BetaScan2: Standardized statistics to detect balancing selection utilizing substitution data

Katherine M. Siewert and Benjamin F. Voight

1 Genomics and Computational Biology Graduate Group, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA
2 Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA
3 Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA
4 Institute for Translational Medicine and Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA

*Corresponding author: bvoight28@gmail.com

December 16, 2018

Abstract

The recently reported statistic $\beta$ detects balanced haplotypes without the need for genomic data from an outgroup species. Here we present an extension to the method that incorporates between-species substitution data into the $\beta$ statistic framework. We show that this approach outperforms existing summary statistics in simulations. We also present the variance of $\beta$ with and without substitution data, allowing calculation of a standardized score. Besides providing a measure of significance, this enables a proper comparison of $\beta$ values across varying underlying parameters, a feature lacking from some related methods.

If multiple alleles at a locus are advantageous, then balancing selection can maintain this diversity for long periods of time (Charlesworth, 2006). Statistics designed to detect balanced loci search for the resulting deviations in the patterns of genetic variation (Tajima, 1989; Hudson et al., 1987).

The previously developed method $\beta$ identifies a specific signature of long-term balancing selection: the build-up of polymorphisms at near-identical frequencies to one another (Siewert and Voight, 2017). $\beta$ is the difference between two estimators of the mutation rate $\hat{\theta}$, one sensitive to the effects of balancing selection ($\hat{\theta}_\beta$) and one that is not (Watterson’s estimator $\hat{\theta}_W$). There are two versions of $\beta$: $\beta^{(1)}$ incorporates outgroup sequence data when available, while $\beta^{(1)*}$ only requires a folded site frequency spectrum.

We 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), $\hat{\theta}_D$. Incorporating this estimator into our $\beta$ framework increases power, owing to the expected reduction of substitutions under balancing selection (Charlesworth, 2006; Hudson et al., 1987). Second, we derive the variance of our statistics, enabling standardization of $\beta$.

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

$$\hat{\theta}_D = \frac{S}{T N_e + \frac{1}{n}} \quad (1)$$

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 $\hat{\theta}_W$ from the
original unfolded statistic with $\hat{\theta}_D$ to define $\beta(2)$:

$$\beta(2) = \hat{\theta}_\beta - \hat{\theta}_D$$

(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 substantially above zero in the presence of long-term balancing selection.

Intuitively, the variances of the $\beta$ statistics depend on the population size, survey sample size, equilibrium frequencies, and mutation rate. We derive the variances of $\beta(1)^*$ and $\beta(2)$ (Supplementary Material), allowing for comparison of scores across samples and genomic loci where these factors may differ. Our expressions for variance match simulation results (Supplementary Fig. 1 and 2). To obtain the variance of $\beta(1)^*$ we used the framework reported in Achaz (2009) (Supplementary Material). The expected value of all $\beta$ statistics is zero, leading to:

$$\beta_{std}(2) = \frac{\beta(2)}{\sqrt{\text{Var}[\beta(2)]}} = \frac{\hat{\theta}_\beta - \hat{\theta}_D}{\sqrt{\text{Var}[\hat{\theta}_\beta] + \text{Var}[\hat{\theta}_D]}}$$

(3)

Full derivations for $\beta_{std}(1)$, $\beta_{std}(1)^*$, and $\beta_{std}(2)$ can be found in the Supplementary Material. Calculating the variance of all three $\beta$ statistics requires an estimate of $\hat{\theta}$, 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 Material. However, the $\beta$ statistics are robust to some specification error (Supplementary Fig. 5 and 6). We recommend $\beta_{std}(1)^*$ in cases without outgroup data to polarize ancestral states, $\beta_{std}(1)$ when ancestral states are known but no informative outgroup is available to call substitutions, and $\beta_{std}(2)$ when substitutions are available.

We measure the power of our statistics using simulations (Supplementary Material). We find that $\beta(2)$ has higher power than either $\beta(1)$ statistic, demonstrating that substitution counts provide additional signal over polymorphism data (Fig. 1a, Supplementary Fig. 3 and 4). When there is mutation rate variation

Figure 1: Power of $\beta$ statistics at (A) different equilibrium frequencies and (B) with mutation rate variation, where one half of neutral and balanced simulation replicates had a mutation rate of $2.5 \times 10^{-8}$ (our default rate), and the remaining half had a rate of $1.25 \times 10^{-8}$. (C) Power of $\beta(2)$ compared to 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).
across simulations, we find that standardization improves power (Fig. 1b).

Next, we compare power to alternative methods: \( NCD_{2} \) (Bitarello et al., 2018), \( NCD_{mid} \) (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). When there is mutation rate variation, we find that \( \beta_{std}^{(2)} \) performs the strongest (Fig. 1c). However, \( T_{2} \), a method that relies on grids of simulations to generate composite likelihoods, performs best when the mutation rate is stable and the parameters underlying the simulations for \( T_{2} \) match the selection scenario. When there is a mismatch between these, the power of \( T_{2} \) is reduced, and \( \beta \) starts to outperform (Supplementary Fig. 3 and 4). Our results were not biased by window size (Supplementary Material). These analyses show that the \( \beta \) framework is a flexible and effective method to detect balancing selection.

Software to calculate all versions of \( \beta \) is available at https://github.com/ksiewert/BetaScan.

References

Achaz, G. 2009. Frequency spectrum neutrality tests: one for all and all for one. Genetics, 183(1): 249–58.

Bitarello, B. D., de Filippo, C., Teixeira, J. C., et al. 2018. Signatures of Long-Term Balancing Selection in Human Genomes. Genome Biology and Evolution, 10(3): 939–955.

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

Cheng, X. and DeGiorgio, M. 2018. Detection of shared balancing selection in the absence of trans-species polymorphism. Molecular Biology and Evolution.

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

Hudson, R. R., Kreitman, M., and Aguadé, M. 1987. A Test of Neutral Molecular Evolution Based on Nucleotide Data. Genetics, 116(1): 153–159.

Siewert, K. M. and Voight, B. F. 2017. Detecting Long-Term Balancing Selection Using Allele Frequency Correlation. Molecular Biology and Evolution, 34(11): 2996–3005.

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