VIBS613 - Homework 6

Daniel Osorio - dcosorioh@tamu.edu
Department of Veterinary Integrative Biosciences
Texas A&M University

1. Describe 4 AFS-based summary statistics.
   Neutrality tests based on the frequency spectrum are commonly used by population geneticists as routine tests to assess the goodness-of-fit of the standard neutral model on their data sets. Neutrality tests compare two estimators of the population mutation parameter θ that characterizes the mutation–drift equilibrium. It is defined as $\theta = 2pN_e\mu$, where p is the ploidy (1 for haploids and 2 for diploids), N_e is the effective population size, and μ is the locus neutral mutation rate. When the standard model is true, the expectations of the several unbiased estimators of θ are equal. Neutrality tests compute the goodness-of-fit of a statistic T, which is the difference between two estimators of θ, normalized by its standard deviation:

$$T = \frac{\theta_1 - \theta_2}{\sqrt{var(\theta_1 - \theta_2)}}$$

   Typical estimators of θ, in a sample of n sequences, are:

   - $\hat{\theta}_S = \frac{S}{a_n}$, proposed in *Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. Theoretical population biology, 7(2), 256-276.*, where S is the number of polymorphic sites and $a_n = \sum_{i=1}^{n-1} \frac{1}{i}$
   - $\hat{\theta}_\pi = \pi$, proposed in *Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics, 123(3), 585-595*, where π is the average pairwise difference between all sequences in the sample.
   - $\hat{\theta}_\eta = \frac{(n-1)}{n}\eta$, proposed in *Fu, Y. X., & Li, W. H. (1993). Statistical tests of neutrality of mutations. Genetics, 133(3), 693-709.*, where η is the total number of polymorphic sites at both frequencies $\frac{i}{n}$ and $1 - \frac{i}{n}$ (singletons derived and ancestral).
   - $\hat{\theta}_H = \sum_{i=1}^{n-1}(\frac{2i^2}{n(n-1)})\xi$, proposed in *Fay, J. C., & Wu, C. I. (2000). Hitchhiking under positive Darwinian selection. Genetics, 155(3), 1405-1413.*, where ξ is the number of polymorphic sites at frequency $\frac{i}{n}$ in the sample ($i \in [1, n-1]$).

   Using these $\hat{\theta}$ estimators, the following neutrality tests can be computed:

   a) **Tajima's** D, is computed as

$$D = \frac{\hat{\theta}_\pi - \hat{\theta}_S}{\sqrt{var(\hat{\theta}_\pi - \hat{\theta}_S)}}$$

   it allow to study the relationship between the two estimates of genetic variation at the DNA level, namely the number of segregating sites and the average number of nucleotide differences estimated from pairwise comparison. Tajima's statistic is a very general way of comparing the allele frequency spectrum against the expectations of

the null model. It was not designed to pick up any particular deviation from the null model, but it will tend to be negative under selective sweeps (and population growth) and positive under balancing selection (or population structure with sampling from many populations).

b) **Fu and Li's** F^*, is computed as

$$F^* = \frac{\hat{\theta}_\pi - \hat{\theta}_\eta}{\sqrt{var(\hat{\theta}_\pi - \hat{\theta}_\eta)}}$$

it detects selection signatures through the comparison of the number of singleton mutations and the mean pairwise difference between sequences

c) **Fu and Li's** D^*, is computed as

$$D^* = \frac{\hat{\theta}_S - \hat{\theta}_\eta}{\sqrt{var(\hat{\theta}_S - \hat{\theta}_\eta)}}$$

it is based on the number of segregating sites at which the rare allele is only represented once (often called singletons). This statistic shares much information with Tajima's D statistic, a negative value indicates an excess of singletons (which would also give a negative Tajima), and a positive value indicates a lack of singletons (which would typically, though not necessarily, give a positive Tajima's D). However, certain population genetic scenarios, particularly selective sweeps, tend to generate an excess of singletons, to which this test is more sensitive than Tajima's D.

d) **Fay and Wu's** H, is computed as

$$H = \frac{\hat{\theta}_\pi - \hat{\theta}_H}{\sqrt{var(\hat{\theta}_\pi - \hat{\theta}_H)}}$$

and provides a way of identifying samples in which there is an excess or dearth of high-frequency derived mutations, an excess of high frequency derived mutations may be a signal of a recent selective sweep and a nearby locus, though there are certain demographic situations (strong population structure with uneven sampling from populations) that can also give the same pattern. Unlike the previous tests, the variance of the test statistic has to be estimated by stochastic simulation.

Adapted from:

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

Ma, Y., Ding, X., Qanbari, S., Weigend, S., Zhang, Q., & Simianer, H. (2015). Properties of different selection signature statistics and a new strategy for combining them. Heredity, 115(5), 426.