# 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

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

$$p(d|g=ra, ado) = p(d|g=ra, dropr) * p(dropr) + p(d|g=ra, dropa) * p(dropa)$$

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.

$$p(d \,|\, g=ra, \text{ado}, \text{drop a}) = \prod_i p(d_i \,|\, g=rr)$$

$$p(d \,|\, g=ra, \text{ado}, \text{drop r}) = \prod_i p(d_i \,|\, g=aa)$$

# 2 Cell-locus Posterior probabilities

Using Bayes' rule:

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

where k is the mutation rate, a learnable parameter. Note: this parameter may be overestimated if the algorithm finds more mutations, increases the rate prior, and so finds more mutations. There may be no reason for this to converge.

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.