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 candidate 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. This approach has an asymptotic complexity of $O(nm^2 + m^3)$, and so especially for data sets with many cells we expect it will yeild results faster even than Monovar. Monovar determines cell genotype posteriors by simultaneously considering the cell in question and the probability of all other cells having allele count $\sigma - g$ and pooling data across all cells. First, we calculate alternate allele priors for a site using read data from all the cells, where $P(\sigma) = P(\sigma \mid D_i)$ can be determined using Bayes' formula using the memoized values computed for Equation **??**. After this step, we make the simplification that $P(D_i \mid g_{ij}) = P(d_{ij} \mid g_{ij})$. Therefore using Bayes' formula and marginalizing on σ:

$$P(g_{ij} \mid D_i) = P(g_{ij} \mid d_{ij}) = \sum_{\sigma=0}^{2m} \left[P(\sigma) \frac{P(d_{ij} \mid g_{ij}) P(g_{ij} \mid \sigma)}{\sum P(d_{ij} \mid g) P(g \mid \sigma)} \right]$$

where $P(d_{ij} \mid g)$ are simple cell genotype likelihoods. For the conditional prior on cell likelihoods, we use a simple binomial approximation that can be pre computed for all sites

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

Next we define a pairwise value $\overline{p}$, the expected frequency with which nucleotides differ between two cells a and b:

$$\overline{p} = \frac{1}{2n} \sum_{i=1}^{n} \left[\sum_{|g_{ia} - g_{ib}| = 2} 2P(g_{ia})P(g_{ib}) + \sum_{|g_{ia} - g_{ib}| = 1} P(g_{ia})P(g_{ib}) \right]$$

Note we assume that if two cells have the same alternate allele count at a locus they have the same phased genotype at that locus. We then compute a distance inspired by the Jukes-Cantor distance:

$$d = -\frac{3}{4} \log \left(1 - \frac{4}{3} \overline{p} \right) \tag{1}$$

Included with the biological cells is a false cell with reference genotype which is included as an outgroup to root the tree. After computing d for all pairs of cells, we implement a simple neighbor-joining algorithm based thereon [?].