CLEAR: Composition of Likelihoods for Evolve And Resequence Experiments

Arya Iranmehr[1*], Ali Akbari[1], Christian Schlötterer[2], Vineet Bafna[3*]

1 Electrical and Computer Engineering, University of California, San Diego, La Jolla, CA, USA
2 Institut für Populationsgenetik, Vetmeduni, Vienna, Austria
3 Computer Science and Engineering, University of California, San Diego, La Jolla, CA, USA

* airanmehr@gmail.com (AI); vbafna@ucsd.edu (VB)

S1 Text Dynamics of Site Frequency Spectrum-based Statistics and Linkage Disequilibrium under Selection

S1.1 Text An approximate logistic function for allele frequency dynamics

Assume that a site is evolving under selection constraints $s, h \in \mathbb{R}$, where s and h denote selection strength and overdominance, respectively. Let ν_t denote the frequency of the site at time $\tau_t \in \mathcal{T}$. Then, ν_{t^+}, the frequency at time $\tau_t + 1$ can be estimated using:

$$\hat{\nu}_{t^+} = \nu_t + \frac{s(h + (1-2h)\nu_t)\nu_t(1-\nu_t)}{1 + s\nu_t(2h + (1-2h)\nu_t))}. \tag{S1}$$

We can show that the dynamic of the favored allele can be modeled via a logistic function, in the case of directional selection ($h = 0.5$). Taking derivatives of Eq. S1, we have

$$\frac{\mathbf{d}\nu_t}{\mathbf{d}t} = \frac{s\nu_t(1-\nu_t)}{2 + 2s\nu_t} \tag{S2}$$

To, solve the differential equation, note that for small s, $2 + 2s\nu_t \approx 2$. Substituting,

$$\nu_t = \frac{1}{1 + \frac{1-\nu_0}{\nu_0} e^{-st/2}} = \sigma(st/2 + \eta(\nu_0)) \tag{S3}$$

where $\sigma(.)$ is the logistic function and $\eta(.)$ is logit function (inverse of the logistic function).

S1.2 Text Dynamic of Tajima's D

In this part we derive dynamic of Tajima's D statistic in *hard sweep* as function of its value at the onset of selection, D_0, selection strength and the frequency of the favored allele at the onset of selection. Let D_0, Π_0, W_0, be Tajima's D, Tajima's estimate of θ, and Watterson's estimate of θ at time zero and $D_0 = \Pi_0 - W_0$. In order to compute, $D_t = \Pi_t - W_t$ we compute Π_t and W_t separately as follows. Let P be the $n \times n$ matrix of pairwise heterozygosity of individuals, then $\Pi = 1/n^2 \sum P_{ij}$. So, if the population

consist of νn identical carrier haplotype (due to lack of recombination), their pairwise hamming distance is zero and should be subtracted from the total Π_t:

$$\Pi_t = (1-\nu_t^2)\Pi_0 \tag{S4}$$

To compute W_t, first remember that $W_t = \frac{m_t}{S_n}$ where m_t is the number of segregating sites at time t and $S_n = \sum_i^n 1/i \approx \log(n)$. Also we have

$$\frac{W_t}{W_0} = \frac{\frac{m_t}{S}}{\frac{m_0}{S}} \quad \Rightarrow W_t = \frac{m_t}{m_0} W0 \tag{S5}$$

Because of hard sweep and lack of recombination assumption, the population at time t consist of $(1-\nu_t)n$ non-carrier haplotypes and $\nu_t n$ identical carrier haplotypes. While not strictly correct, we assume that the $(1-\nu_t)n+1$ individuals are evolving neutrally. Using this assumption, we have

$$\frac{m_t}{m_0} = \frac{\log\left((1-\nu_t)n+1\right)\theta}{\log(n)\theta} \approx \frac{\log\left((1-\nu_t)n\right)}{\log(n)} = \frac{\log(1-\nu_t)+\log(n)}{\log(n)} = 1 + \frac{\log(1-\nu_t)}{\log(n)}. \tag{S6}$$

Finally, by putting Eqs. S4, S5, S6 together, we can explicitly write the dynmics of D statistic as

$$\begin{aligned} D_t &= (1-\nu_t^2)\Pi_0 - (1+\frac{\log(1-\nu_t)}{\log(n)})W_0 \\ &= D_0 - \log(1-\nu_t)\frac{W_0}{\log(n)} - \nu_t^2\Pi_0 \\ &\approx D_0 - \log(1-\sigma(st/2+\eta(\nu_0)))\frac{W_0}{\log(n)} - \sigma(st/2+\eta(\nu_0))^2\Pi_0. \end{aligned} \tag{S7}$$

where σ and η are logistic and logit functions.

S1.3 Text Dynamics of Fay and Wu's H

In any finite population size of n with m segregating sites, allele frequencies take discrete values, i.e., $x_j \in \{\frac{1}{n}, \frac{2}{n}, \ldots, \frac{n-1}{n}\},\ \forall j \in 1, \ldots, m$. We have the following:

$$\|\mathbf{x}\|^2 = \sum_{j=1}^{m} x_j^2 = \sum_{i=1}^{n-1} \left(\frac{i}{n}\right)^2 \xi_i = \frac{(n-1)}{2n} H,$$

where ξ_i is the number of sites with frequency i/n and H is the Fay & Wu's estimate of θ and $\mathbf{x} \in (0,1)^m$ is the vector of allele frequency of a region with m segregating sites. Recently, Ronen *et al.* [1] devised the 1-HAF statistic for identifying selection on static data, and showed that the expected value of 1-HAF statistic is given by:

$$\mathbb{E}[\text{1-HAF}(t)] = n\|\mathbf{x}_t\|^2 \approx n g(\nu_t) \tag{S8}$$

where

$$g(\nu_t) = \theta\nu_t\left(\frac{\nu_t+1}{2} - \frac{1}{(1-\nu_t)n+1}\right) + \theta(1-\nu_t)\left(\frac{n+1}{2n} - \frac{1}{(1-\nu_t)n+1}\right) \tag{S9}$$

The dynamics of Fay & Wu's estimate are given by

$$H_t = \frac{n-1}{2} g(\nu_t) \tag{S10}$$

S1.4 Text Greedy computation of time-series SFS-based statistics

As discussed in Section 2.2, modeling dynamic of Tajima's D (and Fay&Wu's H) requires knowledge of initial carrier frequency ν_0 and the value of D (and H) statistic at the onset of selection, which are often unknown. As these statistics are monotonically decreasing (or increasing for SFSelect) under no demographic changes, we chose to greedily aggregate statistics throughout time. For example, for Tajima's D, we have

$$\mathcal{D} = \sum_{t \in \mathcal{T}} D_t \tag{S11}$$

where the same procedure applies to Fay&Wu's H and SFSelect.

S1.5 Text Linkage Disequilibrium

Nonrandom associations, Linkage Disequilibrium (LD), between polymorphisms are established in the substitution process, broken by recombination events and reinforced by selection. Although LD can not be measured in pooled sequencing data (phased haplotype data is required), it is still worthwhile to examine the behavior of LD as a result of the interaction between recombination and natural selection. In this part we theoretically overview expected LD in short EEs.

Let ρ_0 be the LD at time zero between the favored allele and a segregating site l base-pairs away, then under natural selection we have

$$\rho_t = \alpha_t \beta_t \rho_0 = e^{-rtl} \left(\frac{K_t}{K_0} \right) \rho_0 \tag{S12}$$

where $K_t = 2\nu_t(1-\nu_t)$ is the heterozygousity at the selected site, r is the recombination rate/bp/gen. The *decay factor*, $\alpha_t = e^{-rtl}$, and *growth factor*, β_t (see Eqs. 30-31 in [2]), are result of recombination and selection, respectively. S14 Fig presents the expected theoretical value of LD when $\rho_0 = 0.5$ between favored allele (site at position 500K) and the rest of genome, and $\nu_0 = 0.1$. For neutral evolution (top), LD decays exponentially through space and time, while in natural selection (bottom), LD increases and then decreases. Interestingly, LD increases to its maximum value, 1, for the nearby region (the plateau in S14 Fig bottom) of the favored allele.

In principle, LD increases after the onset of selection, until $\log(\alpha_t) + \log(\beta_t) > 0$, see Eq. S12. Specifically, log of decay term is linear and, using Eq. S3, we write growth factor in term of initial frequency ν_0 and selection strength s. S15 Fig, S16 Fig, S17 Fig, and S18 Fig depict interaction of decay and growth factors for weak and strong selection and soft and hard sweeps. In all the case, LD of the favored allele with a segregating site 50Kbp away, increases in the first 50 generations, which give rise to increasing number of *hitchhikers*.

Increase of LD in a large (100Kbp) region is particularly advantageous to the task of identifying the region under selection, if the composite statistics is used. As a result, $\mathcal{H}$ statistic outperforms existing (single-loci) tools in identifying selection. In contrast, augmentation of LD, increases the number of candidates for the favored allele, which makes it difficult to localize the favored allele.

References

1. Ronen R, Tesler G, Akbari A, Zakov S, Rosenberg NA, Bafna V. Predicting Carriers of Ongoing Selective Sweeps Without Knowledge of the Favored Allele. PLoS Genet. 2015;11(9):e1005527.

2. Stephan W, Song YS, Langley CH. The hitchhiking effect on linkage disequilibrium between linked neutral loci. Genetics. 2006;172(4):2647–2663.