Internal Report CS Bioinformatics Track 14-01 February 2014

Leiden University

Computer Science

Bioinformatics Track

Identifying soft selective sweeps in the human genome from the spatial distribution of strong selective sweeps

Joeri J. Meijsen *BASc*

Identifying soft selective sweeps in the human genome from the spatial distribution of strong selective sweeps

Joeri J. Meijsen[1], Erwin M. Bakker[1] and Oscar Lao Grueso[2]

[1] Leiden Institute of Advanced Computer Science, Leiden University, Leiden
[2] Department of Forensic Molecular Biology, Erasmus Medical Centre Rotterdam, Rotterdam

Abstract

Since the Out of Africa Diaspora, humans have genetically adapted to the environments they encountered in their migration. As a consequence of this genetic adaptation, the frequency of allelic variants providing a selective advantage to a particular environmental factor have increased in these populations where the environmental factor was present. However, with the exception of rare cases such as milk and selection to adult lactose tolerance phenotype, the environmental factor playing as the selective force is usually unknown and its spatial distribution can only be hypothesized by the spatial distribution of the genetic variants that are under selection.
So far, several genomic regions have been suggested to show the fingerprint of strong selective sweeps. Nevertheless, identifying genetic variants under soft selective pressures, expected under the presence of polygenic adaptation is, more complex.
In the present study we address the question of whether the spatial distribution of genetic variants in genomic areas under strong selective sweeps can help identifying other regions in the genome putatively under soft selective pressures. In this study we introduce a novel statistical analysis pipeline called **Geographic Allelic Association among Populations** (GAAP) to indirectly estimate the evidence of soft selective pressures by analysing the functional correlation between the genes close to genetic variants under strong selective sweeps and the genes close to genetic variants showing a similar spatial pattern as the one under strong selective pressures.
Our results show that there is a statistically significant enrichment of similar biological processes between both categories of genes. These results supports the evidence of undetected soft selective pressures in the genome and suggest that this proposed methodology could be used to identify such selective pressures.

Introduction

According to the Recent Out of Africa (RAO) hypothesis and related ones [1] *Homo sapiens* is a relatively newcomer from an evolutionary point of view. Humans evolved ~200,000 years ago in the African continent and spread ~100,000 years ago out of the African continent, colonizing the whole world in a relatively short amount of time [2]. Genetic variation within the species is estimated to be at least 0.5% [3] of which on average approximately 5% is due to differences among populations, ~15% to differences among continents and ~80% to genetic differences within. Moreover, the genetic variation of different types of genetic variants (autosomal single nucleotide polymorphisms (SNPs) , autosomal haplotypes, autosomal Short Tandem Repeats (STRs)) is larger within the African continent than in any other continent [4-7]. Conversely, the amount of linkage disequilibrium (LD) is smaller within Sub-Saharan African populations compared to the rest of the world [8] (see Figure 1)

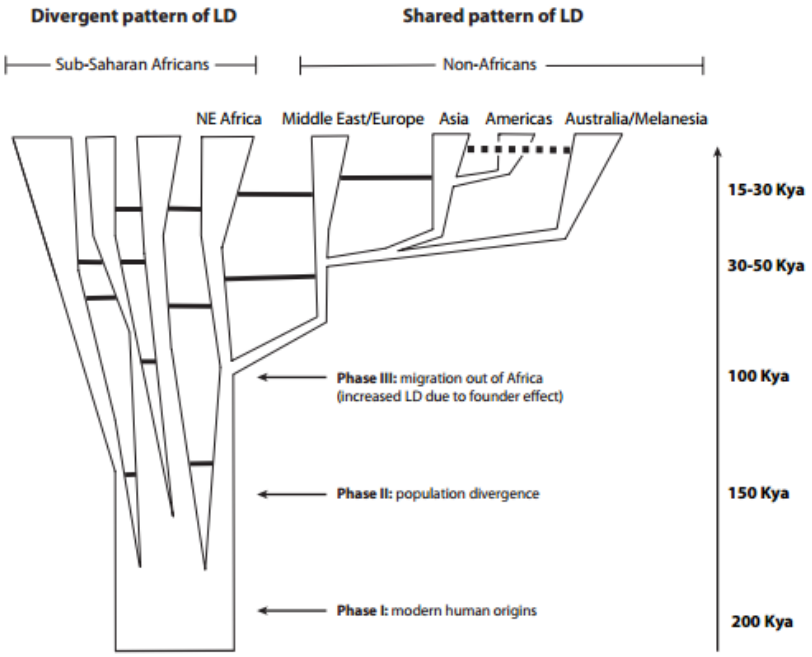

Figure 1. Figure depicting the recent human evolution at a worldwide level and the genomic consequences according to the Out of Africa model [1]. Between ~150-100 kilo years ago (kya), Ancestral Africans experienced an expansion, contraction, migration and admixture which led to large different populations (Phase I and Phase II). Between ~50-100 kya (Phase III), an African subpopulation spread out of the African continent, colonizing the rest of the world. As a result of the Bottleneck effect, the genetic diversity decreased and the amount of LD increased in populations out of the African continent. Solid lines indicate gene flow between populations and the dashed lines indicates recent gene flow from Asia to Australia/Melanesia. Image adapted from Campbell and Tishkoff [2].

Overall, such observations have been explained by traditionally larger effective population sizes within the African continent [2] and the effect of serial founder bottleneck events out of Africa during the human Diaspora [5]. As another consequence of the Out of Africa migration, the amount of genetic differentiation between continental populations correlates with their geographical distance, showing larger differentiation the further the population is geographically located from the suggested RAO's point of origin Addis Ababa (Ethiopia) [9].

During the human Diaspora, populations came in contact to new external factors (such as diet, environmental conditions or diseases among others); it has been suggested that these factors would have worked as a powerful force of evolutionary adaptation [10]. This scenario would have resulted in the increase in frequency of genetic variants distributed through the genome giving individuals a selective advantage of reproducing compared to the non-carriers. So far, several methods for detecting the fingerprint of selection in the genetic variation in the genome of a population or series of populations have been proposed [11,12]. Depending on the timescale of positive selection events, these methods can be differentiated into macroevolutionary (i.e. selective pressures leading to speciation [13]) and microevolutionary (i.e. selective pressures leading to differential population adaptation within the species [13]). Methods focusing on microevolutionary processes use different approaches to detect positive selection in the human genome. Furthermore, depending on how the type of genetic variants are analysed, methods for detecting the fingerprint of recent selective sweeps from microevolution have been classified in three main categories [14]: (i) site-frequency spectrum (SFS) ,(ii) Haplotype and (iii) linkage disequilibrium (LD) methods. SFS based methods aim at identifying genomic regions where the allele frequency pattern in either one, or more populations differs from the overall pattern observed in the whole genome [15]. Examples of SFS methods considering one population include Tajima's D [16] and DH test [17], among others. A particular case of SFS among populations is the fixation index (Fst) [18], which estimates the difference in the allelic frequencies of one SNP among different populations. Haplotype based approaches use the frequency and the length of particular allelic combinations or haplotypes defined in a genomic region to detect

an excess of frequency differentiation and/or longer haplotypic tracks than expected by neutrality. Finally, LD based methods use the expected properties of the decay of LD by recombination with the increase of genomic distance from a particular SNP and its expected allelic frequencies under neutrality [19]. The Extended Haplotype Homozogysity (EHH) [10] and the Integrated Haplotype Score (IHS) [20] among others are examples of LD methods. Nowadays, composite likelihood methods combining the properties of the three type of approaches have been proposed (i.e. Cross Population (XP)-EHH [11]). Nevertheless, the selective sweeps detected when using whole genome data tend to show little overlap among methods [21]. These discrepancies have been explained by different factors [22]. First, the SNP ascertainment bias introduced during marker discovery [23] may lead to false positives in haplotype and LD based tests. Second, in regional comparison methods (i.e. LD) the size of the region can influence the outcome of the result. The third, and most likely reason for discrepancies, may be the different evolutionary assumptions of each model including the different evolutionary timescales of the selective event [11]. Beside these caveats, several regions under strong positive selective sweeps have been proposed in the human genome [4,21,24]. Furthermore, bioinformatics predictions and/or empirical evidence supporting the functional role of genetic variants in regions under positive selection have been reported. In particular, several SNPs showing signals of positive selection lead to amino acid changes which are *in silico* predicted to alter the function of the protein [24]. Functional empirical evidence have been provided in particular cases such as LCT [25], OCA2-HERC2 [26] or EDAR [27] genes among others [28].

However, some authors suggest that classical selective sweeps such as the ones detected by the previously introduced methods are rare in the human genome [29]. Furthermore, it has been suggested that polygenic adaptation of complex traits can be more important than single locus adaptation [30]. Nevertheless, identifying the fingerprint of polygenic adaptation in the human genome is difficult and so far few methods attempt to address some of the aspects of polygenic adaptation in the genome. It has been suggested that polygenic adaptations can produce partial selective sweeps in some of the involved loci, which could be detected by currently available methods [31]. Furthermore, genetic variants under positive selective pressures tend to geographically covariate with the spatial distribution of the selective factor [7], which in turn tends to follow particular geographic patterns. This is particularly evident in the case of skin pigmentation trait, where the skin colour strongly correlates with the latitude at a worldwide level [32]. In general, the geographic distribution of the selective factor is rarely known [33]. Moreover, the currently distribution of putative environmental factors can differ from past ones at the time when the selective pressure took place. For example, in the case of the EDAR gene, the phenotypic changes associated to the 370A SNP suggest a type of selective pressure by environmental conditions (high humidity, especially in summers) that are not currently present in East China. However, the fact that polygenic adaptation could be potentially detected as strong selective sweeps in at least one genomic region, and the suggested dependence between spatial distribution of the selective factor and the genetic variants under positive selective pressure could be in principle exploited to identify new functional genomic variants or regions under polygenic adaptation.

Material and Methods

I_{nA} statistic

The association between the allelic frequencies of two SNPs among a set of populations is quantified by modifying the Informativeness of ancestry (I_n) statistic [34]. Rosenberg's I_n was designed to quantify the amount of differentiation among K populations using N genetic variants. Using Rosenberg's notation, I_n is defined as (1):

$$(1)\quad I_n = \sum_{j=1}^{N}\left(-p_j \log(p_j) + \sum_{i=1}^{K}\frac{p_{ij}\ \log(p_{ij})}{K}\right)$$

Where N represents the number of alleles (N = 2 in this case, since we consider biallelic SNPs), K represents the number of populations, p_{ij} is the frequency of allele *j* in the population *i* and p_j is the mean allele frequency of allele *j* among all the populations:

$$p_j = \frac{\sum_{i=1}^{K} p_{ij}}{K}$$

In case of two biallelic SNPs (N = 2) whose allelic frequencies have been estimated in K populations, an association between these SNPs is defined, if the frequency of the alleles of SNP A predicts the frequency of the alleles in SNP B among the populations. Using the I_n framework, this association can be estimated by (2):

$$(2)\quad I_{nA} = \sum_{j=1}^{N}\left(-p_j \log(p_j) + \sum_{t=1}^{N}\frac{p_{tj}\ \log(p_{tj})}{N}\right)$$

Where p_{tj} is the weighted frequency (p) of allele j of SNP B over all the populations given the allelic frequency (q) of allele t of SNP A:

$$p_{tj} = \frac{\sum_{i=1}^{K} q_{it}p_{it}}{\sum_{i=1}^{K} q_{it}}, \text{where}$$

q_{it} is the allelic frequency *t* of the *A* SNP at population *i* and p_{it} is allelic frequency *j* of the B SNP for each population *i*. The mean allele frequency of allele *j* of SNP A overall populations is given by:

$$p_j = \frac{1}{2}\sum_{t=1}^{2} p_{tj},$$

Because the I_{nA} is an extension of the I_n it inherits its basic properties. Namely, it ranges from 0 to ln(2) where 0 means no association between A and B and ln(2) is obtained when both SNPs covariate perfectly at a population level.

Local Moran's I

We expect that SNPs in close proximity to each other on the genome tend to show similar I_{nA} values compared to the rest of the genome due to linkage disequilibrium (LD) [35]. To measure this local spatial autocorrelation, we use the Local Moran's I (3), based on the global spatial autocorrelation Moran's statistic [36]:

$$(3)\ L_{mi} = \frac{1}{m-1} Z_i \cdot \sum_{j=\, i-\frac{(m-1)}{2}}^{i+\frac{(m-1)}{2}} Z_j \text{, where}$$

i is the SNP of interest, *m* is the number of markers in the window surrounding SNP *i*, *j* are the SNPs in the window and Z is the standardized measurement of the I_{nA} variable:

$$Z_i = \frac{l_{nA_i} - \mu_{l_{nA}}}{\sqrt{\left(\frac{1}{N-1}\sum_{i=1}^{n}\left(l_{nA_i} - \mu_{l_{nA}}\right)^2\right)}} \text{, where}$$

I_{nA} is the informativeness of ancestry of a given SNP *i*, μ_{lnA} the mean I_{nA} over all SNPs and *N* is the number of SNPs considered.

A positive autocorrelation, $L_{mi} > 0$, indicates that the SNP is surrounded by SNPs with similar Z (I_{nA}) values with regards to the rest of the genome. A negative autocorrelation, $L_{mi} < 0$, is found when the opposite occurs. When $L_{mi} = 0$ the SNP is in a neighbourhood with dissimilar Z values based on I_{nA} values.

The I_{nA} and L_{mi} statistics are implemented in a R pipeline called *Geographic Allelic Association among Populations* (GAAP). The complete R package "*PopStat*" is described in the supplementary information and can be freely obtained by request to the first author.

Databases

In this study we used 3 previously published genome-wide SNP data sets and 1 in-house dataset together comprising a total of 2,415 individuals from 85 worldwide populations [4,37,38] (see Table S1,Table S2 and Table S3 in supplemental information for detailed information) and 486,035 SNPs after SNP merging. Data cleaning included exclusion of SNPs with a Minor Allele Frequency (MAF) < 0.01 (519 SNPs), exclusion of individuals from recently admixed populations (Surinam and Mexico) and 196 reported as related individuals. Furthermore, five populations (see Table S3) were renamed and merged with other populations. The final dataset comprised 78 populations and 488,503 SNPs.

Functional Analysis

SNPs in genomic regions showing selective sweeps (aka Proxy SNPs) were ascertained from the genomic regions described in Grossman et al [24] for YRI, CEU and East Asian (CHB and JPT) HapMap [39] populations. Out of the 412 genomic regions, 177 did not contain any so far described gene and were removed from further analyses. 2366 SNPs in the merged dataset could be mapped to 1 or more genomic regions under selective pressures (Figure 2 Step 1). Because this study is only focussing on within gene SNPs, we excluded the SNPs that could not be mapped to genes ascertained from the UCSC database (http://genome.ucsc.edu/) (Figure 2 Step 2). Also SNPs that could be mapped to UCSC genes but could not be mapped to Ensembl Transcript IDs were excluded (Figure 2 Step 3). The GAAP pipeline was applied to each of the 2301 remaining SNPs (Figure 2 Step 4). For each SNP the top 100

SNPs showing the strongest L_{mi} genomic association were identified and selected. From these top 100 SNPs, the SNPs not in the same genomic region showing selective sweeps as the Proxy SNP were selected (aka Association SNPs) (Figure 2 Step 5). In order to recover genes genomically close to each SNP, a +/-100 kilobase (kb) wide window surrounding each Proxy and Associated SNPs was considered, and genes either within or partially overlapping such window were ascertained from the UCSC database(Figure 2 Step 6) and subsequently converted to Ensembl Transcript IDs (Figure 2 Step 7). All Ensembl Transcript IDs were then functionally annotated with the DAVID 6.7 database (http://david.abcc.ncifcrf.gov/) to get the Gene Ontology (GO) [40] terms related to biological processes (Figure 2 Step 8).

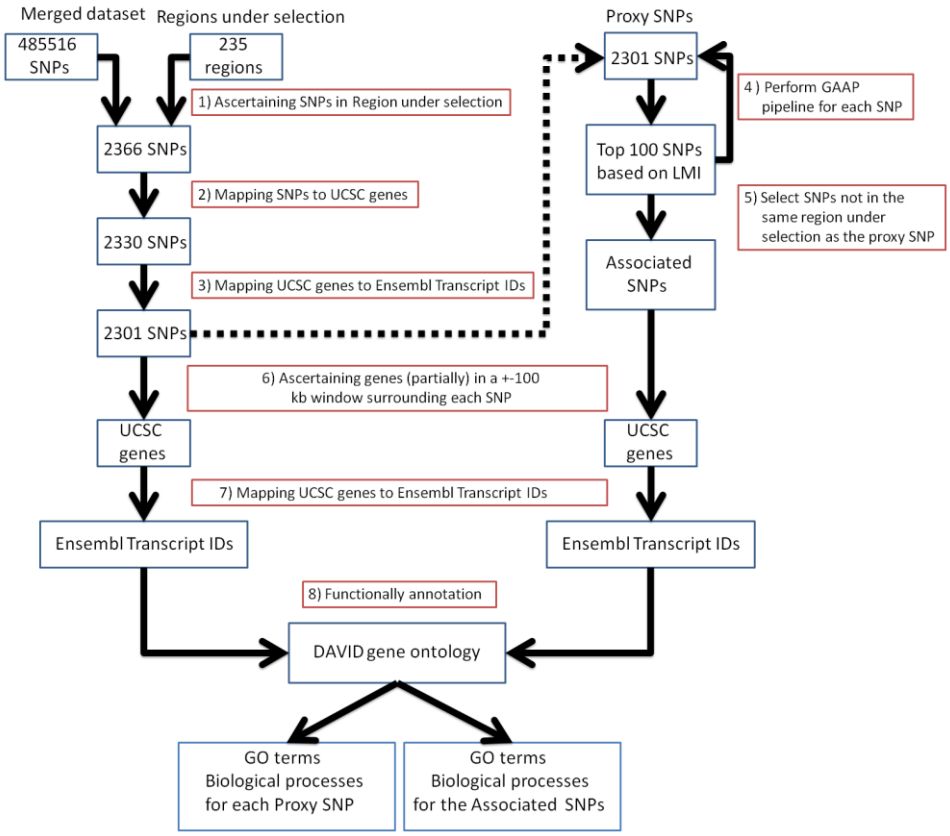

Figure 2. Outline of the Functional Analysis procedure. **1)** Ascertainment of SNP in the Regions under selection. **2)** Mapping ascertained SNP to known UCSC genes. Unmapped SNPs were removed. **3)** Mapping UCSC genes to Ensembl Transcript IDs. SNPs in unmapped UCSC genes were removed. **4)** Performing the GAAP pipeline for each Proxy SNP and selection of top 100 SNPs based on the L_{mi}. **5)** SNPs not in the same genomic region under selection as the Proxy SNPs were selected. **6)** For each Proxy SNP and its Associated SNPs a +-100 kb surrounding window was defined and each UCSC gene (partially) in the window was ascertained. **7)** Mapping UCSC genes to Ensembl Transcript IDs (see step 3). **8)** Performing functional annotation analysis for the Proxy SNP and its Associated SNPs separate to ascertain GO terms related to Biological processes.

Statistical Analyses

From the genes present in each proxy and associated SNPs window, a two by two contingency table was created, counting the number of shared GO terms, GO terms present in either the proxy or the associated SNPs, and GO terms absent in the genes of both the proxy and the associated SNPs (i.e. rest of the genome). The positive association (one tail p-value) of this two by two contingency table was estimated by means of Fisher exact test. The genomic statistical significance of this association was estimated by randomly ascertaining 1000 sets of 100 SNPs from the genome and for each set repeating the process of identifying the genes in the surrounding 200kb window region, retrieving gene transcript IDs and GO terms and computing the two by two contingency table. One tail genomic

p-value for each proxy SNP was then computed by comparing how many times the observed GO association with the genes in the vicinity of the associated SNPs was smaller than the obtained with the randomly sampled SNPs.

All the GO terms from Proxy SNPs showing a significant association with its Associated SNPs were ascertained for further enrichment analysis against the GO terms over the whole genome in order to statistically quantify which GO terms were enriched in the set of proxy SNPs and Associated SNPs showing a p value < 0.05. Positive enrichment was calculated by means of a complementary cumulative distribution function (1 tailed p-value) using the observed GO term counts, total amount of observed GO terms and the frequency of the GO terms in the whole genome. To account for multiple testing problem, a Bonferroni correction was performed on each p-value.

Results and Discussion

Whereas currently available whole genome scan methods have identified a relatively large number of regions in the human genome showing strong signals of selective sweeps [24], detecting the fingerprint of recent polygenetic adaptation in the genome has so far revealed to be highly complex. Previous studies ([41,42] among others) have identified additional regions under selective pressures by analysing the correlation of environmental variables (i.e. latitude) or phenotypes of interest (i.e. height) with the current genetic variation among populations. However, the traits under selection and/or the spatial distribution of the selective force are usually unknown [33]. Nevertheless, it can be assumed that the geographic distribution of genetic variants showing signatures of a strong selective sweep, such as the ones identified by currently available methods, is going to be similar to the one of the selective factor. Since genetic variants under polygenic adaptation should also tend to follow a geographic distribution similar to the spatial distribution of the trait under selection, in this study we hypothesize that genetic variants showing the same spatial distribution as variants identified as being under strong selective sweeps could be also under selective pressures (see Figure 3). These variants, which would not have been detected by means of currently tests for detecting positive selection, would indicate soft selective sweeps and could suggest the presence of polygenic adaptation.

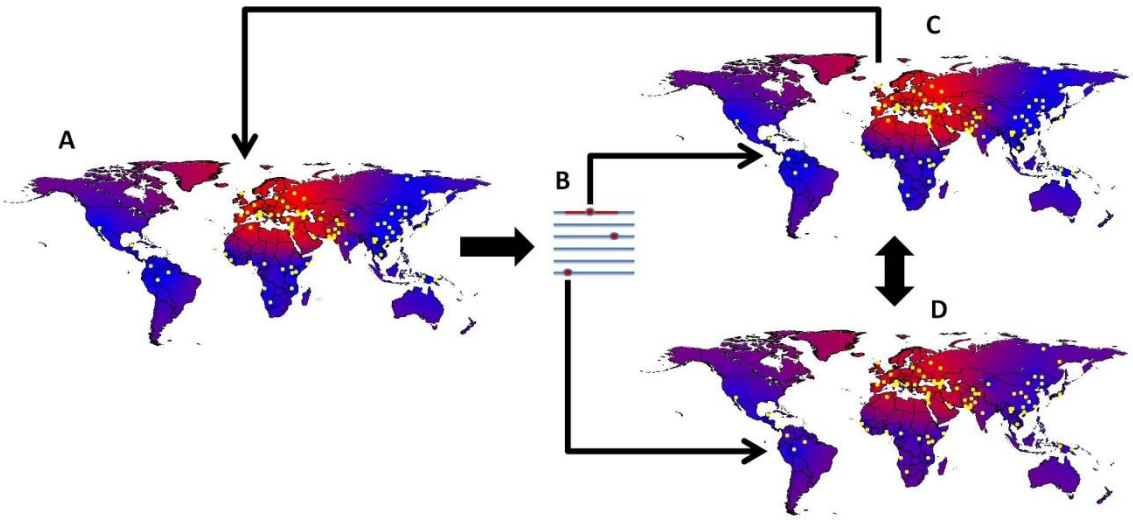

Figure 3. A) Map of the hypothetical geographic distribution of a selective force (usually unknown). **B)** Genetic variant under strong selective sweeps (top variant) and two additional variants under selection which do not show the fingerprint of strong selective sweeps. **C)** Map of the geographical distribution of the allelic frequency of the variant under strong selective sweeps. The distribution strongly resembles the map of the selective force. **D)** Map of the geographical distribution of the allelic frequency of the variant under selective pressure not showing a strong selective sweep. The distribution resembles the map of the selective sweep C.

We recovered the genetic variation of 488,503 SNPs in 78 human populations [4,37,38] and used the spatial distribution of the genetic variation in the suggested genomic regions under strong selective sweeps identified in Grossman et al [24] as proxies of the selective factor.
The sub-continental distribution of these 78 worldwide populations is not at random (Chi square test between observed and expected, p-value = 1.409e-06); North America, South America and Oceania are poorly geographically sampled, whereas Central/South Asia, East Asia and Europe show an excess of sampled populations. In principle, better performance of our approach can be expected by using a more dense and homogeneous geographic coverage. Nevertheless, in this particular case, the expected improvement is limited by the fact that the genomic regions used as proxy of the selective factor were identified in populations (HapMap CEU, CHB and YRI) from the continents that are overrepresented in our dataset. The reason for using such biased geographic signals of selection in contrast to other studies [43] is that Grossman et al [24] study is so far the most comprehensive proposed dataset of signals of recent selection in the human genome; it uses full genome sequence variation from the 1000 Genomes and the composite of multiple signals (CMS) test [43]. Of the 412 genomic regions identified in Grossman et al [24], 177 did not contain any known genes and were excluded from further analyses. Out of the 488,156 SNPs, 2366 SNPs could be mapped to one or more regions. Of these, 2330 SNPs had at least one UCSC gene at a distance <100 kb, and 2301 could be mapped to Ensembl Transcript IDs (see Figure 2 point 6 and 7).

We applied the GAAP pipeline for each of the 2301 proxy SNPs and selected from the genome the 100 SNPs showing the highest genomic association based on the I_{nA} and L_{mi} statistics, excluding SNPs in the same region under selective pressures as the proxy SNP. A first disadvantage of this approach is that the magnitude of the estimated association can differ among the proxy SNPs, so in some cases less than 100 SNPs would be enough (False Positives). A second problem of this procedure is that in some cases more than 100 SNPs can be associated, which results in exclusion of possibly informative SNPs (True Positives). Therefore, the approach we applied to detect soft selective sweeps can be considered as extremely conservative.

For each proxy SNP, and for each of its associated SNP, a 100kb window surrounding the SNP was defined. We found that of the 2301 proxy SNPs, 252 did not contain genes with any known GO terms related to biological processes and were therefore excluded for further analysis.

We identified 664 out of 2049 SNPs with a significant genome association ($p < 0.05$) between GO terms of the proxy and these from the associated SNPs (see Table S4 in supplementary information). The one tail probability of observing at p=0.05 the same or a higher number of significant associations among the analysed SNPs is $P(\geq 664) < 10^{-773}$, thus indicating a statistically significant enrichment of identical biological GO terms among genomic regions whose genetic variation shows a similar geographical pattern than these under strong selective sweeps. Since the presence of population substructure has been interpreted as a signal of positive selection in a locus [44] but it could be that the SNPs present in the regions under selection were not representative of the whole genetic variation of the region, we further analysed to which extent the proxy SNPs showing an empirical significant genomic enrichment of GO terms at $p<0.05$ had a more structured distribution among populations than the non-significant ones. We observe that the I_n distribution of SNPs showing a significant empirical association is statistically significantly higher than the SNPs not showing a statistically significant GO enrichment (Wilcoxon signed-rank test one tail p-value = 0.000152; see Figure 4).

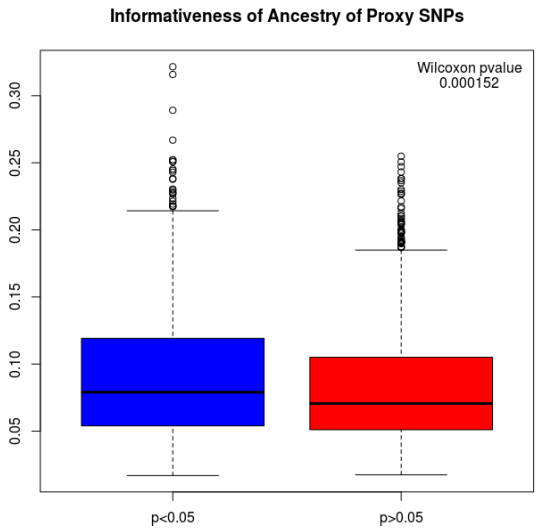

Figure 4. Boxplot of the Informativeness of ancestry (I_n) statistic computed at each Proxy SNPs and the genomic GO term association between the GO terms recovered from the Proxy SNP and its Associated SNPs. Left blue box comprises In values of the Proxy SNPs that show a statistically significant genomic enrichment of GO terms with the associated SNPs, whereas the right red box indicates Proxy SNPs showing a non-significant enrichment. The median I_n value of the blue box is statistically significant higher compared to the one of the red box (Wilcoxon signed-rank test one tail P value = 0.000152).

Taken together, these results suggest a functional association between genes that are located close to SNPs whose genetic variation shows a similar geographic distribution as SNPs in regions showing signals of strong selective sweeps, and the genes that are in these genomic regions under selective pressures.
We then focused on which GO terms were enriched among the genes close to the proxy SNPs and to the genes close to the geographically associated SNPs. Even with the highly conservative Bonferroni P value correction, 164 GO terms are found to be enriched ($p<0.05$) in the Proxy SNPs compared to the whole genome (Table S5 as separate file in the supplementary information). A quick inspection showed that a high amount of enriched GO terms are in some way involved in the Immune system (i.e. GO:0002327, GO:0002702 and GO:0002637), Muscle development (i.e. GO:0048625 and GO:0048745) and Sensory perception (i.e. GO:0007605 and GO:0050953) among others. So far, studies [24,45,46] have reported similar results in the case of the Immune System and Sensory perception. Herráez explained the overrepresentation of these classes due to how humans interact with their environment, especially in the case of pathogens and diet.

Two possible evolutionary explanations could produce these observed results. The first is that different environmental factors showing a similar geographic distribution could apply similar selective pressures. One example of such situation could be two different diseases with a similar geographic distribution but affecting different pathways of the immune system. A second explanation is that the recovered functional signals refer to the same phenotype and pathway, thus indicating the presence of polygenic adaptation. Disentangling between both in humans is currently cumbersome, as the human protein pathways are not really well known and human protein pathway databases still contain numerous incongruence's (i.e. see [47]).

Conclusions

Identifying the fingerprint of polygenic adaptation in the human genome has been proven to be highly difficult. Here we have introduced a novel approach for identifying genetic variants that show a similar geographic distribution as these that are in regions showing evidence of strong selective sweeps. Our results have shown that there is a statistically significant positive relationship of functional categories between genes that are close to genetic variants showing strong selective sweeps and these genes close to variants showing a similar geographic pattern than these under

selective pressures. Given current status of human protein databases, we cannot rule out the possibility that the observed association is due to multiple phenotypes showing the same geographic distribution and affecting the same functional categories of proteins. Nevertheless, an additional explanation is that these recovered signals are due to polygenic adaptation of genes involved in the same phenotype. Further analyses will be required to disentangle between these two hypotheses and to functionally identify the genetic variants in the newly identified regions.
In any case, to the best of our knowledge, this is the first time that empirical evidence of the detection of soft selective sweeps acting in the human genome is provided.

References

1 Stringer, C. Modern human origins: progress and prospects. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **357**, 563-579, doi:10.1098/rstb.2001.1057 (2002).

2 Campbell, M. C. & Tishkoff, S. A. African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. *Annual review of genomics and human genetics* **9**, 403-433, doi:10.1146/annurev.genom.9.081307.164258 (2008).

3 Levy, S. *et al.* The diploid genome sequence of an individual human. *Plos Biol* **5**, e254, doi:10.1371/journal.pbio.0050254 (2007).

4 Li, J. Z. *et al.* Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**, 1100-1104, doi:DOI 10.1126/science.1153717 (2008).

5 Ramachandran, S. *et al.* Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 15942-15947, doi:10.1073/pnas.0507611102 (2005).

6 Rosenberg, N. A. *et al.* Genetic structure of human populations. *Science* **298**, 2381-2385 (2002).

7 Wang, C., Zollner, S. & Rosenberg, N. A. A quantitative comparison of the similarity between genes and geography in worldwide human populations. *PLoS genetics* **8**, e1002886, doi:10.1371/journal.pgen.1002886 (2012).

8 Gabriel, S. B. *et al.* The structure of haplotype blocks in the human genome. *Science* **296**, 2225-2229, doi:10.1126/science.1069424 (2002).

9 Handley, L. J. L., Manica, A., Goudet, J. & Balloux, F. Going the distance: human population genetics in a clinal world. *Trends Genet* **23**, 432-439, doi:DOI 10.1016/j.tig.2007.07.002 (2007).

10 Sabeti, P. C. *et al.* Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**, 832-837, doi:Doi 10.1038/Nature01140 (2002).

11 Sabeti, P. C. *et al.* Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**, 913-U912, doi:Doi 10.1038/Nature06250 (2007).

12 Nielsen, R., Hellmann, I., Hubisz, M., Bustamante, C. & Clark, A. G. Recent and ongoing selection in the human genome. *Nature reviews. Genetics* **8**, 857-868, doi:10.1038/nrg2187 (2007).

13 Reznick, D. N. & Ricklefs, R. E. Darwin's bridge between microevolution and macroevolution. *Nature* **457**, 837-842, doi:10.1038/nature07894 (2009).

14 Zeng, K., Mano, S., Shi, S. & Wu, C. I. Comparisons of site- and haplotype-frequency methods for detecting positive selection. *Molecular biology and evolution* **24**, 1562-1574, doi:10.1093/molbev/msm078 (2007).

15 Nielsen, R. *et al.* Genomic scans for selective sweeps using SNP data. *Genome Res* **15**, 1566-1575, doi:Doi 10.1101/Gr.4252305 (2005).

16 Tajima, F. Statistical-Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. *Genetics* **123**, 585-595 (1989).
17 Zeng, K., Fu, Y. X., Shi, S. H. & Wu, C. I. Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics* **174**, 1431-1439, doi:DOI 10.1534/genetics.106.061432 (2006).
18 Weir, B. S. & Cockerham, C. C. Estimating F-Statistics for the Analysis of Population-Structure. *Evolution* **38**, 1358-1370, doi:Doi 10.2307/2408641 (1984).
19 McVean, G. The structure of linkage disequilibrium around a selective sweep. *Genetics* **175**, 1395-1406, doi:10.1534/genetics.106.062828 (2007).
20 Voight, B. F., Kudaravalli, S., Wen, X. Q. & Pritchard, J. K. A map of recent positive selection in the human genome. *Plos Biol* **4**, 446-458, doi:ARTN e72 DOI 10.1371/journal.pbio.0040072 (2006).
21 Liu, X. *et al.* Detecting and characterizing genomic signatures of positive selection in global populations. *American journal of human genetics* **92**, 866-881, doi:10.1016/j.ajhg.2013.04.021 (2013).
22 de Gruijter, J. M. *et al.* Contrasting signals of positive selection in genes involved in human skin-color variation from tests based on SNP scans and resequencing. *Investigative genetics* **2**, 24, doi:10.1186/2041-2223-2-24 (2011).
23 Clark, A. G., Hubisz, M. J., Bustamante, C. D., Williamson, S. H. & Nielsen, R. Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res* **15**, 1496-1502, doi:Doi 10.1101/Gr.4107905 (2005).
24 Grossman, S. R. *et al.* Identifying recent adaptations in large-scale genomic data. *Cell* **152**, 703-713, doi:10.1016/j.cell.2013.01.035 (2013).
25 Ingram, C. J., Mulcare, C. A., Itan, Y., Thomas, M. G. & Swallow, D. M. Lactose digestion and the evolutionary genetics of lactase persistence. *Human genetics* **124**, 579-591, doi:10.1007/s00439-008-0593-6 (2009).
26 Visser, M., Kayser, M. & Palstra, R. J. HERC2 rs12913832 modulates human pigmentation by attenuating chromatin-loop formation between a long-range enhancer and the OCA2 promoter. *Genome Res* **22**, 446-455, doi:10.1101/gr.128652.111 (2012).
27 Kamberov, Y. G. *et al.* Modeling Recent Human Evolution in Mice by Expression of a Selected EDAR Variant. *Cell* **152**, 691-702, doi:DOI 10.1016/j.cell.2013.01.016 (2013).
28 Sturm, R. A. Molecular genetics of human pigmentation diversity. *Human molecular genetics* **18**, R9-17, doi:10.1093/hmg/ddp003 (2009).
29 Hernandez, R. D. *et al.* Classic selective sweeps were rare in recent human evolution. *Science* **331**, 920-924, doi:10.1126/science.1198878 (2011).
30 Pritchard, J. K. & Di Rienzo, A. Adaptation - not by sweeps alone. *Nature reviews. Genetics* **11**, 665-667, doi:10.1038/nrg2880 (2010).
31 Pritchard, J. K., Pickrell, J. K. & Coop, G. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Current biology : CB* **20**, R208-215, doi:10.1016/j.cub.2009.11.055 (2010).
32 Jablonski, N. G. & Chaplin, G. The evolution of human skin coloration. *Journal of human evolution* **39**, 57-106, doi:10.1006/jhev.2000.0403 (2000).
33 Novembre, J. & Di Rienzo, A. Spatial patterns of variation due to natural selection in humans. *Nature Reviews Genetics* **10**, 745-755, doi:Doi 10.1038/Nrg2632 (2009).
34 Rosenberg, N. A., Li, L. M., Ward, R. & Pritchard, J. K. Informativeness of genetic markers for inference of ancestry. *American journal of human genetics* **73**, 1402-1422, doi:10.1086/380416 (2003).
35 Reich, D. E. *et al.* Linkage disequilibrium in the human genome. *Nature* **411**, 199-204, doi:10.1038/35075590 (2001).
36 Anselin, L. Local Indicators of Spatial Association - Lisa. *Geogr Anal* **27**, 93-115 (1995).
37 Altshuler, D. M. *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52-58, doi:Doi 10.1038/Nature09298 (2010).

38 Yunusbayev, B. *et al.* The Caucasus as an Asymmetric Semipermeable Barrier to Ancient Human Migrations (vol 29, pg 359, 2012). *Molecular biology and evolution* **29**, 1891-1891, doi:DOI 10.1093/molbev/mss141 (2012).
39 Altshuler, D. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073, doi:Doi 10.1038/Nature09534 (2010).
40 Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nature genetics* **25**, 25-29 (2000).
41 Hancock, A. M. *et al.* Adaptations to climate-mediated selective pressures in humans. *PLoS genetics* **7**, e1001375, doi:10.1371/journal.pgen.1001375 (2011).
42 Mendizabal, I., Marigorta, U. M., Lao, O. & Comas, D. Adaptive evolution of loci covarying with the human African Pygmy phenotype. *Human genetics* **131**, 1305-1317, doi:10.1007/s00439-012-1157-3 (2012).
43 Grossman, S. R. *et al.* A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science* **327**, 883-886, doi:10.1126/science.1183863 (2010).
44 Barreiro, L. B., Laval, G., Quach, H., Patin, E. & Quintana-Murci, L. Natural selection has driven population differentiation in modern humans. *Nature genetics* **40**, 340-345, doi:10.1038/ng.78 (2008).
45 Casto, A. M. & Feldman, M. W. Genome-wide association study SNPs in the human genome diversity project populations: does selection affect unlinked SNPs with shared trait associations? *PLoS genetics* **7**, e1001266, doi:10.1371/journal.pgen.1001266 (2011).
46 Herraez, D. L. *et al.* Genetic Variation and Recent Positive Selection in Worldwide Human Populations: Evidence from Nearly 1 Million SNPs. *Plos One* **4**, doi:Artn E7888 Doi 10.1371/Journal.Pone.0007888 (2009).
47 Dall'Olio, G. M., Bertranpetit, J. & Laayouni, H. The annotation and the usage of scientific databases could be improved with public issue tracker software. *Database-Oxford*, doi:ARTN baq035 DOI 10.1093/database/baq035 (2010).

Supplementary information

Sample description	Populations	Sample size	Platform	Reference
HGDP-CEPH	53	940	Illumina 650Y	Li et al, 2008
Caucasus	15	204	Illumina 610K	Yunusbayev et al, 2011
Hapmap phase 3	11	1184	Illumina Human1M/ Affymetrix SNP 6.0	Altshuler et al,200
In house	6	87	Illumina 650Y/ Illumina Human1M	N.A

Table S1

Population	# Samples	Dataset	Population	# Samples	Dataset
Adygei	17	HGDP-CEPH	Mbuti Pygmy	13	HGDP-CEPH
Balochi	24	HGDP-CEPH	Melanesian	10	HGDP-CEPH
Bantu Kenya	11	HGDP-CEPH	Miao	10	HGDP-CEPH
Bantu South East	5	HGDP-CEPH	Mongola	9	HGDP-CEPH
Bantu South West	3	HGDP-CEPH	Mozabite	29	HGDP-CEPH
Basque	24	HGDP-CEPH	Naxi	8	HGDP-CEPH
Bedouin	46	HGDP-CEPH	Orcadian	15	HGDP-CEPH
Biaka Pygmy	21	HGDP-CEPH	Oroqen	9	HGDP-CEPH
Brahui	25	HGDP-CEPH	Palestinian	46	HGDP-CEPH
Burusho	25	HGDP-CEPH	Papuan	17	HGDP-CEPH
Cambodian	10	HGDP-CEPH	Pathan	22	HGDP-CEPH
Colombian	7	HGDP-CEPH	Pima	14	HGDP-CEPH
Dai	10	HGDP-CEPH	Russian	25	HGDP-CEPH
Daur	9	HGDP-CEPH	San	5	HGDP-CEPH
Druze	42	HGDP-CEPH	Sardinian	28	HGDP-CEPH
French	28	HGDP-CEPH	She	10	HGDP-CEPH
Han	23	HGDP-CEPH	Sindhi	24	HGDP-CEPH
Hazara	14	HGDP-CEPH	Surui	8	HGDP-CEPH
Hezhen	8	HGDP-CEPH	Tu	10	HGDP-CEPH
Italian	12	HGDP-CEPH	Tujia	10	HGDP-CEPH
Japanese	28	HGDP-CEPH	Tuscan	8	HGDP-CEPH
Kalash	23	HGDP-CEPH	Uygur	10	HGDP-CEPH
Karitiana	14	HGDP-CEPH	Xibo	9	HGDP-CEPH
Lahu	8	HGDP-CEPH	Yakut	25	HGDP-CEPH
Makrani	25	HGDP-CEPH	Yi	10	HGDP-CEPH
Mandenka	22	HGDP-CEPH	Yoruba	21	HGDP-CEPH
Maya	21	HGDP-CEPH	SurinamK	15	In house
Abkhazians	20	Caucasus	Tajiks	15	Caucasus
Armenians	16	Caucasus	TSI	88	Hapmap 3
ASW	83	Hapmap 3	TurkeyK	11	In house
Balkans	19	Caucasus	Turkmens	15	Caucasus
Belarusians	6	Caucasus	Ukrainians	8	Caucasus

Bulgarians	13	Caucasus	YRI	167	Hapmap 3
CEU	165	Hapmap 3	Kumyks	14	Caucasus
CHB	84	Hapmap 3	Kurds	6	Caucasus
CHD	85	Hapmap 3	Lithuanians	6	Caucasus
Chechnya's	20	Caucasus	LWK	90	Hapmap 3
GIH	88	Hapmap 3	MEX	77	Hapmap 3
IndiaK	16	In house	MKK	171	Hapmap 3
IraqueK	17	In house	MongoliaK	12	In house
JPT	86	Hapmap 3	Mordvins	15	Caucasus
SomaliK	16	In house	Kuban Nogays	16	Caucasus
North Ossetians	15	Caucasus			

Table S2:

ASW: African ancestry in Southwest USA, CEU: Utah residents with Northern and Western European ancestry from the CEPH collection, CHB: Han Chinese in Beijing China, CHD: Chinese in Metropolitan Denver Colorado, GIH: Gujarati Indians in Houston Texas, JPT: Japanese in Tokyo Japan, LWK: Luhya in Webuye Kenya, MXL: Mexican ancestry in Los Angeles California, MKK: Maasai in Kinyawa Kenya
TSI: Toscani in Italia, YRI: Yoruba in Ibadan, Nigeria

Original population name	Population renamed to
CHB	Han
MongoliaK	Mongola
JPT	Japanese
TSI	Tuscan
YRI	Yoruba

Table S3