Original Manuscript

Signatures of long-term balancing selection in human genomes

Bárbara Domingues Bitarello[1,2,*]

Cesare de Filippo[2]

João Carlos Teixeira[2,3]

Joshua M. Schmidt[2]

Philip Kleinert[2,4]

Diogo Meyer[1, ¶,*]

Aida M. Andrés[2, ¶,*]

1, Department of Genetics and Evolutionary Biology, University of São Paulo, São Paulo, Brazil

2, Genetics Department, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

3, Current address: Unit of Human Evolutionary Genetics, Institut Pasteur, Paris, France

4, Current address: Computational Molecular Biology Department, Max Planck Institute for Molecular Genetics, Berlin, Germany

¶, AA and DM co-supervised the study

*Corresponding author(s): barbara_domingues@eva.mpg.de (BDB), diogo@ib.usp.br (DM), aida_andres@eva.mpg.de (AMA).

Abstract

Balancing selection maintains advantageous diversity in populations through different mechanisms. While extensively explored from a theoretical perspective, an empirical understanding of its prevalence and targets lags behind our knowledge of positive selection. Here we describe a simple yet powerful statistic to detect signatures of long-term balancing selection (LTBS) based on the expectation that some types of LTBS result in an accumulation of polymorphic sites at moderate-to-intermediate frequencies. The *Non-Central Deviation* (*NCD*) quantifies the degree to which SNP frequencies within a window of a pre-defined size depart from deterministic expectations under balancing selection. The statistic can be implemented considering only polymorphisms (*NCD1*) or also including also information on fixed differences (*NCD2*), and can detect LTBS under different frequencies of the balanced allele(s). Because of its simplicity, NCD can be applied to single loci or genomic data, and to populations with or without known demographic history. We show that, in humans, *NCD1* and *NCD2* have high power to detect long-term balancing selection, with *NCD2* outperforming all existing methods. We applied *NCD2* to genome-wide data from African and European human populations, and found that 0.6% of the analyzed windows show signatures of LTBS, corresponding to 0.8% of the base pairs and 1.6% of the SNPs in the analyzed genome. This suggests that albeit not prevalent, LTBS affects the evolution of a sizable portion of the genome (it overlapping ~8% of protein-coding genes). These SNPs disproportionally overlap sites with protein-coding and amino-acid altering functions, but not putatively regulatory sites. Our catalog of candidates includes known targets of LTBS, but a majority of them have not been previously identified. As expected, immune-related genes are among those with the strongest signatures, although most candidates are involved in other biological functions, suggesting that balancing selection potentially influences diverse human phenotypes.

Author Summary

With the availability of whole-genome sequences on a population level, genetic variation in humans has been queried for signatures of natural selection. Most of these efforts have focused on positive selection, which results in novel adaptions. Balancing selection, an important form of natural selection that maintains advantageous genetic variants within populations, sometimes for millions of years, has attracted less attention. This is despite the important effects that variants under balancing selection have in phenotypic diversity and susceptibility to disease, as shown by the most eminent target of balancing selection: the Major Histocompatibility Complex Locus (MHC, known as HLA in humans). We developed a statistic that identifies regions of the genome with signatures that are expected under balancing selection. This statistic has very high power to detect long-term

balancing selection in humans, and it is simple enough to be used in a wide variety of species, having the potential to improve our understanding of balancing selection across taxonomic groups. When applied to human data, we find that long-term balancing selection has affected genomic regions that define the sequence of protein-coding genes more often than their regulation, and has targeted genes involved in immunity and a diversity of additional biological functions.

Introduction

Balancing selection refers to a class of selective mechanisms that maintains advantageous genetic diversity in populations. Decades of research have established HLA genes as a prime example of balancing selection [1,2], with thousands of alleles segregating in humans [3], extensive support for functional effects of these polymorphisms (e.g. [4,5]), and various well-documented cases of association between selected alleles and disease susceptibility (e.g. [6,7]). The catalog of well-understood non-HLA targets of balancing selection in humans remains small, but includes genes associated to phenotypes such as auto-immune diseases [8,9], resistance to malaria [10], HIV infection [11] or susceptibility to polycystic ovary syndrome [12]. Thus, besides historically influencing individual fitness, balanced polymorphisms shape current phenotypic diversity and susceptibility to disease.

Balancing selection encompasses several mechanisms (reviewed in [13–15]). These include heterozygote advantage (or overdominance), frequency-dependent selection [16,17], selective pressures that fluctuate in time [18,19] or in space in panmitic populations [20,21], and some cases of pleiotropy [22]. For overdominance, pleiotropy, and some instances of spatially variable selection, a stable equilibrium can be reached [16]. For other mechanisms, the frequency of the selected allele can change in time without reaching a stable equilibrium. Regardless of the mechanism, long-term balancing selection (LTBS) has the potential to leave identifiable signatures in genomic data. These include a local site-frequency spectrum (SFS) with an excess of alleles at intermediate frequencies and, when selection is old enough, an excess of polymorphisms relative to substitutions (reviewed in [15]). In some cases, very ancient balancing selection can maintain trans-

species polymorphisms in sister species [23,24], while transient heterozygote advantage and other types of recent balancing selection [25] will result in signatures difficult to distinguish from incomplete, recent selective sweeps [15].

While balancing selection has been extensively explored from a theoretical perspective, an empirical understanding of its prevalence lags behind our knowledge of positive selection. This stems from technical difficulties in detecting balancing selection, as well as the perception that it may be a rare selective process [26]. In fact, few methods have been developed to identify its targets, and only a handful of studies have sought to uncover them genome-wide in humans [23,24,27–32]. Different approaches have been used to identify genes [28] or genomic regions [31] with an excess of polymorphisms and intermediate frequency alleles, while other studies have identified trans-species polymorphisms between humans and their closest living relatives (chimpanzees and bonobos) [23,24]. Overall, these studies suggested that balancing selection may act on a small portion of the genome, although the limited extent of data available (e.g., exome data [31], small sample size [28]), and stringency of the criteria (e.g., balanced polymorphisms predating human-chimpanzee divergence [23,24]) may underlie the paucity of detected regions.

Here, we developed two statistics that summarize, directly and in a simple way, the degree to which allele frequencies of SNPs in a genomic region deviate from those expected under balancing selection. We then use these statistics to test the null hypothesis of neutral evolution. We showed, through simulations, that one of our statistics outperforms existing methods under a realistic demographic scenario for human populations. We applied this statistic to genome-wide data from four human populations and used both outlier and simulation-based approaches to identify genomic regions bearing signatures of LTBS.

Results

The Non-Central Deviation (NCD) statistic

Background. Owing to linkage, the signature of long-term balancing selection extends to the genetic neighborhood of the selected variant(s); therefore, patterns of polymorphism and divergence in a genomic region can be used to infer whether it evolved under LTBS [13,21]. LTBS leaves two distinctive signatures in linked variation, when compared with neutral expectations. The first is an increase in the ratio of polymorphic to divergent sites: by reducing the probability of fixation of a variant, balancing selection increases the local time to the most recent common ancestor [33]. The HKA test is commonly used to detect this signature [34]. The second signature is an excess of alleles segregating at intermediate frequencies. In humans, the folded SFS – the frequency distribution of minor allele frequencies (MAF) — is typically L-shaped, showing an excess of low-frequency alleles when compared to expectations under neutrality and demographic equilibrium. The abundance of rare alleles is further increased by recent population expansions [35], purifying selection and recent selective sweeps [36]. Regions under LTBS, on the other hand, can show a markedly different SFS, with proportionally more alleles at intermediate frequency (Fig 1A-B). Such a deviation in the SFS is the signature identified by classical neutrality tests such as Tajima's *D* (Taj*D*) and newer statistics such as MWU-high [37].

With heterozygote advantage, the frequency equilibrium (f_{eq}) depends on the relative fitness of each genotype [16]: under symmetric overdominance, i.e. where the two types of homozygotes have the same fitness, $f_{eq} = 0.5$; under asymmetric overdominance, where the fitness of the two homozygotes is different, $f_{eq} \neq 0.5$ (S1 Note). Under frequency-dependent selection and fluctuating selection, while an equilibrium may not be reached (S1 Note), f_{eq} can be thought of as the frequency of the balanced polymorphism at the time of sampling.

NCD statistic. In the tradition of neutrality tests analyzing the SFS directly (e.g. [37–39]), we propose and define the statistic "Non-Central Deviation" (*NCD*) which measures the degree to which the local SFS deviates from a pre-specified allele frequency (the *target frequency*, *tf*) in a genomic region. Under a model of balancing selection, *tf* can be thought of as the expected

frequency of a balanced allele, with the NCD statistic quantifying how far the sampled SNP frequencies are from it. Because bi-allelic loci have complementary allele frequencies, and there is no prior expectation regarding whether ancestral or derived alleles should be maintained at higher frequency, we use the folded SFS (Fig 1B). *NCD* is defined as:

$$NCD(tf) = \sqrt{\frac{\sum_{i=1}^{n}(p_i - tf)^2}{n}} \quad (1)$$

where i is the i-th informative site in a locus, p_i is the MAF for the i-*th* informative site, n is the number of informative sites, and *tf* is the target frequency with respect to which the deviations of the observed alleles frequencies are computed. Thus, *NCD* is a type of standard deviation that quantifies the dispersion of allelic frequencies from *tf*, rather than from the mean of the distribution. Low *NCD* values reflect a low deviation of the SFS from a pre-defined *tf*, as expected under LTBS (Fig 1C and S1 Note).

We propose two *NCD* implementations. *NCD1* uses only on polymorphic sites as informative sites, and *NCD2* also includes the number of fixed differences (FDs) relative to an outgroup species (i.e, all informative sites, ISs = SNPs + FDs, are used to compute the statistic). In *NCD2*, FDs are considered ISs with MAF = 0; thus, the greater the number of FDs, the larger the *NCD2* and the weaker the support for LTBS. From equation 1 it follows that the maximum value for *NCD2*(*tf*) is the *tf* itself (for $tf \geq 0.25$, see S1 Note), which occurs when there are no SNPs and the number of FDs ≥ 1. The maximum *NCD1* value approaches – but never reaches – *tf* when all SNPs are singletons. The minimum value for both *NCD*1 and *NCD*2 is 0, when all SNPs segregate at *tf* and, in the case of *NCD2*, the number of FDs = 0 (S1 and S2 Figs).

Power of NCD to detect LTBS

We evaluated the sensitivity and specificity of *NCD1* and *NCD2* by benchmarking their performance using simulations. Specifically, we considered demographic scenarios inferred for African, European, and Asian human populations, and simulated sequences evolving both under neutrality and LTBS using an overdominance model. We explored the influence of parameters that can affect the power of NCD statistics: time since onset of balancing selection (*Tbs*), frequency equilibrium defined by selection coefficients (f_{eq}), demographic history of the sampled population, *tf* used in *NCD* calculation, length of the genomic region analyzed (*L*) and implementation (*NCD1* or *NCD2*). Box 1 summarizes nomenclature used throughout the text.

Box 1. List of Abbreviations

LTBS, long-term balancing selection.

MAF, minor allele frequency.

SFS, site-frequency spectrum.

FD, fixed differences (between ingroup and outgroup species).

IS, informative sites (polymorphic sites in the ingroup species plus fixed differences between ingroup and outgroup species).

f_{eq}, deterministic equilibrium frequency expected under balancing selection as defined by the selection coefficients.

tf, target frequency: the frequency used in *NCD* as the value to which queried allele frequencies are compared to.

NCD statistics, non-central deviation statistics, with two implementations, *NCD1* and *NCD2*.

NCD1, measures the average departure between polymorphic allele frequencies and a pre-determined frequency (*tf*). *NCD1*(*tf*) is *NCD1* for that given *tf*.

NCD2, measures the average departure between allele frequencies and a pre-determined frequency (*tf*) considering both polymorphisms and fixed differences with an outgroup. *NCD2*(*tf*) is *NCD2* for that given *tf*.

NCD(*tf*), refers to the average of *NCD1*(*tf*) and *NCD2*(*tf*).

For simplicity, we averaged power estimates across *NCD* implementations (*NCD* being the average of *NCD1* and *NCD2*), African and European demographic models (Asian populations were not considered, see below and S2 Note), *L* and *Tbs* (Methods). These averages are helpful in that

they reflect the general changes in power driven by individual parameters. Nevertheless, because they often include conditions for which power is low, they underestimate the power the test can reach under each condition. The complete set of power results is presented in S1 Table, and some key points are discussed below.

Time since the onset of balancing selection (*Tbs*) and sequence length. Signatures of LTBS are expected to be stronger for longer *Tbs*, because time to the most recent common ancestor is older and there will have been more time for linked mutations to accumulate and reach intermediate frequencies. We simulated sequences with variable *Tbs* (1, 3, 5 million years, mya). For simplicity, here we only discuss cases where $tf = f_{eq}$, although this condition is relaxed in later sections. Power to detect LTBS with *Tbs* = 1 mya is low (*NCD(0.5)* = 0.32, averaged across populations and *L* values), and high for 3 (0.74) and 5 mya (0.83) (S3-S8 Figs, S1 Table), suggesting that *NCD* statistics are well powered to detect LTBS starting at least 3 mya. We thus focus subsequent power analyses exclusively on this timescale.

In the absence of epistasis, the long-term effects of recombination result in narrower signatures when *Tbs* is larger [23,24]. Accordingly, we find that, for example, power for *NCD(0.5)* (*Tbs* = 5 mya) is on average 10% higher for 3,000 bp loci than for 12,0000 bp loci (S3-S8 Figs, S1 Table). In brief, our simulations show power is highest for windows of 3 kb centered on the selected site (S2 Note), and we report power results for this length henceforth.

Demography. Power is similar for samples simulated under African and European demographic histories (Table 1), but considerably lower under the Asian one (S1 Table, S3-S8 Figs), possibly due to lower N_e (S2 Note). While power estimates may be influenced by the particular demographic model used, we nevertheless focus on African and European populations, which by showing similar power allow fair comparisons between them.

Simulated and target frequencies. So far, we have only discussed cases where $tf = f_{eq}$, which is expected to favor the performance of *NCD*. Accordingly, under this condition *NCD* has high power: 0.91, 0.85, and 0.79 on average for f_{eq} = 0.5, 0.4, and 0.3, respectively (averaged across *Tbs* and populations, Table 1). However, since in practice there is no prior knowledge about the f_{eq} of balanced polymorphisms, we evaluate the power of NCD when $f_{eq} \neq tf$. When f_{eq} = 0.5, average power is high for tf = 0.5 or 0.4 (above 0.85), but lower for tf = 0.3 (0.50, Table 1). Similar patterns are observed for other simulated f_{eq} (Table 1). Therefore, *NCD* statistics are overall well-powered both when the f_{eq} is the same as tf, but also in some instances of $f_{eq} \neq tf$. In any case, the closest tf is to f_{eq}, the higher the power, so when possible, it is desirable to perform tests across a range of tf.

Table 1. Power for simulations under the African and European demographic models

		Africa						Europe					
		NCD2			***NCD1***			***NCD2***			***NCD1***		
Tbs	f_{eq}	***tf***			***tf***			***tf***			***tf***		
		0.5	**0.4**	**0.3**	**0.5**	**0.4**	**0.3**	**0.5**	**0.4**	**0.3**	**0.5**	**0.4**	**0.3**
5	0.5	0.96	0.94	0.84	0.93	0.91	0.39	0.97	0.95	0.84	0.92	0.85	0.20
5	0.4	0.94	0.94	0.89	0.89	0.89	0.67	0.95	0.94	0.91	0.85	0.84	0.60
5	0.3	0.90	0.91	0.93	0.72	0.80	0.84	0.84	0.86	0.89	0.47	0.57	0.74
3	0.5	0.91	0.88	0.68	0.86	0.80	0.24	0.93	0.89	0.68	0.81	0.69	0.15
3	0.4	0.88	0.86	0.76	0.78	0.78	0.56	0.89	0.88	0.79	0.74	0.71	0.46
3	0.3	0.75	0.77	0.81	0.56	0.64	0.71	0.73	0.76	0.80	0.39	0.48	0.63

Power at false positive rate (FPR) = 5%. Simulations with L = 3 kb. *Tbs*, time in mya since onset of balancing selection; f_{eq}, equilibrium frequency in the simulations. Power on additional conditions is presented on S1 Table.

***NCD* implementations and comparison to other methods.** Power for *NCD2* is greater than for *NCD1* for all tf: f_{eq} = 0.5 (average power of 0.94 for *NCD2(0.5)* vs. 0.88 for *NCD1(0.5),* averaged across populations and *Tbs*; Table 1), f_{eq} = 0.4 (0.90 for *NCD2(0.4)* vs. 0.80 for *NCD1(0.4)*) and f_{eq} = 0.3 (0.86 for *NCD2(0.3)* vs. 0.73 for *NCD1(0.3)*) (Table 1, Fig 2). This illustrates the gain in power by incorporating FDs in the NCD statistic, which is also more powerful than combining *NCD1* and HKA (S1 Table).

We compared the power of *NCD* to two statistics commonly used to detect balancing selection (Taj*D* and HKA), a composite statistic of *NCD1* and HKA (with the goal of quantifying the contribution of FD to *NCD* power), and a pair of composite likelihood-based measures (T_1 and T_2 [31]). The T_2 statistic, similarly to *NCD*2, considers both the SFS and the ratio of polymorphisms to FD. Power results are summarized in Fig 2. When $f_{eq} = 0.5$, *NCD2(0.5)* has the highest power: for example, in Africa (*Tbs* = 5 myr, and 3 kb) *NCD2(0.5)* power is 0.96 (the highest among other tests is 0.94, for T_2) but the difference in power is highest when f_{eq} departs from 0.5. For $f_{eq} = 0.4$, *NCD2(0.4)* power is 0.93 (compared to 0.90 for Taj*D* and T_2, and lower for the other tests). For $f_{eq} = 0.3$, *NCD2(0.3)* power is 0.93 (compared to 0.89 for T_2 and lower for the other tests). These patterns are consistent in the African and European simulations (Fig 2, S10 Fig), where *NCD*2 has greater or comparable power to detect LTBS than other available methods. When focusing on the tests that use only polymorphic sites, *NCD1* has similar power to Tajima's D when $f_{eq} = 0.5$, and it outperforms it when f_{eq} departs from 0.5 (Table S1). Altogether, the advantage of *NCD2* over classic neutrality tests is its high power, especially when f_{eq} departs from 0.5; the advantage over T_2 is its simplicity of implementation and interpretation, and the fact that it can be run in the absence of a demographic model.

Recommendations based on power analyses. Overall, *NCD* performs very well in regions of 3 kb (Table 1, Fig 2) and similarly for African and European demographic scenarios. In fact, *NCD2* outperforms all other methods tested (Fig 2, S10 Fig) and reaches very high power when $tf = f_{eq}$ (always > 0.89 for 5 mya and always > 0.79 for 3 mya). While the f_{eq} of a putatively balanced allele is unknown, the simplicity of the *NCD* statistics makes it trivial to run for several *tf* values, allowing detection of balancing selection for a range of equilibrium frequencies. Also, the analysis can be run in sliding windows to ensure overlap with the narrow signatures of balancing selection. Alternatively, *NCD* could also be computed for 3kb windows centered in each SNP or IS. Because *NCD2* outperforms *NCD1*, we used it for our scan of human populations; *NCD1* is nevertheless a good choice when outgroup data is lacking.

Identifying signatures of LTBS

We aimed to identify regions of the human genome under LTBS. We chose *NCD2(0.5)*, *NCD2(0.4)* and *NCD2(0.3)*, which provide sets of candidate windows that are not fully overlapping (Table 1). We calculated the statistics for 3 kb windows (1.5 kb step size) and tested for significance using two complementary approaches: one testing all windows with respect to neutral expectations, and one identifying outlier windows in the empirical genomic distribution. We analyzed genome-wide data from two African (YRI: Yoruba in Ibadan, Nigeria; LWK: Luhya in Webuye, Kenya) and two European populations (GBR: British, England and Scotland; TSI: Toscani, Italy) [40]. We filtered the data for orthology with the chimpanzee genome (used as the outgroup) and implemented 5 additional filters to avoid technical artifacts (S13 Fig). Finally, we excluded windows with less than 10 IS in any of the populations since these showed a high variance in *NCD2* due to noisy SFS (see empirical patterns in S18 Fig and neutral simulation patterns in S11 Fig).

Simulation and empirical-based sets of windows. After all filters were implemented, we analyzed 1,657,989 windows (~ 81% of the autosomal genome; S13 Fig), overlapping 18,633 protein-coding genes. We defined a *p*-value for each window as the quantile of its *NCD2* value when compared to those from 10,000 neutral simulations under the inferred demographic history of each population and conditioned on the same number of IS. Depending on the population, between 6,226 and 6,854 (0.37-0.41%) of the scanned windows have a lower *NCD2(0.5)* value than any of the 10,000 neutral simulations ($p < 0.0001$). The proportions are similar for *NCD2(0.4)* (0.40-0.45%) and *NCD2(0.3)* (0.33-0.38%) (Table 2). We refer to these sets, whose patterns cannot be explained by the neutral model, as the *significant* windows. In each population, the union of significant windows considering all *tf* values spans, on average, 0.6% of the windows (Table 2) and 0.77% of the base pairs.

Due to our criterion, all significant windows had simulation-based $p < 0.0001$. In order to quantify how far the *NCD2* value of each window is from neutral expectations, we defined $Z_{tf\text{-}IS}$

(Equation 2, see Methods) as the number of standard deviations a window's *NCD2* value lies from the neutral expectation. We defined as *outlier* windows those with the most extreme signatures of LTBS (in the 0.05% lower tail of the $Z_{tf\text{-}IS}$ distribution). This more conservative set contains 829 outlier windows for each population and *tf* value (Table 2), which cover only ~ 0.09% of the base pairs analyzed and largely included in the set of significant windows. Significant and outlier windows are collectively referred to as *candidate windows*.

Table 2. Candidate windows and protein-coding genes across populations

Population	LWK				YRI				GBR				TSI			
tf	0.3	0.4	0.5	U	0.3	0.4	0.5	U	0.3	0.4	0.5	U	0.3	0.4	0.5	U
Significant windows	7,364	6,841	6,226	10,072	8,098	7,420	6,854	10,997	6,173	6,858	6,519	9,521	5,998	6,661	6,403	9,2
Outlier windows	829	829	829	1,248	829	829	829	1,230	829	829	829	1,251	829	829	829	1,2
Significant genes	1,182	1,099	1,026	1,457	1,326	1,187	1,110	1,616	1,026	1,094	1,051	1,414	1,039	1,091	1,056	1,4
Outlier genes	151	158	167	249	142	151	157	232	140	157	159	222	143	157	166	23

Significant and outlier genes and windows, see main text. U, union of windows considering three *tf* values. Total number of queried windows per population is 1,657,989. Union of all candidate genes is 2,348 (significant) and 402 (outlier).

Reliability of candidate windows. Significant windows are enriched both in polymorphic sites (Fig 3A-B) and intermediate-frequency alleles (Fig 3C-D), and the SFS shape reflects the *tf* for which they are significant (Fig 3C-D). Although expected, because these were the patterns used to identify these windows, this shows that significant windows are unusual in both signatures. These striking differences with respect to the background distribution, combined with the fact that neutral simulations do not have *NCD2* values as low as those of the significant windows, precludes relaxation of selective constraint as a an alternative explanation to their signatures [28].

To avoid technical artifacts among significant windows we filtered out regions that are prone to mapping errors (S13 Fig). Also, we find that significant windows have similar coverage to the rest of the genome, i.e, they are not enriched in unannotated, polymorphic duplications (S14 Fig). We also examined whether these signatures could be driven by two biological mechanisms other than LTBS: archaic introgression into modern humans and ectopic gene conversion (among paralogs). These mechanisms can increase the number of polymorphic sites and (in some cases) shift the SFS towards intermediate frequency alleles (S5 Note). We find introgression is an unlikely confounding mechanism, since candidate windows are depleted in SNPs introgressed from Neanderthals (S17 Fig, S5 Note). Also, genes overlapped by significant windows are not predicted to be affected by ectopic gene conversion with neighboring paralogs to an unusually high degree, with the exception of olfactory receptor genes (S16 Fig, S5 Note). Thus, candidate windows represent a catalog of strong candidate targets of LTBS in human populations.

Assigned *tf* values

For both novel and previously known targets of LTBS, an advantage of *NCD* is that it provides an assigned *tf* for each window, which reflects the shape of its SFS. Our simulations suggest that the assigned *tf* is informative about the frequency of the site under balancing selection, so when a window was detected for more than one *tf*, we identified the *tf* value that minimizes $Z_{tf\text{-}IS}$ (S3 Note). On average ~53% of the candidate windows are assigned to *tf* = 0.3, 27% to *tf* = 0.4 and 20% to *tf* = 0.5 (S5 Table).

Non-random distribution across chromosomes. Candidate windows are not randomly distributed across the genome. Chromosome 6 is the most enriched for signatures of LTBS, contributing, for example, 10.2% of significant and 25% of outlier windows genome-wide for LWK while having only 6.4% of analyzed windows (S12 Fig, with qualitatively similar results for the other populations). This pattern can be explained by the MHC region (Fig 4A), rich in genes with well-

supported evidence for LTBS. Specifically, 10 HLA genes are among the strongest candidates for balancing selection in all four populations, most of which have prior evidence of balancing selection (S6 Table, S4 and S6 Notes): *HLA-B, HLA-C, HLA-DPA1, HLA-DPB1, HLA-DQA1, HLA-DQB1, HLA-DQB2, HLA-DRB1, HLA-DRB5, HLA-G* [24,31,41–45].

Biological pathways influenced by LTBS

To gain insight on the biological pathways influenced by LTBS, we focused on protein-coding genes containing at least one candidate window (222-249 outlier and 1,404-1,616 significant genes per population), and investigated their annotations. They are disproportionally expressed in a number of tissues: lung, adipose tissue, adrenal tissue, kidney, and prostate (S4 Table).

Regarding functional categories, significant genes are overrepresented in 28 GO categories, 24 of which are shared by at least two populations and 18 by four populations. Thirteen categories are immune-related according to a list of 386 immune-related keywords from ImmPort (Methods). The more stringent sets of outlier genes are enriched for 28 GO categories (21 shared by all four populations), 18 of which are immune-related. Furthermore, in both sets several of the remaining enriched categories are directly related to antigen presentation although not classified as immune-related (e.g., "ER to golgi transport vesicle membrane", "integral to membrane"). Among the non immune-related categories are "sarcolemma", "epidermis development", "keratin filament" and "negative regulation of blood coagulation" (S2 Table).

When classical HLA genes are removed from the analyses, only two categories remain enriched: "sarcolemma" (in YRI) and "epidermis development" (GBR), but the small set of genes per population hampers power. For the significant windows, "antigen processing and presentation of endogenous peptide antigen via MHC class I" remains significantly enriched (driven by *TAP1*, *TAP2, ERAP1* and *ERAP2*; S2 Table). Also, significant windows are still enriched in categories related to the extracellular space – "extracellular regions", "integral to membrane" (as in [15,28,31]) – and "keratin filament". These categories are not immune-related *per se*, but they

represent physical barriers to the invasion by pathogens. This indicates that LTBS maintains advantageous diversity in numerous defense-related genes other than classical HLA genes.

Overall, 33% of the outlier (and 31% of the significant) genes have at least one immune-related associated GO category, while only 24% of scanned genes do (see Methods). These results collectively suggest that immunity and defense are frequent targets of LTBS, although a large fraction of the candidates for LTBS have non-immune related functions or indirect connections with immunity hitherto unknown.

Functional annotation of SNPs in candidate windows

Because the identification of candidate windows is independent from functional annotation, we were able to test whether LTBS preferentially maintains SNPs at particular types of functional sites. To do so we investigated the overlap of candidate windows with different classes of functional annotations in the human genome, and tested the hypothesis of enrichment of certain classes of sites within our sets of candidate windows, when compared to sets of randomly sampled windows from the genome (S8 Table and Fig 5).

SNPs in outlier windows overlap disproportionally with protein-coding exons in all the populations ($p \leq 0.001$, one-tail test; Fig 5, see Methods). The protein-coding enrichment is even stronger when considering only SNPs within genes, which both in outlier ($p < 0.001$) and significant windows ($p \leq 0.003$) are strongly enriched in protein-coding exons (Fig 5). Within the protein-coding exons, outlier windows in Africa ($p \leq 0.022$) and significant windows in all populations ($p \leq 0.037$) are enriched in non-synonymous SNPs (Fig 5). These observations show that our candidate targets of LTBS tend to be enriched in exonic and potentially functional (amino-acid altering) SNPs.

Conversely, outlier and significant windows have no excess of SNPs annotated as regulatory ($p \geq 0.458$ in all populations, Fig 5). When we explicitly compared protein-coding exons vs. regulatory sites by restricting our analysis to sites in these two categories, outlier windows have an

excess of exonic SNPs ($p \leq 0.003$). The same is true for significant windows ($p \leq 0.016$; Fig 5). When only nonsynonymous and regulatory sites are considered, we see enrichment for LWK and YRI for the outlier windows ($p \leq 0.036$, Fig 5) but not for the significant windows ($p \geq 0.458$ for all populations, Fig 5), although the two analyses that consider nonsynonymous SNPs are likely underpowered due to low SNP counts (S8 Table). Finally, results using more detailed RegulomeDB annotations generally agree with the observation of lack of enrichment of regulatory sites in our candidate windows ($p \geq 0.121$ for a one tail test for enrichment for RegulomeDB1+3 for SNPs with MAF ≥ 0.2) (S6 Note, S8 Table).

Although perhaps limited by the quality of the annotation of regulatory sites and the low power associated to small SNP counts for nonsynonymous variants, we do not have strong evidence that LTBS in human populations has preferentially shaped variation at sites with a role in gene expression regulation. These results suggest that LTBS preferentially affects exons and non-synonymous mutations.

Monoallelic expression. Genes with mono-allelic expression (MAE) – i.e, the random and mitotically stable choice of an active allele for a given locus – have been found to be enriched among those with signatures of balancing selection [46]. Our observations agree with this. For example, 64% and 62% of the outlier and significant genes shared by at least two populations have MAE status according to [46], compared to only 41% for genes without signatures of LTBS ($p < 1.12^{-6}$ Fisher Exact Test, one-sided).

Overlap across populations

On average 86% of outlier windows in a given population are shared with another population (79% for significant windows), and 77% with another population within the same continent (66% for significant ones) (S19 Fig). The sharing is similar when *tf* are considered separately (S20 and S21 Figs). Consequently, there is also considerable overlap of candidate protein-coding genes across

populations: *e.g.* in LWK (*tf* = 0.5), 76.6% of outlier genes are shared with any other population, and 66% are shared with YRI (89% and 77% for significant genes; Fig 4B). In fact, on average 44% of outlier genes for a given population are shared across *all* populations and 78.7% are shared by a same-continent population (50% and 77% for significant genes; S22 Fig).

Candidate genes in more than one population. Instances where signatures of LTBS are not shared between populations may result from changes in selective pressure, which may be important during fast, local adaptation [47]. Still, loci with signatures of LTBS across human populations are more likely to represent stable selection. We considered as "shared" those candidate protein-coding genes (from the union of candidate windows for all *tf*) that are shared by all populations (S6 Table). For the rest, we considered as "African" those shared between YRI and LWK (but neither or only one European population), and "European" those shared between GBR and TSI (but neither or only one African population). We note that these designations do not imply that genes referred to as "African" or "European" are putative targets of LTBS for only one continent (partially because there are some power differences between Africa and Europe, Table 1). The 79 African, 84 European and 102 shared" outlier genes add up to 265 genes in total (~1.5% of all queried genes) and the 458 African, 400 European, and 736 shared significant genes add up to 1,594 (~8.5% of all queried genes; S6 Table).

Discussion

The targets of LTSB in the human genome

Using simulation-based and empirical outlier approaches, we uncovered windows with signatures of LTBS in humans. We showed that these windows are unlikely to be affected by technical artifacts or confounding biological processes other than LTBS, such as introgression from archaic hominins. On average, across populations, 0.6% of the windows in a population are significant: we never observe comparable or more extreme signatures of LTBS in 10,000 neutral simulations.

These windows contain on average 0.77% of the base-pairs and 1.6% of the SNPs in the genome per population, and although they amount to a low proportion of the genome, on average 7.9% of the protein-coding genes in a population contain at least one significant window (considering UTRs, introns and protein coding exons). For the more restrictive set of outlier windows (0.05% of windows), on average 2.1 % of genes in each population show some evidence of selection.

In both sets, we identified many previously known targets of LTBS, but almost 70% of the outlier genes shared by same-continent populations (and 90% of the significant genes) are novel. Many of these candidate genes show strongest evidence for LTBS at *tf* values different from 0.5. This is expected, for instance, under asymmetric overdominance, and highlights the importance of considering selective regimes with different frequencies of the balanced polymorphism.

Functional properties of SNPs in candidate windows

In this study, we confirm cases where protein-coding regions are the likely target of selection, such as *HLA-B* and *HLA-C* [48], as well as cases where regulatory regions are probably targeted, such as *HLA-G*, *UGT2B4, TRIM5* [45,49,50]. Overall, we found a strong enrichment of exonic, and a weaker enrichment of aminoacid-altering SNPs in the candidate windows, suggesting an abundance of potentially functional SNPs within selected regions.

While LTBS has been proposed to play an important role in maintaining genetic diversity that affects gene expression [23,46], we find that regulatory SNPs are underrepresented within the candidate regions. This does not imply that there are no regulatory SNPs under balancing selection, but rather that with existing annotations (which are less precise for regulatory than protein-coding sites) they are not enriched within candidate targets. Overall, we show that LTBS plays an important role in maintaining diversity at the level of protein sequence. This is compatible with two scenarios: (a) direct selection on protein-coding sites or (b) accumulation of functional (including slightly deleterious) variants as a bi-product of balancing selection. Importantly, we show that significant windows are also extreme in their high density of polymorphisms and have a SFS that is

markedly different from neutral expectations, suggesting that relaxed purifying selection and background selection are unlikely to generate their signatures.

Overlap with previous studies. Whereas positive selection scans show a remarkably low overlap with respect to the genes they identify, with as few as 14% of protein-coding loci appearing in more than a single study [51], 34% of our outlier genes (11% of significant ones) had evidence of LTBS in at least one previous study [23,28,31]. Remarkably, 47% of the shared outliers across all four populations (17% of the shared significant ones) have been detected in at least one previous study, and the proportions are similar even when classical HLA genes are removed (39 and 16% overlap, respectively). This is a high degree of overlap, considering the differences in methods and datasets across studies. For example, we find 45% of the genes from [28] among the outliers (and 78% among the significant) and 10 % and 38% of genes from [31] among outlier and significant genes, respectively. Still, the majority of our loci represent novel targets.

Properties of candidate genes

Below we briefly discuss the outlier genes (S6 Table), highlighting the variety of biological functions and known genetic associations (see Methods) potentially shaped by LTBS in humans.

Mono-allelic expression. In agreement with previous findings, we found a significant excess of MAE genes among our outlier candidates. This excess is not driven by HLA genes, which were filtered out in the study originally reporting MAE genes and supports the claim for a biological link between MAE and balancing selection [46]. Heterozygosity in a MAE gene could lead to cell-to-cell heterogeneity within same-cell clusters, which could in turn be potentially advantageous [46,52]), particularly in the case of cell-surface proteins. Some of these MAE genes found in our study, and not previously detected in scans for balancing selection, are involved in immunity/defense barriers (e.g. *IL1RL1*, *IL18R1, FAM114A1, EDARADD, SIRPA, TAS2R14)*,

oxygen transport and hypoxia (e.g. *PRKCE*, *HBE1, HBG2*, *EGLN3),* or reproduction (e.g. *CLDN11).*

Oxygen transport and response to hypoxia. Among the outlier genes with MAE we find members of the beta-globin cluster (*HBE1* and *HBG2,* in the same window) that are involved in oxygen transport and have strong associations to hemoglobin levels and beta thalassemia[53], and *EGLN3,* a regulator of the NF-kß pathway that is significantly upregulated under hypoxia in anti-inflammatory macrophages [54] and also plays a role in skeletal muscle differentiation [55]. The encoded protein hydroxylates the product of *EPAS1*, a gene shown to harbor variants responsible for human adaptation to high altitude in Tibet [56]. Interestingly, in addition to having strong signatures of LTBS in all populations we analyzed, they also have evidence for recent positive selection in Andean (*HBE1*, *HBG2*) or Tibetan (*HBG2*) populations [57–59]. It is plausible that these genes have been under LTBS, and have undergone a shift in selective pressures in high-altitude populations (as in [47]), but further analyses are required to confirm this possibility. Another of our outlier genes, *PRKCE,* is also strongly associated to haemoglobin levels and red blood cell traits.

Immunological function and defense barriers. It has long been argued that genes of immune function are prime candidates for balancing selection. As expected, we detect several classical HLA with known signatures of LTBS. However, many non-HLA candidates from our set of outlier genes have immunological functions. For example, we confirm signatures of LTBS in the *ABO* locus, a well-known case of LTBS in humans [60] (S4 Note), and *TRIM5*, a gene with important antiviral function [49].

Among novel candidates of balancing selection, we find several genes involved in auto-immune disease. For example, *IL1RL1-IL18R1* have strong associations to celiac disease and atopic dermatitis, an auto-immune disease [61]). *HLA-DQB2* mediates superantigen activation of T cells

[62] and is associated both to infectious (hepatitis B) and autoimmune diseases (e.g. lupus [63,64]). Two other significant genes for which there is prior evidence for LTBS [65,66], *ERAP1* and *ERAP2*, are associated with ankylosing spondylitis and psoriasis (e.g [67–69]). Finally, there are several associations to autoimmune disease and susceptibility to infections in the classical HLA genes that we identify. In brief, our results are consistent with the hypothesis that auto-immune disease is linked to natural selection favoring effective immune response against pathogens [9,70].

Another important aspect of defense is the avoidance of poisonous substances. As suggested previously by studies on polymorphism in PTC receptors [71,72], avoidance of bitterness might have been adaptive throughout human evolutionary history because several potentially harmful substances are bitter. The *TAS2R14* gene encodes for a bitter taste receptor, and in humans it has strong associations to taste perception of quinine and caffeine [73], is considered a promiscuous receptor [74–76], and is one of the few bitter taste receptors that binds a vast array of compounds, and for which no common structure has been found [75,77]. This entails diversity in the antigen binding portions of the receptors, which may be enhanced by balancing selection. Indeed, elevated *dN/dS* ratio was reported for a cluster of bitter taste receptors which includes *TAS2R14* [78]. To our knowledge, our study is the first in detecting signatures of LTBS in this gene.

Cognition. Interestingly, several candidate genes are involved in cognitive abilities, or their variation is associated with diversity in related phenotypes. The *KL* (life extension factor klotho) is a gene that has been associated to human longevity [79] and for which signatures of LTBS have been previously reported [31]. In mice, decreased levels of klotho shorten lifespan (reviewed in [80]). In humans, heterozygotes for the KL-VS variant show higher levels of serum klotho and enhanced cognition, independent of sex and age, than wild-type homozygotes. On the other hand, KL-VS homozygotes show decreased lifespan and reduced cognition [81]. If higher cognition is advantageous, overdominance forthis phenotype can explain the signatures of balancing selection we observe (although klotho's the effect in lifespan can also influence).

PDGFD encodes a growth factor that plays an essential role in wound healing and angiogenesis. A comparison between human and mice revealed that the PDGFD-induced signaling is crucial for human (but not mouse) proliferation of the neocortex due to neural stem-cell proliferation [82], a trait that underlies human cognition capacities [83]. This gene has strong associations to coronary artery disease and myocardial infarction, which are related to aging.

Also, among our outliers, a gene with a cognitive-related genetic association is *ROBO2,* a transmembrane receptor involved in axon guidance. Associations with vocabulary growth have been reported for variants in its vicinity [84]. *ROBO2* has signatures of ancient selective sweeps in modern humans after the split with Neanderthals and Denisova [85] on a portion of the gene (chr3:77027850-77034264) almost 40kb apart from the one for which we identified a signature of LTBS (chr3:76985072-76988072). The occurrence of both these signatures highlights the complex evolutionary relevance of this gene.

Associations of genetic diversity in candidate genes with cognition are also supported by case-control and cohort studies linking polymorphisms in the estrogen receptor alpha (ER-α) gene, *ESR1*, to dementia and cognitive decline. Links between ER-α variants and mood outcomes such as anxiety and depression in women have been proposed but lack confirmation (reviewed in [86]). Interestingly, three other of our candidate genes (*PDLIM1,GRIP1, SMYD3)* interact with ER-α at the protein level [87], and two have strong association with suicide risk (*PDLIM1,GRIP1*)[88,89].

In genes like *KL*, where heterozygotes show higher cognitive abilities than homozygotes, cognition may be a driving selective force. This is a possible scenario in other genes, too. Still, given the complexity of brain development and function, it is also possible that cognitive effects of this variation are a byproduct of diversity maintained for other phenotypes. For example, MHC proteins and other immune effectors are believed to affect connectivity and function of the brain {reviewed in [90,91]), with certain alleles being clearly associated with autism disorder ([91–93]).

Reproduction

We see an enrichment for genes preferentially expressed in the prostate, as well as a number of outlier genes involved in the formation of the sperm. For example, *CLDN11* encodes a tight-junction protein expressed in several tissues and crucial for spermatogenesis. Knockout mice for the murine homologue show both neurological and reproductive impairment, i.e, mutations have pleiotropic effects [94,95]. In humans, variants in the gene are strongly associated to prostate cancer.

ESR1, which as mentioned above (in the Cognition section) encodes the ER-α transcription factor activated by estrogen, leads to abnormal secondary sexual characteristics in females when defective [96]. ER-α interacts directly with the product of *BRCA1* and has strong associations to breast cancer [97] and breast size [98]. It also harbors strong associations to menarche (age at onset). In males, it is involved in gonadal development and differentiation, and lack of estrogen and/or this receptor in males can lead to poor sperm viability (reviewed in [99]). Strikingly, this gene also has SNPs strongly associated to a diverse array of phenotypes, including height, bone mineral density (spine and hip), and sudden cardiac arrest [100–102]. Two other genes among our candidates are also part of the estrogen signaling pathway: *PLCB4* and *ADCY5* (which is strongly associated to birth weight). Estrogens are not only involved in reproductive functions (both in male and females), but also in several other processes of neural (see above), muscular or immune nature, and the ER-α-estrogen complex can act directly on promoter regions of other genes, or interact with transcription factors of genes without estrogen-sensitive promotor regions [103]. In this case, balancing selection could be explained by the high level of pleiotropy (if different alleles are beneficial for different functions), including the function in male and female reproduction (if different alleles are beneficial in males than females).

Conclusions

We present two new summary statistics, *NCD1* and *NCD2*, which are both simple and fast to implement on large datasets to identify genomic regions with signatures of LTBS. They have a high

degree of sensitivity for different equilibrium frequencies of the balanced polymorphism and, unlike classical statistics such as Tajima's *D* or the Mann-Whitney U [28,37], allow an exploration of the most likely frequencies at which balancing selection maintains the polymorphisms. This property is shared with the likelihood-based T_1 and T_2 tests [31]. We show that the *NCD* statistics are well-powered to detect LTBS within a complex demographic scenario, such as that of human populations. They can be applied to either single loci or the whole-genome, in species with or without detailed demographic information, and both in the presence and absence of an appropriate outgroup.

More than 85% of our outlier windows are shared across populations, raising the possibility that long-term selective pressures have been maintained after human populations colonized new areas of the globe. Still, about 15% of outlier windows show signatures exclusively in one sampled population and a few of these show opposing signatures of selective regimes between human groups; they are of particular relevance to understand how recent human demography might impact loci evolving under LTBS for millions of years or subsequent local adaptations through selective pressure shifts (e.g. [47]).

Our analyses indicate that, in humans, LTBS may be shaping variation in less than 2 % of variable genomic positions, but that these on average overlap with 7.9% of the protein-coding genes. Although immune-related genes represent a substantial proportion of them, almost 70% of the candidate genes cannot be ascribed to immune-related functions, suggesting that diverse biological functions, and the corresponding phenotypes, contain advantageous genetic diversity.

Methods

Simulations and power analyses

NCD performance was evaluated by simulations with MSMS [104] following the demographic model and parameter values described in [105] for African, European, and East Asian human populations (Fig 2). To obtain the neutral distribution for the NCD statistics, we simulated sequence

data under the following demographic model: generation time of 25 years, mutation rate of 2.5 x 10^{-8} per site and recombination rate of 1 x 10^{-8}, and a human-chimpanzee split at 6.5 mya was added to the model, which was used to obtain the neutral distributions for the NCD statistics. For the simulations with selection, a balanced polymorphism was added to the center of the simulated sequence and modeled to achieve a pre-specified frequency equilibrium (f_{eq} = 0.3, 0.4, 0.5) following an overdominant model (S2 Note). Simulations with and without selection were run for different sequence lengths (3, 6, 12 kb) and times of onset of balancing selection (1, 3, 5 mya). For each combination of parameters, 1,000 simulations, with and without selection, were used to compare the relationship between true (TPR, the power of the statistic) and false (FPR) positive rates for the NCD statistics, represented by ROC curves. For performance comparisons, we used FPR = 0.05. When comparing performance under a given condition, power was averaged across *NCD* implementations, demographic scenarios, *L*, and *Tbs*. When comparing NCD performance to other methods (Tajima's *D* [106], HKA [34], and a combined *NCD1*+HKA test), we simulated under *NCD* optimal conditions: *L* = 3 kb and *Tbs* = 5 mya (S1 Table). Since power for T_1 and T_2 is reported based on windows of 100 informative sites (~ 14 kb for YRI and CEU) up and downstream of the target site [31], we divided simulations of 15 kb into windows of 100 IS, calculated T_1 and T_2 with BALLET [31] and selected the highest T_1 or T_2 value from each simulation to obtain their power for the same set of parameters used for the other simulations.

Human population genetic data

We analyzed genome-wide data from the 1000 Genomes (1000G) Project phase I [40], excluding SNPs only detected in the high coverage exome sequencing in order to avoid SNP density differences between coding and non-coding regions. We queried genomes of individuals from two African (YRI, LWK) and two European populations (GBR, TSI). We did not consider Asian populations due to lower NCD performance for these populations according to our simulations (S1 Table, S7-8 Figs). To equalize sample size, we randomly sampled 50 unrelated individuals from

each population (as in [107]). We dedicated extensive efforts to obtain an unbiased dataset by extensive filtering in order to avoid the inclusion of errors that may bias results. We kept positions that passed mappability (50mer CRG, 2 mismatches [108]), segmental duplication [109,110] and tandem repeats filters [111], as well as the requirement of orthology to chimp (S13 Fig) because *NCD2* requires divergence information (Equation 1). Further, we excluded 3 kb windows: with less than 10 IS in any population (~2% of scanned windows) and less than 500 bp of positions with orthology in chimp (1.6%); the two criteria combined resulted in the exclusion of 2.2% of scanned windows.

Identifying signatures of LTBS

After applying all filters and requiring the presence of at least one informative site, *NCD2* was computed for 1,695,655 windows per population. Because in simulations 3kb windows yielded the highest power for *NCD2* (S3-S6 Figs, Table 1), we queried the 1000G data with sliding windows of 3 kb (1.5 kb step size). Windows were defined in physical distance since the presence of LTBS may affect the population-based estimates of recombination rate. For each window in each population we calculated *NCD2* for three *tf* values (0.3-0.5).

Filtering and correction for number of informative sites. Genome-wide studies of natural selection typically place a threshold on the minimum number of IS necessary (e.g., at least 10 IS in [28], or 100 IS in [31]). We observe considerable variance in the number of IS per 3 kb window in the 1000G data; also, *NCD2* has high variance when the number of IS is low in neutral simulations (S11 and S18 Figs). We thus excluded windows with less than 10 IS in a given population because, for higher values of IS, *NCD2* stabilizes. We then analyzed the 1,657,989 windows that remained in all populations, covering 2,145,937,383 base pairs (S13 Fig). Neutral simulations with different mutation rates were performed in order to retrieve 10,000 simulations for each value of IS (S18 Fig

and Methods). *NCD*2 (*tf* = 0.3, 0.4, 0.5) was calculated for all simulations, allowing the assignment of significant windows and the calculation of $Z_{tf\text{-}IS}$ (Equation 2 below).

Significant and outlier windows. We defined two sets of windows with signatures of LTBS: the *significant* (based on neutral simulations) and *outlier* windows (based on the empirical distribution of $Z_{tf\text{-}IS}$, see below). When referring to both sets, we use the term *candidate windows*. *Significant windows* were defined as those fulfilling the criterion whereby the observed *NCD2* value is lower than all values obtained from 10,000 simulations with the same number of IS. Thus, all significant windows have the same *p*-value ($p < 0.0001$). In order to rank the windows and define outliers, we used a standardized distance measure between the observed *NCD2* (for a queried window) and the mean of the *NCD2* values for the 10,000 simulations with the matching number of IS:

$$Z_{tf-IS} = \frac{NCD2_{tf} - \overline{NCD2}_{tf-IS}}{sd_{tf-IS}} \quad (2)$$

, where $Z_{tf\text{-}IS}$ is the standardized *NCD2*, conditional on the value of IS, $NCD2_{tf}$ is the *NCD2* value with a given *tf* for the *n-th* empirical window, $\overline{NCD2}_{tf-IS}$ is the mean *NCD2* for 10,000 neutral simulations for the corresponding value of IS, and $sd_{tf\text{-}IS}$ is the standard deviation for 10,000 *NCD2* values from simulations with matching IS. $Z_{tf\text{-}IS}$ allows the ranking of windows for a given *tf*, while taking into account the residual effect of IS number on $NCD2_{tf}$, as well as a comparison between the rankings of a window considering different *tf* values. An empirical *p*-value was attributed to each window based on the $Z_{tf\text{-}IS}$ values for each *tf*. Windows with empirical *p*-value < 0.0005 (829 windows) were defined as the *outlier windows.* Outlier windows are essentially a subset of significant windows (except for 5 windows in LWK, 1 window in YRI, 3 windows in GBR, and 4 windows in TSI). Significant and outlier windows for multiple *tf* values had an *assigned tf value*, defined as the one that minimizes the empirical *p*-value for a given window (S3 Note).

Coverage as a proxy for undetected short duplication. To test whether the signatures of LTBS are driven by undetected short duplications, which can produce mapping and SNP call errors, we

analyzed an alternative modern human genome-wide dataset, sequenced to an average coverage of 20x-30x per individual [112,113]. We used an independent data set because read coverage data is low and cryptic in the 1000G, and putative duplications affecting the SFS must be at appreciable frequency and should be present in other data sets. We considered 2 genomes from each of the following populations: Yoruba, San, French, Sardinian, Dai, and Han Chinese. For each sample, we retrieved positions above the 97.5% quantile of the coverage distribution for that sample ("high coverage" positions). For each window with signatures of LTBS, we calculated the proportion of the 3kb window having high coverage in at least two samples and plotted the distributions for different *NCD*2 $Z_{tf\text{-}IS}$ *p*-values. Extreme NCD windows are not enriched in high-coverage regions; in fact, they are depleted of them in some cases (S14 Fig) (Mann-Whitney U two-tail test; $p < 0.02$ for $tf = 0.5$ and $tf = 0.4$ for GBR and TSI).

Enrichment Analyses

Gene (GO) and Phenotype (PO) Ontology, and Tissue-specific expression. We analyzed protein-coding genes overlapped by one or more candidate windows. GO, PO and tissue of expression enrichment analyses were performed using GOWINDA [114], which corrects for gene length-related biases and/or gene clustering (S6 Note). GO/PO accession terms were downloaded from the GO Consortium (http://geneontology.org), and the Human PO (http://human-phenotype-ontology.github.io/). We ran analyses in *mode:gene* (which assumes that all SNPs in a gene are completely linked) and performed 100,000 simulations for FDR (false discovery rate) estimation. Significant GO, PO and tissue-specific categories were defined for a FDR<0.05. In both cases, a minimum of three genes in the enriched category was required.

For tissue-specific expression analysis we used Illumina BodyMap 2.0 [115] expression data for 16 tissues, and considered genes significantly highly expressed in a particular tissue when compared to the remaining 15 tissues using the DESeq package [116], as done in [117]. All three enrichment analyses (GO, PO, and tissue-specific expression) were performed for each population

and set of genes: outliers or significant; different *tf* values (or union of all *tf*); with or without classical HLA genes (S6 Note).

Archaic introgression and ectopic gene conversion. We evaluated two potentially confounding biological factors: ectopic gene conversion and archaic introgression. We verified the proportion of European SNPs in candidate windows that are potentially of archaic origin, and whether candidate genes tend to have elevated number of paralogs in the same chromosome. Details in S5 Note.

SNP annotations and re-sampling procedure. Functional annotations for SNPs were obtained from ENSEMBL-based annotations on the 1000G data (http://www.ensembl.org/info/genome/variation/predicted_data.html). Specifically, we categorized SNPs as: intergenic, genic, exonic, regulatory, synonymous, and non-synonymous. Details on which annotations were allocated to each of these broad categories are presented in S6 Note. Within each category, each SNP was only considered when variable in the population under analysis (S6 Note). For each candidate window, we sum the number of SNPs with each score, and then sum across candidate windows. To compare with non-candidate windows, we performed 1,000 re-samplings of the number of candidate windows (which were merged in case of overlap) from the set of background windows (all windows scanned). For each re-sampled set, we summed the number of SNPs in a particular category and then computed the ratios in Table S8 and Fig 5. We therefore obtained ratios for each re-sampling set, to which we compared the values from candidate windows to obtain empirical *p*-values. Because we considered the sum of scores across windows, and counted each SNP only once, results should be insensitive to window length (as overlapping candidate windows were merged). As before, we performed these analyses for each population and sets of windows: outliers or significant, considering the union of all *tf*.

Genes with monoallelic expression (MAE) and immune-related genes. To test for enrichment for genes with MAE, we quantified the number of outlier and significant genes with MAE and the number that have bi-allelic expression as described in [46]. We compared these proportions to those observed for all scanned genes (one-tailed Fisher's test.) The same procedure was adopted to test for enrichment of immune-related genes among our sets: we used a list of 386 keywords from the Comprehensive List of Immune Related Genes from Immport (https://immport.niaid.nih.gov/immportWeb/queryref/immportgene/immportGeneList.do) and queried how many of the outlier protein-coding genes (402 genes in total across populations and *tf*, of which 378 had at least one associated GO term) had at least one immune-related associated GO category.

All statistical analyses and figures were performed in R [118] (scripts available on https://github.com/bbitarello/NCV_dir_package). Gene Cards (www.genecards.org) and Enrichr [119] were used to obtain basic functional information about genes and STRING v10 [87] was used to obtain information for interactions between genes. The GWAS catalog [120] was used to search for associations included in the discussion (we only report "strong associations", i.e, when there is at least one SNP with $p < 10^{-8}$).

Acknowledgements

We would like to dedicate this manuscript to Scott Williamson, in memoriam, for playing a fundamental role in the conception of NCD. We also thank Warren Kretszchmar for analyses on the properties of related statistics not included here, and Eric Green for his support of that work. We thank Michael DeGiorgio for assistance with BALLET, Felix Key help with 1000 Genomes data sets, Michael Dannemann for assistance in the implementation of expression analyses, Stéphane Peyrégne for comments on the manuscript, and David Reher, members of the Evolutionary Genetics Group (São Paulo), Alex Cagan and Svante Päabo for helpful comments.

Author Contributions

AA, DM and BDB conceived and designed the study. BDB, CDF and PK performed data quality filters. AA, BDB, CDF and DM designed and explored the properties of the statistic. BDB and CDF performed power analyses and ran the genome-wide analysis. JCT and JS performed the enrichment analyses. All authors interpreted the data. AA and DM supervised the project. BDB, DM and AA wrote the manuscript, with contributions from all authors.

References

1. Meyer D, Thomson G. How selection shapes variation of the human major histocompatibility complex: a review. Ann Hum Genet. 2001;65: 1–26.
2. Spurgin LG, Richardson DS. How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. Proc Biol Sci. 2010;277: 979–88.
3. Robinson J, Halliwell JA, McWilliam H, Lopez R, Parham P, Marsh SGE. The IMGT/HLA database. Nucleic Acids Res. 2013;41: D1222-7.
4. Hedrick PW, Whittam TS, Parham P. Heterozygosity at individual amino acid sites: extremely high levels for HLA-A and -B genes. Proc Natl Acad Sci. 1991;88: 5897–5901.
5. Prugnolle F, Manica A, Charpentier M, Guégan JF, Guernier V, Balloux F. Pathogen-driven selection and worldwide HLA class I diversity. Curr Biol. 2005;15: 1022–7.
6. Raychaudhuri S, Sandor C, Stahl E a, Freudenberg J, Lee H-S, Jia X, et al. Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. Nat Genet. Nature Publishing Group; 2012;44: 291–296. doi:10.1038/ng.1076
7. Howell WM. HLA and disease: Guilt by association. Int J Immunogenet. 2014;41: 1–12. doi:10.1111/iji.12088
8. Ferrer-Admetlla A, Bosch E, Sikora M, Marques-Bonet T, Ramirez-Soriano A, Muntasell A, et al. Balancing Selection Is the Main Force Shaping the Evolution of Innate Immunity Genes. J Immunol. 2008;181: 1315–1322. doi:10.4049/jimmunol.181.2.1315
9. Sironi M, Clerici M. The hygiene hypothesis: an evolutionary perspective. Microbes Infect. Elsevier Masson SAS; 2010;12: 421–427. doi:10.1016/j.micinf.2010.02.002
10. Malaria Genomic Epidemiology Network. A novel locus of resistance to severe malaria in a region of ancient balancing selection. Nature. 2015;526: 253–257. doi:10.1038/nature15390
11. Biasin M, Piacentini L, Caputo S Lo, Kanari Y, Magri G, Trabattoni D, et al. Apolipoprotein B mRNA—Editing Enzyme, Catalytic Polypeptide—Like 3G: A Possible Role in the Resistance to HIV of HIV-Exposed Seronegative Individuals. J Infect Dis. 2007;195: 960–964. doi:10.1086/511988
12. Day FR, Hinds DA, Tung JY, Stolk L, Styrkarsdottir U, Saxena R, et al. Causal mechanisms and balancing selection inferred from genetic associations with polycystic ovary syndrome. Nat Commun. 2015;6: 8464. doi:10.1038/ncomms9464
13. Andrés AM. Balancing Selection in the Human Genome. eLS. John Wiley & Sons, Ltd; 2011; 1–8. doi:10.1002/9780470015902.a0022863
14. Fijarczyk A, Babik W. Detecting balancing selection in genomes: Limits and prospects. Mol Ecol. 2015; n/a-n/a. doi:10.1111/mec.13226
15. Key FM, Teixeira JC, de Filippo C, Andrés AM. Advantageous diversity maintained by balancing selection in humans. Curr Opin Genet Dev. 2014;29: 45–51. doi:10.1016/j.gde.2014.08.001
16. Charlesworth B, Charlesworth D. Elements of Evolutionary Genetics. 1st ed. Roberts and Company Publishers; 2010.
17. Clarke B. Balanced polymorphism and the diversity of sympatric species. In: Nichols D, editor. Taxonomy and Geography. Oxford: Systematics Association; 1962.
18. Bergland AO, Behrman EL, O'Brien KR, Schmidt PS, Petrov DA. Genomic Evidence of Rapid and Stable Adaptive Oscillations over Seasonal Time Scales in Drosophila. PLoS Genet. 2014;10: e1004775. doi:10.1371/journal.pgen.1004775
19. Muehlenbachs A, Fried M, Lachowitzer J, Mutabingwa TK, Duffy PE. Natural selection of FLT1 alleles and their association with malaria resistance in utero. Proc Natl Acad Sci. 2008;105: 14488–14491.

doi:10.1073/pnas.0803657105
20. Charlesworth B, Nordborg M, Charlesworth D. The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided population. Genet Res. 1997;70: 155–174.
21. Charlesworth D. Balancing selection and its effects on sequences in nearby genome regions. PLoS Genet. 2006;2: 379–384. doi:10.1371/journal.pgen.0020064
22. Johnston SE, Gratten J, Berenos C, Pilkington JG, Clutton-Brock TH, Pemberton JM, et al. Life history trade-offs at a single locus maintain sexually selected genetic variation. Nature. 2013;502: 93–95. doi:10.1038/nature12489
23. Leffler EM, Pfeifer S, Auton A, Venn O, Bowden R, Bontrop R, et al. Multiple Instances of Ancient Balancing Selection Shared Between Humans and Chimpanzees. Science (80-). 2013;339: 1578–1582. doi:10.1126/science.1234070
24. Teixeira JC, de Filippo C, Weihmann A, Meneu JR, Racimo F, Dannemann M, et al. Long-Term Balancing Selection in LAD1 Maintains a Missense Trans-Species Polymorphism in Humans, Chimpanzees, and Bonobos. Mol Biol Evol. 2015;32: 1186–1196. doi:10.1093/molbev/msv007
25. Sellis D, Callahan BJ, Petrov D a., Messer PW. Heterozygote advantage as a natural consequence of adaptation in diploids. Proc Natl Acad Sci. 2011;108: 20666–20671. doi:10.1073/pnas.1114573108
26. Hedrick PW. What is the evidence for heterozygote advantage selection? Trends Ecol Evol Evol. Elsevier Ltd; 2012;27: 698–704. doi:10.1016/j.tree.2012.08.012
27. Alonso S, Lopez S, Izagirre N, de la Rua C. Overdominance in the Human Genome and Olfactory Receptor Activity. Mol Biol Evol. 2008;25: 997–1001. doi:10.1093/molbev/msn049
28. Andrés AM, Hubisz MJ, Indap A, Torgerson DG, Degenhardt JD, Boyko AR, et al. Targets of balancing selection in the human genome. Mol Biol Evol. 2009;26: 2755–64. doi:10.1093/molbev/msp190
29. Bubb KL, Bovee D, Buckley D, Haugen E, Kibukawa M, Paddock M, et al. Scan of human genome reveals no new Loci under ancient balancing selection. Genetics. 2006;173: 2165–77. doi:10.1534/genetics.106.055715
30. Asthana S, Schmidt S, Sunyaev SR. A limited role for balancing selection. Trends Genet. 2005;21: 30–32. doi:10.1016/j.tig.2004.11.007
31. DeGiorgio M, Lohmueller KE, Nielsen R. A model-based approach for identifying signatures of ancient balancing selection in genetic data. PLoS Genet. 2014;10: e1004561. doi:10.1371/journal.pgen.1004561
32. Rasmussen MD, Hubisz MJ, Gronau I, Siepel A. Genome-Wide Inference of Ancestral Recombination Graphs. Coop G, editor. PLoS Genet. 2014;10: e1004342. doi:10.1371/journal.pgen.1004342
33. Hudson RR, Kaplan NL. The coalescent process in models with selection and recombination. Genetics. 1988;120: 831–840. doi:10.1017/S0016672300029074
34. Hudson RR, Kreitman M, Aguade M. A Test of Neutral Molecular Evolution Based on Nucleotide Data. Genetics. 1987;116: 153–159. Available: http://www.genetics.org/cgi/content/abstract/116/1/153
35. Coventry A, Bull-Otterson LM, Liu X, Clark AG, Maxwell TJ, Crosby J, et al. Deep resequencing reveals excess rare recent variants consistent with explosive population growth. Nat Commun. 2010;1: 131. doi:10.1038/ncomms1130
36. Fu W, Akey JM. Selection and Adaptation in the Human Genome. Annu Rev Genomics Hum Genet. 2013;14: 467–489. doi:10.1146/annurev-genom-091212-153509
37. Nielsen R, Hubisz MJ, Hellmann I, Torgerson D, Andrés AM, Albrechtsen A, et al. Darwinian and demographic forces affecting human protein coding genes. Genome Res. 2009;19: 838–49. doi:10.1101/gr.088336.108
38. Nielsen R, Bustamante C, Clark A, Glanowski S, Sackton T, Hubisz MJ, et al. A Scan for Positively Selected Genes in the Genomes of Humans and Chimpanzees. PLoS Biol. 2005;3: e170. doi:10.1371/journal.pbio.0030170
39. 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. doi:10.1371/journal.pgen.0030090
40. Abecasis GR, Auton A, Brooks LD, DePristo M a, Durbin RM, Handsaker RE, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491: 56–65. doi:10.1038/nature11632
41. Liu X, Fu Y, Liu Z, Lin B, Xie Y, Liu Y, et al. An Ancient Balanced Polymorphism in a Regulatory Region of Human Major Histocompatibility Complex Is Retained in Chinese Minorities but Lost Worldwide. Am J Hum Genet. 2006;78: 393–400. doi:10.1086/500593
42. Meyer D, Single RM, Mack SJ, Erlich HA, Thomson G. Signatures of demographic history and natural selection in the human major histocompatibility complex Loci. Genetics. 2006;173: 2121–2142. doi:10.1534/genetics.105.052837
43. Sanchez-Mazas A. An apportionment of human HLA diversity. Tissue Antigens. 2007;69: 198–202. doi:10.1111/j.1399-0039.2006.00802.x
44. Solberg OD, Mack SJ, Lancaster AK, Single RM, Tsai Y, Sanchez-Mazas A, et al. Balancing selection and heterogeneity across the classical human leukocyte antigen loci: A meta-analytic review of 497 population studies. Hum Immunol. 2008;69: 443–464. doi:10.1016/j.humimm.2008.05.001
45. Tan Z, Shon AM, Ober C. Evidence of balancing selection at the HLA-G promoter region. Hum Mol Genet. 2005;14: 3619–3628. doi:10.1093/hmg/ddi389

46. Savova V, Chun S, Sohail M, McCole RB, Witwicki R, Gai L, et al. Genes with monoallelic expression contribute disproportionately to genetic diversity in humans. Nat Genet. 2016;48: 231–237. doi:10.1038/ng.3493
47. de Filippo C, Key FM, Ghirotto S, Benazzo A, Meneu JR, Weihmann A, et al. Recent Selection Changes in Human Genes under Long-Term Balancing Selection. Mol Biol Evol. 2016; msw023. doi:10.1093/molbev/msw023
48. Hughes AL, Nei M. Pattern of nucleotide substitution at major histocompatibility class I loci reveals overdominant selection. Lett to Nat. 1988;335: 167–170.
49. Cagliani R, Fumagalli M, Biasin M, Piacentini L, Riva S, Pozzoli U, et al. Long-term balancing selection maintains trans-specific polymorphisms in the human TRIM5 gene. Hum Genet. 2010;128: 577–88. doi:10.1007/s00439-010-0884-6
50. Sun C, Huo D, Southard C, Nemesure B, Hennis A, Cristina Leske M, et al. A signature of balancing selection in the region upstream to the human UGT2B4 gene and implications for breast cancer risk. Hum Genet. 2011;130: 767–775. doi:10.1007/s00439-011-1025-6
51. Akey JM. Constructing genomic maps of positive selection in humans: where do we go from here? Genome Res. 2009;19: 711–722. doi:10.1101/gr.086652.108
52. Sung MK, Jang J, Lee KS, Ghim C, Choi JK. Selected heterozygosity at cis -regulatory sequences increases the expression homogeneity of a cell population in humans. Genome Biol. Genome Biology; 2016; 1–15. doi:10.1186/s13059-016-1027-8
53. Danjou F, Zoledziewska M, Sidore C, Steri M, Busonero F, Maschio A, et al. Genome-wide association analyses based on whole-genome sequencing in Sardinia provide insights into regulation of hemoglobin levels. Nat Genet. 2015;47: 1264–1271. doi:10.1038/ng.3307
54. Escribese M, Sierra-Filardi E, Nieto C, Samaniego R, Sánchez-Torres C, Masuyama T, et al. The prolyl hydroxylase PHD3 identifies proinflammatory macrophages and its expression is regulated by activin A. J Immunol. 2012;189: 1946–1954. doi:10.4049/jimmunol.1201064
55. Fu J, Menzies K, Freeman RS, Taubman MB. EGLN3 Prolyl Hydroxylase Regulates Skeletal Muscle Differentiation and Myogenin Protein Stability. J Biol Chem. 2007;282: 12410–12418. doi:10.1074/jbc.M608748200
56. Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZXP, Pool JE, et al. Sequencing of Fifty Human Exomes Reveals Adaptation to High Altitude. Science (80-). 2013;329: 75–78. doi:10.1126/science.1190371.Sequencing
57. Rottgardt I, Rothhammer F, Dittmar M. Native highland and lowland populations differ in γ-globin gene promoter polymorphisms related to altered fetal hemoglobin levels and delayed fetal to adult globin switch after birth. Anthropol Sci. 2010;118: 41–48. doi:10.1537/ase.090402
58. Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZXP, Pool JE, et al. Sequencing of 50 Human Exomes Reveals Adaptation to High Altitude. Science (80-). 2010;329: 75–78. doi:10.1126/science.1190371
59. Bigham A, Bauchet M, Pinto D, Mao X, Akey JM, Mei R, et al. Identifying Signatures of Natural Selection in Tibetan and Andean Populations Using Dense Genome Scan Data. Begun DJ, editor. PLoS Genet. 2010;6: e1001116. doi:10.1371/journal.pgen.1001116
60. Ségurel L, Thompson EE, Flutre T, Lovstad J, Venkat A, Margulis SW, et al. The ABO blood group is a trans-species polymorphism in primates. Proc Natl Acad Sci. 2012;109: 18493–18498. doi:10.1073/pnas.1210603109
61. Hirota T, Takahashi A, Kubo M, Tsunoda T, Tomita K, Sakashita M, et al. Genome-wide association study identifies eight new susceptibility loci for atopic dermatitis in the Japanese population. Nat Genet. 2012;44: 1222–1226. doi:10.1038/ng.2438
62. Lenormand C, Bausinger H, Gross F, Signorino-Gelo F, Koch S, Peressin M, et al. HLA-DQA2 and HLA-DQB2 genes are specifically expressed in human Langerhans cells and encode a new HLA class II molecule. J Immunol. 2012;188: 3903–3911. doi:10.4049/jimmunol.1103048
63. Jiang D-K, Ma X-P, Yu H, Cao G, Ding D-L, Chen H, et al. Genetic variants in five novel loci including CFB and CD40 predispose to chronic hepatitis B. Hepatology. 2015;62: 118–128. doi:10.1002/hep.27794
64. Lee YH, Bae S-C, Choi SJ, Ji JD, Song GG. Genome-wide pathway analysis of genome-wide association studies on systemic lupus erythematosus and rheumatoid arthritis. Mol Biol Rep. 2012;39: 10627–10635. doi:10.1007/s11033-012-1952-x
65. Cagliani R, Riva S, Biasin M, Fumagalli M, Pozzoli U, Lo Caputo S, et al. Genetic diversity at endoplasmic reticulum aminopeptidases is maintained by balancing selection and is associated with natural resistance to HIV-1 infection. Hum Mol Genet. 2010;19: 4705–14. doi:10.1093/hmg/ddq401
66. Andrés AM, Dennis MY, Kretzschmar WW, Cannons JL, Lee-Lin S-Q, Hurle B, et al. Balancing Selection Maintains a Form of ERAP2 that Undergoes Nonsense-Mediated Decay and Affects Antigen Presentation. PLoS Genet. 2010;6: e1001157. doi:10.1371/journal.pgen.1001157
67. Evans DM, Spencer CCA, Pointon JJ, Su Z, Harvey D, Kochan G, et al. Interaction between ERAP1 and HLA-B27 in ankylosing spondylitis implicates peptide handling in the mechanism for HLA-B27 in disease susceptibility. Nat Genet. 2011;43: 761–767. doi:10.1038/ng.873
68. Genetic Analysis of Psoriasis Consortium & the Wellcome Trust Case Control Consortium 2, Strange A, Capon F, Spencer CCA, Knight J, Weale ME, et al. A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1. Nat Genet. 2010;42: 985–90.

doi:10.1038/ng.694
69. Charlesworth B. Effective population size and patterns of molecular evolution and variation. Nat Rev Genet. 2009;10: 195–205. doi:10.1038/nrg2526
70. Corona E, Dudley JT, Butte AJ. Extreme evolutionary disparities seen in positive selection across seven complex diseases. PLoS One. 2010;5: 1–10. doi:10.1371/journal.pone.0012236
71. Wooding S, Kim U, Bamshad MJ, Larsen J, Jorde LB, Drayna D. Natural Selection and Molecular Evolution in PTC, a Bitter-Taste Receptor Gene. Am J Hum Genet. 2004;74: 637–646. doi:10.1086/383092
72. Wooding S, Bufe B, Grassi C, Howard MT, Stone AC, Vazquez M, et al. Independent evolution of bitter-taste sensitivity in humans and chimpanzees. Nature. 2006;440: 930–934. doi:10.1038/nature04655
73. Ledda M, Kutalik Z, Souza Destito MC, Souza MM, Cirillo CA, Zamboni A, et al. GWAS of human bitter taste perception identifies new loci and reveals additional complexity of bitter taste genetics. Hum Mol Genet. 2014;23: 259–267. doi:10.1093/hmg/ddt404
74. Thalmann S, Behrens M, Meyerhof W. Major haplotypes of the human bitter taste receptor TAS2R41 encode functional receptors for chloramphenicol. Biochem Biophys Res Commun. 2013;435: 267–273. doi:10.1016/j.bbrc.2013.04.066
75. Meyerhof W, Batram C, Kuhn C, Brockhoff A, Chudoba E, Bufe B, et al. The Molecular Receptive Ranges of Human TAS2R Bitter Taste Receptors. Chem Senses. 2010;35: 157–170. doi:10.1093/chemse/bjp092
76. Karaman R, Nowak S, Di Pizio A, Kitaneh H, Abu-Jaish A, Meyerhof W, et al. Probing the Binding Pocket of the Broadly Tuned Human Bitter Taste Receptor TAS2R14 by Chemical Modification of Cognate Agonists. Chem Biol Drug Des. 2016;88: 66–75. doi:10.1111/cbdd.12734
77. Behrens M, Brockhoff A, Kuhn C, Bufe B, Winnig M, Meyerhof W. The human taste receptor hTAS2R14 responds to a variety of different bitter compounds. Biochem Biophys Res Commun. 2004;319: 479–485. doi:10.1016/j.bbrc.2004.05.019
78. Kosiol C, Vinar T, da Fonseca RR, Hubisz MJ, Bustamante CD, Nielsen R, et al. Patterns of positive selection in six Mammalian genomes. PLoS Genet. 2008;4: 1–17. doi:10.1371/journal.pgen.1000144
79. Arking DE, Krebsova A, Macek M, Macek M, Arking A, Mian IS, et al. Association of human aging with a functional variant of klotho. Proc Natl Acad Sci. 2002;99: 856–861. doi:10.1073/pnas.022484299
80. Welberg L. Cognition: Klotho spins cognitive fate. Nat Rev Neurosci. 2014;15: 425–425. doi:10.1038/nrn3777
81. Dubal DB, Yokoyama JS, Zhu L, Broestl L, Worden K, Wang D, et al. Life Extension Factor Klotho Enhances Cognition. Cell Rep. The Authors; 2014;7: 1065–1076. doi:10.1016/j.celrep.2014.03.076
82. Lui JH, Nowakowski TJ, Pollen AA, Javaherian A, Kriegstein AR, Oldham MC. Radial glia require PDGFD–PDGFRβ signalling in human but not mouse neocortex. Nature. 2014;515: 264–268. doi:10.1038/nature13973
83. Rakic P. Evolution of the neocortex: a perspective from developmental biology. Nat Rev Neurosci. 2009;10: 724–735. doi:10.1038/nrn2719
84. St Pourcain B, Cents RAM, Whitehouse AJO, Haworth CMA, Davis OSP, O'Reilly PF, et al. Common variation near ROBO2 is associated with expressive vocabulary in infancy. Nat Commun. 2014;5: 4831. doi:10.1038/ncomms5831
85. Peyrégne S, Dannemann M, Prüfer K. Detecting ancient positive selection in humans using extended lineage sorting. BiorXiv. 2016; doi:10.1101/092999
86. Sundermann EE, Maki PM, Bishop JR. A review of estrogen receptor α gene (ESR1) polymorphisms, mood, and cognition. Menopause. 2010;17: 874–886. doi:10.1097/gme.0b013e3181df4a19
87. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 2015;43: D447–D452. doi:10.1093/nar/gku1003
88. Perlis RH, Huang J, Purcell S, Fava M, Rush AJ, Sullivan PF, et al. Genome-Wide Association Study of Suicide Attempts in Mood Disorder Patients. Am J Psychiatry. 2010;167: 1499–1507. doi:10.1176/appi.ajp.2010.10040541
89. Mullins N, Perroud N, Uher R, Butler AW, Cohen-Woods S, Rivera M, et al. Genetic relationships between suicide attempts, suicidal ideation and major psychiatric disorders: a genome-wide association and polygenic scoring study. Am J Med Genet B Neuropsychiatr Genet. 2014;165B: 428–37. doi:10.1002/ajmg.b.32247
90. Shatz CJ. MHC Class I: An Unexpected Role in Neuronal Plasticity. Neuron. 2009;64: 40–45. doi:10.1016/j.neuron.2009.09.044
91. Needleman LA, McAllister AK. The major histocompatibility complex and autism spectrum disorder. Dev Neurobiol. 2012;72: 1288–1301. doi:10.1002/dneu.22046
92. Torres AR, Westover JB, Rosenspire AJ. HLA Immune Function Genes in Autism. Autism Res Treat. 2012;2012: 1–13. doi:10.1155/2012/959073
93. Careaga M, Water J, Ashwood P. Immune dysfunction in autism: A pathway to treatment. Neurotherapeutics. 2010;7: 283–292. doi:10.1016/j.nurt.2010.05.003
94. Gow A, Southwood CM, Li JS, Pariali M, Riordan GP, Brodie SE, et al. CNS myelin and sertoli cell tight junction strands are absent in Osp/claudin-11 null mice. Cell. 1999;99: 649–59. Available: https://www.ncbi.nlm.nih.gov/pubmed/10612400/
95. Wu X, Peppi M, Vengalil MJ, Maheras KJ, Southwood CM, Bradley M, et al. Transgene-Mediated Rescue of Spermatogenesis in Cldn11-Null Mice1. Biol Reprod. 2012;86. doi:10.1095/biolreprod.111.096230

96. Quaynor SD, Stradtman EW, Kim H-G, Shen Y, Chorich LP, Schreihofer DA, et al. Delayed puberty and estrogen resistance in a woman with estrogen receptor α variant. N Engl J Med. 2013;369: 164–71. doi:10.1056/NEJMoa1303611
97. Michailidou K, Hall P, Gonzalez-Neira A, Ghoussaini M, Dennis J, Milne RL, et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. Nat Genet. 2013;45: 353–61, 361–2. doi:10.1038/ng.2563
98. Eriksson N, Benton GM, Do CB, Kiefer AK, Mountain JL, Hinds DA, et al. Genetic variants associated with breast size also influence breast cancer risk. BMC Med Genet. 2012;13: 53. doi:10.1186/1471-2350-13-53
99. Lazari MFM, Lucas TFG, Yasuhara F, Gomes GRO, Siu ER, Royer C, et al. Estrogen receptors and function in the male reproductive system. Arq Bras Endocrinol Metabol. 2009;53: 923–933.
100. Rivadeneira F, Styrkársdottir U, Estrada K, Halldórsson B V, Hsu Y-H, Richards JB, et al. Twenty bone-mineral-density loci identified by large-scale meta-analysis of genome-wide association studies. Nat Genet. 2009;41: 1199–1206. doi:10.1038/ng.446
101. Aouizerat BE, Vittinghoff E, Musone SL, Pawlikowska L, Kwok P-Y, Olgin JE, et al. GWAS for discovery and replication of genetic loci associated with sudden cardiac arrest in patients with coronary artery disease. BMC Cardiovasc Disord. 2011;11: 29. doi:10.1186/1471-2261-11-29
102. Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. Nat Genet. 2014;46: 1173–1186. doi:10.1038/ng.3097
103. Heldring N, Pike A, Andersson S, Matthews J, Cheng G, Treuter E, et al. Estrogen Receptors : How Do They Signal and What Are Their Targets. Physiol Rev. 2007;87: 905–931. doi:10.1152/physrev.00026.2006.
104. Ewing G, Hermisson J. MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. Bioinformatics. 2010;26: 2064–2065. doi:10.1093/bioinformatics/btq322
105. Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, et al. Demographic history and rare allele sharing among human populations. Proc Natl Acad Sci U S A. 2011;108: 11983–8. doi:10.1073/pnas.1019276108
106. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. Genetics Soc America; 1989;123: 585–595.
107. Key FM, Peter B, Dennis MY, Huerta-Sánchez E, Tang W, Prokunina-Olsson L, et al. Selection on a Variant Associated with Improved Viral Clearance Drives Local, Adaptive Pseudogenization of Interferon Lambda 4 (IFNL4). PLoS Genet. 2014;10: e1004681. doi:10.1371/journal.pgen.1004681
108. Derrien T, Estellé J, Marco Sola S, Knowles DG, Raineri E, Guigó R, et al. Fast Computation and Applications of Genome Mappability. Ouzounis CA, editor. PLoS One. 2012;7: e30377. doi:10.1371/journal.pone.0030377
109. Cheng Z, Ventura M, She X, Khaitovich P, Graves T, Osoegawa K, et al. A genome-wide comparison of recent chimpanzee and human segmental duplications. Nature. 2005;437: 88–93. doi:10.1038/nature04000
110. Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F, et al. Personalized copy number and segmental duplication maps using next-generation sequencing. Nat Genet. 2009;41: 1061–1067. doi:10.1038/ng.437
111. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999;27: 573–580.
112. Meyer M, Kircher M, Gansauge M-T, Li H, Racimo F, Mallick S, et al. A High-Coverage Genome Sequence from an Archaic Denisovan Individual. Science (80-). 2012;338: 222–226. doi:10.1126/science.1224344
113. Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, et al. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature. 2013;505: 43–49. doi:10.1038/nature12886
114. Kofler R, Schlötterer C. Gowinda: Unbiased analysis of gene set enrichment for genome-wide association studies. Bioinformatics. 2012;28: 2084–2085. doi:10.1093/bioinformatics/bts315
115. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, et al. The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. Genome Res. 2012;22: 1775–1789. doi:10.1101/gr.132159.111
116. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11: R106. doi:10.1186/gb-2010-11-10-r106
117. Sankararaman S, Mallick S, Dannemann M, Prüfer K, Kelso J, Pääbo S, et al. The genomic landscape of Neanderthal ancestry in present-day humans. Nature. 2014;507: 354–7. doi:10.1038/nature12961
118. Development Core Team R. R: A language and environment for statistical computing. [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2009. Available: http://www.r-project.org
119. Kuleshov M V., Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 2016;44: W90–W97. doi:10.1093/nar/gkw377
120. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 2014;42: D1001–D1006. doi:10.1093/nar/gkt1229

Figure Captions

Fig 1. A schematic representation of site frequency spectra (SFS) under neutrality and selection, which motivates the NCD statistic. (**A**) Unfolded SFS (ranging from 0 to 1) of derived allele frequencies (DAF) for loci under neutrality (grey) or containing one site under balancing selection with frequency equilibrium (f_{eq}) of 0.5 (blue), 0.4 (orange) and 0.3 (pink). (**B**) Folded SFS (ranging from 0 to 0.5) for minor allele frequencies (MAF). Colors as in **A**. (**C**) Distribution of *NCD* (Non-Central Deviation) expected under neutrality (grey) and under selection assuming $tf = f_{eq}$. Colors as in **A.** *x*-axis shows minimum and maximum values that *NCD* can have for a given *tf* value.

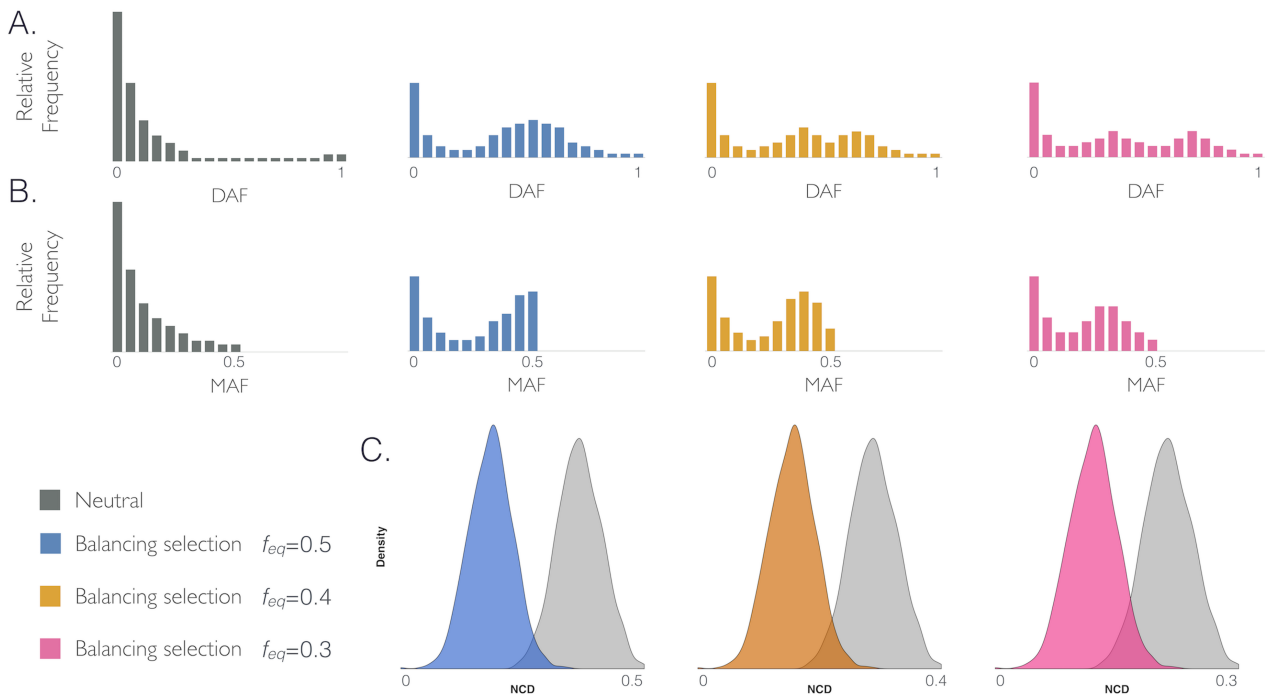

Fig 2. Power to detect balancing selection for *NCD(0.5)* and other tests. The ROC curves summarize the power (in function of the false positive rate) to detect LTBS for simulations where the balanced polymorphism was modeled to achieve f_{eq} of **(A)** 0.3, **(B)** 0.4, and **(C)** 0.5. Plotted values are for the African demography, *Tbs* = 5 mya. L = 3 kb, except for T_1 and T_2 where L = 100 ISs, following [31] (see Methods). For *NCD* calculations, $tf = f_{eq}$. European demography yields similar results (S10 Fig).

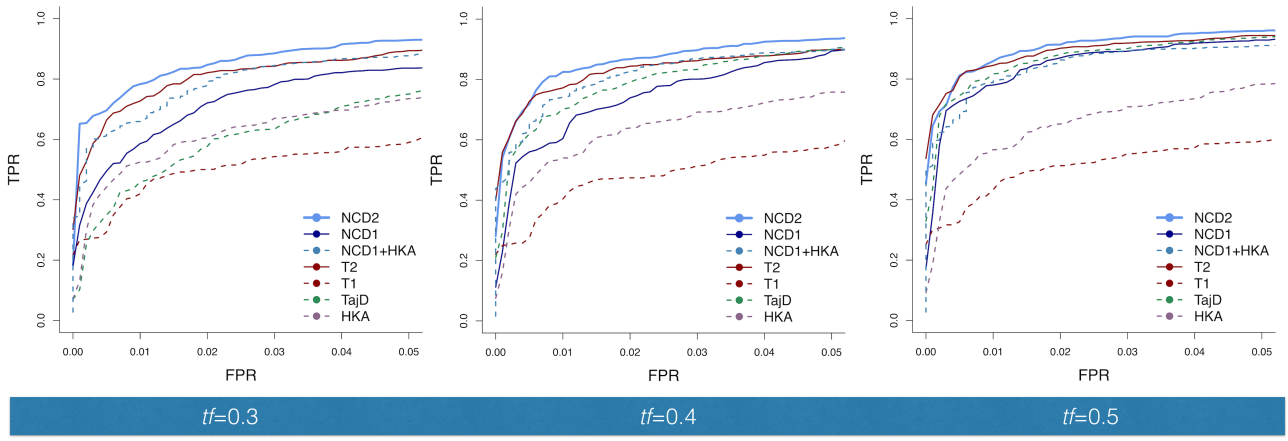

Fig 3. Polymorphism-to-divergence and SFS. (A,B) SNP/(FD+1) for LWK **(A)** and GBR **(B)** populations. SNP/(FD+1) measures the proportion of polymorphic-to-divergent sites for the union of significant windows for all *tf* (purple, green) compared to all scanned windows (gray). **(C-D)** SFS in LWK **(C)** and GBR **(D)** of all scanned windows in chr1 (gray), significant windows for *NCD2(0.5)* (blue), *NCD2(0.4)* (orange), *NCD2(0.3)* (pink). DAF, derived allele frequency.

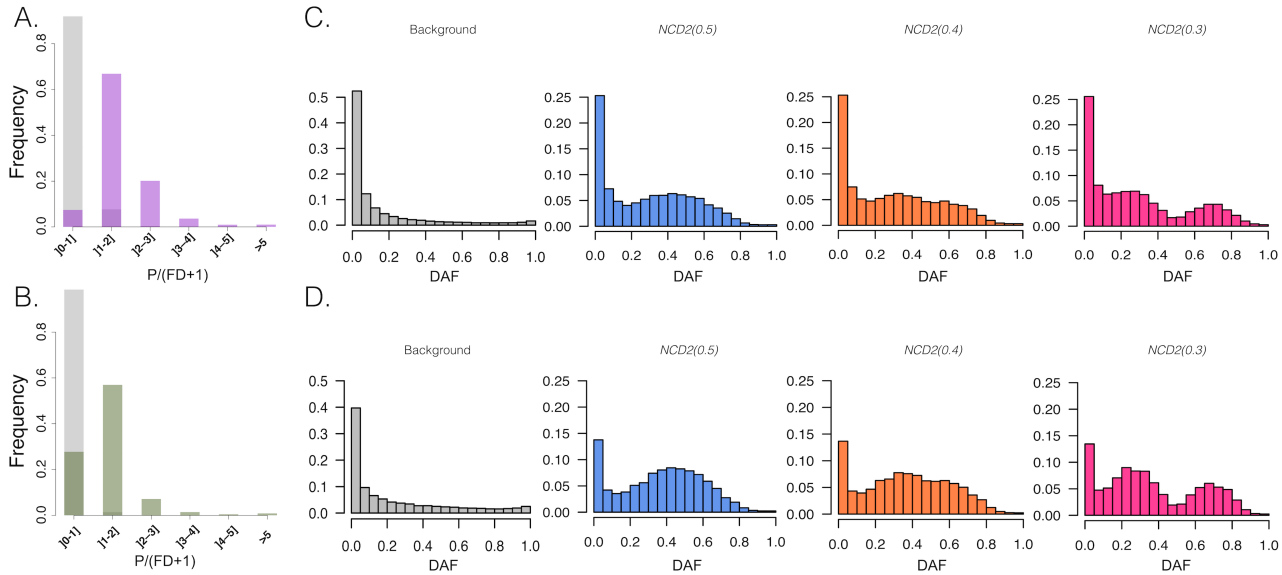

Fig 4. Manhattan plot and population sharing. (A) Manhattan plot of all scanned windows, for one analysis (*NCD2(0.5)* and LWK). *y*-axis, *p*-value (log-scale) based on $Z_{tf\text{-}IS}$. *x*-axis, ordered location of analyzed windows on the genome. Each point is a scanned (grey and black), significant (blue) or outlier (pink) window. Names of outlier protein-coding genes are provided, sorted by name. We note that significant windows were defined based of simulations, not on $Z_{tf\text{-}IS}$. ($Z_{tf\text{-}IS}$ is used to rank even those with $p < 0.0001$) **(B)** Venn diagram showing the overlap in signatures of the 167 outlier genes annotated in **(A)** with other populations.

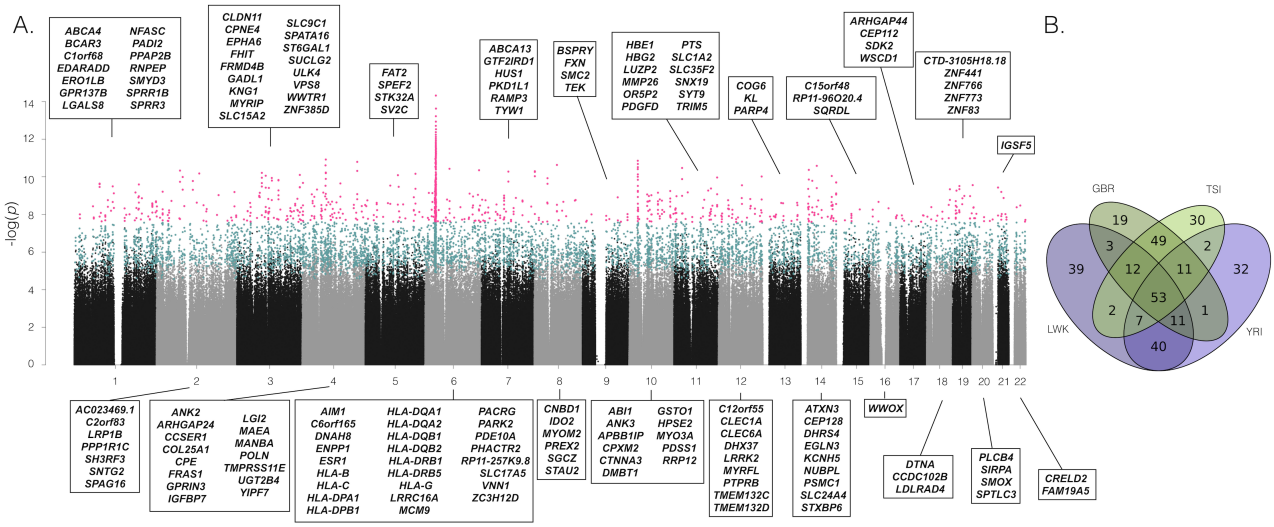
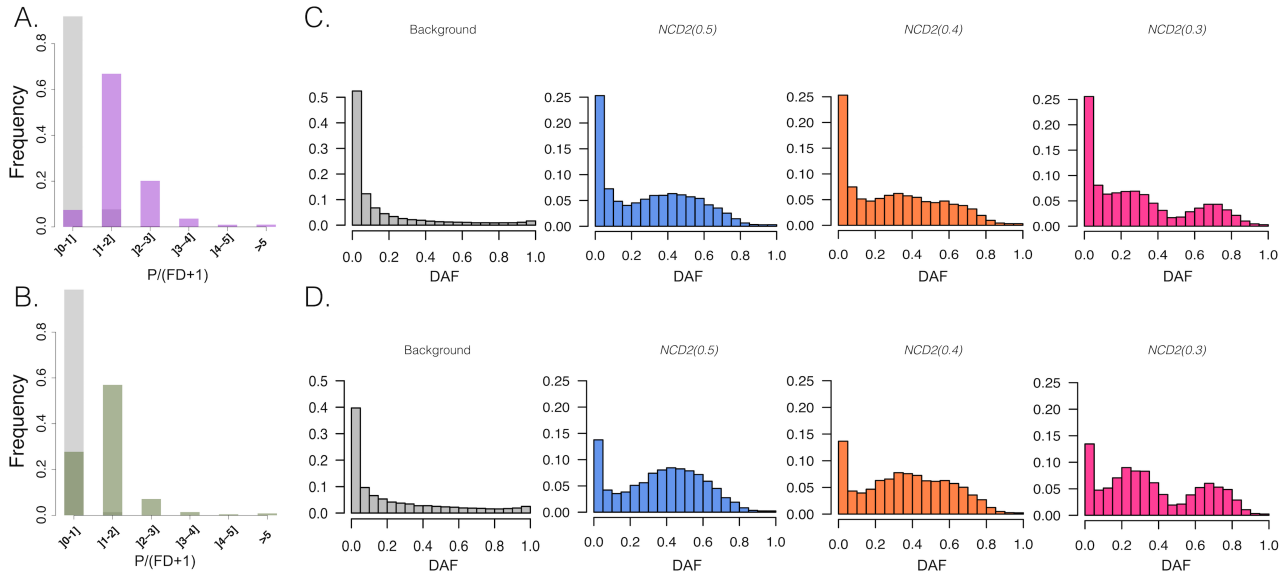
Fig 5. Enrichment of classes of sites amongst candidate windows. Dashed lines mark the p = 0.975 (bottom) and p = 0.025 (top) thresholds for the one-tailed p-values (hypothesis: enrichment). NSyn, nonsynonymous; all, Genic plus Intergenic plus Regulatory. The annotation is based on Ensembl variant predictor (S6 Note). p< 0.001 was treated as 0.001 to avoid infinite values.