Genotyping equations

The genotype calling can be split into 2 parts:

1. The likelihood of the reads distributed at the position of interest - the likelihood is obtained from preexisting software like samtools
2. The prior for all the genotypes - This is where we implement a population genetics based prior.

Population Genetics based genotype priors

Pop-gen based genotype priors can be modeled hierarchically, in three parts as follows:

1. Probability that the site is a variant site or not -
   $P(var|\theta) := P(\text{site is variable})$
2. Given that the site is a variant site, the frequency of the variant can be obtained using the neutral frequency spectrum expectation - frequency of the non-reference (alternate) allele $:= p_a \sim SFS_{neutral}$
3. Given the frequency of the alternate allele, p_a, we compute the probability of the the genotypes using HWE,i.e.
   $P(G = (a, b)) = 2^{I(a!=b)} p_a p_b$.

Genotype calling

Let $C_i = (C_a, C_c, C_g, C_t)$ be the vector of base counts at current position for individual i, and let $P(C_i|G = g)$ be the likelihood of genotype g computed using samtools. $P(G = g|C_i)$, the posterior probability can be computed using the priors mentioned in the previous section.

$$\begin{aligned} P(G = g|C_i) &\propto P(C_i|G = g)P(G = g) \\ &\propto P(C_i|G = g)P(var|\theta)P(p_v|var)P(G = g|p_v) \end{aligned} \qquad (1)$$

Here, $f(\theta)$ can be calculated using the expected SFS. Given n diploid samples and a population scaled mutation rate of θ, we can compute the total expected number of variant sites to be $E(S) = l\theta \sum_{k=1}^{2n-1} 1/k$. So, $P(var|\theta) = (l - E(S))/l = 1 - \theta \sum_{k=1}^{2n-1} 1/k$, where l is the total length of the region.

Similarly, we can use the neutral SFS to compute the frequency of the variant in the population,
$P(p_v = 1/m|var) = (1/m)/(\sum_{k=1}^{2n-1} 1/k), \forall m \in 1, 2 \ldots 2n-1.$
Algorithm for computing the posterior of the genotypes:

- Select an initial value for θ
- Compute the probability of being variant as $P(var|\theta)$
- Sample an allele frequency for the alternate allele, p_a, for the variant from the neutral SFS
- Assign this allele frequency to the reference or alternate allele randomly, setting the other allele frequency to be $1 - p_a$
- Using the allele frequencies, compute the prior genotype probabilities using HWE.
- Compute the posterior using the genotype priors and the likelihood.

Prior probabilities:

$$P(G=(g_1,g_2)|\theta)=P(G=(g_1,g_2)|v_i=0)P(v_i=0)+P(G=(g_1,g_2)|v_i=1)P(v_i=1) \tag{2}$$

where v_i is an indicator variable for site i, which is 1 if the site is variable and 0 otherwise. If site i is not variable, i.e. $v_i = 0$, the only possible genotype can be the reference homozygote. If the site is a variable site, we need to consider the alternate allele. In this work, we limit our analysis to two alleles at a time, where one allele is always the reference base.
For a variant site, we need the second allele. Since we do not know the other allele, we consider all three possible alternate alleles.