Title: Mixture Density Regression reveals frequent recent adaptation in the human genome

Authors: Diego F. Salazar-Tortosa[1,2], Yi-Fei Huang[3,4], David Enard[1]

Affiliations:

[1]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ, USA.

[2]PROFITH 'PROmoting FITness and Health through physical activity' research group, Sport and Health University Research Institute (iMUDS), University of Granada, Granada, Spain.

[3]Department of Biology, Pennsylvania State University, University Park, PA, USA.

[4]Huck Institutes of the Life Sciences, Pennsylvania State University, University Park, PA, USA.

Corresponding Authors: David Enard: denard@email.arizona.edu; Yi-Fei Huang: yuh371@psu.edu

Running head: MDR-based analysis of positive selection.

Keywords: human evolution, recent adaptation, mixture distributions, selection determinants, recombination rate.

Abstract

How much genome differences between species reflect neutral or adaptive evolution is a central question in evolutionary genomics. In humans and other mammals, the prevalence of adaptive versus neutral genomic evolution has proven particularly difficult to quantify. The difficulty notably stems from the highly heterogenous organization of mammalian genomes at multiple levels (functional sequence density, recombination, etc.) that complicates the interpretation and distinction of adaptive vs. neutral evolution signals. Here, we introduce Mixture Density Regressions (MDRs) for the study of the determinants of recent adaptation in the human genome. MDRs provide a flexible regression model based on multiple Gaussian distributions. We use MDRs to model the association between recent selection signals and multiple genomic factors likely to affect positive selection, if the latter was common enough in the first place to generate these associations. We find that a MDR model with two Gaussian distributions provides an excellent fit to the genome-wide distribution of a common sweep summary statistic (iHS), with one of the two distributions likely capturing the positively selected component of the genome. We further find several factors associated with recent adaptation, including the recombination rate, the density of regulatory elements in immune cells and testis, GC-content, gene expression in immune cells, the density of mammal-wide conserved elements, and the distance to the nearest virus-interacting gene. These results support that strong positive selection was relatively common in recent human evolution and highlight MDRs as a powerful tool to make sense of signals of recent genomic adaptation.

Introduction

The characteristics of the human genome can influence both the occurrence and the detection of recent positive selection in a specific genomic region. Similar to other mammals, the human genome has a complex organization, with highly heterogeneous recombination rates and distribution of functional elements (regulatory or coding) along

chromosomes. This inherent heterogeneity in the factors that are likely to influence both the occurrence of positive selection and our ability to detect it, has complicated the study of the prevalence of recent adaptation in the human genome, and has made the topic controversial (Enard et al. 2014; Hernandez et al. 2011). Even though there is growing evidence that recent positive selection in the form of selective sweeps may have been relatively common during recent human evolution (Enard and Petrov 2020; Souilmi et al. 2021), further resolution of the controversy will depend on our ability to understand what factors govern the local genomic rate of recent adaptation along human chromosomes. Indeed, if signatures of selective sweeps do not occur randomly, as expected if all sweep signals are false positives only reflecting genetic drift, but instead associate non-randomly with functional elements in the genome, then this provides evidence that sweeps had to be common enough in the first place to generate the association.

What factors then are *a priori* expected to matter for recent adaptation? Among all genomic factors, recombination is a key process determining the patterns of linkage between alleles across the human genome (The International HapMap Consortium 2005), that can strongly influence the probability to detect recent selection in the form of selective sweeps. The lower the recombination rate, the larger the genomic region where neutral variants will hitchhike to higher frequencies along with an advantageous variant, before recombination breaks down their linkage (Sabeti et al. 2006; Pritchard et al. 2010). The higher the recombination rate, the smaller the genomic region affected by the sweep. Different sweep sizes in low and high recombination regions result in a higher statistical power to detect sweeps in low than in high recombination regions (O'Reilly et al. 2008; Booker et al. 2020; Johnson and Voight 2018). In parallel, larger neutral haplotypes make low recombination regions more prone to detect false positive sweeps (O'Reilly et al. 2008; Booker et al. 2020). This heterogeneity, together with other factors such as background selection, has greatly contributed to a persistent debate around the prevalence of selective sweeps in the human genome.

If recent adaptation and selective sweeps in particular were relatively common during recent human evolution, we expect that other factors on top of recombination should be relevant for their occurrence and detection in the human genome. Generally,

we expect that selective sweeps should occur more frequently around functional segments of the genome where adaptive mutations are expected to take place (Enard et al. 2014). Therefore, we should find enough signals of positive association between selective sweeps and the overall density of functional elements, either coding or non-coding. We also expect specific functions to be associated with an increased occurrence of selective sweeps. For example, the presence of genes coding for proteins that interact with pathogens, and viruses in particular (Virus-Interacting Proteins or VIPs), should influence the frequency of sweeps in a given genomic region, independently of other genomic features (Enard and Petrov 2020; Souilmi et al. 2021). More generally, genomic regions with immune genes are expected to have experienced more sweeps (Barreiro et al. 2009; Enard et al. 2016; Deschamps et al. 2016; Enard and Petrov 2018). Similarly, reproduction-related functions show signals of positive selection (Voight et al. 2006; Nielsen et al. 2005), thus tissues related with these functions should be associated with the frequency of sweeps.

Several summary statistics are currently available to detect recent genomic adaptation. Among all possible choices, the statistics that use the structure of haplotypes along chromosomes are the most appropriate to investigate the prevalence and determinants of recent positive selection. First, haplotype-based statistics have good statistical power to detect strong and recent incomplete sweeps (Voight et al. 2006; Ferrer-Admetlla et al. 2014; Garud et al. 2015; Enard and Petrov 2020). Second, they are not confounded by other processes such as background selection (BGS) (Enard et al. 2014; Schrider 2020). This insensitivity to BGS is a particularly important attribute when testing the specific association between recent positive selection and multiple genomic factors that can also correlate with BGS. In this regard, the integrated Haplotype Score (iHS) (Voight et al. 2006) is a haplotype-based statistic that has been extensively tested (Voight et al. 2006; Barreiro et al. 2009; Enard et al. 2014; Johnson and Voight 2018), and can detect selection signals both from *de novo* mutations and standing genetic variation, provided that selection started from not too high initial allele frequency (Enard and Petrov 2020). This summary statistic can be used to scan many genomes given the availability of fast implementations (Maclean et al. 2015).

Here, we model the genomic determinants of recent positive selection in the human genome. We measure recent selective sweeps with the iHS statistic from five human populations represented in the 1,000 Genomes phase 3 dataset (The 1000 Genomes Project Consortium 2015). The iHS statistic has more power to detect recent and incomplete sweeps, while it has reduced power to detect complete sweeps or sweeps more than 30,000 years old (Sabeti et al. 2006; Enard and Petrov 2020). Therefore, we restrict our analyses to selection that occurred after the main human migration out of Africa (Henn et al. 2012). We use Mixture Density Regressions (MDRs) to test the association between iHS and several genomic factors that are possible determinants of selection. MDRs can fit a mixture of several distributions to the observed statistic. In our case, we fit a mixture of two Gaussian distributions to the observed iHS, expecting that one distribution matches the drift component of iHS, while the second one fits its selected component. Then, we look at the genomic factors that influence iHS in the regression model, but only considering the selected component (Methods). This differs from other approaches trying to understand the determinants of selection in two important ways. First, previous studies have focused on the detection of partial correlations between selection and single genomic factors (e.g., Lohmueller et al. (2011)). This approach is likely limited to model the impact of multiple genomic factors that could influence selection simultaneously. Second, they analyze the overall correlation without distinguishing between the selected and non-selected components of the genome. Therefore, the true impact of positive selection could be diluted among values of iHS that are not influenced by selective sweeps (i.e., drift component). Given that the impact of hitchhiking on the human genome is likely localized and not expected to influence the entire genome, an approach specifically distinguishing a selected component should have more power to reveal the impact of positive selection.

As expected according to previous studies, we find a strong association between recombination rate and recent positive selection signals in the human genome. In addition to recombination, we find that other factors, such as the density of regulatory elements in immune cells and testis, GC-content, gene expression in immune cells, the density of mammal-wide conserved elements and the distance to VIPs, all show strong associations with recent positive selection. These results imply that recent positive selection had to be

common enough to create these associations, and further highlight how Mixed Density Regressions (MDRs) can be used to clarify the prevalence and determinants of recent positive selection based on genome scans.

Results

We consider multiple genomic factors to model the association with recent selection signals. These factors are possible determinants of selection according to previous evidence (Siepel et al. 2005; Nielsen et al. 2005; O'Reilly et al. 2008; Enard et al. 2014; Luisi et al. 2015; Booker et al. 2020; Enard and Petrov 2020; Voight et al. 2006). We use a gene-centric perspective where each individual protein-coding gene in the human genome is associated with a set of factors measured by centering a genomic window on each gene (Methods). This approach is likely to detect the influence of factors that differ between genes, rather than only differences between genic and intergenic regions. The set of factors in the MDR model includes several genomic features, like the length of the gene at the center of the window or recombination estimates from the latest deCODE genetic map (Methods). We obtain other genomic factors, like the GC-content and the densities of genes, coding sequences and conserved elements. Moreover, we include the density of transcription factors binding sites according to the hypersensitivity to DNaseI and chromatin immunoprecipitation (ChIP-seq) experiments (The ENCODE Project Consortium 2012). In the latter case, we also obtain the density of regulatory elements for specific subsets of cell lines: testis and immune cells. Table 1 provides a complete list of the factors considered, and the Methods section details how these factors were obtained.

We also consider other functional factors like gene expression, including the average gene expression across 53 GTEx v7 tissues (GTEx Consortium 2015). The expression in testis and immune cells could impact the frequency of sweeps, thus we also include them as independent expression variables (Nielsen et al. 2005; Voight et al. 2006). We also consider the number of protein-protein interactions (PPIs) of each gene, given that this variable has been previously associated with the rate of sweeps (Luisi et al. 2015). Finally, we consider the influence of viruses on genomic adaptation, as previous

evidence suggests that they have acted as a strong selective pressure during recent human evolution (Enard and Petrov 2020; Souilmi et al. 2021). We use the distance of each gene to the closest gene coding for a protein that interacts with viruses. More details for all the factors considered are given in the Methods.

Genomic factors that are calculated across the genomic windows can vary depending on the window size used. Therefore, we infer the association between recent selection and factors measured in genomic windows of varying sizes (Methods). For example, we measure the association between the average iHS within 50kb windows and the average recombination rate within the same 50kb windows, or between the average iHS and recombination within the same 1000kb windows (Methods). Using different window sizes also relaxes assumptions about the expected strength of selection, as larger windows are likely more sensitive to strong selection compared to smaller windows, and vice versa (Enard and Petrov 2020). In total, we use five different window sizes (50, 100, 200, 500 and 1,000 kb) centered at the genomic center of Ensembl v99 coding genes (Yates et al. 2020). Using fixed size windows irrespective of gene length avoids a priori biasing measures of iHS by gene length. We also account for gene length in our model through the inclusion of multiple functional densities along with the genomic length of the gene (between transcription start and stop) at the center of the window as potential explanatory variables.

We estimate recent selection signals using the iHS statistic in five populations represented in the 1,000 Genomes Project (The 1000 Genomes Project Consortium 2015). Each population represents a different continent: Yoruba for Africa, Han Chinese for East Asia, Utah residents with Northern and Western European ancestry, Toscani for Southern Europe, and Peruvians for Americas. We select Peruvians given that they show the highest percentage of Native American ancestry among populations included in the 1,000 Genomes Project (The 1000 Genomes Project Consortium 2015; Harris et al. 2018). The iHS statistic measures how far the haplotypes carrying a specific derived allele extend upstream and downstream of this focal allele, compared to how far haplotypes carrying the ancestral allele extend (Voight et al. 2006). As such, the approach provides one value of iHS for each biallelic Single Nucleotide Polymorphism (SNP) in

the genome with clearly defined derived and ancestral alleles (Methods). Although isolated extreme values of iHS at a single SNP have often been used as a signature of selection, we recently found that measuring the average absolute value of iHS across all the SNPs in a genomic window provides high power to detect recent selection (Enard and Petrov 2020). The average iHS in an entire window is also likely to better associate with other factors similarly measured as averages in the same windows. To account for the fact that the variance of the average iHS depends on the number of SNPs with individual iHS values in a window, we include this number as a predictor in the tested MDR models.

MDR fit to recent selection signals

For each separate human population and window size, we model the association between genomic factors and iHS using a MDR with a mixture of two Gaussian distributions. We design one distribution to fit the non-selective, drift component of iHS that is not expected to correlate with genomic factors. In contrast, the second distribution is designed to fit the selective component of iHS and it is expected to associate with the tested genomic factors (Methods). We apply this approach to account for the fact that only a part of the genome is expected to have undergone recent positive selection. In this context, a very small second distribution and non-significant associations with the tested predictive factors would suggest a very small contribution of recent positive selection. The patterns of iHS would then be entirely dominated by genetic drift and past demography. We use this modeling approach together with an optimization algorithm to get the overall mixture distribution with the best fit for the observed iHS in each population and window size (Methods).

Overall, the MDR models fit well the distribution of observed iHS (Figure 1). iHS does not follow a Gaussian distribution, and our flexible approach combining a mixture of two Gaussian distributions and linear regression clearly fits better the iHS distribution than would a single distribution from a classic linear model. We find a clear separation between the selected and non-selected components, with the selected component shifted

toward higher iHS values (Figure 1). Together, the close fit and separation of the two gaussian distributions likely provide an improved ability of our model to estimate the contributions of different factors to iHS.

Patterns of selection across populations and window sizes

We find different patterns of selection across five populations from four different continents (Figure 1). The African Yoruba population shows the largest selected component (Figure 1A). The selected component is clearly shifted towards high iHS values, indicating stronger selection signals that stand out more from the non-selected component. The selection components are less pronounced but still substantial in the four other populations (Figure 1B-1E). The Peruvian population shows the smallest selected component (Figure 1E). These results suggest that the visible (not necessarily the actual) contribution of selection to the whole variation of iHS is smaller in populations that were exposed to bottlenecks during migrations Out of Africa and consequently, more subject to genetic drift (Henn et al. 2012; The 1000 Genomes Project Consortium 2015). In other words, our results might be explained by the fact that selection is more visible in Yoruba due to a reduced effect of genetic drift.

With respect to window size, we find the largest selective components when using larger, 1,000 kb windows (Supplementary Data 1). This is an expected result, because it is easier to distinguish strong selection from genetic drift and background noise in larger windows. Smaller windows are less sensitive to strong selection, while larger windows can include a great accumulation of large iHS values. In addition, small windows are more influenced not only by the content of a genomic window, but also by genetically linked surrounding regions. These surrounding areas are in linkage disequilibrium with the genomic areas inside smaller windows as their center is closer to the edges. Therefore, smaller windows do not differentiate well between inside and outside genomic regions (Enard et al. 2014; Enard and Petrov 2020). Given these results, we mainly focus the rest of our analyses on the Yoruba population and 1,000 kb windows.

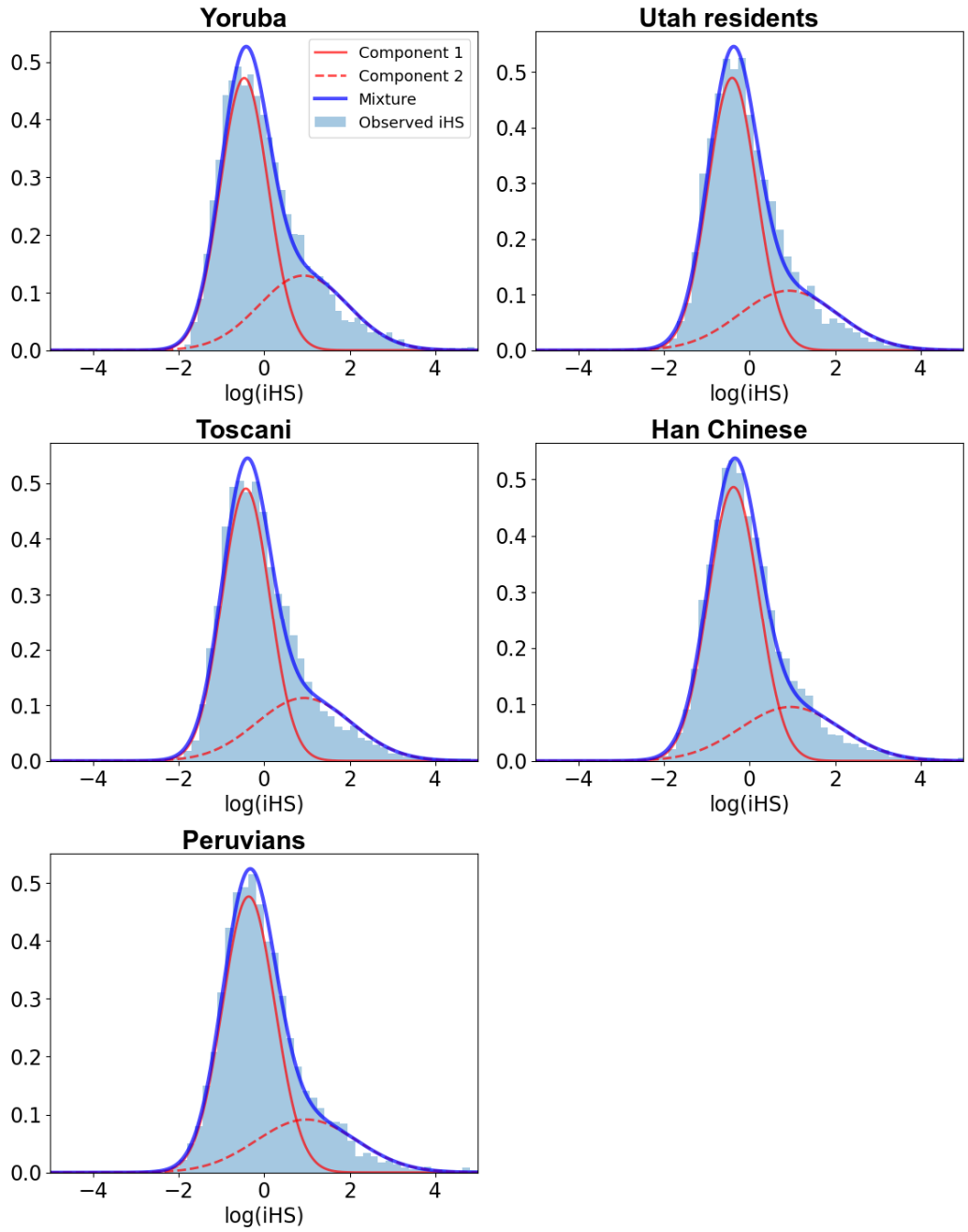

Figure 1. Mixture of Gaussian distributions fitting observed iHS (1,000 kb windows) for the five studied populations: A) Africa - Yoruba; B) Europe - Utah residents with Northern and Western European ancestry; C) Europe - Toscani; D) East Asia - Han Chinese; E) America - Peruvians. The figure shows the non-selected and selected components (Component 1 and 2, respectively) for each population. In the latter component, iHS linearly depends on the genomic factors considered. Legend: Light blue = Observed iHS; Dark blue = Mixture model; Full red curve = Component 1 of the mixture model without selection; Dashed red curve = Component 2 of the mixture model with selection.

Determinants of strong positive selection in the Yoruba population

Several genomic factors show significant associations with iHS in the selected component (Table 1). Unsurprisingly, we find the strongest association with recombination rate, which negatively correlates with iHS (slope=-2.44; p-value<1E-16; Table 1). This is congruent with the eroding effect of recombination on selection signatures. Recombination can break haplotypes generated by selection, making it more difficult to detect the selective events (O'Reilly et al. 2008; Booker et al. 2020). These results confirm the role of recombination as a key determinant of the probability of detecting selection with iHS (Johnson and Voight 2018). We also find an association between GC-content and recent positive selection. When the effect of recombination is not accounted for, that is, recombination is not included in the model, GC-content negatively correlates with iHS (slope=-0.29; p-value=1.22E-03). This is expected given that GC-content is positively associated with long-term recombination (Duret and Arndt 2008). However, adding the control for recombination reveals a positive association between GC-content and iHS (slope=0.55; p-value=1.34E-06; Table 1). Given that recombination already explains a great proportion of iHS variability including that shared with GC-content, the positive association between GC-content and iHS could be independent from recombination. Accounting for the eroding effect of recombination on haplotypes could make more visible a positive (direct or indirect) influence of GC-content on selection. Indeed, this association remains significant when considering only low-recombination regions, further suggesting its independence from the eroding effect of recombination (see below).

GC-content might be a better proxy of overall functional density than individual functional factors included in the regression model, which could explain its positive association with iHS. We test this hypothesis by removing different genomic factors from the model (see Supplementary Data 2 for modeling results after removing different sets of factors). The removal of GC-content increases the association of iHS with both coding density (slope=0.22 vs 0.16; p-value=1.32E-02 vs. 8.41E-02) and DNAseI hypersensitivity, which is a measure of the overall density of transcription-factor binding

sites (slope=0.44 vs. -0.047; p-value=1.77E-07 vs. 7.15E-01). Since GC-content has been positively associated with the density of coding sequences and the hypersensitivity to DNAseI (Lander et al. 2001; Di Filippo and Bernardi 2008), removing GC-content may thus make more visible a positive association between iHS and functional densities. We further test if this is also the case for more tissue-specific functional densities such as the density of regulatory elements in testis or immune cells that are expected to exhibit more positive selection (Methods), but we find no such evidence. The removal of GC-content does not affect to the association between selection and regulatory density in testis (slope=-0.765 vs. -0.795; p-value<1E-16 in both cases) and immune cells (slope=0.072 vs. 0.047; p-value=4.5E-01 vs. 6.23E-01). As shown later in this manuscript, fine-scale patterns of recombination seem to hinder the detection of associations for these regulatory variables. Finally, we find that the removal of coding density and DNAseI hypersensitivity increases the significance for the association between GC-content and iHS (p-value=2.40E-14 vs. 1.34E-06). From this model, we then further remove additional functional factors whose association with selection is also affected by the removal of GC-content, namely, gene number, gene length and ChIP-seq regulatory density (Methods). Accordingly, the removal of these functional factors leads to a highly significant association between GC-content and recent positive selection (slope=0.49; p-value<1E-16). In summary, the removal of different functional genomic factors from the model supports that the positive association between positive selection and GC-content is mediated by the role of this factor as proxy of overall functional density. However, we cannot exclude the implication of other processes related to recombination that would require more detailed analyses (for example, see fine scale patterns related to regulatory density below).

Table 1. Slopes and p-values of the association between iHS and genomic factors for the Yoruba population in 1,000 kb windows.

Covariate	Slope	p-value
Intercept	-1.426	<1E-16
Recombination rate	-2.435	<1E-16

GC-content	0.55	1.34E-06
Density of conserved elements	0.159	6.40E-04
Number PPIs	-0.07	4.25E-02
Distance to VIPs	-0.17	4.16E-06
Gene expression	-0.139	3.77E-02
Gene expression in immune cells	0.265	1.51E-05
Gene expression in testis	0.001	9.90E-01
Regulatory density (ChIP-seq)	-0.147	2.37E-01
Regulatory density in immune cells (ChIP-seq)	0.047	6.24E-01
Regulatory density in testis (ChIP-seq)	-0.795	<1E-16
Regulatory density (DNaseI)	-0.047	7.15E-01
Gene length	-0.059	1.12E-01
Gene number	-0.086	3.46E-01
Coding density	0.159	8.41E-02
Number iHS data points	-0.007	8.19E-01

PPIs: Protein-Protein Interactions. VIPs: Virus Interacting Proteins. ChIP-seq: Density of regulatory elements according to chromatin immunoprecipitation experiments. DNaseI: Density of regulatory elements according to the hypersensitivity to DNaseI. iHS: Integrated Haplotype Score.

Another genomic factor associated with selection is the density of conserved elements (Methods), showing a positive association with iHS (slope=0.16; p-value=6.4E-04; Table 1). This is an expected result since coding sequences and non-coding regulatory elements tend to be conserved (Siepel et al. 2005). The density of conserved elements is thus likely to be a good proxy of overall (coding and non-coding) functional density. The inclusion of this factor in the model could then, however, explain the lack of association for other factors related to functional density such as coding density. Indeed, the removal of conserved elements density makes significant the positive association between coding density and selection (slope=0.26 vs. 0.16; p-value=2.35E-03 vs. 8.41E-02; Supplementary Data 2). In addition, the simultaneous removal of conserved elements density and GC-content, which can also act as a proxy of functional density, increases even more the positive association of coding density with iHS (slope=0.29; p-value=7.76E-04). These, and previous results about GC-content suggest that the lack of association for multiple functional factors could be explained, at least partially, by the

simultaneous consideration of better measures of overall functional density in our models.

The number of protein-protein interactions is another factor that correlates with iHS. This factor associates negatively with selection, suggesting the existence of a slight depletion of selection for regions with higher number of PPIs (slope=-0.07; p-value=4.25E-02; Table 1). This is opposite to that found by Luisi et al. (2015), as they reported higher recent positive selection on central proteins within the human interactome. Higher number of PPIs has been also associated with higher positive selection in the chimpanzee lineage according to the McDonald-Kreitman (MK) test. However, this factor does not seems to be a determinant of positive selection when other genomic factors are simultaneously considered using a MK regression approach (Huang 2021). Given the inconsistency with previous evidence and the low strength of the association, the weak negative relationship between selection and the number of PPIs found in this study should be taken with caution.

The modeling approach presented in this study can also be used to test the influence of specific selective pressures. We illustrate this feature by showing that positive selection is associated with multiple variables related to viral interaction (Table 1). One of the factors more highly associated with iHS is the distance to VIPs (slope=-0.17; p-value=4.16E-06). Selection decreases further away from VIPs, in other words, we find an enrichment of selection around VIPs. Viruses have acted as a strong selective pressure during human evolution, shaping genomic adaptation, and previous studies have also found more positive selection at VIPs (Enard and Petrov 2020; Souilmi et al. 2021). Gene expression is another functional factor significantly associated with selection (Table 1). The average expression across 53 tissues from GTEx (GTEx Consortium 2015) is only slightly and negatively associated with iHS (slope=-0.14; p-value=3.77E-02), while gene expression in immune cells shows a strong and positive association with selection (slope=0.26, p-value=1.51E-05). Given their relevance in the response to pathogens, genomic regions highly expressed in immune cells do represent expected targets of positive selection.

The densities of regulatory elements in tissues expected to experience more positive selection (testis and immune cells) show surprising patterns of association with iHS (Table 1). In particular, a higher density of transcription factor binding sites in testis correlates with weaker iHS selection signals (slope=-0.79; p-value<1E-16). This is a counterintuitive result, given that more regulatory density would give more options to modulate gene expression and hence more room for positive selection to act. An explanation could be the following: hotspots of recombination are usually regions where the chromatin is more open, and hence more accessible for transcription factors to bind (Spruce et al. 2020). If the density of transcription factor binding sites in testis (where meiosis and recombination occur) is higher in areas with high recombination, then selection signals within these sites would be erased by recombination (O'Reilly et al. 2008; Booker et al. 2020). Fine scale patterns of recombination and selection support this hypothesis, as recombination tends to increase close to regulatory elements, while iHS tends to decrease (Supplementary Figures 1, 2, 3). Therefore, high levels of local recombination could erode haplotypes close to the selected variants, although the average recombination in the gene window covering that region is low. In other words, the 1,000 kb gene windows we use could have a low recombination rate on average, while containing local peaks of recombination close to regulatory elements. Our original model considers the average recombination in each 1,000 kb gene-centered window, so it is likely unable to detect the effect of these fine-scale patterns of recombination (see below for further analyses in this respect).

Robustness of recent selection patterns to varying genomic window sizes

Next, we ask if the trends observed with 1,000 kb windows in the Yoruba genomes also hold or not when using smaller window sizes (Supplementary Data 1). Some patterns visible in large 1000kb windows are also visible when using smaller genomic window sizes, while some are not (100 kb vs 1,000 kb windows; Supplementary Data 1). The strong negative association between recombination rate and selection is found regardless of window size. On the contrary, the distance to VIPs for example shows a strong association with iHS in large windows but not in smaller windows (100 kb windows:

slope=-0.02; p-value=5.11E-01). This is consistent with previous evidence showing that VIPs are particularly enriched in strong selective sweeps, for which larger windows are more sensitive (Enard and Petrov 2020). This is also the case for the density of conserved elements (100 kb windows: slope=0.1; p-value=9.3E-02) and especially regulatory elements in testis, with a slope going from -0.70 in 1,000 kb windows to -0.03 in 100 kb windows. These results suggest that both factors may correlate with strong rather than weak selection. In the case of testis regulatory density, this pattern might be explained by the fact that smaller windows take better into account the fine scale patterns of recombination, which limits the negative influence of using average recombination (see below for fine-scale analyses of recombination). Note, however, that all these differences might also be explained by the fact that smaller windows do not differentiate well between inside and outside of the genomic regions they are supposed to delineate. That is, they may be more influenced by the surrounding, genetically linked genomic regions (Enard et al. 2014; Enard and Petrov 2020).

Robustness of the recent selection patterns across human populations

We next ask whether the selection patterns observed in the Yoruba genomes are consistent across different human populations. Association patterns with iHS are largely similar across the five populations studied. Recombination rate and GC-content show strong associations with iHS in 1000kb window across all populations (Tables 1 and 2). Similarly, functional factors consistently correlate with selection across populations. This is the case of VIP distance, gene expression in immune cells and density of regulatory elements in testis. Most patterns are significant across all groups, but to varying extents. One notable difference is that the density of conserved elements associates with selection in 1000kb windows for all populations except for the Peruvian population sample, which might reflect the impact of past bottlenecks on the visibility of the selective patterns (Tables 1 and 2). Regulatory density across multiple tissues according to ChIP-seq experiments and hypersensitivity to DNaseI along with coding density correlate with selection in all populations except in Yoruba (Tables 1 and 2). Despite we detect a masking effect of GC-content and conserved elements on functional factors (see above),

it is still surprising that the association between selection and some functional factors is not readily visible in the Yoruba population, while it is in others. The selected component is larger (i.e., selection is more visible) in Yoruba than in other populations. Therefore, the association between functional density and iHS should be more visible in this compared to other populations that have been affected by large bottlenecks connected to the Out of Africa migration. We will return to this apparent contradiction later in the Discussion. See Supplementary Data 1 for the complete results across populations and window sizes.

Table 2. Slopes and p-values of the association between iHS and genomic factors for four of the five studied populations in 1,000 kb windows: Utah residents, Toscani, Han Chinese and Peruvians.

Covariate	Utah residents		Toscani		Han Chinese		Peruvians	
	Slope	**p-value**	**Slope**	**p-value**	**Slope**	**p-value**	**Slope**	**p-value**
Intercept	-1.77	<1E-16	-1.66	<1E-16	-1.78	<1E-16	-1.55	<1E-16
Recombination rate	-2.80	<1E-16	-2.70	<1E-16	-2.51	<1E-16	-1.74	<1E-16
GC-content	0.75	1.35E-08	0.52	2.47E-05	1.17	<1E-16	0.35	2.63E-03
Density of conserved elements	0.38	7.74E-13	0.34	2.70E-11	0.31	4.20E-10	0.06	2.43E-01
Number PPIs	-0.05	1.95E-01	-0.04	2.67E-01	-0.09	1.92E-02	-0.04	3.00E-01
Distance to VIPs	-0.34	8.45E-14	-0.27	1.92E-10	-0.14	5.35E-04	-0.18	2.77E-05
Gene expression	-0.34	9.70E-06	-0.25	4.92E-04	-0.25	1.14E-03	-0.26	2.56E-04
Gene expression in immune cells	0.41	6.87E-09	0.37	5.34E-08	0.31	9.51E-06	0.25	1.56E-04
Gene expression in testis	-0.03	5.45E-01	-0.04	4.21E-01	0.07	2.06E-01	0.05	2.61E-01
Regulatory density (ChIP-seq)	0.59	3.45E-06	0.74	3.06E-10	0.49	3.92E-05	0.33	7.64E-03
Regulatory density in immune cells (ChIP-seq)	-0.05	6.01E-01	-0.08	3.94E-01	-0.16	1.26E-01	0.15	1.32E-01
Regulatory density in testis (ChIP-seq)	-1.03	<1E-16	-0.86	<1E-16	-0.39	1.65E-09	-0.11	4.76E-02
Regulatory density (DNaseI)	-0.61	5.45E-05	-0.64	8.24E-06	-0.84	1.66E-08	-0.55	8.27E-05
Gene length	-0.05	2.84E-01	-0.02	5.59E-01	0.00	9.13E-01	0.02	5.46E-01
Gene number	-0.52	1.28E-06	-0.40	5.91E-05	-0.46	6.23E-04	-0.02	7.96E-01
Coding density	0.54	8.72E-06	0.59	3.77E-08	0.40	1.93E-03	0.20	2.73E-02
Number iHS data points	0.32	<1E-16	0.33	<1E-16	0.16	7.29E-08	0.08	1.51E-03

Genomic determinants of selection in low-recombination regions

Given the strong influence of recombination on selection across populations and window sizes, we further assess whether the rest of significant associations with other genomic factors are truly independent of recombination in the MDR, and whether insufficient accounting of recombination can explain the observed selected component in the MDR, instead of actual recent positive selection. The risk indeed exists that because we use an imperfect measure of recombination in our model (due to measurements errors in recombination maps and variation across populations), the effect of recombination on iHS may not be properly removed. The selected component of the distribution of iHS might then reflect an effect of recombination, rather than actual signals of recent positive selection.

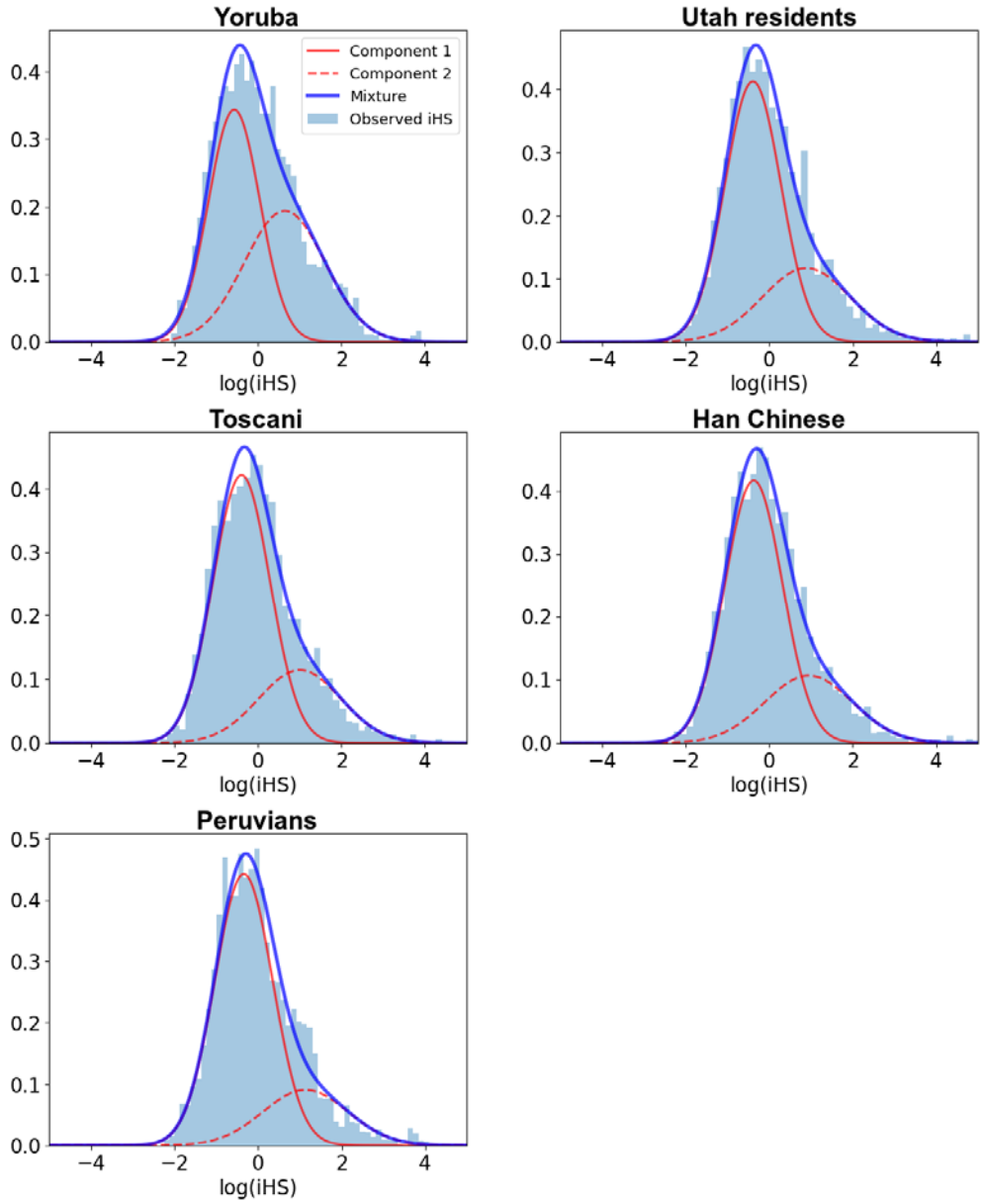

Figure 2. Mixture of Gaussian distributions fitting observed iHS (1,000 kb windows) in low-recombination regions for the five studied populations: A) Africa - Yoruba; B) Europe - Utah residents with Northern and Western European ancestry; C) Europe - Toscani; D) East Asia - Han Chinese; E) America - Peruvians. The figure shows the non-selected and selected components (Component 1 and 2, respectively) for each population considering only genes in low-recombination regions. In the latter component, iHS linearly depends on the genomic factors considered. Legend: Light blue = Observed iHS; Dark blue = Mixture model; Full red curve = Component 1 of the mixture model without selection; Dashed red curve = Component 2 of the mixture model with selection.

Because of this concern, we test if our results are robust to recombination by recalculating the association between all the genomic factors and iHS, but this time considering only genes with more similar recombination rates overall, in particular lower recombination rates. Indeed, by using genes with more homogeneous recombination rates and a reduced range of recombination values, we should also reduce any indirect effect of recombination. We focus on genes with low recombination rates because we can predict that if the significant associations found are due to actual positive selection, then the positive selection component predicted by the MDR should become more prominent. This is expected because positive selection is more readily detectable in low recombination regions, where the eroding effect of recombination is reduced (O'Reilly et al. 2008; Booker et al. 2020). Conversely, if the selected component and the significant associations are in fact due to residual correlations with recombination but not actual selection, then the selective component should become less prominent when focusing on regions with more homogeneous, lower recombination rates. Note, however, that for specific associations between iHS and genomic factors, a decrease of significance could be caused by a reduction in gene sample size. The consideration of only a subset of genes could reduce the power to detect association for genomic factors that correlate with recombination. Focusing only on lower recombination genes may restrict the range of values taken by these genomic factors, which could in turn limit our ability to detect positive associations with iHS.

We select those genes with a recombination rate equal to or lower than the lowest tertile of recombination across genes (0.837 cM/Mb). Larger windows are more sensitive to strong selection and the estimates of genomic factors are more precise in these windows. In addition, the Yoruba population again shows the largest selected component across all the studied populations (Figure 2). Therefore, in the main text we only describe results with 1,000 kb windows in Yoruba. See Supplementary Data 3 for full results of low-recombination regions across other populations and windows sizes.

As expected from reducing the range of recombination values, running the model with similar levels of recombination substantially decreases the regression slope between recombination rate and iHS compared to the whole set of genes (slope: -1.01 vs. -2.44;

Table 3). In contrast, the selected component (second distribution in the model) becomes larger (Figure 2 compared to Figure 1). This is consistent with the expectation that recent positive selection is more readily detectable in regions where recombination is less frequent, and hence long haplotypes caused by selection can increase in frequency and remain unbroken for a longer amount of evolutionary time. Congruent with this, the distance to VIPs is more associated with selection in low-recombination regions (slope=-0.34 vs. -0.17 in the entire gene set; p-value=2.18E-05 vs. 4.16E-06). An expected result given that VIPs distance does not correlate strongly with recombination (Spearman's ρ=0.028; p-value=4.40E-04; correlation calculated across 1000 kb windows), thus its range of values is not restricted by focusing only on low recombination genes. Despite GC-content is strongly correlated with recombination (Spearman's ρ=0.30; p-value<1E-16; correlation calculated across 1000 kb windows), its association with iHS remains very strong and positive at low recombination genes (slope=0.93 vs. 0.55; p-value=1.17E-06 vs. 1.34E-06; Table 3). This latter association suggests that the positive trend of GC-content may be independent of recombination, given that it increases when focusing only on low-recombination regions. Regulatory density in testis maintains a negative association with iHS (slope=-0.76 vs. -0.79; p-value=1.10E-11 vs. <1E-16; Table 3), while regulatory density across multiple tissues becomes significant and negatively associated with selection in low-recombination regions (ChIP-seq; slope=-0.63 vs. -0.15, p-value=7.22E-04 vs. 2.37E-01). As previously explained, a higher accessibility of transcription factors to DNA in recombination hotspots should mask the influence of regulatory density on selection. Indeed, if the mismatch between recombination at local and gene-window scale is the cause of this pattern, focusing on low-recombination regions should increase this confounding effect, by increasing the discrepancy between low window-wide average recombination rates and more elevated fine-scale recombination rates around regulatory elements. We specifically focus on gene windows having overall low recombination, but they could still have fine-scale local peaks of recombination, resulting in an increased mismatch between scales. Results from this and previous sections suggest that a fine control of recombination rate is needed in order to limit its confounding effect on the detection of recent positive selection. In the

next section we further analyze the association of iHS with genomic factors, but this time controlling for fine-scale patterns of recombination.

Table 3. Slopes and p-values of the association between iHS and genomic factors for the Yoruba population in 1,000 kb windows. These associations were calculated considering only low-recombination regions.

Covariate	Slope	p-value
Intercept	-0.153	3.56E-01
Recombination rate	-1.005	<1E-16
GC-content	0.929	1.17E-06
Density of conserved elements	0.085	3.14E-01
Number PPIs	-0.095	9.66E-02
Distance to VIPs	-0.344	2.18E-05
Gene expression	-0.113	3.31E-01
Gene expression in immune cells	0.275	1.14E-02
Gene expression in testis	-0.087	2.68E-01
Regulatory density (ChIP-seq)	-0.633	7.22E-04
Regulatory density in immune cells (ChIP-seq)	0.217	2.17E-01
Regulatory density in testis (ChIP-seq)	-0.757	1.09E-11
Regulatory density (DNaseI)	0.229	3.09E-01
Gene length	-0.094	1.46E-01
Gene number	-0.265	7.58E-02
Coding density	-0.01	9.49E-01
Number iHS data points	-0.01	8.37E-01

PPIs: Protein-Protein Interactions. VIPs: Virus Interacting Proteins. ChIP-seq: Density of regulatory elements according to chromatin immunoprecipitation experiments. DNaseI: Density of regulatory elements according to the hypersensitivity to DNaseI. iHS: Integrated Haplotype Score.

Although some factors remain significant, in other cases there is a decrease of significance (Table 3). For instance, gene expression in immune genes shows a similar slope (slope=0.27 vs. 0.26), but decreased significance (p-value=1.14E-02 vs. 1.51E-05). The association with selection of the density of conserved elements decreases even more,

completely losing significance (slope=0.085 vs. 0.16, p-value=3.14E-01 vs. 6.4E-04). This could be caused by the smaller sample of genes included, as focusing only on low-recombination genes could restricts the range of values taken by these factors. This is supported by the fact that immune gene expression and especially the density of conserved elements are correlated with recombination (Spearman's ρ=-0.044 and -0.17; p-value = 3.94E-08 and <1E-16; correlation calculated across 1000 kb windows).

Fine scale patterns of recombination around regulatory elements

We further analyze the influence of local patterns of recombination around regulatory elements in our results. Given that recombination increases closer to the regulatory elements (Supplementary Figures 1, 2, 3), it is likely that whole window estimates of recombination are not an accurate measure of its effect on the association between iHS and regulatory element densities. To address this limitation, we use three variables related to the recombination around regulatory elements in immune cells, testis and across multiple tissues, respectively. We calculate the more local, average recombination around regulatory elements within each gene window, up to a maximum distance of 5kb from each side of regulatory elements. Each variable is included in the original model of Yoruba 1000 kb separately, and replaces the previous, window-wide average recombination. Given that some gene windows can overlap with more regulatory elements than others, we also consider the number of recombination data points obtained for regulatory elements inside each gene window. In this way, we account for the fact that recombination may vary more broadly in windows that have less regulatory elements. See Supplementary Data 4 for further details about these calculations.

Recombination sharply increases around regulatory elements in immune cells while iHS sharply decreases, with a steeper slope compared to other tissues (Supplementary Figures 1, 2, 3), which might have confounded the expected positive association with selection in the original regression model with window-wide recombination rates. The control for local patterns of recombination around immune regulatory elements should increase our power to detect this association. In line with this

prediction, we find that the association between immune regulatory density and iHS becomes positive and highly significant when considering the local patterns of recombination around immune regulatory elements (slope=0.68 vs. 0.047, p-value=4.60E-14 vs. 6.24E-01; Supplementary Data 4). In order to assess whether this increase is caused just by the lack of consideration of recombination patterns at a window-wide scale, we repeat the analyses but including the original window-wide recombination variable. After including recombination rate at window-wide scale, the positive association remains stronger compared to the original model (slope=0.28 vs. 0.047, p-value=5.72E-03 vs. 6.24E-01).

As explained above, we hypothesize that the confounding effect of recombination is caused by a mismatch between different scales. Gene windows with low recombination can have local peaks of recombination around transcription factor binding sites (Supplementary Figures 1, 2, 3), as these peaks are more accessible to transcription factors (Spruce et al. 2020). Therefore, recombination could erode signals of positive selection around regulatory elements even if the region has an overall low recombination. In contrast, regions with high recombination at the window-wide level should suffer less from this confounding effect, as recombination rate is then high both at a local and window-wide scale. Therefore, focusing on high-recombination regions should improve our ability to detect the expected positive association between selection and regulatory density. This approach, however, has the caveat of reducing the power to detect positive selection due the higher probability that recombination breaks selected haplotypes. We nevertheless repeat the analyses focusing on gene windows with a recombination rate equal or higher than the second tertile (1.552 cM/Mb). Focusing on high-recombination regions, regulatory density in immune cells shows a stronger association with selection in a model including recombination rate around both gene windows and immune regulatory elements compared to the same model run across all genes (slope=0.85 vs, 0.28, p-value=1.27E-08 vs. 5.72E-03). Interestingly, regulatory density in testis also shows a positive association with selection in a model including recombination around both testis regulatory elements and gene windows and focused on high-recombination regions. Note that this regulatory variable shows one of the strongest negative associations with selection in the original model (slope=0.58 vs. -0.79, p-value=4.98E-05 vs. <1E-16), an

unexpected result given the role of regulatory sequences in positive selection (Enard et al. 2014). These results suggest that, indeed, a higher regulatory density can increase the probability for positive selection to occur, especially in some tissues. However, this signal is distorted by fine-scale patterns of recombination, which have to be carefully taken into account.

Discussion

We find several functional factors independently associated with recent positive selection. For instance, viral interactions along with gene expression and regulatory density in immune cells are strongly and positively associated with selection. Viruses have acted as key drivers of adaptation in humans and VIPs show higher expression in lymphocytes compared to non-VIPs (Halehalli and Nagarajaram 2015; Enard et al. 2016; Enard and Petrov 2018). These results confirm the existence of strong viral and immune selective pressures during recent human evolution (Enard and Petrov 2020; Souilmi et al. 2021; Enard and Petrov 2018). We also detect a positive association between testis regulatory density and selection, which is, congruent with the fact that reproduction-related functions show signals of positive selection (Nielsen et al. 2005; Voight et al. 2006). Note, however, that this latter result has been more difficult to reveal in our analyses, as it is only visible after controlling for fine-scale patterns of recombination around testis regulatory elements and after focusing on high-recombination regions, thus it is more questionable. Finally, the density of conserved elements and GC-content, which are related to overall (coding and non-coding) functional density (Lander et al. 2001; Di Filippo and Bernardi 2008; Siepel et al. 2005), also show a positive association with selection. In summary, recent and positive selection is associated with the distribution of functional regions across the human genome.

If sweep signals are false positives only reflecting genetic drift, the signatures of sweeps should occur randomly across the genome. In contrast, if selective sweeps were relatively common during recent human evolution, we expect that other factors on top of recombination should associate with sweep signals across the human genome. In that

case, we expect that selective sweeps should occur more frequently around functional (coding or non-coding) elements where adaptive mutations are expected to take place (Enard et al. 2014), also around genomic regions associated with specific functions. We find strong evidence for the latter scenario, as multiple functional factors are associated with recent positive selection. In other words, the distribution of sweep signals is not random, and is associated with the functional characteristics of the human genome. Therefore, the present study provides evidence to further clarify the controversy about the frequency of positive selection during recent human evolution (Enard et al. 2014; Hernandez et al. 2011). Our results support that selective sweeps were at least common enough during recent human evolution to generate associations between selection and functional factors.

Our MDR approach shows a good fit to the distribution of iHS across the human genome, with its two components (selection and genetic drift) showing a clear differentiation. This suggests that the consideration of a more biologically meaningful model, which assumes two Gaussian distributions instead of one, works better to analyze recent adaptation than simple correlations. The associations found using this approach are more informative than classic correlations as this model represents more directly genome evolution and considers multiple factors simultaneously. These characteristics make it possible to reveal expected, but also unexpected patterns. For instance, GC-content is strongly and positively associated with iHS. Regions with high GC-content also tend to have high long-term recombination rates (Duret and Arndt 2008), which would predict a negative association between iHS and GC content, not positive. A possible explanation for this pattern is that the MDR approach can control for recombination as a confounder by simultaneously analyzing recombination rate and other genomic factors. This can unravel new patterns, like a positive association between GC-content and sweep signals, which may be independent from the eroding effect of recombination. This is supported by the fact that GC-content only shows a negative association with selection after removing recombination from the model. GC-content could be a better proxy of overall functional density than individual functional factors (Lander et al. 2001; Di Filippo and Bernardi 2008), explaining its positive association with selection. Results from MDRs support this hypothesis as the removal of functional factors like coding density, gene number or

regulatory density strongly increases the significance of this association. This could also explain another surprising result, the lack of association between iHS and coding regions, which are expected targets of selection. The removal of GC-content and the density of conserved elements, which is another factor associated with overall functional density (Siepel et al. 2005), unmasks a significant positive association between coding density and selection. These results suggest that considering measures of overall functional density (i.e., GC-content and conserved elements density) could absorb the expected positive association with selection for individual functional factors like coding density. These results illustrate the utility of our approach. The simultaneous consideration of multiple predictors lets us unravel novel associations between genomic factors and adaptation. In addition, it enables easier testing of different hypotheses by modifying the set of predictors considered. Therefore, the MDR approach is useful to model recent genomic adaptation, and to analyze its prevalence and determinants.

We find other unexpected results using the MDR approach. For instance, we do not find a positive association between selection and testis regulatory density in the original model. This is an unexpected result given that a higher regulatory density would lead to a larger mutational target for adaptation through changes in expression. Given that genomic regions with local peaks of recombination are usually more accessible, transcription factors are more likely to bind DNA in these regions (Spruce et al. 2020). Therefore, any adaptative signal associated with regulatory density would be likely eroded by recombination (Sabeti et al. 2006; O'Reilly et al. 2008; Booker et al. 2020). In support of this hypothesis, we find higher recombination and lower iHS closer to regulatory elements. Indeed, when fine-scale patterns of recombination are considered as new predictors, regulatory density in immune cells and testis become strongly and positively associated with selection. As previously mentioned, the positive association of testis regulatory density only emerges after focusing on high-recombination regions, thus it should be taken with more caution. These results support the role of regulatory sequences as targets of positive selection independently of other genomic features (Enard et al. 2014), along with the relevance of testis and specially immune cells in recent human evolution (Nielsen et al. 2005; Voight et al. 2006). Our results also illustrate how the detection of selection can be hindered by fine-scale characteristics of the genome, like

the sharp increase of recombination we observe around regulatory elements. This might mask selection signals around regulatory elements, but not only because of the breaking of haplotypes caused by recombination. An additional mechanism might be the existence of biased gene conversion in these recombinant regions, which could specifically favor GC mutations (gBGC) (Galtier et al. 2001; Duret and Arndt 2008; Dutta et al. 2018). This could lead to an increase in the frequency of GC mutations that are associated with short haplotypes (due to the eroding effect of recombination). BGC might therefore make it even more difficult to detect recent positive selection around regulatory regions, where it is expected to happen (Enard et al. 2014). Note, however, that the present study does not provide direct evidence about the implication of gBGC in the observed patterns. It can be difficult to disentangle the eroding effect of recombination from GC-biased gene conversion on haplotypes because gBGC is caused and therefore very correlated with recombination. Distinguishing the specific effect of biased-gene conversion may require precise recombination maps for African populations, where we find the clearest selection signals. Although we have not found evidence about the implication of gBGC yet, the fact that this hypothesis has emerged from implementing our model illustrates its utility to propose new hypotheses about the different processes that influence the detection and occurrence of positive selection.

Although the MDR approach provides an excellent fit to the observed selection statistic, we may still not detect associations with specific factors because they do not have a simple linear relationship with recent selection. Although more biologically meaningful than a model assuming just one distribution, our approach is still linear, thus more complex relationships may not be detected. For example, the unexpected lack of significance for coding density could be explained not only by the consideration of better proxies of overall functional density, but also by the existence of non-linear relationships with selection. Advantageous mutations are more likely to appear in regions with high coding density, thus a high density of coding sequences should favour the appearance of selection (Enard et al. 2014). However, high coding density can also mean a higher frequency of deleterious mutations. These mutations can interfere with adaptation when they are in linkage with advantageous mutations. This would be especially relevant for genomic regions with low recombination, where the linkage between deleterious and

advantageous mutations is more likely (Di et al. 2021). Therefore, a higher coding density could favor or hinder positive selection depending on the circumstances, complicating the detection of an association between coding density and selection with our approach. The existence of this genetic interference could also explain the different results observed across populations in relation to coding density. Despite having the largest selected component, the Yoruba population does not show a clear link between selection and coding density. This association becomes visible only after removing proxies of overall functional density (i.e., density of conserved elements and GC-content). In contrast, this pattern is more readily visible in the other populations studied despite showing smaller selected components. A potential, but still speculative explanation at this point could be the fact that bottlenecks related to the migration out of Africa may have removed segregating recessive deleterious variants, which may reduce their interference over advantageous mutations, thus making it easier for positive selection to act on coding regions (Di et al. 2021). In contrast, there were no similar bottlenecks to decrease the genetic interference in the Yoruba, and regions of the genome with high a coding density may still experience stronger interference from recessive deleterious mutations. Interestingly, despite this potential limitation of positive selection, we still see a larger selective component in the entire Yoruba genome, suggesting that selection is either more abundant and/or more visible in this population. Note, however, this is a potential explanation that is speculative at this point and will require further investigation.

Conclusions

We found multiple functional factors highly associated with recent selection using MDRs with two distributions meant to reflect the selected and drift components of the human genome. Our results suggest that positive selection was relatively common in recent human evolution and highlight MDRs as a novel approach to quantify the prevalence of recent adaptation and its determinants in the human and other genomes.

Methods

Window sizes and genes coordinates

All analyses were performed using hg19 genomic coordinates and protein-coding gene annotations from Ensembl v99 (Yates et al. 2020). We considered windows of different sizes (50, 100, 200, 500 and 1,000 kb). Each window was centered at the genomic center of a gene, in the middle between the most upstream transcription start site and the most downstream transcription end site. We used a fixed window size to avoid biases related to gene length, as larger genes are more likely to overlap with high local iHS values just by chance compared to shorter genes. This would bias power in favour of larger genes, increasing the probability to detect sweeps in them just because of their length. In addition, the consideration of different window sizes provides the opportunity to detect different types of sweeps as larger windows are specifically sensitive to strong selective sweeps compare to smaller windows (Enard and Petrov 2020).

Genomic features

Multiple genomic factors were calculated inside the gene windows. We considered the following factors that are likely to influence the frequency of sweeps:

- Length of the gene at the center of each window.
- Number of genes overlapped with each window. We used this value as an estimate of gene density.
- Recombination rate. For each window, we calculated the genetic distance between the edges of the window and then divided it by the physical distance between them. Genetic position of each window edge was obtained from the deCODE 2019 genetic map (Halldorsson et al. 2019). In case no data was available for a window edge, we selected the closest genetic position data points within 50 kb at each side of the window edge to estimate its genetic position. We used linear interpolation to this end, considering the genetic and physical position of the two points around the

window edge. In other words, we assumed that genetic distance increased linearly between the two selected data points, and hence we can estimate the genetic position of the window edge based on its physical position respect to the other two positions. If no genetic position data was available at 50 kb or closer to the window edge, no genetic position was calculated and the whole window was discarded for recombination calculations and subsequent analyses. Note that we searched for genetic position data points only up to 50 kb from each window edge to avoid that the total distance between points at each edge could be too large (specially for larger windows). In that scenario, linear interpolation could be inappropriate because we would use very distant data to calculate recombination rate. For example, if a 1,000 kb window had no genetic position for any of its edges and we looked for genetic positions up to 1,000 kb at each side, it would possible that the selected data points of genetic position downstream and upstream of the window were separated by 3,000 kb. In addition, the recombination rate would not correspond with the size of the window.

- Density of coding sequences. We used Ensembl v99 coding sequences. The density was calculated as the proportion of coding bases respect to the whole length of the window without considering gaps (gap locations obtained from the UCSC Genome Browser; https://genome.ucsc.edu/; http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/). A similar approach was used for the rest of density estimates.

- Density of mammalian phastCons conserved elements (Siepel et al. 2005), also downloaded from the UCSC Genome Browser. Given that each conserved segment had a score, we considered as conserved only those segments above a given threshold. In order to minimize the inclusion of non-conserved elements, we used a threshold that only considers 10% of the genome as conserved, as it is unlikely that more than 10% of the whole genome is constrained (Siepel et al. 2005).

- Density of regulatory elements. We calculated several variables related to the density of binding sites for transcription factors. In all cases, we calculated the proportion of sequences that are considered binding sites within a window. The data

were also obtained from the UCSC Genome Browser (The ENCODE Project Consortium 2012).

  - Density of DNaseI hypersensitive sites. We considered as binding sites those segments with a score higher than a given threshold. The selected threshold was a score value according to which only 10% of the genome is considered accessible to DNaseI, and hence a binding site for transcription factors. In this way, we minimize the probability to consider non-binding sites (i.e., false positives).

  - Density of binding sites according to the technique of chromatin immunoprecipitation followed by sequencing (ChIP-seq). We calculated a threshold to consider a given segment as a binding site following the same approach used for DNaseI hypersensitive sites. This database includes 1264 experiments representing 338 transcription factors in 130 cell types. We calculated the density of binding sites using the whole dataset, but also considering two subsets, one for experiments performed on cells lines of testis, and another for experiments performed on immune cells (lymphocytes in most cases).

- GC-content, calculated as a percentage per window. It was also obtained from the UCSC Genome Browser.

- Gene expression: We used the log (base 2) of TPM (Transcripts Per Million), which was obtained from GTEx (GTEx Consortium 2015) (https://www.gtexportal.org/home/). We considered the average gene expression across 53 GTEx v7 tissues, but also expression in immune cells (lymphocytes) and testis.

- Number of protein-protein interactions (PPIs) in the human protein interaction network (Luisi et al. 2015). We used the log (base 2) of the number of PPIs of each gene. We summed 1 to the original dataset to avoid problems applying the logarithm to genes with no interactions (i.e., PPI=0).

- Proportion of genes that interact with viruses (VIPs). We used a previously published dataset (Enard and Petrov 2020) that includes around 4500 VIPs with evidence of physical interaction with viruses. 1920 of these VIPs were manually curated from the virology literature, while the remaining 2600 VIPs were identified using high-throughput methods and retrieved from the VirHostnet 2.0 database and additional studies. See the original publication for further details (Enard and Petrov 2020). The variable included in the model was calculated as the distance of each gene to the closest gene that interacts with viruses.

iHS calculation

In order to calculate iHS, we used polymorphism data from the 1,000 Genomes Project phase 3 (The 1000 Genomes Project Consortium 2015). We calculated iHS for 5 populations: Yoruba, Utah residents with Northern and Western European ancestry, Toscani, Han Chinese and Peruvians. We used the hapbin software (Maclean et al. 2015) in order to perform fast scans of iHS across the genomes of these populations. Note that the genetic maps used as input in hapbin were calculated using the deCODE 2019 genetic map (Halldorsson et al. 2019). In case that no data was available for the focal variant, we again used linear interpolation to estimate its genetic position. We followed the same approach used to calculate the genetic position of window edges during recombination calculations. In this case, we searched for genetic position data points at both sides of the focal variant until 1,000 kb. We discarded variants with no data points around, as they were present in genomic regions with low density of recombination data. The genetic maps based on the deCODE 2019 map were used as an input in order to calculate the iHS for the 5 populations.

The raw iHS values obtained with hapbin were standardized by frequency. For each population, we divided variants in 50 bins of frequency (from 0 to 1) considering the whole genome. We then calculated the mean and standard deviation of raw iHS within each bin, using them to standardize raw iHS values: (raw iHS - mean iHS) / sd iHS. The mean of standardized iHS was calculated per window, considering the absolute value of

iHS. In other words, we considered as high selection both very positive and very negative iHS values, which indicates selection in favor of the haplotypes carrying the derived and the ancestral allele, respectively. We also obtained the number of iHS values per window to have an estimate of data density. Mean iHS in windows with low number of iHS values can be more influenced by outliers (very low or very high iHS values). Therefore, we included the number of iHS values per window as predictor in the models to control for this.

Modeling

We modeled the association between genomic factors and iHS in each population and window size by combining a Gaussian mixture distribution with linear regression. Mixture distributions are useful to model data that cannot be fully described by a single distribution and that are likely generated by multiple processes. In this case, a mixture of two Gaussian distributions could fit better a scenario where selection influences some genomic regions but not others, and hence a bimodal distribution of iHS would be observed across the genome. Therefore, we posit that sweep signals in the human genome may have a selected and a non-selected component. According to this, we assumed that observed log iHS followed a mixture of two Gaussian distributions representing the two selection components. Because selective sweeps often result in elevated iHS, we assumed that the Gaussian distribution for the selected component had a higher mean than that for the non-selected component. Furthermore, to model the effects of genomic factors on iHS, we assumed that the probability of the selected component was a linear combination of genomic factors followed by a sigmoid transformation. Specifically, for each factor, we calculated the product of its observed value with a slope indicating its association with iHS. We summed the product over all factors for a given gene to model the accumulative effect of genomic factors on the distribution of iHS. Then, we added an intercept indicating the baseline level of selection when all genomic factors are equal to 0 and transformed the result to a probability (0 to 1) using a sigmoid function. This probability was obtained for each gene and can be regarded as the overall influence of all genomic factors on the frequency of recent selective sweeps. These steps are described by

the equation $p = \sigma(\sum b_i * x_i + a) = \frac{1}{1+\exp(-\sum b_i * x_i - a)}$, where p is a probability summarizing the influence of all studied genomic factors on iHS, σ is the sigmoid function, b_i and x_i are the slope and the observed value for genomic factor *i* in a given gene, respectively, and a is the intercept.

To obtain the likelihood of log iHS in each gene, we calculated a weighted sum of Gaussian density functions, $P(Y) = p * f(Y|\mu_1, \sigma_1) + (1-p) * f(Y|\mu_0, \sigma_0)$, where Y and $P(Y)$ are the observed log iHS and its likelihood in a given gene, respectively, p is the probability of the selected component calculated in the previous step, $f(Y|\mu_1, \sigma_1)$ is the Gaussian density function with mean μ_1 and standard deviation σ_1 describing the selected component, and $f(Y|\mu_0, \sigma_0)$ is the Gaussian density function with mean μ_0 and standard deviation σ_0 describing the non-selected component. Finally, we summed the logarithm of $P(Y)$ over all genes to obtain the log likelihood of the entire dataset.

We estimated model parameters by maximizing the log likelihood of the entire dataset. These parameters included the mean and standard deviation of each Gaussian distribution, along with the intercept and the slopes representing the association between iHS and each genomic factor. For each genomic factor, we tested the significance of its slope by comparing the log likelihood between the full model with all genomic factors and a nested null model without the selected factor. We assumed that two times the difference in log likelihood between the two models followed a chi-square distribution with one degree of freedom.

Data access

Genetic polymorphism data was obtained from the 1,000 Genomes Project (The 1000 Genomes Project Consortium 2015). Data for the genomic factors were mostly obtained from the UCSC genome browser (https://genome.ucsc.edu/), along with other public sources. See Methods section for specific details about the source of each dataset.

Competing interest statement

The authors declare no competing interests.

Acknowledgments

We thank the members of the Enard, Masel and Gutenkunst labs for their helpful comments. DFST was supported by a Marie S. Curie Global Fellowship within the European Union research and innovation framework programme (2014-2020; ClimAHealth: 101030971). Y-FH was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number R35GM142560 and by startup funds from Pennsylvania State University. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the European Union.

Author contributions

Y-FH, DE and DFST designed the study. DE and DFST collected the data. Y-FH and DFST performed the analyses. DE and DFST wrote the paper and interpreted the results. Y-FH, DE and DFST edited and approved the paper.

References

Barreiro LB, Ben-Ali M, Quach H, Laval G, Patin E, Pickrell JK, Bouchier C, Tichit M, Neyrolles O, Gicquel B, et al. 2009. Evolutionary dynamics of human toll-like receptors and their different contributions to host defense. *PLoS Genet* **5**.

Booker TR, Yeaman S, Whitlock MC. 2020. Variation in recombination rate affects detection of outliers in genome scans under neutrality. *Mol Ecol* **29**: 4274–4279. https://doi.org/10.1111/mec.15501.

Deschamps M, Laval G, Fagny M, Itan Y, Abel L, Casanova JL, Patin E, Quintana-Murci L. 2016. Genomic Signatures of Selective Pressures and Introgression from Archaic Hominins at Human Innate Immunity Genes. *Am J Hum Genet* **98**: 5–21. http://dx.doi.org/10.1016/j.ajhg.2015.11.014.

Di C, Murga Moreno J, Salazar-Tortosa DF, Lauterbur ME, Enard D. 2021. Decreased adaptation at human disease genes as a possible consequence of interference between advantageous and deleterious variants. *Elife* **10**: e69026. https://doi.org/10.7554/eLife.69026.

Di Filippo M, Bernardi G. 2008. Mapping DNase-I hypersensitive sites on human isochores. *Gene* **419**: 62–65.

Duret L, Arndt PF. 2008. The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genet* **4**: e1000071. https://doi.org/10.1371/journal.pgen.1000071.

Dutta R, Saha-Mandal A, Cheng X, Qiu S, Serpen J, Fedorova L, Fedorov A. 2018. 1000 human genomes carry widespread signatures of GC biased gene conversion. *BMC Genomics* **19**: 256.

Enard D, Cai L, Gwennap C, Petrov DA. 2016. Viruses are a dominant driver of protein adaptation in mammals. *Elife* **5**: e12469.

Enard D, Messer PW, Petrov DA. 2014. Genome-wide signals of positive selection in human evolution. *Genome Res* **24**: 885–895. https://doi.org/10.1101/gr.164822.113.

Enard D, Petrov DA. 2020. Ancient RNA virus epidemics through the lens of recent adaptation in human genomes. *Philos Trans R Soc London Ser B* **375**: 20190575. http://dx.doi.org/10.1098/rstb.2019.0575.

Enard D, Petrov DA. 2018. Evidence that RNA Viruses Drove Adaptive Introgression between Neanderthals and Modern Humans. *Cell* **175**: 360–371. https://doi.org/10.1016/j.cell.2018.08.034.

Ferrer-Admetlla A, Liang M, Korneliussen T, Nielsen R. 2014. On detecting incomplete

soft or hard selective sweeps using haplotype structure. *Mol Biol Evol* **31**: 1275–1291. https://doi.org/10.1093/molbev/msu077.

Galtier N, Piganeau G, Mouchiroud D, Duret L. 2001. GC-Content Evolution in Mammalian Genomes: The Biased Gene Conversion Hypothesis. *Genetics* **159**: 907–911. https://doi.org/10.1093/genetics/159.2.907.

Garud NR, Messer PW, Buzbas EO, Petrov DA. 2015. Recent Selective Sweeps in North American Drosophila melanogaster Show Signatures of Soft Sweeps. *PLoS Genet* **11**: e1005004. https://doi.org/10.1093/molbev/msu077.

GTEx Consortium. 2015. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans eds. K.G. Ardlie, D.S. Deluca, A. V Segrè, T.J. Sullivan, T.R. Young, E.T. Gelfand, C.A. Trowbridge, J.B. Maller, T. Tukiainen, M. Lek, et al. *Science (80-)* **348**: 648–660. https://doi.org/10.1126/science.1262110.

Halehalli R, Nagarajaram HA. 2015. Molecular principles of human virus protein-protein interactions. *Bioinformatics* **31**: 1025–1033.

Halldorsson B V., Palsson G, Stefansson OA, Jonsson H, Hardarson MT, Eggertsson HP, Gunnarsson B, Oddsson A, Halldorsson GH, Zink F, et al. 2019. Characterizing mutagenic effects of recombination through a sequence-level genetic map. *Science (80-)* **363**. https://doi.org/10.1126/science.aau1043.

Harris DN, Song W, Shetty AC, Levano KS, Cáceres O, Padilla C, Borda V, Tarazona D, Trujillo O, Sanchez C, et al. 2018. Evolutionary genomic dynamics of Peruvians before, during, and after the Inca Empire. *Proc Natl Acad Sci U S A* **115**: E6526–E6535.

Henn BM, Cavalli-Sforza LL, Feldman MW. 2012. The great human expansion. *Proc Natl Acad Sci U S A* **109**: 17758–17764. https://doi.org/10.1073/pnas.1212380109.

Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A, McVean G, Sella G, Przeworski M. 2011. Classic selective sweeps were rare in recent human evolution.

Science (80-) **331**: 920–924. https://doi.org/10.1126/science.1198878.

Huang Y-F. 2021. Dissecting genomic determinants of positive selection with an evolution-guided regression model. *Mol Biol Evol* msab291. ttps://doi.org/10.1093/molbev/msab291.

Johnson KE, Voight BF. 2018. Patterns of shared signatures of recent positive selection across human populations. *Nat Ecol Evol* **2**: 713–720. http://dx.doi.org/10.1038/s41559-018-0478-6.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, Fitzhugh W, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921.

Lohmueller KE, Albrechtsen A, Li Y, Kim SY, Korneliussen T, Vinckenbosch N, Tian G, Huerta-Sanchez E, Feder AF, Grarup N, et al. 2011. Natural selection affects multiple aspects of genetic variation at putatively neutral sites across the human genome. *PLoS Genet* **7**: e1002326. https:doi.org/10.1371/journal.pgen.1002326.

Luisi P, Alvarez-Ponce D, Pybus M, Fares MA, Bertranpetit J, Laayouni H. 2015. Recent positive selection has acted on genes encoding proteins with more interactions within the whole human interactome. *Genome Biol Evol* **7**: 1141–1154. https://doi.org/10.1093/gbe/evv055.

Maclean CA, Chue Hong NP, Prendergast JGD. 2015. hapbin: An efficient program for performing haplotype-based scans for positive selection in large genomic datasets. *Mol Biol Evol* **32**: 3027–3029. https://doi.org/10.1093/molbev/msv172.

Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB, Hubisz MJ, Fledel-Alon A, Tanenbaum DM, Civello D, White TJ, et al. 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol* **3**: e170.

O'Reilly PF, Birney E, Balding DJ. 2008. Confounding between recombination and selection, and the Ped/Pop method for detecting selection. *Genome Res* **18**: 1304–1313. https://doi.org/10.1101/gr.067181.107.

Pritchard JK, Pickrell JK, Coop G. 2010. The Genetics of Human Adaptation: Hard Sweeps, Soft Sweeps, and Polygenic Adaptation. *Curr Biol* **20**: R208–R215. http://doi.org/10.1016/j.cub.2009.11.055.

Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, Palma A, Mikkelsen TS, Altshuler D, Lander ES. 2006. Positive natural selection in the human lineage. *Science (80-)* **312**: 1614–1620. https://science.sciencemag.org/content/312/5780/1614.

Schrider DR. 2020. Background selection does not mimic the patterns of genetic diversity produced by selective sweeps. *Genetics* **216**: 499–519.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LDW, Richards S, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**: 1034–1050. https://doi.org/10.1101/gr.3715005.

Souilmi Y, Lauterbur ME, Tobler R, Huber CD, Johar AS, Varasteh Moradi S, Johnston WA, Krogan NJ, Alexandrov K, Enard D. 2021. An ancient viral epidemic involving host coronavirus interacting genes more than 20,000 years ago in East Asia. *Curr Biol* **31**: 1–11. https://doi.org/10.1016/j.cub.2021.05.067.

Spruce C, Dlamini S, Ananda G, Bronkema N, Tian H, Paigen K, Carter GW, Baker CL. 2020. HELLS and PRDM9 form a pioneer complex to open chromatin at meiotic recombination hot spots. *Genes Dev* **34**: 398–412. https://doi.org/10.1101/gad.333542.119.

The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. *Nature* **526**: 68–74. https://doi.org/10.1038/nature15393.

The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57–74. https://doi.org/10.1038/nature11247.

The International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature* **437**: 1299–1320. https://doi.org/10.1038/nature04226.

Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. *PLoS Biol* **4**: 0446–0458.

Yates AD, Achuthan P, Akanni W, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, Azov AG, Bennett R, et al. 2020. Ensembl 2020. *Nucleic Acids Res* **48**: D682–D688. https://doi.org/10.1093/nar/gkz966.