Selection regimes through hierarchical MK approaches

Jesus Murga Moreno

October 23, 2019

1 Estimation of sligthly deleterious mutations

Here we present b estimator, which calculate the fraction/excess of slightly deleterious mutations at selected sites. This estimation bring us a direct evidence of the strength of purifying selection at selected sites, altogether with the estimation of effectively neutral mutations and strongly deleterious mutations (defined as deleterious mutations not segregating at samples because of fitness). In addition b represents the variable biasing P_i/P_0 ratio. Then b is define as the excess of slightly deleterious mutations normalized by the total number of sites. I

$$b = \sum_{j=0}^{1} \frac{P_{wd(j)}}{P_{0(j)}} \cdot \frac{m_0}{m_i} \tag{1}$$

Assuming beneficial alleles get fixed quickly and barely contribute to polymorphism, the site frequency spectrum (SFS) of selected sites could be defined by one-side gamma distribution as described in REFS, where mutations will be effectively neutral or deleterious. EXPLANATION WHY GAMMA. The distribution is defined by two parameters: k and a, governing shape and mean of the distribution.The kurtosis will determinate the number of slightly deleterious mutations segregating on selected sites.

1.1 b through αMK

α will be underestimated depending on the presence and proportion of slightly deleterious. In the absence of slightly deleterious mutations in that frequency, α should be approximately the asymptotic value in a given frequency:

$$\alpha_{(j)} \approx \alpha_a \tag{2}$$

Under this assumption we could redefine the $P_{i(j)}/P_{0(j)}$ ratio, where the slightly deleterious mutations ($P_{wd(j)}$) tend to be 0 in that frequency.

$$\begin{gathered} P_{wd(j)} \rightarrow 0 \\ \frac{P_{i(j)} - P_{wd(j)}}{P_{0(j)}} \end{gathered} \tag{3}$$

At frequencies where slightly deleterious mutations are segregating, we could estimate $P_{wd(j)}$ from equation (2), assuming again $\alpha_{(j)}$ should be α_a. Then we redefined the $\alpha_{(j)}$ with equation (3) to include the expected count of slightly deleterious mutations biasing $\alpha_{(j)}$.

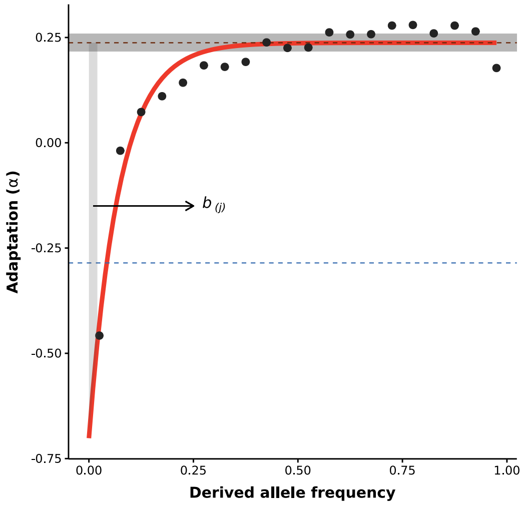
$$\alpha_a = 1 - \frac{(P_{i(j)} - P_{wd(j)}) \cdot D_0}{P_{0(j)} \cdot D_i} \tag{4}$$

We understand $P_{wd(j)}$ as the expected number of slightly deleterious mutations biasing α at a given frequency, distorting the frequency spectra of selected sites and therefore underestimating $\alpha_{(j)}$.

$$P_{wd(j)} = P_{i(j)} - \frac{(1 - \alpha_a) \cdot P_0 \cdot D_i}{D_0} \tag{5}$$

We just expect slightly deleterious at low frequency, although it depends on several important factors as the distribution of fitness effect (DFE), the effective population size (N_e) or recent demography events. Nevertheless, we estimated P_{wd} at the range $\alpha_{(0)} \rightarrow \alpha_{(1)}$, following the α_a estimation at the αMK methodology. At the moment $\alpha_{(j)}$ is within the confidence intervals estimated at αMK, $\alpha_{(j)}$ will range around α_a values. At these frequencies we don't expect slightly deleterious mutations and depending on $\alpha_{(j)}$ values we could estimate positive or negatives values of $P_{wd(j)}$ if $\alpha_{(j)}$ is above or bellow the asymptote. However, once $\alpha_{(j)}$ is at the confidence intervals, $P_{wd(j)}$ values are minimal and tend to be 0. In these cases the $\alpha_{(j)}$ fluctuation is canceling the total count at $P_{wd(j|\alpha_{(CIlow)}} \rightarrow P_{wd(1)}$. In these way we avoid several assumptions of the DFE at selected sites, not considering where slightly deleterious mutations should segregate *a priori*, maximizing the estimation.

$b_{(j)}$ definition could be easily visualize as the area under the asymptotic curve. At frequencies where P_i have an excess of slightly deleterious mutation, the difference between $\alpha_{(j)}$ and α_a should be greater than $\alpha_{(j+1)}$,



assuming slightly deleterious mutation tend to segregate low along the site frequency spectrum.

1.2 α and b estimation through FWW imputation method

Fay et al. proposed to remove low polymorphism segregating, because slightly deleterious mutations tend to segregate at low frequencies. Charlesworth, J et al. investigate how the removal of low-frequency polymorphism affects the estimation of α depending on the continuous function defining the DFE, in addition to a non-arbitrary cutoff to remove low-frequency polymorphism. As they described, only if adaptive mutations are frequent and DFE is leptokurtic the method is quite efficient to compare α_{true} with α_{FWW}, independently of the cutoff even the threshold were 15%. In addition to these problem, remove low-frequency polymorphism imply major performance issues, since we expect to find at these frequencies the highest proportion of segregating sites. Even though the sample size were enough, we would found a decrease in particualr analyzable cases, which could lead to underestimate adaptation too.

Assuming the limitation of this hierarchical MK approach, regarding DFE, we modified the methodology in order to improve the performance. Since one the main problem of the methodology is the amount of data lost removing low-frequency polymorphism, we try to impute slightly deleterious mutations from SFS to exclude them. Following Charlesworth, J et al, Eyre-Walker et al. (REFS), we considered slightly deleterious mutations should be segregating below 15% frequencies. Then we divided the SFS in two taking on the cutoff as threshold. Then $P_{i(j>15\%)}/P_{0(j>15\%)}$ (ratio used at original FWW methodology to estimate alpha) should generate an α value near to $\alpha true$. At this situation we are assuming slightly deleterious mutations shouldn't be segregating below the threshold, then $P_{wd(j>15\%}$ tend to be 0. We modified the polymorphism ratio to include this assumption. α_{FWW} is:

$$P_{wd(j>15\%)} \rightarrow 0$$

$$\alpha_{FWW>15\%} = 1 - \frac{(P_{i(j>15\%)} - P_{wd(j>15\%)}) \cdot D_0}{P_{0(j>15\%)} \cdot D_i} \tag{6}$$

$$\alpha_{FWW>15\%} \approx \alpha_{true}$$

If there are segregating slightly deleterious mutations at $P_{i(j<15\%)}$, then the ratio above the cutoff will be greater than the ratio below.

$$\frac{P_{i(j<15\%)}}{P_{0(j<15\%)}} > \frac{P_{i(j>15\%)}}{P_{0(j>15\%)}} \tag{7}$$

Assuming $P_{wd(j>15\%)}$ tend to 0, we could impute the $P_{wd(j<15\%)}$ based on the expected ratio without slightly mutations segregating at the sample, and redefine $P_{i(j<15\%)}$ without slightly deleterious mutations.

$$P_{wd(j<15\%)} \neq 0$$
$$P_{wd(j>15\%)} \rightarrow 0$$

$$\frac{P_{i(j<15\%)} - P_{wd(j<15\%)}}{P_{0(j<15\%)}} = \frac{P_{i(j>15\%)} - P_{wd(j>15\%)}}{P_{0(j>15\%)}} \tag{8}$$

$$P_{wd(j<15\%)} = P_{i(j<15\%)} - \frac{P_{i(j>15\%)} - P_{wd(j>15\%)}}{P_{0(j>15\%)}}$$

$$P_{i(j<15\%)} = P_{i(j<15\%)} - P_{wd(j<15\%)}$$

In this way we are removing the slightly deleterious polymorphism included at the SFS at the same way presented at Fay et al., but not excluding the vast majority of the information of the analysis, which in fact will be always segregating below the threshold, independently of DFE.

We understand P_{wd} as the expected number of slightly deleterious mutations biasing, underestimating α and distorting the frequency spectra of selected sites. Again with P_{wd} we could define b as the excess of slightly deleterious mutations normalized by the total number of sites. I

$$b = \frac{P_{wd}}{P_0} \cdot \frac{m_i}{m_i} \tag{9}$$

1.3 Testing b estimator

To test the sensitivity and accuracy of b we perform multiple simulations through SLiM software. Forward in-time simulations allow us to compare the real excess of slightly deleterious mutations with our estimator. In order to compare both, we estimate the real excess of slightly deleterious mutations (*true b*) dividing the SFS (P_i) through their fitness coefficients (s) on: (1) effectively neutral mutations ($-1 < N_e s < 1$), (2) slightly deleterious mutations ($-10 < N_e s < 1$) and (3) strongly deleterious mutations ($N_e s < -10$). The *true b* is defined like in equation (1), where P_{wd} is the sum of slightly and strongly deleterious mutations subset from the simulated site frequency spectrum. Although we don't expect mutations segregating at the range $N_e s < -10$, it will depends on the DFE and N_e, since DFE goes $-\infty \rightarrow \infty$. We observed some situations were linkage could sweep some of these mutations despite of their low fitness coefficients, although the total number are minimal and tend to segregate at really low frequencies. In this way we compare our estimators with the real excess of slightly deleterious mutations.

We performed a total of 13 simulations. Simulations were performed over a 10Mb with two genomic element (synonymous and non-synonymous sites) simulating a typical coding sequence structure, in a proportion of 1/2 (eg. 004) with a mutation rate of $1e-9$, a recombination rate of $1e-7$ and a dominance coefficient of 0.5 over $2e5$ generations with a $10N_e$ burnin period. We defined synonymous sites as neutral sites introducing only neutral mutations with fixed probability. On the other hand non-synonymous sites were used as selected sites. The spectra of selected sites were defined using a one-side gamma distribution with a k parameter controlling the shape and a parameter controlling the mean (a parameter refers to selection coefficient mean). We based all the scenarios on the same SLiM recipe from which we change determinate parameters to test effects on b. b values were checked changing in several ways the DFE of selected sites (table 1). In this way we increase or decrease the number of deleterious sites segregating at P_i using leptokurtics and platykurtic distributions and changes in the strength of selection. Simulations were divided in two categories: neutral and adaptive, where we introduce the presence of adaptive mutations and test the presence of high α values on b estimation. In addition we replicate the estimations from two different effective population size ($N_e = 100$ and $N_e = 1000$).

1.4 b estimation on presence of recent positive selection

Linkage and recent beneficial alleles segregating on selected sites could lead to underestimate α too. A new methodology combining *abc* frameworks and αMK approaches developed by Uricchio et al. is able to re-estimate α values taking into account these variables. Because of *abc* scripts are currently set up to run on the Stanford cluster, we couldn't apply b estimation at genetics scenarios were adaptation is lead by selective sweeps. At these sites frequency spectrum will be biased due to the proportion of weakly beneficial (determined by α_w in the methodology) and deleterious alleles. Taking into account α_w on the site frequency spectrum, we would expect similar results on accuracy and sensitivity since b only depends on the asymptotic value of α. In any case, at these kind of complex linkage scenarios would require an extensive exploration to determine possibly errors in b estimations.