BIOB480/BIOE548 notes 10/15/2024

Introduction

- Grading: sometime before I retire. Will keep you posted.

Miller & Waits 2002

See `15_slides.pdf` for questions.

The coalescent

Before our initial discussion of N_e recedes too far into the past, we should touch on how it is estimated. A large number of methods exist, relying on data on the loss of heterozygosity in subsequent generations, the rate of decay in linkage disequilibrium, the loss of allelic diversity, and estimates of drift / mutation equilibrium (for long temporal time scales). Perhaps most useful is Watterson's estimator, which relies on a parameter θ ("theta"), defined as $\theta = 4N_e\mu$.

θ is something known as the population mutation parameter, which gives us the expected number of segregating (=different) alleles in a popultion of a given size and given mutation rate. To understand this, recall that any two alleles in a single generation are identical by descent in the previous generation at probability $\frac{1}{2Ne}$. This observation is the basis of a body of work known as coalescent theory, which is concerned with finding the time to the most recent common ancestor ("MRCA") of a pair of alleles. We can expand the probability of IBD to get at the probability any two alleles *coalesce* in t generations, which is simply the probability they *don't* coalesce in the previous generation multiplied by the probability of coalescence in the generation of interest:

$$P_{coalesce}(t) = (1 - \frac{1}{2N_e})^{t-1}(\frac{1}{2N_e})$$

This form should look familiar, as it can be approximated by the exponential distribution when N_e is sufficiently large:

$$P_{coalesce}(t) = \frac{1}{2N_e}e^{-\frac{t-1}{2N_e}}$$

This is helped because the expected value (think weighted average) of the exponential distribution is $\frac{1}{\lambda}$, where λ is the rate parameter of the distribution. In this case, that's $\frac{1}{2N_e}$, which means that the expected time to coalescence for any two alleles is $1 \div \frac{1}{2N_e} = 2N_e$ generations. For diploid individuals, the expected number of mutations on these haplotypes (or the expected number of mutant alles out of all $2N_e$) is simply this value multiplied by 2μ—the 2 here indicating ploidy, as each copy of the allele has a chance of mutating. Put this all together, and the expected number of segregating sites in any two randomly selected nonrecombining haplotypes is $\theta = 4N_e\mu$. In practice, θ is usually estimated with a metric called π (or θ_π), which is calculated as the average number of nucleotide differences per site in all possible pairs of the sample population. Given this, effective population size can be estimated from $\theta/4\mu$.

As a measure of genetic diversity, θ is closely related to heterozygosity. Using coalescent theory, we can define expected (or mean) heterozygosity as the probability of a mutation in a given generation divided by the total probability of an "event" happening in that generation (either mutation or coalescence):

$$\bar{H} = \frac{2\mu}{2\mu + \frac{1}{2N_e}} = \frac{2\mu}{\frac{4\mu N_e + 1}{2N_e}} = \frac{2\mu}{1} \cdot \frac{2N_e}{4\mu N_e + 1} = \frac{4\mu N_e}{4\mu N_e + 1}$$

Since $\theta = 4\mu N_e$, $\bar{H} = \frac{\theta}{\theta + 1}$!