1 Genotype Likelihoods from Reads

$$\begin{aligned}
L(g=1) &= p(d \,|\, g=1) \\
&= p(d, g=ra \,|\, g=1) + p(d, g=rb \,|\, g=1) + p(d, g=rc \,|\, g=1) \\
&= \sum_a p(d \,|\, g=ra) \times p(g=ra \,|\, g=1)
\end{aligned}$$

Here we do not favour any heterozygous genotype, and all have likelihood 1/3. This may be changed to reflect empirical or dbSNP data

$$\begin{aligned}
L(g=1) &= 1/3 \sum_a p(d \,|\, g=ra) \\
&= 1/3 \sum_a p(d \,|\, g=ra, \text{ado}) \times p(\text{ado}) + p(d \,|\, g=ra, \text{no ado})(1-p(\text{ado})) \\
p(d|g=ra, ado) &= p(d|g=ra, dropr) * p(dropr) + p(d|g=ra, dropa) * p(dropa)
\end{aligned}$$

Here we assume either allele is equally likely to be dropped in an ado event and $p(\text{drop r}) = p(\text{drop a}) = 0.5$. This is unlikely to change.

$$\begin{aligned}
p(d \,|\, g=ra, \text{ado}, \text{drop a}) &= \prod_i p(d_i \,|\, g=rr) \\
p(d \,|\, g=ra, \text{ado}, \text{drop r}) &= \prod_i p(d_i \,|\, g=aa)
\end{aligned}$$

2 Cell-locus Posterior probabilities

Using Bayes' rule:

$$\begin{aligned}
p(g=k \,|\, d) &= \frac{p(d \,|\, g=k)\, p(g=k)}{p(d)} \\
&= \frac{p(d \,|\, g=k)\, \binom{2}{k} f_1^k\, (1-f_1)^{2-k}}{p(d)}
\end{aligned}$$

where f_1 is the alternate allele frequency at that site. This implies HWE, which may or may not be a valid assumption. Since $p(d)$ is the same for all values of k at a cell-locus, we do not need to find it and can simply normalise.

2.1 Priors

Above is the current implementation. The alernate allele frequency may be estimated by EM at each site. Other options exist such as:

Marginalizing by site allele count

As done by Zafar et al., first probabilities for the number of alternate alleles l at the site are calculated using dynamic programing.

$$\begin{aligned}
p(g=k) &= \sum_{l'=k}^{2m-2+k} p(g=k \,|\, l=l')\, p(l=l') \\
&=
\end{aligned}$$

Site frequency spectrum

Phylogeny aware prior

As done by Singer et al.. Similar to Zafar et al. except we cosider the number of affected cells, K rather than the number of affected alleles. This prior includes the probability of a given site containing a mutation (λ) as well as a distribution of the number of cells affected. For $P(K=0)$ is simply $1-\lambda$. for $K \neq 0$:

$$p(K=k) = \lambda \frac{\binom{m}{k}^2}{(2k-1)\binom{2m}{2k}}$$

3 Probabilistic Hamming distance

The Hamming distance between two sequences s, p both length n is given by

$$\sum_i^n (1 - \delta_{s_i p_i})$$

where δ_{ab} is the Kronecker delta. Since we have a probabilistic tree, we use a similar metric but weighted by the posterior probabilities of the genotypes at each locus:

$$\sum_i^n \sum_{(a,b)\in\{0,1\}\times\{0,1\}} (1-\delta_{ab}) p(s_i p_i = ab)$$

Since $(1-\delta_{ab})$ vanishes when $s_i = p_i$, the distance reduces to

$$\sum_i^n p(s_i = 0\,,\, p_i = 1) + p(s_i = 1\,,\, p_i = 0)$$

If we assume independce (TODO) we have

$$D_{s,p} = \sum_i^n p(s_i = 0) p(p_i = 1) + p(s_i = 1) p(p_i = 0)$$