PBG 200A Notes

Sam Fleischer

November 2, 2016

1 Coalescent process with 3 lineages

Probability three coalesce at time t is $\dfrac{3}{2N}\left(1-\dfrac{3}{2N}\right)^t$. That is,

$$t_3 \sim \text{Geom}\left(\frac{3}{2N}\right) \qquad \text{which implies} \qquad \mathbb{E}[t_3] = \frac{2N}{3}. \tag{1}$$

Then, $t_2 \sim \text{Geom}\left(\dfrac{1}{2N}\right)$]. Then

$$T_{\text{MRCA}} = T_3 + T_2 \tag{2}$$
$$\implies \mathbb{E}[T_{\text{MRCA}}] = \mathbb{E}[T_3] + \mathbb{E}[T_2] \tag{3}$$

$$T_{\text{Tot}} = 3T_3 + 2T_2\mathbb{E}[T_{\text{Tot}}] = 3\mathbb{E}[T_3] + 2\mathbb{E}[T_2] \tag{4}$$

2 Process with k lineages

Probability all k coalesce at time t is $\dfrac{\binom{k}{2}}{2N}\left(1-\dfrac{\binom{k}{2}}{2N}\right)^t$. So the coalescent time t_k is a Geometric distribution

$$t_k \sim \text{Geom}\left(\frac{\binom{k}{2}}{2N}\right) \qquad \text{which implies} \qquad \mathbb{E}[t_k] = \frac{2N}{\binom{k}{2}}. \tag{5}$$

When there are n individuals,

$$\mathbb{E}[T_{\text{MRCA}}] = \sum_{k=n}^{n} \mathbb{E}[T_k] = \sum_{k=n}^{2} \frac{2N}{\binom{k}{2}} = \underbrace{\sum_{k=n}^{2} \frac{4N}{k(k-1)}}_{\text{telescoping sum}} = 4N\left(1-\frac{1}{n}\right) \tag{6}$$

Also,

$$\mathbb{E}[T_{\text{Tot}}] = \sum_{k=n}^{2} k\mathbb{E}[T_k] = \sum_{k=n} k\mathbb{E}[T_k] = \sum_{k=n}^{2} k\frac{2N}{\binom{k}{2}} = \sum_{k=n}^{2} \frac{4N}{k-1} \tag{7}$$

Let S be the total number of segregating sites. Then

$$\mathbb{E}[S] = \mu\mathbb{E}[T_{\text{Tot}}] = \underbrace{4N\mu}_{\theta} \underbrace{\sum_{k=n}^{2} \frac{1}{k=1}]}_{w} \tag{8}$$

The total number of mutations we see is the total amount of time in the genealogy.

3 Frequency of mutations

For a constant population size, the expected count of sites at mutation frequency i is $\frac{\theta}{i}$. For an expanded population, this is more exaggerated (more like $\frac{\theta}{i^2}$) since the total time at smaller alleles is less. In a bottlenecked population, however, this is less exagerated (more like $\frac{\theta}{\sqrt{i}}$).