1 Introduction

2 Methods

2.1 Preprocessing and parallelism

2.2 Cell genotype likelihoods

2.3 Mutated site priors

2.4 Welltype site priors

2.5 Variant candidate site calling

2.6 Building a cell phylogeny

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 $O(nm^3 \log(m))$ [1]. 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 for Equation **??**. 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:

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

After computing d for all pairs of cells, we implement a simple neighbor-joining algorithm based thereon [2].

2.7 Parameter reestimation and second cell phylogeny

is this necessary? Overall mutation rate, $P(H_t \mid H)$, $P(H)$, λ, $P(SNP_T \mid SNP)$, ...
To some extent can be derived from papers like 21 breasts and metastatic. Can be reestimated?
do we just reestimate parameters or update priors on cell loci? updating priors could be fun, but maybe wait til completed to see if it brings additional benefit.

2.8 Mutation tree inference

Pairwise test from cell tree. Maximum parsimony inspired? Minimal way to create perfect phylogeny from cell phylogeny?

2.9 Genotyping single cells

Use probabilistic mutation tree as a prior, DFS

2.10 Additional computational methods

Stirlings approximation for $T(a, b)$. Pre computation /memoization of priors where possible. Log space.

3 Results

4 Discussion

Two methods for leaves affected: power law and uniform branches. Uniform branches (sciphi) used for sampled vs power law for considering whole tumour. Cannot use power law for sample? Why? Why not?

References

[1] Jochen Singer, Jack Kuipers, Katharina Jahn, and Niko Beerenwinkel. Single-cell mutation identification via phylogenetic inference. *Nature communications*, 9(1):5144, 2018.

[2] Richard Durbin, Sean R Eddy, Anders Krogh, and Graeme Mitchison. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press, 1998.

A Clonal mutations and aneuploidy

P(clonal): average of counts in those papiers
P(H) = 0.9 because according to Gao 90% of all tumours are aneuploid hence $(1-0.91)^{24} \approx 0.1$ assuming 24 chromosomes are independent. This assumes all aneuploid are haplod... maybe /2? High $P(H)$ could lead to FP errors... Or would it? remember we want to find homozygous variants.

B Parameter tuning

Grid search, geometric ?? Do we do this or use empirical values?