# Simulating GBS data for pooled samples

Lucas Nell

*Spring 2017*

Nei and Li's (1979) measure of nucleotide diversity, $\theta_\pi$, is calculated using the following equation:

$$\theta_\pi = \sum_{ij} x_i x_j \pi_{ij} \tag{1}$$

Here, $x_i$ and $x_j$ represent the frequencies of the $i$th and $j$th unique sequences respectively and $\pi_{ij}$ represents the proportion of divergent sequence between the $i$th and $j$th unique sequences.

If I assume that all lines will be unique sequencesa safe assumption if whole genomes are consideredthen the above equation can be expressed as follows:

$$\theta_\pi = \frac{1}{n^2} \sum_{ij} \pi_{ij} \tag{2}$$

Then, since the number of total pairwise combinations between $n$ sequences is $\binom{n}{2}$, we can calculate $\bar{\pi}$, the mean proportional sequence divergence between any two sequences, as such:

$$\bar{\pi} = \frac{\sum_{ij} \pi_{ij}}{\binom{n}{2}} \tag{3}$$

Some simple arithmetic gives us...

$$\sum_{ij} \pi_{ij} = \binom{n}{2} \bar{\pi} \tag{4}$$

Now I insert this into equation 2:

$$\theta_\pi = \frac{1}{n^2} \binom{n}{2} \bar{\pi} \tag{5}$$

Solving for $\bar{\pi}$ yields the following:

$$\bar{\pi} = \frac{\theta_\pi n^2}{\binom{n}{2}} \tag{6}$$