The most accurate phylogenetic structure of the sampled tumour cells could be found by searching through the entire tree space and finding a tree that maximizes likelihood or posterior probability. If s mutant sites are called in the previous step there are $(2m-3)!!(2m-1)^s$ trees in the search space making this approach infeasible, leading a previous phylogeny aware approach to adopt a more efficient Markov chain Monte Carlo (MCMC) algorithm. This more efficient approach results in an overall asymptotic complexity of $O(nm^3 \log(m))$ [**?**]. The Monovar algorithm has an overall asymptotic complexity of $O(nm^3)$.

We use a simple neighbour-joining algorithm to infer a cell phylogeny based on the sites called as variant candidates. While this approach still has an asymptotic complexity of $O(nm^3)$ we expect it will yeild results faster even than Monovar on real and simulated data. While Monovar determines cell genotype posteriors by simultaneously considering the cell in question and the probability of all other cells having allele count $\sigma - g$, for efficiency we simply use a binomial model:

$$P(g_{ij} \mid D_i) = \sum_{\sigma=1}^{2m} \binom{2}{g} \left(\frac{\sigma}{2m}\right)^g \left(1 - \frac{\sigma}{2m}\right)^{2-g} P(\sigma \mid D_i)$$

where $P(\sigma \mid D_i)$ can be determined using Bayes' formula using the memoized values computed in the dynamic programming algorithm used to call candidate loci. Next we define a pairwise value $\overline{p}$, the expected frequency with which nucleotides differ between two cells a and b:

$$\overline{p} = \frac{1}{2n} \sum_{i=1}^{n} [P(g_{ia}=0)P(g_{ib}=1) + 2P(g_{ia}=0)P(g_{ib}=2) + P(g_{ia}=1)P(g_{ib}=0) + P(g_{ia}=1)P(g_{ib}=2) + 2P(g_{ia}=2)P(g \tag{1}$$

Note we assume that if two cells are heterozygous at a locus they have the same phased genotype.