bioRxiv preprint doi: https://doi.org/10.1101/112870; this version posted March 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Detecting Long-term Balancing Selection using Allele Frequency Correlation

Katherine M. Siewert, Benjamin F. Voight

Correspondence to: bvoight@upenn.edu

Abstract

Balancing selection preserves multiple alleles at a locus over long evolutionary time periods. A characteristic signature of balancing selection is an excess number of intermediate frequency polymorphisms nearby the balanced site. However, the expected distribution of allele frequencies at these loci has not been extensively detailed. We use simulations to show that new mutations that arise very close to a site targeted by long term balancing selection accumulate at frequencies nearly identical to that of the frequency of the balanced allele to which they are linked. To capture this unique signature, we propose a new summary statistic, β. Compared to existing summary statistics, simulation studies show that our statistic has improved power to detect balancing selection, and is reasonably powered in non-equilibrium demographic models or when recombination or mutation rate varies. We compute β on 1000 Genomes Project data, to identify loci potentially subjected to long-term balancing selection in humans. We report two balanced haplotypes - localized to the genes WFS1 and CADM2 - that are strongly linked to association signals for complex traits. Our approach is computationally efficient and applicable to species that lack appropriate outgroup sequences, allowing for well-powered analysis of selection in the wide variety of species for which population data is rapidly being generated.

Introduction

Large-scale, high-quality genomic data has spurred new methods to detect regions of the genome that have been subjected to natural selection in a wide variety of species [1, 2, 3]. One type of pressure, balancing selection, occurs when allelic diversity in a population is advantageous, causing the maintenance of more than one allele at a locus. It can be due to overdominance, in which the fitness of heterozygotes at a locus is higher than either type of homozygote; or frequency, temporally, or spatially-dependent selection [4]. A classic case of this type of selection occurs at the hemoglobin-β locus in populations located in malaria-endemic regions. Homozygotes for one allele at this locus have sickle-cell anemia, and homozygotes for the other allele have an increased risk of malaria. In contrast, heterzoygoyes are protected from malaria, and at most have a mild case of sickle-cell anemia [5, 6].

New targets of balancing selection could help us better understand the role and frequency of this type of selection in evolution, uncover traits that have been preserved for long evolutionary time periods, and potentially aid in interpreting regions previously associated with phenotypes of interest. In addition, theory predicts that signatures of balanced selection will usually be confined to a relatively narrow region of a few kilobases [7], which would facilitate rapid causal variant identification and mechanistic studies.

Patterns of genetic variation around a locus targeted by balancing selection are distorted, and can be used to identify such events indirectly from population-level data. At a neutral locus, lineages

bioRxiv preprint doi: https://doi.org/10.1101/112870; this version posted March 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

are free to drift in and out of the population, so the time to the most recent common ancestor (TMRCA) is expected to be of only moderate age. In contrast, by maintaining both alleles at the locus, balancing selection dramatically increases the TMRCA relative to a neutral model. This elevates the levels of polymorphism around the balanced locus, and leads to a corresponding reduction in substitutions (*i.e.*, fixed differences relative to an outgroup).

Each of these signatures have been harnessed to identify signals of balancing selection in population data, genome-wide. These methods include Tajimas D [8], which detects distortions in the site frequency spectrum. Another commonly used method, the HKA test [9], is sensitive to excess levels of polymorphisms and a deficit of substitutions compared with neutrality. While these methods are easily implemented and widely applicable, their power under certain demographic scenarios or equilibrium frequencies is modest. If the selection began prior to the divergence of two species, then both species can share a balanced haplotype [4], an event unlikely under strict neutrality [7]. Several recent studies have utilized primate outgroups to identify these haplotypes [10, **?**, 11]. While specific, this approach only detects ancient selection that persists in both species.

More powerful methods to detect balancing selection have been developed, though challenges have limited their broader application. DeGeorgio *et al.* proposed two model-based summaries, which generate a composite likelihood of a site being under balancing selection [12]. However, the most powerful measure (T2) depends on the existence of a closely-related outgroup sequence, and on knowledge of the underlying demographic history from which an extensive grid of simulations must first be generated. New advances in estimating population-scale coalescent trees have also been harnessed to detect regions of the genome showing an unusually old TMRCA, but genome-wide application may be computationally prohibitive [13].

Despite these methodological advances, it is worth noting that the exact distortion on the site frequency spectrum for nearby sites has not been precisely quantified to date. The key insight motivating our work was the observation that the frequencies of the excess variants found nearby the selected site often closely match its frequency, which we confirm using simulations. Motivated by this observation, and inspired by the structure of summary-spectrum based statistics [8, 14], we developed a new summary statistic that takes advantage of this unique signature. This statistic is computationally efficient and does not require knowledge of the ancestral state or an outgroup sequence. Using simulations, we show our approach has equivalent or elevated power compared to other similar approaches, and retains power over a range of population genetic models and assumptions

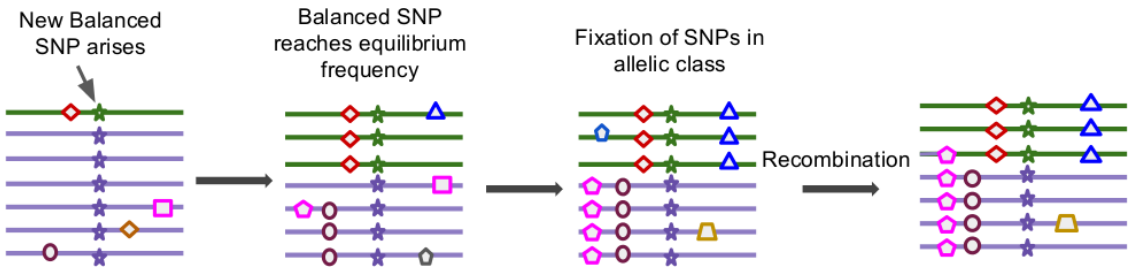

Figure 1: Model of allelic class build-up. (1) A new SNP (green star) arises in the population and is subject to balancing selection. (2) It sweeps up to its equilibrium frequency. (3) New SNPs enter the population linked to one of the two balanced alleles and some drift up in frequency. However, unlike in the neutral case, their maximum frequency is that of the balanced allele they are found in, so they build-up at this frequency (e.g. red diamond or brown circle). (4) Recombination decouples SNPs from the balanced site (pink pentagon), allowing them to experience further genetic drift.

bioRxiv preprint doi: https://doi.org/10.1101/112870; this version posted March 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

(i.e., demography, mutation, or recombination). Using our method, we report a genome-wide scan in humans using 1000 Genomes Project data [15], focused on regions of high sequence quality. We report signals of balancing selection at two loci (*WFS1* and *CAMD2*) with functional evidence supporting those as the target gene, as well as at several previously known loci.

Theory and Model

Allelic Class Build-up

We begin with an idealized model to generate the expected distribution of allele frequencies of these linked sites over time. Consider a new mutation that arises within a population. Under neutrality, this mutation either drifts out and is lost, or becomes fixed (*i.e.*, a substitution). However, under balancing selection the variant is retained (Fig. 1), and rather than drifting to complete fixation, its frequency only reaches the equilibrium frequency of the balanced allele it arose in linkage with. Without a recombination event, variants are maintained at frequencies within these allelic classes for long periods of time [4, 16]. Eventually, recombination decouples variants from the balanced allele, which then allows drift to fixation. Even after recombination, the frequency of the variants previously fixed in their allelic class will remain close to that of their previous class until enough time has passed for genetic drift to significantly change their frequencies. Therefore we do expect to see the classic signature of an increase in polymorphisms levels within the immediate vicinity of a balanced site. However, these alleles are not expected to be just at any intermediate frequency, but instead at nearly identical frequencies to the balanced site.

To assess the validity of this model, we used Wright-Fisher forward simulations to measure the allele frequency of neutral sites in a window around a site targeted by balancing selection (**Methods**). As expected, we recapitulated the excess of polymorphism and reduction in substitutions expected under this model **(Fig. S1)**. Within the region not experiencing recombination since the start of selection, we observed an excess frequency of alleles at identical (or nearly so) frequencies to the balanced allele, relative to neutrality **(Fig. 2a)**. For larger windows, there is excess correlation, but it is diluted by recombination **(Fig. 2b)**. These data suggest that the presence of many variants with correlated alleles frequencies is a precise signature of balancing selection.

A Measure for Allele Frequency Correlation

To capture this signature, we sought to derive a measurement sensitive to the correlation in allele frequencies between nearby sites. Let n be the number of chromosomes sampled, f_i be the frequency of SNP i, f_0 be frequency of the core SNP, $g(f)$ return the folded allele frequency, m be the maximum possible allele frequency difference between the core SNP and SNP i, and p be a scaling constant (see **Supplementary note**). The similarity in frequency, d_i can be measured by:

$$g(f) = \min(f, n - f) \tag{1}$$

$$m = \max\left(g(f_0), \frac{n}{2} - g(f_0)\right) \tag{2}$$

$$d_i = \left(\frac{m - |g(f_0) - g(f_i)|}{m}\right)^p \tag{3}$$

where d_i can range from 0 if a SNP has the maximum frequency difference with the core SNP, to 1 if SNP i is at the same frequency as the core SNP. In a region under long-term balancing selection, d_i between a core SNP and the surrounding variants is expected to be elevated. We use the folded site frequency spectrum in calculating d_i, as the frequency difference between the core variant and a nearby variant is independent of whether the derived or ancestral allele of the nearby allele is in linkage with the derived or ancestral core allele. However, d_i alone is not sufficient to detect balancing selection, as its value will be sensitive to changes in the mutation rate in the surrounding region, and it does not take into account the probability of seeing each allele frequency under neutrality.

bioRxiv preprint doi: https://doi.org/10.1101/112870; this version posted March 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

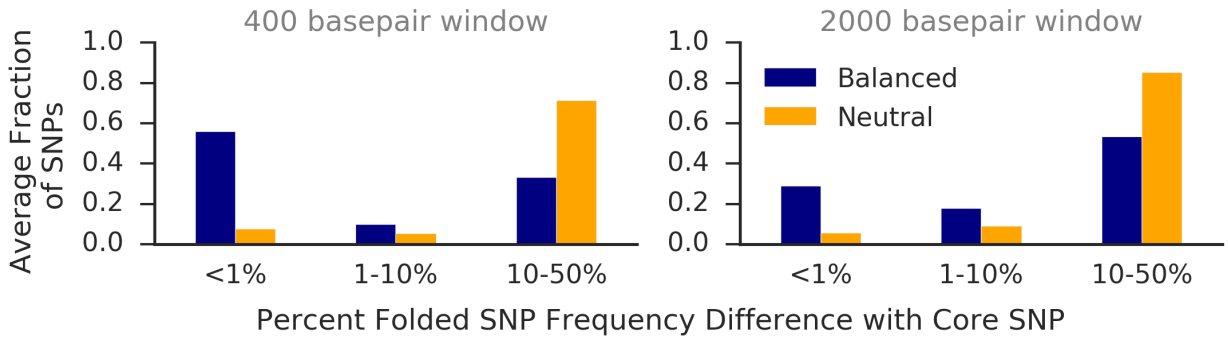

Figure 2: Accumulation of SNPs at frequencies identical to the balanced allele is diluted due to recombination (A) Folded frequency difference between the core SNP and each SNP in a 400 base-pair (bp) window surrounding the balanced site (blue). In orange is the frequency difference between a neutral SNP within frequency 0.1 of the equilibrium frequency and each other SNP in the window. Recombination is not expected to have occurred in this region since the start of selection [7]. (B) Frequency differences in 2,000 bp windows, where significant recombination is expected to have occurred since the start of selection.

Capturing Alleleic Class Build-up

We propose a statistic, β, that uses our measure of allele frequency correlation, combined with a measure of the overall mutation rate, to detect balancing selection. Our approach is inspired by previous summary statistics of the site frequency spectrum [8, 14]. These methods compute the difference between two estimators of θ, the population mutation rate parameter, one of which is more sensitive to characteristics of the site frequency spectrum distorted in the presence of natural selection. We propose to calculate β at each SNP in a region of interest to identify regions in which there is an excess of variants near the core SNP's allele frequency, as evidence of balancing selection.

It has been previously shown that the mutation rate in a region can be estimated as: $\hat{\theta}_i = S_i * i$, where S_i is the total number of derived variants found i times in the window from a sample of n chromosomes in the population [17]. An estimator of θ can then be obtained by taking a weighted average of θ_i. In our method, we weigh by the similarity in allele frequency to the core SNP, as measured by d_i. If there is an excess of variants at frequencies close to the core SNP allele frequency, then our new estimator, θ_β, will be elevated. We propose:

$$\beta = \hat{\theta}_\beta - \hat{\theta}_w \tag{4}$$

$$\hat{\theta}_\beta = \frac{\sum_{i=1}^{n-1} i d_i S_i}{\sum_{i=1}^{n-1} d_i} \tag{5}$$

$$\hat{\theta}_w = \frac{\sum_{i=1}^{n-1} S_i}{\sum_{i=1}^{n-1} \frac{1}{i}} \tag{6}$$

where S_i is the total number of derived variants found i times in the window from a sample of n chromosomes in the population, with θ_w simply as Watterson's estimator [18]. This is, in effect, a weighted average of SNP counts based on their frequency similarity to the core SNP. Note that here, we exclude the core site from our estimation of θ_w and θ_β.

To better understand the properties of β, we used simulations to examine its distribution with and without a balanced SNP (**Fig. S2**). As expected, under long-term balancing selection β is greater than

bioRxiv preprint doi: https://doi.org/10.1101/112870; this version posted March 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

zero. We note that the mean value of β in neutral simulations increased slightly with higher equilibrium frequencies. This behavior is expected because higher frequency alleles will tend to have a longer TMRCA, increasing diversity. The mean value of β was lowest for both balanced and neutral SNPs of frequency 0.5, which we posit is due to the fact that this allele frequency requires the most time for mutations to drift up to the equilibrium frequency to fix in their allelic class.

While our statistic can be calculated on any window size, previous work has suggested that the effects of balancing selection localize to a relatively narrow region surrounding the balanced site [7]. Ultimately, the optimal window size depends on the recombination rate, as it breaks up allelic classes. We present some mathematical formulations to give guidance on window sizes that are reasonable (**Online Supplement**).

Table 1: Power to detect ancient balancing selection at a 1% False Discovery Rate

	Older Selection*			Newer Selection†		
Method	0.25	0.5	0.75	0.25	0.5	0.75
Beta	0.64	0.67	0.58	0.29	0.36	0.23
HKA	0.40	0.24	0.35	0.13	0.08	0.08
Tajima's D	0.11	0.35	0.11	0.04	0.15	0.04
T1	0.52	0.44	0.58	0.17	0.11	0.15
T2‡	0.75	0.79	0.68	0.31	0.46	0.24

* Selection started 250,000 generations prior; †100,000 prior.

‡ T2 is not a summary statistic method, but instead relies on comparisons with the results of large numbers of simulations. It has ideal power, but is not applicable in many cases.

Results

Power Analysis

We used forward simulations [19] to calculate the power of our approach to detect balancing selection, relative to other commonly utilized statistics. Initially, we considered an equilibrium demographic model, varied over a range of balancing selection equilibrium frequencies and onset times (**Methods**). We computed the power of β, Tajima's D, HKA, and T1 to distinguish between these selective scenarios and neutrality. As a reference, we also measured the likelihood-based statistic, T2.

Compared to other summaries, β had the greatest performance across all parameter combinations (**Table 1**, **Tables S1-7**). As expected, β performs slightly below T2 under these same conditions. However, unlike T2, our method does not require an outgroup sequence, or grids of simulations which are computationally expensive.

We next investigated the power of β under non-neutral demographic scenarios (**Methods**) compatible with recent human history [12]. We found that β also performs well under bottleneck and expansion models. Under an expansion scenario, performance of all methods decreased, consistent with results from previous studies [12], possibly due to the larger population size increasing the expected time until an allele can fix in allelic class. The effect of a population bottleneck on power was less drastic, and led to a slight increase in power to detect newer selection (**Tables S2 and S3**).

We next examined the power for β with variable mutation rate, recombination rates, and sample sizes (**Methods**). As expected, the power of β was positively correlated with mutation rate (**Table S7 and S9**), and negatively correlated with recombination rate (**Table S6 and S8**). A higher mutation rate provides more variants that can accumulate within an allelic class, whereas a lower recombination rate allows for longer haplotypes upon which mutations can accumulate. Finally, β has reasonable power down to very small sample sizes, achieving near maximum power with as few as 20 sampled chromosomes **(Fig. S4)**.

In our initial formulation, knowledge of the ancestral state for each site is required. Therefore, we developed a version of the statistic based on a folded site frequency spectrum (Online Supplement). Simulations show that the power of the folded statistic is similar to the unfolded version at intermediate al-

bioRxiv preprint doi: https://doi.org/10.1101/112870; this version posted March 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

lele frequencies, but suffers at extreme frequencies. However, even at these frequencies, it still outperforms Tajima's D, the only available test that does not require knowledge of ancestral state or an outgroup **(Online Supplement)**.

Genome-wide Scan in Human Populations

We applied the unfolded version of β to population data obtained by the 1000 Genomes Project (Phase 3) to detect signatures of balancing selection [15]. We calculated the value of β in 1kb windows for each SNP in all 26 populations. We focused on regions that passed sequencing accessibility filters and whose center SNP had a minimum frequency of 15% in at least one population. We defined extreme β scores as those in the top 1% genome-wide in the population under consideration (**Methods**). As we are specifically interested in selection that may have started before the split of modern populations, we further filtered for loci that were top-scoring in at least half of the populations tested (**Methods**).

We identified 8,702 autosomal variants and 317 on the X-chromosome that were shared among at least half (≥ 13) of the 1000 Genomes populations we analyzed **(Supplementary Data)**. Together, these variants comprise 2,453 distinct autosomal and 86 X-chromosomal loci. There signatures overlapped 692 autosomal and 29 X-chromosomal genes.

Characterization of Identified Signals

Trans-species haplotypes (trHaps) are two or more variants that occur on the same haplotype and are also shared between humans and a primate outgroup (in our case, chimpanzee). trHaps are unlikely to occur by chance and indicate balancing selection [**?**] independently from the signature captured by β. If β captures true signatures of balancing selection, one would expect an enrichment of high β values at trHaps. We found that β was in fact predictive of trHap status, even after including adjustments for distance to nearest gene ($P < 2 \times 10^{-16}$, **Methods**).

Our scan identified several loci that have been previously implicated as putative targets of balancing selection **(Supplementary Data)**. Several major signals occurred on chromosome 6 nearby the HLA, a region presumed to be subjected to adaptive selective pressure. In particular, we found a strong signal in the MHC at a locus influencing response to Hepatitis B infection, rs3077 [20, 12, 21]. In addition, several additional top sites in our scan matched those from DeGiorgio *et al.* [12]. These include previously found sites that tag phenotypic associations [22], including *MYRIP*, involved with sleep-related phenotypes [23], and *BICC1*, associated with corneal astigmatism [24].

A signature of selection at the *CADM2* locus

One of our top-scoring regions fell within an intron of the cell adhesion molecule 2 gene, *CADM2*. This locus contains a haplotype with β scores for the respective variants falling in the top 0.25 percentile in 17 of the the 1000 Genomes populations, and scoring in the top 0.75 percentile across all 26 populations (Fig. 3). This site was also a top scoring SNP in the CEU population based on the T2 statistic [12].

To elucidate the potential mechanisms contributing to the signal in this region, we overlapped multiple genomic datasets to identify potential functional variants that were tightly linked with our haplotype signature. First, one variant that perfectly tags (EUR $r^2 = 1.0$) our signature, rs17518584, has been genome-wide significantly associated with cognitive functions, including information processing speed [25, 26]. Second, multiple variants in this region co-localized (EUR r^2 between $0.9 - 1$ with rs17518584) with eQTLs of *CADM2* in numerous tissues (Lung, Adipose, Skeletal Muscle, Heart-Left Ventricle), though notably not in brain [27]. That said, several SNPs with regulatory potential (RegulomeDB scores of 3a or higher) are also strongly tagged by our high-scoring haplotype (EUR r^2 between $0.9 - 1.0$ with rs17518584), which include regions of open chromatin in Cerebellum and other cell types [28]. Several SNPs on this haplotype, particularly rs1449378 and rs1449379, fall in enhancers in several brain tissues, including the hippocampus [29, 28]. Taken collectively, these data suggest that our haplotype tags a region of regulatory potential

bioRxiv preprint doi: https://doi.org/10.1101/112870; this version posted March 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

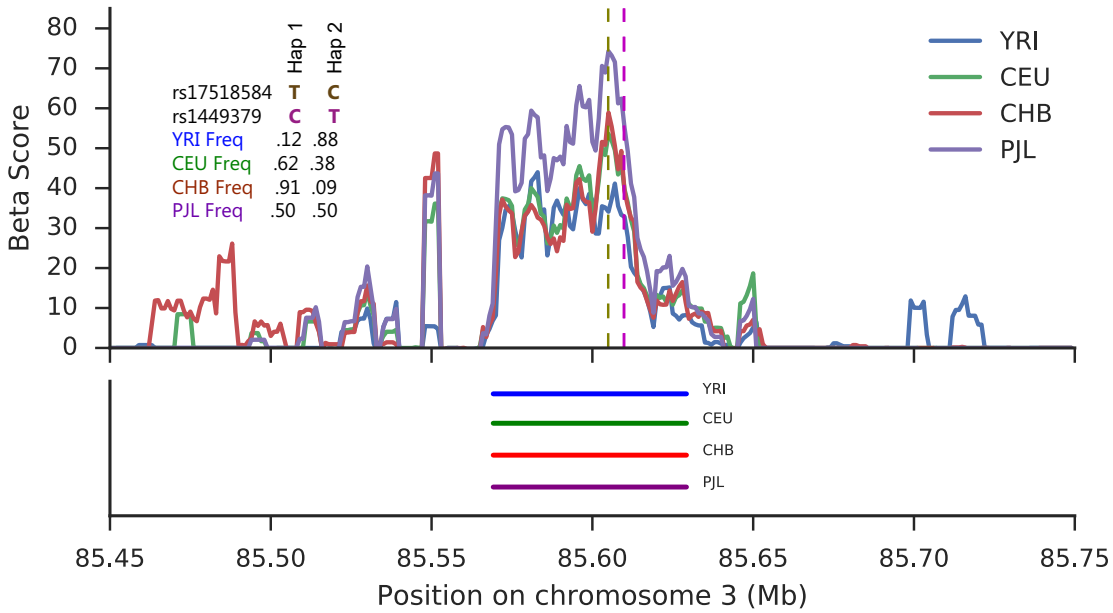

Figure 3: Signal of balancing selection at CADM2. The signal of selection is located in an intron of CADM2. (a) rs17518584 is the lead GWAS SNP for intellectual traits, and is marked by the brown vertical dashed line. The purple dashed line marks two regulatory variants found on the balanced haplotype. Beta Scores were calculated using a rolling average with windows of size 5 kb, including only SNPs at the same frequency as the balanced SNP in the average. In addition, we show the allele frequencies of the GWAS and a top-scoring Beta SNP in each representative population. (b) Approximate haplotype lengths for each population.

that may influence the expression of *CADM2*, and potentially implicates cognitive or neuronal phenotypes in the selective pressure at this site.

A signature of balancing selection near the diabetes associated locus, *WFS1*

We identified a novel region of interest within the intron of *WFS1*, a transmembrane glycoprotein localized primarily to the endoplasmic reticulum (ER). *WFS1* functions in protein assembly [30] and is an important regulator in the unfolded protein and ER Stress Response pathways [31]. A haplotype in this region (approximately 3.5 kb) contains approximately 26 variants, 3 of which are in high-quality windows and are high-scoring β in all populations.

Our identified high-scoring haplotype tags several functional and phenotypic variant associations. First, one variant that perfectly tags our signature (EUR $r^2 = 1.0$), rs4458523, has been previously associated with type 2 diabetes [32, 33]. Second, multiple variants in this region are associated with change in expression of *WFS1* in numerous tissues [27]; these variants are strongly tagged by our high-scoring haplotype (EUR r^2 between $0.85 - 0.9$ with rs4458523). Finally, several SNPs with regulatory potential (RegulomeDB scores of 2b or higher) are also strongly tagged by our high-scoring haplotype (EUR r^2 between $0.9 - 1.0$ with rs4458523). Taken collectively, these data suggest that our haplotype tags a region of strong regulatory potential that is likely to influence the expression of *WFS1*.

bioRxiv preprint doi: https://doi.org/10.1101/112870; this version posted March 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

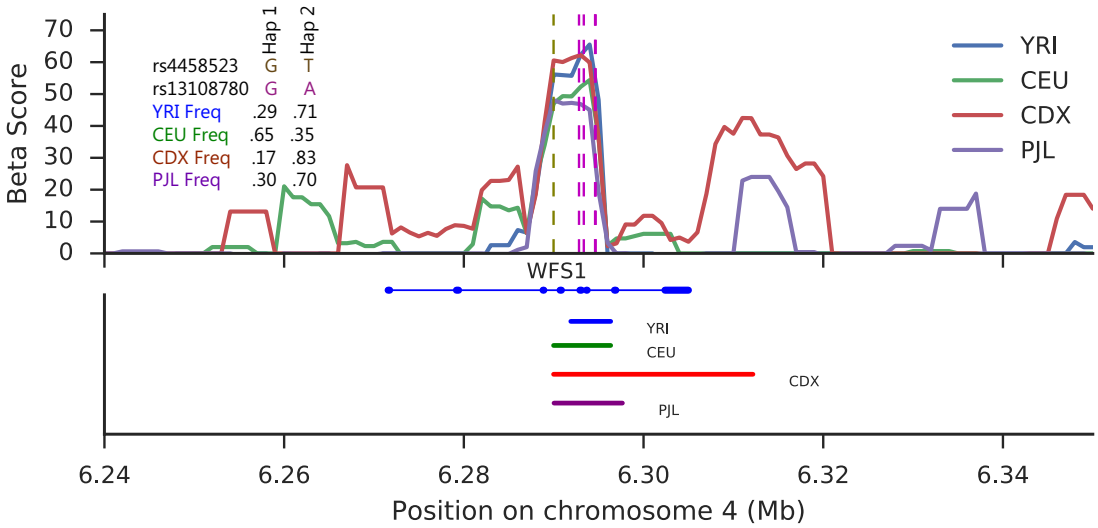

Figure 4: Signal of balancing selection at the WFS1 gene. (a) rs4458523 is the lead GWAS SNP for diabetes, and is marked by the brown vertical dashed line in the top section. The purple dashed line marks 5 regulatory variants variants found on the balanced haplotype. In addition, we show the allele frequencies of the GWAS and a top-scoring Beta SNP in each representative population. (b) Approximate haplotype lengths for each population.

Discussion

Informed by previous theory for allelic-class build-up [16], we developed a summary statistic sensitive to changes in the site frequency spectrum near a target of balancing selection. We show that our statistic has equivalent or improved power to discover these selective events relative to previous summary statistics. Moreover, while our method does not require knowledge of ancestral states for each variant from outgroup sequences, this information can improve power at extreme equilibrium frequencies.

Our method is designed to capture regions of the genome with an increased TMRCA. This type of signature could also be induced by introgression with other species. This alternative is not immediately distinguished or modeled by existing approaches, though signals of trHaps may be robust to this type of genomic signature. Therefore, it is not unreasonable to evaluate interesting regions for evidence of genomic introgression when possible, or to examine populations which do not have strong evidence of historical introgression. Alternatively, regions that contain a strong β signal across many populations can be prioritized, as we do here.

Although our method outperforms existing summary statistic methods, it is not as powerful as computationally-intensive approaches that use simulations to calculate likelihoods of observed data. This does suggest the possibility that there remains additional signal that our method does not fully capture. One additional feature we did consider was an expected reduction in the number of substitutions, but addition of this information, while significant, did not substantially improve discriminatory power (**Online Methods**). Alternative possibilities could include (i) consideration of sequences further beyond the balanced variant than the ancestral region, or (ii) deviations in the frequency spectrum beyond the closely-matched frequency to the balanced SNP we considered here. Further improvements could capture the signal of excess variation that has been decoupled from its allelic class, but has not yet drifted out or

bioRxiv preprint doi: https://doi.org/10.1101/112870; this version posted March 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

been fixed in the population.

Although it is impossible to know the true selective pressures in these cases, these results suggest that balancing selection may be responsible for the maintenance of important phenotypes in human populations. At the *CADM2* locus, functional regulatory variants seem to connect our haplotype signature to changes in gene expression and, coupled with enhancer annotation information and human phenotype associations, suggests that brain or brain-related biology may be the target of selection. Intriguingly, a recent report also noted a strong signature of selection at this locus in canine [34], suggesting a possibility of convergent evolution. That said, the phenotypes that have resulted in a historical fitness trade-off at this locus are far from obvious.

Similarly, speculation on the potential phenotypes subject to balancing selection at *WFS1* should also be interpreted cautiously. It is known that autosomal recessive, loss of function mutations in this gene cause Wolfram Syndrome, characterized by diabetes insipidus, diabetes mellitus, optic atrophy, and deafness. Studies that have investigated the role of this gene in diabetes have determined that this gene is a component of the unfolded protein response [31] and is involved with ER maintenance in pancreatic β-cells. Furthermore, deficiency of *WFS1* results in increased ER stress, impairment of cell cycle, and ultimately increased apoptosis of beta-cells [35]. These observations provide a straightforward, mechanistic hypothesis for the common variant associated in the region: that this common variant lowers the expression of *WFS1* in pancreas in some manner. However, the eQTLs that co-localized with the diabetes association link the risk increasing allele with *higher* levels of expression of *WFS1*, at least in non-pancreas tissue. Thus, the functional mechanism may be more complex. How the unfolded protein response could connect to historical balancing selection is also not immediately obvious. A possibility derives from recent work suggesting that these pathways respond not only to stimulus from nutrients or ER stress, but also from sensing pathogens [36]. This could suggest the possibility that expression of *WFS1* is optimized in part to respond to pathogen exposure at a population level.

Our approach is powered to detect balancing selection when clear outgroup sequences are not immediately available, when detailed demographies for the study population are not known in detail, and with a relatively small sample of chromosomes from the population. Given the increasing ease of collecting population genetic data from non-model organisms, our approach should provide an efficient way to characterize balancing selection in these populations.

An implementation of both the folded and unfolded versions of β is available for download at: https://github.com/ksiewert/BetaScan.

Methods

Simulations

Simulations were performed using the simulation software SLiM 2.0 [19]. We generated two types of simulations: without a balanced loci and with a balanced loci. Paremeters are provided in the Supplementary methods.

Empirical Site Analysis

To apply our method to 1000 genomes data, we first downloaded data for each of the 26 populations in phase 3 of the project (obtained May 02, 2013). We then calculated allele frequencies separately for each population, and calculated β in 1 kb sized windows centered around each SNP for each population. Because poorly sequenced regions can artificially inflate the number of SNPs in a region, we then filtered out regions that contained one or more basepairs that were ruled as poor quality in the 1000 Genomes phase III strict mask file. For further confirmation that the signal was not a result of poor mapping quality, we overlapped SNPs of interest with hg19 human RepeatMasker regions, downloaded from the UCSC Table Browser on February 9th, 2017. We then removed all core SNPs from consideration that were found within a Repeat, similar to what was done in [37]. We further removed SNPs from consideration that were not of common frequency (at or above 15%) in at least one population. We then found the top 1%

bioRxiv preprint doi: https://doi.org/10.1101/112870; this version posted March 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

of these high quality SNPs in each population. After filtering, there were 1,803,299 SNPs in the genome that remained.

The lowest significance cut-off of any population, ASW, corresponds to a Beta score of 49.47, which scores in the top 0.5% of neutral simulations with an equilibrium frequency of 0.5 (**Fig. S3**). To find top-scoring sites that are also GWAS hits, we obtained LD proxies for all SNPs that were in the top 1% of at least half of the populations in European populations, using a cut-off of r^2 of 0.9 and a maximum distance of 50kb and a minimum minor allele frequency of 5%. We then overlapped these LD proxies with GWAS hits obtained from the GWAS Catalog to get our final list of putatively balanced GWAS hits [22]. Gene names and locations were downloaded from Ensembl BioMart on November 26th, 2016.

For our trSNP comparison, we used the Human/Chimp shared haplotypes from Leffler et al. [**?**]. Using logistic regression, we then modeled the outcome of a SNP being part of a trHap as dependent on the Beta Score and distance to nearest gene.

We would like to acknowledge the members of the Voight Lab for their helpful suggestions and support and Philipp Messer for his comments on this manuscript. This work was supported through grants from the National Institutes of Health (NIDDK R01DK101478) and a fellowship from the Alfred P. Sloan Foundation (BR2012-087) to BFV. KMS was supported in part by National Institutes of Health T32HG000046-17.

bioRxiv preprint doi: https://doi.org/10.1101/112870; this version posted March 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

References

[1] Vitti JJ, Grossman SR, Sabeti PC (2013) Detecting Natural Selection in Genomic Data. *Annual Review of Genetics* 47(1):97–120.

[2] Xu L, et al. (2015) Genomic signatures reveal new evidences for selection of important traits in domestic cattle. *Molecular biology and evolution* 32(3):711–25.

[3] Singh ND, Jensen JD, Clark AG, Aquadro CF (2012) Inferences of Demography and Selection in an African Population of Drosophila melanogaster. *Genetics* 193(1).

[4] Charlesworth D (2006) Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genetics* 2(4):379–384.

[5] Aidoo M, et al. (2002) Protective effects of the sickle cell gene against malaria morbidity and mortality. *The Lancet* 359(9314):1311–1312.

[6] Luzzatto L (2012) Sickle cell anaemia and malaria. *Mediterranean journal of hematology and infectious diseases* 4(1):e2012065.

[7] Gao Z, Przeworski M, Sella G (2015) Footprints of ancient-balanced polymorphisms in genetic variation data from closely related species. *Evolution* 69(2):431–446.

[8] Tajima F (1989) Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. *Genetics* 123(3):585.

[9] Hudson RR, Kreitman M, Aguadé M (1987) A Test of Neutral Molecular Evolution Based on Nucleotide Data. *Genetics* 116(1):153–159.

[10] Andrés AM, et al. (2009) Targets of balancing selection in the human genome. *Molecular Biology and Evolution* 26:2755–2764.

[11] Teixeira JC, et al. (2015) Long-Term Balancing Selection in LAD1 Maintains a Missense Trans-Species Polymorphism in Humans, Chimpanzees, and Bonobos. *Molecular biology and evolution* pp. msv007–.

[12] DeGiorgio M, Lohmueller KE, Nielsen R (2014) A Model-Based Approach for Identifying Signatures of Ancient Balancing Selection in Genetic Data. *PLoS Genetics* 10(8):e1004561.

[13] Rasmussen MD, Hubisz MJ, Gronau I, Siepel A (2014) Genome-wide inference of ancestral recombination graphs. *PLoS genetics* 10(5):e1004342.

[14] Fay JC, Wu CI (2000) Hitchhiking under positive Darwinian selection. *Genetics* 155(3):1405–13.

[15] The 1000 Genomes Consortium (2015) A global reference for human genetic variation. *Nature* 526(7571):68–74.

[16] Hey J (1991) A multi-dimensional coalescent process applied to multi-allelic selection models and migration models. *Theoretical population biology* 39(1):30–48.

[17] Fu Y (1995) Statistical Properties of Segregating Sites. *Theoretical Population Biology* 48(2):172–197.

[18] Watterson G (1975) On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology* 7(2):256–276.

[19] Haller BC, Messer PW (2017) SLiM 2: Flexible, Interactive Forward Genetic Simulations. *Molecular biology and evolution* 34(1):230–240.

[20] Jiang DK, et al. (2015) Genetic variants in five novel loci including CFB and CD40 predispose to chronic hepatitis B. *Hepatology* 62(1):118–128.

[21] Thursz MR, Thomas HC, Greenwood BM, Hill AV (1997) Heterozygote advantage for HLA class-II type in hepatitis B virus infection. *Nature Genetics* 17(1):11–12.

[22] Welter D, et al. (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic acids research* 42(Database issue):D1001–6.

bioRxiv preprint doi: https://doi.org/10.1101/112870; this version posted March 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

[23] Gottlieb DJ, O'Connor GT, Wilk JB (2007) Genome-wide association of sleep and circadian phenotypes. *BMC Medical Genetics* 8(Suppl 1):S9.

[24] Lopes MC, et al. (2013) Identification of a Candidate Gene for Astigmatism. *Investigative Opthalmology & Visual Science* 54(2):1260.

[25] Davies G, et al. (2015) Genetic contributions to variation in general cognitive function: a meta-analysis of genome-wide association studies in the CHARGE consortium (N=53,949). *Molecular psychiatry* 20(2):183–92.

[26] Ibrahim-Verbaas CA, et al. (2016) GWAS for executive function and processing speed suggests involvement of the CADM2 gene. *Molecular Psychiatry* 21(2):189–197.

[27] The GTEx Consortium (2015) The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* 348(6235):648–660.

[28] Boyle AP, et al. (2012) Annotation of functional variation in personal genomes using RegulomeDB. *Genome Research* 22(9):1790–1797.

[29] Ernst J, Kellis M (2012) ChromHMM: automating chromatin-state discovery and characterization. *Nature Methods* 9(3):215–216.

[30] Takei D, et al. (2006) WFS1 protein modulates the free Ca2+ concentration in the endoplasmic reticulum. *FEBS Letters* 580(24):5635–5640.

[31] Fonseca SG, et al. (2005) WFS1 is a novel component of the unfolded protein response and maintains homeostasis of the endoplasmic reticulum in pancreatic beta-cells. *The Journal of biological chemistry* 280(47):39609–15.

[32] Voight BF, et al. (2010) Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nature Genetics* 42(7):579–589.

[33] Mahajan A, et al. (2014) Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nature Genetics* 46(3):234–244.

[34] Freedman AH, et al. (2016) Demographically-Based Evaluation of Genomic Regions under Selection in Domestic Dogs. *PLOS Genetics* 12(3):e1005851.

[35] Yamada T, et al. (2006) WFS1-deficiency increases endoplasmic reticulum stress, impairs cell cycle progression and triggers the apoptotic pathway specifically in pancreatic -cells. *Human Molecular Genetics* 15(10):1600–1609.

[36] Nakamura T, et al. (2010) Double-stranded RNA-dependent protein kinase links pathogen sensing with stress and metabolic homeostasis. *Cell* 140(3):338–48.

[37] Bubb KL, et al. (2006) Scan of human genome reveals no new Loci under ancient balancing selection. *Genetics* 173(4):2165–77.