BMC Genomics

RESEARCH ARTICLE Open Access

Evolutionary forces shaping genomic islands of population differentiation in humans

Tamara Hofer[1,2*], Matthieu Foll[1,2] and Laurent Excoffier[1,2]

Abstract

Background: Levels of differentiation among populations depend both on demographic and selective factors: genetic drift and local adaptation increase population differentiation, which is eroded by gene flow and balancing selection. We describe here the genomic distribution and the properties of genomic regions with unusually high and low levels of population differentiation in humans to assess the influence of selective and neutral processes on human genetic structure.

Methods: Individual SNPs of the Human Genome Diversity Panel (HGDP) showing significantly high or low levels of population differentiation were detected under a hierarchical-island model (HIM). A Hidden Markov Model allowed us to detect genomic regions or islands of high or low population differentiation.

Results: Under the HIM, only 1.5% of all SNPs are significant at the 1% level, but their genomic spatial distribution is significantly non-random. We find evidence that local adaptation shaped high-differentiation islands, as they are enriched for non-synonymous SNPs and overlap with previously identified candidate regions for positive selection. Moreover there is a negative relationship between the size of islands and recombination rate, which is stronger for islands overlapping with genes. Gene ontology analysis supports the role of diet as a major selective pressure in those highly differentiated islands. Low-differentiation islands are also enriched for non-synonymous SNPs, and contain an overly high proportion of genes belonging to the 'Oncogenesis' biological process.

Conclusions: Even though selection seems to be acting in shaping islands of high population differentiation, neutral demographic processes might have promoted the appearance of some genomic islands since i) as much as 20% of islands are in non-genic regions ii) these non-genic islands are on average two times shorter than genic islands, suggesting a more rapid erosion by recombination, and iii) most loci are strongly differentiated between Africans and non-Africans, a result consistent with known human demographic history.

Background

A number of studies investigated patterns of divergence between closely related, sympatric species and identified so-called 'islands of speciation' in the genome, where the divergence was particularly strong [1-3]. These highly differentiated genomic islands might emerge due to divergent selection acting on the two species [4,5]. Divergent genomic regions cannot only be observed between species, but also between populations within species. Humans are the least differentiated of the extant primate species [6] and most genetic variation is found in populations rather than between populations [7]. Levels of population differentiation are determined both by demographic factors such as genetic drift and gene flow, which can increase or respectively decrease population differentiation [8], and by selective processes, which can also promote [e.g. [9]] or lower [e.g. [10]] differentiation [11-13]. While it is usually believed that demographic forces globally act on the whole genome, it has been shown that pure neutral processes can affect allele frequencies at specific loci during range expansions [14,15], and increase allele frequencies in newly colonised areas [16-18]. This allelic surfing phenomenon depends on local demographic patterns [16] but can be also be enhanced by spatial bottlenecks [19,20].

In the genomics era, large-scale human datasets such as the HapMap project [21-23], the Human Genome Diversity Panel [HGDP; [7,24]], or recently the 1,000

* Correspondence: tamara.hofer@iee.unibe.ch
[1]Computational and Molecular Population Genetics Lab, Institute of Ecology and Evolution, University of Bern, 3012 Bern, Switzerland
Full list of author information is available at the end of the article

Genomes project [25] can be analysed to reconstruct the demographic history of populations or to find signatures of selection on the genome. Rosenberg *et al.* [26] investigated the genetic relationship of human populations and found that populations from the same continent share more ancestry than random populations [see also [7]]. Coop et al. [27] found most evidence for selection between continental groups suggesting that these are ancient adaptations that potentially occurred during the colonisation of continents. There are rather few examples for very local patterns of selection in humans [see e.g. [28-31]]. Recently, Hernandez *et al.* [32] argued that reduced diversity and increased population differentiation in exons could partly result from background selection rather than from selective sweeps, but alternative forms of adaptation, such as selection on standing variation or on multiple beneficial alleles could contribute to population differentiation [see e.g. [33,34]].

In this study we aimed at identifying regions in the human genome with elevated or decreased levels of population differentiation potentially indicative of past episodes of selection. We used the HGDP-CEPH Human Genome Diversity Panel [7] including 660,664 SNPs typed in 53 populations to reliably infer population differentiation. Indeed, whereas the HapMap [21-23] or the 1,000 genome [25] panels provide more detailed genomic information than the HGDP SNP panel, they have been tested in far fewer populations making them actually less powerful for detecting outlier SNPs. We used a novel method to identify loci with unusual F_{ST} values that takes into account hierarchical structure of human populations [35]. Using the significance of individual SNP F_{ST} values as observations, we used a HMM to identify genomic regions with average, high or low population differentiation hidden states. Based on a large set of populations, we thus provide an extensive map of significantly differentiated genomic islands, whose properties such as size, location in the genome, recombination rates or overlap with genic regions, were assessed to collect evidence for the respective effects of selection and neutral demographic processes.

Results

Selection test

By simulating the joint null distribution of F_{ST} and heterozygosity between populations (H_{BP}) under both a Finite Island Model (FIM) [36] and a Hierarchical Island model (HIM) [35], we obtained the significance of SNP-specific F_{ST}s under these two models (Figure 1). We find as many as 21.80% of all SNPs with significant *p*-values at the 1% level under a FIM, while this proportion reduces to 1.5% under a HIM. The excess of significant loci observed under the FIM likely includes many false positives, due to the unrealistic assumption of even levels of differentiation between all pairs of populations [35]. Indeed, the F_{ST} distribution simulated under the HIM much better fits the observed distribution of F_{ST} values than that simulated under the FIM (see Additional file 1). We thus do not find much evidence of selection under the HIM, but neither F_{ST} nor their associated *p*-values are randomly distributed along chromosomes (run-tests *p*-value < 0.001 for all chromosomes). It implies that there are genomic regions with generally elevated or decreased levels of population differentiation, which we have tried to identify using a model-free HMM approach. To this aim, F_{ST} *p*-values were first transformed into *z*-values, also called normal scores, as observations for the HMM. *Z*-values are expected to follow a normal distribution under a proper null model, and interestingly, this is only true for *z*-values obtained under the HIM, while *z*-values computed under the FIM have a right-skewed distribution (Figure 1). The shape of these distributions thus gives us further evidence that the HIM better fits the data than a FIM, since we expect that most SNPs are actually neutral. Given the unrealistic assumption of the FIM and its associated non-Normal distribution of the *z*-values, all results presented below will be based on the HIM only.

Genomic islands

We used a two-step HMM approach to identify regions in the genome with significant population differentiation than average (see Methods and Additional file 2). We first broadly defined genomic regions with generally increased or decreased population differentiation using standard HMM algorithms. In a second step we controlled the False Discovery Rate (FDR) of SNPs and retained only those regions that contained at least one SNP with a genome-wide FDR of 0.001 (hereafter called FDR SNPs). Hereafter, we shall call *high-differentiation islands* (HDIs) those genomic regions with significantly high levels of population differentiation, and *low-differentiation island*s (LDIs) those regions with significant low levels of population differentiation.

Under the HIM we detected 625 HDIs as well as 197 LDIs (Table 1, Figure 2, and Additional file 3).

We find that the number of HDIs per chromosomes correlates with the number of SNPs (R = 0.859, *p*-value < 0.001), but the number of LDIs does not (R = 0.226, *p*-value = 0.301). Chromosomes 3 and X appear to have a higher density of regions with LDIs than the other chromosomes (see Figure 2), which, for the X chromosome, is not due to the overall higher observed level of differentiation since our HIM test controls for that (see below). The average heterozygosity in HDIs (0.23) is significantly below the genome-wide average of 0.28 (*t*-test, *p*-value < 0.001), whereas LDIs present a significantly higher heterozygosity of 0.31 (*t*-test, *p*-value < 0.001).

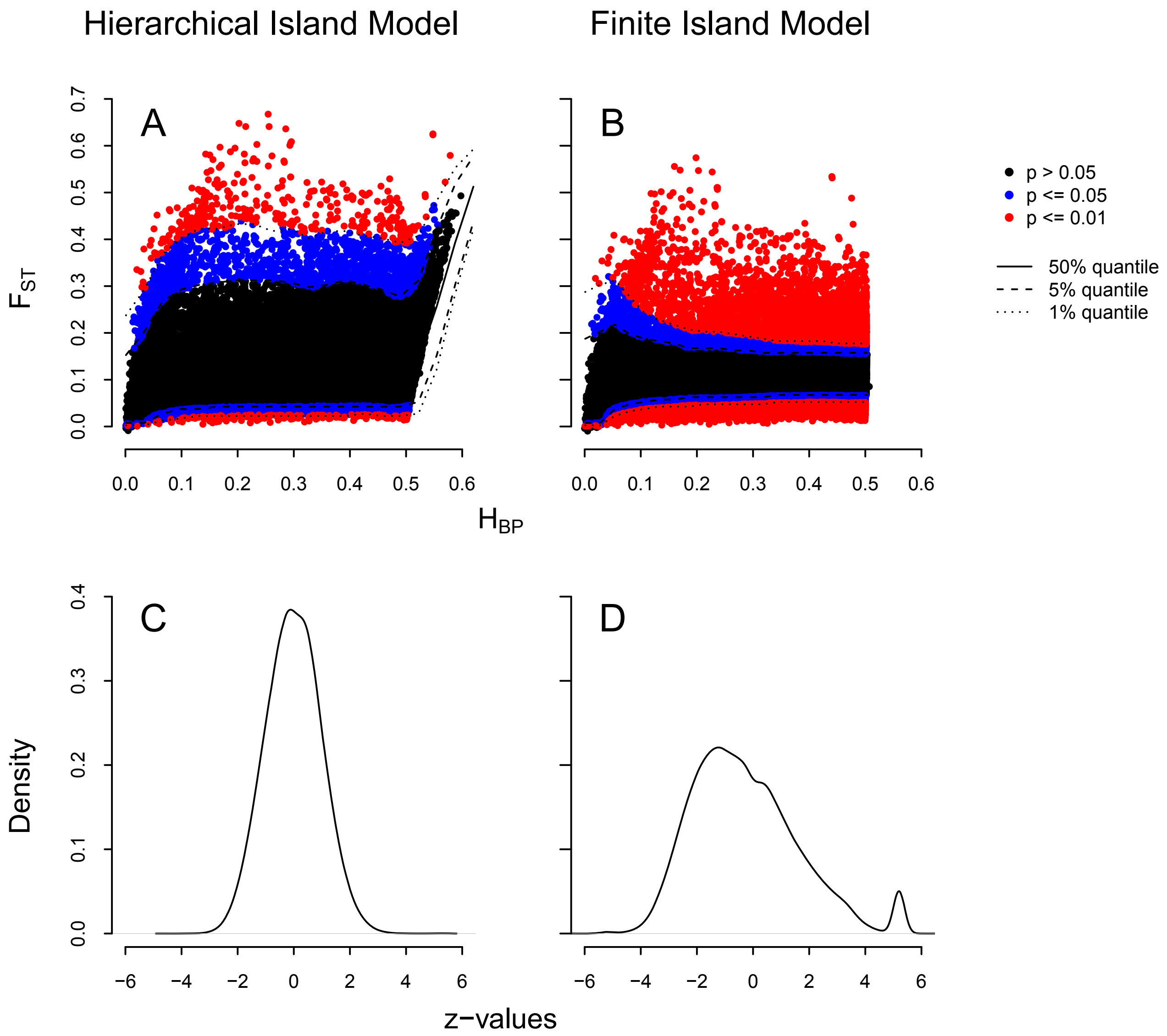

Figure 1 Detection of outlier SNPs based on F_{ST} . Joint distributions of F_{ST} and H_{BP} for SNPs on chromosome 4 (A, B) and genome-wide distribution of SNP *z*-values (C, D). Locus specific *p*-values and *z*-values were computed either based on the hierarchical island model (A, C) or the finite island model (B, D). Chromosome 4 is representative of the joint distribution of these statistics on all other chromosomes.

Table 1 Properties of genomic islands.

	High-differentiation islands (HDIs)		Low-differentiation islands (LDIs)	
	Mean	Min-max	Mean	Min-max
No. of islands	625		197	
Length (bp)	465,756	11,037 - 15,210,531	368,902	25,709 - 2,021,127
No. of SNPs	73.61	4 - 322	90.76	10 - 430
No. of genes	5.35	0 - 92	3.45	0 - 37
Mean Recombination rate	1.04	0.00 - 6.31	1.31	0.04 - 7.48
Mean F_{ST}	0.23	0.17 - 0.43	0.09	0.06 - 0.14
Mean heterozygosity	0.23	0.09 - 0.33	0.31	0.18 - 0.42
Mean *z*-value	0.88	0.32 - 2.63	-0.77	-1.41 - -0.32

Properties of islands with significant levels of population differentiation as detected by the HMM approach

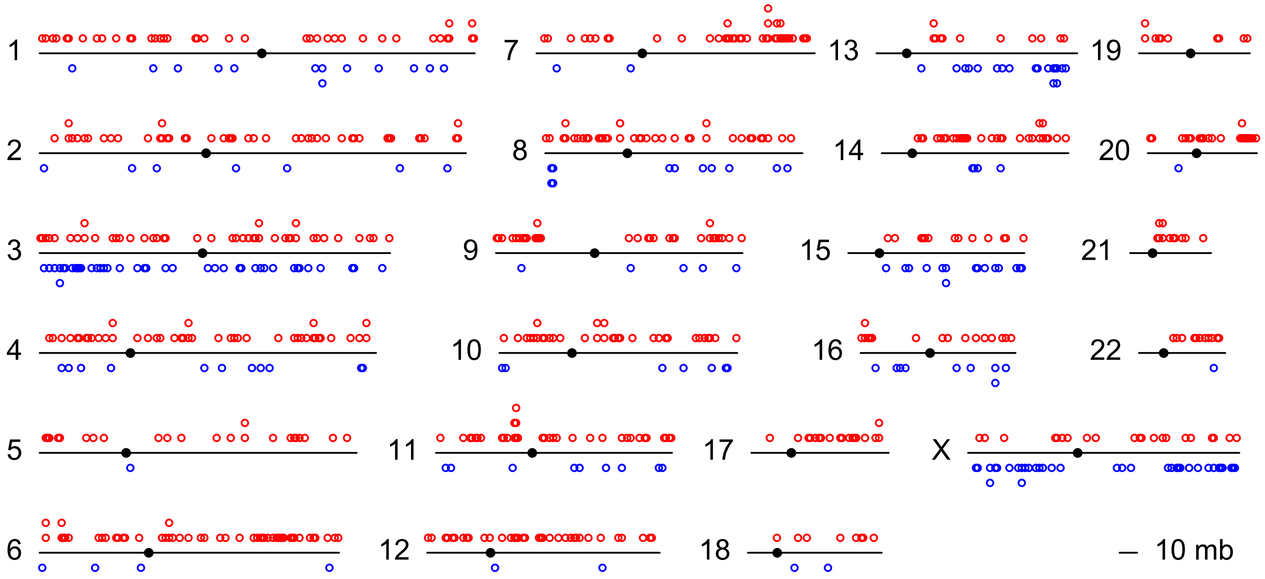

Figure 2 Distribution of HDIs and LDIs. Genomic distribution of HDIs (red) and LDIs (blue). Chromosomes are represented by lines with a filled circle at the centromere. Each circle represents a significant genomic island, as identified by a Hidden Markov Model.

Previous outlier detection studies sometimes refrained from simultaneously analysing the autosomes and the X-chromosome at once, since demographic histories between autosomal and sex-chromosomes differ [37]. In our case, each chromosome was analysed separately, and individual SNP loci were analysed based on their transformed p-value instead of their absolute F_{ST} value, allowing for a comparison of chromosomes with different average F_{ST}. Note however that the largest HDI is located on the X-chromosome, in the low-recombination centromeric region between positions 55.9 and 67.0 Mb.

FDR SNPs in HDIs are found mainly differentiated between Africans vs. non-Africans (71.8%), and less between Americans vs. non-Americans (11.2%), Eurasians vs. non-Eurasians (9.6%), or East Asians vs. non-East Asians (7.5%; see Figures 3 and 4). More precisely, many FDR SNPs have high ancestral frequencies in Africa and low frequencies everywhere else.

Finally, it is worth noting that at odds with genomic islands results, the analysis of individual SNPs with significant population differentiation do not show any enrichment for genic or non-synonymous SNPs, leaving us with no direct evidence for selection at this level.

Genic regions

We find that HDIs are enriched for genic regions. Indeed, 81.3% of high-differentiation regions overlap with at least one gene, which is significantly more than expected based on the size and the number of regions (p-value = 0.033). Contrastingly, LDIs are not enriched for genic regions, even though 76.6% of them overlap with at least one gene (p-value = 0.630). In keeping with these results, HDIs are significantly enriched for genic SNPs (p-value < 0.001), while LDIs are significantly depleted for genic SNPs (p-value < 0.001; Table 2 and Additional file 4). Interestingly, both HDIs and LDIs are significantly enriched for non-synonymous SNPs, which have a slightly significant higher F_{ST} than other SNPs in HDIs (p-value = 0.030), and a significantly lower average F_{ST} than other SNPs in LDIs (p-value = 0.004). We define here non-genic islands as those that do not overlap with any genic region (18.7% of HDIs and 23.4% of LDIs). The average distance of non-genic islands to the closest gene is less than 200 kb, which is not closer than expected (permutation test, p-value > 0.1 for both HDIs and LDIs). Remarkably, all LDIs as well as all but 4 HDIs overlap with transcription factor binding sites, which is significantly more than expected by chance for HDIs (permutation test, p-value = 0.002), but not for LDIs (p-value = 0.073), suggesting that non-genic HDIs might still be under some functional constraints.

We find that both HDIs and LDIs have a significant lower recombination rate than the rest of the genome (weighted t-test p-value < 0.001 for HDIs and p-value = 0.039 for LDIs). Since we find that the average size of non-genic islands is only about half the size of genic islands, we used an ancova to determine the association between the size of islands and local recombination rate, controlling for their genic or non-genic state. We find that HDI size is negatively related to recombination rate in genic HDIs (test of slope p-value < 0.001; Additional

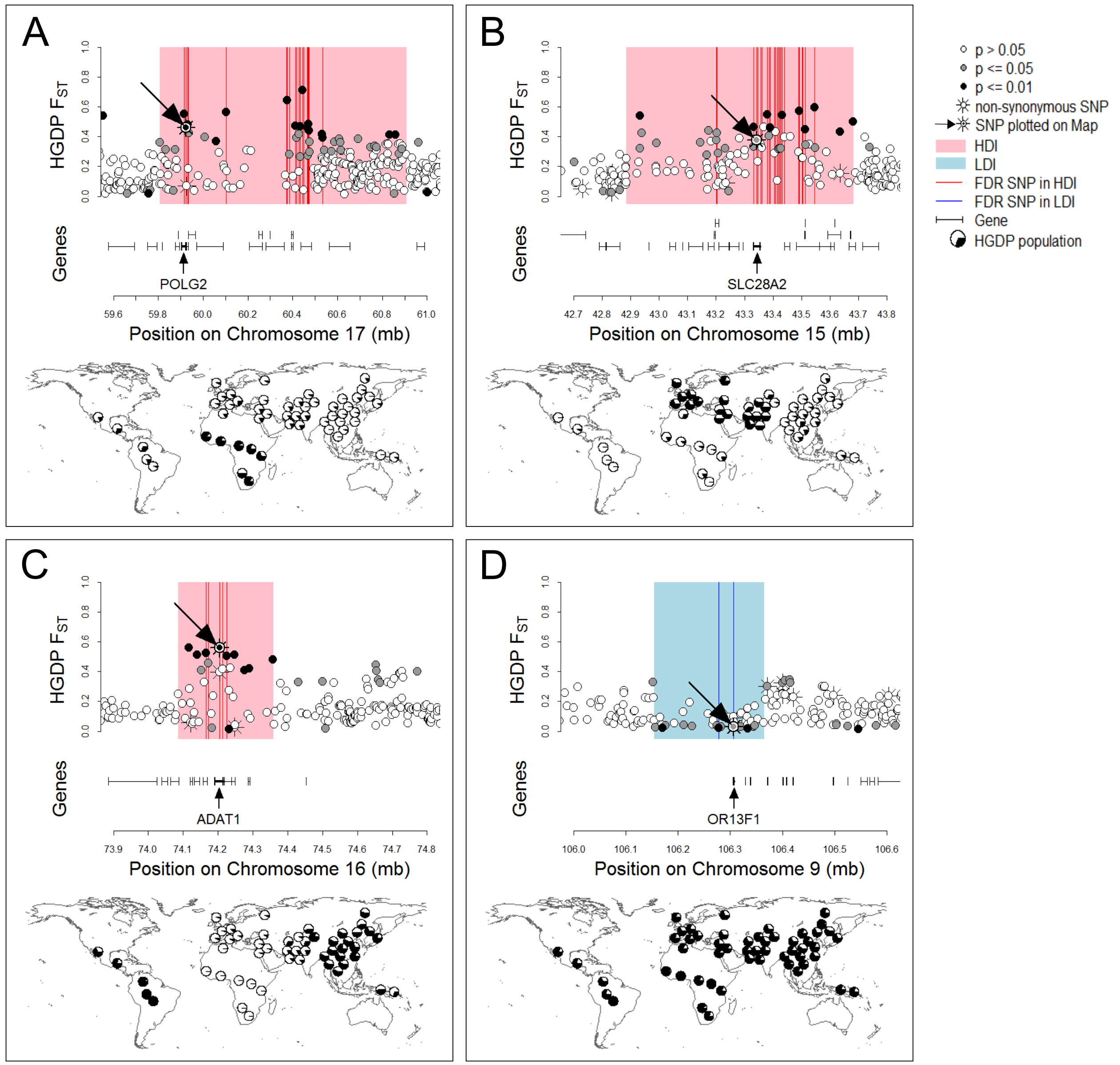

Figure 3 Examples of genomic islands. Each panel shows a genomic island that includes a significantly differentiated non-synonymous SNP. On top, the HDI or LDI is shown in details with all the SNPs and genes located in that region. SNPs are coloured according to their significance level in the selection test and FDR-SNPs are indicated by red vertical lines. The oblique arrow points at the non-synonymous SNP, and the vertical arrow indicates the gene embedding this SNP. The allele frequencies of the non-synonymous SNP in the HGDP populations are shown on the map below. Most common patterns of significant differentiation are found between continental groups, such as African vs. non-Africans (A), between Eurasian vs. non-Eurasians (B), or between East Asian and American vs. African and Eurasian populations (C). SNPs in low-differentiation islands tend to have similar allele frequencies in all populations (D).

file 5) as well as in non-genic HDIs (p-value = 0.021). However, the regression slope is significantly steeper in non-genic than in genic HDIs (ancova test of slope difference p-value = 0.017), which suggests that recombination is more efficient in eroding HDIs in non-genic regions, potentially due to an absence of selective constraints in these regions. Contrastingly, island sizes are negatively correlated with recombination rate in genic LDIs (p-value < 0.001), but not in non-genic LDIs (p-value = 0.34) and the regression slopes of genic and non-genic LDIs are not significantly different (p-value = 0.077). Note also that HDIs overlap less with recombination hotspots than expected by chance (p-value < 0.001), unlike LDIs (p-value = 0.192). There are hotspots

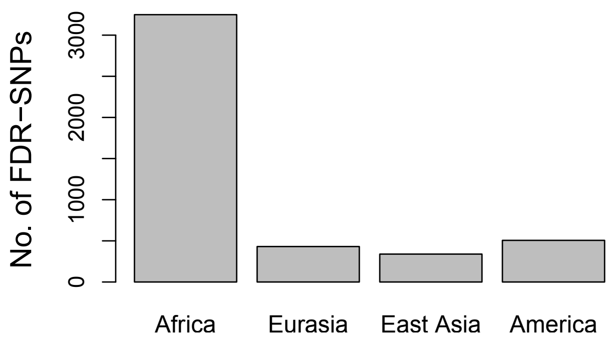

Figure 4 Continental groups showing the largest degree of genetic differentiation. For each FDR-SNP, we determined the allele frequency difference between populations in each continental group (Africa, Eurasia, East-Asia and America) and the rest of the world, and we assigned it to the continental group that showed the highest frequency difference. The histogram reports the number of FDR-SNPs in HDIs assigned to each group.

separating 9 out of 35 pairs of HDIs that are within 200 kb of each other, supporting the view that the HMM detected distinct signals of selection close to each other. We then tested for an enrichment or depletion of genes that overlap with HDIs or LDIs in biological processes using the PANTHER gene ontology database [38,39]. After Bonferroni correction, only two biological processes ('Lipid and fatty acid binding', p-value = 0.006; 'Glycogen metabolism', p-value = 0.046) are enriched for genes in HDIs, and a single process ('Oncogenesis', p-value = 0.005) is enriched for genes located in LDIs.

Discussion

A limited role for selection in humans

Using a hierarchical island model (HIM) to describe patterns of differentiation within and between human continental groups, we identified SNPs from the HGDP panel that presented unusual levels of population differentiation. The slight excess (< 1%) of significant SNPs under the HIM suggests a limited role for adaptive or balancing selection in the human genome, in agreement with a recent analysis of the 1,000 genome project [32]. Note that this excess could also be due to our inability to take into account the exact details of past human history [40], but it is important to underline that outlier loci are not randomly distributed along chromosomes, which motivated us to use a HMM to detect islands of high or low differentiation (HDIs and LDIs respectively). The fact that the joint distribution of F_{ST} and H_{BP} generated under the HIM much better fits the observed distribution than that obtained under the FIM (Figure 1 and Additional file 1) suggests that the HIM captures key aspects of human demography and that the identified outlier loci are more likely to be enriched for true signal of selection. The choice of the right demographic model thus appears crucial for the proper identification of selection signals.

The action of neutral and selective processes in the genome

Several features of outlier SNPs and differentiation islands point towards the action of selection: i) HDIs are enriched for both genic and non-synonymous SNPs, ii) non-synonymous SNPs are more differentiated than other SNPs within HDIs, consistent with directional selection, and non-synonymous SNPs are less differentiated than synonymous SNPs within LDIs, consistent with purifying selection (see Additional file 6 for a list of non-synonymous SNPs with significant differentiation located in HDIs or LDIs), iii) non-synonymous SNPs are enriched in LDIs, compatible with balancing selection at

Table 2 Enrichment of genic and non-synonymous SNPs in genomic islands.

SNP subset		Genic	Non-synonymous
High population differentiation	p-value(selection test) < = 0.01[a]	0.6576	0.1307
	Most likely states (Viterbi)[b]	< 0.0001[+]	0.0002[+]
	FDR < = 0.001[c]	< 0.0001[+]	0.0374[+]
	Islands[d]	< 0.0001[+]	0.0053[+]
Low population differentiation	p-value(selection test) < = 0.01[a]	0.2143	0.6402
	Most likely states (Viterbi)[b]	0.0303[-]	0.5888
	FDR < = 0.001[c]	0.0007[-]	0.6543
	Islands[d]	< 0.0001[-]	0.0083[+]

Fisher exact test for enrichment or depletion of genic and non-synonymous SNPs in genomic islands defined by the selection test or by the HMM approach

[a] subset of SNPs that are significant in the selection test at the 1% level

[b] subset of SNPs that belong to the high-differentiation or low-differentiation state as identified by the Viterbi algorithm in the first step of the HMM approach

[c] subset of SNPs that are assigned to the high-differentiation or low-differentiation state under a genome-wide FDR = 0.001 in the second step of the HMM approach

[d] subset of SNPs that are located within HDIs or LDIs

[+] enrichment of SNP category in SNP class

[-] under-representation of SNP category in SNP class

a few SNPs and the accumulation of neutral mutations at other nearby sites, iv) LDIs are generally smaller than HDIs, in agreement with balancing selection giving more time for recombination to erode LDIs and fast selective sweeps creating large HDIs, v) the negative correlation between HDI size and recombination rate is stronger for genic than for non-genic HDIs, consistent with recent positive selection on genes and vi) HDIs overlap more often with genes than expected by chance, and there is a significant excess of transcription factor binding sites in HDIs. Note, however, that the HMM has more power to highlight regions with low recombination rates (such as functional regions), where neighbouring SNPs are more likely to have similar F_{ST} values (but see Additional file 7).

Some other features of the islands are better explained by past demography than by selection, like i) levels of heterozygosity in LDIs that are comparable with genome-wide levels, unlike what would be predicted by balancing or background selection, ii) most high-F_{ST} SNPs with low FDR that are mainly differentiated between Africa and non-Africa, and between America and non-America, which is compatible with the action of surfing after spatial bottlenecks (Figure 4), or iii) 23.4% of LDIs and 18.6% of HDIs do not overlap with any annotated functional gene. Interestingly, these non-genic islands (both HDIs and LDIs) are about two times shorter than islands overlapping with genes, which is consistent with the assumption that they are neutral and therefore more easily eroded by recombination. But note that most of the non-genic islands overlap with transcription factor binding sites, which are, however, less constrained than genic regions [41].

It thus appears likely that genomic islands with unusual levels of differentiation have been shaped both demographic and selective events, which still appear very challenging to disentangle without a higher density of markers and populations.

Recently selected biological processes

We find evidence for local adaptation to food sources and nutrition as the biological processes of 'Lipid and fatty acid binding' and 'Glycogen metabolism' are enriched for genes located in HDIs. This result is in line with earlier studies showing that diet differs between populations and should present strong selective pressures [42,43]. Interestingly the process of 'Oncogenesis' is enriched for genes in LDIs. This process includes genes that normally regulate cell growth and differentiation [44] and cancer/testis genes that seem to be under rapid diversifying selection between human and chimpanzees, especially on the X-chromosome [45]. The X-chromosome is enriched for LDIs, supporting the interpretation that cancer/testis genes are under diversifying selection in humans, leading to balanced polymorphisms. On the other hand, we might have more power for the detection of LDIs on the X-chromosome due to its higher level of differentiation as compared to the autosomes. However, chromosome 3 has a level of population differentiation that is comparable to that of the other autosomes, but is also enriched for LDIs. A literature search did not reveal any neutral explanation for the high prevalence of LDIs on chromosome 3 leaving the possibility that chromosome 3 is enriched for targets of balancing selection.

Comparison with other genome scan studies

HDIs contain several candidate genes for local adaptation identified in previous studies, such as *TRPV6* [46], *ASPM* [47], *prodynorphin* [PDYN; [48]], the *duffy blood group* locus involved in malaria resistance [DARC; [49]], or the *ectodysplasin A receptor* [EDAR; [50]]. Additionally, HDIs overlap with genes linked to skin pigmentation including the *melanocortin 1 receptor* gene [MC1R; [51]], *KITLG* [52], *SLC24A5* [53], *tyrosinase-related protein 1* gene [TYRP1; [54]], and *OCA2* [55]. LDIs also overlap with genes that were previously shown to be under balancing selection such as *HLA-C* [56] and *dystrophin* [DMD; [57]]. While HDIs are more sensitive towards local adaptation, LDIs may fail to highlight genes that are under balancing selection in only a few populations. More generally, we tested the overlap of HDIs and LDIs with regions identified as being under positive and balancing selection in previous genome scans (Additional file 8). We actually find a clear overlap between HDIs and regions under positive selection detected with methods based on levels of population subdivision [58-60], analyses of the site frequency spectrum [61-64], or tests relying on patterns of LD and haplotype variability [23,28,65-67]. Interestingly, HDIs do not overlap with studies that aimed at detecting old episodes of selection based on the ratio of polymorphism relative to divergence among species [68-70], which suggests that we detect genomic islands that have appeared more recently, after the out-of-Africa event 50-60 Ky ago [13]. However, the fact that some genomic regions are detected with several selection tests does not really demonstrate the action of selection, but rather that the identified signal is strong enough to be picked up by various methods. It is also worth noting that several candidate regions for balancing selection identified in previous studies did not overlap with LDIs (Additional file 8), potentially because balancing selection events are rare in the human genome and very hard to detect [71,72], in agreement with the fact that we detect less LDIs than HDIs. The difficulty to detect balancing selection is further illustrated by the low concordance between the 5 previous studies aiming at detecting balancing selection, where only 1.3% of all identified candidate regions were detected in more than one study.

Future HMM approaches could be extended to the analysis of next-generation sequencing data instead of a limited number of linked markers. With deeper coverage and the inclusion of additional populations allowing precise estimation of levels of population differentiation, the 1,000 Genomes project [25] could provide a unique and very powerful tool to refine the delineation of islands of differentiation.

Conclusions

The proper detection of loci with unusual levels of population differentiation requires an appropriate model of human genetic structure, such as the HIM model used here. While we find little direct evidence of selection at the level of individual SNPs, the identification of genomic islands of differentiation under an HMM approach pooling information over linked SNPs reveals more powerful. We find that several properties of genomic islands overlapping with gene regions are difficult to explain without the action of selection, but that past demographic events such as gene surfing are probably involved in their occurrence in non-genic regions.

Methods

Data

We analysed the HGDP-CEPH Human Genome Diversity Panel including a total of 660,918 SNPs typed in 53 populations worldwide [[7]; ftp://ftp.cephb.fr/hgdp_supp1/]. For subsequent hierarchical analyses, the 53 populations were grouped into the 5 major geographic regions defined by Rosenberg *et al.* [7,26]: Africa, Eurasia, East Asia, Oceania, and America. We removed 12 SNPs that have only missing data, 4 SNPs that were not typed at all in a population, 50 SNPs that were monomorphic in all populations, and we discarded 188 SNPs that were located on the Y-chromosome, on the pseudoautosomal region of the X and Y chromosome, or on mitochondrial DNA, leaving us with 660,664 SNPs for subsequent analyses.

Selection test

We used the hierarchical selection test [35] implemented in ARLEQUIN ver 3.5 [73] to identify loci with significant levels of population differentiation. For each chromosome, ARLEQUIN generated the joint null distribution of global F_{ST} [74,75] and heterozygosity between populations (H_{BP}) based on 50,000 coalescent simulations under a hierarchical island model (HIM) or under a finite island model (FIM). Beaumont and Nichols [36] proposed to simulate the joint distribution of F_{ST} and heterozygosity under a FIM, which they assumed to be robust under a wide range of conditions. However, Excoffier et al. [35] recently showed that the presence of hierarchical structure among sampled populations leads to an excess of false positives if the data is analysed under the assumption of a FIM. Instead the underlying continental structure of human populations [7,26] needs to be taken into account, which can be done by using a HIM [76]. In our study the simulated HIM consisted in 10 groups of 100 demes and the FIM of a single group of 100 demes. Migrations rates within and between groups were estimated from the observed F-statistics (Excoffier et al. 2009). Obtained null distributions were used to compute *p*-values for the individual SNPs and corresponding *z*-values (i.e. standard scores) with the quantile function *qnorm* implemented in the statistical software R [77]. For instance, a positive *z*-value of 1.64 corresponds to a *p*-value of 0.05 for high population differentiation whereas a *z*-value of -1.64 indicates a *p*-value of 0.05 for low differentiation. We used a run test [78] to test for a non-random distribution of F_{ST} and *p*-values along chromosomes. Note that previous studies using outlier approaches sometimes restricted their analyses to autosomes because the differentiation level of the X-chromosome is higher due to its reduced effective size [e.g. [27,79,80]]. However, our model-based approach allows us to take into account the specificities of both autosomes and sex chromosomes and to compare them in the same analysis.

Hidden Markov model

Sliding window approaches have often been used to find clustered values of some statistic along a sequence [see e.g. [63,81]], but this approach has some drawbacks. Indeed, the choice of the correct window size and increment is not trivial, and it might have a strong impact on the number and size of detected clusters. Additionally, random fluctuations of the test statistic in a delimited window might lead to the detection of a cluster when there is none [82]. Hidden Markov Models (HMM) are widely used in biology for sequence analyses, since they explicitly model dependencies among neighbouring markers [83-85] and they have recently been introduced to identify genomic regions influenced by selection [e.g. [86,87]]. To define HDIs and LDIs, we thus applied a HMM as implemented in the R package 'HiddenMarkov' [88]. We used a HMM with 3 hidden states for low (LDI), intermediate, and high (HDI) levels of population differentiation, respectively. A HMM is characterised by different parameters, such as the distribution of the test statistic under each state, emission probabilities, and transition probabilities. We used *z*-values as the observed test statistic for the HMM. The distribution of *z*-values under each of the 3 states was assumed to be Gaussian with a given mean and standard deviation estimated from the data. Emission probabilities specifying how likely it is to observe a given value under each state, and the transition matrix defining how likely it is to pass from one state into another state were also estimated from the data. We imposed a constrained transition matrix disallowing direct transitions between HDI and LDI states. The Baum-Welch algorithm [89] was used to estimate the parameters for the HMM for each chromosome independently. The algorithm was launched 1,000 times from different starting values and the resulting parameter estimates with the highest likelihood were retained for the final model (Additional file 9). Note that when we tried using a model with an unconstrained transition matrix, the transition probabilities between these two states were in all cases below 0.07.

We used a two-step approach to identify HDIs and LDIs (see Additional file 2 for an illustration of the method). In a first step the Viterbi algorithm [90] was used to determine the most likely sequence of states on a given chromosome and to define regions with different levels of population differentiation. We then identified SNPs that were significantly assigned to either the high-differentiation state or the low-differentiation state by computing their Local Index of Significance [LIS; [91,92]], which is the probability estimated from the HMM that this SNP does not belong to a given state. LIS takes into account linkage between neighbouring loci since it is based on the local dependence structure of the HMM. Following the approach of Wei *et al.* [92] we used the LIS to identify SNPs that were significantly associated with either high- or low-differentiation under a genome-wide False Discovery Rate (FDR) of 0.001 and that are called here FDR-SNPs. These FDR-SNPs are not necessarily the most significant SNPs as detected from the selection test alone, but they are rather located in regions with globally high or low levels of population differentiation.

Finally, we combined the output of the Viterbi algorithm and the FDR procedure by only retaining those regions that contained at least one high- or low-differentiation FDR-SNP to define the most significant HDI and LDIs, respectively.

Recombination rate and recombination hotspots

We used standardized, sex-averaged DeCode recombination rates [93] to assess the impact of recombination on genomic islands of high or low population differentiation on the autosomes. We applied a weighted t-test [94] to test for differences in recombination rates between presumably neutral regions, HDIs and LDIs, respectively. Following Kong *et al.* [93] we defined recombination hotspots as those regions on the recombination map that have a standardized recombination rate greater than 10. We then determined the number of HDIs and LDIs that overlap with one or more recombination hotspots. To determine the significance of this value we randomly permuted islands 10,000 times across the whole genome.

Functional analyses

Functional genes as annotated in ENSEMBL 54 [95] overlapping with HDIs and LDIs were identified. We determined the number of HDIs and LDIs that overlap with at least one gene (i.e. genic islands) and tested the significance of this value using 10,000 random permutations of islands across the whole genome. In this procedure, positions of HDIs and LDIs and regions in-between were permuted simultaneously to compute the null distribution of the overlap between islands and genes. We used an ancova model to test for an association between the length of islands and their average recombination rate, controlling for their genic or non-genic status. In this analysis, we excluded 3 HDIs that span centromeres on chromosomes 12, 16, and 18 as the size of these regions is likely overestimated due to the absence of typed SNPs and the very low recombination rate in these regions. Furthermore, we tested whether biological processes of the PANTHER gene ontology database [38,39] were enriched for genes in significantly differentiated regions.

SNPs were assigned to functional categories, such as genic and non-synonymous, based on information from ENSEMBL 54 [95]. First we tested if non-synonymous SNPs had a different F_{ST} from other SNPs within HDIs and within LDIs using a Mann-Whitney U test. We then used a Fisher exact-test to investigate whether certain functional categories were enriched among SNPs in HDIs and LDIs. We additionally tested the enrichment of functional categories among SNPs identified in earlier steps of the analyses pipeline: i) the SNPs that were significant at the 1% level in the selection test, ii) the SNPs assigned to high-differentiation or low-differentiation states by the Viterbi algorithm, iii) and FDR-SNPs.

To test if non-genic islands are closer to genes than expected, we permuted non-genic HDIs and LDIs 1,000 times across the genome, conditioning on the fact that the permuted islands do not overlap with any gene. Using the same procedure, we also tested if non-genic islands overlapped more with transcription factor binding sites than expected by chance. To make this test, we used the list of transcription factor binding sites in the human genome available from ENCODE through the UCSC table browser [96,97]; table wgEncodeRegTfbsClustered].

Comparison with previous studies

Previous genome scan studies have detected many candidate regions for both positive and balancing selection. We determined how many HDIs and LDIs overlapped with candidate regions discovered by other studies. We converted the positions of candidate regions from these studies into NCBI Build 36-reference system with the liftOver tool available on the UCSC web page if necessary [97]. For studies that identified genes instead of genomic regions we used the transcription start and transcription end of these genes as the limits of the candidate regions. Empirical p-values of the overlap with previous studies were obtained from 10,000 random permutations of HDI and LDI positions in the genome.

Additional material

Additional file 1: Comparison of observed and simulated F_{ST} distributions. Q-Q plots of observed F_{ST} values and F_{ST} values simulated under the FIM and the HIM.

Additional file 2: The steps involved in the HMM approach to detect HDIs and LDIs. Illustration of the procedure used to identify islands with significant high and low population differentiation in the human genome.

Additional file 3: List of all detected HDIs and LDIs. The properties of all HDIs and LDIs detected under the hierarchical island model.

Additional file 4: Test of enrichment of SNP categories in SNP subsets. Fisher exact test for enrichment or depletion of SNPs with a given consequence to transcript (Ensembl) in genomic islands defined by the selection test or by the HMM approach.

Additional file 5: Island size and recombination rate. Plots illustrating the relationship between island size and local recombination rate in genic and non-genic HDIs and LDIs.

Additional file 6: List of non-synonymous SNPs located in HDIs or LDIs with a p-value ≤ 0.05. Candidate loci that might be affected by local adaptation or balancing selection, due to their effect on the transcript, significant level of population differentiation, and location within HDIs or LDIs, respectively.

Additional file 7: Overlap of HDIs and LDIs to candidate regions for selection identified in previous studies. Histogram of recombination rate of all genomic bins, bins overlapping to HDIs, and bins overlapping to LDIs.

Additional file 8: Overlap of HDIs and LDIs to candidate regions for selection identified in previous studies. Number of candidate regions for selection detected by previous studies that overlap with the 625 HDIs and 197 LDIs identified in this study [98,100-117].

Additional file 9: Estimated parameters for the HMM. Parameters of the HMM for each chromosome estimated by the Baum-Welch algorithm. (S8_HMM_parameters.xls can be viewed with Microsoft Excel or Excel Viewer).

Acknowledgements
We would like to thank L. Dümbgen for sharing his expertise about statistical analyses as well as a script that we used to perform a run-test. T. Oleksyk and I. Hellmann kindly shared databases of candidate regions identified by previous studies. Additional thanks go to D. Wegmann for insightful discussions about the analyses and I. Duperret for assistance with computational and database issues. This work was supported by the Swiss National Science Foundation grant No 3100-126074 to LE.

Author details
[1]Computational and Molecular Population Genetics Lab, Institute of Ecology and Evolution, University of Bern, 3012 Bern, Switzerland. [2]Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland.

Authors' contributions
TH developed the data analysis pipeline, performed statistical analyses, interpreted the results and wrote the manuscript. MF and LE were involved in project design, statistical analyses, and manuscript editing. All authors have read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Received: 4 October 2011 Accepted: 22 March 2012
Published: 22 March 2012

References

1. Harr B: Genomic islands of differentiation between house mouse subspecies. *Genome Res* 2006, 16(6):730-737.
2. Nosil P, Egan SP, Funk DJ: Heterogeneous genomic differentiation between walking-stick ecotypes: "Isolation by adaptation" and multiple roles for divergent selection. *Evolution* 2008, 62(2):316-336.
3. Turner TL, Hahn MW: Locus- and population-specific selection and differentiation between incipient species of Anopheles gambiae. *Mol Biol Evol* 2007, 24:2132-2138.
4. Feder JL, Nosil P: The efficacy of divergence hitchhiking in generating genomic islands during ecological speciation. *Evolution* 2010, 64(6):1729-1747.
5. Michel AP, Sim S, Powell THQ, Taylor MS, Nosil P, Feder JL: Widespread genomic divergence during sympatric speciation. *P Natl Acad Sci USA* 2010, 107(21):9724-9729.
6. Kaessmann H, Wiebe V, Weiss G, Paabo S: Great ape DNA sequences reveal a reduced diversity and an expansion in humans. *Nature Genet* 2001, 27(2):155-156.
7. Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, *et al*: Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 2008, 319(5866):1100-1104.
8. Wright S: Evolution in Mendelian populations. *Genetics* 1931, 16(2):0097-0159.
9. Xue YL, Zhang XL, Huang N, Daly A, Gillson CJ, MacArthur DG, Yngvadottir B, Nica AC, Woodwark C, Chen Y, *et al*: Population differentiation as an indicator of recent positive selection in humans: an empirical evaluation. *Genetics* 2009, 183(3):1065-1077.
10. Schierup MH, Charlesworth D, Vekemans X: The effect of hitch-hiking on genes linked to a balanced polymorphism in a subdivided population. *Genet Res* 2000, 76(1):63-73.
11. Charlesworth D: Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genet* 2006, 2(4):379-384.
12. Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG: Recent and ongoing selection in the human genome. *Nat Rev Genet* 2007, 8(11):857-868.
13. Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, Palma A, Mikkelsen TS, Altshuler D, Lander ES: Positive natural selection in the human lineage. *Science* 2006, 312(5780):1614-1620.
14. Hofer T, Ray N, Wegmann D, Excoffier L: Large allele frequency differences between human continental groups are more likely to have occurred by drift during range expansions than by selection. *Ann Hum Genet* 2009, 73:95-108.
15. Currat M, Excoffier L, Maddison W, Otto SP, Ray N, Whitlock MC, Yeaman S: Comment on "Ongoing adaptive evolution of ASPM, a brain size determinant in homo sapiens"and"microcephalin, a gene regulating brain size, continues to evolve adaptively in humans". *Science* 2006, 313(5784):2.
16. Klopfstein S, Currat M, Excoffier L: The fate of mutations surfing on the wave of a range expansion. *Mol Biol Evol* 2006, 23(3):482-490.
17. Hallatschek O, Nelson DR: Life at the front of an expanding population. *Evolution* 2010, 64(1):193-206.
18. Edmonds CA, Lillie AS, Cavalli-Sforza LL: Mutations arising in the wave front of an expanding population. *P Natl Acad Sci USA* 2004, 101(4):975-979.
19. Burton OJ, Travis JMJ: Landscape structure and boundary effects determine the fate of mutations occurring during range expansions. *Heredity* 2008, 101(4):329-340.

20. Travis JMJ, Munkemuller T, Burton OJ, Best A, Dytham C, Johst K: **Deleterious mutations can surf to high densities on the wave front of an expanding population.** *Mol Biol Evol* 2007, **24(10)**:2334-2343.
21. Altshuler D, Brooks LD, Chakravarti A, Collins FS, Daly MJ, Donnelly P, Gibbs RA, Belmont JW, Boudreau A, Leal SM, *et al*: **A haplotype map of the human genome.** *Nature* 2005, **437(7063)**:1299-1320.
22. Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu FL, Bonnen PE, de Bakker PIW, Deloukas P, Gabriel SB, *et al*: **Integrating common and rare genetic variation in diverse human populations.** *Nature* 2010, **467(7311)**:52-58.
23. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, *et al*: **A second generation human haplotype map of over 3.1 million SNPs.** *Nature* 2007, **449(7164)**:851-U853.
24. Cann HM, de Toma C, Cazes L, Legrand M-F, Morel V, Piouffre L, Bodmer J, Bodmer WF, Bonne-Tamir B, Cambon-Thomsen A, *et al*: **A human genome diversity cell line panel.** *Science* 2002, **296(5566)**:261b-262b.
25. Altshuler DL, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, Collins FS, De la Vega FM, Donnelly P, Egholm M, *et al*: **A map of human genome variation from population-scale sequencing.** *Nature* 2010, **467(7319)**:1061-1073.
26. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW: **Genetic structure of human populations.** *Science* 2002, **298(5602)**:2381-2385.
27. Coop G, Pickrell JK, Novembre J, Kudaravalli S, Li J, Absher D, Myers RM, Cavalli-Sforza LL, Feldman MW, Pritchard JK: **The role of geography in human adaptation.** *PLoS Genet* 2009, **5(6)**:16.
28. Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, Absher D, Srinivasan BS, Barsh GS, Myers RM, Feldman MW, *et al*: **Signals of recent positive selection in a worldwide sample of human populations.** *Genome Res* 2009, **19(5)**:826-837.
29. Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZXP, Pool JE, Xu X, Jiang H, Vinckenbosch N, Korneliussen TS, *et al*: **Sequencing of 50 human exomes reveals adaptation to high altitude.** *Science* 2010, **329(5987)**:75-78.
30. Saunders MA, Hammer MF, Nachman MW: **Nucleotide variability at G6pd and the signature of malarial selection in humans.** *Genetics* 2002, **162(4)**:1849-1861.
31. Bersaglieri T, Sabeti P, Patterson N, Vanderploeg T, Schaffner S, Drake J, Rhodes M, Reich D, Hirschhorn J: **Genetic signatures of strong recent positive selection at the lactase gene.** *Am J Hum Genet* 2004, **74(6)**:1111-1120.
32. Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A, McVean G, Project G, Sella G, Przeworski M: **Classic selective sweeps were rare in recent human evolution.** *Science* 2011, **331(6019)**:920-924.
33. Pritchard JK, Pickrell JK, Coop G: **The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation.** *Curr Biol* 2010, **20(4)**: R208-R215.
34. Ralph P, Coop G: **Parallel adaptation: one or many waves of advance of an advantageous allele?** *Genetics* 2010, **186(2)**:647-668.
35. Excoffier L, Hofer T, Foll M: **Detecting loci under selection in a hierarchically structured population.** *Heredity* 2009, **103(4)**:285-298.
36. Beaumont MA, Nichols RA: **Evaluating loci for use in the genetic analysis of population structure.** *P Roy Soc B-Biol Sci* 1996, **263(1377)**:1619-1626.
37. Keinan A, Mullikin JC, Patterson N, Reich D: **Accelerated genetic drift on chromosome X during the human dispersal out of Africa.** *Nature Genet* 2009, **41(1)**:66-70.
38. Thomas PD, Campbell MJ, Kejariwal A, Mi HY, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A: **PANTHER: a library of protein families and subfamilies indexed by function.** *Genome Res* 2003, **13(9)**:2129-2141.
39. Thomas PD, Kejariwal A, Guo N, Mi HY, Campbell MJ, Muruganujan A, Lazareva-Ulitsky B: **Applications for protein sequence-function evolution data: mRNA/protein expression analysis and coding SNP scoring tools.** *Nucleic Acids Res* 2006, **34**:W645-W650.
40. Hermisson J: **Who believes in whole-genome scans for selection?** *Heredity* 2009, **103(4)**:283-284.
41. Ponting CP, Lunter G: **Signatures of adaptive evolution within human non-coding sequence.** *Hum Mol Genet* 2006, **15**:R170-R175.
42. Patin E, Quintana-Murci L: **Demeter's legacy: rapid changes to our genome imposed by diet.** *Trends Ecol Evol* 2008, **23(2)**:56-59.
43. Hancock AM, Witonsky DB, Ehler E, Alkorta-Aranburu G, Beall C, Gebremedhin A, Sukernik R, Utermann G, Pritchard J, Coop G, *et al*: **Human adaptations to diet, subsistence, and ecoregion are due to subtle shifts in allele frequency.** *P Natl Acad Sci USA* 2010, **107**:8924-8930.
44. Croce CM: **Molecular origins of cancer: oncogenes and cancer.** *N Engl J Med* 2008, **358(5)**:502-511.
45. Stevenson BJ, Iseli C, Panji S, Zahn-Zabal M, Hide W, Old LJ, Simpson AJ, Jongeneel CV: **Rapid evolution of cancer/testis genes on the X chromosome.** *Bmc Genomics* 2007, **8**:129.
46. Soejima M, Tachida H, Koda Y: **Sequence analysis of human TRPV6 suggests positive selection outside africa.** *Biochem Genet* 2009, **47(1-2)**:147-153.
47. Mekel-Bobrov N, Gilbert SL, Evans PD, Vallender EJ, Anderson JR, Hudson RR, Tishkoff SA, Lahn BT: **Ongoing adaptive evolution of ASPM, a brain size determinant in homo sapiens.** *Science* 2005, **309(5741)**:1720-1722.
48. Rockman MV, Hahn MW, Soranzo N, Zimprich F, Goldstein DB, Wray GA: **Ancient and recent positive selection transformed opioid cis-regulation in humans.** *PLoS Biol* 2005, **3(12)**:e387.
49. Hamblin MT, Thompson EE, Di Rienzo A: **Complex signatures of natural selection at the Duffy blood group locus.** *Am J Hum Genet* 2002, **70(2)**:369-383.
50. Hillmer AM, Freudenberg J, Myles S, Herms S, Tang K, Hughes DA, Brockschmidt FF, Ruan YJ, Stoneking M, Nothen MM: **Recent positive selection of a human androgen receptor/ectodysplasin A2 receptor haplotype and its relationship to male pattern baldness.** *Hum Genet* 2009, **126(2)**:255-264.
51. Gerstenblith MR, Goldstein AM, Fargnoli MC, Peris K, Landi MT: **Comprehensive evaluation of allele frequency differences of MC1R variants across populations.** *Hum Mutat* 2007, **28(5)**:495-505.
52. Lao O, de Gruijter JM, van Duijn K, Navarro A, Kayser M: **Signatures of positive selection in genes associated with human skin pigmentation as revealed from analyses of single nucleotide polymorphisms.** *Ann Hum Genet* 2007, **71(3)**:354-369.
53. Lamason RL, Mohideen M-APK, Mest JR, Wong AC, Norton HL, Aros MC, Jurynec MJ, Mao X, Humphreville VR, Humbert JE, *et al*: **SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans.** *Science* 2005, **310(5755)**:1782-1786.
54. Izagirre N, Garcia I, Junquera C, de la Rua C, Alonso S: **A scan for signatures of positive selection in candidate loci for skin pigmentation in humans.** *Mol Biol Evol* 2006, **23(9)**:1697-1706.
55. Norton HL, Kittles RA, Parra E, McKeigue P, Mao X, Cheng K, Canfield VA, Bradley DG, McEvoy B, Shriver MD: **Genetic evidence for the convergent evolution of light skin in Europeans and East Asians.** *Mol Biol Evol* 2007, **24(3)**:710-722.
56. Solberg OD, Mack SJ, Lancastera AK, Single RM, Tsai Y, Sanchez-Mazas A, Thomson G: **Balancing selection and heterogeneity across the classical human leukocyte antigen loci: a meta-analytic review of 497 population studies.** *Hum Immunol* 2008, **69(7)**:443-464.
57. Nachman MW, Crowell SL: **Contrasting evolutionary histories of two introns of the duchenne muscular dystrophy gene, DMD, in humans.** *Genetics* 2000, **155(4)**:1855-1864.
58. Akey JM, Zhang G, Zhang K, Jin L, Shriver MD: **Interrogating a high-density SNP map for signatures of natural selection.** *Genome Res* 2002, **12(12)**:1805-1814.
59. Nielsen R, Hubisz MJ, Hellmann I, Torgerson D, Andres AM, Albrechtsen A, Gutenkunst R, Adams MD, Cargill M, Boyko A, *et al*: **Darwinian and demographic forces affecting human protein coding genes.** *Genome Res* 2009, **19(5)**:838-849.
60. Oleksyk TK, Zhao K, De La Vega FM, Gilbert DA, O'Brien SJ, Smith MW: **Identifying selected regions from heterozygosity and divergence using a light-coverage genomic dataset from two human populations.** *Plos One* 2008, **3(3)**:e1712.
61. Williamson SH, Hubisz MJ, Clark AG, Payseur BA, Bustamante CD, Nielsen R: **Localizing recent adaptive evolution in the human genome.** *PLoS Genet* 2007, **3(6)**:e90.
62. Chen H, Patterson N, Reich D: **Population differentiation as a test for selective sweeps.** *Genome Res* 2010, **20(3)**:393-402.
63. Carlson CS, Thomas DJ, Eberle MA, Swanson JE, Livingston RJ, Rieder MJ, Nickerson DA: **Genomic regions exhibiting positive selection identified from dense genotype data.** *Genome Res* 2005, **15(11)**:1553-1565.

64. Hellmann I, Mang Y, Gu ZP, Li P, de la Vega FM, Clark AG, Nielsen R: **Population genetic analysis of shotgun assemblies of genomic sequences from multiple individuals.** *Genome Res* 2008, **18(7)**:1020-1029.
65. Voight BF, Kudaravalli S, Wen X, Pritchard JK: **A map of recent positive selection in the human genome.** *PLoS Biol* 2006, **4(3)**:e72.
66. Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll SA, Gaudet R, *et al*: **Genome-wide detection and characterization of positive selection in human populations.** *Nature* 2007, **449(7164)**:913.
67. Tang K, Thornton KR, Stoneking M: **A new approach for using genome scans to detect recent positive selection in the human genome.** *PLoS Biol* 2007, **5(7)**:1587-1602.
68. Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, Glanowski S, Tanenbaum DM, White TJ, Sninsky JJ, Hernandez RD, *et al*: **Natural selection on protein-coding genes in the human genome.** *Nature* 2005, **437(7062)**:1153-1157.
69. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai WW, Fritz MHY, *et al*: **A draft sequence of the Neandertal genome.** *Science* 2010, **328(5979)**:710-722.
70. Mikkelsen TS, Hillier LW, Eichler EE, Zody MC, Jaffe DB, Yang SP, Enard W, Hellmann I, Lindblad-Toh K, Altheide TK, *et al*: **Initial sequence of the chimpanzee genome and comparison with the human genome.** *Nature* 2005, **437(7055)**:69-87.
71. Bubb KL, Bovee D, Buckley D, Haugen E, Kibukawa M, Paddock M, Palmieri A, Subramanian S, Zhou Y, Kaul R, *et al*: **Scan of human genome reveals no new loci under ancient balancing selection.** *Genetics* 2006, **173(4)**:2165-2177.
72. Asthana S, Schmidt S, Sunyaev S: **A limited role for balancing selection.** *Trends Genet* 2005, **21(1)**:30-32.
73. Excoffier L, Lischer HEL: **Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows.** *Mol Ecol Resour* 2010, **10(3)**:564-567.
74. Weir BS, Cockerham CC: **Estimating F-statistics for the analysis of population structure.** *Evolution* 1984, **38(6)**:1358-1370.
75. Excoffier L, Smouse PE, Quattro JM: **Analysis of molecular variance inferred from metric distances among DNA haplotypes–application to human mitochondrial-DNA restriction data.** *Genetics* 1992, 131(2):479-491.
76. Slatkin M, Voelm L: **F(ST) in a Hierarchical island model.** *Genetics* 1991, **127(3)**:627-629.
77. R Development Core Team: **R: A language and environment for statistical computing.** Vienna, Austria: R Foundation for Statistical Computing; 2011.
78. Wald A, Wolfowitz J: **On a test whether two samples are from the same population.** *Ann Math Stat* 1940, **11**:147-162.
79. Kelley JL, Madeoy J, Calhoun JC, Swanson W, Akey JM: **Genomic signatures of positive selection in humans and the limits of outlier approaches.** *Genome Res* 2006, **16(8)**:980-989.
80. Myles S, Davison D, Barrett J, Stoneking M, Timpson N: **Worldwide population differentiation at disease-associated SNPs.** *BMC Med Genomics* 2008, **1**:10.
81. Weir BS, Cardon LR, Anderson AD, Nielsen DM, Hill WG: **Measures of human population structure show heterogeneity among genomic regions.** *Genome Res* 2005, **15(11)**:1468-1476.
82. Schmid K, Yang ZH: **The Trouble with Sliding Windows and the Selective Pressure in BRCA1.** *Plos One* 2008, **3(11)**:e3746.
83. Rabiner LR: **A tutorial on hidden markov-models and selected applications in speech secognition.** *Proc IEEE* 1989, **77(2)**:257-286.
84. Felsenstein J, Churchill GA: **A hidden Markov model approach to variation among sites in rate of evolution.** *Mol Biol Evol* 1996, **13(1)**:93-104.
85. Boys RJ, Henderson DA, Wilkinson DJ: **Detecting homogeneous segments in DNA sequences by using hidden Markov models.** *J Roy Stat Soc C-App* 2000, **49**:269-285.
86. Boitard S, Schlotterer C, Futschik A: **Detecting selective sweeps: a new approach based on hidden Markov models.** *Genetics* 2009, **181(4)**:1567-1578.
87. Kern AD, Haussler D: **A population genetic hidden Markov model for detecting genomic regions under selection.** *Mol Biol Evol* 2010, **27(7)**:1673-1685.
88. Harte D: **HiddenMarkov: Hidden Markov models.** Wellington, New Zealand: Statistics Research Associates; 2009.
89. Baum LE, Petrie T, Soules G, Weiss N: **A maximization technique occuring in statistical analyses of probabilistic functions of markov chains.** *Ann Math Stat* 1970, **41(1)**:164-171.
90. Viterbi AJ: **Error bounds for convolutional codes and an asymptotically optimum decoding algorithm.** *Ieee T Inform Theory* 1967, **13(2)**:260-269.
91. Sun WG, Cai TT: **Large-scale multiple testing under dependence.** *J Roy Stat Soc B* 2009, **71**:393-424.
92. Wei Z, Sun WG, Wang K, Hakonarson H: **Multiple testing in genome-wide association studies via hidden Markov models.** *Bioinformatics* 2009, **25(21)**:2802-2808.
93. Kong A, Thorleifsson G, Gudbjartsson DF, Masson G, Sigurdsson A, Jonasdottir A, Walters GB, Jonasdottir A, Gylfason A, Kristinsson KT, *et al*: **Fine-scale recombination rate differences between sexes, populations and individuals.** *Nature* 2010, **467(7319)**:1099-1103.
94. Bland JM, Kerry SM: **Weighted comparison of means.** *BMJ* 1998, **316(7125)**:129.
95. Hubbard TJP, Aken BL, Ayling S, Ballester B, Beal K, Bragin E, Brent S, Chen Y, Clapham P, Clarke L, *et al*: **Ensembl 2009.** *Nucleic Acids Res* 2009, **37**:D690-D697.
96. Rosenbloom KR, Dreszer TR, Pheasant M, Barber GP, Meyer LR, Pohl A, Raney BJ, Wang T, Hinrichs AS, Zweig AS, *et al*: **ENCODE whole-genome data in the UCSC genome browser.** *Nucleic Acids Res* 2010, **38(suppl 1)**: D620-D625.
97. Rhead B, Karolchik D, Kuhn RM, Hinrichs AS, Zweig AS, Fujita PA, Diekhans M, Smith KE, Rosenbloom KR, Raney BJ, *et al*: **The UCSC genome browser database: update 2010.** *Nucleic Acids Res* 2010, **38(suppl_1)**: D613-D619.
98. Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ, *et al*: **Detecting recent positive selection in the human genome from haplotype structure.** *Nature* 2002, **419**:832-837.
99. Huttley GA, Smith MW, Carrington M, O'Brien SJ: **A scan for linkage disequilibrium across the human genome.** *Genetics* 1999, **152**:1711-1722.
100. Wang ET, Kodama G, Baidi P, Moyzis RK: **Global landscape of recent inferred Darwinian selection for Homo sapiens.** *Proc Natl Acad Sci USA* 2006, **103**:135-140.
101. Voight BF, Kudaravalli S, Wen X, Pritchard JK: **A map of recent positive selection in the human genome.** *PLoS Biol* 2006, **4**:e72.
102. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, *et al*: **A second generation human haplotype map of over 3.1 million SNPs.** *Nature* 2007, **449**:851-U853.
103. Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll SA, Gaudet R, *et al*: **Genome-wide detection and characterization of positive selection in human populations.** *Nature* 2007, **449**:913.
104. Tang K, Thornton KR, Stoneking M: **A new approach for using genome scans to detect recent positive selection in the human genome.** *PLoS Biol* 2007, **5**:1587-1602.
105. Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, Absher D, Srinivasan BS, Barsh GS, Myers RM, Feldman MW, Pritchard JK: **Signals of recent positive selection in a worldwide sample of human populations.** *Genome Res* 2009, **19**:826-837.
106. Carlson CS, Thomas DJ, Eberle MA, Swanson JE, Livingston RJ, Rieder MJ, Nickerson DA: **Genomic regions exhibiting positive selection identified from dense genotype data.** *Genome Res* 2005, **15**:1553-1565.
107. Williamson SH, Hubisz MJ, Clark AG, Payseur BA, Bustamante CD, Nielsen R: **Localizing recent adaptive evolution in the human genome.** *PLoS Genet* 2007, **3**:e90.
108. Hellmann I, Mang Y, Gu ZP, Li P, de la Vega FM, Clark AG, Nielsen R: **Population genetic analysis of shotgun assemblies of genomic sequences from multiple individuals.** *Genome Res* 2008, **18**:1020-1029.
109. Nielsen R, Hubisz MJ, Hellmann I, Torgerson D, Andres AM, Albrechtsen A, Gutenkunst R, Adams MD, Cargill M, Boyko A, *et al*: **Darwinian and demographic forces affecting human protein coding genes.** *Genome Res* 2009, **19**:838-849.
110. Altshuler D, Brooks LD, Chakravarti A, Collins FS, Daly MJ, Donnelly P, Gibbs RA, Belmont JW, Boudreau A, Leal SM, *et al*: **A haplotype map of the human genome.** *Nature* 2005, **437**:1299-1320.
111. Chen H, Patterson N, Reich D: **Population differentiation as a test for selective sweeps.** *Genome Res* 2010, **20**:393-402.
112. Grossman SR, Shylakhter I, Karlsson EK, Byrne EH, Morales S, Frieden G, Hostetter E, Angelino E, Garber M, Zuk O, *et al*: **A composite of multiple signals distinguishes causal variants in regions of positive selection.** *Science* 2010, **327**:883-886.

113. Akey JM, Zhang G, Zhang K, Jin L, Shriver MD: **Interrogating a high-density SNP map for signatures of natural selection.** *Genome Res* 2002, **12**:1805-1814.
114. Mikkelsen TS, Hillier LW, Eichler EE, Zody MC, Jaffe DB, Yang SP, Enard W, Hellmann I, Lindblad-Toh K, Altheide TK, *et al*: **Initial sequence of the chimpanzee genome and comparison with the human genome.** *Nature* 2005, **437**:69-87.
115. Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, Glanowski S, Tanenbaum DM, White TJ, Sninsky JJ, Hernandez RD, *et al*: **Natural selection on protein-coding genes in the human genome.** *Nature* 2005, **437**:1153-1157.
116. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai WW, Fritz MHY, *et al*: **A draft sequence of the Neandertal genome.** *Science* 2010, **328**:710-722.
117. Andres AM, Hubisz MJ, Indap A, Torgerson DG, Degenhardt JD, Boyko AR, Gutenkunst RN, White TJ, Green ED, Bustamante CD, *et al*: **Targets of balancing selection in the human genome.** *Mol Biol Evol* 2009, **26**:2755-2764.

doi:10.1186/1471-2164-13-107
Cite this article as: Hofer *et al.*: **Evolutionary forces shaping genomic islands of population differentiation in humans.** *BMC Genomics* 2012 **13**:107.