Statistical Methods

immediate

July 12, 2022

Likelihood

The likelihood in Eyalshiv et al. (2016) has the form,

$$\log \mathcal{L}(\theta) = \sum_{v \in \mathcal{V}} \sum_{i \neq j \in \mathcal{S}} \log(P(O_{i,j}(v)|\theta)) \tag{1}$$

where $\mathcal{V}$ is the set of putatively neutral sites, $\mathcal{S}$ is the set of samples, and θ are the BGS parameters. The indicator variable $O_{i,j}(v)$ is 1 if samples i and j are different at site v, and zero otherwise. Thus as they specify in the paper,

$$P(O_{i,j}(v)|\theta) = \begin{cases} \pi(v|\theta), & O_{i,j}(v) = 1 \\ 1 - \pi(v|\theta), & O_{i,j}(v) = 0 \end{cases}. \tag{2}$$

The total size of the set of samples $\mathcal{S}$ is $n_{\mathcal{S}} = |\mathcal{S}|$. Assuming all sites are biallelic, we can simplify the inner summation by counting the number of possible same and different pairwise combinations. If a site v's vector of allele counts is $[c_1, c_2]$, the total number of pairwise combinations with the same alleles is

$$n_s(v) = \binom{c_1}{2} + \binom{c_2}{2} \tag{3}$$

and the number of different pairwise combinations is

$$n_d(v) = n_T - n_s(v) \tag{4}$$

where $n_T = n_{\mathcal{S}}(n_{\mathcal{S}} - 1)/2$ is the total number of pairwise combinations across the sample set $\mathcal{S}$. Note that these site-specific counts allow us to use allelic counts directly, and can vary across sites. Our likelihood is then,

$$\log \mathcal{L}(\theta) = \sum_{v \in \mathcal{V}} \left[\log(\pi(v|\theta)) n_{\mathrm{D}}(v) + \log(1 - \pi(v|\theta)) n_{\mathrm{S}}(v) \right]. \tag{5}$$

In practice, we calculate these values across bins. For each bin, we treat $\pi(v|\theta)$ as fixed, assuming that at this scale, the variation in expected diversity across sites is minimal. For a particular chromosome, we have two classes of sites: those included in the diversity calculation and those ignored. The former sites are all putatively neutral and have reliably called genotypes, and the other sites are possibly non-neutral or do have reliably called genotypes. The total log-likelihood is the sum of bin likelihoods, $\mathcal{L}(b)$

$$\mathcal{L}(\theta) = \sum_b \mathcal{L}(b|\theta) \tag{6}$$

The likelihood within a bin is then,

$$\log \mathcal{L}(b|\theta) = \log(\bar{\pi}(b|\theta)) \sum_{v \in \mathcal{V}_b} n_D(v) + \log(1 - \bar{\pi}(b|\theta)) \sum_{v \in \mathcal{V}_b} n_S(v) \tag{7}$$

$$= \log(\bar{\pi}(b|\theta)) Y_D(b) + \log(1 - \bar{\pi}(b|\theta)) Y_S(b) \tag{8}$$

where the two sum terms as $Y_D(b)$ and $Y_S(b)$ are data reductions at the bin level.

If the data are such that only polymorphic sites are considered, we can adapt this by partitioning the set $\mathcal{V}$ of neutral sites into polymorphic ($\mathcal{P}$) and fixed sites ($\mathcal{F}$), i.e. $\mathcal{V} = \mathcal{P} \cup \mathcal{F}$ and $\mathcal{P} \cap \mathcal{F} = \varnothing$. For all $v \in \mathcal{F}$, $n_d(v) = 0$ and $n_s(v) = n_T$.

Then,

$$\log \mathcal{L}(\theta) = \sum_{v \in \mathcal{V}} [\log(\pi(v|\theta)) n_{\mathrm{D}}(v) + \log(1 - \pi(v|\theta)) n_{\mathrm{S}}(v)] \tag{9}$$

$$= \sum_{v \in \mathcal{P}} [\log(\pi(v|\theta)) n_{\mathrm{D}}(v) + \log(1 - \pi(v|\theta)) n_{\mathrm{S}}(v)] + \sum_{v \in \mathcal{F}} \log(1 - \pi(v|\theta)) n_T(v) \tag{10}$$

$$\tag{11}$$

$$\log \mathcal{L}(b|\theta) = \log(\bar{\pi}(b|\theta)) \sum_{v \in \mathcal{P}_b} n_{\mathrm{D}}(v) + \log(1 - \bar{\pi}(b|\theta)) \left(\sum_{v \in \mathcal{P}_b} n_{\mathrm{S}}(v) + \sum_{v \in \mathcal{F}_b} n_T(v) \right). \tag{12}$$

Note that if we assume that the total number of combinations at each fixed site is constant, e.g. $n_T = n_T(v)$ for all v, then we can use $\sum_v n_T(v) = n_T |\mathcal{F}_b|$.

Windowed Diversity

Although we use the components of diversity, n_t and n_s, to calculate the likelihood, it is still of interest to calculate diversity from these in a window. The raw allele count data a $L \times 2$ matrix,

1 B Scores