# BetaScan2: Standardized Statistics to Detect Balancing Selection Utilizing Substitution Data

Katherine M. Siewert [1] and Benjamin F. Voight[2,3,4,*]

[1]Genomics and Computational Biology Graduate Group, Perelman School of Medicine, University of Pennsylvania

[2]Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania

[3]Department of Genetics, Perelman School of Medicine, University of Pennsylvania

[4]Institute for Translational Medicine and Therapeutics, Perelman School of Medicine, University of Pennsylvania

*Corresponding author: E-mail: bvoight@pennmedicine.upenn.edu.

**Accepted:** January 18, 2020

## Abstract

Long-term balancing selection results in a build-up of alleles at similar frequencies and a deficit of substitutions when compared with an outgroup at a locus. The previously published $\beta^{(1)}$ statistics detect balancing selection using only polymorphism data. We now propose the $\beta^{(2)}$ statistic which detects balancing selection using both polymorphism and substitution data. In addition, we derive the variance of all $\beta$ statistics, allowing for their standardization and thereby reducing the influence of parameters which can confound other selection tests. The standardized $\beta$ statistics outperform existing summary statistics in simulations, indicating $\beta$ is a well-powered and widely applicable approach for detecting balancing selection. We apply the $\beta^{(2)}$ statistic to 1000 Genomes data and report two missense mutations with high $\beta$ scores in the *ACSBG2* gene. An implementation of all $\beta$ statistics and their standardization are available in the BetaScan2 software package at https://github.com/ksiewert/BetaScan.

**Key words:** balancing selection, selection scans, human evolution, selection statistics.

## Introduction

Balancing selection can maintain multiple alleles at a locus. Several scenarios can result in this type of selection, including heterozygote advantage or frequency-dependent selection. Recently, there have been significant methodological advances in the detection of balancing selection in population-level sequencing data (Siewert and Voight, 2017; DeGiorgio et al., 2014; Cheng and DeGiorgio, 2018; Bitarello et al., 2018). As a result of this recent interest, a number of specific balanced loci in a wide variety of species have been reported with strong observational or experimental evidence (Wheat et al., 2010; Network, 2015; Schweizer et al., 2018; Sano et al., 2018). These results suggest there may be more loci yet to be discovered.

Selection can maintain multiple balanced alleles for very long periods of time, referred to as long-term balancing selection. By doing so, selection increases the time to most recent common ancestor (TMRCA) (Charlesworth, 2006). The resulting signature can be detected to identify putatively balanced loci. Classically, this signature was defined as an excess of heterozygosity and a deficit of substitutions. This is due to variants building up around the balanced alleles through time but being prevented from fixing in the population (Hudson et al., 1987; Tajima, 1989). However, power analyses demonstrate that these classic methods have low power.

More recent methods detect a more precise signature in the site frequency spectrum, and as a result, are higher powered. Specifically, they detect a build-up of variants at highly similar frequencies (Siewert and Voight, 2017; Bitarello et al., 2018). This build-up is most likely a main contribution to the signal that is captured implicitly in methods that rely on simulations of balancing selection to generate an expected site frequency spectrum (DeGiorgio et al., 2014; Cheng and DeGiorgio, 2018). A build-up occurs because two allelic classes are maintained in a population, accumulating variation throughout time. These variants can fix within each haplotype class, and will therefore all be at the same frequency until their frequency is altered by recombination. This signature is not independent of excess heterozygosity, but is instead a more specific signature of the same underlying process.

Simulation-based methods have been shown to be the highest-powered methods, however they require additional types of data rendering them inapplicable in some situations (e.g., large grids of simulations of balancing selection, a sequenced genome from a closely related outgroup and/or genome-wide data in order to estimate the background site frequency spectrum; DeGiorgio et al., 2014; Cheng and DeGiorgio, 2018). In contrast, the $\beta^{(1)}$ statistics are nearly as high powered, yet do not require additional types of data, enabling wide-spread application.

The previously developed method $\beta$ detects the signature of alleleic class build-up around each SNP it is applied to (i.e., each core SNP) (Siewert and Voight, 2017) by looking for an excess of genomically proximate SNPs that have a very similar allele frequency to the core SNP. Conceptually, $\beta$ is a weighted sum of SNP counts around each core SNP, where SNPs are weighted more if they have very similar frequencies to the core SNP. By calculating a $\beta$ score in a window around each SNP, one can find loci in which the pattern of allele frequencies is consistent with the action of balancing selection. Mathematically, $\beta$ is the difference between two estimators of the mutation rate $\widehat{\theta}$, one sensitive to excess frequency similarity ($\widehat{\theta}_\beta$) and one that is not (Watterson's estimator $\widehat{\theta}_W$) (Watterson, 1975). We previously reported two versions of $\beta$: $\beta^{(1)}$ incorporates outgroup sequence data to call ancestral allele states, whereas $\beta^{(1)*}$ only requires a folded site frequency spectrum.

We now report two improvements to $\beta$. The first is a new estimator based on the number of fixed differences with an outgroup species (i.e., substitutions), $\widehat{\theta}_D$. Incorporating this estimator into our $\beta$ framework increases power, owing to the expected reduction of substitutions under balancing selection (Hudson et al., 1987; Charlesworth, 2006). Second, we derive the variance of our statistics, enabling normalization of $\beta$. This allows $\beta$ scores to be properly compared across a range of parameters which can affect its distribution, a feature lacking from other summary statistics, with the exception of Tajima's *D* (Tajima, 1989).

## Results and Discussion

We measure the power of our statistics using simulations conducted using SLiM 2 (supplementary information, Supplementary Material online) (Haller and Messer, 2017). We find that $\beta^{(2)}$ has higher power than either $\beta^{(1)}$ statistic, with the greatest gain in power at equilibrium frequencies 25% and 75%, demonstrating that substitution counts provide additional signal over polymorphism data (fig. 1*A*, supplementary figs. 5 and 6,Supplementary Material online). When there is mutation rate variation across simulations, we find that standardization improves power (fig. 1*B*).

Next, we compare power with alternative methods: NCD2 (Bitarello et al., 2018), NCD$_{\mathrm{mid}}$ (Cheng and DeGiorgio, 2018), *T*1 and *T*2 (DeGiorgio et al., 2014), Tajima's *D* (Tajima, 1989), and the HKA test (Hudson et al., 1987). *T*2 and $\beta$ (fig. 1*C*) perform similarly. However, $\beta$ does not require large grids of simulations as input and can be applied when an outgroup is not available. These two statistics significantly outperform NCD2, Tajima's *D*, and the HKA test (fig. 1*C*, supplementary figs. 5 and 6, Supplementary Material online). Our results were not biased by window size (supplementary fig. 7, Supplementary Material online) or power comparison method (supplementary information, Supplementary Material online).

The power to detect balancing selection depends on underlying population parameters, including mutation rate, recombination rate and effective population size. We found that consistent with prior reports, the power of all methods increased with lower recombination rate (supplementary fig. 8, Supplementary Material online) and higher mutation rate (supplementary fig. 9, Supplementary Material online) (DeGiorgio et al., 2014; Siewert and Voight, 2017). This is because a higher mutation rate results in more SNPs fixing in allelic class, whereas a lower mutation rate results in less recombination decoupling SNPs from the balanced SNP. We found that power decreases with increasing population size (supplementary fig. 10, Supplementary Material online). This is expected because a higher population size increases the expected TMRCA at a neutral locus, resulting in a smaller difference in TMRCA between balanced and neutral loci.

We next scan the human genome for evidence of balancing selection, applying the $\beta^{(2)}_{\mathrm{std}}$ method (supplementary information, Supplementary Material online) to the YRI, CEU, and CHB populations from the 1000 genomes project (The 1000 Genomes Consortium, 2015) (supplementary fig. 11, Supplementary Material online). To prepare files for input into BetaScan we used the glactools toolkit, which can output in BetaScan format with the option to include substitutions (Renaud, 2018) (supplementary information, Supplementary Material online). As expected, there is a high correlation between the three $\beta$ statistics, the unstandardized and standardized statistics (supplementary tables 1 and 2, Supplementary Material online), and scores in the three populations (supplementary tables 3 and 4, Supplementary Material online).

We next tested if $\beta^{(2)}$ captures genomic regions which are likely true positives. Trans-species haplotypes between human and chimpanzee are an independent measure of balancing selection and perhaps the closest available thing to a set of true positive loci under long-term balancing selection. We found that SNPs in trans-species haplotypes from Leffler et al. (2013) had a significantly higher $\beta^{(2)}_{\mathrm{std}}$ value than SNPs which are not (Mann–Whitney *U* *P*-value $= 1.08 \times 10^{-14}$). We also compared the values of $\beta^{(2)}_{\mathrm{std}}$ versus T2 in SNPs in trans-haplotypes. We found that after removing rare variants, the mean percentile for SNPs in these haplotypes was 0.63 for $\beta^{(2)}_{\mathrm{std}}$ and 0.66 for T2. Before removing rare variants, the mean percentile for $\beta^{(2)}_{\mathrm{std}}$ was 0.64 and T2 was 0.72. This indicates that $\beta^{(2)}_{\mathrm{std}}$ may have additional noise when calculated on rare variants in real data due to demography that is not captured in the theoretical site frequency spectrum. These effects will be captured in the empirical neutral distribution used by *T*2, increasing its power. As with the $\beta^{(1)}$ statistics, we therefore

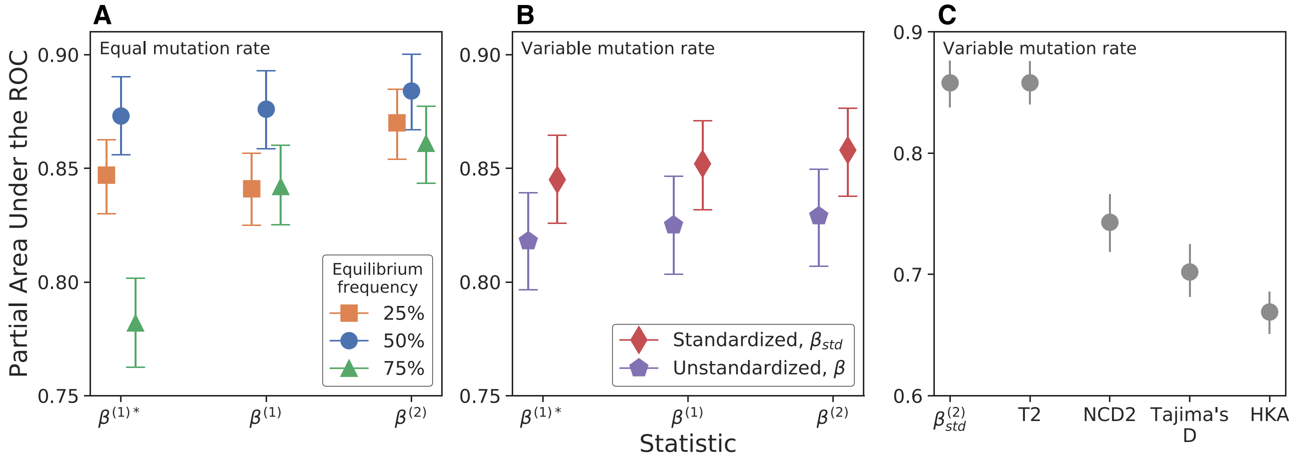

Fig. 1.—Partial area under the receiver operator curve (ROC) from a false positive rate of 0 to 0.05 in simulations for each statistic under (*A*) different equilibrium frequencies and (*B*) with mutation rate variation, where one half of neutral and balanced simulation replicates had a mutation rate of 2.5 × $10^{-8}$ (our default rate), and the remaining half had a rate of 1.25 × $10^{-8}$. (*C*) Power of $\beta^{(2)}$ compared with other methods for detecting balancing selection. An equilibrium frequency of 50% was used for (*B*) and (*C*). The values of each statistic were compared between simulations containing only neutral variants (True Negatives) or with a balanced variant (True Positives). Confidence intervals show the 2.5th to 97.5th percentile for 1,000 sets of bootstrapped simulation replicates.

recommend only applying $\beta^{(2)}$ to SNPs which are not rare. Because balanced variants are unlikely to be maintained at extreme equilibrium frequencies (Ewens and Thomson, 1970; Nei and Roychoudhury, 1973), this should not hurt power.

If $\beta^{(2)}_{\text{std}}$ is capturing true balanced loci, we would expect scores to be higher in regions with a demonstrated functional importance, such as eQTLS or GWAS SNPs. To test this hypothesis, we overlapped our top SNPs with the GWAS catalog (Buniello et al., 2019) (supplementary information, Supplementary Material online), and significant eQTLs from the GTeX project (The GTEx Consortium, 2015). Using logistic regression, we found that both standardized and unstandardized $\beta^{(2)}$ were positively predictive of a SNP being an eQTL (supplementary table 5, Supplementary Material online) or in the GWAS catalog (supplementary table 6, Supplementary Material online). This remained true after taking into account potential confounding factors such as minor allele frequency and distance to the nearest gene. Reassuringly, included in our top hits are previously discovered loci, including several loci in the HLA region (supplementary fig. 11, Supplementary Material online).

Two missense SNPs in the gene *ACSBG2*, rs17856650 and rs17856651, have standardized $\beta$ scores in the top 99.9 percentile in the CEU population (fig. 2), top 99.5 percentile in the CHB population and top 99 percentile in the YRI population. Confirming this finding, this gene was one of the top 100 loci detected using the *T2* statistic (DeGiorgio et al., 2014), but has been previously uncharacterized. In all three populations, the percentile of the standardized $\beta^{(2)}$ score is slightly higher than the unstandardized, demonstrating the power of standardization in regions with a lower mutation rate, such as within genes. The Acyl-CoA Synthetase Bubblegum Family Member 2 (ACSBG2) gene is involved in fatty acid metabolism (Pei et al., 2006). ACSBG2 may play a role in spermatogenesis (Fraisl et al., 2006), a process that has been previously associated with balancing selection (Bitarello et al., 2018). In addition, this gene was highlighted as harboring potentially deleterious lineage-specific nonsynonymous single nucleotide changes in bonobo (Han et al., 2019). However, the reported bonobo changes are at a different location in the gene than the two missense SNPs in humans.

The results presented in this manuscript demonstrate the power and wide-applicability of the BetaScan2 suite of statistics. In addition to the ability to standardize $\beta$ and the implementation of $\beta^{(2)}$, BetaScan2 can calculate $\beta_{\text{std}}$ on a predefined window of interest, instead of requiring the use of a sliding window across the genome. This locus-based calculation is made possible by standardization allowing comparison of scores across different window sizes, and enables the use of $\beta$ on species for which only a small fraction of the genome is available. BetaScan2 is freely available at https://github.com/ksiewert/BetaScan.

## Materials and Methods

We first derive a closed-form solution for the expected number of substitutions under the neutral model (supplementary information, Supplementary Material online). This expression leads to an estimator of the mutation rate $\widehat{\theta}_D$ based upon the number of substitutions *D*:

$$\widehat{\theta}_D = \frac{D}{T/(2N_e) + 1/n}. \tag{1}$$

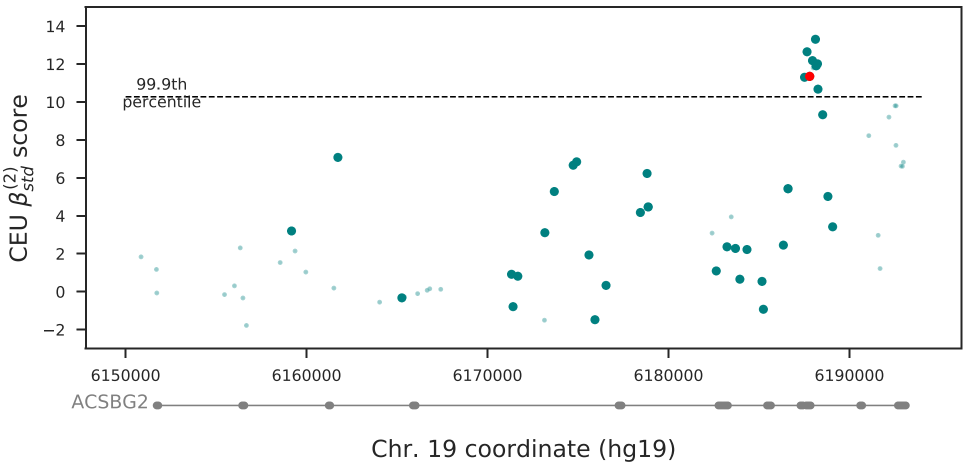

**Fig. 2.**—The ACSBG2 locus shows evidence of balancing selection in the CEU population. The red dot indicates two missense mutations. The large circles indicate SNPs at the frequency of the putatively balanced haplotype, whereas the small circles indicate SNPs at other frequencies. Only SNPs passing our filtering for quality are plotted.

Here $T$ is the divergence time between the two species, $N_e$ is the effective population size, and $n$ is the number of surveyed chromosomes. To incorporate information from substitutions, we replace $\widehat{\theta}_W$ from the original unfolded statistic with $\widehat{\theta}_D$ to define $\beta^{(2)}$:

$$\beta^{(2)} = \widehat{\theta}_\beta - \widehat{\theta}_D. \quad (2)$$

Under long-term balancing selection, variants nearby the selected site are maintained at similar allele frequencies rather than fixing in the population, resulting in an increased estimate of $\theta_\beta$ and reduced estimate of $\theta_D$. Therefore, $\beta$ is expected to score above zero in the presence of long-term balancing selection, whereas it will be below or around zero under neutrality.

Intuitively, the variances of the $\beta$ statistics depend on the population size, survey sample size, equilibrium frequency, and mutation rate. We derive the variances of $\beta^{(1)*}$ and $\beta^{(2)}$ (supplementary material, Supplementary Material online), allowing for comparison of scores across samples, window sizes and genomic loci where these factors may differ. Our expressions for variance match simulation results (supplementary figs. 1 and 2, Supplementary Material online). To obtain the variance of $\beta^{(1)}$ we used the framework reported in Achaz (2009) (supplementary information, Supplementary Material online). We note that the expected value of all $\beta$ statistics is zero. This leads to, for $\beta^{(2)}$ for instance:

$$\beta^{(2)}_{\text{std}} = \frac{\beta^{(2)}}{\sqrt{\text{Var}[\beta^{(2)}]}} = \frac{\widehat{\theta}_\beta - \widehat{\theta}_D}{\sqrt{\text{Var}[\widehat{\theta}_\beta] + \text{Var}[\widehat{\theta}_D]}}. \quad (3)$$

Full derivations for $\beta^{(1)}_{\text{std}}$, $\beta^{(1)*}_{\text{std}}$, and $\beta^{(2)}_{\text{std}}$ can be found in the supplementary information, Supplementary Material online. Calculating the variance of all three $\beta$ statistics requires an estimate of $\widehat{\theta}$, the underlying mutation rate. The variance of $\beta^{(2)}$ also requires an estimate of the speciation time. We discuss techniques for estimation of these parameters in the supplementary information, Supplementary Material online. However, the $\beta$ statistics are robust to some specification error (supplementary figs. 3 and 4, Supplementary Material online). We recommend $\beta^{(1)*}_{\text{std}}$ in cases without outgroup data to polarize ancestral states, $\beta^{(1)*}_{\text{std}}$ when ancestral states are known but no informative outgroup is available to call substitutions, and $\beta^{(2)}_{\text{std}}$ when substitutions are available.

## Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

## Acknowledgments

The authors would like to thank Gabriel Renaud for his gracious assistance integrating glactools with BetaScan, along with the members of the Voight Lab for their helpful feedback. This study was supported by the US National Institutes of Health (R01 DK101478 to B.F.V. and a T32 HG000046 and T32 DK110919 to K.M.S.), and a Linda Pechenik Montague Investigator award (to B.F.V.).

## Literature Cited

Achaz G. 2009. Frequency spectrum neutrality tests: one for all and all for one. Genetics 183(1):249–258.

Bitarello BD, et al. 2018. Signatures of long-term balancing selection in human genomes. Genome Biol Evol. 10(3):939–955.

Buniello A, et al. 2019. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 47(D1):D1005–D1012.

Charlesworth D. 2006. Balancing selection and its effects on sequences in nearby genome regions. PLoS Genet. 2(4):e64–384.

Cheng X, DeGiorgio M. 2018. Detection of shared balancing selection in the absence of trans-species polymorphism.Mol Biol Evol. 36(1):177–199.

DeGiorgio M, Lohmueller KE, Nielsen R. 2014. A model-based approach for identifying signatures of ancient balancing selection in genetic data. PLoS Genet. 10(8):e1004561.

Ewens WJ, Thomson G. 1970. Heterozygote selective advantage. Ann Hum Genet. 33(4):365–376.

Fraisl P, et al. 2006. A novel mammalian bubblegum-related acyl-CoA synthetase restricted to testes and possibly involved in spermatogenesis. Arch Biochem Biophys. 451(1):23–33.

Haller BC, Messer PW. 2017. SLiM 2: flexible, interactive forward genetic simulations. Mol Biol Evol. 34(1):230–240.

Han S, Andre AM, Marques-Bonet T, Kuhlwilm M. 2019. Genetic variation in pan species is shaped by demographic history and harbors lineage-specific functions. Genome Biol Evol. 11(4):1178–1191.

Hudson RR, Kreitman M, Aguadé M. 1987. A test of neutral molecular evolution based on nucleotide data. Genetics 116(1):153–159.

Leffler EM, et al. 2013. Multiple instances of ancient balancing selection shared between humans and chimpanzees. Science 340(6127):1578–1582.

Nei M, Roychoudhury AK. 1973. Probability of fixation and mean fixation time of an overdominant mutation. Genetics 74(2):371–380.

Network MGE. 2015. A novel locus of resistance to severe malaria in a region of ancient balancing selection. Nature 526(7572): 253–257.

Pei Z, Jia Z, Watkins PA. 2006. The second member of the human and murine "bubblegum" family is a testis- and brainstem-specific Acyl-CoA synthetase. J Biol Chem. 281(10):6632–6641.

Renaud G. 2018. Glactools: a command-line toolset for the management of genotype likelihoods and allele counts. Bioinformatics 34(8):1398–1400.

Sano EB, Wall CA, Hutchins PR, Miller SR. 2018. Ancient balancing selection on heterocyst function in a cosmopolitan cyanobacterium. Nat Ecol Evol. 2(3):510–519.

Schweizer RM, et al. 2018. Natural selection and origin of a melanistic allele in North American Gray Wolves. Mol Biol Evol. 35(5):1190–1209.

Siewert KM, Voight BF. 2017. Detecting long-term balancing selection using allele frequency correlation. Mol Biol Evol. 34(11):2996–3005.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123(3):585–595.

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

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

Watterson G. 1975. On the number of segregating sites in genetical models without recombination. Theor Popul Biol. 7(2):256–276.

Wheat CW, Haag CR, Marden JH, Hanski I, Frilander MJ. 2010. Nucleotide polymorphism at a gene (Pgi) under balancing selection in a butterfly metapopulation. Mol Biol Evol. 27(2):267–281.

**Associate editor:** Laura A. Katz