# Supplementary Information: Detecting Long-term Balancing Selection using Allele Frequency Correlation

Katherine M. Siewert, Benjamin F. Voight

## Methods

### Simulations

Simulations were performed using the simulation software SLiM 2.0 [1]. The size of each population was $N_e = 10,000$. We first simulated these scenarios under parameters suitable for human populations, with mutation and recombination rates of $\pi = r = 2.5$ x $10^{-8}$. Similar to that in [2], we did so under a model of constant population size, of a population bottleneck (population size drops to $N_e = 5,500$ from generations 320,000 to 328,000), and of expansion (population expands to $N_e = 20,000$ at generation 302,000). Two populations (corresponding to human and chimpanzee) were simulated, diverging 250,000 generations prior to the ending of the simulation. 100,000 generations of burn-in was simulated prior to this speciation event.

We simulated both neutral regions, and regions with a balanced variant in the center. In each simulations of balancing selection, we introduced an overdominant balanced allele. In the balanced case, we introduced a balanced SNP in the human population either at the time of speciation, or 150,000 generations after speciation. We conditioned our simulations on the balanced SNP remaining in the population until the final generation. If the mutation was lost, we repeated the simulation until maintenance was achieved. Each balanced SNP had an overdominance coefficient $h$ and selection coefficient $s$. The fitness of the heterozygote is then $1 + hs$, and the fitness of the ancestral and derived homozygotes are 1 and $1 + s$, respectively. We simulated two different $s$ values: $10^{-2}$ (our default) and $10^{-4}$. We simulated six different equilibrium frequencies: $.17, .25, .5, .75, .83$. More extreme equilibrium frequencies were not possible to simulate, because the frequency with which the balanced variant drifted out of the population was very high. Previous studies have not conditioned on maintenance of the balanced polymorphism in the population [2]. However, without conditioning on maintenance, the power analysis captures the fixation/loss probabilities at that equilibrium frequency rather than just the power of the method to detect any balanced loci at the frequency that are maintained.

To increase simulation speed, we rescaled our simulation by a factor of 10

for specified power analyses in the supporting information [3]. A minimum of 2000 simulation replicates was performed for each parameter set. We simulated 10kb regions for each simulation replicate.

### Power Analysis

To calculate the power of each method, we compared the score of the balanced variant in balanced simulations with the score of SNPs matched for equilibrium frequency in neutral simulations. For each neutral simulation replicate, we randomly identified one SNP in the region at a frequency within 10 percent of the equilibrium frequency of the corresponding simulations with a balanced SNP. All power calculations were performed with p=20 for Beta, and w=20 for T1 and T2, unless otherwise specified.

## Choice of $p$ parameter

The power of our method lies in capturing allele frequency correlations. The parameter $p$ controls how similar of allele frequencies to the core site are captured. As $p$ approaches infinity, the only sites that contribute towards $\theta_B$ are those that exactly match the frequency of the core SNP. At $p = 0$, all SNPs contribute the same amount to the estimate of $\hat{\theta}_B$, and so $\hat{\theta}_B$ becomes equivalent to $\hat{\theta}_w$. Simulations show that our method is fairly robust to choice of $p$ (Table S4). The optimal $p$ will depend on the data set at hand. If allele frequency estimates are known to be inaccurate, then a lower $p$ may be more optimal, because variants fixed in allelic class may not accurately be called at exactly the same frequency as the core SNP. In addition, allowing variants at very similar frequency to the core SNP contribute to $\hat{\theta}_B$ allows it to capture SNPs that are very close to fixing in allelic class, or were once fixed in class, but are no longer due to recombination followed by a small amount of drift. In our analysis, we chose a $p = 20$, which gives the most weight to exact frequency matches, and a small amount of weight to very near, but not exact frequencies.

## Substitutions and $\beta$

We also wanted to explore whether taking the number of fixed differences with an outgroup (substitutions) would increase power. Substitutions are used by the HKA, T1 and T2 tests. However, we observed that the number of substitutions does not greatly increase predictive capabilities over that of just $\beta$ values (AIC with logistic regression of 2523 with just Beta values, 6246 with just substitutions and 2434 combined). Thus, we decided to focus our method only on polymorphism data.

## Size of the ancestral region

We want to derive the expected length of the ancestral region around the balanced SNP, i.e., what is the distribution of region sizes around the balanced site where the coalescent tree looks identical to that of the balanced loci, ignoring mutation. The ancestral region is the region starting at the balanced loci, moving outwards in either direction until an observable recombination event has occurred in the history of the sample. An observable recombination event is one in which there was recombination between allelic classes. This concept is similar to that in Gao *et al.* [4]; however we are not concerned with an outgroup species, which simplifies the derivation.

The ancestral region roughly corresponds to the optimal window size to calculate $\beta$ on, because it contains the region in which alleles can fix in allelic class, and have not been decoupled from selection due to recombination. In reality, this may slightly underestimate the optimal window size, because it is possible for a position to "re-fix" in allelic class. In this scenario, a recombination event occurred, then a new mutation arose and drifted up to the balanced frequency.

The probability of recombination between allelic classes is equal to the total coalescent branch length in the allelic class multiplied by the probability of recombination onto the other allelic class. Because we are detecting long-term selection, most of the coalescent branch length will fall into the portion between coalescence within each allelic class and coalescence of the two allelic classes. We can therefore put an upper bound on the size of the ancestral region. The probability of any recombination event occurring at a certain position at any time point in $T$ generations is $\rho$*T, where $\rho$ is the individual recombination rate. The probability of a recombination occurring between a chromosome from allelic class 1 and any chromosome from allelic class 2, given that a recombination event occurs in a chromosome from class 1, is just the frequency of allelic class 2. Similarly, the probability that if a recombination event occurs in class 2, it is with any chromosome from class 1 is just the frequency of allelic class 1. Let $\lambda$ be the rate of observable recombination, in units of basepairs, where $p$ and $q$ are the frequencies of the 2 allelic classes, which must sum to 1 by definition.

$$\lambda = T\rho p + T\rho q$$
$$\lambda = T\rho$$

The distribution of ancestral segments on either side of the balanced loci is then exponential with rate parameter $T\rho$.

For our analysis of the 1000 Genomes Project, we are focusing on detecting events that occurred after a split with chimpanzee, but that are old enough that our method has power. Assuming a recombination rate of 2.5e-8 and a split time of $250,000$ generations prior with selection starting at the same time, the 95 quantile on either side is then 479. The most recent events we can hope to detect are closer to $100,000$ generations prior to present, giving a 95th quantile

of 1198 bases on either side of the core SNP. Therefore, for our analysis, we choose 500 basepairs on either side, for a total window size of 1kb.

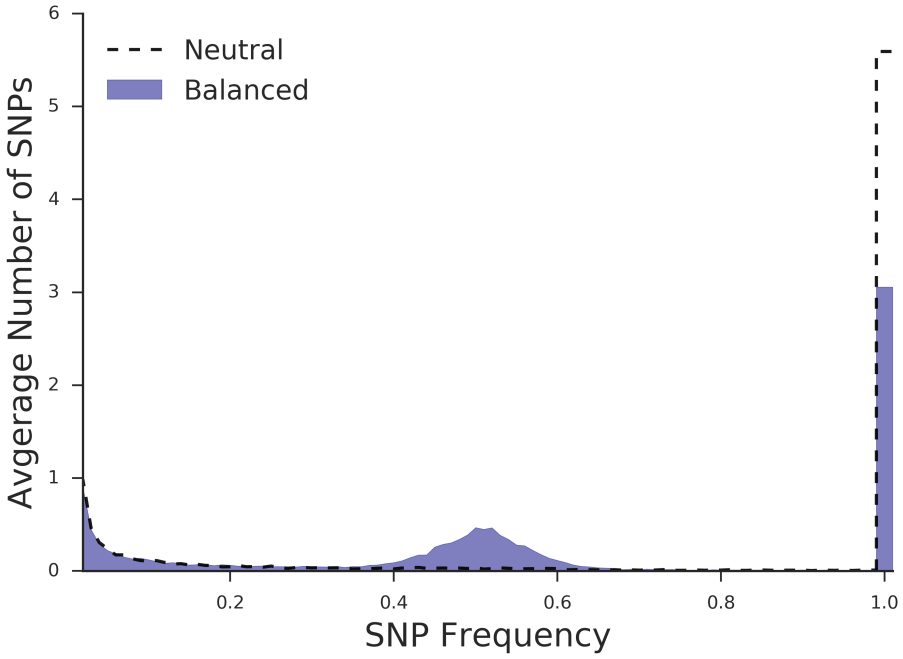


Figure 1: Site frequency spectrum of derived alleles in balanced or neutral cases, with core variant removed. Window size is 500 basepairs on either side of the core site, with sample size 100 chromosomes.

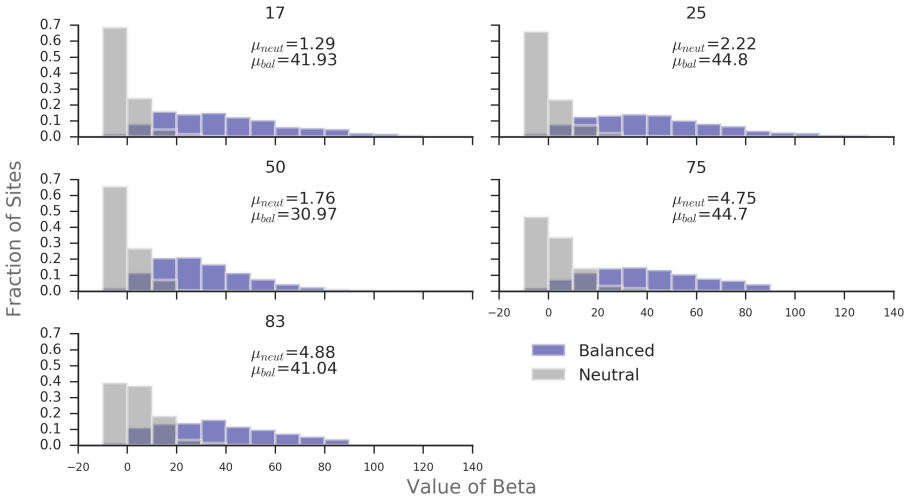


Figure 2: Distribution of Beta in 1kb windows around a core SNP at different equilibrium frequencies.

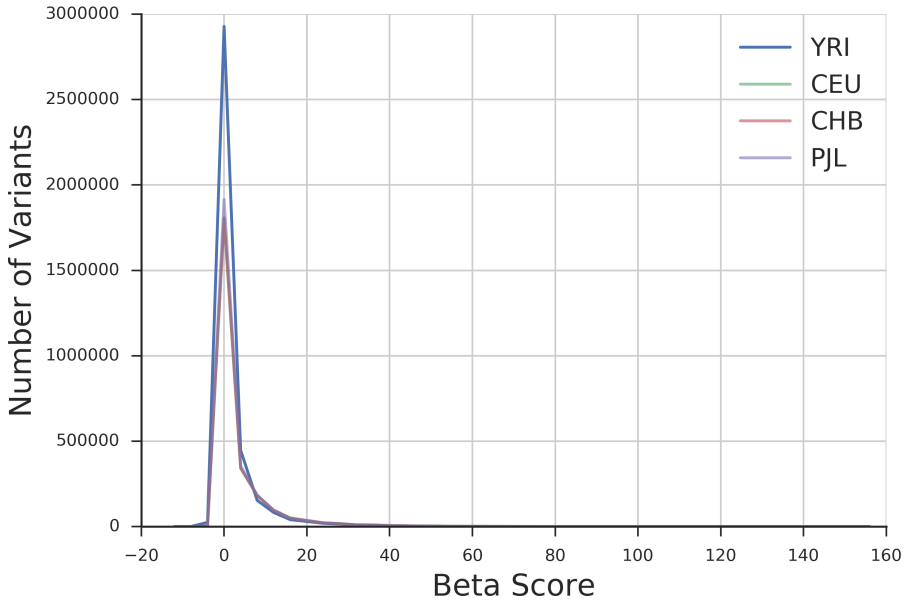


Figure 3: Distribution of Beta in 4 populations. Beta scores binned in units of 4.

## Derivation of Unfolded $\theta_B$

Let $n$ be the number of chromosomes samples, $d_i$ be the similarity measure (see main text) and $S_i$ be the number of variants at frequency $i$ in the sample. For ease of calculation, assume no covariance between sites:

$$\begin{aligned} E[\sum_{i=1}^{n-1} id_i S_i] &= \sum_{i=1}^{n-1} E[id_i S_i] \\ &= \sum_{i=1}^{n-1} id_i E[S_i] \\ &= \sum_{i=1}^{n-1} id_i \frac{1}{i}\theta \\ \hat{\theta}_\beta &= \frac{\sum_{i=1}^{n-1} id_i S_i}{\sum_{i=1}^{n-1} d_i} \end{aligned}$$

## Derivation of Folded $\theta_B$

$$\begin{aligned} E[\sum_{i=1}^{n-1} d_i S_i] &= \sum_{i=1}^{n-1} E[d_i S_i] \\ &= \sum_{i=1}^{n-1} d_i E[S_i] \\ &= \sum_{i=1}^{n-1} d_i \frac{1}{i}\theta \\ \hat{\theta}_\beta &= \frac{\sum_{i=1}^{n-1} d_i S_i}{\sum_{i=1}^{n-1} d_i \frac{1}{i}} \end{aligned}$$

# Power Analysis

Table 1: Power of methods to detect ancient balancing selection for various demographies and selection equilibrium frequencies.

| | Equilibrium | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Older Selection[1] | | | | | Newer Selection[2] | | | | |
| Method | 0.17 | 0.25 | 0.5 | 0.75 | 0.83 | .17 | 0.25 | 0.5 | 0.75 | .83 |
| Beta unfolded | 0.65 | 0.64 | 0.67 | 0.58 | 0.5 | 0.28 | 0.29 | 0.36 | 0.23 | 0.18 |
| Beta folded | 0.59 | 0.65 | 0.67 | 0.52 | 0.35 | 0.29 | 0.32 | 0.36 | 0.19 | 0.11 |
| HKA | 0.43 | 0.42 | 0.29 | 0.36 | 0.46 | 0.14 | 0.13 | 0.08 | 0.08 | 0.15 |
| Tajima's D | 0.10 | 0.08 | 0.35 | 0.12 | 0.02 | 0.01 | 0.03 | 0.14 | 0.04 | 0.02 |
| T1 | 0.69 | 0.52 | 0.44 | 0.58 | 0.67 | 0.30 | 0.17 | 0.11 | 0.15 | 0.28 |
| T2 | 0.77 | 0.75 | 0.79 | 0.68 | 0.67 | 0.36 | 0.31 | 0.46 | 0.24 | 0.28 |

Table 2: Power of methods to detect ancient balancing selection under model of population bottleneck.

| | Older Selection | | | Newer Selection | | |
|---|---|---|---|---|---|---|
| Method | 0.25 | 0.5 | 0.75 | 0.25 | 0.5 | 0.75 |
| Beta unfolded | 0.69 | 0.49 | 0.41 | 0.40 | 0.18 | 0.15 |
| Beta folded | 0.7 | 0.48 | 0.43 | 0.42 | 0.17 | 0.16 |
| HKA | 0.42 | 0.29 | 0.33 | 0.12 | 0.10 | 0.08 |
| Tajima's D | 0.17 | 0.48 | 0.19 | 0.07 | 0.23 | 0.11 |
| T1 | 0.63 | 0.44 | 0.47 | 0.22 | 0.17 | 0.12 |
| T2 | 0.74 | 0.69 | 0.54 | 0.32 | 0.35 | 0.18 |

Table 3: Power of methods to detect ancient balancing selection under model of population expansion

| | Older Selection | | | Newer Selection | | |
|---|---|---|---|---|---|---|
| Method | 0.25 | 0.5 | 0.75 | 0.25 | 0.5 | 0.75 |
| Beta unfolded | 0.50 | 0.56 | 0.42 | 0.19 | 0.51 | 0.14 |
| Beta folded | 0.52 | 0.54 | 0.43 | 0.19 | 0.49 | 0.15 |
| HKA | 0.28 | 0.16 | 0.20 | 0.05 | 0.13 | 0.05 |
| Tajima's D | 0.09 | 0.26 | 0.15 | 0.04 | 0.25 | 0.06 |
| T1 | 0.05 | 0.04 | 0.07 | 0.01 | 0.05 | 0.01 |
| T2 | 0.58 | 0.57 | 0.52 | 0.21 | 0.54 | 0.17 |

Table 4: Power of methods to detect selection based on value of parameter p.

| | Older Selection | | | Newer Selection | | |
|---|---|---|---|---|---|---|
| p | .25 | .5 | .75 | .25 | .5 | .75 |
| 1 | .55 | .68 | .46 | .27 | .32 | .18 |
| 10 | .66 | .68 | .53 | .34 | .37 | .19 |
| 20 | .64 | .66 | .51 | .31 | .35 | .19 |
| 50 | .63 | ..64 | .50 | .31 | .33 | .17 |
| 100 | .63 | .61 | .47 | .31 | .31 | .16 |

Table 5: Power of methods to detect ancient balancing selection for various window sizes

| | Older Selection | | | Newer Selection | | |
|---|---|---|---|---|---|---|
| Window Size (bp) | .25 | .5 | .75 | .25 | .5 | .75 |
| 200 | .55 | .55 | .34 | .22 | .22 | .1 |
| 500 | .67 | .63 | .48 | .30 | .28 | .15 |
| 1000 | .64 | .66 | .51 | .31 | .35 | .19 |
| 2000 | .57 | .54 | .43 | .30 | .26 | .18 |
| 5000 | .47 | .35 | .32 | .25 | .15 | .13 |

Table 6: Power of methods to detect ancient balancing selection for elevated recombination rate $2.5e-7$, Rescaled Simulations

| | Older Selection | | | Newer Selection | | |
|---|---|---|---|---|---|---|
| Method | .25 | .5 | .75 | .25 | .5 | .75 |
| Beta Unfolded | .13 | .11 | .14 | .07 | .06 | .07 |
| Beta Folded | .17 | ..11 | ..11 | ..08 | .07 | .06 |
| HKA | .06 | .00 | .05 | .05 | .004 | .03 |
| Tajima's D | .03 | .06 | .02 | .03 | .05 | .02 |
| T1 | .14 | .01 | .09 | .07 | .03 | .03 |
| T2 | .22 | .14 | .15 | .11 | .05 | .1305 |

Table 7: Power of methods to detect ancient balancing selection for elevated mutation rate 2.5e-7, Rescaled Simulations

| | Older Selection | | | Newer Selection | | |
|---|---|---|---|---|---|---|
| Method | .25 | .5 | .75 | .25 | .5 | .75 |
| Beta | .81 | .73 | .79 | .45 | .39 | .43 |
| HKA | .78 | .19 | .73 | .23 | .01 | .17 |
| Tajima's D | .22 | .83 | .31 | .05 | .61 | .1 |
| T1 | .89 | .66 | .88 | .19 | .06 | .2 |
| T2 | .95 | .94 | .95 | .31 | .47 | .31 |

Table 8: Power of methods to detect ancient balancing selection for lowered recombination rate 2.5e-7, Rescaled Simulations

| | Older Selection | | | Newer Selection | | |
|---|---|---|---|---|---|---|
| Method | .25 | .5 | .75 | .25 | .5 | .75 |
| Beta | .93 | .97 | .90 | .35 | .63 | .35 |
| HKA | .81 | .62 | .83 | .19 | .09 | .15 |
| Tajima's D | .20 | .89 | .35 | .05 | .46 | .09 |
| T1 | .94 | .77 | .90 | .29 | .12 | .20 |
| T2 | .98 | .94 | .96 | .43 | .44 | .39 |

Table 9: Power of methods to detect ancient balancing selection for lowered mutation rate 2.5e-7, Rescaled Simulations

| | Older Selection | | | Newer Selection | | |
|---|---|---|---|---|---|---|
| Method | .25 | .5 | .75 | .25 | .5 | .75 |
| Beta Unfolded | .17 | .16 | .12 | .07 | .08 | .06 |
| Beta Folded | .2 | .14 | .13 | .07 | .07 | .07 |
| HKA | .10 | .03 | .08 | .04 | .01 | .03 |
| Tajima's D | .04 | .11 | .06 | .02 | .05 | .03 |

T1 or T2 could not be calculated with -w 20 due to there not being 20 mutations in the simulated windows.

Table 10: Power of Beta to detect ancient balancing selection with s=1e-4 and $h = 100$ (equilibrium frequency of 0.5)

| Method | Older Selection | Newer Selection |
|---|---|---|
| Beta Unfolded | .74 | .42 |
| Beta Folded | .74 | .41 |
| HKA | .41 | .12 |
| Tajima's D | .37 | .13 |
| T1 | .54 | .16 |
| T2 | .84 | .48 |

Table 11: Power of Beta to detect Frequency Dependent Selection with equilibrium frequency 0.5 and $h = .01$, Rescaled Simulations

| Selection start time | Power |
|---|---|
| 100,000 | .71 |
| 250,000 | .40 |

## Power with different sample sizes

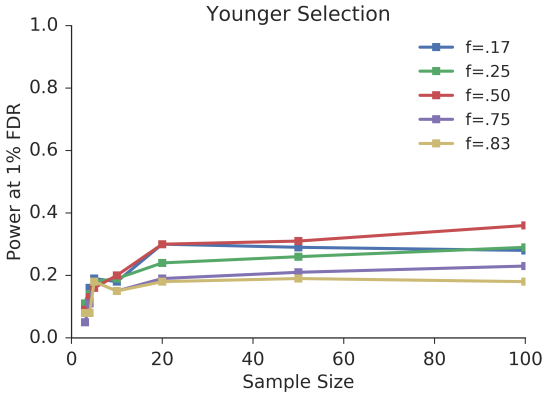


Figure 4: Power of $\beta$ at a 1 percent false discovery rate to detect selection 100,000 generations old.

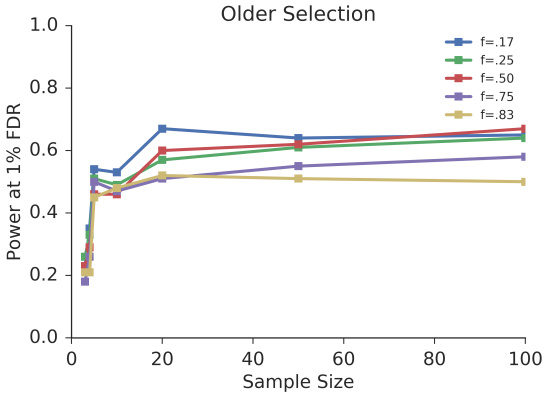


Figure 5: Power of $\beta$ at a 1 percent false discovery rate to detect selection 250,000 generations old.

# References


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

[2] 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.

[3] Hoggart CJ, et al. (2007) Sequence-Level Population Simulations Over Large Genomic Regions. *Genetics* 177(3).

[4] 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.