Expanded sequence context model reveals extensive variation in human mutation rate

Supplementary Mathematics

Rachael Aikens

May 2017

1 Mutation rate inference from polymorphism

1.1 Estimation

Suppose we would like to calculate the private mutation rate μ_m of a specific mutation type m within a certain population. Let c denote the context from which mutation type m is derived (e.g. if m = 'GAA→A', then c = 'GGA'), and let the polymorphism probability, θ_m, be the probability that a given c type context is a type m private polymorphism in the population.
From population genetics, it is known that mutation rate is proportional to polymorphism probability. Thus, we can assume that there is some constant k, in this population which relates mutation rate to polymorphism as follows:

$$\mu_m = k\theta_m,$$

for any mutation type m.
Let us assume that the total mutation rate per base pair per generation in the population is 1.2×10^{-8}. This means that

$$1.2 \times 10^{-8} = k\Theta,$$

where Θ represents the probability that any given base (from any context) is polymorphic. Rearranging gives

$$k = 1.2 \times 10^{-8}\Theta^{-1},$$

so, for any m,

$$\mu_m = 1.2 \times 10^{-8}\Theta^{-1}\theta_m.$$

From private polymorphism data, we can estimate the probability that a given c context in the genome is polymorphic as

$$\hat{\theta}_m = \frac{n_m}{N_c},$$

where n_m is the number of observed private type m polymorphisms, and N_c is the number of c type contexts that appear in the genome. Moreover, Θ can be estimated as

$$\hat{\Theta} = \frac{\sum\limits_m n_m}{\sum\limits_c N_c}.$$

Combining these two estimates gives our inferred mutation rate, $\hat{\mu}_m = 1.2 \times 10^{-8}\hat{\Theta}^{-1}\hat{\theta}_m$

1.2 Confidence Intervals

Notice that, $n_m \sim \text{Binom}(\theta_m, N_c)$. Because N_c is very large, we can use the following normal-approximation 95% confidence interval for $\hat{\theta}_m$:

$$\hat{\theta}_m \pm 1.96\sqrt{\frac{\hat{\theta}_m(1-\hat{\theta}_m)}{N_c}}.$$

However, we would really like to be able to calculate a 95% CI for μ_m, this is harder because $\hat{\Theta}$ is not known with certainty. In fact, it has approximate sampling distribution

$$Normal(\Theta \sum_c N_c, \frac{\Theta(1-\Theta)}{\sum_c N_c}).$$

Notice that the variance in this sampling distribution is upwardly bounded by $\frac{1}{4\sum_c N_c}$, and that $\sum_c N_c$ is roughly the size of the human genome. With this in mind, we consider Θ to be estimated with negligible error. Under this assumption, an approximate 95% CI for μ_m is given by

$$\left(1.2\times10^{-8}\hat{\Theta}\left(\hat{\theta}_m - 1.96\sqrt{\frac{\hat{\theta}_m(1-\hat{\theta}_m)}{N_c}}\right), 1.2\times10^{-8}\hat{\Theta}\left(\hat{\theta}_m + 1.96\sqrt{\frac{\hat{\theta}_m(1-\hat{\theta}_m)}{N_c}}\right)\right).$$

1.3 Generalization

Assuming a constant mutation rate of 1.2×10^{-8} across chromosomes or subpopulations allows us to generalize these calculations to give chromosome-specific or subpopulation-specific estimates.

1.4 Weaknesses

There are a couple weaknesses to this approach. First, it assumes that all measurements of n_m and N_c have been made without error. Second, in all of the analyses for this paper, n_m has been ascertained by filtering out all singletons and multiallelic variants. Third, the assumption that mutation rate across populations and chromosomes is steady at 1.2×10^{-8} may not be very fair.