

1 **Original Manuscript**

2
3

4 **Signatures of long-term balancing selection in human genomes**

5
6

7 Bárbara Domingues Bitarello^{1,2,*}

8 Cesare de Filippo²

9 João Carlos Teixeira^{2,3}

10 Joshua M. Schmidt²

11 Philip Kleinert^{2,4}

12 Diogo Meyer^{1, ¶,*}

13 Aida M. Andrés^{2, ¶,*}

14

15 1, Department of Genetics and Evolutionary Biology, University of São Paulo, São Paulo, Brazil

16

17 2, Genetics Department, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

18

19 3, Current address: Unit of Human Evolutionary Genetics, Institut Pasteur, Paris, France

20

21 4, Current address: Computational Molecular Biology Department, Max Planck Institute for

22 Molecular Genetics, Berlin, Germany

23 ¶, AA and DM co-supervised the study

24

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

26 aida_andres@eva.mpg.de (AMA).

1 **Abstract**

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

24

25 **Author Summary**

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

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

6

7 **Introduction**

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

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

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

4 While balancing selection has been extensively explored from a theoretical perspective, an
5 empirical understanding of its prevalence lags behind our knowledge of positive selection. This
6 stems from technical difficulties in detecting balancing selection, as well as the perception that it
7 may be a rare selective process [26]. In fact, few methods have been developed to identify its
8 targets, and only a handful of studies have sought to uncover them genome-wide in humans
9 [23,24,27–32]. Different approaches have been used to identify genes [28] or genomic regions [31]
10 with an excess of polymorphisms and intermediate frequency alleles, while other studies have
11 identified trans-species polymorphisms between humans and their closest living relatives
12 (chimpanzees and bonobos) [23,24]. Overall, these studies suggested that balancing selection may
13 act on a small portion of the genome, although the limited extent of data available (e.g., exome data
14 [31], small sample size [28]), and stringency of the criteria (e.g., balanced polymorphisms predating
15 human-chimpanzee divergence [23,24]) may underlie the paucity of detected regions.

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

23

24 **Results**

25 **The Non-Central Deviation (NCD) statistic**

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

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

22

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

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

$$5 \quad NCD(tf) = \sqrt{\frac{\sum_{i=1}^n (p_i - tf)^2}{n}} \quad (1)$$

6 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
7 number of informative sites, and tf is the target frequency with respect to which the deviations of
8 the observed alleles frequencies are computed. Thus, *NCD* is a type of standard deviation that
9 quantifies the dispersion of allelic frequencies from tf , rather than from the mean of the distribution.
10 Low *NCD* values reflect a low deviation of the SFS from a pre-defined tf , as expected under LTBS
11 (Fig 1C and S1 Note).

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

21

22 **Power of NCD to detect LTBS**

1 We evaluated the sensitivity and specificity of *NCD1* and *NCD2* by benchmarking their
2 performance using simulations. Specifically, we considered demographic scenarios inferred for
3 African, European, and Asian human populations, and simulated sequences evolving both under
4 neutrality and LTBS using an overdominance model. We explored the influence of parameters that
5 can affect the power of NCD statistics: time since onset of balancing selection (*Tbs*), frequency
6 equilibrium defined by selection coefficients (f_{eq}), demographic history of the sampled population,
7 *tf* used in *NCD* calculation, length of the genomic region analyzed (*L*) and implementation (*NCD1*
8 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)*.

9

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

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

5

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

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

20

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

26

1 **Simulated and target frequencies.** So far, we have only discussed cases where $tf = f_{eq}$, which is
 2 expected to favor the performance of *NCD*. Accordingly, under this condition *NCD* has high power:
 3 0.91, 0.85, and 0.79 on average for $f_{eq} = 0.5, 0.4,$ and $0.3,$ respectively (averaged across *Tbs* and
 4 populations, Table 1). However, since in practice there is no prior knowledge about the f_{eq} of
 5 balanced polymorphisms, we evaluate the power of *NCD* when $f_{eq} \neq tf$. When $f_{eq} = 0.5,$ average
 6 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
 7 are observed for other simulated f_{eq} (Table 1). Therefore, *NCD* statistics are overall well-powered
 8 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
 9 to f_{eq} , the higher the power, so when possible, it is desirable to perform tests across a range of tf .

10

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

12

		Africa						Europe					
<i>Tbs</i>	f_{eq}	<i>NCD2</i>			<i>NCD1</i>			<i>NCD2</i>			<i>NCD1</i>		
		<i>tf</i>			<i>tf</i>			<i>tf</i>			<i>tf</i>		
		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

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

16

17 ***NCD* implementations and comparison to other methods.** Power for *NCD2* is greater than for
 18 *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
 19 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}
 20 = 0.3 (0.86 for *NCD2*(0.3) vs. 0.73 for *NCD1*(0.3)) (Table 1, Fig 2). This illustrates the gain in
 21 power by incorporating FDs in the *NCD* statistic, which is also more powerful than combining
 22 *NCD1* and HKA (S1 Table).

1 We compared the power of *NCD* to two statistics commonly used to detect balancing
2 selection (Taj D and HKA), a composite statistic of *NCDI* and HKA (with the goal of quantifying
3 the contribution of FD to *NCD* power), and a pair of composite likelihood-based measures (T_1 and
4 T_2 [31]). The T_2 statistic, similarly to *NCD2*, considers both the SFS and the ratio of polymorphisms
5 to FD. Power results are summarized in Fig 2. When $f_{\text{eq}} = 0.5$, *NCD2(0.5)* has the highest power: for
6 example, in Africa ($Tbs = 5$ myr, and 3 kb) *NCD2(0.5)* power is 0.96 (the highest among other tests
7 is 0.94, for T_2) but the difference in power is highest when f_{eq} departs from 0.5. For $f_{\text{eq}} = 0.4$,
8 *NCD2(0.4)* power is 0.93 (compared to 0.90 for Taj D and T_2 , and lower for the other tests). For $f_{\text{eq}} =$
9 0.3, *NCD2(0.3)* power is 0.93 (compared to 0.89 for T_2 and lower for the other tests). These patterns
10 are consistent in the African and European simulations (Fig 2, S10 Fig), where *NCD2* has greater or
11 comparable power to detect LTBS than other available methods. When focusing on the tests that
12 use only polymorphic sites, *NCDI* has similar power to Tajima's D when $f_{\text{eq}} = 0.5$, and it
13 outperforms it when f_{eq} departs from 0.5 (Table S1). Altogether, the advantage of *NCD2* over
14 classic neutrality tests is its high power, especially when f_{eq} departs from 0.5; the advantage over T_2
15 is its simplicity of implementation and interpretation, and the fact that it can be run in the absence
16 of a demographic model.

17
18 **Recommendations based on power analyses.** Overall, *NCD* performs very well in regions of 3 kb
19 (Table 1, Fig 2) and similarly for African and European demographic scenarios. In fact, *NCD2*
20 outperforms all other methods tested (Fig 2, S10 Fig) and reaches very high power when $tf = f_{\text{eq}}$
21 (always > 0.89 for 5 mya and always > 0.79 for 3 mya). While the f_{eq} of a putatively balanced allele
22 is unknown, the simplicity of the *NCD* statistics makes it trivial to run for several tf values, allowing
23 detection of balancing selection for a range of equilibrium frequencies. Also, the analysis can be run
24 in sliding windows to ensure overlap with the narrow signatures of balancing selection.
25 Alternatively, *NCD* could also be computed for 3kb windows centered in each SNP or IS. Because
26 *NCD2* outperforms *NCDI*, we used it for our scan of human populations; *NCDI* is nevertheless a
27 good choice when outgroup data is lacking.

1 **Identifying signatures of LTBS**

2 We aimed to identify regions of the human genome under LTBS. We chose $NCD2(0.5)$, $NCD2(0.4)$
3 and $NCD2(0.3)$, which provide sets of candidate windows that are not fully overlapping (Table 1).
4 We calculated the statistics for 3 kb windows (1.5 kb step size) and tested for significance using
5 two complementary approaches: one testing all windows with respect to neutral expectations, and
6 one identifying outlier windows in the empirical genomic distribution. We analyzed genome-wide
7 data from two African (YRI: Yoruba in Ibadan, Nigeria; LWK: Luhya in Webuye, Kenya) and two
8 European populations (GBR: British, England and Scotland; TSI: Toscani, Italy) [40]. We filtered
9 the data for orthology with the chimpanzee genome (used as the outgroup) and implemented 5
10 additional filters to avoid technical artifacts (S13 Fig). Finally, we excluded windows with less than
11 10 IS in any of the populations since these showed a high variance in $NCD2$ due to noisy SFS (see
12 empirical patterns in S18 Fig and neutral simulation patterns in S11 Fig).

13
14 **Simulation and empirical-based sets of windows.** After all filters were implemented, we analyzed
15 1,657,989 windows (~ 81% of the autosomal genome; S13 Fig), overlapping 18,633 protein-coding
16 genes. We defined a p -value for each window as the quantile of its $NCD2$ value when compared to
17 those from 10,000 neutral simulations under the inferred demographic history of each population
18 and conditioned on the same number of IS. Depending on the population, between 6,226 and 6,854
19 (0.37-0.41%) of the scanned windows have a lower $NCD2(0.5)$ value than any of the 10,000 neutral
20 simulations ($p < 0.0001$). The proportions are similar for $NCD2(0.4)$ (0.40-0.45%) and $NCD2(0.3)$
21 (0.33-0.38%) (Table 2). We refer to these sets, whose patterns cannot be explained by the neutral
22 model, as the *significant* windows. In each population, the union of significant windows
23 considering all tf values spans, on average, 0.6% of the windows (Table 2) and 0.77% of the base
24 pairs.

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

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

7

8 **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

9 Significant and outlier genes and windows, see main text. U, union of windows considering three tf values.

10 Total number of queried windows per population is 1,657,989. Union of all candidate genes is 2,348
 11 (significant) and 402 (outlier).

12

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

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

13

14 **Assigned tf values**

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

21

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

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

5

6 **Biological pathways influenced by LTBS**

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

11 Regarding functional categories, significant genes are overrepresented in 28 GO categories,
12 24 of which are shared by at least two populations and 18 by four populations. Thirteen categories
13 are immune-related according to a list of 386 immune-related keywords from ImmPort (Methods).
14 The more stringent sets of outlier genes are enriched for 28 GO categories (21 shared by all four
15 populations), 18 of which are immune-related. Furthermore, in both sets several of the remaining
16 enriched categories are directly related to antigen presentation although not classified as immune-
17 related (e.g., “ER to golgi transport vesicle membrane”, “integral to membrane”). Among the non
18 immune-related categories are “sarcolemma”, “epidermis development”, “keratin filament” and
19 “negative regulation of blood coagulation” (S2 Table).

20 When classical HLA genes are removed from the analyses, only two categories remain
21 enriched: “sarcolemma” (in YRI) and “epidermis development” (GBR), but the small set of genes
22 per population hampers power. For the significant windows, “antigen processing and presentation
23 of endogenous peptide antigen via MHC class I” remains significantly enriched (driven by *TAP1*,
24 *TAP2*, *ERAP1* and *ERAP2*; S2 Table). Also, significant windows are still enriched in categories
25 related to the extracellular space – “extracellular regions”, “integral to membrane” (as in
26 [15,28,31]) – and “keratin filament”. These categories are not immune-related *per se*, but they

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

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

8

9 **Functional annotation of SNPs in candidate windows**

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

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

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

1 excess of exonic SNPs ($p \leq 0.003$). The same is true for significant windows ($p \leq 0.016$; Fig 5).
2 When only nonsynonymous and regulatory sites are considered, we see enrichment for LWK and
3 YRI for the outlier windows ($p \leq 0.036$, Fig 5) but not for the significant windows ($p \geq 0.458$ for all
4 populations, Fig 5), although the two analyses that consider nonsynonymous SNPs are likely
5 underpowered due to low SNP counts (S8 Table). Finally, results using more detailed RegulomeDB
6 annotations generally agree with the observation of lack of enrichment of regulatory sites in our
7 candidate windows ($p \geq 0.121$ for a one tail test for enrichment for RegulomeDB1+3 for SNPs with
8 $MAF \geq 0.2$) (S6 Note, S8 Table).

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

14

15 **Monoallelic expression.** Genes with mono-allelic expression (MAE) – i.e, the random and
16 mitotically stable choice of an active allele for a given locus – have been found to be enriched
17 among those with signatures of balancing selection [46]. Our observations agree with this. For
18 example, 64% and 62% of the outlier and significant genes shared by at least two populations have
19 MAE status according to [46], compared to only 41% for genes without signatures of LTBS ($p <$
20 1.12^{-6} Fisher Exact Test, one-sided).

21

22 **Overlap across populations**

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

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

5
6 **Candidate genes in more than one population.** Instances where signatures of LTBS are not
7 shared between populations may result from changes in selective pressure, which may be important
8 during fast, local adaptation [47]. Still, loci with signatures of LTBS across human populations are
9 more likely to represent stable selection. We considered as “shared” those candidate protein-coding
10 genes (from the union of candidate windows for all tf) that are shared by all populations (S6 Table).
11 For the rest, we considered as “African” those shared between YRI and LWK (but neither or only
12 one European population), and “European” those shared between GBR and TSI (but neither or only
13 one African population). We note that these designations do not imply that genes referred to as
14 “African” or “European” are putative targets of LTBS for only one continent (partially because
15 there are some power differences between Africa and Europe, Table 1). The 79 African, 84
16 European and 102 shared” outlier genes add up to 265 genes in total (~1.5% of all queried genes)
17 and the 458 African, 400 European, and 736 shared significant genes add up to 1,594 (~8.5% of all
18 queried genes; S6 Table).

19

20 **Discussion**

21 **The targets of LTBS in the human genome**

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

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

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

11

12 **Functional properties of SNPs in candidate windows**

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

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

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

3

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

14

15 **Properties of candidate genes**

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

18

19 **Mono-allelic expression.** In agreement with previous findings, we found a significant excess of
20 MAE genes among our outlier candidates. This excess is not driven by HLA genes, which were
21 filtered out in the study originally reporting MAE genes and supports the claim for a biological link
22 between MAE and balancing selection [46]. Heterozygosity in a MAE gene could lead to cell-to-
23 cell heterogeneity within same-cell clusters, which could in turn be potentially advantageous
24 [46,52]), particularly in the case of cell-surface proteins. Some of these MAE genes found in our
25 study, and not previously detected in scans for balancing selection, are involved in
26 immunity/defense barriers (e.g. *IL1RL1*, *IL18R1*, *FAM114A1*, *EDARADD*, *SIRPA*, *TAS2R14*),

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

3

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

17

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

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

1 [62] and is associated both to infectious (hepatitis B) and autoimmune diseases (e.g. lupus [63,64]).
2 Two other significant genes for which there is prior evidence for LTBS [65,66], *ERAPI* and
3 *ERAP2*, are associated with ankylosing spondylitis and psoriasis (e.g [67–69]). Finally, there are
4 several associations to autoimmune disease and susceptibility to infections in the classical HLA
5 genes that we identify. In brief, our results are consistent with the hypothesis that auto-immune
6 disease is linked to natural selection favoring effective immune response against pathogens [9,70].

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

17
18 **Cognition.** Interestingly, several candidate genes are involved in cognitive abilities, or their
19 variation is associated with diversity in related phenotypes. The *KL* (life extension factor klotho) is
20 a gene that has been associated to human longevity [79] and for which signatures of LTBS have
21 been previously reported [31]. In mice, decreased levels of klotho shorten lifespan (reviewed in
22 [80]). In humans, heterozygotes for the *KL-VS* variant show higher levels of serum klotho and
23 enhanced cognition, independent of sex and age, than wild-type homozygotes. On the other hand,
24 *KL-VS* homozygotes show decreased lifespan and reduced cognition [81]. If higher cognition is
25 advantageous, overdominance for this phenotype can explain the signatures of balancing selection
26 we observe (although klotho's the effect in lifespan can also influence).

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

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

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

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

25

26 **Reproduction**

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

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

23

24 **Conclusions**

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

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

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

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

21

22 **Methods**

23

24 **Simulations and power analyses**

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

1 data under the following demographic model: generation time of 25 years, mutation rate of 2.5×10^{-8}
2 per site and recombination rate of 1×10^{-8} , and a human-chimpanzee split at 6.5 mya was added to
3 the model, which was used to obtain the neutral distributions for the NCD statistics. For the
4 simulations with selection, a balanced polymorphism was added to the center of the simulated
5 sequence and modeled to achieve a pre-specified frequency equilibrium ($f_{eq} = 0.3, 0.4, 0.5$)
6 following an overdominant model (S2 Note). Simulations with and without selection were run for
7 different sequence lengths (3, 6, 12 kb) and times of onset of balancing selection (1, 3, 5 mya). For
8 each combination of parameters, 1,000 simulations, with and without selection, were used to
9 compare the relationship between true (TPR, the power of the statistic) and false (FPR) positive
10 rates for the NCD statistics, represented by ROC curves. For performance comparisons, we used
11 $FPR = 0.05$. When comparing performance under a given condition, power was averaged across
12 *NCD* implementations, demographic scenarios, L , and Tbs . When comparing NCD performance to
13 other methods (Tajima's D [106], HKA [34], and a combined *NCDI*+HKA test), we simulated
14 under *NCD* optimal conditions: $L = 3$ kb and $Tbs = 5$ mya (S1 Table). Since power for T_1 and T_2 is
15 reported based on windows of 100 informative sites (~ 14 kb for YRI and CEU) up and
16 downstream of the target site [31], we divided simulations of 15 kb into windows of 100 IS,
17 calculated T_1 and T_2 with BALLET [31] and selected the highest T_1 or T_2 value from each
18 simulation to obtain their power for the same set of parameters used for the other simulations.

19

20 **Human population genetic data**

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

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

9

10 **Identifying signatures of LTBS**

11 After applying all filters and requiring the presence of at least one informative site, *NCD2* was
12 computed for 1,695,655 windows per population. Because in simulations 3kb windows yielded the
13 highest power for *NCD2* (S3-S6 Figs, Table 1), we queried the 1000G data with sliding windows of
14 3 kb (1.5 kb step size). Windows were defined in physical distance since the presence of LTBS may
15 affect the population-based estimates of recombination rate. For each window in each population
16 we calculated *NCD2* for three *tf* values (0.3-0.5).

17

18 **Filtering and correction for number of informative sites.** Genome-wide studies of natural
19 selection typically place a threshold on the minimum number of IS necessary (e.g., at least 10 IS in
20 [28], or 100 IS in [31]). We observe considerable variance in the number of IS per 3 kb window in
21 the 1000G data; also, *NCD2* has high variance when the number of IS is low in neutral simulations
22 (S11 and S18 Figs). We thus excluded windows with less than 10 IS in a given population because,
23 for higher values of IS, *NCD2* stabilizes. We then analyzed the 1,657,989 windows that remained in
24 all populations, covering 2,145,937,383 base pairs (S13 Fig). Neutral simulations with different
25 mutation rates were performed in order to retrieve 10,000 simulations for each value of IS (S18 Fig

1 and Methods). $NCD2$ ($tf = 0.3, 0.4, 0.5$) was calculated for all simulations, allowing the assignment
2 of significant windows and the calculation of Z_{tf-IS} (Equation 2 below).

3
4 **Significant and outlier windows.** We defined two sets of windows with signatures of LTBS: the
5 *significant* (based on neutral simulations) and *outlier* windows (based on the empirical distribution
6 of Z_{tf-IS} , see below). When referring to both sets, we use the term *candidate windows*. *Significant*
7 *windows* were defined as those fulfilling the criterion whereby the observed $NCD2$ value is lower
8 than all values obtained from 10,000 simulations with the same number of IS. Thus, all significant
9 windows have the same p -value ($p < 0.0001$). In order to rank the windows and define outliers, we
10 used a standardized distance measure between the observed $NCD2$ (for a queried window) and the
11 mean of the $NCD2$ values for the 10,000 simulations with the matching number of IS:

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

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

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

1 analyzed an alternative modern human genome-wide dataset, sequenced to an average coverage of
2 20x-30x per individual [112,113]. We used an independent data set because read coverage data is
3 low and cryptic in the 1000G, and putative duplications affecting the SFS must be at appreciable
4 frequency and should be present in other data sets. We considered 2 genomes from each of the
5 following populations: Yoruba, San, French, Sardinian, Dai, and Han Chinese. For each sample, we
6 retrieved positions above the 97.5% quantile of the coverage distribution for that sample (“high
7 coverage” positions). For each window with signatures of LTBS, we calculated the proportion of
8 the 3kb window having high coverage in at least two samples and plotted the distributions for
9 different $NCD2$ Z_{tf-IS} p -values. Extreme NCD windows are not enriched in high-coverage regions;
10 in fact, they are depleted of them in some cases (S14 Fig) (Mann-Whitney U two-tail test; $p < 0.02$
11 for $tf = 0.5$ and $tf = 0.4$ for GBR and TSI).

12

13 **Enrichment Analyses**

14 **Gene (GO) and Phenotype (PO) Ontology, and Tissue-specific expression.** We analyzed
15 protein-coding genes overlapped by one or more candidate windows. GO, PO and tissue of
16 expression enrichment analyses were performed using GOWINDA [114], which corrects for gene
17 length-related biases and/or gene clustering (S6 Note). GO/PO accession terms were downloaded
18 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
19 completely linked) and performed 100,000 simulations for FDR (false discovery rate) estimation.
20 Significant GO, PO and tissue-specific categories were defined for a $FDR < 0.05$. In both cases, a
21 minimum of three genes in the enriched category was required.

22
23 For tissue-specific expression analysis we used Illumina BodyMap 2.0 [115] expression data
24 for 16 tissues, and considered genes significantly highly expressed in a particular tissue when
25 compared to the remaining 15 tissues using the DESeq package [116], as done in [117]. All three
26 enrichment analyses (GO, PO, and tissue-specific expression) were performed for each population

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

3

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

8

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

25

1 **Genes with monoallelic expression (MAE) and immune-related genes.** To test for enrichment
2 for genes with MAE, we quantified the number of outlier and significant genes with MAE and the
3 number that have bi-allelic expression as described in [46]. We compared these proportions to those
4 observed for all scanned genes (one-tailed Fisher’s test.) The same procedure was adopted to test
5 for enrichment of immune-related genes among our sets: we used a list of 386 keywords from the
6 Comprehensive List of Immune Related Genes from Import
7 (<https://import.niaid.nih.gov/importWeb/queryref/importgene/importGeneList.do>) and
8 queried how many of the outlier protein-coding genes (402 genes in total across populations and *tf*,
9 of which 378 had at least one associated GO term) had at least one immune-related associated GO
10 category.

11 All statistical analyses and figures were performed in R [118] (scripts available on
12 https://github.com/bbitarello/NCV_dir_package). Gene Cards (www.genecards.org) and Enrichr
13 [119] were used to obtain basic functional information about genes and STRING v10 [87] was used
14 to obtain information for interactions between genes. The GWAS catalog [120] was used to search
15 for associations included in the discussion (we only report “strong associations”, i.e, when there is
16 at least one SNP with $p < 10^{-8}$).

17

18 **Acknowledgements**

19

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

27

1 Author Contributions

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

7

8 References

- 9 1. Meyer D, Thomson G. How selection shapes variation of the human major histocompatibility complex: a
10 review. *Ann Hum Genet.* 2001;65: 1–26.
- 11 2. Spurgin LG, Richardson DS. How pathogens drive genetic diversity: MHC, mechanisms and
12 misunderstandings. *Proc Biol Sci.* 2010;277: 979–88.
- 13 3. Robinson J, Halliwell JA, McWilliam H, Lopez R, Parham P, Marsh SGE. The IMGT/HLA database. *Nucleic
14 Acids Res.* 2013;41: D1222–7.
- 15 4. Hedrick PW, Whittam TS, Parham P. Heterozygosity at individual amino acid sites: extremely high levels for
16 HLA-A and -B genes. *Proc Natl Acad Sci.* 1991;88: 5897–5901.
- 17 5. Prugnolle F, Manica A, Charpentier M, Guégan JF, Guernier V, Balloux F. Pathogen-driven selection and
18 worldwide HLA class I diversity. *Curr Biol.* 2005;15: 1022–7.
- 19 6. Raychaudhuri S, Sandor C, Stahl EA, Freudenberg J, Lee H-S, Jia X, et al. Five amino acids in three HLA
20 proteins explain most of the association between MHC and seropositive rheumatoid arthritis. *Nat Genet. Nature
21 Publishing Group;* 2012;44: 291–296. doi:10.1038/ng.1076
- 22 7. Howell WM. HLA and disease: Guilt by association. *Int J Immunogenet.* 2014;41: 1–12. doi:10.1111/iji.12088
- 23 8. Ferrer-Admetlla A, Bosch E, Sikora M, Marques-Bonet T, Ramirez-Soriano A, Muntasell A, et al. Balancing
24 Selection Is the Main Force Shaping the Evolution of Innate Immunity Genes. *J Immunol.* 2008;181: 1315–
25 1322. doi:10.4049/jimmunol.181.2.1315
- 26 9. Sironi M, Clerici M. The hygiene hypothesis: an evolutionary perspective. *Microbes Infect. Elsevier Masson
27 SAS;* 2010;12: 421–427. doi:10.1016/j.micinf.2010.02.002
- 28 10. Malaria Genomic Epidemiology Network. A novel locus of resistance to severe malaria in a region of ancient
29 balancing selection. *Nature.* 2015;526: 253–257. doi:10.1038/nature15390
- 30 11. Biasin M, Piacentini L, Caputo S Lo, Kanari Y, Magri G, Trabattoni D, et al. Apolipoprotein B mRNA—
31 Editing Enzyme, Catalytic Polypeptide—Like 3G: A Possible Role in the Resistance to HIV of HIV-Exposed
32 Seronegative Individuals. *J Infect Dis.* 2007;195: 960–964. doi:10.1086/511988
- 33 12. Day FR, Hinds DA, Tung JY, Stolk L, Stykarsdottir U, Saxena R, et al. Causal mechanisms and balancing
34 selection inferred from genetic associations with polycystic ovary syndrome. *Nat Commun.* 2015;6: 8464.
35 doi:10.1038/ncomms9464
- 36 13. Andrés AM. *Balancing Selection in the Human Genome.* eLS. John Wiley & Sons, Ltd; 2011; 1–8.
37 doi:10.1002/9780470015902.a0022863
- 38 14. Fijarczyk A, Babik W. Detecting balancing selection in genomes: Limits and prospects. *Mol Ecol.* 2015; n/a-
39 n/a. doi:10.1111/mec.13226
- 40 15. Key FM, Teixeira JC, de Filippo C, Andrés AM. Advantageous diversity maintained by balancing selection in
41 humans. *Curr Opin Genet Dev.* 2014;29: 45–51. doi:10.1016/j.gde.2014.08.001
- 42 16. Charlesworth B, Charlesworth D. *Elements of Evolutionary Genetics.* 1st ed. Roberts and Company Publishers;
43 2010.
- 44 17. Clarke B. Balanced polymorphism and the diversity of sympatric species. In: Nichols D, editor. *Taxonomy and
45 Geography.* Oxford: Systematics Association; 1962.
- 46 18. Bergland AO, Behrman EL, O'Brien KR, Schmidt PS, Petrov DA. Genomic Evidence of Rapid and Stable
47 Adaptive Oscillations over Seasonal Time Scales in *Drosophila*. *PLoS Genet.* 2014;10: e1004775.
48 doi:10.1371/journal.pgen.1004775
- 49 19. Muehlenbachs A, Fried M, Lachowitz J, Mutabingwa TK, Duffy PE. Natural selection of FLT1 alleles and
50 their association with malaria resistance in utero. *Proc Natl Acad Sci.* 2008;105: 14488–14491.

- 1 doi:10.1073/pnas.0803657105
- 2 20. Charlesworth B, Nordborg M, Charlesworth D. The effects of local selection, balanced polymorphism and
3 background selection on equilibrium patterns of genetic diversity in subdivided population. *Genet Res.* 1997;70:
4 155–174.
- 5 21. Charlesworth D. Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genet.*
6 2006;2: 379–384. doi:10.1371/journal.pgen.0020064
- 7 22. Johnston SE, Gratten J, Berenos C, Pilkington JG, Clutton-Brock TH, Pemberton JM, et al. Life history trade-
8 offs at a single locus maintain sexually selected genetic variation. *Nature.* 2013;502: 93–95.
9 doi:10.1038/nature12489
- 10 23. Leffler EM, Pfeifer S, Auton A, Venn O, Bowden R, Bontrop R, et al. Multiple Instances of Ancient Balancing
11 Selection Shared Between Humans and Chimpanzees. *Science (80-).* 2013;339: 1578–1582.
12 doi:10.1126/science.1234070
- 13 24. Teixeira JC, de Filippo C, Weihmann A, Meneu JR, Racimo F, Dannemann M, et al. Long-Term Balancing
14 Selection in LAD1 Maintains a Missense Trans-Species Polymorphism in Humans, Chimpanzees, and
15 Bonobos. *Mol Biol Evol.* 2015;32: 1186–1196. doi:10.1093/molbev/msv007
- 16 25. Sellis D, Callahan BJ, Petrov D a., Messer PW. Heterozygote advantage as a natural consequence of adaptation
17 in diploids. *Proc Natl Acad Sci.* 2011;108: 20666–20671. doi:10.1073/pnas.1114573108
- 18 26. Hedrick PW. What is the evidence for heterozygote advantage selection? *Trends Ecol Evol Evol.* Elsevier Ltd;
19 2012;27: 698–704. doi:10.1016/j.tree.2012.08.012
- 20 27. Alonso S, Lopez S, Izagirre N, de la Rúa C. Overdominance in the Human Genome and Olfactory Receptor
21 Activity. *Mol Biol Evol.* 2008;25: 997–1001. doi:10.1093/molbev/msn049
- 22 28. Andrés AM, Hubisz MJ, Indap A, Torgerson DG, Degenhardt JD, Boyko AR, et al. Targets of balancing
23 selection in the human genome. *Mol Biol Evol.* 2009;26: 2755–64. doi:10.1093/molbev/msp190
- 24 29. Bubb KL, Bovee D, Buckley D, Haugen E, Kibukawa M, Paddock M, et al. Scan of human genome reveals no
25 new Loci under ancient balancing selection. *Genetics.* 2006;173: 2165–77. doi:10.1534/genetics.106.055715
- 26 30. Asthana S, Schmidt S, Sunyaev SR. A limited role for balancing selection. *Trends Genet.* 2005;21: 30–32.
27 doi:10.1016/j.tig.2004.11.007
- 28 31. DeGiorgio M, Lohmueller KE, Nielsen R. A model-based approach for identifying signatures of ancient
29 balancing selection in genetic data. *PLoS Genet.* 2014;10: e1004561. doi:10.1371/journal.pgen.1004561
- 30 32. Rasmussen MD, Hubisz MJ, Gronau I, Siepel A. Genome-Wide Inference of Ancestral Recombination Graphs.
31 Coop G, editor. *PLoS Genet.* 2014;10: e1004342. doi:10.1371/journal.pgen.1004342
- 32 33. Hudson RR, Kaplan NL. The coalescent process in models with selection and recombination. *Genetics.*
33 1988;120: 831–840. doi:10.1017/S0016672300029074
- 34 34. Hudson RR, Kreitman M, Aguade M. A Test of Neutral Molecular Evolution Based on Nucleotide Data.
35 *Genetics.* 1987;116: 153–159. Available: <http://www.genetics.org/cgi/content/abstract/116/1/153>
- 36 35. Coventry A, Bull-Otterson LM, Liu X, Clark AG, Maxwell TJ, Crosby J, et al. Deep resequencing reveals
37 excess rare recent variants consistent with explosive population growth. *Nat Commun.* 2010;1: 131.
38 doi:10.1038/ncomms1130
- 39 36. Fu W, Akey JM. Selection and Adaptation in the Human Genome. *Annu Rev Genomics Hum Genet.* 2013;14:
40 467–489. doi:10.1146/annurev-genom-091212-153509
- 41 37. Nielsen R, Hubisz MJ, Hellmann I, Torgerson D, Andrés AM, Albrechtsen A, et al. Darwinian and
42 demographic forces affecting human protein coding genes. *Genome Res.* 2009;19: 838–49.
43 doi:10.1101/gr.088336.108
- 44 38. Nielsen R, Bustamante C, Clark A, Glanowski S, Sackton T, Hubisz MJ, et al. A Scan for Positively Selected
45 Genes in the Genomes of Humans and Chimpanzees. *PLoS Biol.* 2005;3: e170.
46 doi:10.1371/journal.pbio.0030170
- 47 39. Williamson SH, Hubisz MJ, Clark AG, Payseur BA, Bustamante CD, Nielsen R. Localizing recent adaptive
48 evolution in the human genome. *PLoS Genet.* 2007;3: e90. doi:10.1371/journal.pgen.0030090
- 49 40. Abecasis GR, Auton A, Brooks LD, DePristo M a, Durbin RM, Handsaker RE, et al. An integrated map of
50 genetic variation from 1,092 human genomes. *Nature.* 2012;491: 56–65. doi:10.1038/nature11632
- 51 41. Liu X, Fu Y, Liu Z, Lin B, Xie Y, Liu Y, et al. An Ancient Balanced Polymorphism in a Regulatory Region of
52 Human Major Histocompatibility Complex Is Retained in Chinese Minorities but Lost Worldwide. *Am J Hum*
53 *Genet.* 2006;78: 393–400. doi:10.1086/500593
- 54 42. Meyer D, Single RM, Mack SJ, Erlich HA, Thomson G. Signatures of demographic history and natural
55 selection in the human major histocompatibility complex Loci. *Genetics.* 2006;173: 2121–2142.
56 doi:10.1534/genetics.105.052837
- 57 43. Sanchez-Mazas A. An apportionment of human HLA diversity. *Tissue Antigens.* 2007;69: 198–202.
58 doi:10.1111/j.1399-0039.2006.00802.x
- 59 44. Solberg OD, Mack SJ, Lancaster AK, Single RM, Tsai Y, Sanchez-Mazas A, et al. Balancing selection and
60 heterogeneity across the classical human leukocyte antigen loci: A meta-analytic review of 497 population
61 studies. *Hum Immunol.* 2008;69: 443–464. doi:10.1016/j.humimm.2008.05.001
- 62 45. Tan Z, Shon AM, Ober C. Evidence of balancing selection at the HLA-G promoter region. *Hum Mol Genet.*
63 2005;14: 3619–3628. doi:10.1093/hmg/ddi389

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

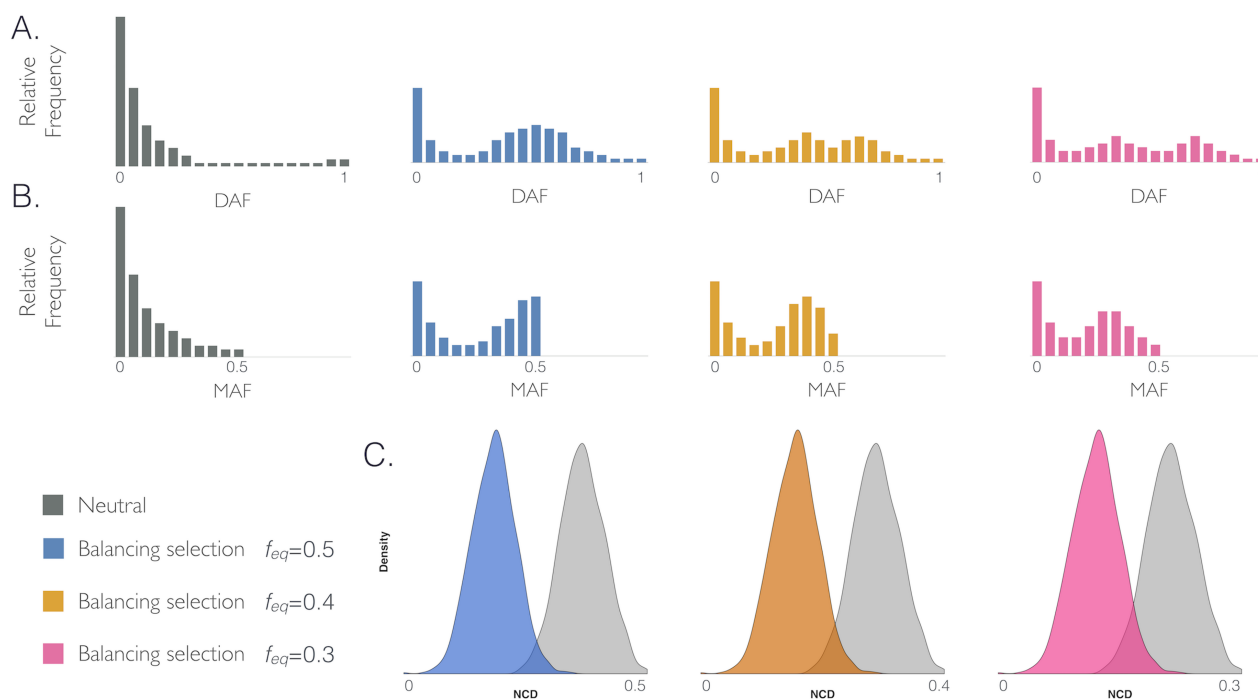
- 1 doi:10.1038/ng.694
- 2 69. Charlesworth B. Effective population size and patterns of molecular evolution and variation. *Nat Rev Genet.* 2009;10: 195–205. doi:10.1038/nrg2526
- 3
- 4 70. Corona E, Dudley JT, Butte AJ. Extreme evolutionary disparities seen in positive selection across seven
- 5 complex diseases. *PLoS One.* 2010;5: 1–10. doi:10.1371/journal.pone.0012236
- 6 71. Wooding S, Kim U, Bamshad MJ, Larsen J, Jorde LB, Drayna D. Natural Selection and Molecular Evolution in
- 7 PTC, a Bitter-Taste Receptor Gene. *Am J Hum Genet.* 2004;74: 637–646. doi:10.1086/383092
- 8 72. Wooding S, Bufo B, Grassi C, Howard MT, Stone AC, Vazquez M, et al. Independent evolution of bitter-taste
- 9 sensitivity in humans and chimpanzees. *Nature.* 2006;440: 930–934. doi:10.1038/nature04655
- 10 73. Ledda M, Kutalik Z, Souza Destito MC, Souza MM, Cirillo CA, Zamboni A, et al. GWAS of human bitter taste
- 11 perception identifies new loci and reveals additional complexity of bitter taste genetics. *Hum Mol Genet.*
- 12 2014;23: 259–267. doi:10.1093/hmg/ddt404
- 13 74. Thalmann S, Behrens M, Meyerhof W. Major haplotypes of the human bitter taste receptor TAS2R41 encode
- 14 functional receptors for chloramphenicol. *Biochem Biophys Res Commun.* 2013;435: 267–273.
- 15 doi:10.1016/j.bbrc.2013.04.066
- 16 75. Meyerhof W, Batram C, Kuhn C, Brockhoff A, Chudoba E, Bufo B, et al. The Molecular Receptive Ranges of
- 17 Human TAS2R Bitter Taste Receptors. *Chem Senses.* 2010;35: 157–170. doi:10.1093/chemse/bjp092
- 18 76. Karaman R, Nowak S, Di Pizio A, Kitaneh H, Abu-Jaish A, Meyerhof W, et al. Probing the Binding Pocket of
- 19 the Broadly Tuned Human Bitter Taste Receptor TAS2R14 by Chemical Modification of Cognate Agonists.
- 20 *Chem Biol Drug Des.* 2016;88: 66–75. doi:10.1111/cbdd.12734
- 21 77. Behrens M, Brockhoff A, Kuhn C, Bufo B, Winnig M, Meyerhof W. The human taste receptor hTAS2R14
- 22 responds to a variety of different bitter compounds. *Biochem Biophys Res Commun.* 2004;319: 479–485.
- 23 doi:10.1016/j.bbrc.2004.05.019
- 24 78. Kosiol C, Vinar T, da Fonseca RR, Hubisz MJ, Bustamante CD, Nielsen R, et al. Patterns of positive selection
- 25 in six Mammalian genomes. *PLoS Genet.* 2008;4: 1–17. doi:10.1371/journal.pgen.1000144
- 26 79. Arking DE, Krebsova A, Macek M, Macek M, Arking A, Mian IS, et al. Association of human aging with a
- 27 functional variant of klotho. *Proc Natl Acad Sci.* 2002;99: 856–861. doi:10.1073/pnas.022484299
- 28 80. Welberg L. Cognition: Klotho spins cognitive fate. *Nat Rev Neurosci.* 2014;15: 425–425. doi:10.1038/nrn3777
- 29 81. Dubal DB, Yokoyama JS, Zhu L, Broestl L, Worden K, Wang D, et al. Life Extension Factor Klotho Enhances
- 30 Cognition. *Cell Rep. The Authors;* 2014;7: 1065–1076. doi:10.1016/j.celrep.2014.03.076
- 31 82. Lui JH, Nowakowski TJ, Pollen AA, Javaherian A, Kriegstein AR, Oldham MC. Radial glia require PDGFR β –
- 32 PDGFR β signalling in human but not mouse neocortex. *Nature.* 2014;515: 264–268. doi:10.1038/nature13973
- 33 83. Rakic P. Evolution of the neocortex: a perspective from developmental biology. *Nat Rev Neurosci.* 2009;10:
- 34 724–735. doi:10.1038/nrn2719
- 35 84. St Pourcain B, Cents RAM, Whitehouse AJO, Haworth CMA, Davis OSP, O’Reilly PF, et al. Common
- 36 variation near ROBO2 is associated with expressive vocabulary in infancy. *Nat Commun.* 2014;5: 4831.
- 37 doi:10.1038/ncomms5831
- 38 85. Peyrégne S, Dannemann M, Prüfer K. Detecting ancient positive selection in humans using extended lineage
- 39 sorting. *BiorXiv.* 2016; doi:10.1101/092999
- 40 86. Sundermann EE, Maki PM, Bishop JR. A review of estrogen receptor α gene (ESR1) polymorphisms, mood,
- 41 and cognition. *Menopause.* 2010;17: 874–886. doi:10.1097/gme.0b013e3181df4a19
- 42 87. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: protein-
- 43 protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* 2015;43: D447–D452.
- 44 doi:10.1093/nar/gku1003
- 45 88. Perlis RH, Huang J, Purcell S, Fava M, Rush AJ, Sullivan PF, et al. Genome-Wide Association Study of
- 46 Suicide Attempts in Mood Disorder Patients. *Am J Psychiatry.* 2010;167: 1499–1507.
- 47 doi:10.1176/appi.ajp.2010.10040541
- 48 89. Mullins N, Perroud N, Uher R, Butler AW, Cohen-Woods S, Rivera M, et al. Genetic relationships between
- 49 suicide attempts, suicidal ideation and major psychiatric disorders: a genome-wide association and polygenic
- 50 scoring study. *Am J Med Genet B Neuropsychiatr Genet.* 2014;165B: 428–37. doi:10.1002/ajmg.b.32247
- 51 90. Shatz CJ. MHC Class I: An Unexpected Role in Neuronal Plasticity. *Neuron.* 2009;64: 40–45.
- 52 doi:10.1016/j.neuron.2009.09.044
- 53 91. Needleman LA, McAllister AK. The major histocompatibility complex and autism spectrum disorder. *Dev*
- 54 *Neurobiol.* 2012;72: 1288–1301. doi:10.1002/dneu.22046
- 55 92. Torres AR, Westover JB, Rosenspire AJ. HLA Immune Function Genes in Autism. *Autism Res Treat.*
- 56 2012;2012: 1–13. doi:10.1155/2012/959073
- 57 93. Careaga M, Water J, Ashwood P. Immune dysfunction in autism: A pathway to treatment. *Neurotherapeutics.*
- 58 2010;7: 283–292. doi:10.1016/j.nurt.2010.05.003
- 59 94. Gow A, Southwood CM, Li JS, Pariali M, Riordan GP, Brodie SE, et al. CNS myelin and sertoli cell tight
- 60 junction strands are absent in *Osp/claudin-11* null mice. *Cell.* 1999;99: 649–59. Available:
- 61 <https://www.ncbi.nlm.nih.gov/pubmed/10612400/>
- 62 95. Wu X, Peppi M, Vengalil MJ, Maheras KJ, Southwood CM, Bradley M, et al. Transgene-Mediated Rescue of
- 63 Spermatogenesis in *Cldn11-Null* Mice. *Biol Reprod.* 2012;86. doi:10.1095/biolreprod.111.096230

- 1 96. Quaynor SD, Stradtman EW, Kim H-G, Shen Y, Chorich LP, Schreihof DA, et al. Delayed puberty and
2 estrogen resistance in a woman with estrogen receptor α variant. *N Engl J Med*. 2013;369: 164–71.
3 doi:10.1056/NEJMoal303611
- 4 97. Michailidou K, Hall P, Gonzalez-Neira A, Ghoussaini M, Dennis J, Milne RL, et al. Large-scale genotyping
5 identifies 41 new loci associated with breast cancer risk. *Nat Genet*. 2013;45: 353–61, 361–2.
6 doi:10.1038/ng.2563
- 7 98. Eriksson N, Benton GM, Do CB, Kiefer AK, Mountain JL, Hinds DA, et al. Genetic variants associated with
8 breast size also influence breast cancer risk. *BMC Med Genet*. 2012;13: 53. doi:10.1186/1471-2350-13-53
- 9 99. Lazari MFM, Lucas TFG, Yasuhara F, Gomes GRO, Siu ER, Royer C, et al. Estrogen receptors and function in
10 the male reproductive system. *Arq Bras Endocrinol Metabol*. 2009;53: 923–933.
- 11 100. Rivadeneira F, Styrkarsdottir U, Estrada K, Halldórsson B V, Hsu Y-H, Richards JB, et al. Twenty bone-
12 mineral-density loci identified by large-scale meta-analysis of genome-wide association studies. *Nat Genet*.
13 2009;41: 1199–1206. doi:10.1038/ng.446
- 14 101. Aouizerat BE, Vittinghoff E, Musone SL, Pawlikowska L, Kwok P-Y, Olgin JE, et al. GWAS for discovery and
15 replication of genetic loci associated with sudden cardiac arrest in patients with coronary artery disease. *BMC*
16 *Cardiovasc Disord*. 2011;11: 29. doi:10.1186/1471-2261-11-29
- 17 102. Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, et al. Defining the role of common variation in
18 the genomic and biological architecture of adult human height. *Nat Genet*. 2014;46: 1173–1186.
19 doi:10.1038/ng.3097
- 20 103. Heldring N, Pike A, Andersson S, Matthews J, Cheng G, Treuter E, et al. Estrogen Receptors : How Do They
21 Signal and What Are Their Targets. *Physiol Rev*. 2007;87: 905–931. doi:10.1152/physrev.00026.2006.
- 22 104. Ewing G, Hermisson J. MSMS: a coalescent simulation program including recombination, demographic
23 structure and selection at a single locus. *Bioinformatics*. 2010;26: 2064–2065.
24 doi:10.1093/bioinformatics/btq322
- 25 105. Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, et al. Demographic history and rare
26 allele sharing among human populations. *Proc Natl Acad Sci U S A*. 2011;108: 11983–8.
27 doi:10.1073/pnas.1019276108
- 28 106. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*.
29 *Genetics Soc America*; 1989;123: 585–595.
- 30 107. Key FM, Peter B, Dennis MY, Huerta-Sánchez E, Tang W, Prokunina-Olsson L, et al. Selection on a Variant
31 Associated with Improved Viral Clearance Drives Local, Adaptive Pseudogenization of Interferon Lambda 4
32 (IFNL4). *PLoS Genet*. 2014;10: e1004681. doi:10.1371/journal.pgen.1004681
- 33 108. Derrien T, Estellé J, Marco Sola S, Knowles DG, Raineri E, Guigó R, et al. Fast Computation and Applications
34 of Genome Mappability. Ouzounis CA, editor. *PLoS One*. 2012;7: e30377. doi:10.1371/journal.pone.0030377
- 35 109. Cheng Z, Ventura M, She X, Khaitovich P, Graves T, Osoegawa K, et al. A genome-wide comparison of recent
36 chimpanzee and human segmental duplications. *Nature*. 2005;437: 88–93. doi:10.1038/nature04000
- 37 110. Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F, et al. Personalized copy number
38 and segmental duplication maps using next-generation sequencing. *Nat Genet*. 2009;41: 1061–1067.
39 doi:10.1038/ng.437
- 40 111. Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res*. 1999;27: 573–
41 580.
- 42 112. Meyer M, Kircher M, Gansauge M-T, Li H, Racimo F, Mallick S, et al. A High-Coverage Genome Sequence
43 from an Archaic Denisovan Individual. *Science (80-)*. 2012;338: 222–226. doi:10.1126/science.1224344
- 44 113. Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, et al. The complete genome sequence of a
45 Neanderthal from the Altai Mountains. *Nature*. 2013;505: 43–49. doi:10.1038/nature12886
- 46 114. Kofler R, Schlötterer C. Gowinda: Unbiased analysis of gene set enrichment for genome-wide association
47 studies. *Bioinformatics*. 2012;28: 2084–2085. doi:10.1093/bioinformatics/bts315
- 48 115. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, et al. The GENCODE v7 catalog of human
49 long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Res*. 2012;22:
50 1775–1789. doi:10.1101/gr.132159.111
- 51 116. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol*. 2010;11: R106.
52 doi:10.1186/gb-2010-11-10-r106
- 53 117. Sankararaman S, Mallick S, Dannemann M, Prüfer K, Kelso J, Pääbo S, et al. The genomic landscape of
54 Neanderthal ancestry in present-day humans. *Nature*. 2014;507: 354–7. doi:10.1038/nature12961
- 55 118. Development Core Team R. R: A language and environment for statistical computing. [Internet]. Vienna,
56 Austria: R Foundation for Statistical Computing; 2009. Available: <http://www.r-project.org>
- 57 119. Kuleshov M V., Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive
58 gene set enrichment analysis web server 2016 update. *Nucleic Acids Res*. 2016;44: W90–W97.
59 doi:10.1093/nar/gkw377
- 60 120. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS Catalog, a curated
61 resource of SNP-trait associations. *Nucleic Acids Res*. 2014;42: D1001–D1006. doi:10.1093/nar/gkt1229
62

1 Figure Captions

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

10



11

12

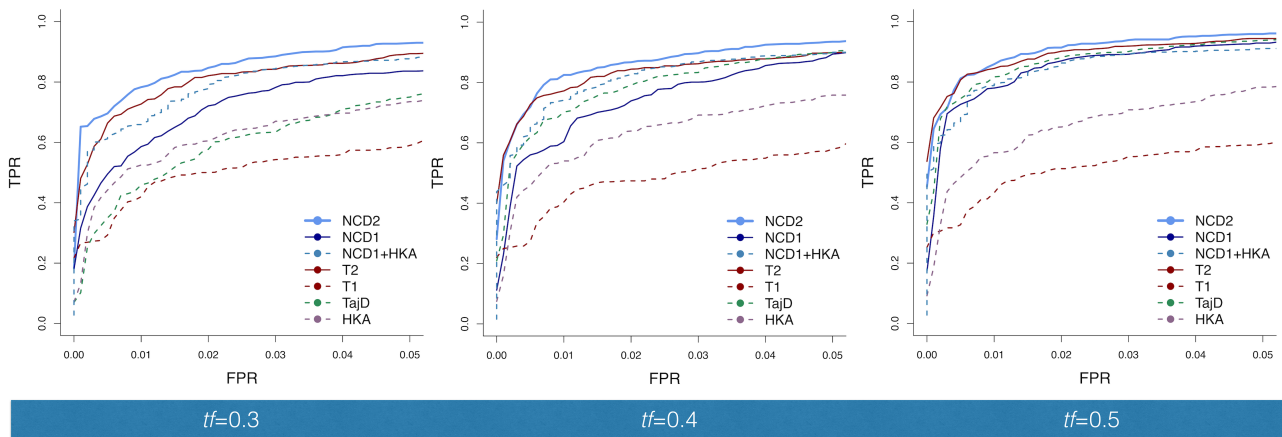
13

14

15

16

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



7 $tf=0.3$

8 $tf=0.4$

9 $tf=0.5$

10

11

12

13

14

15

16

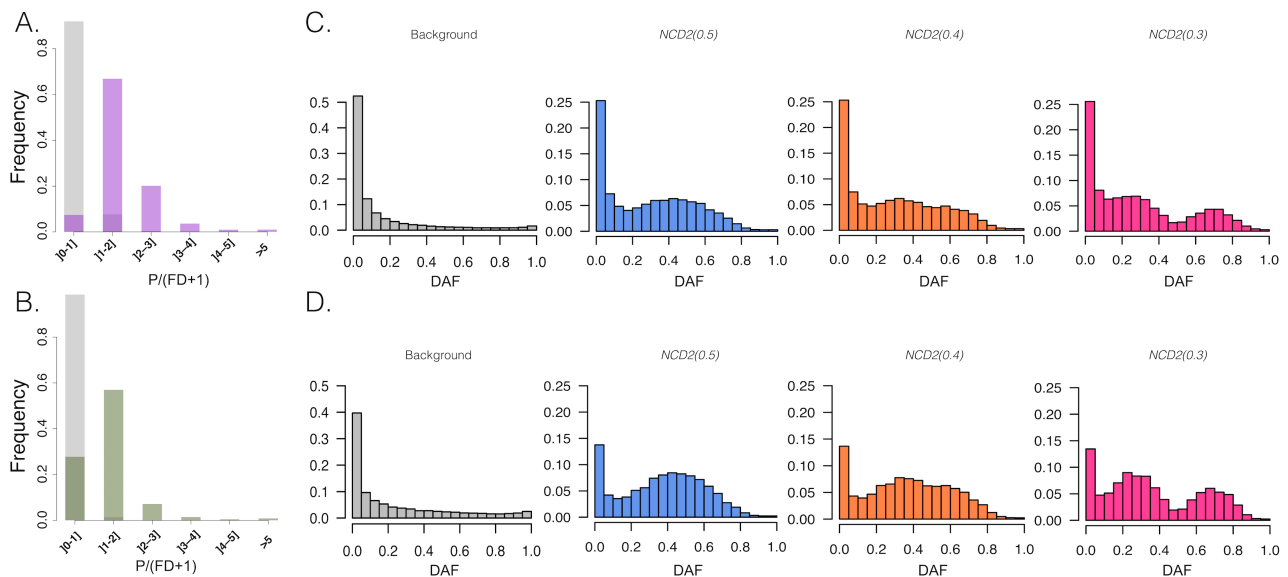
17

18

19

20

1 **Fig 3. Polymorphism-to-divergence and SFS. (A,B)** SNP/(FD+1) for LWK **(A)** and GBR **(B)**
 2 populations. SNP/(FD+1) measures the proportion of polymorphic-to-divergent sites for the union
 3 of significant windows for all t_f (purple, green) compared to all scanned windows (gray). **(C-D)**
 4 SFS in LWK **(C)** and GBR **(D)** of all scanned windows in chr1 (gray), significant windows for
 5 $NCD2(0.5)$ (blue), $NCD2(0.4)$ (orange), $NCD2(0.3)$ (pink). DAF, derived allele frequency.



6

7

8

9

10

11

12

13

14

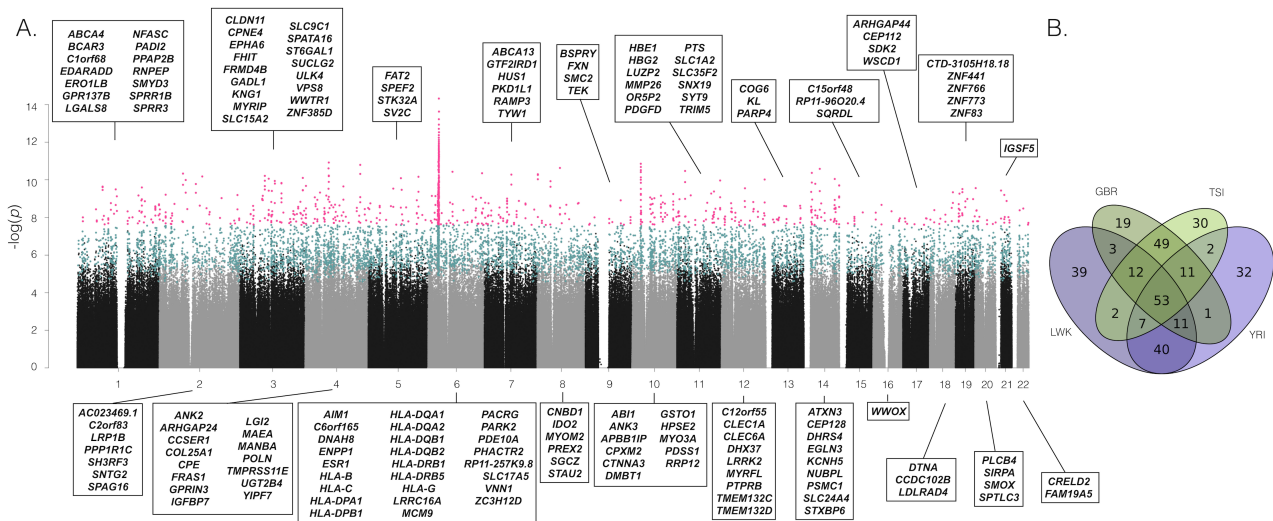
15

16

17

18

1 **Fig 4. Manhattan plot and population sharing.** (A) Manhattan plot of all scanned windows, for
 2 one analysis (*NCD2(0.5)* and LWK). *y*-axis, *p*-value (log-scale) based on Z_{tf-IS} . *x*-axis, ordered
 3 location of analyzed windows on the genome. Each point is a scanned (grey and black), significant
 4 (blue) or outlier (pink) window. Names of outlier protein-coding genes are provided, sorted by
 5 name. We note that significant windows were defined based of simulations, not on Z_{tf-IS} . (Z_{tf-IS} is
 6 used to rank those with $p < 0.0001$) (B) Venn diagram showing the overlap in signatures of the
 7 167 outlier genes annotated in (A) with other populations.



8

9

10

11

12

13

14

15

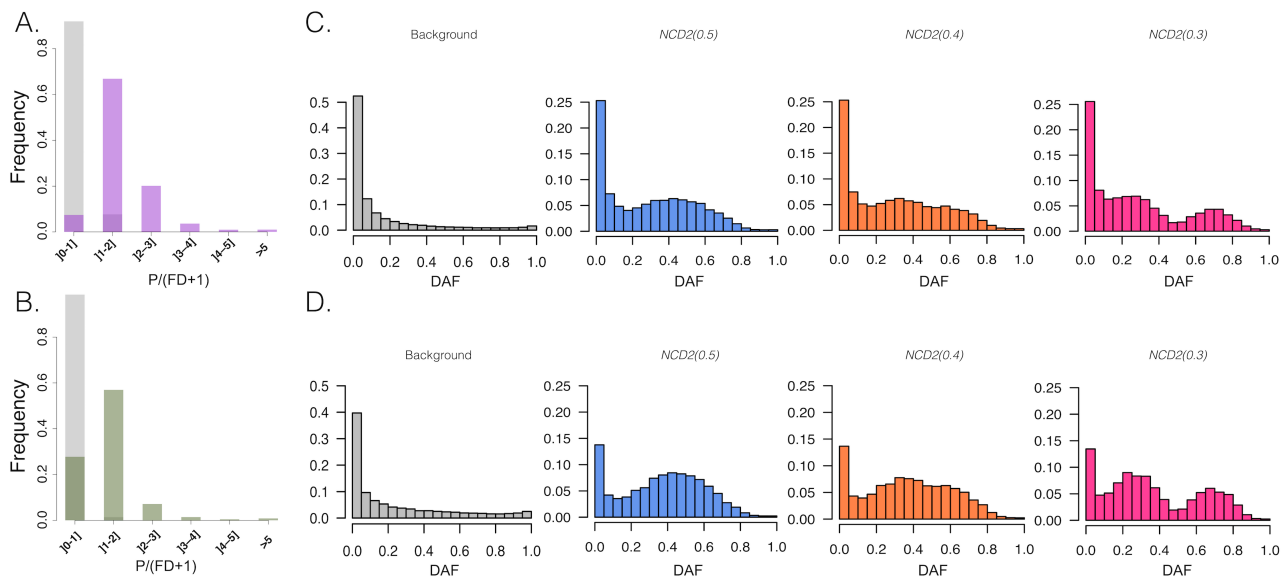
16

17

18

19

1 **Fig 5. Enrichment of classes of sites amongst candidate windows.** Dashed lines mark the $p =$
2 0.975 (bottom) and $p = 0.025$ (top) thresholds for the one-tailed p -values (hypothesis: enrichment).
3 NSyn, nonsynonymous; all, Genic plus Intergenic plus Regulatory. The annotation is based on
4 Ensembl variant predictor (S6 Note). $p < 0.001$ was treated as 0.001 to avoid infinite values.
5



6