To TEST improve algorithmic efficiency we wish only to consider sites with a non-trivial posterior probability of containing a somatic mutation. Furthermore it has been shown that combining low coverage sequencing data across samples at a locus can decrease false positive rates [**?**]. We therefore must reject loci where the posterior probability of mutation is low. For a given locus i:

$$P(SNV_i \mid D_i) = 1 - P\left(\sum_{j=1}^{m} g_{ij} = 0 \mid D_i\right) = 1 - P(\sigma = 0 \mid D_i) \tag{1}$$

Using Bayes' formula:

$$P\left(\sigma = 0 \mid D_i\right) = \frac{P(D_i \mid \sigma = 0)P(\sigma = 0)}{\sum_{\sigma'=0}^{2m}[P(D_i \mid \sigma = \sigma')P(\sigma')]} \tag{2}$$

The value of $P(D_i \mid \sigma = 0)$ is simply the product of the cell likelihoods of homozygous reference calculated above. The priors $P(\sigma)$ are those determined by Equation (**??**). To compute the denominator, howevever, we must compute the likelihood for each alternate allele count across a locus. There are various permutations of cell genotypes that may give rise to an alternate allele count of σ, so this is not as simple as the special case where $\sigma = 0$.

Let the phased genotypes of all m cells at a site be represented by $\vec{G} = (G_1, G_2, \ldots, G_m)$ where $G_j \in [0,1] \times [0,1]$ is the phased genotype for cell j (0 = reference, 1 = alternate). Furthermore let the unphased genotype vector be $\vec{g} = (g_1, g_2, \ldots, g_m)$ be such that $g_j = \|G_j\|_1$. Our likelihood for σ can therefore be considered

$$P(D_i \mid \sigma_i) = \sum_{\vec{G}} P(D_i \mid \vec{G}) P(\vec{G} \mid \sigma_i) \tag{3}$$

We assume that all phased genotype vectors with a total alternate allele count of σ are equally probable. Since there are $\binom{2m}{\sigma}$ different phased genotype vectors with total alternate allele count σ, then for any such $\vec{G}$:

$$P(\vec{G} \mid \sigma) = \binom{2m}{\sigma}^{-1}$$

Since we do not consider phased sequencing data, we must reproduce Equation (3) in an unphased form. To begin, we see that the likelihood $P(D_i \mid \vec{G}) = P(D_i \mid \vec{g})$ if $\vec{g}$ is the unphased vector that corresponds to $\vec{G}$, since our cell genotype likelihoods do not consider phasing. Note that there are 2^{χ} phased genotype vectors that correspond to any given unphased genotype vector $\vec{g}$, where $\chi(\vec{g})$ is the number of heterozygous cells in the vector. Using this multiplicity, we can now reproduce Equation (3) without reference to phasing.

$$P(D_i \mid \sigma_i) = \sum_{\vec{g}} \frac{2^{\chi(\vec{g})}}{\binom{2m}{\sigma_i}} P(D_i \mid \vec{g}) = \sum_{\vec{g}} \frac{2^{\chi(\vec{g})}}{\binom{2m}{\sigma_i}} \prod_{j=1}^{m} P(D_{ij} \mid g_j)$$

Let the function $\delta(\vec{g}, \sigma) = 1$ if $\|\vec{g}\| = \sigma$ otherwise it evaluates to 0. We can now write the above in a more suggestive form:

$$P(D_i \mid \sigma_i) = \binom{2m}{\sigma_i}^{-1} \sum_{g_1=0}^{2} \sum_{g_2=0}^{2} \cdots \sum_{g_m=0}^{2} \delta((g_1, \ldots g_m), \sigma_i) \left[\prod_{j=1}^{m} \binom{2}{g_j} P(D_{ij} \mid g_j)\right] \tag{4}$$

As has been done previously, we can employ a dynamic programming approach to compute these likelihoods for σ from cell genotype likelihoods [**?**, **?**, **?**]. If we let $F(k,l)$ be the subproblem objective given by

$$F(k,l) = \begin{cases} \sum_{g_1=0}^{2} \sum_{g_2=0}^{2} \cdots \sum_{g_k=0}^{2} \delta((g_1, \ldots g_k), l) \left[\prod_{j=1}^{k} \binom{2}{g_j} P(D_{ij} \mid g_j)\right] & 0 \leq l \leq 2k \\ 0 & \text{else} \end{cases} \tag{5}$$

We can consider creating a genotype vector of length k from a vector of length $k-1$ by adding one new cell with an alternate allele count of $0, 1$ or 2. Hence our recurrence relation can be given by

$$F(k,l) = F(k-1,l)P(D_{ik} \mid g_k = 0) + 2F(k-1,l-1)P(D_{ik} \mid g_k = 1) + F(k-1,l-2)P(D_{ik} \mid g_k = 2) \tag{6}$$

Note that two possible phased genotypes correspond to the heterozygous case, hence the factor of 2 in the second term. The base case where $k = 1$ corresponds to a single cell

$$F(1,0) = P(D_{i1} \mid g_1 = 0), \;\; F(1,1) = 2P(D_{i1} \mid g_1 = 1), \;\; F(1,2) = P(D_{i1} \mid g_1 = 2)$$

The values for $F(k,l)$ are memoized in an array and the likelihood given in Equation 4 can be given by

$$P(D_i \mid \sigma_i) = \frac{F(m, \sigma_i)}{\binom{2m}{\sigma_i}} \tag{7}$$

In this way we can determine the likelihood of all $0 \leq \sigma \leq 2m$ which when the priors $P(\sigma)$ compose the sum in Equation (2).

Sites which have a posterior probability of being variant greater than 0.5(???) will be called as variant.