Model Building

Rachael 'Rocky' Aikens, Voight Lab

July 28, 2017

Methodology

In order to develop a formal statistical framework for understanding global polymorphism, we designed a series of multinomial models after Aggarwala and Voight, which capture different levels of mutation rate variation. First, we defined cosmopolitan SNPs to be those which are shared between two or more of the African, European, South Asian, and East Asian 1,000 genomes samples. For a given population at a given sequence context, the probability of recurrent mutation is assumed to be zero. Under these assumptions, we have seven mutually exclusive possible events:

- site is not polymorphic,
- it is a private polymorphism for that population (with three possible alternate alleles), or
- it is a cosmopolitan polymorphism (with three possible alternate alleles).

If the context appears N times in the genome, polymorphism in this population follows a multinomial distribution with size N and parameters c_1, c_2, and c_3 for the probabilities of each cosmopolitan polymorphism, and p_1, p_2, and p_3 for the three private polymorphisms.

If the mutation rate at this context had not changed in recent evolutionary time for this population, then we would expect the probabilities of each private polymorphism type to be proportional to the probabilities of the corresponding cosmopolitan polymorphism types. It remains to estimate this proportionality constant, which we denote α. In a null model (H_0), mutation rate has not changed at any context, so α for a given population is just the ratio of total private polymorphisms to total cosmopolitan polymorphisms over all contexts. Alternatively (H_1), if mutation rate has changed at specific contexts but the relative substitution probabilities for the alternative alleles is fixed, then a unique α must be estimated from the private to cosmopolitan ratio of polymorphisms at each context. Finally, in a model which allows for maximal polymorphism variation (H_2), mutation rate may have changed even between different polymorphism types at same context (e.g. C/T, C/A, and C/G polymorphism at a C context). In this model, the private substitution rates are not proportional to the cosmopolitan rates even at a context-specific level, and the private rates must be estimated independently for each possible mutation.

X chromosome

Because of certain features of demography and sampling, we expect that there will be fewer polymorphisms observed on the X chromosome than on the autosomes, even if the mutation rates are the same. For these reasons, it is desirable to exclude the X chromosome from the model for polymorphism we apply to the autosomes. The code that follows will perform likelihood testing and parameter estimation both without the X chromosome. The same code can be run on the count dataframes including X in the same way.

Likelihood testing

Once we have estimated the necessary parameters for each of these three models from the 1,000 genomes dataset, we can compare the fit of the observed data under each model (H_0, H_1, or H_2) using a log-likelihood ratio test. If Λ represents the ratio of the likelihoods of a null model to an alternative, the test statistic $-2\ln(\Lambda)$ is known to approximately follow a chi-squared distribution with degrees of freedom equal to the difference in the number of parameters in the null versus the alternative. This testing framework allows us to

ask broad questions about what level of mutation rate variability best explains observed polymorphism data at a given sequence context level.

I've written a function **likelihood.test** that essentially does these calculations for you. Like most other functions of mine, it uses count dataframes with a reference dataframe of genome wide counts.

Parameter Estimation

The function **likelihood.test** generates dataframes of the counts of each type of private and cosmopolitan polymorphism, plus likelihood calculations for each of the three hypothesis (H_0, H_1, and H_2). Since H_2, the model allowing for maximal level of mutation rate variation, is the domimant one, we'd like to be able to calculate and report the parameters of these models for future use. For each context type, we can do that simply by dividing the counts of private and cosmopolitan substitutions by the number of total genome wide sites in our inclusion regions. The function **estimate.H2** does that conversion.

These likelihood ratio test results and parameter estimates are saved in the results subdirectory.