Yang 1997). According to the LRT, 52 genes exhibited a signature of positive selection at the 0.01 level of significance and up to 88 at the 0.05 level. The following criteria were applied to this 88-gene set to ensure appropriate robustness for the evidence of accelerated protein evolution in the human lineage: i) alignments with large gaps or excessive mismatches were manually discarded; ii) the estimated ω value in the human branch under the alternative model was checked to be greater than one; iii) at least one codon with a posterior probability ≥ 95% of belonging to the site class of positive selection on the human lineage was required for inclusion in the study; and iv) only alignments with known gene symbols were considered. Only a total of 11 genes were left after applying these strict criteria.

SNPs

We selected one SNP every 5-10 kb inside each candidate gene and up to around 30 kb in both 5' and 3' flanking regions, plus additional SNPs every 40 kb up to around 200 kb in both flanking regions. Preference was given to SNPs with a minor allele frequency (MAF) over 10%, based on the HapMap (Release 20 Jan 2006) and dbSNP (Build 125 Oct 2005) databases. Additionally, most coding non-synonymous SNPs (CNS) within each candidate gene and other functional SNPs identified using the PupaSuite web-based SNP analysis tool (Conde et al. 2006) were included, regardless of their allele frequency or validation status. A total of 223 SNPs out of 270 (82.6%) were successfully genotyped using the SNPlex Genotyping System from Applied Biosystems following the manufacturer's standard protocol. Allele separation was performed on an Applied Biosystems 3730 analyzer and both quality metrics and allele calling were reviewed using GeneMapper Software 3.5. Data from the Infinium Human Hap650Y BeadChip genotyped in the HGDP-CEPH panel (Li et al.

2008) was downloaded in bulk from the Stanford Human Genome Center website. For each candidate region, SNP genotype data were then extracted for 2 Mb regions centered on each gene of interest and merged with our previously obtained genotypes in order to maximize marker density. For *CA14* it was not possible to obtain a centered 2 Mb region because the 650K Bead Chip lacked appropiate SNP coverage along ~500 kb; we therefore obtained a slightly off-centered 2 Mb region for *CA14*. In total, 4,814 SNPs spanning ~22 Mb of the genome were analyzed (Table 1).

Data handling and analysis

Unless otherwise stated, data storage, quality control and data analysis were carried out using the SNPator web-based SNP data analysis platform (Morcillo-Suarez et al. 2008). For specific calculations and plotting purposes we used the R statistical software package (version 2.4.0, http://www.r-project.org).

Allele frequency analysis

Across every genomic region and population, we analyzed the distribution of the minor allele frequencies (MAF) and the derived allele frequencies (DAF) of the corresponding SNPs. The proportion of SNPs with allele frequencies higher or lower than a defined threshold (MAF < 0.10 for the MAF analysis and DAF > 0.80 for the DAF analysis) was calculated within sliding windows of 100 kb in size every 20 kb and plotted against distance over the full 2 Mb regions (see Figure 1). Thresholds were chosen to maximize sensitivity to selection as suggested by Walsh et al. (2006), and we required a minimum of 5 SNPs per window. Non-polymorphic SNPs in the overall 39 populations were not considered in the threshold analyses of minor and derived allele frequencies. However, SNPs fixed in a population but polymorphic elsewhere were counted as having MAF < 0.10 and DAF > 0.80

when applicable. The definition of minor allele was specific for each population rather than global and uniform across populations. We considered as outliers those regions in which multiple windows were found in the top 5% of the distribution of the respective proportions obtained independently for each population and considering all 2 Mb regions together. Given their small sample size (< 10 individuals), the San population was not included in the MAF threshold analysis. Ancestral states for all analyzed SNPs were obtained from the chimpanzee and/or the macaque genome sequences (panTro2, Mar. 2006 assembly and rheMac2, Jan. 2006 assembly, respectively). For 31 SNPs neither of the human alleles corresponded to the chimpanzee or macaque sequence and therefore these were not included the DAF threshold analysis. Genotyping of two chimpanzee samples confirmed the ancestral state of 192 SNPs.

F_{ST} calculation

The proportion of the variance explained by populational differences was measured through the molecular fixation index F_{ST}, by means of a locus by locus Analysis of Molecular Variance (AMOVA) (Excoffier, Smouse, and Quattro 1992) using the Arlequin software package version 3.11 (Excoffier, Laval, and Schneider 2005). Global F_{ST} values were obtained, taking into account all 39 worlwide populations studied as a single group. Empirical percentiles of global F_{ST} values were calculated based on the distribution of all F_{ST} values obtained over the eleven 2 Mb regions analyzed. We defined as outliers to be considered for further analysis those values within the top 5% of the empirical distribution.

Haplotype and long-range haplotype analysis

Haplotypes centered on each candidate gene were estimated using FastPHASE (Scheet and Stephens 2006). Linkage disequilibrium (LD) blocks were explored using Haploview version 4.1 (Barrett et al. 2005). The Long Range Haplotype (LRH) test (Sabeti et al. 2002)

was carried out using the SWEEP software package (version 1.1). We defined cores as the longest non-overlapping core haplotypes with at least one SNP and not more than 20 SNPs. For each identified core haplotype, we calculated the EHH and the relative EHH (REHH) at a genetic distance of 0.3 cM in both directions and plotted these against the core haplotype frequency. Distributions of EHH and REHH values were obtained for all main geographical regions from the relevant populational phased haplotype data, considering together the eleven 2 Mb regions centered on the genes of interest. Core haplotypes were placed in 5% frequency bins and the respective EHH and REHH values were log-transformed for each bin in order to obtain approximately normally distributed values. Empirical p values for the LRH test were obtained by using the mean and standard deviation of the empirical distribution of the respective scores in each continental region. The LRH test was not performed in Oceania because the populations in the continent showed reduced background distribution. To account for multiple testing, we estimated the false positive discovery rate (pFDR) (Storey and Tibshirani 2003) and calculated the q value for the scores within each frequency bin using the package q value (version 1.1) for R. The q value for a particular p value is defined as the expected proportion of false positives among all significant p values when calling that p value significant. We used a q value cutoff of 0.05 for assigning significance.

In order to allow for multiple populational and/or continental comparisons of EHH, we slightly modified the method introduced by Tang, Thornton, and Stoneking (2007) based on the ln(Rsb) statistic. The integrated extended haplotype homozygosity of individual SNP sites (iEHHS or iES) was calculated for every SNP site and population directly from genotype data using a home-coded script (PopMX package by Tang K, Bauchet M, Theunert C, unpublished data). Each iEHHS value was first normalized to the median of all values within each population, resulting in the EHHS' as following:

$$EHHS' = \frac{EHHS}{median(EHHS)}$$

EHHS' from each individual population was then divided by the average EHHS' across populations weighted for population sizes as following:

$$XP - Rsb = \frac{EHHS'}{ave(EHHS')}$$

Where,

$$ave(EHHS') = \frac{\sum(EHHS'_i \cdot n_i)}{N}$$

Note that ave(EHHS') takes sample size into consideration. Here n_i refers to the number of individuals in a population *i* and *N* is the total number of samples used from the HGDP–CEPH panel. To estimate significance values for our results, we obtained a background distribution of XP-Rsb values for 642,690 genome-wide distributed SNPs using genotype data for the same panel generated elsewhere (Li et al. 2008). For each population and each SNP site, we obtained a p-value by ranking its XP-Rsb value across the whole genome in that population and determine its quantile. We then log transformed the p-values and plotted them against position within each 2 Mb region, searching for clusters of significant values inside or around our candidate genes at both population and continental levels. We calculated XP-Rsb for every SNP site and each population (vs. all other HGDP-CEPH populations) as well as for each main geographical region (versus the remaining regions represented in the panel).

Identification of functional variants

We limited the search of possible functional variants along the *VPS37C* genomic region to an LD block spanning ~420 kb, where the strongest signals of selection were concentrated. For this purpose, we explored *in silico* the functional relevance of all the genotyped SNPs in our mixed dataset as well as of all available HapMap SNPs within the same region (HapMap data Release 23a/phaseII). The PupaSuite web-based tool (Conde et al.

2006) was used to detect all SNPs with a potential phenotypic effect, including coding non-synonymous SNPs (CNS), SNPs disrupting miRNAs and their targets, as well as those SNPs located at triplexes or altering exonic splicing enhancers, exonic splicing silencers or transcription factor binding sites. The impact exerted by the amino acid substitution of each CNS was evaluated by means of Grantham's physicochemical distances (Grantham 1974), the damaging probabilities predicted by PolyPhen (Ramensky, Bork, and Sunyaev 2002) and by the codon-level selective constraints for prediction of functional altering mutations as estimated in PupaSuite (Arbiza et al. 2006). The haplotype extension of rs2229177 was explored on HapMap Phase II data using the Haplotter web-based application (Voight et al. 2006). We also downloaded HapMap data (release 22/phaseII) to look for tagSNPs in the *CD5* gene ±100 kb for both CEU and CHB+JPT samples using Tagger with the default parameters given by the authors (de Bakker et al. 2005). Multi-species alignment of the *CD5* sequence was visualized within the Ensembl genome browser (release 50, July 2008) selecting all available sequences from 23 eutherian mammals. The amino acid sequence and structure information of the CD5 protein were obtained from the UniProt Knowledgebase (UniProtKB, entry P06127) and both the ModBase database (Pieper et al. 2004) and the Protein Data Bank (PDB, entry 1by2) (Berman et al. 2000), respectively.

Results

Selection of genes

Table 1 summarizes the 11 fast-evolving genes in the human lineage for which we analyzed SNP genotype data from 39 globally-distributed populations. Most of them had p-values smaller than 0.01 as determined by a likelihood ratio test of positive selection specifically in the human lineage on a phylogeny containing 5 mammal species (see Materials and Methods). None of the analyzed regions overlap, and although some map to the same

chromosome, they can be treated as 11 independent genomic regions. Six of the 11 selected genes had more than 2 codons putatively affected by selection-driven amino acid changes, and 2 cases (both of them olfactory receptors) showed up to 7 codons putatively affected by selection, according to the posterior probability analysis of the Bayesian empirical inference. Interestingly, some of the positively selected codons contain non-synonymous polymorphic positions in extant human populations (rs6597801 in *LHPP*, rs2961160 and rs2961161 in *OR2A14*, rs754382 in *VPS37C*, and rs1943639 in *OR5G1P*). Given recent criticisms on the reliability of the site-prediction methods (Nozawa, Suzuki, and Nei 2009), these results alone should be interpreted with caution. Notably, 10 of the 11 genes analyzed here were also inferred to have undergone positive selection in the human lineage in at least one of three other publications based on more closely related and/or a larger number of species and similar branch-site tests of positive selection (Arbiza, Dopazo, and Dopazo 2006; Bakewell, Shi, and Zhang 2007; Nickel, Tefft, and Adams 2008). Next, we explored these gene regions for different signatures of recent positive selection in worldwide human populations using SNP data.

Allele frequency threshold analyses

The distribution of minor and derived allele frequencies around a given genomic region may suggest particular selective pressures acting on it. In particular, an excess of high frequency derived alleles may indicate positive selection, whereas the presence of an excess of low frequency variants could reflect either purifying selection or a recent selective sweep.

In the MAF threshold analysis, for each population within each of the 7 main geographical regions analyzed, we plotted the proportion of SNPs with MAF < 0.10 within multiple 100 kb sliding windows along 2 Mb regions centered on each candidate gene (Figures S1–S7). The general trend is characterized by a limited number of scattered outlier

windows in different populations within each continental region along the 2 Mb regions analyzed. The American and Oceanian populations displayed many frequency fluctuations resulting in many consecutive windows with extreme proportions (either high or low) of SNPs with MAF < 0.10, a pattern which is probably due to their high levels of genetic drift and isolation. As for the genes of interest, only *VPS37C* and *OR2A14* concentrated an excess of rare alleles in East Asian, Oceanian and/or in American populations, respectively (Figures 1a and S5-S7).

In the DAF threshold analysis, for each population within each of the 7 main geographical regions analyzed, we computed the proportion of SNPs with DAF > 0.8 within multiple 100 kb sliding windows along 2 Mb regions centered on each candidate gene (Figures S8-S14). Again, there are several clusters of windows with an excess of high-frequency derived alleles along the 2 Mb regions analyzed but just a few of them seem to involve the positively-selected genes. Within those, both the *VPS37C* and *USP2* genomic regions stand out for displaying in several populations a significant excess of high-frequency derived alleles, either in the genes or just nearby. The strongest signals for *VPS37C* are found in East Asia (Figure 1b) and in some Central South Asian populations (Figure S11). Without clearly including the *VPS37C* gene, the same pattern of an excess of high-frequency derived alleles is detectable in almost any population outside Sub-Saharan Africa, extending almost 400 kb from the 3' flanking region. The signal for the *USP2* genomic region is found in populations from all main geographical regions except Oceania and America, and in all cases involves the 3' region of the *USP2* gene but not the candidate itself.

Population differentiation

Local adaptation may cause unusually large allele frequency differences between populations at the selected loci, and consequently accentuate their levels of population

differentiation. Here, we used F_{ST} to measure differentiation among all 39 worldwide populations for the 4,814 SNPs distributed across the eleven 2 Mb regions analyzed (Figure S15). None of the candidate genes have unusually high F_{ST} values, either in the gene or nearby, except *VPS37C* (Figure 1c). A total of 64 SNPs above the 95th percentile (23.3% of the top 5% values) were found in the 2 Mb region centered on *VPS37C.* The highest individual F_{ST} value was 0.4252 (rs17156025) and the average F_{ST} within the 64 highly differentiated SNPs was 0.2889.

Long unbroken haplotypes

Recent selective sweeps can produce a distinctive signature on the haplotype structure of chromosomes consisting of an allele (or haplotype) that has both high frequency and long-range associations with alleles at nearby loci (Sabeti et al. 2006). In order to try to detect such a signature in the candidate regions, we applied 2 complementary approaches based on the Extended Haplotype Homozygosity (EHH) measure (see a similar strategy in Pickrell et al. 2009). The first approach compares the EHH decay between the alleles (EHHA) of a site or core-haplotype within a given population and has strong power for identifying alleles that have been driven to intermediate frequencies during a recent selective sweep (Sabeti et al. 2002). In contrast, the second approach aims to detect nearly or recently completed local selective sweeps by comparing the EHH profile at individual SNP sites (EHHS) between populations (Tang, Thornton, and Stoneking 2007). As to the first approach, we applied the long range haplotype (LRH) test (Sabeti et al. 2002) by measuring for each core haplotype detected in our data the relative EHH (REHH) at a genetic distance of 0.3 cM in both directions from each core. Figure 2 shows the distributions of REHH values versus frequency for all of the populations analyzed within each main geographical region (except Oceania). Table 2 lists the corresponding significant core haplotypes after correction for multiple

testing. Both Europe and Middle East-North Africa presented 2 high frequency core haplotypes as outliers (Figure 2). Three of them remained significant after multiple test correction (Table 2), but none of them mapped directly upon any of the candidate genes. However, the long range homozygosity associated with the significant core haplotype found in Middle East-North Africa is mantained near the *VPS37C* gene (Figure 3). The corresponding haplotype bifurcation plots for the 2 main haplotypes found in the core show unusual long range LD for the ACG core, given its frequency (0.821). Notably, the strongest signal for this significant core is reached at 0.42 cM where the REHH goes up to 46.4 (Figure S16). Central South Asia and East Asia showed 3 significant outliers but in low-frequency bins (Figure 2). Interestingly, 2 of those involved the candidate *LHPP* gene. Given their low frequency the maintenance of these haplotypes over the region is less clear when looking at the bifurcation and EHH decay patterns (data not shown); nonetheless they remained significant inside their frequency bins, which could reflect a partial ongoing selective sweep on the way to higher frequencies.

An obvious caveat of the previous analysis is that the intra-population comparison has low power when the selected allele variant is at high frequency, and becomes impossible when the variant is fixed. For this reason, in our second approach, we applied a slight modification of the ln(Rsb) statistic developed by Tang, Thornton, and Stoneking (2007) designed to detect local selective sweeps by means of inter-population comparisons of EHH. Here we are analyzing 39 different populations and a minimum of 7 groups when pooling populations into their main geographical regions, a number for which the ln(Rsb) statistic was not initially designed. To tackle this problem we modified the original formulation to XP-Rsb by comparing each individual population's iEHHS against a weighted cross-population average for each SNP position (see details in Materials and Methods). Figure S17 shows the -log p-value of XP-Rsb along the eleven 2 Mb genomic regions analyzed for each main

geographical region. Although there are some clusters of significant p-values, only 2 were located near any of the candidate genes. In particular, East Asia showed a significant EHH differentiation pattern when compared to the other geographic regions around the *GFRA3* and *VPS37C* gene regions (Figure 1d). In order to identify which population(s) within each geographic region might account for these signals, we also computed XP-Rsb between the 39 worldwide populations. Detailed results for the 11 full 2 Mb genomic regions are shown in Figures S18–S28; previously observed signals at the regional level could be attributed to specific populations, and some new signals were detected. For example, Cambodians, Han and Japanese are behind the EASIA signal previously observed in the *GFRA3* gene region. As for the *VPS37C* gene (Figure 1e), we found that significant XP-Rsb values were obtained in this gene region for 3 out of 6 East Asian populations, namely North East China, Han and Japanese, with Han Chinese accounting for most of the significant values. Not surprisingly, the signal is less significant than the one observed at the regional level, since the latter is decomposed into different individual population signals. New signals were also found for South Chinese in *HDHD3*, Pathan and Pima in *LHPP* and Japanese and Yakut in *OR5G1P*. Despite encompassing the genes of interest, these last cases have their highest significance values far outside them.

Insights on the VPS37C genomic region

Out of the eleven 2 Mb regions centered on our candidate genes, *VPS37C* consistently exhibits significant signatures of positive selection, especially in Asians. Most of the signals extend along more than 0.5 Mb and comprise several genes besides *VPS37C* (see top part of Figure 1). In order to identify which allelic variants could be responsible for the observed pattern, we first characterized the haplotype composition in this 0.5 Mb region, and then searched for variants with functional relevance on the putatively selected haplotype. In

particular, we focussed our analysis on a ~420 kb region of relatively strong linkage disequilibrium, delimited by two hotspots of recombination, and containing 54 SNPs. The haplotype frequency distribution across the 39 worldwide populations analyzed for this narrowed *VPS37C* region (Figure 4) reveals a total of 692 different haplotypes. While up to 80% of the Sub-Saharan African haplotypes were found in single chromosomes and most of the Eurasian populations had 10-20% unique haplotypes, one particular 54-SNP based haplotype stands out as having relatively high frequencies in North West China (59 %) and most East Asian (67 %) and American populations (60 %).

We functionally characterized not only the 54 SNPs contained in the analyzed dataset, but also all available SNPs in HapMap (Frazer et al. 2007) within the same region. In order to explain the observed significant pattern of selection any potentially causative genetic variant should be: i) functionally relevant, ii) particularly frequent in Asians but not elsewhere and iii) embedded within the extended haplotype in the major allele state. A total of 12 coding non-synonymous SNPs (CNS) affecting 7 different genes were found in the target region, most of which were not typed in our mixed dataset but in HapMap. Table 3 summarizes their allele frequencies in the HapMap populations, the amino acid replacements they involve, and their inferred functional effects as predicted by different methods (see Materials and Methods). Only one CNS (rs2229177 in *CD5*) showed high frequencies in Asians with intermediate frequencies elsewhere, and a relevant functional effect (predicted as possibly damaging by PolyPhen and pathological by PupaSuite). Since this SNP was not genotyped in the HGDP panel, we explored Haplotter Phase II data for rs2229177 and confirmed that the derived state (T) sits in a long unbroken haplotype that is maintained at very high frequencies for approximately 400 kb in the Asian sample (data not shown). Moreover, taking advantage of the strong linkage disequilibrium in the region, we looked for tagSNPs capturing rs2229177 variation in the HapMap populations. Several SNPs tag rs222917 with $r^2 = 1.0$,

four of which (i.e. rs4245224, rs10897141, rs610777 and rs628831) were typed in the HGDP panel. All 4 major alleles (as for Asian populations) are embedded within the main 54 SNP based haplotype found in Asians. Moreover, 2 of them (i.e. rs4245224 and rs628831) reproduced exactly the same derived allele frequencies of rs2229177 from the 4 HapMap populations in our equivalent samples from the HGDP-CEPH panel (i.e. Yoruba, French, Han and Japanese) which allows us to infer the possible worldwide frequency distribution of this CNS. When searching for other SNP functional categories (see Materials and methods for details), we compiled a total of 62 additional SNPs with potential phenotypic effects (data not shown) but only one, rs1787904, appeared to be differentiated at high frequencies in Asians (CHB: 0.988, JPT: 1, YRI: 0.542 and CEU: 0.567) and linked to the putatively selected haplotype (data not shown). Although this substitution maps within an intron of the *VPS37C* gene, it is located in a triplex sequence (Goni, de la Cruz, and Orozco 2004; Conde et al. 2006) that is within 10 kb from the 3' end of the *CD5* gene, and hence could modify *CD5* expression.

Discussion

We have addressed the question of whether a robust subset of 11 fast-evolving genes in the human lineage are still evolving adaptatively within extant human populations. To do so, we looked for signatures of recent adaptation in worldwide human populations along eleven 2 Mb genomic regions, each centered on a gene identified as exhibiting accelerated evolution on the human lineage. Most of the candidate gene regions did not show clear evidence of undergoing recent selection within the worldwide human diversity panel. The absence of recent signatures of selection on most positively-selected genes included in this study may imply that once they had acquired a specific human function, they became functionally constrained against further change. Note that false negative results could arise

from insufficient statistical power of the methods employed to detect selection. In particular, EHH tests based on alleles tend to miss selection near fixation whereas EHH tests for sites are less powerful for partial sweeps (see Sabeti et al. 2006; Voight et al. 2006 and Tang et al. 2007 for discussion of the power of EHH in relation to the frequency of the selected variant). As for demography, Tang et al. (2007) demonstrated the robustness of the lnRsb statistic accross a wide range of demographic models and Pickrell et al. (2009) shown that XP-EHH (which is similar to XP-Rsb) is not particularly sensitive to demographic effects either. However, when using empirical distributions, rates of false positive and false negative results will be affected by the overall number of recently selected genes. This is an unknown (and highly debated) quantity, with some authors (Pickrell et al. 2009) finding few genes under positive selection, while others (Hawks et al. 2007) find large numbers.

Only one genomic region, *VPS37C,* showed significant signals across all the signatures of selection we explored. The observed pattern in this region is consistent with the action of recent positive selection in East Asian populations. Given that population-specific empirical distributions based either on the whole genome or a fraction of it (i.e., 2 Mb windows around the original 11 genes) were used, this result is difficult to be explained alternatively by demographic processes such as bottlenecks, that would mimic the genomic signatures of selection, albeit in all of the genome. Despite the strong evidence in favour of a selective sweep occurring along the *VPS37C* genomic region, it is difficult to pinpoint the source, since there were different clusters of significant signals across a ~ 0.5 Mb region. While there are highly differentiated loci throughout the candidate region, some signals (such as those displayed by the MAF and DAF threshold analyses or the presence of significant core haplotypes in the LRH test) do appear to be concentrated from the vicinity of the *VPS37C* gene up to ~ 400 kb upstream. On the contrary, the highest concentration of significant p-values in the XP-Rsb analysis starts at *VPS37C* but only extends 100 kb

downstream. Despite the limitations of these methods to accurately locate the target of selection, the observed pattern does suggest that it might be somewhere in or around the *VPS37C* gene and that part of the extended signal is due to high LD.

In agreement with these results, a previous genome-wide scan reported the *VPS37C* gene region among the 101 regions with the strongest evidence for a recent selective sweep in Chinese (Williamson et al. 2007). This study used a composite likelihood ratio test, which provides fine-scale estimates of the position of the selected site, and which for this region was mapped to a 200-SNP window centered on the *VPS37C* gene. As suggested by the same authors (Williamson et al. 2007), since the VPS37C protein is recruited by HIV and other viruses to promote viral budding from infected cells (Stuchell et al. 2004; Eastman et al. 2005), it might play an important role in pathogen interactions. However, the identification of *VPS37C* as the actual gene responsible for the signal of selection in this region remains to be confirmed, and all known genes within ±100 kb were not rejected as alternative candidates.

Here, in a detailed analysis of the haplotype composition of the region we identified a paticular 54-SNP based haplotype at relatively high frequencies in Asia and America. This haplotype spans a ~420 kb block, and encompasses all of the different signals observed in this genomic region. The functional characterization of all known allelic variants linked to this putatively selected haplotype suggested 2 candidates, a non-synonymous coding SNP located in the last exon of *CD5* (rs2229177) and a substitution altering a triplex-forming target sequence within its 3' end regulatory region (rs1787904). The *CD5* gene is located less than 5 kb from *VPS37C* and codes for a 495-amino-acid-long transmembrane receptor expressed in the T-cell surface. Topologically, it comprises a large extracellular domain (amino acids 25-372) containing 3 repeats of a cysteine-rich region (SRCR domains), followed by a single-pass transmembrane domain (amino acids 373-402) and a short cytoplasmic region (amino acids 403-495). The aforementioned CNS (rs2229177) leads to an Ala-Val substitution at

position 471 in the cytoplasmatic part of the protein, which has been reported as essential for the function of the receptor (Pena-Rossi et al. 1999; Bhandoola et al. 2002). Alanine is encoded by the ancestral state (C) while Valine (encoded by T) is the derived state, which characterises the major form of the protein in Asian and American populations. In contrast with the much more variable extracellular region, this cytoplasmatic part of the receptor is highly conserved across species based on multiple sequence alignments. Berland and Wortis (2002) reported that only 5 out of the 96 amino acids of this region differ among 5 mammalian sequences (human, mouse, sheep, bovine and rat). Moreover, all other eutherian mammals (23 species compared) conserved the ancestral state at the polymorphic rs2229177 position. Despite the availability of several 3D-structure models for this receptor, none of them includes the cytoplasmatic region where the A471V substitution is located. The lack of a complete experimental template prevents any conclusive prediction of the structural or functional impact of this substitution.

The *CD5* gene codes for a glycoprotein that acts as a transmembrane receptor in regulating T-cell proliferation. Specifically, CD5 functions as a negative regulator of T-Cell Receptor (TCR) signaling during intrathymic T cell development. Experimental studies have reported that CD5 mediated down-regulation does not require the CD5 extracellular domain and, consequently, does not involve CD5 binding of an extracellular ligand (Bhandoola et al. 2002). In contrast, the cytoplasmic portion of the molecule is required to act as an inhibitory receptor (Pena-Rossi et al. 1999). It has been pointed out that autoimmune disorders may result from the disruption of inhibitory receptors, particularly in their conserved intracellular motifs which are responsible for transducing signals to distinct pathways (Ravetch and Lanier 2000). Additional evidence for a functional role for rs2229177 (A471V) comes from a genetic association study in which it was shown that homozygosity for the ancestral allele in A471V is associated with a poorer prognosis in patients of chronic lymphocytic leukemia (CLL)

(Sellick et al. 2008). Given the function of the *CD5* gene and its role in the immune system physiopathology, it is tempting to speculate a possible protective effect for the putatively selected haplotype in Asians, although the exact mechanism by which the silencing of such a regulatory receptor would have been favoured by selection remains elusive. Additionally, other unknown variation linked to the same haplotype cannot be discarded as the actual functional variant responsible for the observed signals of selection.

Our results also demonstrate how both EHH-based approaches complement each other, as predicted by their estimated power to detect selection depending on the frequency of the selected allele in the population (Sabeti et al. 2007). Here, we found signals for the LRH test in populations where the putatively selected haplotype is segregating at intermediate frequencies (i.e. Middle East-North Africa), while for XP-Rsb we found evidence for selection involving the same haplotype in East Asia, where it has nearly reached fixation. Finally, we illustrate a case in which the ancient selective event of *VPS37C* during early human evolution and its apparent recent selective sweep are actually two independent phenomena. For a set of 11 genes with the most robust signs of adaptation in the human lineage, we could not find evidence of non-neutral evolution occurring after the advent of the human species. Based on these data, it appears that population-specific adaptation in humans may have been an independent process, involving different sets of genes than those that participated in defining our species.

Acknowledgments

This work was supported by Dirección General de Investigación, Ministerio de Educación y Ciencia, Spain [BFU2005-00243 and BFU2008-010406/BMC to E.B.] and by the Direcció General de Recerca, Generalitat de Catalunya [2005SGR00608 to J.B]. A Moreno-Estrada was supported by a fellowship from the Consejo Nacional de Ciencia y Tecnología (CONACyT), México and received a travel grant from the Boehringer Ingelheim Foundation, Germany. SNP genotyping services were provided by the Spanish "Centro Nacional de Genotipado" (www.cegen.org).

Electronic Database Information

The Uniform Resource Locators (URLs) for data presented herein are as follows:

Stanford HGDP SNP Genotyping Data: http://shgc.stanford.edu/hgdp/index.html

SNPator web application: http://bioinformatica.cegen.upf.es

R statistical software package: http://www.r-project.org

SWEEP software package: http://www.broad.mit.edu/mpg/sweep/index.html

PolyPhen: http://genetics.bwh.harvard.edu/pph

SNPeffect: http://snpeffect.vib.be

Haplotter: http://hg-wen.uchicago.edu/selection/haplotter.htm

PupaSuite: http://pupasuite.bioinfo.cipf.es/

Tagger: http://www.broad.mit.edu/mpg/tagger/server.html

UniProtKB: http://www.uniprot.org/

Protein Data Bank: http://www.pdb.org/pdb/home/home.do

ModBase database: http://modbase.compbio.ucsf.edu/modbase-cgi/index.cgi

HapMap Genome Browser: http://www.hapmap.org/

References

Akey JM, Zhang G, Zhang K, Jin L, Shriver MD. 2002. Interrogating a high-density SNP map for signatures of natural selection. Genome Res. 12:1805-1814.

Arbiza L, Dopazo J, Dopazo H. 2006. Positive selection, relaxation, and acceleration in the evolution of the human and chimp genome. PLoS Comput Biol. 2:e38.

Arbiza L, Duchi S, Montaner D, Burguet J, Pantoja-Uceda D, Pineda-Lucena A, Dopazo J, Dopazo H. 2006. Selective pressures at a codon-level predict deleterious mutations in human disease genes. J Mol Biol. 358:1390-1404.

Bakewell MA, Shi P, Zhang J. 2007. More genes underwent positive selection in chimpanzee evolution than in human evolution. Proc Natl Acad Sci USA 104:7489-7494.

Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L. 2008. Natural selection has driven population differentiation in modern humans. Nat Genet. 40:340-345.

Barrett JC, Fry B, Maller J, Daly MJ. 2005. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21:263-265.

Berland R, Wortis HH. 2002. Origins and functions of B-1 cells with notes on the role of CD5. Annu Rev Immunol. 20:253-300.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The Protein Data Bank. Nucleic Acids Res. 28:235-242.

Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, Hirschhorn JN. 2004. Genetic signatures of strong recent positive selection at the lactase gene. Am J Hum Genet. 74:1111-1120.

Bhandoola A, Bosselut R, Yu Q, Cowan ML, Feigenbaum L, Love PE, Singer A. 2002. CD5-mediated inhibition of TCR signaling during intrathymic selection and development does not require the CD5 extracellular domain. Eur J Immunol. 32:1811-1817.

Bustamante CD, Fledel-Alon A, Williamson S, et al. (14 co-authors). 2005. Natural selection on protein-coding genes in the human genome. Nature 437:1153-1157.

Cann HM, de Toma C, Cazes L, et al. (41 co-authors). 2002. A human genome diversity cell line panel. Science 296:261-262.

Clark AG, Glanowski S, Nielsen R, et al. (17 co-authors). 2003. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. Science 302:1960-1963.

Conde L, Vaquerizas JM, Dopazo H, Arbiza L, Reumers J, Rousseau F, Schymkowitz J, Dopazo J. 2006. PupaSuite: finding functional single nucleotide polymorphisms for large-scale genotyping purposes. Nucleic Acids Res. 34:W621-625.

de Bakker PI, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D. 2005. Efficiency and power in genetic association studies. Nat Genet. 37:1217-1223.

Eastman SW, Martin-Serrano J, Chung W, Zang T, Bieniasz PD. 2005. Identification of human VPS37C, a component of endosomal sorting complex required for transport-I important for viral budding. J Biol Chem. 280:628-636.

Excoffier L, Laval G, Schneider S. 2005. Arlequin ver. 3.0: An integrated software package for population genetics data analysis. Evol Bioinform Online 1:47-50.

Excoffier L, Smouse PE, Quattro JM. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics 131:479-491.

Frazer KA, Ballinger DG, Cox DR, et al. (233 co-authors). 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851-861.

Goni JR, de la Cruz X, Orozco M. 2004. Triplex-forming oligonucleotide target sequences in the human genome. Nucleic Acids Res. 32:354-360.

Grantham R. 1974. Amino acid difference formula to help explain protein evolution. Science 185:862-864.

Han Y, Gu S, Oota H, Osier MV, Pakstis AJ, Speed WC, Kidd JR, Kidd KK. 2007. Evidence of positive selection on a class I ADH locus. Am J Hum Genet. 80:441-456.

Hawks J, Wang ET, Cochran GM, Harpending HC, Moyzis RK. 2007. Recent acceleration of human adaptive evolution. Proc Natl Acad Sci U S A 104:20753-20758.

Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR. 2005. Whole-genome patterns of common DNA variation in three human populations. Science 307:1072-1079.

Hughes DA, Tang K, Strotmann R, Schoneberg T, Prenen J, Nilius B, Stoneking M. 2008. Parallel selection on TRPV6 in human populations. PLoS ONE 3:e1686.

Kimura R, Fujimoto A, Tokunaga K, Ohashi J. 2007. A practical genome scan for population-specific strong selective sweeps that have reached fixation. PLoS ONE 2:e286.

Kosiol C, Vinar T, da Fonseca RR, Hubisz MJ, Bustamante CD, Nielsen R, Siepel A. 2008. Patterns of positive selection in six Mammalian genomes. PLoS Genet 4:e1000144.

Li JZ, Absher DM, Tang H, et al. (11 co-authors). 2008. Worldwide human relationships inferred from genome-wide patterns of variation. Science 319:1100-1104.

Morcillo-Suarez C, Alegre J, Sangros R, et al. (17 co-authors). 2008. SNP analysis to results (SNPator): a web-based environment oriented to statistical genomics analyses upon SNP data. Bioinformatics 24:1643-1644.

Moreno-Estrada A, Casals F, Ramirez-Soriano A, Oliva B, Calafell F, Bertranpetit J, Bosch E. 2008. Signatures of selection in the human olfactory receptor OR5I1 gene. Mol Biol Evol. 25:144-154.

Myles S, Tang K, Somel M, Green RE, Kelso J, Stoneking M. 2008. Identification and analysis of genomic regions with large between-population differentiation in humans. Ann Hum Genet. 72:99-110.

Nickel GC, Tefft D, Adams MD. 2008. Human PAML browser: a database of positive selection on human genes using phylogenetic methods. Nucleic Acids Res. 36:D800-808.

Nielsen R, Bustamante C, Clark AG, et al. (12 co-authors). 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol. 3:e170.

Nielsen R, Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148:929-936.

Nozawa M, Suzuki Y, Nei M. 2009. Reliabilities of identifying positive selection by the branch-site and the site-prediction methods. Proc Natl Acad Sci U S A. 106:6700-6705.

Pena-Rossi C, Zuckerman LA, Strong J, Kwan J, Ferris W, Chan S, Tarakhovsky A, Beyers AD, Killeen N. 1999. Negative regulation of CD4 lineage development and responses by CD5. J Immunol. 163:6494-6501.

Pickrell JK, Coop G, Novembre J, et al. (11 co-authors). 2009. Signals of recent positive selection in a worldwide sample of human populations. Genome Res. 19:826-837.

Pieper U, Eswar N, Braberg H, et al. (15 co-authors). 2004. MODBASE, a database of annotated comparative protein structure models, and associated resources. Nucleic Acids Res. 32:D217-222.

Ramensky V, Bork P, Sunyaev S. 2002. Human non-synonymous SNPs: server and survey. Nucleic Acids Res. 30:3894-3900.

Ravetch JV, Lanier LL. 2000. Immune inhibitory receptors. Science 290:84-89.

Rosenberg NA. 2006. Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives. Ann Hum Genet. 70:841-847.

Sabeti PC, Reich DE, Higgins JM, et al. (17 co-authors). 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature 419:832-837.

Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, Palma A, Mikkelsen TS, Altshuler D, Lander ES. 2006. Positive natural selection in the human lineage. Science 312:1614-1620.

Sabeti PC, Varilly P, Fry B, et al. (244 co-authors). 2007. Genome-wide detection and characterization of positive selection in human populations. Nature 449:913-918.

Scheet P, Stephens M. 2006. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet. 78:629-644.

Sellick GS, Wade R, Richards S, Oscier DG, Catovsky D, Houlston RS. 2008. Scan of 977 nonsynonymous SNPs in CLL4 trial patients for the identification of genetic variants influencing prognosis. Blood 111:1625-1633.

Sikora M, Ferrer-Admetlla A, Mayor A, Bertranpetit J, Casals F. 2008. Evolutionary analysis of genes of two pathways involved in placental malaria infection. Hum Genet. 123:343-357.

Storey JD, Tibshirani R. 2003. Statistical significance for genomewide studies. Proc Natl Acad Sci USA 100:9440-9445.

Stuchell MD, Garrus JE, Muller B, Stray KM, Ghaffarian S, McKinnon R, Krausslich HG, Morham SG, Sundquist WI. 2004. The human endosomal sorting complex required for transport (ESCRT-I) and its role in HIV-1 budding. J Biol Chem. 279:36059-36071.

Tang K, Thornton KR, Stoneking M. 2007. A New Approach for Using Genome Scans to Detect Recent Positive Selection in the Human Genome. PLoS Biol. 5:e171.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673-4680.

Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. PLoS Biol. 4:e72.

Walsh, EC, Sabeti P, Hutcheson HB, et al. (16 co-authors). 2006. Searching for signals of evolutionary selection in 168 genes related to immune function. Hum Genet. 119:92-102.

Williamson SH, Hubisz MJ, Clark AG, Payseur BA, Bustamante CD, Nielsen R. 2007. Localizing recent adaptive evolution in the human genome. PLoS Genet. 3:e90.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci. 13:555-556.

Yang Z, JP Bielawski. 2000. Statistical methods for detecting molecular adaptation. Trends Ecol Evol. 15:496-503.

Yang Z, Nielsen R, Goldman N, Pedersen AM. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155:431-449.

Yang Z, Wong WS, Nielsen R. 2005. Bayes empirical bayes inference of amino acid sites under positive selection. Mol Biol Evol. 22:1107-1118.

Zhang J, Kumar S, Nei M. 1997. Small-sample tests of episodic adaptive evolution: a case study of primate lysozymes. Mol Biol Evol. 14:1335-1338.

Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol. 22:2472-2479.

Table 1. Summary of analyzed regions

Gene[a]	Chr	Size (kb)	Biological function	LRT[b] P value	Positively selected codons[c]	SNP coverage	
						400 kb[d]	2 Mb[e]
GFRA3	5	22.2	Glycosyl-phosphatidyl inositol receptor	0.0012	304	20	266
PTGER4	5	13.8	Skin immune responses	0.0014	8, 235	20	493
HDHD3	9	2.6	Glycobiological function	0.0014	28	19	663
LHPP	10	152.3	Hydrolase activity	0.0017	69, 78, **94** (rs6597801)	26	655
CA14	1	7.3	Hydration of carbon dioxide	0.0032	131, 153, 204	17	175
OR2A14	7	0.9	Olfactory receptor	0.0077	4, 14, 75, **132** (rs2961160), 157, **163** (rs2961161), 256	16	371
VPS37C	11	31.2	Endosomal transport important for viral budding	0.0091	70, **197** (rs754382), 248, 275	22	420
MRC2	17	64.9	Binding and transmission of HIV	0.0111	158	23	261
USP2	11	25.4	Cell cycle regulation and apoptosis	0.0161	140	22	408
OR5G1P	11	0.8	Olfactory receptor	0.0168	4, 43, 48, 54, 108, **183** (rs1943639), 193	17	423
ADI1	2	21.6	Suppression of tumor cell invasion in tissues	0.0256	62	21	679

[a] Gene of interest

[b] Likelihood Ratio Test

[c] According to the Bayes Empirical Bayes analysis, all codons with posterior probability greater than 95% for being positively selected in the human lineage are listed. Codons involving polymorphic positions within human populations are in bold and the corresponding SNP ID presented in brackets

[d] Number of genotyped SNPs covering a 400 kb region centered on the respective gene

[e] Total number of analyzed SNPs in 2 Mb around each gene including both our genotyped SNPs and those publicly available from the 650K SNP array typed on the HGDP panel (Li et al. 2008)

Table 2. Core haplotypes with significant REHH values across the eleven 2 Mb genomic regions analyzed

Genomic region[a]	Geographical region	Genes in core region[b]	Distance (cM)[c]	Core haplotype	REHH	Frequency	*P* value	*q* value
VPS37C	MENA	*C11orf79, C11orf66, SYT7*	0.33	ACG[d]	24.87	0.821	2×10^{-4}	0.0124
VPS37C	EUR	*C11orf11, C11orf9, C11orf10, FEN1*	0.30	A[e]	34.52	0.735	0.8×10^{-4}	0.0080
PTGER4	EUR		0.26	A[f]	18.71	0.605	1.6×10^{-4}	0.0342
OR2A14	CSASIA	*TPK1*	0.30	TCT[g]	22.79	0.197	0.6×10^{-4}	0.0191
LHPP	CSASIA	*LHPP*	0.35	CGTC[h]	27.12	0.183	0.2×10^{-4}	0.0153
LHPP	EASIA	*LHPP, FAM53B*	-0.30	AGGAGGGA[i]	25.76	0.114	0.8×10^{-4}	0.0487

[a] Genomic regions are identified with the name of the candidate gene they contain

[b] Genes within ±100 kb around the core are considered

[c] Genetic distance (cM) from the core at which the signal has been captured. (–) indicates downstream direction, otherwise upstream

[d] rs3019187, rs2957858, rs12295977

[e] rs174534

[f] rs1876142

[g] rs6946827, rs17287011, rs990282

[h] rs7917600, rs7070581, rs4962607, rs11245137

[i] rs12411439, rs3781458, rs1006368, rs3781453, rs17152175, rs7099298, rs3781452, rs6597848

Table 3. Functional characterization and HapMap frequencies for the coding non-synonymous SNPs present in the 54-SNP *VPS37C* region.

SNP[a]	Position[b]	Gene	Alleles[c]	Derived allele frequencies				Aa change	Aa pos	Grantham Distance[d]	PolyPhen		Phenotypic effect[g]
				YRI	CEU	CHB	JPT				prediction[e]	score[f]	
rs2241002	60643489	*CD5*	C/**T**	0.300	0.186	0.044	0.045	P/L	224	98	PR D	2.547	Pathological
rs637186	60649182	*CD5*	G/**A**	0	0.083	0	0	H/R	461	29	benign	0.4	
rs2229177	60649811	*CD5*	C/**T**	0.475	0.576	0.988	1	A/V	471	64	PO D	1.542	Pathological
rs4297482	60656155	*VPS37C*	A/**C**	0	0	0	0	S/A	261	99	benign	1.323	Pathological
rs754382	60656343	*VPS37C*	C/**T**	0.058	0.300	0.011	0	L/S	198	145	benign	1.441	
rs3750982	60783066	*VWCE*	G/**C**	0	0	0	0	P/R	842	103	PR D	2.074	
rs2260655	60865550	*DAK*	G/**A**	0.342	1	0.989	0.966	A/T	185	58	benign	0.668	Pathological
rs11605407	60875103	*CYBASC3*	A/**C**	0	0	0	0	V/G	214	109	PO D	2.64	Pathological
rs11557691	60935017	*CPSF7*	G/**T**	0	0	0	0	D/Y	464	160	benign	N/A	Pathological
rs1064377	60939747	*CPSF7*	G/**C**	0	0	0	0	A/P	351	27	benign	1.452	Pathological
rs11230707	61010417	*C11orf66*	C/**A**	0.142	0	0	0	T/N	238	65	PO D	1.536	
rs12787061	61013931	*C11orf66*	G/**C**	0.008	0.018	0	0	S/T	382	58	benign	0.335	

[a] For the two SNPs in bold, genotype data for the HGDP panel are also available

[b] SNP positions are based on NCBI build 36

[c] Observed alleles are indicated as ancestral/**derived**

[d] Mean chemical distance for the corresponding amino acid pair

[e] PR D, probably damaging; PO D, possibly damaging.

[f] Ratio of the likelihood of a given amino acid occurring at a particular position to the likelihood of this amino acid occurring at any position

[g] Phenotypic effect of nonsynonymous coding SNPs as predicted by selective pressures estimated at the codon level. Pathological effects imply residues with $\omega < 0.1$.

Legends to Figures

Figure 1: Summary of signatures of selection in the *VPS37C* region. Depicted on top is the gene track along 2 Mb centered on *VPS37C* (highlighted in yellow) and below is a summary of the results of the different types of analyses in this region (*VPS37C* is delimited by the vertical bars in each plot). a and b, Proportion of SNPs with MAF < 0.10 and DAF > 0.80, respectively, within 100 kb sliding windows separated by 20 kb steps in East Asian populations. Solid dots represent values above the 95th percentile for each population, whereas open dots are values below the 95th percentile. c, F_{ST} values between all 39 populations for each SNP over the region. Solid dots represent the top 5% values of the overall F_{ST} distribution obtained across all analyzed regions. d and e, -log p-values of the XP-Rsb statistic for the seven continental regions and the six populations within EASIA, respectively. Horizontal dashed lines indicate statistical significance at the 0.05 level. Solid dots represent the lowest 5% p-values of the genome-wide Rsb distribution within each population (see methods for details).

Figure 2: Distribution of REHH against frequency for populations within six of the main geographical regions studied. Core haplotypes within ±100 kb of the candidate genes (black dots) are plotted over the background distribution of cores from the eleven full 2 Mb regions (gray dots) analyzed. REHH is shown at a distance of about 0.3 cM for all populations. Dashed lines indicate 0.95, 0.99 and 0.999 percentiles of REHH considering all cores. Cores that remained significant after multiple test correction (q < 0.05) are indicated with a black open diamond.

Figure 3: Bifurcation plots and EHH decay over physical distance for the two main haplotypes observed at the significant 3-SNP core of the *VPS37C* genomic region in MENA. On top, boxes represent genes, vertical gray lines are SNPs, vertical blue lines denote those constituting the core and vertical red lines indicate non-synonymous SNPs. Underlined SNPs represent other cores within the region. Gene symbols are shown (from right to left) for the gene containing the core, the gene of interest in the region and the gene with the putatively selected variant in the region.

Figure 4: Worldwide frequency distribution of non-unique haplotypes based on 54 SNPs in the *VPS37C* region. Light blue segments represent the putatively selected haplotype whereas other haplotypes are indicated in other colours.

Legends to Supplementary Figures

Figures S1 – S7: Distribution of low-frequency minor alleles. The proportion of SNPs with minor allele frequency (MAF) less than 0.10 within 100 kb sliding windows is plotted for each genomic region and population. The vertical gray rectangles delimit the physical coordinates of the gene of interest in each region. Solid colored circles represent the top 5% values of the distribution within each population across all regions. Open circles are values below the top 5% and hence are considered non-significant for that population. Gaps are the consequence of sliding windows having less than 5 SNPs, which has been set as the minimum for computing allele frequency proportions.

Figures S8 – S14: Distribution of high-frequency derived alleles. The proportion of SNPs with derived allele frequency (DAF) greater than 0.80 within 100 kb sliding windows is plotted for each genomic region and population. The vertical gray rectangles delimit the physical coordinates of the gene of interest in each region. Solid colored circles represent the top 5% values of the distribution within each population across all regions. Open circles are values below the top 5% and hence are considered non-significant for that population. Gaps are the consequence of sliding windows having less than 5 SNPs, which has been set as the minimum for computing allele frequency proportions.

Figure S15: Global F_{ST} for 39 populations across the eleven genomic regions analyzed. The vertical gray rectangles delimit the physical coordinates of the gene of interest in each region. Solid blue circles represent SNPs with F_{ST} values above the 95th percentile

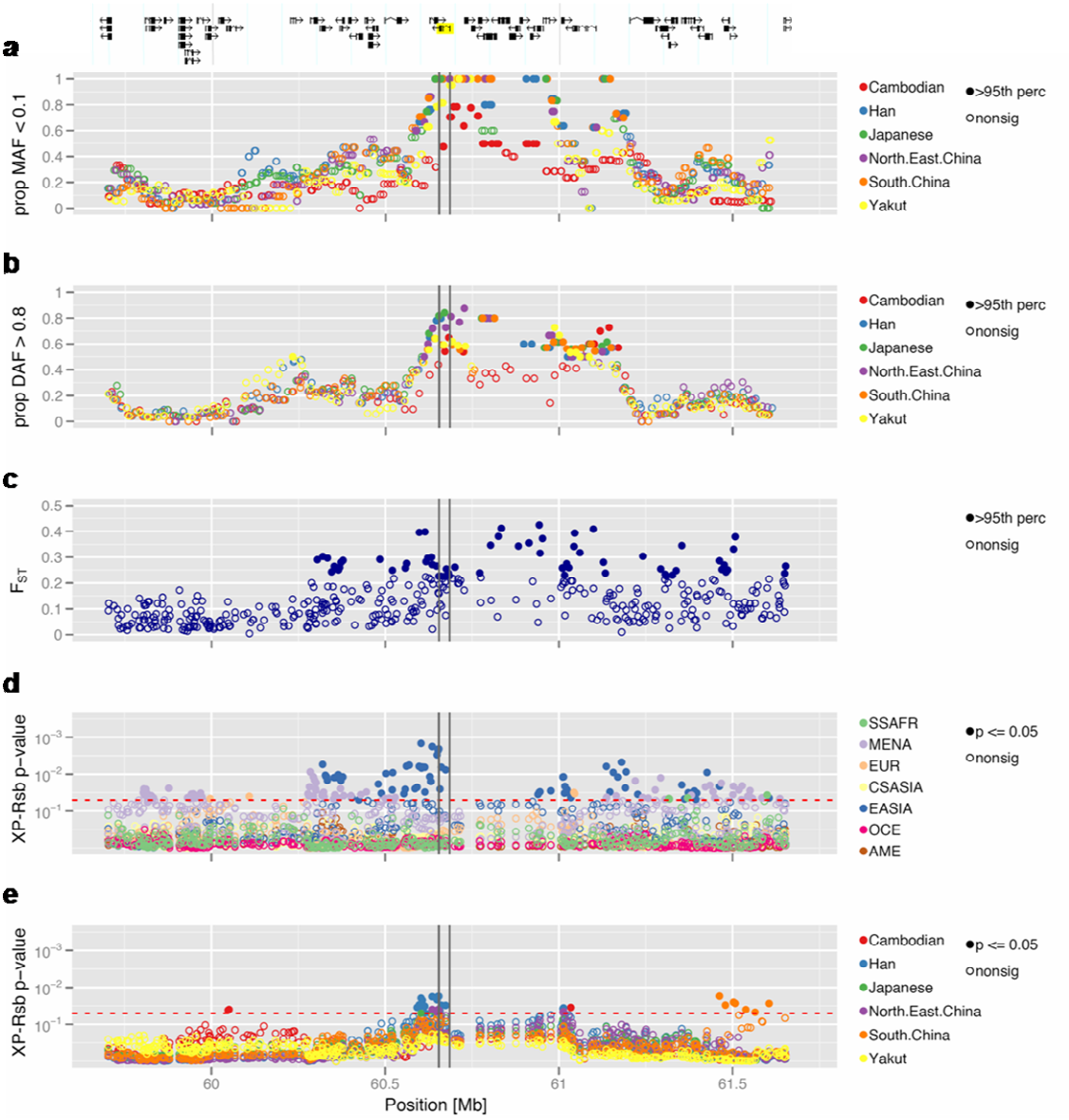
of the distribution across all regions. Open blue circles are SNPs below the 95th percentile and hence are considered non-significant.

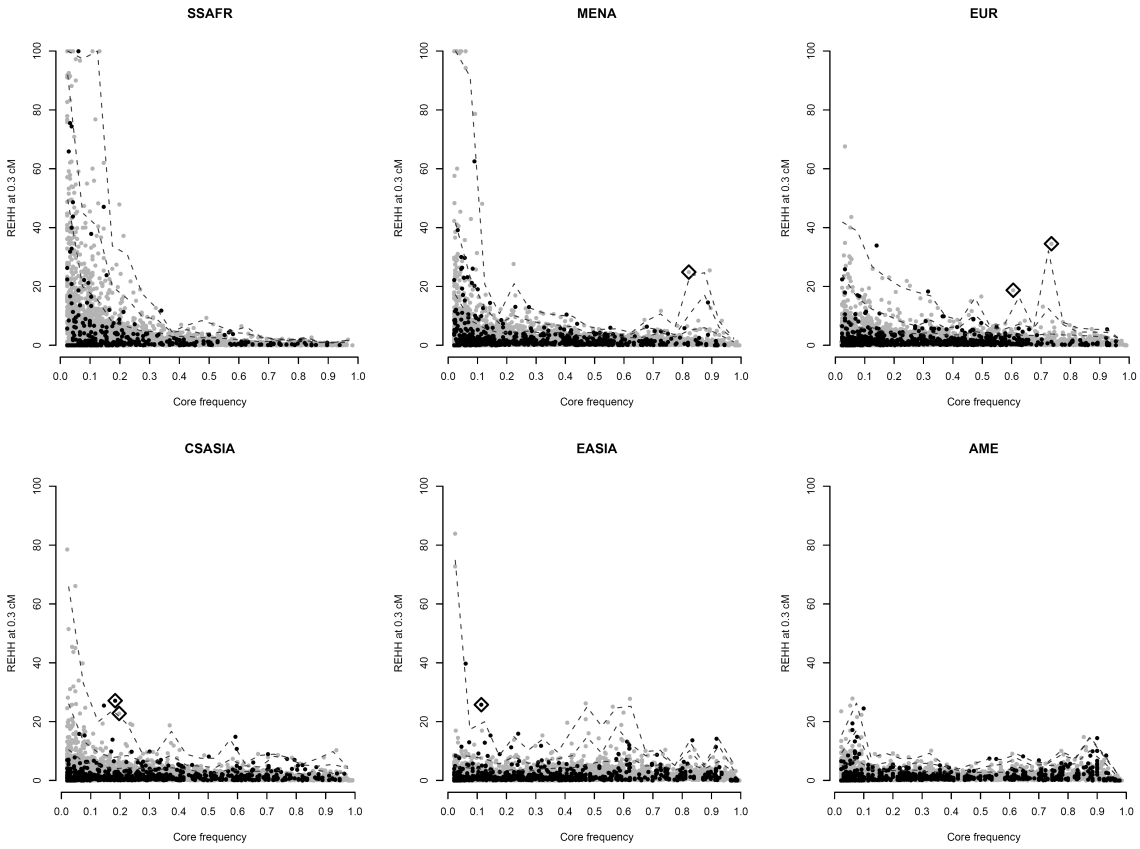
Figure S16: Distribution of REHH against frequency for populations within Middle East-North Africa. Core haplotypes within ±100 kb of the candidate genes (black dots) are plotted over the background distribution of cores from the eleven full 2 Mb regions (gray dots) analyzed. REHH is shown at a distance of about 0.42 cM for all populations included. Dashed lines indicate 0.95, 0.99 and 0.999 percentiles of REHH considering all cores. Cores that remained significant after multiple test correction ($q < 0.05$) are indicated with a black open diamond.

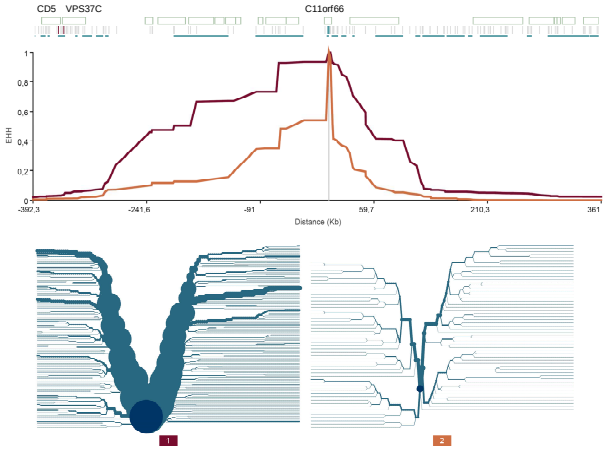
Figure S17: Distribution of –log p-values of XP-Rsb for each major geographic area across the eleven genomic regions analyzed. The genomewide significance of XP-Rsb at every SNP site for each population is plotted against physical distance over the candidate regions. Gray rectangles indicate the location of the gene of interest in each region. Dotted and dashed lines show 0.05 and 0.01 significance levels, respectively. Values above the latter are additionally represented with solid color circles, while open circles indicate values below the 0.01 significance level.

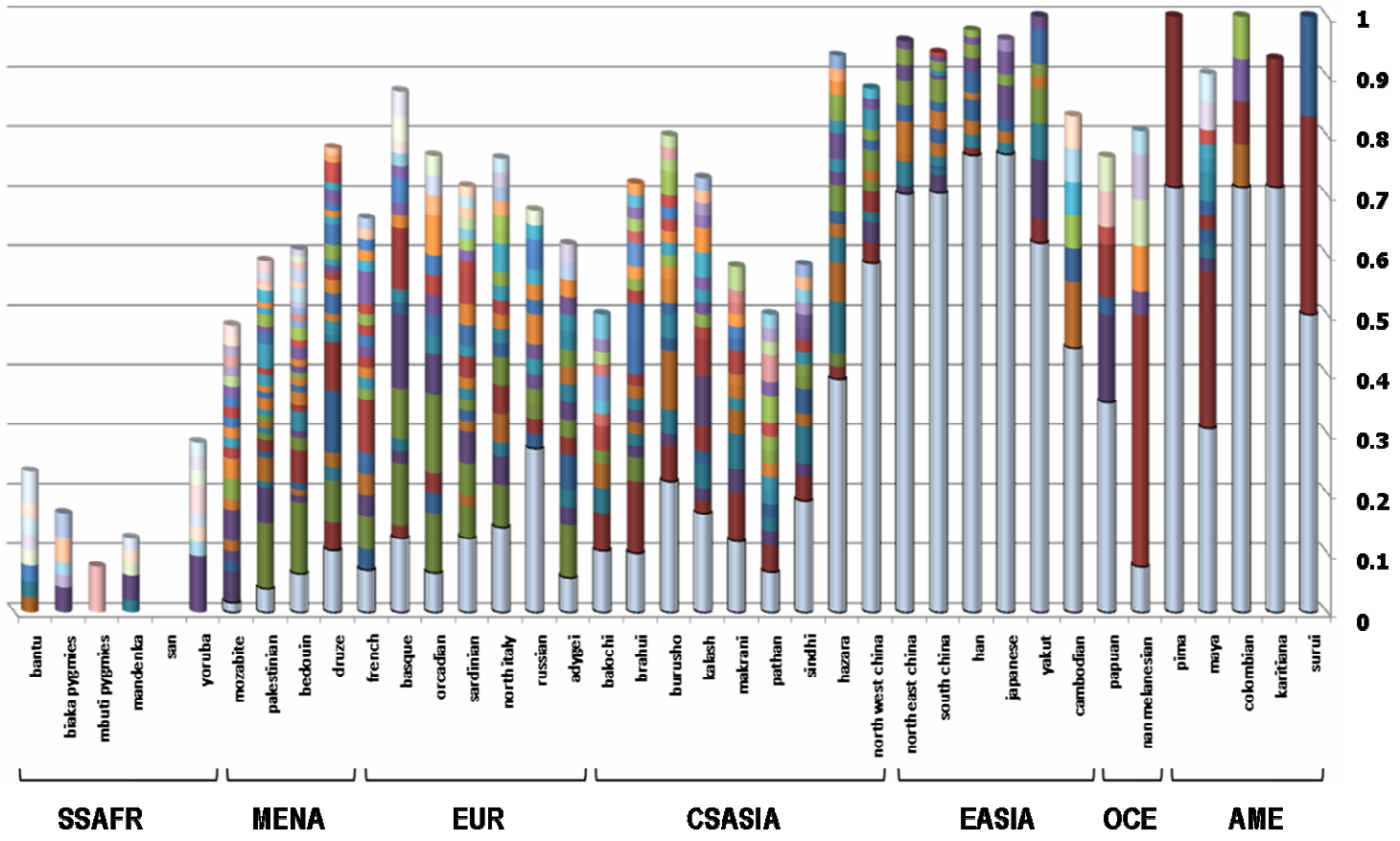
Figures S18 – S28: Distribution of populational –log p-values of XP-Rsb grouped across main geographical regions. The genome-wide significance of XP-Rsb at every SNP site and population is plotted against distance over each candidate region. Gray rectangles indicate the location of the gene of interest in each region. Dotted and dashed lines show 0.05 and 0.01 significance levels respectively. Values above the latter are

additionally represented with solid color circles while open circles indicate values below the 0.01 significance level.


a
prop MAF < 0.1
b
prop DAF > 0.8
c
F_{ST}
d
XP-Rsb p-value
e
XP-Rsb p-value
Position [Mb]
Cambodian
Han
Japanese
North.East.China
South.China
Yakut
SSAFR
MENA
EUR
CSASIA
EASIA
OCE
AME
>95th perc
nonsig
p <= 0.05


SSAFR
MENA
EUR
CSASIA
EASIA
AME
REHH at 0.3 cM
Core frequency
0
20
40
60
80
100
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0


CD5 VPS37C
C11orf66
EHH
Distance (Kb)
1
2


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
bantu
biaka pygmies
mbuti pygmies
mandenka
san
yoruba
mozabite
palestinian
bedouin
druze
french
basque
orcadian
sardinian
north italy
russian
adygei
balochi
brahui
burusho
kalash
makrani
pathan
sindhi
hazara
north west china
north east china
south china
han
japanese
yakut
cambodian
papuan
nan melanesian
pima
maya
colombian
karitiana
surui
SSAFR
MENA
EUR
CSASIA
EASIA
OCE
AME